ANÁLISE NUMÉRICA

MICHEL P. J. CARPENTIER

Fevereiro 1993

Departamento de Matemática • Instituto Superior Técnico • U. T. L.

Índice

Prefácio

Com estes apontamentos de Análise Numérica pretendemos apresentar um elemento básico para o estudo das disciplinas de Análise Numérica que são actualmente lecionadas no IST.

Na resolução de um problema em engenharia e em outras ciências é geralmente possível distinguir três fases: (1) a formulação matemática do problema, (2) a escolha de um método ou combinações de métodos analíticos e numéricos para resolver o problema formulado, e (3) o cálculo numérico da solução. Basicamente um método numérico é um conjunto ordenado de operações aritméticas e lógicas, fundamentado em teoremas da análise matemática, que conduz à solução de um problema matemático. A um método numérico está pois associado um algoritmo. A construção de métodos numéricos e a escolha apropriada destes métodos para a resolução de um determinado problema constitui o campo da análise numérica.

O aparecimento dos computadores veio possibilitar cálculos numéricos até então humanamente impossíveis, o que teve como reflexo o desenvolvimento de novos métodos numéricos e a adaptação dos já existentes à nova realidade. Desde então a análise numérica desenvolveu-se como ciência própria contribuindo para a resolução dos mais variados problemas. Baseada na análise matemática clássica, a análise numérica desenvolveu por sua vez uma teoria própria: a análise de erros. Um método numérico deve ser acompanhado de um estudo sobre majorações de erros, convergência, escolha do passo, estabilidade, etc.; e para cada problema deve-se saber algo acerca do seu bom ou mau condicionamento.

Dado que a escolha de um método numérico se situa entre a formulação matemática do problema e a concretização dos cálculos, geralmente efectuados com uma máquina, a análise numérica requere por um lado conhecimentos teóricos sobre o problema a resolver e por outro experiência computacional. Para escolher dentro de vários métodos numéricos o mais indicado à resolução de um determinado problema devemos saber como estes se deduzem e por conseguinte os seus domínios de aplicação e suas limitações; como já referimos devemos fazer as respectivas análises de erros e devemos, em certos casos, estimar o número de operações a efectuar em cada método. Uma vez escolhido um método numérico devemos construir um algoritmo associado a este. Este algoritmo deverá conter os principais passos do método e permitir ao programador traduzi-lo num programa inteligível para o computador. Na hipótese de vários métodos numéricos poderem conduzir à resolução do mesmo problema com a precisão exigida pode não ser indiferente a escolha de um deles. Critérios computacionais tais como maior rapidez de execução, menor ocupação da memória e menor complexidade computacional podem ser decisivos na escolha.

Vemos assim que, hoje em dia, um método numérico é indissociável da teoria

que conduziu à sua criação e da implementação da seu algoritmo no computador. Dado que a análise numérica abarca campos tão variados não é de admirar que este facto tenha reflexos na estrutura de disciplinas introdutórias nesta área.

Com estas disciplinas pretendemos dominar, quer do ponto de vista teórico quer do ponto de vista prático, a maioria dos métodos numéricos mais utilizados nas aplicações. Isto implica, como já se deixou entender atrás, que a apresentação dos métodos deverá ser acompanhada dos teoremas que os suportam e ilustrada frequentemente por algoritmos de modo a permitir uma melhor compreensão dos métodos e uma mais fácil passagem para a fase de computação. Ao apresentar vários métodos para resolver o mesmo problema matemático deveremos comparar os seus campos de aplicação e vantagens e inconvenientes de cada um. Os métodos deverão ser testados oportunadamente no computador utilizando programas feitos pelos utilizadores e subrotinas pertencentes a bibliotecas instaladas no computador (NAG, CERN, SSP, LINPACK, MINPACK, EISPACK).

Nestes apontamentos são abordados a maioria dos temas que, na literatura, habitualmente são incluídos num curso introdutório à análise numérica. No entanto, dada a vastidão da análise numérica não se incluiu muitos temas, tais como valores e vectores próprios, optimização, equações diferenciais ordinárias com condições de fronteira, e equações diferenciais parciais, sob pena de ter de tratar superficialmente os assuntos e também porque alguns dos temas requerem conhecimentos ainda não adquiridos pelos alunos.

Entenda-se que para os assuntos escolhidos bastarão conhecimentos de análise matemática elementar, álgebra linear e prática de programação. Dado a inclusão nesta disciplina de métodos de resolução numérica de equações diferenciais ordinárias, seria conveniente um conhecimento prévio da teoria destas equações. Visto que, com o actual currículo, os alunos só mais tarde é que adquirem este conhecimento, foi necessário incluir alguns conceitos básicos de equações diferenciais, nomeadamente a existência e unicidade das suas soluções. Penso que a falta de um conhecimento mais aprofundado das equações diferenciais, que poderá ocasionar algumas dificuldades, é largamente compensada pela supressão de uma lacuna (a resolução numérica de equações diferenciais) que se fazia sentir na preparação dos alunos.

A matéria distribui-se por sete capítulos. A escolha da teoria dos erros como 1 capítulo, onde se inclui a aritmética do computador, justifica-se, dado que os métodos numéricos são para implementar no computador e que os resultados obtidos por esses métodos vêm afectados de erros que de alguma forma é necessário controlar. Apesar de algumas noções de interpolação dadas no capítulo 4 serem utilizadas em certos métodos de resolução de equações não lineares no capítulo 2, optou-se por esta ordem por várias razões. Em primeiro lugar, de todos os métodos numéricos os de resolução de uma equação não linear são, sem dúvida, dos mais atraentes para os alunos, justificando assim a sua localização no início do programa. Leva os alunos a familiarizarem-se desde o início com métodos iterativos e noções relacionadas com estes, tais como: convergência, critério de paragem, precisão dos resultados, eficiência computacional, comparação de métodos, etc. Os métodos iterativos mais elementares são de fácil implementação no computador, e portanto, do ponto de vista pedagógico, é vantajoso ensaiá-los no computador o mais cedo possível. Pelo contrário, começando pela interpolação seria mais difícil conseguir de imediato ensaios no computador, dado que a aproximação de funções requer previamente uma base teórica relativamente aprofundada.

Depois desta escolha, a ordenação dos restantes capítulos torna-se relativamente simples. Assim, depois do capítulo 2, é lógico estudar a resolução de sistemas de equações, primeiro os lineares e depois os não lineares, que é o objectivo do capítulo 3. Os capítulos 4 e 5 são dedicados à teoria da aproximação; a interpolação polinomial no capítulo 4 e o ajustamento de funções utilizando o critério dos mínimos quadrados no capítulo 5. Os capítulos 6 e 7 que tratam repectivamente do cálculo de integrais e integração de equações diferenciais baseiam-se fundamentalmente na interpolação polinomial, fazendo também uso de assuntos dos restantes capítulos.

Capítulo 1

Teoria dos erros

Neste capítulo examinaremos o efeito dos erros de dados e de arredondamento no resultado de um cálculo. Os erros de truncatura ou de método serão tratados em pormenor nos restantes capítulos por altura do estudo de cada método.

Este capítulo destina-se principalmente a fazer sentir a limitação do computador e por conseguinte a necessidade de escolher algoritmos numericamente estáveis. Desta forma, e sendo os computadores o principal meio de cálculo em análise numérica, é consequentemente útil perceber como eles operam. Nas secções seguintes vamos ver como os números são representados no computador e estudar os erros provenientes desta representação.

1.1 Representação dos números no computador

Raramente a base usada no computador é decimal. A maioria dos computadores utiliza o sistema numérico binário ou algumas das suas variantes, tais como a base 8 ou a base 16. Como exemplo da representação binária e da sua conversão a base decimal, consideremos

$$(1101.101)_2 = 2^3 + 2^2 + 2^0 + 2^{-1} + 2^{-3} = (13.625)_{10}$$

O primeiro número é escrito em base binária e o último em decimal.

A maioria dos computadores possuem uma representação para os números inteiros e outra para os números reais. Actualmente, a representação dos reais é feita, quase exclusivamente, sob a forma de vírgula flutuante, que constitui uma adaptação da notação científica dos números. A representação em vírgula flutuante, como veremos mais tarde, permite representar números reais com ordens de grandezas muito diferentes.

A título de exemplo da representação dos números em notação científica consideremos novamente o número 13.625 que se pode exprimir sob a forma

$$0.13625 \times 10^2$$

ou em base binária

$$(0.1101101)_2 \times 2^4$$

Com generalidade, seja β a base do sistema de números utilizada (geralmente $\beta = 2, 8, 10$ ou 16), então um número x representado em notação científica tem a forma

$$x = \sigma \times m \times \beta^t \qquad \sigma = \pm 1$$

em que m é um número real designado por *mantissa* e t é um número inteiro designado por *expoente*. Fixada a base β, esta representação não é única pois, por exemplo, $x = 1$ pode ser escrito na base 2 de várias maneiras,

$$x = 1 \times 2^0 = 0.1 \times 2^1 = 0.01 \times 2^{10} = 10 \times 2^{-1} = 100 \times 2^{-10}$$

(os números correspondentes s mantissas e aos expoentes estão escritos em base 2). Para resolver esta ambiguidade é usual adoptar a seguinte convenção: para um número $x \neq 0$ toma-se a mantissa m de modo a que

$$\beta^{-1} = (0.1)_\beta \leq m < 1$$

Diz-se neste caso que se usa uma representação *normalizada*. Nesta representação a mantissa de um número diferente de zero tem a seguinte forma

$$m = 0.a_1 a_2 \ldots \qquad a_1 \neq 0$$

em que os a_i são todos algarismos na base β ($0 \leq a_i \leq \beta - 1$). Assim, p.ex., o número 1 tem a seguinte forma

$$1 = (0.1)_\beta \times \beta^1$$

É claro que a notação científica, tal como acabamos de apresentar, não pode ser implementada em computador, pois para cobrir todos os números reais a mantissa e o expoente exigiriam um número infinito de algarismos. Assim a notação científica é modificada no sentido de se utilizar um número finito n de algarismos para a mantissa m e confinar o expoente t a um intervalo $[t_1, t_2]$, obtendo-se a representação em *vírgula flutuante*.

Para simplificar a exposição usaremos a notação $VF(\beta, n, t_1, t_2)$ para designar o sistema de representação em vírgula flutuante constituído por todos os números reais x da forma

$$x = \pm m \beta^t \qquad \beta^{-1} \leq m \leq 1 - \beta^{-n} \qquad t_1 \leq t \leq t_2$$

e ainda $x = 0$. Assim m tem a seguinte forma

$$m = 0.a_1 a_2 \ldots a_n \qquad a_1 \neq 0$$

Por exemplo o número

$$x = 0.13625 \times 10^2$$

é tal que $x \in VF(10, 5, -3, 3)$ mas $x \notin VF(10, 4, -3, 3)$ ou $x \notin VF(10, 5, -3, 1)$. O último caso é conhecido na literatura inglesa por "overflow", *i.e.*, o número é demasiadamente grande para ser representado naquele sistema.

Notando que $a_1 \neq 0$, vemos que a mantissa pode tomar $(\beta - 1)\beta^{n-1}$ valores diferentes e portanto o sistema $VF(\beta, n, t_1, t_2)$ é constituído pelo conjunto de

$$N = (t_2 - t_1 + 1)(\beta - 1)\beta^{n-1}$$

Figura 1.1: O sistema $VF(2, 4, -1, 2)$

números racionais positivos compreendidos entre

$$L_1 = \beta^{-1} \times \beta^{t_1}$$

e

$$L_2 = (1 - \beta^{-n}) \times \beta^{t_2}$$

e ainda pelo conjunto dos números simétricos aos anteriores e o número zero.

Exemplo 1.1 *Concretizar* $VF(2, 4, -1, 2)$.

$VF(2, 4, -1, 2)$ designa o sistema de vírgula flutuante de base $\beta = 2$, cuja mantissa m possui $n = 4$ algarismos e cujo expoente t pode variar entre $t_1 = -1$ e $t_2 = 2$. Assim a mantissa pode tomar neste sistema os seguintes valores:

$$\begin{aligned}
(0.1000)_2 &= 0.5000 \\
(0.1001)_2 &= 0.5625 \\
(0.1010)_2 &= 0.6250 \\
(0.1011)_2 &= 0.6875 \\
(0.1100)_2 &= 0.7500 \\
(0.1101)_2 &= 0.8125 \\
(0.1110)_2 &= 0.8750 \\
(0.1111)_2 &= 0.9375
\end{aligned}$$

Atendendo que β^t pode tomar os valores $2^{-1}, 2^0, 2^1, 2^2$, vemos que $N = 32$, $L_1 = 0.25$, $L_2 = 3.75$ e portanto $VF(2, 4, -1, 2)$ é constituído por 65 números que representamos na Figura 1.1. ■

No VAX, em precisão *simples* (32 bits), o sistema numérico utilizado é o

$$VF(2, 24, -(2^7 - 1), 2^7 - 1) = VF(2, 24, -127, 127)$$

Para este sistema os

$$N = (2^8 - 1) \times 2^{23} = 2\,139\,095\,040$$

números racionais positivos estão compreendidos entre

$$L_1 = 2^{-1} \times 2^{-127} \approx 0.29 \times 10^{-38}$$

e

$$L_2 = (1 - 2^{-24}) \times 2^{127} \approx 1.7 \times 10^{38}$$

Em precisão *dupla* (64 bits) o sistema numérico é o

$$VF(2, 56, -(2^7 - 1), 2^7 - 1) = VF(2, 56, -127, 127)$$

com

$$N = (2^8 - 1) \times 2^{55} \approx 0.92 \times 10^{19}$$
$$L_1 = 2^{-128} \approx 0.29 \times 10^{-38}$$
$$L_2 = (1 - 2^{-56}) \times 2^{127}) \approx 1.7 \times 10^{38}$$

ou, no caso da variante *"G_floating"* (64 bits), o

$$VF(2, 53, -(2^{10} - 1), 2^{10} - 1) = VF(2, 53, -1023, 1023)$$

com

$$N = (2^{11} - 1) \times 2^{52} \approx 0.92 \times 10^{19}$$
$$L_1 = 2^{-1024} \approx 0.56 \times 10^{-308}$$
$$L_2 = (1 - 2^{-53}) \times 2^{1023} \approx 0.90 \times 10^{308}$$

Em precisão *quádrupla* (128 bits) o sistema é o

$$VF(2, 113, -(2^{14} - 1), 2^{14} - 1) = VF(2, 113, -16383, 16383)$$

com

$$N = (2^{15} - 1) \times 2^{112} \approx 1.7 \times 10^{38}$$
$$L_1 = 2^{-16384} \approx 0.84 \times 10^{-4932}$$
$$L_2 = (1 - 2^{-113}) \times 2^{16383} \approx 0.59 \times 10^{4932}$$

1.2 Erros na representação em vírgula flutuante

Como acabamos de ver, a representação em vírgula flutuante só permite a representação exacta de alguns números reais, o que implica que em geral tenhamos que aceitar representar números reais com um certo erro. Põe-se a seguinte questão: dado um número real x qual o número em VF[1], que denotaremos por $fl(x)$, que o representa? Se $x \in VF$ então fazemos $fl(x) = x$. Se $x \notin VF$ então temos que efectuar o *arredondamento* de x. Há dois tipos de arredondamento: o arredondamento por *corte* que consiste em escolher para $fl(x)$ o número mais perto de x entre 0 e x e o arredondamento *simétrico* que consiste em escolher para $fl(x)$ o número mais perto de x.

Assim, por exemplo, no sistema $VF(10, 2, -2, 2)$ o número

$$x = 0.13625 \times 10^2$$

será representado por

$$fl(x) = 0.13 \times 10^2 \text{ ou } fl(x) = 0.14 \times 10^2$$

consoante o arredondamento for por corte ou simétrico. Com generalidade, representando x na notação científica

$$x = \sigma \times (0.a_1 a_2 \ldots a_n a_{n+1} \ldots) \times \beta^t \qquad \sigma = \pm 1 \qquad a_1 \neq 0 \tag{1.1}$$

[1]Por comodidade passamos a designar $VF(\beta, n, t_1, t_2)$ por VF sempre que não seja necessário explicitar os argumentos.

e supondo que a base β é par (normalmente β=2,8,10 ou16), temos para o arredondamento por corte

$$fl(x) = \sigma \times (0.a_1 a_2 \ldots a_n) \times \beta^t \tag{1.2}$$

e para o arredondamento simétrico

$$\begin{aligned} fl(x) &= \sigma \times (0.a_1 a_2 \ldots a_n) \times \beta^t & 0 \leq a_{n+1} < \beta/2 \\ fl(x) &= \sigma \times (0.a_1 a_2 \ldots a_n + 0.00\ldots 1) \times \beta^t & \beta/2 \leq a_{n+1} < \beta \end{aligned}$$

em que $0.00\ldots 1 = \beta^{-n}$.

Exemplo 1.2 *Achar as representações de: (1) π em $VF(10, 7, -38, 38)$, (2) $(0.4)_{10}$ em $VF(2, 4, -1, 2)$, e (3) $(0.1)_{10}$ no VAX.*

1. Como
$$\pi = 3.1415926535\ldots$$
temos que
$$fl(\pi) = 0.3141592 \times 10$$
ou
$$fl(\pi) = 0.3141593 \times 10$$
consoante o arredondamento for por corte ou simétrico.

2. Como
$$(0.4)_{10} = (0.01100110011\ldots)_2$$
temos que
$$fl((0.4)_{10}) = (0.1100)_2 \times 2^{-1} = (0.375)_{10}$$
ou
$$fl((0.4)_{10}) = (0.1101)_2 \times 2^{-1} = (0.40625)_{10}$$
consoante o arredondamento for por corte ou simétrico.

3. O VAX trabalha em $VF(2, 24, -127, 127)$, como vimos atrás, e com arredondamento simétrico. Sendo assim, como
$$(0.1)_{10} = (0.000110011\ldots)_2$$
temos que
$$\begin{aligned} fl((0.1)_{10}) &= (0.11001100110011001100110 1)_2 \times 2^{-(11)_2} \\ &\approx (0.100000001490116)_{10} \end{aligned}$$

■

Acabamos de ver que em geral um número não é representado exactamente no computador sendo afectado de um certo erro. Para ser mais explicíto vamos adoptar a seguinte nomenclatura. Sendo x o valor exacto de uma grandeza e $\tilde{x}$ um seu valor aproximado, define-se *erro* de $\tilde{x}$ (em relação a x) como sendo o valor:

$$e_x = x - \tilde{x}$$

Quando $x \neq 0$ podemos calcular e_x/x que designamos por δ_x. Podemos então escrever

$$x - \tilde{x} = x\delta_x \qquad \tilde{x} = x(1 - \delta_x) \tag{1.3}$$

Ao valor absoluto de e_x e de δ_x correspondem respectivamente o erro absoluto e o erro relativo. Assim, define-se *erro absoluto* de $\tilde{x}$ como sendo

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

e define-se *erro relativo* de $\tilde{x}$ como sendo

$$|\delta_x| = \frac{|x - \tilde{x}|}{|x|}$$

Muitas vezes, designaremos também por erro relativo a quantidade δ_x, definida por (1.3). Para além destas noções convém-nos ainda introduzir a de *algarismo significativo*. Para isso consideremos $\tilde{x}$ representado na notação científica normalizada, dada por (1.1), na base decimal.

Definição 1.1 *Diz-se que um dos algarismos a_i de $\tilde{x}$ é significativo se*

$$|e_x| \leq \frac{1}{2}10^{t-i}$$

Assim se

$$|e_x| \leq \frac{1}{2}10^{t-n}$$

o número $\tilde{x}$ tem n algarismos significativos relativamente a x.

Exemplo 1.3 *Determinar os vários tipos de erros de $\tilde{\pi} = 3.141592$ e o número de algarismos significativos relativamente a $\pi = 3.1415926535\ldots$.*

Temos que

$$\begin{aligned} e_\pi &= 0.6535 \times 10^{-6} \\ \delta_\pi &= e_\pi/\pi = 0.2080 \times 10^{-6} \end{aligned}$$

Dado que $|e_\pi| > 0.5 \times 10^{-6}$ o último algarismo de $\tilde{\pi}$ não é significativo, e portanto $\tilde{\pi}$ só possui 6 algarismos significativos. Se $\tilde{\pi}$ fosse 3.141593 já teria todos os seus algarismos significativos. ■

Podemos agora ver qual é o erro absoluto cometido num arredondamento por corte. Atendendo a (1.1) e (1.2) temos que

$$\begin{aligned} e_{ar} = x - fl(x) &= \sigma \times (0.00\ldots0a_{n+1}\ldots) \times \beta^t \\ &= \sigma \times (0.a_{n+1}\ldots) \times \beta^{t-n} \end{aligned}$$

Por conseguinte temos a seguinte majoração para o *erro absoluto de arredondamento por corte*

$$|e_{ar}| = |x - fl(x)| \leq \beta^{t-n}$$

De uma forma semelhante, obtém-se para o *erro absoluto de arredondamento simétrico* a majoração

$$|e_{ar}| = |x - fl(x)| \leq \frac{1}{2}\beta^{t-n} \tag{1.4}$$

Vemos que na base 10 o arredondamento simétrico garante que todos os n algarismos de $fl(x)$ sejam significativos relativamente a x.

As majorações apresentadas têm o inconveniente de serem função de β^t, *i.e.*, dependerem da ordem de grandeza de x. Para obviar a este inconveniente vamos obter majorações para o erro relativo. De (1.4) temos que

$$\begin{aligned} |\delta_{ar}| = \frac{|e_{ar}|}{|x|} &\leq \frac{1}{2}\beta^{t-n}\frac{1}{|x|} = \frac{1}{2}\frac{\beta^{t-n}}{(0.a_1a_2\ldots)_\beta \times \beta^t} \\ &\leq \frac{1}{2}\frac{\beta^{-n}}{(0.1)_\beta} = \frac{1}{2}\frac{\beta^{-n}}{\beta^{-1}} \end{aligned}$$

No caso do arredondamento por corte obteríamos o mesmo resultado parte o factor 1/2. Portanto o *erro relativo de arredondamento* δ_{ar} satisfaz a seguinte majoração

$$\begin{aligned} &|\delta_{ar}| = |\frac{x - fl(x)}{x}| \leq u \\ &u = \beta^{1-n} \quad \textit{arredondamento por corte} \\ &u = \frac{1}{2}\beta^{1-n} \quad \textit{arredondamento simétrico} \end{aligned} \tag{1.5}$$

Atendendo a (1.3) podemos escrever

$$e_{ar} = x - fl(x) = x\delta_{ar} \qquad fl(x) = x(1-\delta_{ar}) \qquad |\delta_{ar}| \leq u$$

O u só depende da base β, do número de algarismos n utilizados na mantissa e do tipo de arredondamento (por corte ou simétrico). Portanto é uma constante característica da máquina utilizada e designá-la-emos por *unidade de arredondamento* da máquina. Temos assim que, no VAX (arredondamento simétrico)

$$u = 2^{-24} \approx 0.596 \times 10^{-7}$$

em precisão *simples* e

$$u = 2^{-56} \approx 0.139 \times 10^{-16}$$

em precisão *dupla* ($u = 2^{-53} \approx 0.111 \times 10^{-15}$ na versão *"G_floating"*) e

$$u = 2^{-113} \approx 0.096 \times 10^{-33}$$

em precisão *quádrupla*. Pode-se portanto garantir a $fl(x)$ quase sempre 7 algarismos significativos (relativamente a x) em precisão *simples* e sempre 16 algarismos significativos em precisão *dupla* (15 na versão *"G_floating"*) e 33 em precisão *quádrupla*.

1.3 Operações aritméticas em vírgula flutuante

Vejamos sumariamente como é efectuada uma operação aritmética no computador. Sejam x e y dois números pertencentes a VF e portanto $fl(x) = x$ e $fl(y) = y$. Pretendemos calcular

$$z = \varphi(x, y) = x\omega y$$

em que ω é uma das operações aritméticas $+, -, \times, :$. Designando por $\tilde{\omega}$ a operação efectuada no computador e por $\tilde{z}$ o resultado desta operação obtemos

$$\tilde{z} = \tilde{\varphi}(x, y) = x\tilde{\omega}y$$

que geralmente não coincide com z. Na maioria dos computadores os cálculos são efectuados de forma a que $\tilde{z} = fl(z)$. Portanto, dado dois números x e y com representação exacta no computador temos que

$$\varphi(x, y) - \tilde{\varphi}(x, y) = \varphi(x, y) - fl(\varphi(x, y)) = \varphi(x, y)\delta_{ar} \qquad |\delta_{ar}| \leq u \tag{1.6}$$

Vamos exemplificar considerando um computador que funciona em base decimal com o sistema $VF(10, 7, -38, 38)$ e com arredondamento simétrico. Sejam

$$x = 99.91232 \qquad y = 0.2345271 \qquad z = x + y$$

Então

$$fl(x) = 0.9991232 \times 10^2 \qquad fl(y) = 0.2345271 \times 10^0$$

Antes de efectuar a soma temos que exprimir os dois números de forma a que os seus expoentes coincidam; assim, desnormalizamos a mantissa do número com menor expoente de forma a que este expoente se reduza ao outro. Portanto, temos que

$$y = 0.002345271 \times 10^2$$

em que a mantisssa passou a ter 9 algarismos em vez de 7; os algarismos adicionais (neste caso 7 e 1) são designados por *algarismos de guarda*. Podemos agora somar as mantissas obtendo

$$z = (0.9991232 + 0.002345271) \times 10^2 = 1.001468471 \times 10^2$$

De notar que agora a mantissa tem 10 algarismos (3 algarismos de guarda). O resultado obtido é exacto, e só agora é que se faz o arredondamento. Assim, obtemos finalmente

$$fl(z) = 0.1001468 \times 10^3$$

Se não tivessemos guardados em y os algarismos de guarda 7 e 1, excedentes na mantissa depois da desnormalização e tivessemos arredondado para

$$y = 0.0023453 \times 10^2$$

então obteríamos

$$x\tilde{+}y = 0.1001469 \times 10^3 \neq fl(x + y)$$

Vemos com este exemplo que para garantir (1.6) as operações deverão ser efectuadas utilizando algarismos de guarda.

Vejamos outro exemplo: calcular

$$z = x - y \qquad x = 0.1004567 \qquad y = 0.09952144$$

Estamos neste caso a subtrair 2 números quase iguais o que é conhecido por *cancelamento subtractivo*. Obtemos

$$z = 0.00093526$$

e portanto

$$fl(z) = 0.9352600 \times 10^{-3}$$

Vemos que no caso de cancelamento subtractivo, esta operação no computador é exacta, *i.e.*, não há erro de arredondamento. No entanto, se os números x e y estão afectados de erros provenientes de cálculos anteriores sua determinação, então o cancelamento é uma operação extremamente perigosa visto que o erro relativo do resultado pode ser muito superior aos erros associados a x e a y. Suponhamos que os valores exactos de x e y são

$$x = 0.100456683 \qquad y = 0.0995214437$$

e que pretendemos calcular

$$z = x - y = 0.9352393 \times 10^{-3}$$

Como

$$\tilde{x} = fl(x) = 0.1004567$$

e

$$\tilde{y} = fl(y) = 0.09952144$$

o valor calculado será

$$\tilde{z} = fl(\tilde{x} - \tilde{y}) = 0.9352600 \times 10^{-3}$$

Vemos assim que os 3 últimos algarismos não são significativos.

Ainda com um computador que funciona em base decimal com o sistema $VF(10, 7, -38, 38)$ e com arredondamento simétrico, calculemos

$$1 + u \qquad 1 + v$$

em que

$$u = 1/2 \times 10^{1-7} = 0.5 \times 10^{-6}$$

é a unidade de arredondamento e

$$v = 0.9u = 0.45 \times 10^{-6}$$

No 1 caso temos

$$\begin{aligned} fl(1+u) &= fl((0.1 + 0.00000005) \times 10^1) = fl(0.10000005 \times 10^1) = \\ &= 0.1000001 \times 10^1 = 1.000001 \end{aligned}$$

No 2 caso temos

$$\begin{aligned} fl(1+v) &= fl((0.1 + 0.000000045) \times 10^1) = fl(0.100000045 \times 10^1) = \\ &= 0.1000000 \times 10^1 = 1 \end{aligned}$$

Deste exemplo podemos facilmente ver que em geral a unidade de arredondamento é o número u para o qual

$$\begin{aligned} fl(1+x) &\neq 1 \quad se \;\; x \geq u \\ fl(1+x) &= 1 \quad se \;\; 0 \leq x < u \end{aligned}$$

Por outras palavras, u é o menor número positivo que adicionado a 1 conduz no computador a um resultado diferente de 1. Baseado neste facto apresentamos um algoritmo que permite determinar o u de um computador que trabalha em base binária.

Algoritmo 1.1 *(Determinação da unidade de arredondamento u)*

INICIALIZAÇÃO:
$x = 1$

CICLO: PARA $k = 0, 1, \ldots$ FAZER

fazer $x = x/2,\ y = x + 1$

SE $y = 1$ ENTÃO fazer $u = 2x$, SAÍDA

FIM DO CICLO

1.4 Propagação dos erros

1.4.1 Propagação dos erros em operações elementares

Admitimos que, ao efectuar uma operação aritmética ω no computador, a relação (1.6) é válida desde que os operandos x e y sejam representados exactamente no computador. Em geral quando se pretende determinar

$$z = \varphi(x, y) = x\omega y$$

no computador, o que é de facto calculado é

$$\tilde{z} = \tilde{\varphi}(\tilde{x}, \tilde{y})$$

em que $\tilde{\varphi}$ designa a função φ calculada no computador, como vimos anteriormente, e $\tilde{x} = fl(\tilde{x})$ e $\tilde{y} = fl(\tilde{y})$ são valores aproximados de x e y. Isto pode provir de várias situações: (1) x não é representado exactamente no computador, (2) x é o valor de uma grandeza física não sendo possível a sua medição exacta, (3) x é o resultado de cálculos anteriores onde os erros se vão propagando. Temos assim que o erro total cometido é

$$e_z = \varphi(x, y) - \tilde{\varphi}(\tilde{x}, \tilde{y}) = [\varphi(x, y) - \varphi(\tilde{x}, \tilde{y})] + [\varphi(\tilde{x}, \tilde{y}) - \tilde{\varphi}(\tilde{x}, \tilde{y})] \tag{1.7}$$

O 2 termo entre parênteses rectos é o erro de arredondamento que temos vindo a estudar e calcula-se usando (1.6), ficando

$$\varphi(\tilde{x}, \tilde{y}) - \tilde{\varphi}(\tilde{x}, \tilde{y}) = \varphi(\tilde{x}, \tilde{y})\delta_{ar} \approx \varphi(x, y)\delta_{ar} \qquad |\delta_{ar}| \leq u \tag{1.8}$$

em que a aproximação $\varphi(\tilde{x}, \tilde{y}) \approx \varphi(x, y)$ justifica-se por desprezarmos no cálculo de erros termos de ordem superior (supõe-se erros relativos pequenos). O 1 termo entre parênteses rectos é o *erro propagado*, *i.e.*, a contribuição dos erros dos operandos para o erro do resultado da operação.

Vejamos como se pode determinar o erro propagado. Recorrendo ao teorema de Lagrange podemos escrever

$$\begin{aligned} e_\varphi &= \varphi(x, y) - \varphi(\tilde{x}, \tilde{y}) \\ &= [\varphi(x, y) - \varphi(\tilde{x}, y)] + [\varphi(\tilde{x}, y) - \varphi(\tilde{x}, \tilde{y})] \\ &= \varphi'_x(\overline{x}, y)(x - \tilde{x}) + \varphi'_y(\tilde{x}, \overline{y})(y - \tilde{y}) \\ &\approx \varphi'_x(x, y)(x - \tilde{x}) + \varphi'_y(x, y)(y - \tilde{y}) \end{aligned}$$

em que $\overline{x} \in (x, \tilde{x})$, $\overline{y} \in (y, \tilde{y})$, φ'_x e φ'_y são as derivadas parciais de φ em ordem a x e a y, respectivamente, e as aproximações $\varphi'_x(\overline{x}, y) \approx \varphi'_x(x, y)$ e $\varphi'_y(\tilde{x}, \overline{y}) \approx \varphi'_y(x, y)$ justificam-se por desprezarmos novamente termos de ordem superior. Podemos então escrever

$$e_\varphi = \varphi'_x(x, y)e_x + \varphi'_y(x, y)e_y \tag{1.9}$$

Este resultado podia obter-se recorrendo ao desenvolvimento em série de Taylor em torno de (x, y) e desprezando termos de ordem superior.

Podemos exprimir a relação anterior utilizando erros relativos.

$$\delta_\varphi = \frac{e_\varphi}{\varphi(x, y)} = \frac{x\varphi'_x(x, y)}{\varphi(x, y)}.\frac{e_x}{x} + \frac{y\varphi'_y(x, y)}{\varphi(x, y)}.\frac{e_y}{y}$$

ou ainda

$$\delta_\varphi = p_x\delta_x + p_y\delta_y \tag{1.10}$$

onde

$$p_x = \frac{x\varphi'_x(x, y)}{\varphi(x, y)} \qquad p_y = \frac{y\varphi'_y(x, y)}{\varphi(x, y)} \tag{1.11}$$

Os pesos p_x e p_y são designados na literatura por *números de condição* e relacionam o erro relativo δ_φ da função φ respectivamente com o erro relativo δ_x da variável x e o erro relativo δ_y da variável y.

Para obtermos o erro propagado em cada uma das operações aritméticas bastará determinar os respectivos números de condição. Assim teremos sucessivamente.

Multiplicação: $\varphi(x, y) = xy$. Temos que

$$\varphi'_x(x, y) = y \qquad \varphi'_y(x, y) = x$$

e portanto

$$p_x = 1 \qquad p_y = 1$$

Logo

$$\delta_\varphi = \delta_{xy} = \delta_x + \delta_y$$

Divisão: $\varphi(x, y) = x/y$. Temos que

$$\varphi'_x(x, y) = 1/y \qquad \varphi'_y(x, y) = -x/y^2$$

e portanto

$$p_x = 1 \qquad p_y = -1$$

Logo

$$\delta_\varphi = \delta_{x/y} = \delta_x - \delta_y$$

Adição e subtracção: $\varphi(x, y) = x \pm y$. Temos que

$$\varphi'_x(x, y) = 1 \qquad \varphi'_y(x, y) = \pm 1$$

e portanto

$$p_x = x/(x \pm y) \qquad p_y = \pm y/(x \pm y)$$

Logo

$$\delta_\varphi = \delta_{x\pm y} = \frac{x}{x \pm y}\delta_x \pm \frac{y}{x \pm y}\delta_y$$

Nota-se que se utilizarmos (1.9) obtem-se a seguinte relação muito simples

$$e_{x\pm y} = e_x \pm e_y$$

que, neste caso, é exacta mesmo quando não se desprezam termos de ordem superior.

Estamos agora em condições de obtermos o erro total cometido numa operação algébrica $\varphi(x, y)$. Com efeito, basta atender a (1.7), (1.8) e (1.10) e utilizarmos erros relativos para obtermos o erro relativo total

$$\delta_z = \frac{\varphi(x,y) - \tilde{\varphi}(\tilde{x},\tilde{y})}{\varphi(x,y)} = \delta_\varphi + \delta_{ar} = p_x\delta_x + p_y\delta_y + \delta_{ar} \qquad |\delta_{ar}| \leq u \tag{1.12}$$

Acabamos de ver que para as 4 operações aritméticas os números de condição são:

$$\begin{aligned}
\varphi(x,y) &= xy && \Longrightarrow \quad p_x = 1 && p_y = 1 \\
\varphi(x,y) &= x/y && \Longrightarrow \quad p_x = 1 && p_y = -1 \\
\varphi(x,y) &= x \pm y && \Longrightarrow \quad p_x = x/(x \pm y) && p_y = \pm y/(x \pm y)
\end{aligned}$$

Vemos que na multiplicação e na divisão os erros relativos das parcelas não são aumentados. Se x e y ou x e $-y$ tiverem o mesmo sinal respectivamente para a adição ou subtracção, então os pesos são em valores absolutos inferiores a 1. Nestes casos os erros não se propagam fortemente no resultado, sendo as operações numericamente estáveis. Se x e y ou x e $-y$ tiverem sinais contrários respectivamente para a adição ou subtracção, então os pesos são em valores absolutos superiores a 1 podendo no caso de $x \approx -y$, na adição ou $x \approx y$, na subtracção, atingir valores muito elevados; é o caso do cancelamento subtractivo, já referido atrás. Neste último caso os erros são muito ampliados sendo a operação numericamente instável.

Exemplo 1.4 *Determinar o número de algarismos significativos que se pode garantir a xy, x/y, $x+y$ e $x-y$ sendo $\tilde{x} = 1010$ e $\tilde{y} = 1000$, sabendo que todos os algarismos (em número de 4) de $\tilde{x}$ e $\tilde{y}$ são significativos. Desprezar os erros de arredondamento.*

Os erros absolutos $|e_x|$ e $|e_y|$ são inferiores a 0.5. Portanto os erros relativos têm as majorações[2]

$$|\delta_x| = |e_x|/|x| \approx |e_x|/|\tilde{x}| \leq 0.5/1010 = 0.495 \times 10^{-3}$$

$$|\delta_y| = |e_y|/|y| \approx |e_y|/|\tilde{y}| \leq 0.5/1000 = 0.5 \times 10^{-3}$$

Assim temos que

$$|\delta_{xy}| = |\delta_x + \delta_y| \leq |\delta_x| + |\delta_y| \leq 0.995 \times 10^{-3}$$

e

$$|\delta_{x/y}| = |\delta_x - \delta_y| \leq |\delta_x| + |\delta_y| \leq 0.995 \times 10^{-3}$$

Como

$$\tilde{x}\tilde{y} = 1.010 \times 10^6$$

[2]Como se desconhece os valores exactos de x e y utiliza-se os seus valores aproximados $\tilde{x}$ e $\tilde{y}$. Este procedimento é habitual e equivale a desprezar no cálculo dos erros termos de ordens superiores.

e

$$\tilde{x}/\tilde{y} = 1.010$$

temos

$$|e_{xy}| = |\delta_{xy}||xy| \approx |\delta_{xy}||\tilde{x}\tilde{y}| \leq 0.995 \times 10^{-3} \times 1.010 \times 10^{6} = 1.005 \times 10^{3}$$

e

$$|e_{x/y}| = |\delta_{x/y}||x/y| \approx |\delta_{x/y}||\tilde{x}/\tilde{y}| \leq 0.995 \times 10^{-3} \times 1.010 = 1.005 \times 10^{-3}$$

Portanto $\tilde{x}\tilde{y}$ e $\tilde{x}/\tilde{y}$ têm pelo menos 3 algarismos significativos. Por outro lado

$$|e_{x\pm y}| = |e_x \pm e_y| \leq |e_x| + |e_y| \leq 0.5 + 0.5 = 1$$

e como

$$\tilde{x} + \tilde{y} = 2010$$

e

$$\tilde{x} - \tilde{y} = 10$$

concluimos que $\tilde{x} + \tilde{y}$ tem também 3 algarismos significativos enquanto que $\tilde{x} - \tilde{y}$ tem apenas 1 algarismo significativo. ■

Voltemos ao problema do cálculo aproximado de $\varphi(x, y)$ através de $\tilde{\varphi}(\tilde{x}, \tilde{y})$, com $\tilde{x} = fl(\tilde{x})$ e $\tilde{y} = fl(\tilde{y})$. Até agora temos considerado φ como representando uma das 4 operações algébricas, para as quais admitimos que se verifica (1.6) ou (1.8), *i.e.*,

$$\tilde{\varphi} = fl(\varphi) \tag{1.13}$$

Neste caso o termo $\varphi(\tilde{x}, \tilde{y}) - \tilde{\varphi}(\tilde{x}, \tilde{y})$ em (1.7) deve-se a um único arredondamento efectuado pelo computador. A uma operação que satisfaz (1.13) designamos por *operação elementar*. As funções tais como $\varphi(x) = \sqrt{x}$, $\sin x$, $\cos x$, $\exp x$, $\ln x$, etc serão também consideradas como operações elementares, se bem que (1.13) não seja rigorosamente satisfeita para estas funções.

Para podermos utilizar (1.7) no caso destas funções necessitamos de determinar o erro propagado $\varphi(x) - \varphi(\tilde{x})$. Como $\varphi(x)$ só depende de x as relações (1.9), (1.10) e (1.11) obtidas para $\varphi(x, y)$ reduzem-se respectivamente a

$$\begin{aligned} e_\varphi &= \varphi(x) - \varphi(\tilde{x}) = \varphi'(x)e_x \\ \delta_\varphi &= p_x\delta_x \\ p_x &= \frac{x\varphi'(x)}{\varphi(x)} \end{aligned} \tag{1.14}$$

em que $\varphi'(x)$ é uma derivada total e p_x é o número de condição que relaciona o erro relativo δ_φ da função φ com o erro relativo δ_x da variável x. O erro total será dado por

$$\delta_z = \frac{\varphi(x) - \tilde{\varphi}(\tilde{x})}{\varphi(x)} = \delta_\varphi + \delta_{ar} = p_x\delta_x + \delta_{ar} \qquad |\delta_{ar}| \leq u \tag{1.15}$$

que é um caso particular de (1.12).

A seguir apresentamos o número de condição para algumas das operações consideradas elementares

$$
\begin{aligned}
\varphi(x) = \sqrt{x} \quad &\Longrightarrow \quad p_x = x\frac{\frac{1}{2\sqrt{x}}}{\sqrt{x}} = \frac{1}{2} \\
\varphi(x) = \sin x \quad &\Longrightarrow \quad p_x = x\frac{\cos x}{\sin x} = x\cot x \\
\varphi(x) = \cos x \quad &\Longrightarrow \quad p_x = x\frac{-\sin x}{\cos x} = -x\tan x \\
\varphi(x) = \exp x \quad &\Longrightarrow \quad p_x = x\frac{\exp x}{\exp x} = x \\
\varphi(x) = \ln x \quad &\Longrightarrow \quad p_x = x\frac{\frac{1}{x}}{\ln x} = \frac{1}{\ln x}
\end{aligned}
\tag{1.16}
$$

Vemos que $\sqrt{\ }$ reduz sempre o erro relativo; é pois uma operação bem condicionada. O sin é mal condicionado para $x \approx k\pi,\ k \neq 0$ (com k inteiro) e $|x| \gg 1$, visto que para estes valores de x o número de condição p_x atinge em módulo valores elevados. De igual modo o cos é mal condicionado para $x \approx (2k+1)\pi/2$ e $|x| \gg 1$, o exp para $|x| \gg 1$ e o ln para $x \approx 1$.

Exemplo 1.5 *Determinar um majorante para o erro absoluto de* $\tilde{z} = \tilde{\cos}\tilde{x}$ *quando* $\tilde{x} = 1.5$ *com um erro relativo de 0.001 e o cálculo de* cos *é efectuado por uma máquina usando o sistema* $VF(10, 4, -10, 10)$ *com arredondamento por corte.*

O valor calculado para o coseno é

$$\tilde{z} = \tilde{\cos}\tilde{x} = fl(0.07073720) = 0.07073$$

Vamos obter sucessivamente majorações para os erros relativo e absoluto de $\tilde{z}$. De (1.14) e (1.16) temos

$$\delta_{\cos} = -x\tan x\ \delta_x \approx -\tilde{x}\tan\tilde{x}\ \delta_x$$

e de (1.15) temos

$$|\delta_z| = |\delta_{\cos} + \delta_{ar}| \leq |\delta_{\cos}| + |\delta_{ar}|$$

Então, das expressões anteriores e de (1.5) temos a majoração

$$|\delta_z| \leq |-1.5\tan 1.5 \times 0.001| + |10^{1-4}| = 0.0212 + 0.001 = 0.0222$$

Como

$$|e_z| = |z||\delta_z| \approx |\tilde{z}||\delta_z|$$

temos finalmente

$$|e_z| \leq 0.07073 \times 0.0222 = 0.00157$$

■

1.4.2 Mau condicionamento e estabilidade

Até agora só considerámos operações elementares isoladas. Suponhamos que pretendemos calcular

$$z = f(a,b) = a^2 - b^2 = (a+b)(a-b)$$

Podemos fazê-lo utilizando sucessivas operações elementares. A sequência como é efectuada estas operações constitui um *algoritmo*. Assim para este exemplo podemos utilizar os 2 algoritmos seguintes:

$$\begin{array}{ll} \textit{Algoritmo } 1: & \textit{Algoritmo } 2: \\ z_1 = \varphi_1(a) = a \times a & z_1 = \varphi_1(a,b) = a + b \\ z_2 = \varphi_2(b) = b \times b & z_2 = \varphi_2(a,b) = a - b \\ z_3 = \varphi_3(z_1, z_2) = z_1 - z_2 & z_3 = \varphi_3(z_1, z_2) = z_1 \times z_2 \end{array} \tag{1.17}$$

Em qualquer destes algoritmos z_3 será o valor $z = f(a,b)$ pretendido supondo que todos os cálculos foram efectuados sem erros. Quando efectuamos os cálculos com uma máquina será que é indiferente o algoritmo escolhido? Vamos ver com um exemplo numérico que de facto o resultado depende da forma como são efectuadas as contas.

Exemplo 1.6 *Comparar os 2 algoritmos dados em (1.17) para determinar* $z = a^2 - b^2$ *com* $a = 0.3237$ *e* $b = 0.3134$*, utilizando uma máquina com* $VF(10, 4, -10, 10)$ *e arredondamento simétrico. O resultado exacto é* $z = a^2 - b^2 = 0.656213 \times 10^{-2}$.

$$\begin{array}{ll} \textit{Algoritmo } 1: & \textit{Algoritmo } 2: \\ \tilde{z}_1 = a \tilde{\times} a = 0.1048 & \tilde{z}_1 = a \tilde{+} b = 0.6371 \\ \tilde{z}_2 = b \tilde{\times} b = 0.9822 \times 10^{-1} & \tilde{z}_2 = a \tilde{-} b = 0.1030 \times 10^{-1} \\ \tilde{z}_1 \tilde{-} \tilde{z}_2 = 0.6580 \times 10^{-2} & \tilde{z}_1 \tilde{\times} \tilde{z}_2 = 0.6562 \times 10^{-2} \\ & \\ \delta_z = -0.27232 \times 10^{-2} & \delta_z = 0.00198 \times 10^{-2} \end{array}$$

■

Vemos assim que um dos algoritmos conduziu a um resultado bastante mais preciso do que o outro. Para entender o que se passa vamos determinar uma expressão para o erro correspondente a cada algoritmo. Para isso obtém-se sucessivamente o erro de cada operação elementar que constitui o algoritmo em estudo. Assim temos para o *Algoritmo 1:*

$$\begin{array}{lll} z_1 = \varphi_1(a) = a \times a & \Longrightarrow & \delta_{z_1} = \delta_a + \delta_a + \delta_{ar_1} \\ z_2 = \varphi_2(b) = b \times b & \Longrightarrow & \delta_{z_2} = \delta_b + \delta_b + \delta_{ar_2} \\ z_3 = \varphi_3(z_1, z_2) = z_1 - z_2 & \Longrightarrow & \delta_{z_3} = \frac{z_1}{z_3}\delta_{z_1} - \frac{z_2}{z_3}\delta_{z_2} + \delta_{ar_3} \end{array}$$

em que δ_{ar_1},δ_{ar_2} e δ_{ar_3} são os erros de arredondamento cometidos em cada operação elementar. Portanto

$$\delta_{z_3} = 2(\frac{z_1}{z_3}\delta_a - \frac{z_2}{z_3}\delta_b) + \frac{z_1}{z_3}\delta_{ar_1} - \frac{z_2}{z_3}\delta_{ar_2} + \delta_{ar_3}$$

Atendendo que $z_1 = a^2$, $z_2 = b^2$ e $z_3 = a^2 - b^2$ obtemos

$$\begin{aligned} \delta_z &= \delta_f + \delta_{ar} \\ \delta_f &= \frac{2a^2}{a^2-b^2}\delta_a - \frac{2b^2}{a^2-b^2}\delta_b \\ \delta_{ar} &= \frac{a^2}{a^2-b^2}\delta_{ar_1} - \frac{b^2}{a^2-b^2}\delta_{ar_2} + \delta_{ar_3} \end{aligned} \tag{1.18}$$

Da mesma forma para o *Algoritmo 2:*

$$\begin{aligned} z_1 = \varphi_1(a,b) = a+b \quad &\Longrightarrow \quad \delta_{z_1} = \frac{a}{a+b}\delta_a + \frac{b}{a+b}\delta_b + \delta_{ar_1} \\ z_2 = \varphi_2(a,b) = a-b \quad &\Longrightarrow \quad \delta_{z_2} = \frac{a}{a-b}\delta_a - \frac{b}{a-b}\delta_b + \delta_{ar_2} \\ z_3 = \varphi_3(z_1,z_2) = z_1 \times z_2 \quad &\Longrightarrow \quad \delta_{z_3} = \delta_{z_1} + \delta_{z_2} + \delta_{ar_3} \end{aligned}$$

Portanto

$$\delta_{z_3} = (\frac{a}{a+b} + \frac{a}{a-b})\delta_a + (\frac{b}{a+b} - \frac{b}{a-b})\delta_b + \delta_{ar_1} + \delta_{ar_2} + \delta_{ar_3}$$

Obtemos finalmente

$$\begin{aligned} \delta_z &= \delta_f + \delta_{ar} \\ \delta_f &= \frac{2a^2}{a^2-b^2}\delta_a - \frac{2b^2}{a^2-b^2}\delta_b \\ \delta_{ar} &= \delta_{ar_1} + \delta_{ar_2} + \delta_{ar_3} \end{aligned} \tag{1.19}$$

De (1.18) e (1.19) vemos que o erro relativo δ_z é constituído por dois termos δ_f e δ_{ar} com características diferentes. O primeiro, δ_f, depende dos erros relativos de a e b, argumentos de f, mas não depende dos erros de arredondamento δ_{ar_1}, δ_{ar_2} e δ_{ar_3} provenientes das sucessivas operações elementares efectuadas pelo computador. Por isso este termo, δ_f, é independente do algoritmo escolhido e pode obter-se directamente utilizando (1.10), como facilmente se depreende. Por outro lado o segundo termo, δ_{ar}, é explicitamente independente dos erros dos argumentos de f, e depende dos erros de arredondamento e consequentemente do algoritmo escolhido.

Vamos ver que estes dois tipos de erros, δ_f e δ_{ar}, estão relacionados respectivamente com dois conceitos: condicionamento e estabilidade numérica. O conceito de condicionamento intervem sempre quando um problema consiste em obter um resultado z a partir de um conjunto de dados $x_1, x_2, \ldots, x_m$.

Definição 1.2 *Um problema diz-se bem condicionado se pequenos erros nos dados produzem pequenos erros no resultado ou de forma equivalente, se uma pequena mudança nos dados produz uma pequena mudança no resultado. Caso contrário o problema é mal condicionado.*

Consideremos o caso em que $z \in \mathbb{R}$ relaciona-se com $x_1, x_2, \ldots, x_m$ através de uma função f duas vezes continuamente diferenciável tal que $z = f(x_1, x_2, \ldots, x_m)$. Neste caso podemos, recorrendo ao desenvolvimento em série de Taylor e desprezando termos de ordem superior, obter

$$e_f = f(x_1, x_2, \ldots, x_m) - f(\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_m) = \sum_{i=1}^{m} f'_{x_i}(x_1, x_2, \ldots, x_m) e_{x_i}$$

que é uma generalização de (1.9). Escrevendo esta relação em termos de erros relativos obtemos

$$\delta_f = \sum_{i=1}^{m} p_{x_i} \delta_{x_i} \tag{1.20}$$

em que

$$p_{x_i} = \frac{x_i f'_{x_i}(x_1, x_2, \ldots, x_m)}{f(x_1, x_2, \ldots, x_m)}$$

Como já referimos atrás os p_{x_i} são os números de condição associados s variáveis x_i e indicam como o erro relativo em x_i afecta o erro relativo em f. Se um desses números em valor absoluto for grande então o problema é mal condicionado. Se forem todos pequenos então o problema é bem condicionado.

No exemplo considerado atrás, quer para o *Algoritmo 1* quer para o *Algoritmo 2*, obtivemos a seguinte expressão

$$\delta_f = \frac{2a^2}{a^2 - b^2}\delta_a - \frac{2b^2}{a^2 - b^2}\delta_b$$

Notamos que podiamos ter obtido esta expressão utilizando (1.20). Para $b \approx a$, como $2a^2/|a^2 - b^2|$ e $2b^2/|a^2 - b^2|$ podem tomar valores muito elevados, vemos que $f(a, b) = a^2 - b^2$ é mal condicionada.

O cálculo da função f no computador é efectuado em vários passos; de forma que a cada passo k corresponde uma operação elementar φ_k a que está associado um erro de arredondamento δ_{ar_k}. Assim a função f é a composição destes φ_k, que dependem do algoritmo escolhido. Admitindo que o cálculo de f é efectuado com s passos, então o erro de arredondamento δ_{ar} acumulado nestes s passos tem a forma geral

$$\delta_{ar} = \sum_{k=1}^{s} q_k \delta_{ar_k} \qquad q_s = 1 \tag{1.21}$$

em que os pesos $q_1, q_2, \ldots, q_{k-1}$ dependem do algoritmo escolhido. Se para um certo passo t corresponde um φ_t mal condicionado, então os erros de arredondamento $\delta_{ar_1}, \delta_{ar_2}, \ldots, \delta_{ar_{t-1}}$ introduzidos nas operações elementares anteriores podem contribuir fortemente para o erro final; geralmente um ou mais dos $q_1, q_2, \ldots, q_{t-1}$ tomará valores absolutos grandes.

O erro total δ_z da função $z = f(x_1, x_2, \ldots, x_m)$ é dado por

$$\delta_z = \delta_f + \delta_{ar}$$

em que δ_f e δ_{ar} são dados respectivamente por (1.20) e (1.21). Sendo assim, temos que

$$\delta_z = \sum_{i=1}^{m} p_{x_i} \delta_{x_i} + \sum_{k=1}^{s} q_k \delta_{ar_k} \tag{1.22}$$

Definição 1.3 *Um algoritmo diz-se numericamente instável se pelo menos um dos pesos p_{x_i} ou q_k, em (1.22), for grande em valor absoluto. Caso contrário, o algoritmo diz-se numericamente estável.*

Com este definição, é obvio que num problema mal condicionado qualquer que seja o algoritmo escolhido ele será instável.

Para $b \approx a$, o *Algoritmo 1* é instável devido ao cálculo de $z_3 = z_1 - z_2$, sendo

$$|q_1| \gg 1 \text{ e } |q_2| \gg 1$$

com

$$q_1 = z_1/z_3 \text{ e } q_2 = -z_2/z_3$$

e o *Algoritmo 2* é instável devido ao cálculo de $z_2 = a - b$, sendo

$$|p_a| \gg 1 \text{ e } |p_b| \gg 1$$

com

$$p_a = a/(a-b) + a/(a+b) \text{ e } p_b = -b/(a-b) + b/(a+b)$$

No entanto comparando (1.18) com (1.19) podemos dizer que o *Algoritmo 1* é mais instável do que o *Algoritmo 2* visto que $|\delta_{ar}|$ é muito maior para o 1 algoritmo. A instabilidade do *Algoritmo 2* deve-se ao facto de $|\delta_f| \gg |\delta_a|$ e $|\delta_f| \gg |\delta_b|$.

Recorrendo a (1.18) e (1.19), podemos obter majorações para o erro relativo de z quando utilizamos respectivamente os *Algoritmo 1* e *Algoritmo 2*. Atendendo que

$$|\delta_{ar_i}| \leq u \qquad i = 1, 2, 3$$

em que u é a unidade de arredondamento, designando por $|\delta_d|$ e r as quantidades definidas respectivamente por

$$|\delta_d| = \max(|\delta_a|, |\delta_b|)$$

e

$$r = \frac{a^2 + b^2}{|a^2 - b^2|}$$

obtemos então para o *Algoritmo 1*

$$|\delta_z| \leq 2r|\delta_d| + (r+1)u$$

e para o *Algoritmo 2*

$$|\delta_z| \leq 2r|\delta_d| + 3u$$

Para os valores escolhidos no Exemplo 1.6 temos $u = 0.5 \times 10^{-3}$ e $r = 30.9352$. Se δ_a e δ_b forem nulo, caso em que a e b são conhecidos exactamente e são representados também exactamente no computador, então o resultado utilizando o *Algoritmo 2* é muito mais preciso do que utilizando o *Algoritmo 1*. É o caso considerado no Exemplo 1.6 em que o majorante de $|\delta_z|$ é respectivamente 1.59676×10^{-2} e 0.15×10^{-2} (comparar com os valores -0.27232×10^{-2} e 0.00198×10^{-2} de δ_z, obtidos anteriormente naquele exemplo). Se a e b são conhecidos exactamente mas não são representados exactamente no computador então temos que $|\delta_d| \leq u$ e portanto temos para o *Algoritmo 1*

$$|\delta_z| \leq (3r+1)u$$

e para o *Algoritmo 2*

$$|\delta_z| \leq (2r+3)u$$

Concluímos que mesmo para este caso é preferível o *Algoritmo 2*. Finalmente se a e b vêm afectados de erros de dados muito superior a u então $|\delta_f| \gg |\delta_{ar}|$ e é praticamente indiferente escolher um ou outro dos algoritmos.

Vamos a seguir dar um exemplo de um algoritmo para o cálculo de uma função bem condicionada que é numericamente instável e apresentar algoritmos alternativos que sejam estáveis.

Exemplo 1.7 *Dado um $x \approx 0$ pretende-se determinar o valor $z = f(x) = 1 - \cos x$. Verificar que este problema é bem condicionado mas que o cálculo de z a partir de $1 - \cos x$ conduz a um algoritmo numericamente instável. Apresentar 2 algoritmos alternativos que sejam estáveis.*

Como o número de condição de f é

$$p_x = xf'(x)/f(x) = x \sin x/(1 - \cos x)$$

e que $p_x \approx 2$ para $x \approx 0$ concluímos que o problema é bem condicionado. Como $\cos x \approx 1$ para $x \approx 0$ vamos ter um cancelamento subtractivo quando efectuarmos $1 - \cos x$ e portanto é de esperar que o algoritmo seja numericamente instável. De facto temos

$$\begin{aligned} z_1 = \varphi_1(x) = \cos x \quad &\Longrightarrow \quad \delta_{z_1} = -x \tan x \delta_x + \delta_{ar_1} \\ z_2 = \varphi_2(z_1) = 1 - z_1 \quad &\Longrightarrow \quad \delta_{z_2} = \frac{-z_1}{1 - z_1}\delta_{z_1} + \delta_{ar_2} \end{aligned}$$

Por conseguinte

$$\delta_z = x\frac{\sin x}{1 - \cos x}\delta_x + \frac{-\cos x}{1 - \cos x}\delta_{ar_1} + \delta_{ar_2}$$

Vemos que o número de condição

$$q_1 = \frac{-\cos x}{1 - \cos x}$$

pode tomar em módulo valores extremamente elevados quando for $x \approx 0$ e porconseguinte o algoritmo é numericamente instável. Notando que

$$1 - \cos x = \frac{\sin^2 x}{1 + \cos x} = 2\sin^2(x/2)$$

deixamos ao leitor o trabalho de apresentar 2 algoritmos estáveis e de verificar qual destes é o mais estável. ■

Vejamos agora um exemplo do cálculo de uma função bem condicionada utilizando um algoritmo que é estável apesar de ter associada a um dos passos (o primeiro) uma operação elementar mal condicionada.

Exemplo 1.8 *Dado um $x \approx 1$ pretende-se determinar o valor $z = f(x)$, com $f(x) = \exp(1 - x)$. Verificar que este problema é bem condicionado e que o cálculo de z a partir de $\exp(1 - x)$ conduz a um algoritmo numericamente estável apesar de $z_1 = \varphi_1(x) = 1 - x$ ser mal condicionada quando $x \approx 1$ (cancelamento subtractivo).*

Temos

$$z_1 = \varphi_1(x) = 1 - x \quad \Longrightarrow \quad \delta_{z_1} = \frac{-x}{z_1}\delta_x + \delta_{ar_1}$$
$$z_2 = \varphi_2(z_1) = exp(z_1) \quad \Longrightarrow \quad \delta_{z_2} = z_1\delta_{z_1} + \delta_{ar_2}$$

Por conseguinte

$$\delta_z = -x\delta_x + (1-x)\delta_{ar_1} + \delta_{ar_2}$$

Dado que $|p_x| \approx 1$ ($p_x = -x$), f é bem condicionada. Dado que $|q_1| \approx 0$ ($q_1 = 1 - x$) e $q_2 = 1$ o algoritmo é estável apesar do número de condição $-x/(1-x)$ associado a φ_1 tomar valores absolutos elevados. ■

1.5 Problemas

1. Considere o cálculo de $w = \sqrt{x+4} - 2$ para $|x| \ll 1$.
   (a) Será um problema bem condicionado? Justifique. Verifique que expressão que define w corrresponde um algoritmo instável.
   (b) Apresente um algoritmo estável para o cálculo de w e mostre que ele é de facto estável.

2. As raízes z_1 e z_2 da equação de 2 grau $x^2 + 2kx + c = 0$ são dadas por

$$z_1 = \sqrt{k^2 - c} - k \qquad z_2 = -\sqrt{k^2 - c} - k$$

   (a) Obtenha a expressão do erro relativo δ_{z_1} de z_1.
   (b) Justifique a afirmação seguinte: "Quando $k \gg \sqrt{|c|}$, o valor que se obtém para z_1 é pouco preciso, o que não se observa com z_2 ". Escreva um algoritmo alternativo do dado e que, nas condições anteriores, conduza a um valor mais preciso.

3. Na resolução de equações utilizando a aceleração de Aitken é necessário calcular

$$w_1 = (x_2 - x_1) - (x_1 - x_0)$$

   ou

$$w_2 = (x_2 - 2x_1) + x_0$$

   A expressão do erro relativo de w_2 é

$$\delta_{w_2} = \delta_f + \delta_2 + \delta_{ar_3}$$

   com

$$\delta_f = (x_2\delta_{x_2} - 2x_1\delta_{x_1} + x_0\delta_{x_0})/w$$
$$\delta_2 = (-2x_1\delta_{ar_1} + (x_2 - 2x_1)\delta_{ar_2})/w$$

   (a) Determine δ_{w_1}.
   (b) Compare δ_{w_1} com δ_{w_2} e diga, justificando sucintamente, qual dos 2 algoritmos prefere para determinar $w = w_1 = w_2$, supondo que as iteradas já estão perto da solução z da equação a resolver, *i.e.*, $x_2 \approx x_1 \approx z$.

4. Dado um $x \approx 0$, pretende-se determinar o valor de $z = f(x) = x - \sin x$ com um computador que funciona em base decimal com 7 algarismos na mantissa e arredondamento simétrico.

   (a) Obtenha a expressão do erro $\delta_z = \delta_f + \delta_{ar}$.

   (b) Em face desta expressão diga, justificando, se o problema é bem condicionado e se o algoritmo é numericamente estável. Note que

   $$1 - cosx = x^2/2(1 + O(x^2)) \qquad z = x^3/6(1 + O(x^2))$$

   Obtenha um majorante do erro relativo de z admitindo que $x = 0.01$ é representado exactamente nesse computador ($z \approx (1./6) \times 10^{-6}$).

5. Conhecido $y = \cos x$, sabe-se que $\sin x$ pode ser obtido através da expressão $z = \sqrt{1 - y^2}$. Obtenha a expressão do erro δ_z de z. Para valores de x próximos de 0, o que pode afirmar quanto precisão do resultado? Justifique.

6. A expressão

   $$P = (\sin a + \cos a + 1)h$$

   permite obter o valor do perímetro P de um triângulo rectângulo, sendo h o comprimento da hipotenusa e a um dos ângulos agudos. Obtenha a expressão do erro δ_P de P.

7. Suponha a seguinte função

   $$z = f(x) = \frac{\sqrt[3]{x}}{x^2 + \ln x}$$

   Sabe-se que a unidade de arredondamento do computador é de 0.5×10^{-7}.

   (a) Indique a expressão que permita calcular o erro δ_z de z.

   (b) Compare esta expressão com a que se obtem utilizando $\delta_f = p_x \delta_x$ em que p_x é o número de condição de f. Justifique devidamente.

8. Considere a seguinte expressão:

   $$c = \sqrt{a^2 + b^2} \sin \beta$$

   Apresente a expressão que lhe permite calcular o erro δ_c de c.

9. Na resolução da equação $x^2 + 2x - 0.0012345 = 0$ somos conduzidos ao cálculo de

   $$x_1 = \sqrt{1.0012345} - 1$$

   Suponha que utiliza um computador decimal que opera com 6 dígitos na mantissa e arredondamento simétrico. Obtenha a expressão do erro δ_{x_1} de x_1 e diga quantos algarismos significativos pode garantir a x_1. Indique uma fórmula que permita determinar x_1 com maior precisão.

10. Considere a divisão sintética de um polinómio $p_3(x)$ por $x - z$.

$$\begin{array}{c|cccc} & a_3 & a_2 & a_1 & a_0 \\ z & & za_3 & zb_2 & zb_1 \\ \hline & a_3 & b_2 & b_1 & b_0 \approx 0 \end{array}$$

Pretende-se determinar o erro dos coeficientes b_1 e b_2 provocados pelo erro da raiz z de $p_3(x)$.

(a) Obtenha as expressões dos erros de b_1 e b_2.

(b) Considerando os a_k exactos determine um majorante do erro de b_2.

11. É usual no computador determinar $\sqrt{c}$ $(c > 0)$ a partir da sucessão

$$x_{m+1} = (x_m + c/x_m)/2 \qquad m = 0, 1, \ldots \qquad x_0 > \sqrt{c}$$

Suponha que o computador opera na base 10 com 7 algarismos na mantissa e arredondamento simétrico.

(a) Obtenha a expressão do erro $\delta_{x_{m+1}}$ em função de δ_{x_m}, δ_c e dos erros de arredondamento.

(b) Admitindo que m é suficientemente grande de modo a que $x_m \approx \sqrt{c}$, e supondo que $c \approx 4$ é representado exactamente no computador, determine quantos algarismos significativos pode garantir a aproximação x_{m+1} de $\sqrt{c}$.

12. Sendo $x \gg 1$, pretende-se calcular $z = f(x) = x - (x^2 - 1)^{1/2}$ com um computador que funciona em base decimal com 9 algarismos na mantissa e arredondamento simétrico. Nestas condições a expressão do erro é

$$\delta_z \approx \delta_f - x^2[\delta_{ar_1} + \delta_{ar_2} + 2\delta_{ar_3}] + \delta_{ar_4}$$

(a) Determine δ_f. Diga se o problema é bem condicionado e se o algoritmo é estável. Justifique.

(b) Supondo que $x = 10^3$ é representado exactamente nesse computador, determine o número de algarismos significativos que pode garantir ao valor calculado $\tilde{z}$. Apresenta um algoritmo estável, justificando a sua escolha.

13. Considere o cálculo de $z = f(x, y, u) = u(x - y)$ que pode ser feito utilizando 2 algoritmos diferentes:

$$w_1 = (x - y) \times u \qquad w_2 = u \times x - u \times y$$

A expressão do erro para o primeiro algoritmo é $\delta_{w_1} = \delta_f + \delta_{ar_1} + \delta_{ar_2}$ com

$$\delta_f = \delta_u + \frac{x}{x - y}\delta_x - \frac{y}{x - y}\delta_y$$

(a) Determine δ_{w_2}.

(b) Supondo agora que u, x e y são representados exactamente no computador, encontra condições para as quais um (qual?) dos algoritmos é preferível ao outro. Em que condições é indiferente utilizar um ou outro destes algoritmos.

14. Considere o cálculo de $z = f(x, y) = x^2 - y^2$ que pode ser feito utilizando 3 algoritmos diferentes:

$$w_1 = x \times x - y \times y \quad w_2 = (x + y) \times (x - y) \quad w_3 = (x + y) \times x - (x + y) \times y$$

As expressões dos erros para os 2 primeiros algoritmos são

$$\delta_{w_1} = \delta_f + \delta_1 \quad \delta_{w_2} = \delta_f + \delta_2$$

com

$$\begin{aligned}
\delta_f &= \frac{2x^2}{x^2 - y^2}\delta_x - \frac{2y^2}{x^2 - y^2}\delta_y \\
\delta_1 &= \frac{x^2}{x^2 - y^2}\delta_{ar_1} - \frac{y^2}{x^2 - y^2}\delta_{ar_2} + \delta_{ar_3} \\
\delta_2 &= \delta_{ar_1} + \delta_{ar_2} + \delta_{ar_3}
\end{aligned}$$

(a) Determine δ_{w_3}.

(b) Supondo agora que x e y são representados exactamente no computador, encontra para cada algoritmo condições para as quais este algoritmo é melhor do que os outros.

Capítulo 2

Equações não lineares

2.1 Introdução

Neste capítulo apresentamos os principais métodos iterativos para a resolução de equações não lineares a uma variável (o estudo de sistemas de equações não lineares é adiado para o Capítulo 3). Obter soluções de uma equação é um dos problemas que surge com maior frequência nas aplicações, especialmente quando a equação é algébrica. Por isso, reservamos a parte final deste capítulo à determinação dos zeros de polinómios.

Depois de apresentarmos um exemplo simples e alguns conceitos dos métodos iterativos faremos o estudo dos métodos da *bissecção, falsa posição, secante* e de *Newton*. A seguir faremos o estudo geral dos métodos iterativos do ponto fixo a uma variável. Esta generalização tem como objectivos enquadrar os vários métodos anteriores e prepararmo-nos para o estudo, no Capítulo 3, dos métodos iterativos para a resolução de sistemas de equações. Aproveitando este contexto mais geral estudaremos uma técnica de aceleração dos métodos de convergência linear, e apresentaremos critérios de paragem das iterações. Finalmente, particularizaremos o método de Newton à resolução de equações polinomiais e chamaremos à atenção de algumas dificuldades levantadas na resolução numérica das mesmas.

Se pretendermos obter as soluções de uma equação algébrica do segundo grau, $ax^2 + bx + c = 0$, basta utilizar a conhecida fórmula resolvente

$$x = [-b \pm \sqrt{b^2 - 4ac}]/(2a)$$

Para equações algébricas de grau 3 ou 4 ainda é possível obter fórmulas resolventes, se bem que mais complicadas; mas para a equação geral de grau n com $n > 4$ Galois demonstrou em 1832 que é impossível obter fórmulas resolventes que permitam determinar directamente as soluções da equação algébrica.

As equações algébricas de grau superior a quatro bem como as equações transcendentes (p. ex. $x = 2\ sin\ x$) são resolvidas numericamente por métodos iterativos. Num método iterativo começa-se com uma aproximação inicial x_0 da raiz z (solução da equação) a partir da qual se obtém outra aproximação x_1, desta obtém-se outra aproximação x_2 e assim sucessivamente. Um método iterativo consiste portanto em obter uma sucessão de valores $x_0, x_1, x_2, \ldots$ que se pretende que tenha como limite

uma raiz z da equação. O processo iterativo prolonga-se até uma aproximação x_m satisfazer a precisão pretendida à partida.

Uma equação não linear pode genericamente escrever-se sob a forma

$$f(x) = 0 \tag{2.1}$$

com $f : D \to \mathbb{R}$, em que $D \subset \mathbb{R}$ é o domínio de f. Para ilustrar o conceito de um método iterativo para a resolução de (2.1), começamos com um exemplo. Consideremos a resolução de

$$f(x) = x^2 - a$$

para um dado $a > 0$. Este problema tem uma aplicação prática no computador para a obtenção da raiz quadrada de um número.

Seja $x_0 \approx \sqrt{a}$ uma aproximação da solução da equação. Então

$$x_0 \leq \sqrt{a} \leq a/x_0 \quad ou \quad a/x_0 \leq \sqrt{a} \leq x_0$$

É de admitir então que

$$x_1 = (x_0 + a/x_0)/2$$

constitui uma melhor aproximação de $\sqrt{a}$ do que x_0. Assim, um possível método iterativo para determinar $\sqrt{a}$ consiste em utilizar a relação anterior indefinidamente, i. e., fazer[1]

$$x_{m+1} = (x_m + a/x_m)/2 \qquad m = 0, 1, 2, \ldots \tag{2.2}$$

Para este método ser de alguma utilidade a sucessão $x_0, x_1, x_2, \ldots$ deverá convergir para o zero $z = \sqrt{a}$ de f, *i.e.*, o erro

$$e_m = z - x_m$$

deverá tender para zero quando m tender para infinito. Começemos por relacionar o erro em duas iterações sucessivas. Temos que

$$e_{m+1} = \sqrt{a} - (x_m + a/x_m)/2 = (2\sqrt{a}x_m - x_m^2 - a)/(2x_m)$$

e portanto

$$e_{m+1} = -e_m^2/(2x_m) \tag{2.3}$$

Se $x_0 > 0$ então de (2.2) vemos que $x_m > 0$ para qualquer $m \geq 0$ e portanto de (2.3) concluimos que $e_m \leq 0$ para qualquer $m > 0$, *i.e.*, $x_m \geq \sqrt{a}$ para qualquer $m > 0$. Escrevemos (2.3) na forma

$$e_{m+1} = e_m(1 - \sqrt{a}/x_m)/2$$

Dado que $0 \leq (1 - \sqrt{a}/x_m) < 1$ para $m > 0$ vemos que

$$|e_{m+1}| \leq |e_m/2| \qquad m = 1, 2, \ldots$$

e portanto $e_m \to 0$. Concluimos assim que a sucessão $\{x_m\}$ dada por (2.2) converge para $z = \sqrt{a}$ se $x_0 \in (0, \infty)$.

Depois de verificar a convergência de um método iterativo temos que examinar a sua rapidez de convergência.

[1] Mais tarde veremos que este método iterativo corresponde à aplicação do método de Newton.

Definição 2.1 *Seja $\{x_m\}$ uma sucessão que converge para z e $e_m = z - x_m$. Se existirem reais $p \geq 1$ e $K_\infty > 0$ tais que*

$$\lim_{m\to\infty} |e_{m+1}|/|e_m|^p = K_\infty \tag{2.4}$$

então a sucessão $\{x_m\}$ converge para z com convergência de ordem p. A K_∞ chama-se coeficiente assintótico de convergência. Se $p = 1$ a convergência é linear; se $p > 1$ a convergência é supralinear e no caso de $p = 2$ a convergência é quadrática.

O coeficiente assintótico de convergência K_∞ mede a rapidez de convergência quando m for suficientemente grande, *i.e.*,

$$|e_{m+1}| \approx K_\infty |e_m|^p \qquad m \gg 1$$

Pode ser difícil determinar K_∞. De (2.4) facilmente se verifica que existe $K > K_\infty$ tal que

$$|e_{m+1}| \leq K|e_m|^p \qquad m \geq 0 \tag{2.5}$$

A desigualdade (2.5) pode ser útil quando se pretende obter uma majoração do erro e_{m+1} e é às vezes mais fácil de obter do que o coeficiente assintótico de convergência; neste caso K constitui um majorante de K_∞. No entanto K pode tomar valores bastantes superiores a K_∞ e portanto é importante a determinação do coeficiente assintótico de convergência.

Usando (2.3) e a Definição 2.1 vemos que o método definido por (2.2) é de convergência quadrática, com $K_\infty = 1/(2\sqrt{a})$. O exemplo considerado ilustra a construção de um método iterativo para resolver uma equação não linear e a análise da sua convergência. Esta análise inclui a prova da convergência sob certas condições (no exemplo, $x_0 \in (0, \infty)$) e a determinação da ordem da convergência.

2.2 Métodos da bissecção e da falsa posição

Os dois métodos apresentados a seguir baseiam-se no seguinte corolário do teorema do valor intermédio:

Teorema 2.1 *Dado um intervalo $I = [a, b]$ e uma função f que satisfaz às condições:*

$$f \text{ contínua em } [a,b] \tag{2.6}$$

$$f(a)f(b) < 0 \tag{2.7}$$

então a equação

$$f(x) = 0$$

tem pelo menos uma solução z em (a, b). ■

É fácil de ver que (2.7) não garante por si só a existência de uma raiz z. Para aplicar os métodos da bissecção e da falsa posição basta que se verifique (2.6) e (2.7). Se f satisfazer, além disso, a condição:

$$f' \text{ existe, é contínua e não se anula em } (a, b)$$

conclui-se então, usando o teorema de Rolle, que a solução z é única em (a, b).

2.2.1 Método da bissecção

Suponhamos que a função f satisfaz às condições (2.6) e (2.7), então esta função possui pelo menos um zero z no intervalo $I = [a, b]$. O método da *bissecção* consiste em construir subintervalos $I_m = [a_m, b_m] \subset I = [a, b]$ por divisões sucessivas a meio e relativamente aos quais também se verifica a condição (2.7). Deste modo vamos confinando um zero da função f a intervalos tão pequenos quanto desejarmos. Concretizando, ponhamos

$$I_m = [a_m, b_m] \qquad m = 0, 1, 2, \ldots$$

em que $a_0 = a$ e $b_0 = b$ e seja x_{m+1} o ponto médio do intervalo I_m, *i.e.*,

$$x_{m+1} = \frac{a_m + b_m}{2}$$

Consideremos as três situações seguintes . Se $f(x_{m+1}) = 0$ então x_{m+1} é um zero e o processo iterativo termina. Se $f(a_m)$ e $f(x_{m+1})$ tiverem sinais diferentes faça-se

$$a_{m+1} = a_m \quad e \quad b_{m+1} = x_{m+1}$$

de modo a que o novo subintervalo contenha um zero de f. Se pelo contrário forem $f(b_m)$ e $f(x_{m+1})$ a ter sinais diferentes então ponha-se

$$a_{m+1} = x_{m+1} \quad e \quad b_{m+1} = b_m$$

Deste modo garante-se que o novo subintervalo $I_{m+1} = [a_{m+1}, b_{m+1}]$ contém pelo menos um zero de f. Pela forma como foi construído, este subintervalo tem um comprimento que é metade do subintervalo que o precedeu. Portanto

$$b_{m+1} - a_{m+1} = \frac{1}{2}(b_m - a_m) \tag{2.8}$$

Por outro lado, sendo z um zero localizado em I_{m+1}, temos que

$$|z - x_{m+1}| \leq |x_{m+1} - x_m|$$

A Figura 2.1 ilustra esquematicamente a aplicação do método.

Façamos agora o estudo da convergência do método da bissecção. Como no intervalo I_m está localizado pelo menos um zero z e x_{m+1} é o ponto médio deste intervalo então

$$|e_{m+1}| \leq (b_m - a_m)/2 \qquad m = 0, 1, 2, \ldots$$

em que $e_m = z - x_m$. Aplicando sucessivamente (2.8) obtemos

$$|e_m| \leq (b - a)/2^m \qquad m = 0, 1, 2, \ldots \tag{2.9}$$

em que, para o caso de $m = 0$, se considera que x_0 é um dos extremos de I. Vemos assim que o método é convergente. Da Definição 2.1 não podemos concluir que o método da bissecção é de convergência linear. No entanto podemos compara-lo com os método de convergência linear. Vimos da Definição 2.1 que se um método possui convergência linear então

$$|e_m| \leq K|e_{m-1}| \quad m = 1, 2, \ldots$$

Figura 2.1: Exemplo do método da bissecção

em que K é um majorante do coeficiente assintótico de convergência K_∞. Aplicando m vezes esta desigualdade obtemos

$$|e_m| \leq K^m |e_0|$$

Notando que $|e_0| = |z - x_0| \leq (b - a)$ se $x_0 \in [a, b]$ e $z \in [a, b]$, vemos que quando um método converge linearmente então existe um K tal que

$$|e_m| \leq K^m (b - a)$$

Comparando com (2.9), vemos que o método da bissecção comporta-se de uma forma parecida com métodos de convergência linear com o coeficiente assintótico de convergência igual a 1/2.

Apresentamos a seguir um algoritmo e um exemplo de aplicação do método da bissecção.

Algoritmo 2.1 *(Método da bissecção)*

NOTA: *Condição suficiente de convergência do método:* f é contínua em $[a, b]$ e tal que $f(a)f(b) < 0$.

INICIALIZAÇÃO:
$a_0 = a$, $b_0 = b$, $x_0 = b$

CICLO: PARA $m = 0, 1, ...$ FAZER

$x_{m+1} = \dfrac{a_m + b_m}{2}$

SE $|x_{m+1} - x_m| \leq \epsilon$ OU $f(x_{m+1}) = 0$
ENTÃO fazer $raiz = x_{m+1}$, SAÍDA

SE $f(x_{m+1})f(a_m) < 0$
ENTÃO fazer $a_{m+1} = a_m,\ b_{m+1} = x_{m+1}$
SENÃO fazer $a_{m+1} = x_{m+1},\ b_{m+1} = b_m$

FIM DO CICLO.

Tabela 2.1: Exemplo do método da bissecção para $x^2 - 2 = 0$

m	a_{m-1}	b_{m-1}	x_m	$f(x_m)$	e_m
-1			0.000000	-2.000000	1.414214
0			2.000000	2.000000	-0.585787
1	x_{-1}	x_0	1.000000	-1.000000	0.414214
2	x_1	x_0	1.500000	0.250000	-0.085787
3	x_1	x_2	1.250000	-0.437500	0.164214
4	x_3	x_2	1.375000	-0.109375	0.039214
5	x_4	x_2	1.437500	0.066406	-0.023287
6	x_4	x_5	1.406250	-0.022461	0.007964
7	x_6	x_5	1.421875	0.021729	-0.007662
8	x_6	x_7	1.414063	-0.000427	0.000151
9	x_8	x_7	1.417969	0.010635	-0.003756
10	x_8	x_9	1.416016	0.005100	-0.001803
11	x_8	x_{10}	1.415039	0.002336	-0.000826
12	x_8	x_{11}	1.414551	0.000954	-0.000337
13	x_8	x_{12}	1.414307	0.000263	-0.000093
14	x_8	x_{13}	1.414185	-0.000082	0.000029
15	x_{14}	x_{13}	1.414246	0.000091	-0.000032
16	x_{14}	x_{15}	1.414215	0.000004	-0.000002
17	x_{14}	x_{16}	1.414200	-0.000039	0.000014
18	x_{17}	x_{16}	1.414207	-0.000017	0.000006
19	x_{18}	x_{16}	1.414211	-0.000006	0.000002

Exemplo 2.1 *Determinar, pelo método da bissecção (Algoritmo 2.1), a raiz positiva da equação $f(x) = x^2 - 2 = 0$ com um erro inferior a $\epsilon = 5E - 6$, usando o computador[2] em precisão simples. Tomar $a = 0$, $b = 2$.*

Este exemplo é escolhido pela sua simplicidade e por se conhecer a solução exacta que é $z = \sqrt{2} = 1.41421356$ (com 9 algarismos significativos). Tomando $I = [0, 2]$ verificamos que $f(0)f(2) < 0$ e como f é contínua em I temos a garantia que o método converge. Apresenta-se na Tabela 2.1 os resultados. Desta tabela vê-se que x_8 é uma melhor aproximação de z do que $x_9, x_{10}, x_{11}, x_{12}$. Comparando $f(x_7) = 0.021729$ com $f(x_8) = -0.000427$ era de esperar que z estivesse muito mais perto de x_8 do que de $x_9 = (x_7 + x_8)/2$. Assim, se além do sinal utilizarmos o próprio valor de $f(x_m)$ é de

[2]Rede do CIIST constituida por VAXs 750 e 780.

esperar que obtenhamos métodos com convergência mais rápida do que a do método da bissecção. Um desses métodos é o da falsa posição que estudaremos a seguir. ■

Este exemplo permite realçar duas características importantes do método da bissecção. A primeira é a de que a convergência poderá ser bastante lenta. A segunda é a de que a convergência é certa quando f satisfaz às condições (2.6) e (2.7), *i.e.*, quando f é contínua num intervalo $[a, b]$ e $f(a)f(b) < 0$. Como é frequente verificar-se estas condições, o método da bissecção é por vezes utilizado como fase preparatória de localização de zeros num intervalo mais ou menos pequeno que torne possível o arranque de outros métodos mais rápidos mas cuja convergência pode estar condicionada a uma relativamente boa aproximação inicial do zero.

2.2.2 Método da falsa posição

Como já referimos, a única informação sobre a função f que o método da bissecção utiliza é o seu sinal, o que é muito pouco. É de esperar que, se recorrermos a mais informação sobre o andamento de f, possamos conceber métodos mais rápidos. Tal como para o método da bissecção, consideremos no método da *falsa posição* uma função f que satisfaz, no intervalo $I_m = [a_m, b_m]$, às condições (2.6) e (2.7). Nesse intervalo $f(x)$ é aproximado pela secante

$$p_1(x) = f(b_m) + \frac{f(b_m) - f(a_m)}{b_m - a_m}(x - b_m) \tag{2.10}$$

que passa pelos pontos $(a_m, f(a_m))$ e $(b_m, f(b_m))$ (como veremos mais tarde p_1 é um polinómio interpolador de f). O zero x_{m+1} de p_1 constitui uma aproximação do zero z de f. Então, se fizermos em (2.10) $x = x_{m+1}$ obtemos

$$p_1(x_{m+1}) = 0 = f(b_m) + \frac{f(b_m) - f(a_m)}{b_m - a_m}(x_{m+1} - b_m) \tag{2.11}$$

e portanto

$$x_{m+1} = b_m - f(b_m)\frac{b_m - a_m}{f(b_m) - f(a_m)} \tag{2.12}$$

Fazendo $a_{m+1} = x_{m+1}$ ou $b_{m+1} = x_{m+1}$ e mantendo o outro extremo do intervalo inalterado consoante for $f(x_{m+1})f(b_m) < 0$ ou $f(a_m)f(x_{m+1}) < 0$, e procedendo de igual modo podemos obter nova aproximação do zero z. Apresentamos a seguir um algoritmo do método.

Algoritmo 2.2 *(Método da falsa posição)*

NOTA: *Condição suficiente de convergência do método:* f é contínua em $[a, b]$ e tal que $f(a)f(b) < 0$.

INICIALIZAÇÃO:
$a_0 = a$, $b_0 = b$, $x_0 = b$

CICLO: PARA $m = 0, 1, ...$ FAZER

$$x_{m+1} = b_m - f(b_m)\frac{b_m - a_m}{f(b_m) - f(a_m)}$$

SE $|x_{m+1} - x_m| \leq \epsilon$ OU $f(x_{m+1}) = 0$
ENTÃO fazer $raiz = x_{m+1}$, SAÍDA

SE $f(x_{m+1})f(a_m) < 0$
ENTÃO fazer $a_{m+1} = a_m,\ b_{m+1} = x_{m+1}$
SENÃO fazer $a_{m+1} = x_{m+1},\ b_{m+1} = b_m$

FIM DO CICLO.

Notando que x_{m+1} é a abcissa do ponto de intersecção do eixo dos x com a recta que passa pelos pontos $(a_m, f(a_m))$ e $(b_m, f(b_m))$, construimos um esboço, representado na Figura 2.2, que permite visualizar o método da falsa posição. Desta

Figura 2.2: Exemplo do método da falsa posição

figura vemos que para o caso nela representado tem-se $a_{m+1} = a_m$ para $\forall m \geq 1$. Isto deve-se ao facto de $f(x)$ não mudar de convexidade em $[a_1, b_1]$, *i.e.*, $f''(x) \neq 0$ em (a_1, b_1). Em geral, sempre que tenhamos $f''(x) \neq 0$ num intervalo aberto I_r com extremos a_r e b_r, sendo $r \geq 0$, então verifica-se que

$$f''(x)f(a_r) > 0 \qquad x \in I_r \implies a_{m+1} = a_m \qquad \forall m \geq r$$

ou

$$f''(x)f(b_r) > 0 \qquad x \in I_r \implies b_{m+1} = b_m \qquad \forall m \geq r$$

Assim se $f''(x) \neq 0$ em (a, b) o método iterativo da falsa posição exprime-se por

$$x_{m+1} = x_m - f(x_m)\frac{x_m - w_0}{f(x_m) - f(w_0)} \qquad m = 0, 1, \ldots \tag{2.13}$$

em que $w_0 = a$, $x_0 = b$ se $f''(x)f(a) > 0$ e $w_0 = b$, $x_0 = a$ se $f''(x)f(b) > 0$ com $x \in (a, b)$.

Exemplo 2.2 *Determinar, pelo método da falsa posição (Algoritmo 2.2), a raiz positiva da equação $f(x) = x^2 - 2 = 0$ com um erro inferior a $\epsilon = 5E - 6$, usando o computador em precisão simples. Tomar $a = 0$, $b = 2$.*

Este exemplo coincide com o escolhido anteriormente na ilustração do método da bissecção. Os resultados estão apresentados na Tabela 2.2. Neste caso $b_m = 2$, $\forall m \geq 0$. Notando que a razão $|e_m/e_{m-1}|$ dos sucessivos erros é praticamente constante, conclui-se da definição 2.1 que a convergência é linear com um coeficiente assintótico de convergência $K_\infty \approx 0.17$. Vemos que neste exemplo o método da falsa posição conduz a uma convergência linear mais rápida do que o da bissecção, visto o coeficiente assintótico de convergência ser inferior a 1/2. Nem sempre ao método da falsa posição corresponde uma convergência mais rápida que ao método da bissecção; se no exemplo considerado tivessemos escolhidos $b = 6$ um majorante para o coeficiente assintótico de convergência seria $0.62 > 0.5$. ■

Tabela 2.2: Exemplo do método da falsa posição para $x^2 - 2 = 0$

m	a_{m-1}	b_{m-1}	x_m	$f(x_m)$	e_m	e_m/e_{m-1}
-1			0.000000	-2.000000	1.414214	
0			2.000000	2.000000	-0.585787	
1	x_{-1}	x_0	1.000000	-1.000000	0.414214	-0.7071
2	x_1	x_0	1.333333	-0.222222	0.080880	0.1953
3	x_2	x_0	1.400000	-0.040000	0.014214	0.1757
4	x_3	x_0	1.411765	-0.006920	0.002449	0.1723
5	x_4	x_0	1.413793	-0.001189	0.000420	0.1717
6	x_5	x_0	1.414141	-0.000204	0.000072	0.1716
7	x_6	x_0	1.414201	-0.000035	0.000012	0.1722
8	x_7	x_0	1.414211	-0.000006	0.000002	0.1747
9	x_8	x_0	1.414213	-0.000001	$3.82E-7$	0.1760

Para compreender o comportamento da convergência do método da falsa posição, é necessário uma fórmula que relacione os erros em iterações sucessivas. Vimos que o método da falsa posição baseia-se na aproximação da função f pelo polinómio interpolador p_1 dado por (2.10). No estudo da interpolação polinomial veremos que

$$f(x) = p_1(x) + \frac{1}{2} f''(\xi_m)(x - a_m)(x - b_m) \tag{2.14}$$

em que p_1 é dado por (2.10) e ξ_m é um real compreendido entre o mínimo de $\{a_m, b_m, x\}$ e o máximo de $\{a_m, b_m, x\}$ o que passamos a partir de agora a representar por $\xi_m \in \text{int}(a_m, b_m, x)$. Note-se que se fizermos tender b_m para a_m, então (2.14) reduz-se a conhecida fórmula de Taylor

$$f(x) = f(a_m) + f'(a_m)(x - a_m) + \frac{1}{2} f''(\xi_m)(x - a_m)^2$$

em que $\xi_m \in \text{int}(a_m, x)$. Fazendo em (2.14) $x = z$, sendo z um zero de f, temos

$$f(z) = 0 = f(b_m) + \frac{f(b_m) - f(a_m)}{b_m - a_m}(z - b_m) + \frac{1}{2} f''(\xi_m)(z - a_m)(z - b_m) \tag{2.15}$$

em que $\xi_m \in \text{int}(a_m, b_m, z)$. Utilizando novamente (2.11)

$$p_1(x_{m+1}) = 0 = f(b_m) + \frac{f(b_m) - f(a_m)}{b_m - a_m}(x_{m+1} - b_m)$$

e subtraindo a (2.15), obtemos

$$\frac{f(b_m)-f(a_m)}{b_m-a_m}(z-x_{m+1})+\frac{1}{2}f''(\xi_m)(z-a_m)(z-b_m)=0 \tag{2.16}$$

Atendendo ao teorema de Lagrange, temos que

$$\frac{f(b_m)-f(a_m)}{b_m-a_m}=f'(\eta_m)$$

em que $\eta_m \in (a_m, b_m)$. Portanto, de (2.16) temos finalmente que

$$z-x_{m+1}=-\frac{f''(\xi_m)}{2f'(\eta_m)}(z-a_m)(z-b_m) \tag{2.17}$$

em que $\eta_m \in (a_m, b_m)$ e $\xi_m \in \mathrm{int}(a_m, b_m, z)$.

Como vimos, quando $f''(x) \neq 0$ em (a, b), um dos extremos a_m ou b_m imobiliza-se. Tal como em (2.13) designamos este extremo por w_0. Assim, com esta condição, se em (2.17) substituirmos a_m e b_m por w_0 e x_m, obtemos

$$e_{m+1}=[-\frac{f''(\xi_m)}{2f'(\eta_m)}(z-w_0)]e_m$$

em que $\eta_m \in \mathrm{int}(w_0, x_m)$, $\xi_m \in \mathrm{int}(w_0, x_m)$ e $e_m = z - x_m$. Com a condição considerada, $f''(x) \neq 0$ em (a, b), o termo entre chavetas é sempre diferente de zero e portanto a convergência do método é só linear. Como já referimos atrás, se w_0 estiver muito afastado de z, o coeficiente assintótico de convergência pode ser maior do que 1/2.

Como acabamos de ver, o método da falsa posição tende a desenvolver uma convergência lenta quando um dos extremos a_m ou b_m se imobiliza. Para evitar este inconveniente pode usar-se a seguinte modificação. Suponhamos para exemplificar que $b_m = b_{m-1}$. Então, em vez de se empregarem os pontos $(a_m, f(a_m))$ e $(b_m, f(b_m))$ para traçar a secante, utilizam-se os pontos $(a_m, f(a_m))$ e $(b_m, f(b_m)/2)$. A figura 2.3 mostra qualitativamente o efeito produzido por esta modificação. Convém realçar que a secante continua a ser definida por um ponto acima e outro abaixo do eixo dos x tal como na versão original. Para melhor compreensão deste método apresentamos a seguir um algoritmo e um exemplo.

Algoritmo 2.3 *(Método da falsa posição modificado)*

NOTA: *Condição suficiente de convergência do método:* f é contínua em $[a, b]$ e tal que $f(a)f(b) < 0$.

INICIALIZAÇÃO:

$a_0 = a$, $b_0 = b$, $x_0 = b$, $F = f(a_0)$, $G = f(b_0)$

CICLO: PARA $m = 0, 1, ...$ FAZER

$x_{m+1} = b_m - G\dfrac{b_m - a_m}{G - F}$

SE $|x_{m+1} - x_m| \leq \epsilon$ OU $f(x_{m+1}) = 0$

Figura 2.3: Exemplo do método da falsa posição modificado

ENTÃO fazer $raiz = x_{m+1}$, SAÍDA

SE $f(x_{m+1})f(a_m) < 0$
ENTÃO fazer $a_{m+1} = a_m,\ b_{m+1} = x_{m+1},\ G = f(x_{m+1})$
SE TAMBÉM $f(x_m)f(x_{m+1}) > 0$
ENTÃO fazer $F = F/2$

SENÃO fazer $a_{m+1} = x_{m+1},\ b_{m+1} = b_m,\ F = f(x_{m+1})$
SE TAMBÉM $f(x_m)f(x_{m+1}) > 0$
ENTÃO fazer $G = G/2$

FIM DO CICLO.

Tabela 2.3: Exemplo do método da falsa posição modificado para $x^2 - 2 = 0$

m	a_{m-1}	b_{m-1}	x_m	$f(x_m)$	e_m	e_m/e_{m-1}
-1			0.000000	-2.000000	1.414214	
0			2.000000	2.000000	-0.585787	
1	x_{-1}	x_0	1.000000	-1.000000	0.414214	-0.7071
2	x_1	x_0	1.333333	-0.222222	0.080880	0.1953
3	x_2	x_0	1.454545	0.115703	-0.040332	-0.4987
4	x_2	x_3	1.413043	-0.003308	0.001170	-0.0290
5	x_4	x_3	1.414197	-0.000047	0.000016	0.0141
6	x_5	x_3	1.414230	0.000045	-0.000016	-0.9681
7	x_5	x_6	1.414214	$-1.19E-7$	$2.42E-8$	-0.0015
8	x_7	x_6	1.414214	$-1.19E-7$	$2.42E-8$	1.0000

Exemplo 2.3 *Determinar, pelo método da falsa posição modificado (Algoritmo 2.3),*

a raiz positiva da equação $f(x) = x^2 - 2 = 0$ *com um erro inferior a* $\epsilon = 5E - 6$, *usando o computador em precisão simples. Tomar* $a = 0$, $b = 2$.

Comparando os resultados apresentados na Tabela 2.3, com os anteriormente obtidos quando da utilização do método da falsa posição (Tabela 2.2), notamos um ligeiro aumento na rapidez de convergência. ■

2.3 Métodos da secante e de Newton

2.3.1 Método da secante

Tal como no método da falsa posição, no *método da secante* o zero z de f é aproximado pelo zero x_{m+1} da secante p_1 que passa pelos pontos $(a_m, f(a_m))$ e $(b_m, f(b_m))$ e cuja equação é dada por (2.10). O valor de x_{m+1} é, como vimos, dado por (2.12). A diferença deste método relativamente ao anterior é que não se impõe que $f(a_m)f(b_m) < 0$ e portanto x_{m+1} pode não estar contido em (a_m, b_m) o que pode ter como consequência que o método da secante divirja em certos casos. Fazendo em (2.12) $a_m = x_{m-1}$ e $b_m = x_m$, obtemos o método da secante

$$x_{m+1} = x_m - f(x_m)\frac{x_m - x_{m-1}}{f(x_m) - f(x_{m-1})} \qquad m = 0, 1, \ldots \tag{2.18}$$

em que x_{-1} e x_0 são dados. Quando converge, o método da secante, normalmente, converge mais rapidamente do que o da falsa posição devido ao facto de a_m e b_m tenderem ambos para z. Ilustramos o método da secante na Figura 2.4.

Figura 2.4: Exemplo do método da secante

Para obter uma fórmula dos erros basta utilizar (2.17) e substituir nessa relação a_m e b_m por x_{m-1} e x_m. Obtemos

$$e_{m+1} = -\frac{f''(\xi_m)}{2f'(\eta_m)}e_{m-1}e_m \tag{2.19}$$

em que $\eta_m \in \text{int}(x_{m-1}, x_m)$ e $\xi_m \in \text{int}(x_{m-1}, x_m, z)$. Com esta fórmula é possível fazer um estudo da convergência do método da secante.

Teorema 2.2 *Seja f uma função duas vezes continuamente diferenciável numa vizinhança de um zero z e tal que $f'(z) \neq 0$. Então se os valores iniciais x_{-1} e x_0 são escolhidos suficientemente perto de z, a sucessão $\{x_m\}$ definida por (2.18) converge para z.*

Demonstração: Escolhemos uma vizinhança $I = [z - \rho, z + \rho]$ $(\rho > 0)$ de z suficientemente pequena de modo que

$$M = \frac{\max_{x \in I} |f''(x)|}{2 \min_{x \in I} |f'(x)|}$$

existe e

$$M\rho \leq \delta < 1 \tag{2.20}$$

Então, atendo a (2.20), para todos os $x_{m-1}, x_m \in I$ temos que

$$|e_{m+1}| \leq M|e_{m-1}||e_m| \tag{2.21}$$

Notando que $|e_m| \leq \rho$ quando $x_m \in I$, de (2.20) e (2.21) temos que

$$|e_{m+1}| \leq \delta|e_m| \tag{2.22}$$

e

$$|e_{m+1}| < \rho$$

e portanto $x_{m+1} \in I$. Concluimos por indução que se $x_{-1}, x_0 \in I$ e a condição (2.20) é satisfeita então: para qualquer $m \geq -1$, $x_m \in I$ e (2.22) é verificada. Se utilizarmos (2.22) m vezes obtemos

$$|e_m| \leq \delta^m |e_0|$$

donde concluimos que $\lim_{m \to \infty} |e_m| = 0$, visto que por hipótese $\delta < 1$. ■

Teorema 2.3 *Nas condições do Teorema 2.2 a ordem de convergência do método da secante é $p = (1 + \sqrt{5})/2 \approx 1.618$.* ■

Teorema 2.4 *Se $f \in C^k([a, b])$, $z \in (a, b)$ e $f(z) = f'(z) = \cdots = f^{(k-1)}(z) = 0$ e $f^{(k)}(z) \neq 0$, i.e., se f tiver um zero de multiplicidade k, então o método da secante converge linearmente desde que x_{-1} e x_0 estejam suficientemente perto de z e localizados ambos de um mesmo lado de z.* ■

Podemos agora apresentar um algoritmo para o método em estudo.

Algoritmo 2.4 *(Método da secante)*

NOTA: *Condição suficiente de convergência do método:* $f \in C^2([a,b])$, tem um zero z em $[a,b]$ e $|z - x_m| < \min_{x\in[a,b]} |f'(x)| / \max_{x\in[a,b]} |f''(x)|$ para $m = -1, 0$ (no caso de $f'(z) = 0$ e $f''(z) \neq 0$ o método converge linearmente desde que o intervalo $[a,b]$ seja suficientemente pequeno e que $z \notin [x_{-1}, x_0]$).

INICIALIZAÇÃO:

x_{-1} , $x_0 \in [a,b]$ com $x_{-1} \neq x_0$

CICLO: PARA $m = 0, 1, ...$ FAZER

$$x_{m+1} = x_m - f(x_m)\frac{x_m - x_{m-1}}{f(x_m) - f(x_{m-1})}$$

SE $|x_{m+1} - x_m| \leq \epsilon$

ENTÃO fazer $raiz = x_{m+1}$, SAÍDA

FIM DO CICLO.

Exemplo 2.4 *Determinar, pelo método da secante (Algoritmo 2.4), a raiz positiva da equação* $f(x) = x^2 - 2 = 0$ *com um erro inferior a* $\epsilon = 5E - 6$, *usando o computador em precisão simples. Tomar* $a = 0$, $b = 2$. *Considerar os casos: (i)* $x_{-1} = a$, $x_0 = b$ *(ii)* $x_{-1} = (a+b)/2$, $x_0 = 1.001x_{-1}$.

Tabela 2.4: Exemplo do método da secante para $x^2 - 2 = 0$

m	x_m	e_m	$\frac{\lvert e_m\rvert}{\lvert e_{m-1}\rvert^{1.618}}$	m	x_m	e_m	$\frac{\lvert e_m\rvert}{\lvert e_{m-1}\rvert^{1.618}}$
-1	0.000000	1.414214		-1	1.000000	0.414214	
0	2.000000	-0.585787		0	1.001000	0.413213	
1	1.000000	0.414214	0.9841	1	1.499762	-0.085548	0.3575
2	1.333333	0.080880	0.3366	2	1.400078	0.014136	0.7551
3	1.428571	-0.014358	0.8399	3	1.413797	0.000417	0.4102
4	1.413793	0.000420	0.4032	4	1.414216	-0.000002	0.6241
5	1.414211	0.000002	0.6299	5	1.414214	$2.42E-8$	36.588
6	1.414214	$2.42E-8$	35.277				

O caso (i) coincide com os exemplos apresesentados anteriormente para os método da bissecção e da falsa posição. Dado que no método da secante não se impõe $f(a)f(b) < 0$, é natural escolher para x_{-1} e x_0 valores a meio do intervalo $[a,b]$ que constituem melhores aproximações de z do que pelo menos um dos extremos do intervalo $[a,b]$; é esta a opção considerada no caso (ii). Dos resultados obtidos, que apresentamos na Tabela 2.4, depreendemos que o método da secante é de convergência supra linear, com $p = 1.618$ [3], e portanto converge mais rapidamente que os métodos anteriores. ■

[3] A sucessão $|e_m|/|e_{m-1}|^{1.618}$ converge para o coeficiente assintótico de convergência $K_\infty = |0.5f''(z)/f'(z)|^{1/1.618} = (0.5/\sqrt{2})^{1/1.618} = 0.5259$, contudo o último valor desta sucessão na tabela difere substancialmente daquele limite devido a influência dos erros de arredondamento.

2.3.2 Método de Newton

Tal como nos métodos precedentes, aproximamos o gráfico de $y = f(x)$ na vizinhança de x_m por uma função linear (recta), mas agora utilizando uma tangente em vez de uma secante. Tomamos o zero x_{m+1} da função linear para aproximação do zero z da função dada f. Obtém-se assim o *método iterativo de Newton* (Figura 2.5)

$$x_{m+1} = x_m - \frac{f(x_m)}{f'(x_m)} \qquad m = 0, 1, \ldots \tag{2.23}$$

De notar que o método da secante está relacionado com o de Newton visto que

Figura 2.5: Método de Newton

$$f'(x_m) \approx \frac{f(x_m) - f(x_{m-1})}{x_m - x_{m-1}}$$

quando $x_{m-1} \approx x_m$.

O método de Newton é o mais conhecido para determinar as raizes de uma equação. Não é sempre o melhor método para um determinado problema, mas a sua fórmula simples e a sua grande rapidez de convergência fazem dele o método que geralmente se utiliza em primeiro lugar para resolver uma equação.

Uma outra maneira de obter (2.23), é utilizar a fórmula de Taylor. Desenvolvemos $f(x)$ em torno de x_m,

$$f(x) = f(x_m) + f'(x_m)(x - x_m) + \frac{1}{2} f''(\xi_m)(x - x_m)^2$$

com $\xi_m \in \text{int}(x, x_m)$. Fazendo $x = z$, com $f(z) = 0$, e resolvendo em ordem a z obtemos

$$z = x_m - \frac{f(x_m)}{f'(x_m)} - \frac{f''(\xi_m)}{2f'(x_m)}(z - x_m)^2 \tag{2.24}$$

com $\xi_m \in \text{int}(z, x_m)$. Desprezando o último termo, vamos obter a aproximação x_{m+1} dada por (2.23). Então, subtraindo (2.23) a (2.24), chegamos à seguinte fórmula dos erros

$$e_{m+1} = -\frac{f''(\xi_m)}{2f'(x_m)} e_m^2 \tag{2.25}$$

com $\xi_m \in \text{int}(z, x_m)$, em que se pode notar a semelhança com a fórmula dos erros do método da secante (2.19). A fórmula (2.25) vai ser usada para concluir que o método de Newton, sob certas condições, tem uma convergência quadrática, $p = 2$.

Teorema 2.5 *Seja f uma função duas vezes continuamente diferenciável numa vizinhança de um zero z e tal que $f'(z) \neq 0$. Então se o valor inicial x_0 é escolhido suficientemente perto de z, a sucessão $\{x_m\}$ definida por (2.23) converge para z. Além disso,*

$$\lim_{m\to\infty} \frac{z - x_{m+1}}{(z - x_m)^2} = -\frac{f''(z)}{2f'(z)} \tag{2.26}$$

e portanto o método de Newton tem convergência quadrática.

Demonstração: Escolhemos uma vizinhança $I = [z - \rho, z + \rho]$ ($\rho > 0$) de z suficientemente pequena de modo que

$$K = \frac{\max_{x\in I} |f''(x)|}{2 \min_{x\in I} |f'(x)|}$$

existe e

$$K\rho \leq \delta < 1 \tag{2.27}$$

Então, atendo a (2.25), para todo o $x_m \in I$ temos que

$$|e_{m+1}| \leq K e_m^2 \tag{2.28}$$

Notando que $|e_m| \leq \rho$ quando $x_m \in I$, de (2.27) e (2.28) temos que

$$|e_{m+1}| \leq \delta |e_m| \tag{2.29}$$

e

$$|e_{m+1}| < \rho$$

e portanto $x_{m+1} \in I$. Concluimos por indução que se $x_0 \in I$ e a condição (2.27) é satisfeita então: para qualquer $m \geq 0$, $x_m \in I$ e (2.29) é verificada. Se utilizarmos (2.29) m vezes obtemos

$$|e_m| \leq \delta^m |e_0|$$

donde concluimos que $\lim_{m\to\infty} |e_m| = 0$, visto que por hipótese $\delta < 1$. Em (2.25) ξ_m está entre z e x_m e portanto $\xi_m \to z$ quando $m \to \infty$. Logo

$$\lim_{m\to\infty} \frac{z - x_{m+1}}{(z - x_m)^2} = -\lim_{m\to\infty} \frac{f''(\xi_m)}{2f'(x_m)} = -\frac{f''(z)}{2f'(z)}$$

e portanto (2.26) fica provada. ■

Teorema 2.6 *Se $f \in C^k([a,b])$, $z \in (a,b)$ e $f(z) = f'(z) = \cdots = f^{(k-1)}(z) = 0$ e $f^{(k)}(z) \neq 0$, i.e., se f tiver um zero de multiplicidade k, então o método de Newton converge linearmente desde que $x_0 \in [a,b]$ esteja suficientemente perto de z.* ■

Apresentamos a seguir um algoritmo e um exemplo para o método em estudo.

Algoritmo 2.5 *(Método de Newton)*

NOTA: *Condição suficiente de convergência do método:* $f \in C^2([a,b])$, tem um zero z em $[a,b]$ e $|z - x_0| < \min_{x\in[a,b]} |f'(x)| / \max_{x\in[a,b]} |f''(x)|$ (no caso de $f'(z) = 0$ e $f''(z) \neq 0$ o método converge linearmente desde que o intervalo $[a,b]$ seja suficientemente pequeno).

INICIALIZAÇÃO:
$x_0 \in [a,b]$

CICLO: PARA $m = 0, 1, ...$ FAZER

$$x_{m+1} = x_m - \frac{f(x_m)}{f'(x_m)}$$

SE $|x_{m+1} - x_m| \leq \epsilon$
ENTÃO fazer $raiz = x_{m+1}$, SAÍDA
FIM DO CICLO.

Exemplo 2.5 *Determinar, pelo método de Newton (Algoritmo 2.5), a raiz positiva da equação $f(x) = x^2 - 2 = 0$ com um erro inferior a $\epsilon = 5E-6$, usando o computador em precisão simples. Tomar $a = 0$, $b = 2$. Considerar o caso $x_0 = \frac{a+b}{2}$.*

Tabela 2.5: Exemplo do método de Newton para $x^2 - 2 = 0$

m	x_m	e_m	e_m/e_{m-1}^2
0	1.000000	0.414214	
1	1.500000	−0.085786	−0.5000
2	1.416667	−0.002453	−0.3333
3	1.414216	−0.000002	−0.3526
4	1.414214	$2.42E-8$	5377

Dos resultados obtidos, que apresentamos na Tabela 2.5, depreendemos que o método de Newton é de convergência quadrática [4] e portanto converge mais rapidamente que os métodos anteriores. ■

[4] A sucessão e_m/e_{m-1}^2 converge para $-0.5f''(z)/f'(z) = -0.5/\sqrt{2} = -0.3536$, contudo o último valor desta sucessão na tabela difere substancialmente daquele limite devido a influência dos erros de arredondamento.

2.3.3 Critério de convergência para os métodos da secante e de Newton

Dado que os Teoremas 2.2 e 2.5 são de difícil aplicação, vamos apresentar um conjunto de *condições suficientes* para a convergência dos métodos da secante e de Newton que são de verificação mais simples.

Teorema 2.7 *Seja $x_0 \in [a,b]$ e f uma função duas vezes continuamente diferenciável no intervalo $[a,b]$ que satisfaz as seguintes condições:*

1. $f(a)f(b) < 0$
2. $f'(x) \neq 0,\ x \in [a,b]$
3. $f''(x)$ *não muda de sinal em* $[a,b]$, *i.e.,* $f''(x) \geq 0$ *ou* $f''(x) \leq 0$ *com* $x \in [a,b]$
4. (a) $|f(a)/f'(a)| < |b-a|$ *e* $|f(b)/f'(b)| < |b-a|$

   ou

   (b) $f(x_0)f''(x) \geq 0$ *com* $x \in [a,b]$.

Então o método de Newton converge para a única solução z de $f(x) = 0$ em $[a,b]$. ■

Teorema 2.8 *Seja $x_{-1}, x_0 \in [a,b]$ com $x_{-1} \neq x_0$ e f uma função que satisfaz às condições do Teorema 2.7 com a condição 4b substituida por*

$$f(x_{-1})f''(x) \geq 0 \text{ e } f(x_0)f''(x) \geq 0 \text{ com } x \in [a,b]$$

Então o método da secante converge para a única solução z de $f(x) = 0$ em $[a,b]$. ■

Figura 2.6: Convergência dos métodos de Newton e da secante

Vamos fazer alguns comentários sobre estas *condições suficientes* de convergência dos métodos de Newton e da secante. As condições 1 e 2 garantem-nos que existe uma e uma só solução de $f(x) = 0$ em $[a, b]$. A condição 3 implica que f é cncava ou convexa. Qualquer das duas condições 4, em conjunto com as anteriores, garantem que as iteradas x_m , $m = 1, 2, \ldots$ pertencem ao intervalo $[a, b]$; no caso de ser satifeita a condição 4b as iteradas localizam-se todas de um mesmo lado do zero z, constituindo um sucessão monótona convergente para z (ver Figura 2.6).

2.4 Iteração do ponto fixo

Até agora temos estudados métodos de resolução de equações não lineares escritas na forma

$$f(x) = 0 \tag{2.30}$$

Para fazermos um estudo geral desses métodos vamos transformar a equação (2.30) numa equação da forma

$$x = g(x) \tag{2.31}$$

em que qualquer solução z de (2.31) é uma solução de (2.30) e vice-versa. Podemos, p.ex., escolher para g qualquer função da forma

$$g(x) = x - \phi(x)f(x)$$

em que $\phi(x) \neq 0$ é uma função limitada num intervalo que contenha as soluções de (2.30). Se z for solução de (2.30) será também solução de (2.31) e por conseguinte podemos escrever

$$z = g(z) \tag{2.32}$$

Porque a aplicação g transforma z em si mesmo, z é denominado *ponto fixo* de g e o método iterativo baseado em (2.31), que a seguir apresentamos, é denominado *método do ponto fixo*. Consiste em construir a sucessão $x_0, x_1, x_2, \ldots$ tal que

$$x_{m+1} = g(x_m) \qquad m = 0, 1, \ldots \tag{2.33}$$

sendo x_0 uma aproximação inicial de z. Vemos que o método de Newton pode considerar-se um método do ponto fixo bastando para isso escolher

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

Também o método da falsa posição, quando $f''(x) \neq 0$ num intervalo (a, b) que contenha z, pode considerar-se um método do ponto fixo; recorrendo a (2.13) vemos que neste caso é

$$g(x) = x - f(x)\frac{x - w_0}{f(x) - f(w_0)}$$

em que $w_0 = a$ se $f''(x)f(a) > 0$ e $w_0 = b$ se $f''(x)f(b) > 0$ com $x \in (a, b)$.

Exemplo 2.6 *Considerar a equação $x^2 - a = 0$ com $a > 0$. Utilizando o método do ponto fixo e usando o computador em precisão simples, efectuar 4 iterações com as seguintes escolhas para a função iteradora g e para o valor inicial x_0:*

1. $g(x) = x - (x^2 - a)$ *e* $x_0 = 1.5$
2. $g(x) = x - 0.6(x^2 - a)$ *e* $x_0 = 4$
3. $g(x) = x - 0.6(x^2 - a)$ *e* $x_0 = 2$
4. $g(x) = (x + a/x)/2$ *e* $x_0 = 2$

Tomar $a = 2$ *(neste caso* $z = \sqrt{2} = 1.414214$*).*

Dos resultados obtidos, que apresentamos na Tabela 2.6, depreendemos que a convergência e sua rapidez dependem da função iteradora g e do valor inicial x_0 escolhidos (nos casos 1 e 2 o método diverge, no caso 3 o método converge lentamente, no caso 4 o método converge rapidamente). ■

Tabela 2.6: Exemplo do método do ponto fixo para $x^2 - 2 = 0$

	Caso1	Caso2	Caso3	Caso4
m	x_m	x_m	x_m	x_m
0	1.500000	4.000000	2.000000	2.000000
1	1.250000	−4.400001	0.800000	1.500000
2	1.687500	−14.81600	1.616000	1.416667
3	0.839844	−145.3244	1.249126	1.414216
4	2.134506	−12815.63	1.512936	1.414214

2.4.1 Convergência do método iterativo do ponto fixo

Vamos a seguir apresentar alguns teoremas que nos esclarecem sobre a convergência do método iterativo do ponto fixo. Antes porém vamos introduzir a seguinte definição

Definição 2.2 *Uma função* g *diz-se Lipschitziana no intervalo* $I = [a, b]$ *se existir uma constante* $L > 0$ *tal que*

$$|g(y) - g(x)| \leq L|y - x|, \qquad \forall y, x \in I \tag{2.34}$$

No caso em que $L < 1$ *a função diz-se contractiva.*

Nota-se que uma função Lipschitziana é contínua. Por outro lado, se g for continuamente diferenciável em I, pelo teorema de Lagrange temos que, para $y, x \in I$,

$$g(y) - g(x) = g'(\xi)(y - x) \qquad \xi \in \text{int}(y, x)$$

e portanto se $|g'(\xi)| < 1$ para todo o $\xi \in I$, a função g é contractiva. Neste caso, a constante de Lipschitz L em (2.34) pode ser dada por

$$L = \max_{x \in I} |g'(x)| \tag{2.35}$$

O teorema seguinte é uma das versões mais simples do *teorema* conhecido na literatura pelo *do ponto fixo*.

Teorema 2.9 *Se existir um intervalo $I = [a, b]$ no qual a função g é contractiva e $g(I) \subset I$, i.e., $a \leq g(x) \leq b$, então:*

1. *g possui pelo menos um ponto fixo z em I*
2. *z é o único ponto fixo em I*
3. *para qualquer x_0 em I a sucessão $\{x_m\}$ gerada por (2.33) converge para z*
4. $|z - x_m| \leq \frac{1}{1-L}|x_{m+1} - x_m| \quad , \quad |z - x_{m+1}| \leq \frac{L}{1-L}|x_{m+1} - x_m|$
5. $|z - x_m| \leq L^m|z - x_0| \leq \frac{L^m}{1-L}|x_1 - x_0|$

Demonstração: (**1**) Se $g(a) = a$ ou $g(b) = b$ então g tem um ponto fixo em I. Suponhamos então que $g(a) \neq a$ e $g(b) \neq b$ e consideremos a função contínua $h(x) = g(x) - x$. Por termos suposto que $g(I) \subset I$, tem-se que $h(a) > 0$ e $h(b) < 0$ e portanto do teorema do valor intermédio concluimos que h tem que possuir um zero em (a, b). Fica provado que g possui pelo menos um ponto fixo z em I. (**2**) Suponhamos que w é um ponto fixo em I. Então por ser g contractiva de (2.34) temos

$$|z - w| = |g(z) - g(w)| \leq L|z - w|$$

$$(1 - L)|z - w| \leq 0$$

Como $1 - L > 0$, conclui-se que $w = z$ (unicidade). (**3**) Começamos por notar que se $x_m \in I$ então $x_{m+1} = g(x_m) \in I$, visto que $g(I) \subset I$. Para mostrar que $\lim_{m\to\infty} x_m = z$ atendemos a que $|z - x_{m+1}| = |g(z) - g(x_m)|$, por (2.32) e (2.33), $|g(z) - g(x_m)| \leq L|z - x_m|$, por (2.34). Então se $x_0 \in I$, temos

$$|z - x_{m+1}| \leq L|z - x_m| \qquad x_m \in I \tag{2.36}$$

De (2.36) temos $|z - x_m| \leq L|z - x_{m-1}|$ e por indução segue-se que

$$|z - x_m| \leq L^m|z - x_0| \qquad m = 0, 1, \ldots \tag{2.37}$$

Quando $m \to \infty$, $L^m \to 0$ ($L < 1$ por hipótese) e portanto $\lim_{m\to\infty} x_m = z$. (**4**) De (2.36) temos que

$$|z - x_m| - |x_{m+1} - x_m| \leq |(z - x_m) - (x_{m+1} - x_m)| = |z - x_{m+1}| \leq L|z - x_m|$$

Logo temos que $(1 - L)|z - x_m| \leq |x_{m+1} - x_m|$ e portanto, como $L < 1$, obtemos a primeira desigualdade

$$|z - x_m| \leq \frac{1}{1-L}|x_{m+1} - x_m| \tag{2.38}$$

Utilizando (2.38) e novamente (2.36) temos a outra desigualdade pretendida

$$|z - x_{m+1}| \leq \frac{L}{1-L}|x_{m+1} - x_m| \tag{2.39}$$

(**5**) A primeira desigualdade coincide com (2.37). Tomando $m = 0$ em (2.38) e recorrendo a (2.37) temos a segunda desigualdade

$$|z - x_m| \leq \frac{L^m}{1-L}|x_1 - x_0| \tag{2.40}$$

■

As desigualdades (2.37) e (2.40) permitem obter majorações (apriori) do erro $e_m = z - x_m$ da iterada m sem efectuar as m iterações. As desigualdades (2.38) e (2.39) permitem obter majorações (aposteriori) dos erros e_m e e_{m+1} depois de ter efectuado $m+1$ iterações.

Na prática pode ser difícil verificar se g satisfaz ou não a condição (2.34).Por isso, quando g for diferenciável, pode-se substituir esta condição por outra mais forte, obtendo-se o seguinte:

Teorema 2.10 *Seja g continuamente diferenciável em $I = [a, b]$ com $|g'| < 1$ em I e $g(I) \subset I$, então são verificadas todas as proposições do Teorema 2.9 e tem-se ainda que:*

> $\lim_{m\to\infty} e_{m+1}/e_m = g'(z)$, *e portanto no caso de $g'(z) \neq 0$ a convergência é linear sendo o coeficiente assintótico de convergência $K_\infty = |g'(z)|$. (Convergência supralinear só é possível se $g'(z) = 0$. Este caso será tratado à frente, Teorema 2.14)*

Demonstração: Visto que a primeira hipótese do teorema implica, como já referimos, que g seja contractivo e que L seja dado por (2.35), concluimos que o Teorema 2.9 é aplicável. Por outro lado basta subtrair (2.33) a (2.32) e recorrer mais uma vez ao teorema de Lagrange para termos

$$e_{m+1} = g'(\xi_m)e_m \qquad \xi_m \in \mathrm{int}(z, x_m) \tag{2.41}$$

em que $e_m = z - x_m$. Por conseguinte, no caso de convergência, temos

$$\lim_{m\to\infty} e_{m+1}/e_m = g'(z) \tag{2.42}$$

visto que g' é contínua. Logo $|g'(z)|$ é o coeficiente assintótico de convergência. ■

Vemos que se $|g'(z)| > 1$ não há convergência visto que para x_m suficientemente perto de z tem-se $|e_{m+1}| > |e_m|$, atendendo a (2.41). Portanto somos conduzidos ao seguinte:

Teorema 2.11 *Seja g continuamente diferenciável em $I = [a, b]$ com $|g'(x)| > 1$ para $x \in I$. Admitindo a existência de um ponto fixo z interior a I, então $|g'(z)| > 1$ e a sucessão $\{x_m\}$ definida por (2.33) não poderá convergir para z (excluindo o caso de ser $x_m = z$ a partir de uma certa iteração).* ■

Nas hipóteses do Teorema 2.10, se $0 < g'(x) < 1$ para $x \in I$, então a convergência é *monótona*, *i.e.*, $x_0 < x_1 < \cdots < z$ ou $x_0 > x_1 > \cdots > z$ com $x_0 \in I$ e se $-1 < g'(x) < 0$ para $x \in I$, então a convergência é *alternada*, *i.e.*, as sucessivas iteradas $x_0, x_1, \ldots$ vão estando alternadamente à esquerda e à direita da raiz z.

Para visualizar melhor o problema da convergência do método do ponto fixo é útil uma representação gráfica. O ponto fixo z é a abcissa do ponto de intersecção da recta $y = x$ com o gráfico de $y = g(x)$. Para passar da iterada x_0 para a iterada x_1 parte-se do ponto $(x_0, g(x_0))$ pertencente ao gráfico de g. Como $x_1 = g(x_0)$, traçamos um segmento horizontal partindo daquele ponto até ao ponto (x_1, x_1) pertencente à recta $y = x$. A abcissa deste ponto é a iterada x_1. Para obter x_2 traça-se um segmento vertical partindo de (x_1, x_1) até ao ponto (x_1, x_2) pertencente ao gráfico de g e procede-se como anteriormente. Estão representados na Figura 2.7 os vários casos de convergência ou divergência considerados atrás.

Figura 2.7: Casos de convergência e divergência do método do ponto fixo

Para verificar se a condição $g(I) \subset I$ é satisfeita, começamos por ver se $g(a) \in [a, b]$ e $g(b) \in [a, b]$. Se estas duas últimas condições forem satisfeitas e g

for monótona (e portanto g' não mudar de sinal em I) então $g(I) \subset I$. Se g tiver máximos ou mínimos localizados em pontos t_i ($g'(t_i) = 0$) então deverá verificar-se $g(t_i) \in I$ para todos esses extremantes. A condição $|g'| < 1$ em I (g contractiva em I) é equivalente a $g'(I) \subset [-1, 1]$. Portanto para verificar se esta condição é satisfeita, começamos por ver se $g'(a) \in [-1, 1]$ e $g'(b) \in [-1, 1]$. Se estas duas últimas condições forem satisfeitas e g' for monótona (e portanto g'' não mudar de sinal em I) então $|g'| < 1$ em I. Se g' tiver máximos ou mínimos localizados em pontos t_i ($g''(t_i) = 0$) então deverá verificar-se $|g'(t_i)| < 1$ para todos esses extremantes.

Se $x_0 \in I$, em que I é um intervalo para o qual g satisfaz às condições do Teorema 2.10, então podemos garantir a convergência do método. É evidente que para haver convergência basta a existência do referido intervalo, mesmo que não saibamos qual é. O seguinte teorema tem um carácter local, *i.e.*, garante convergência da sucessão se x_0 estiver numa certa vizinhança de z, mas não indica explicitamente um intervalo onde escolher x_0.

Teorema 2.12 *Seja z ponto fixo de g e suponhamos que g é continuamente diferenciável numa vizinhança de z e que $|g'(z)| < 1$, então a sucessão gerada por (2.33) converge para z desde que x_0 esteja suficientemente perto de z*

Demonstração: Dado que $|g'(z)| < 1$ e g' é contínua numa vizinhança de z é sempre possível escolher um intervalo $I = [z - \rho, z + \rho]$ para o qual

$$\max_{x \in I} |g'(x)| = L < 1$$

Então g é contractiva. Para $x \in I$ temos que $|z - x| \leq \rho$ o que implica

$$|z - g(x)| = |g(z) - g(x)| \leq L|z - x| < \rho$$

e portanto $g(I) \subset I$. Se $x_0 \in I$ estão satisfeitas as condições do Teorema 2.9 e por conseguinte o método iterativo converge. ■

Até agora supusemos que g é contractiva. Se a existência de um ponto fixo z já foi previamente provada então somos conduzidos ao seguinte teorema.

Teorema 2.13 *Se z for um ponto fixo de g e I um intervalo tal que $z \in I$, $g(I) \subset I$ e para $x \in I$*

$$|g(z) - g(x)| \leq L|z - x| \tag{2.43}$$

com $L < 1$, então são verificadas as proposições 2, 3, 4 e 5 do Teorema 2.9.

Demonstração: A demonstração é igual ao do Teorema 2.9. ■

Notamos que a desigualdade (2.43) não implica que g seja contractiva em I visto que um dos pontos deve ser z (comparar com (2.34)). Se escolhermos para o intervalo $I = [z - \rho, z + \rho]$, com $\rho > 0$, então pode-se demonstrar que a condição (2.43) implica que $g(I) \subset I$. Com esta escolha para I, deste Teorema 2.13 tira-se como corolário o Teorema 2.12.

Um exemplo de uma função iteradora g que está nas condições do Teorema 2.13 é $g(x) = x^2 + 0.125$ com $I = [0, 0.8]$. Neste caso é fácil ver que em I a função iteradora g tem um único ponto fixo $z = (1 - \sqrt{2}/2)/2 = 0.1464466$ onde $g'(z) = 0.2928932$. Contudo, $g'(x) > 1$ para $x > 0.5$ e portanto g não é contractiva em I.

Como $g(0) = 0.125 > 0$, $g(0.8) = 0.765 < 0.8$ e g é crescente então $g(I) \subset I$. Pode-se mostrar que (2.43) é verificada com

$$L = \frac{g(0.8) - g(z)}{0.8 - z} = 0.9464466$$

Como se vê, nesse exemplo a condição $g(I) \subset I$ é satisfeita mas não a condição $|g'(x)| < 1$ para todos os valores de x do intervalo I. No entanto existe uma vizinhança de z para a qual $|g'(x)| < 1$.

Para ilustrar a importância da condição $g(I) \subset I$ consideremos o seguinte exemplo.

Exemplo 2.7 *Considerar a função iteradora* $g(x) = -0.5x^3 + 1.5x^2 - 2x + 2.8125$. *Utilizando o método do ponto fixo, efectuar 9 iterações com as seguintes escolhas para o valor inicial: (i)* $x_0 = 0.45$*; (ii)* $x_0 = 1.55$.

Os valores iniciais foram escolhidos de modo a pertencer a um intervalo $I = [.45, 1.55]$ para o qual $|g'| < 1$. De facto, pode-se verificar que $-1 < g'(x) \leq -0.5$ para $x \in (1 - \sqrt{3}/3, 1 + \sqrt{3}/3) = (0.422650, 1.577350)$. No entanto, não é satisfeita a condição $g(I) \subset I$ visto que $g(0.45) > 1.55$. Isto tem por consequência, *neste exemplo*, que a sucessão de iteradas inicializada em $x_0 = 0.45$ não converge para o ponto fixo $z = 1.5$, como se depreende da Tabela 2.7. Se tivessemos considerado o intervalo $I = [1.410, 1.574]$, então para qualquer valor inicial $x_0 \in I$ o método iterativo convergiria, visto que $g(I) \subset I$, dado que g é monotona decrescente e $g(1.41) = 1.573040$ e $g(1.574) = 1.430941$ pertencem a I. Nota-se que as condições de convergência *são suficientes* mas *não necessárias*. Nomeadamente, para a função g considerada neste exemplo, se $x_0 \in I = [1, 2]$ o método converge apesar de não estar satisfeita nenhuma das condições: $|g'| < 1$ em I e $g(I) \subset I$. ■

Nota: Dado que as condições $|g'| < 1$ em I e $g(I) \subset I$ são suficientes *mas não*

Tabela 2.7: Exemplo do método do ponto fixo

m	x_m	x_m
0	0.450000	1.550000
1	2.170687	1.454313
2	0.424938	1.538459
3	2.195117	1.465211
4	0.361446	1.529554
5	2.261962	1.473472
6	0.176651	1.522693
7	2.503250	1.479751
8	−0.637618	1.517414
9	4.827183	1.484533

necessárias para garantir a convergência de uma sucessão $\{x_m\}$ para um ponto fixo $z \in I$ (Teorema 2.10), se uma delas (ou as duas) não se verificar não podemos concluir nada, excepto se $|g'| > 1$ em todo o intervalo I (Teorema 2.11). Mas se soubermos

que existe uma vizinhança de z onde $|g'| < 1$, então é sempre possível escolher um novo intervalo I_0 para o qual já são verificadas as duas condições: $|g'| < 1$ em I_0 e $g(I_0) \subset I_0$ (Teorema 2.12). Podemos dizer que neste caso o primeiro intervalo I foi mal escolhido.

De (2.42) vemos que a convergência de um método iterativo do ponto fixo será tanto mais rápida quanto menor for $|g'(z)|$. Se $g'(z) = 0$ a convergência deixa de ser linear para passar a supralinear, *i.e.*, a ordem de convergência passa a ser $p > 1$ (ver Definição 2.1). Vejamos um teorema que considere este caso.

Teorema 2.14 *Seja z um ponto fixo de g e $g^{(p)}$ contínua numa vizinhança de z com $p \geq 2$. Seja ainda*

$$g'(z) = \cdots = g^{(p-1)}(z) = 0 \qquad g^{(p)}(z) \neq 0 \tag{2.44}$$

Então para x_0 suficientemente perto de z a sucessão $\{x_m\}$, com $x_{m+1} = g(x_m)$, converge para z e

$$\lim_{m\to\infty} e_{m+1}/e_m^p = (-1)^{p-1} g^{(p)}(z)/p! \tag{2.45}$$

e portanto a convergência é supralinear de ordem p sendo o coeficiente assintótico de convergência $K_\infty = |g^{(p)}(z)/p!|$.

Demonstração: Dado que $|g'(z)| < 1$, estamos nas condições do Teorema 2.12 e portanto existe um intervalo $I = [z - \rho, z + \rho]$ tal que para $x_0 \in I$ está garantida a convergência. Utilizemos agora a fórmula de Taylor

$$g(x_m) = g(z) + g'(z)(x_m - z) + \cdots + \frac{g^{(p-1)}(z)}{(p-1)!}(x_m - z)^{p-1} + \frac{g^{(p)}(\xi_m)}{(p)!}(x_m - z)^p$$

com $\xi_m \in \mathrm{int}(x_m, z)$. Atendendo a (2.44) temos que

$$x_{m+1} = z + \frac{g^{(p)}(\xi_m)}{(p)!}(x_m - z)^p \tag{2.46}$$

De (2.46) e atendendo que $g^{(p)}$ é contínua e que $\lim_{m\to\infty} \xi_m = z$ obtemos (2.45). e portanto a convergência é de ordem p. ■

Façamos

$$K = \max_{x\in I} |g^{(p)}(x)/p!| \tag{2.47}$$

Logo de (2.46) e (2.47) obtemos a majoração (2.5) $|e_{m+1}| \leq K|e_m|^p$.

Podemos utilizar este teorema para verificar, mais uma vez, que o método de Newton tem em geral convergência quadrática. Com efeito, para este método temos que

$$g(x) = x - \frac{f(x)}{f'(x)} \tag{2.48}$$

e portanto

$$g'(x) = \frac{f(x)f''(x)}{[f'(x)]^2}$$

$$g''(x) = \frac{f''(x)}{f'(x)} + f(x)\frac{f'''(x)f'(x) - 2[f''(x)]^2}{[f'(x)]^3}$$

Logo, se $f'(z) \neq 0$ (z zero simples de f) e $f''(z) \neq 0$, temos que

$$g'(z) = 0 \qquad g''(z) = f''(z)/f'(z) \neq 0$$

Este último resultado e (2.45) conduz a (2.26) que dá o coeficiente assintótico de convergência do método de Newton. Vamos ver a seguir que o método de Newton deixa de ter convergência quadrática quando o zero de f for múltiplo, e uma forma de voltar a obter um método quadrático para este caso.

2.4.2 Modificação do método de Newton para zeros múltiplos

Diz-se que a função f possui um zero z de multiplicidade $k > 1$ se

$$f(x) = (x - z)^k h(x) \tag{2.49}$$

com $h(z) \neq 0$ e h contínua em z. Restringiremos k a um inteiro. Se h é suficientemente diferenciável em z, então (2.49) implica que

$$f(z) = f'(z) = \cdots = f^{(k-1)}(z) = 0 \qquad f^{(k)}(z) \neq 0 \tag{2.50}$$

Por outro lado se f for k vezes continuamente diferenciável numa vizinhança de z, então (2.50) implica (2.49).

Vamos utilizar (2.49) para determinar $g'(z)$ para o caso de usarmos o método de Newton em que g é, como vimos atrás, dado por (2.48). Temos

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

que substituida em (2.48) dá

$$g(x) = x - \frac{(x - z)h(x)}{kh(x) + (x - z)h'(x)}$$

Derivando

$$g'(x) = 1 - \frac{h(x)}{kh(x) + (x - z)h'(x)} - (x - z)\frac{d}{dx}\left[\frac{h(x)}{kh(x) + (x - z)h'(x)}\right]$$

conclui-se que

$$g'(z) = 1 - \frac{1}{k} \neq 0 \qquad \textit{para} \;\; k > 1 \tag{2.51}$$

Portanto o método de Newton é um método linear (no caso de um zero múltiplo) com um coeficiente assintótico de convergência $(k - 1)/k$.

Exemplo 2.8 *Determinar, pelo método de Newton (Algoritmo 2.5), o zero duplo do polinómio* $p(x) = x^3 - 3.08x^2 + 3.1236x - 1.03968$ *com um erro inferior a* $\epsilon = 5E - 6$, *usando o computador em precisão dupla.*[5] *Escolher para aproximação inicial* $x_0 = 1$.

Tabela 2.8: Exemplo do método de Newton para zeros múltiplos

m	x_m	e_m	e_m/e_{m-1}
0	1.000000	0.140000	
1	1.107692	0.032308	0.2308
2	1.124741	0.015259	0.4723
3	1.132554	0.007446	0.4880
4	1.136319	0.003681	0.4943
5	1.138170	0.001830	0.4972
6	1.139087	0.000913	0.4986
7	1.139544	0.000456	0.4993
8	1.139772	0.000228	0.4997
9	1.139886	0.000114	0.4998
10	1.139943	0.000057	0.4999
11	1.139972	0.000028	0.5000
12	1.139986	0.000014	0.5000
13	1.139993	0.000007	0.5000
14	1.139996	0.000004	0.5000

Dos resultados obtidos, que apresentamos na Tabela 2.8, depreendemos que o método de Newton é de convergência linear com um coeficiente assintótico de convergência 1/2, o que era de prever de (2.51) visto que $z = 1.14$ é um zero duplo do polinómio p considerado (p tem um zero simples em 0.8). ■

Para melhorar o método de Newton, temos que obter uma função iteradora g para a qual $g'(z) = 0$. Inspirado na dedução de (2.51), escolhemos para nova função iteradora

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

Então, repetindo a dedução de (2.51) com este novo g, concluimos, como pretendíamos, que $g'(z) = 0$. Do Teorema 2.14 temos então que o *método de Newton modificado para um zero de multiplicidade k*

$$x_{m+1} = x_m - k\frac{f(x_m)}{f'(x_m)}$$

tem convergência quadrática.

Exemplo 2.9 *Determinar,pelo método de Newton modificado, o zero duplo do polinómio*

$$p(x) = x^3 - 3.08x^2 + 3.1236x - 1.03968$$

com um erro inferior a $\epsilon = 5E{-}6$, *usando o computador em precisão dupla.*[6] *Escolher para aproximação inicial* $x_0 = 1$.

[5]Mais tarde veremos a razão de não utilizarmos neste caso precisão simples.

[6]Mais tarde veremos a razão de não utilizarmos neste caso precisão simples.

Dos resultados obtidos, que apresentamos na Tabela 2.9, depreendemos que o método de Newton modificado é novamente de convergência quadrática.[7] ■

Tabela 2.9: Exemplo do método de Newton modificado para zeros múltiplos

m	x_m	e_m	e_m/e_{m-1}^2
0	1.000000	0.140000	
1	1.215385	-0.075385	-3.846
2	1.146271	-0.006271	-1.104
3	1.140056	-0.000056	-1.431
4	1.140000	$-4.66E-9$	-1.472
5	1.140000	$-3.97E-8$	$-1.83E9$

2.4.3 Critérios de paragem

Ao utilizar um método iterativo necessitamos de um critério de paragem, *i.e.*, decidir quando uma iterada x_{m+1} já constitui uma aproximação suficientemente boa da solução z. Este critério depende da precisão exigida previamente, que se traduz numa dada tolerância ϵ para o erro da aproximação obtida. O critério mais utilizado é terminar a iteração quando

$$|x_{m+1} - x_m| \leq \epsilon \tag{2.52}$$

A seguir vamos ver quando é que esta condição garante que $|e_{m+1}| = |z - x_{m+1}| \leq \epsilon$.

Começamos por admitir que estamos nas condições do Teorema 2.9 (do ponto fixo) e portanto que para $m = 0, 1, \ldots$ se verifica a desigualdade (2.36)

$$|e_{m+1}| \leq L|e_m|$$

com $L < 1$. Na demonstração do Teorema 2.9 (do ponto fixo) a partir desta desigualdade obtivemos a majoração (2.38)

$$|e_m| \leq \frac{1}{1-L}|x_{m+1} - x_m|$$

e a majoração (2.39)

$$|e_{m+1}| \leq \frac{L}{1-L}|x_{m+1} - x_m|$$

É fácil ver que quando for $L \leq 1/2$, se a condição (2.52) for satisfeita, $|e_m| \leq 2\epsilon$ e $|e_{m+1}| \leq \epsilon$. Portanto sempre que em cada iteração o erro da iterada vem reduzido pelo menos a metade do erro da iterada anterior, o critério de paragem (2.52) é satisfatório. Se for $L > 1/2$ então a convergência poderá ser lenta verificando-se $|e_{m+1}| \gg \epsilon$ apesar de $|x_{m+1} - x_m| \leq \epsilon$. Nota-se que o que acabamos de referir é válido para qualquer

[7] A sucessão e_m/e_{m-1}^2 converge para $-g''(z)/2 = -h'(z)/[kh(z)] = -1/[2(1.14 - 0.8)] = -1.471$, contudo o último valor desta sucessão na tabela difere substancialmente daquele limite devido à influência dos erros de arredondamento.

método iterativo verificando a desigualdade (2.36), mesmo que a sucessão $\{x_m\}$ não seja gerada por $x_m = g(x_{m-1})$.

Se imposermos as condições mais restritivas do Teorema 2.10 temos a igualdade (2.41)

$$e_{m+1} = g'(\xi_m)e_m \qquad \xi_m \in \mathrm{int}(z, x_m)$$

que, como vimos, conduz a desigualdade (2.36) com L dado por (2.35)

$$L = \max_{x \in I} |g'(x)|$$

Se escrevermos a igualdade (2.41) na forma

$$z - x_{m+1} = g'(\xi_m)(z - x_{m+1} + x_{m+1} - x_m)$$

temos

$$[1 - g'(\xi_m)](z - x_{m+1}) = g'(\xi_m)(x_{m+1} - x_m)$$

Se $g'(\xi_m) \neq 1$, então obtemos a seguinte igualdade

$$e_{m+1} = \frac{g'(\xi_m)}{1 - g'(\xi_m)}(x_{m+1} - x_m) \tag{2.53}$$

Substituindo (2.41) em (2.53) temos ainda

$$e_m = \frac{1}{1 - g'(\xi_m)}(x_{m+1} - x_m) \tag{2.54}$$

Se tomarmos $L = \max_{x \in I} |g'(x)|$, então (2.53) e (2.54) conduzem as majorações (2.39) e (2.38).

Havendo convergência, a partir de uma certa iterada, x_m estará suficientemente perto de z e $\xi_m \approx z$. Logo, supondo que $g'(z) \neq 0$ (caso da convergência linear), $g'(\xi_m)$ será praticamente constante e $g'(\xi_m) \approx g'(z)$. Então de (2.53) e (2.54) temos as seguintes estimativas para os erros

$$e_{m+1} \approx \frac{g'(z)}{1 - g'(z)}(x_{m+1} - x_m)$$

e

$$e_m \approx \frac{1}{1 - g'(z)}(x_{m+1} - x_m)$$

Dado que em geral não é conhecido o valor de $g'(z)$ vamos obter uma estimativa de $g'(z)$ a partir de 3 iteradas consecutivas. Com efeito,

$$x_{m+1} - x_m = g(x_m) - g(x_{m-1}) = g'(\eta_m)(x_m - x_{m-1})$$

com $\eta_m \in (x_m, x_{m-1})$ e portanto

$$\frac{x_{m+1} - x_m}{x_m - x_{m-1}} = g'(\eta_m) \approx g'(z) \tag{2.55}$$

que é a estimativa pretendida. Dado que a convergência neste caso é linear o coeficiente assintótico de convergência K_∞ é dado por $K_\infty = |g'(z)|$ e portanto

$$K_\infty \approx \frac{|x_{m+1} - x_m|}{|x_m - x_{m-1}|}$$

No caso de uma convergência supralinear tem-se $g'(z) = 0$. Recorrendo a (2.54) e (2.53) conclui-se que $|e_m| \approx |x_{m+1} - x_m|$ e $|e_{m+1}| \ll |x_{m+1} - x_m|$ quando x_m estiver suficientemente perto de z. É o caso do método de Newton quando z for um zero simples. Neste caso se ao fim de $m+1$ iterações $|x_{m+1} - x_m| \leq \epsilon$ então $|e_{m+1}| \ll \epsilon$ e provavelmente a iterada x_m obtida na iteração anterior já teria um erro $|e_m|$ inferior a ϵ, o que contudo só é detectado depois de determinar a iterada x_{m+1}.

Admitindo que o critério dado por (2.52) é aceitável podemos, para o método iterativo do ponto fixo, apresentar o seguinte algoritmo.

Algoritmo 2.6 *(Método iterativo do ponto fixo)*

NOTA: *Condição suficiente de convergência do método:* g é continuamente diferenciável e $|g'| < 1$ em $I = [a, b]$ e $g(I) \subset I$.

INICIALIZAÇÃO:
$x_0 \in [a, b]$

CICLO: PARA $m = 0, 1, ...$ FAZER
$x_{m+1} = g(x_m)$
SE $|x_{m+1} - x_m| \leq \epsilon$
ENTÃO fazer $raiz = x_{m+1}$, SAÍDA
FIM DO CICLO.

2.4.4 Aceleração de Aitken

A convergência do método iterativo do ponto fixo é em geral linear, o que, como pudemos constatar, pode significar que em termos práticos é frequentemente lenta. No entanto é possível acelerar a convergência por um processo devido a Aitken.

Supondo que o método converge, então para m suficientemente grande, no caso de $g'(z) \neq 0$ (convergência linear), temos que

$$z - x_{m+1} \approx g'(z)(z - x_m) \tag{2.56}$$

Se utilizarmos a aproximação de $g'(z)$ dada por (2.55) e substituirmos em (2.56) podemos obter uma aproximação a_{m+1} de z que satisfaz a relação

$$a_{m+1} - x_{m+1} = \frac{x_{m+1} - x_m}{x_m - x_{m-1}}(a_{m+1} - x_m)$$

Explicitando a_{m+1}, obtemos a *fórmula de aceleração de Aitken*

$$a_{m+1} = x_{m-1} - \frac{(x_m - x_{m-1})^2}{(x_{m+1} - x_m) - (x_m - x_{m-1})} \tag{2.57}$$

É de esperar que a_{m+1} constitua uma melhor aproximação de z do que x_{m+1}. Notemos que, para aplicar a aceleração de Aitken, necessitamos de 3 iteradas x_{m-1}, x_m, x_{m+1}. Assim, partindo de uma iterada x_{m-1} calculamos duas novas iteradas $x_m = g(x_{m-1})$ e $x_{m+1} = g(x_m)$ e depois podemos utilizar a aceleração (2.57) para determinar a_{m+1}.

Exemplo 2.10 *Considerar novamente os casos 3 e 4 do Exemplo 2.6 em que, com o valor inicial $x_0 = 2$, se pretendia determinar o zero $z = \sqrt{2} = 1.414214$ da equação $x^2 - 2 = 0$ utilizando respectivamente as funções iteradoras $g(x) = x - 0.6(x^2 - 2)$ e $g(x) = (x + 2/x)/2$. Pretende-se agora ilustrar a aceleração de Aitken.*

Dos resultados obtidos para o caso 3, que apresentamos na Tabela 2.10, vemos que a_m constitui uma melhor aproximação da solução z do que x_m (nomeadamente para $m = 7$ em que $|z - a_7| \approx |z - x_7|/27$) e que

$$(x_m - x_{m-1})/(x_{m-1} - x_{m-2})$$

tende para

$$g'(z) = 1 - 1.2\sqrt{2} = -0.6971$$

Para o caso 4, em que a sucessão converge quadraticamente para z, da Tabela 2.11 vemos que a_m constitui uma pior aproximação de z do que x_m (nomeadamente para $m = 3$ em que $|z - a_3| \approx 36|z - x_3|$) e que

$$(x_m - x_{m-1})/(x_{m-1} - x_{m-2})$$

tende para $g'(z) = 0$. ■

Tabela 2.10: Exemplo da aceleração de Aitken para uma sucessão com convergência linear

m	x_m	a_m	$z - x_m$	$z - a_m$	$\frac{x_m - x_{m-1}}{x_{m-1} - x_{m-2}}$
0	2.000000		−0.585787		
1	0.800000		0.614214		
2	1.616000	1.285714	−0.201787	0.128499	−0.6800
3	1.249126	1.362914	0.165087	0.051300	−0.4496
4	1.512936	1.402587	−0.098723	0.011627	−0.7191
5	1.339550	1.408313	0.074663	0.005901	−0.6572
6	1.462913	1.411630	−0.048700	0.002584	−0.7115
7	1.378844	1.412916	0.035369	0.001297	−0.6815

É lógico que seja agora a_{m+1}, e não x_{m+1}, o novo valor de partida para obter 2 novas iteradas. Assim somos levados ao seguinte algoritmo em que a condição de aplicabilidade será explicada a seguir.

Algoritmo 2.7 *(Método de Steffensen)*

Tabela 2.11: Exemplo da aceleração de Aitken para uma sucessão com convergência quadrática

m	x_m	a_m	$z - x_m$	$z - a_m$	$\frac{x_m - x_{m-1}}{x_{m-1} - x_{m-2}}$
0	2.000000		-0.585787		
1	1.500000		-0.085787		
2	1.416667	1.400000	-0.002453	0.014213	0.1667
3	1.414216	1.414141	-0.000002	0.000072	0.0294
4	1.414214	1.414214	$2.42E-8$	$2.42E-8$	0.0009

NOTA: *Condição de aplicabilidade do método:* g é continuamente diferenciável, tem um único ponto fixo z em $[a, b]$ e $[a, b]$ é um intervalo suficientemente pequeno de modo a que o método convirja (se $g'(z) < 1$ ou $g'(z) > 1$ o método converge pelo menos quadraticamente, se $g'(z) = 1$ o método converge linearmente; se $g'(z) = 0$ não se deve utilizar o método).

INICIALIZAÇÃO:
$x_0 \in [a, b]$

CICLO: PARA $m = 0, 1, ...$ FAZER
$t_0 = x_m$
$t_1 = g(t_0)$
$t_2 = g(t_1)$

$$x_{m+1} = t_0 - \frac{(t_1 - t_0)^2}{t_2 - 2t_1 + t_0}$$

SE $|x_{m+1} - x_m| \leq \epsilon$
ENTÃO fazer $raiz = x_{m+1}$, SAÍDA
FIM DO CICLO.

Este algoritmo ilustra o *método de Steffensen.* Este método pode considerar-se como um método do ponto fixo com a função iteradora

$$G(x) = x - \frac{(g(x) - x)^2}{g(g(x)) - 2g(x) + x}$$

Notamos que x_m, $g(x_m)$ e $g(g(x_m))$ são 3 iteradas consecutivas de um método iterativo cuja função iteradora seja g. Se g for continuamente diferenciável numa vizinhança de z, pode-se verificar que $\lim_{x \to z} G(x) = z$ e portanto podemos definir por continuidade G em z escrevendo G(z)=z. Se $g'(z) \neq 1$ então $G'(z) = 0$, e portanto a convergência do método de Steffensen é pelo menos quadrática, o que consiste de facto numa aceleração da convergência quando $g'(z) \neq 0$ (caso em que o método iterativo original converge linearmente ou não converge). Se $g'(z) = 0$, então o método original já tem convergência supra linear e pode-se demonstrar que o método de Steffensen conduz, neste caso, a um maior número de cálculos dos valores de g do que o método original, para atingir uma dada precisão.

Exemplo 2.11 *Considerar novamente os casos 1 e 3 do Exemplo 2.6 em que, com os valores iniciais respectivamente* $x_0 = 1.5$ *e* $x_0 = 2$*, se pretendia determinar o zero*

$z = \sqrt{2} = 1.414214$ da equação $x^2 - 2 = 0$ utilizando respectivamente as funções iteradoras $g(x) = x - (x^2 - 2)$ e $g(x) = x - 0.6(x^2 - 2)$. Pretende-se agora, utilizando o método de Steffensen (Algoritmo 2.7) e usando o computador em precisão simples, determinar o zero com um erro inferior a $\epsilon = 5E - 6$.

Tabela 2.12: Exemplo do método de Steffensen num caso em que $|g'(z)| > 1$

m	x_m	e_m	e_m/e_{m-1}^2
0	1.500000	-0.085786	
1	1.409091	0.005123	0.6961
2	1.414197	0.000017	0.6414
3	1.414214	$2.42E-8$	85.421
4	1.414214	$2.42E-8$	$4.13E7$

Tabela 2.13: Exemplo do método de Steffensen num caso em que $|g'(z)| < 1$

m	x_m	e_m	e_m/e_{m-1}^2
0	2.000000	-0.585787	
1	1.285714	0.128499	0.3745
2	1.410531	0.003683	0.2230
3	1.414210	0.000003	0.2479
4	1.414214	$2.42E-8$	2141.2

Dos resultados obtidos, que apresentamos nas Tabelas 2.12 e 2.13, depreendemos que o método de Steffensen converge quadraticamente.[8] ■

2.5 Zeros de polinómios

A determinação dos zeros de polinómios é um problema frequente nas aplicações, pelo que foram desenvolvidos algoritmos especialmente adaptados a este tipo de funções. Nesta secção consideraremos apenas polinómios reais, *i.e.*,

$$p(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0 \tag{2.58}$$

onde os coeficientes $a_0, a_1, \ldots, a_n$ são reais. Como sabemos, um polinómio de grau n tem exactamente n zeros, contando um zero de multiplicidade k como k zeros. Além disso, um polinómio real pode ter zeros complexos, os quais aparecem neste caso sob a forma de pares conjugados.

[8] Apesar das sucessões e_m/e_{m-1}^2 serem convergentes, os últimos valores destas sucessões nas tabelas diferem substancialmente dos restantes devido a influência dos erros de arredondamento.

2.5.1 Localização dos zeros

Tal como sucede com as demais funções, um bom conhecimento da localização dos zeros é uma vantagem apreciável no sentido de assegurar uma rápida convergência dos diversos métodos iterativos, sendo por vezes uma condição determinante. É possível, dada a forma particular dos polinómios, obter sem grande dificuldade alguma informação relativamente à natureza e localização dos zeros deste tipo de funções.

Para tal consideremos um polinómio escrito sob a forma (2.58) e designemos por v o número de variações de sinal dos respectivos coeficientes não nulos e por n_+ o número de zeros reais positivos.

Regra de Descartes (zeros positivos) *O número n_+ de zeros reais positivos de um polinómio real p não excede o número v de variações de sinal dos seus coeficientes não nulos e o valor $v - n_+$ é par.*

Consideremos agora o polinómio $q(x) = p(-x)$. Como é fácil de ver, se z for um zero de p, então

$$0 = p(z) = p(-(-z)) = q(-z)$$

pelo que concluimos que $-z$ é um zero de q. Daqui resulta a regra de Descartes para zeros negativos.

Regra de Descartes (zeros negativos) *O número n_- de zeros reais negativos de um polinómio real p não excede o número v de variações de sinal dos coeficientes não nulos do polinómio $q(x) = p(-x)$ e o valor $v - n_-$ é par.*

Ilustremos a aplicação desta regra.

Exemplo 2.12 *Utilizando a regra de Descartes, determinar a natureza dos zeros do polinómio*

$$p(x) = 8x^3 + 4x^2 - 34x + 15$$

Notando que os sinais dos coeficientes deste polinómio são + + - + , o número v de variações de sinal é $v = 2$. Então, pela regra de Descartes, temos que

$$n_+ \leq 2 \qquad e \qquad v - n_+ = par$$

pelo que $n_+ = 2$ ou $n_+ = 0$. O número v de variações de sinal do polinómio

$$q(x) = -8x^3 + 4x^2 + 34x + 15$$

é $v = 1$, donde se tira que $n_- = 1$. Resumindo, podemos dizer que o polinómio p tem em alternativa:

1. 2 zeros positivos e 1 zero negativo
2. 2 zeros complexos conjugados e 1 zero negativo ■

A regra de Descartes só nos fornece uma indicação sobre o número de zeros reais nos intervalos $(-\infty, 0)$ e $(0, \infty)$. Para obter um intervalo que contenha todos os

zeros reais podemos fazer uso da seguinte regra:

Regra do máximo *Todos os zeros z, reais ou complexos, de um polinómio p, escrito sob a forma (2.58), verificam a seguinte desigualdade $|z| < r$, em que*

$$r = 1 + \max_{0 \le k \le n-1} |\frac{a_k}{a_n}|$$

Em particular, se o polinómio tiver zeros reais, eles estarão contidos no intervalo $(-r, r)$.

Exemplo 2.13 *Aplicar a regra do máximo ao polinómio*

$$p(x) = 8x^3 + 4x^2 - \mathbf{34}x + 15$$

Neste caso temos

$$r = 1 + |-34/8| = 5.25$$

Portanto, os zeros reais de $p(x) = 0$, se existirem, situam-se no intervalo (-5.25,5.25).
■

Vamos a seguir considerar uma regra que permite obter uma indicação sobre o número de zeros reais num dado intervalo (a, b).

Regra de Budan *Sejam v_a e v_b os números de variação de sinal das sucessões*

$$p(a), p'(a), p''(a), \ldots, p^{(n)}(a)$$

$$p(b), p'(b), p''(b), \ldots, p^{(n)}(b)$$

respectivamente. Então o número $n(a, b)$ de zeros reais do polinómio p no intervalo (a, b) não excede o valor $v_a - v_b$ e $(v_a - v_b) - n(a, b)$ é par.

Exemplo 2.14 *Utilizando a regra de Budan, localiza os zeros do polinómio*

$$p(x) = 8x^3 + 4x^2 - 34x + 15$$

Começamos por determinar as sucessivas derivadas do polinómio p.

$$\begin{aligned} p(x) &= 8x^3 + 4x^2 - 34x + 15 \\ p'(x) &= 24x^2 + 8x - 34 \\ p''(x) &= 48x + 8 \\ p'''(x) &= 48 \end{aligned}$$

Escolhemos para valores de a e b sucessivamente os inteiros entre -3 e 2. Então podemos construir o seguinte quadro:

x	-3	-2	-1	0	1	2
$p(x)$	$-$	$+$	$+$	$+$	$-$	$+$
$p'(x)$	$+$	$+$	$-$	$-$	$-$	$+$
$p''(x)$	$-$	$-$	$-$	$+$	$+$	$+$
$p'''(x)$	$+$	$+$	$+$	$+$	$+$	$+$
v_x	3	2	2	2	1	0
$v_a - v_b$	1			1	1	

Notamos que $v_{-3} = n$ e $v_2 = 0$, o que nos indica que todas as raízes reais estão no intervalo $[-3, 2]$. Do quadro conclui-se que existe 1 raiz real em cada um dos intervalos $[-3, -2]$, $[0, 1]$ e $[1, 2]$. De facto este polinómio tem zeros em -2.5,0.5 e 1.5. De notar, que se existisse um par de raízes complexas ou uma raiz dupla, haveria sempre um intervalo $[a, b]$ de largura arbitrariamente pequena para o qual $v_a - v_b$ seria par. ■

2.5.2 Método de Newton para equações algébricas

Consideremos a utilização do método de Newton na resolução da equação algébrica

$$p(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0 = 0 \qquad a_n \neq 0 \tag{2.59}$$

Com esta finalidade vamos apresentar formas eficientes de calcular $p(x)$ e $p'(x)$. O polinómio $p(x)$ pode escrever-se na seguinte forma

$$p(x) = ((\cdots((a_n x + a_{n-1})x + a_{n-2})x + \cdots + a_2)x + a_1)x + a_0 \tag{2.60}$$

conhecida por *forma de Horner*. Com esta forma são necessárias n adições e n multiplicações, o que constitui uma nítida melhoria se compararmos com as n adições e $2n - 1$ multiplicações que são necessárias na forma anterior (2.59).

Definimos, agora, os seguintes coeficientes:

$$\begin{aligned} b_n &= a_n \\ b_i &= b_{i+1}x + a_i \qquad i = n-1, n-2, \ldots, 0 \end{aligned}$$

Da forma de Horner (2.60), é fácil ver que

$$p(x) = b_0$$

Introduzindo o polinómio

$$q(t) = b_n t^{n-1} + b_{n-1}t^{n-2} + \cdots + b_2 t + b_1$$

temos que

$$\begin{aligned} (t-x)q(t) + b_0 &= (t-x)(b_n t^{n-1} + b_{n-1}t^{n-2} + \cdots + b_2 t + b_1) + b_0 \\ &= b_n t^n + (b_{n-1} - b_n x)t^{n-1} + \cdots + (b_1 - b_2 x)t + (b_0 - b_1 x) \\ &= a_n t^n + a_{n-1}t^{n-1} + \cdots + a_1 t + a_0 \end{aligned}$$

Logo

$$p(t) = (t-x)q(t) + b_0 \tag{2.61}$$

onde $q(t)$ é o quociente e b_0 o resto da divisão de $p(t)$ por $t - x$. Se x é um zero de $p(t)$, então $b_0 = 0$ e $p(t) = (t-x)q(t)$.

A relação (2.61) permite-nos obter $p'(x)$. Com efeito derivando em ordem a t, temos que

$$\begin{aligned} p'(t) &= (t-x)q'(t) + q(t) \\ p'(x) &= q(x) \end{aligned}$$

Para calcular $p'(x)$ basta portanto aplicar a forma de Horner a $q(x)$.

Estamos agora em condições de escrever o seguinte:

Algoritmo 2.8 *(Método de Newton adaptado à polinómios; método de Bierge-Vieta)*

CICLO: PARA $m = 0, 1, ...$ FAZER
INICIALIZAÇÃO:
$x = x_m$, $b_n = a_n$, $c_n = b_n$
CICLO: PARA $i = n - 1$ ATÉ 1 FAZER
$b_i = b_{i+1}x + a_i$
$c_i = c_{i+1}x + b_i$
FIM DO CICLO i.
$b_0 = b_1 x + a_0$
FIM DO CÁLCULO DOS POLINÓMIOS:
$p(x_m) = b_0$, $p'(x_m) = c_1$
$x_{m+1} = x_m - b_0/c_1$
SE $|x_{m+1} - x_m| \leq \epsilon$
ENTÃO fazer $raiz = x_{m+1}$, SAÍDA
FIM DO CICLO m.

Podemos visualizar o ciclo que determina os valores de $p(x_m)$ e $p'(x_m)$ com o seguinte quadro

$$
\begin{array}{c|ccccccc}
 & a_n & a_{n-1} & a_{n-2} & \cdots & a_2 & a_1 & a_0 \\
x & & b_n x & b_{n-1}x & \cdots & b_3 x & b_2 x & b_1 x \\
\hline
 & b_n & b_{n-1} & b_{n-2} & \cdots & b_2 & b_1 & \mathbf{b_0} \\
x & & c_n x & c_{n-1}x & \cdots & c_3 x & c_2 x & \\
\hline
 & c_n & c_{n-1} & c_{n-2} & \cdots & c_2 & \mathbf{c_1} &
\end{array}
$$

Exemplo 2.15 *Aplique o método de Newton ao polinómio*

$$p(x) = 3x^5 + 3x^4 - 4x^3 + 2x - 120$$

efectuando só uma iteração; $x_0 = 2$.

$$
\begin{array}{c|cccccc}
 & 3 & 3 & -4 & 0 & 2 & -120 \\
2 & & 6 & 18 & 28 & 56 & 116 \\
\hline
 & 3 & 9 & 14 & 28 & 58 & \mathbf{-4} \\
2 & & 6 & 30 & 88 & 232 & \\
\hline
 & 3 & 15 & 44 & 116 & \mathbf{290} &
\end{array}
$$

$$x_1 = 2 - (-4/290) = 2.013793$$

■

2.6 Precisão no cálculo de zeros múltiplos

Quando determinamos um zero de uma função no computador há sempre um intervalo de incerteza relativamente a este zero; e isto agrava-se quando a raiz é multipla. Para entendermos melhor este facto, consideremos o seguinte exemplo:

Exemplo 2.16 *Calcular o polinómio* $p(x) = x^3 - 3.08x^2 + 3.1236x - 1.03968 =$ $= (x - 0.8)(x - 1.14)^2$*:*

1. *na forma (2.59) de potências decrescentes de* x*, utilizando o computador em precisão simples*
2. *na forma de Horner (2.60), utilizando o computador em precisão simples*
3. *na forma de Horner (2.60), utilizando o computador em precisão dupla.*

O polinómio é o mesmo dos Exemplos 2.8 e 2.9 que ilustravam a determinação de um zero duplo pelo método de Newton sem e com modificação. Pelos resultados apresentados na Tabela 2.14 entende-se agora porque é que na vizinhança de um zero múltiplo é conveniente utilizar a precisão dupla do computador. É de referir também que, para este caso, a forma de Horner não diminui substancialmente os erros de arredondamentos (utilizando a precisão simples), reduzindo contudo o número de operações a efectuar. ■

Tabela 2.14: Valores do polinómio $p(x) = x^3 - 3.08x^2 + 3.1236x - 1.03968$

x	$Caso1$	$Caso2$	$Caso3$
1.1390	$4.768E-7$	$4.768E-7$	$3.390E-7$
1.1392	$0.000E+0$	$3.576E-7$	$2.171E-7$
1.1394	$2.384E-7$	$1.192E-7$	$1.222E-7$
1.1396	$0.000E+0$	$0.000E+0$	$5.434E-8$
1.1398	$0.000E+0$	$1.192E-7$	$1.359E-8$
<u>1.1400</u>	$0.000E+0$	$2.384E-7$	$0.000E+0$
1.1402	$-2.384E-7$	$2.384E-7$	$1.361E-8$
1.1404	$-2.384E-7$	$1.192E-7$	$5.446E-8$
1.1406	$0.000E+0$	$2.384E-7$	$1.226E-7$
1.1408	$4.768E-7$	$3.576E-7$	$2.181E-7$
1.1410	$2.384E-7$	$3.576E-7$	$3.410E-7$

2.7 Problemas

1. Considere o polinómio de Legendre $P_3(x) = 2.5x^3 - 1.5x$.

   (a) A raiz positiva desse polinómio localiza-se em $[0.7, 1]$.

   i. Utilizando o método iterativo do ponto fixo, para determinar esta raiz, qual a função iteradora que utilizaria: $x = 5/3\ x^3$ ou $x = \sqrt[3]{0.6x}$? Justifique.

   ii. Sem efectuar iterações diga, justificando, ao fim de quantas iterações poderia garantir que a iterada tivesse 4 algarismos significativos.

iii. Verifique que, com uma determinada escolha da aproximação inicial x_0, obteria a sucessão de valores aproximados seguinte: $x_1 = 0.84343$, $x_2 = 0.796894$. Qual foi o valor escolhido para x_0?

iv. Utilizando os valores referidos em 1(a)iii, qual a aproximação de Aitken a_2?

(b) A raiz negativa desse polinómio pertence a $[-1, -0.7]$. Pretende-se determina-la pelo método de Newton para equações algébricas. Diga qual é o valor inicial x_0 (garantindo convergência), a $1^{\underline{a}}$ iterada x_1 e a sua significância.

2. Seja a equação algébrica $f(x) = x^4 - x - 1 = 0$.

(a) Justificando, utilize as regras de Descartes, Budan e do máximo para concluir sobre a existência das raízes reais da equação e obter intervalos de comprimento unitário que as contenham.

(b) Utilizando o método de Newton para equações algébricas, efectue uma iteração para obter uma aproximação da raiz negativa. Escolha para aproximação inicial um dos extremos do intervalo obtido em 2a e justifica a sua escolha.

(c) Para aplicar o método iterativo do ponto fixo podemos escrever a equação na forma: $x = x + Af(x)$. Designando por z a raiz localizada no intervalo $[1.2, 1.3]$, mostra que, para iteradas x_{k-1} e x_k localizadas neste intervalo, se tem

$$|z - x_k| < 0.21|x_k - x_{k-1}|$$

quando $A = -0.15$. Tomando $x_0 = 1.2$ calcule uma aproximação de z com 3 algarismos significativos, efectuando o menor número possível de iterações.

3. A equação $x^2 = a$ com $a > 0$ pode escrever-se na forma $x = g(x)$ com $g(x) = a/x$.

(a) Mostre que o método iterativo do ponto fixo é divergente para qualquer estimativa inicial $x_0 = \sqrt{a}$.

(b) Aplique o método de Steffensen e verifique que se obtem convergência, tomando por exemplo $a = 2$, $x_0 = 1$ e efectuando as iterações necessárias para conseguir uma aproximação de $\sqrt{2}$ com 3 algarismos significativos.

(c) Mostre que a fórmula iterativa de Steffensen é neste caso idêntica à obtida pelo método de Newton.

4. A sucessão $\{x_m\}$ definida por

$$x_{m+1} = +\sqrt{2 + x_m} \qquad m = 0, 1, \ldots$$

com $x_0 \geq 0$ converge para um certo valor z.

(a) Determine z e mostre que de facto a sucessão é convergente.

(b) Mostre que $|z - x_{m+1}| \leq |x_{m+1} - x_m|$.

5. Considere a equação algébrica $f(x) = 2x^3 - 2x^2 - 8x + 9 = 0$. Utilizando a regra de Descartes, o que pode concluir acerca do tipo de raízes desta equação?

6. Sendo $f(x) = g(x) - x$:

   (a) Mostre que o método de Steffensen reduz-se ao seguinte método iterativo

$$x_{m+1} = x_m - \frac{f^2(x_m)}{f(x_m + f(x_m)) - f(x_m)} \qquad m = 0, 1, \ldots \tag{2.62}$$

   (b) Mostre que podemos obter cada iteração em (2.62) calculando $y_m = g(x_m)$ e aplicando em seguida o método da secante.

   (c) Utilizando 6b e o formulário mostre que para o método de Steffensen temos

$$z - x_{m+1} = -\frac{1}{2}\,\frac{g''(\xi_m)g'(\mu_m)}{g'(\eta_m) - 1}\,(z - x_m)^2$$

   com

$$\eta_m \in \mathrm{int}(x_m, x_m + f(x_m)) \quad \xi_m \in \mathrm{int}(x_m, x_m + f(x_m), z) \quad \mu_m \in \mathrm{int}(z, x_m)$$

7. Pretende-se utilizar o método de Newton para equações algébricas para obter uma aproximação do zero $z \in [1, 2]$ do polinómio $x^4 + 8x^3 + 16x - 32$. Verifique se todas as *condições suficientes* de convergência estão garantidas escolhendo para valor inicial x_0 um dos extremos (qual ?) do intervalo $[1, 2]$. Efectuando uma iteração, obtenha x_1.

8. Dada a equação polinomial $p(x) = x^4 + 2x^3 + 2x - 1 = 0$:

   (a) Mostre que a equação tem uma e uma só raiz positiva e determine uma sua aproximação efectuando uma iteração pelo método de Newton para equações algébricas (considere $x_0 = 0.5$). Refira-se, sucintamente, às vantagens da aplicação, nesta questão, deste método em relação ao de Newton.

   (b) Ao pretender-se resolver a questão anterior utilizando os métodos da falsa posição e da secante tomando $x_0 = 0.1$ e $x_1 = 1$ obteve-se:

$$\begin{array}{ll} f(x_0) = -0.7979 & f(x_1) = 4 \\ x_2 = 0.2496718 & f(x_2) = -0.4656436 \\ x_3 = 0.3279103 & f(x_3) = -0.2621004 \end{array}$$

   Determine x_4 utilizando cada um daqueles métodos.

   (c) Como classifica cada um dos métodos considerados em 8b quanto à rapidez de convergência. Utilizando o método de Aitken na pesquisa da raiz, qual dos dois métodos produz uma efectiva aceleração de convergência? Justifique.

9. Considere a seguinte função $f(x) = \exp(-x) + 2x - 2$.

   (a) Quantas raízes tem a equação $f(x) = 0$? Determine um intervalo, com a amplitude de uma décima, que contenha a maior raiz.

(b) Ao pretender-se obter a raiz localizada no intervalo $[-2,-1]$ utilizando o método da falsa posição obteve-se:

$$\begin{array}{ll} f(-2) = 1.389056 & f(-1) = -1.281718 \\ x_1 = -1.479905 & f(x_1) = -0.567281 \end{array}$$

Determine a 2ª iterada x_2. Ao utilizar o método da secante, quais os valores iniciais x_{-1} e x_0 que teria que escolher de modo a que as duas primeiras iteradas x_1 e x_2 fossem as mesmas das obtidas atrás? Justifique.

10. Para obter-se uma das raízes de uma certa equação utilizou-se um método iterativo de convergência linear e obteve-se para as 4 primeiras iteradas os seguintes valores

$$x_0 = 1.11198 \quad x_1 = 1.13133 \quad x_2 = 1.13423 \quad x_3 = 1.13465$$

Obtenha uma estimativa para o erro da última iterada x_3.

11. Seja a equação $P_4(x) = 0$ em que $P_4(x)$ é o polinómio de Legendre de 4º grau ortogonal no intervalo [-1,1], dado por

$$P_4(x) = \frac{1}{8}(35x^4 - 30x^2 + 3)$$

(a) Por aplicação das regras de Descartes e Budan conclua sobre a existência de raízes reais da equação e obtenha intervalos de comprimento unitário que as contenham.

(b) A equação pode ser reescrita na forma: $x = g(x)$ com $g(x) = x + AP_4(x)$. Designando por z a raiz contida no intervalo $[0.3, 0.4]$, mostre que com $A = 0.5$ o método iterativo, baseado nesta forma, converge se tomarmos para iterada inicial x_0 um valor no intervalo $[0.3, 0.4]$ e que a convergência é monótona. Tomando $x_0 = 0.4$, efectue uma iteração e obtenha x_1. Sem efectuar mais iterações determine o número de algarismos significativos de x_1?

(c) Utilizando o método de Newton para equações algébricas, efectue uma iteração para obter uma aproximação da raiz z contida em $[0.8, 0.9]$. Escolha para aproximação inicial um dos extremos do intervalo $[0.8, 0.9]$ e justifica a sua escolha.

12. Diga para que servem os métodos da bissecção, da falsa posição e de Newton e compare-os, indicando as vantagens e inconvenientes relativos de cada um.

13. Diga para que servem os métodos da bissecção, secante e iterativo do ponto fixo e compare-os, indicando as vantagens e inconvenientes relativos de cada um.

14. Imagine que precisava de calcular a raiz quinta de um número a e que o computador não possuía o operador potenciação (**); como poderia resolver o problema, usando um dos métodos numéricos abordados nas aulas? Seja específico, indicando nomeadamente: o problema que se quer resolver; o método que vai usar para resolver esse problema (e porquê) ; as fórmulas concretas (aplicadas a este caso) que constituem o método de resolução.

15. Dada a equação paramétrica $\ln x + A = x^3$, em que A é um parâmetro real, indique o(s) valor(es) de A para o(s) qual(is) a equação tem uma e uma só raiz para esse valor de A.

16. Qual a finalidade do método de Aitken e quais as restrições da sua aplicabilidade?

17. Pretende-se determinar as raízes da equação $35x^4 - 30x^2 + 3 = 0$.

    (a) Aplique o teorema de Budan aos pontos 0, 0.8, 1 e tire conclusões.

    (b) Calcule, com 3 algarismos significativos, a menor raiz positiva z_1 aplicando o método de Newton para equações algébricas com $x_0 = 0.340$. Com esta aproximação inicial garantia convergência? Justifique.

    (c) Para o cálculo da maior raiz z_2 efectue uma iteração utilizando o método iterativo do ponto fixo com $x_0 = 0.861$ e escrevendo a equação na forma:

    $$x = x - (1/80)(35x^4 - 30x^2 + 3)$$

    Sem efectuar mais iterações mostre que pode garantir 3 algarismos significativos para o valor obtido.

18. O método de Bailly, aplicado à resolução de uma equação $f(x) = 0$, pode ser obtido através do desenvolvimento em série de Taylor da função f, considerando apenas os termos até à 2ª ordem. Com base no exposto, mostre que o algoritmo do método de Bailly se traduz na fórmula iteradora:

    $$x_{m+1} = x_m - \frac{f(x_m)}{f'(x_m) - \dfrac{f(x_m)f''(x_m)}{2f'(x_m)}}$$

    Neste método, de que ordem será de esperar que seja o erro de truncatura? Justifique.

19. Na obtenção de uma raiz quadrada num computador de aritmética binária é usual reduzir o problema à determinação de $\sqrt{a}$, com $1/4 \leq a < 1$, utilizando a sucessão

    $$x_{m+1} = (x_m + a/x_m)/2 \qquad m = 0, 1, \ldots$$

    Esta sucessão obtem-se utilizando o método de Newton. Verifique esta afirmação. Mostre que

    $$e_{m+1} = -e_m^2/(2x_m)$$

    com $e_m = \sqrt{a} - x_m$.

20. Considere a equação polinomial $f(x) = 3x^3 + 40x - 80 = 0$. Pretende-se determinar a solução $z \in I$ desta equação com $I \equiv [1, 2]$. Com esta finalidade considere-se um método iterativo cuja função iteradora seja $g(x) = -0.075x^3 + 2$. Verifique se estão garantidas todas as *condições suficientes* de convergência do método, qualquer que seja $x_0 \in I$.

21. Com um certo método iterativo de convergência linear obteve-se as seguintes iteradas: $x_{10} = 2.00300$, $x_{11} = 2.00270$ e $x_{12} = 2.00243$. Justificando, determine um majorante para $|z - x_{12}|$ em que z é o limite da sucessão $\{x_m\}$.

22. Para uma certa função f ao utilizar o método da falsa posição obteve-se os seguintes resultados:

$$f(0) = -2 \qquad f(2) = 2$$

$$\begin{array}{ll} x_1 = 1 & f(x_1) = -1 \\ x_2 = 4/3 & f(x_2) = -2/9 \end{array}$$

Determine a 3ª iterada x_3. Ao utilizar o método da secante, quais os valores iniciais x_{-1} e x_0 que teria que escolher de modo a que as duas primeiras iteradas x_1 e x_2 fossem as mesmas das obtidas atrás? Justifique.

23. Mostre que o método de Newton fornece o seguinte processo iterativo para o cálculo de $1/a$

$$x_{m+1} = x_m(2 - a\, x_m) \qquad m = 0, 1, \ldots$$

Prove que a condição *suficiente* e *necessaria* de convergência é $0 < ax_0 < 2$.

24. Considere o polinómio $f(x) = x^3 + 8x - 12$ e a equação algébrica $f(x) = 0$.

(a) A equação tem no intervalo $[0, 2]$ a sua única raiz positiva z. Verifique a veracidade desta afirmação, recorrendo as regras de Descartes e de Budan.

(b) Se se pretendesse determinar aquela raiz z pelo método de Newton para equações algébricas, qual dos extremos do referido intervalo seria mais adequado como aproximação inicial? Justifique. Com essa escolha de aproximação inicial, faça uma iteração do método.

(c) Pretende-se agora determinar aquela raiz z por um método iterativo cuja função iteradora seja $g(x) = Af(x) + x$, sendo A uma constante real. Estabelece *condições suficientes* a verificar por A de modo a garantir que, tomando como aproximação inicial x_0 qualquer valor em $[1, 2]$, o método iterativo é monotonicamente convergente para a referida raiz z e

$$|z - x_m| \leq |x_m - x_{m-1}|$$

Capítulo 3

Sistemas de equações

A resolução de sistemas de equações lineares é fundamental quer em problemas de aplicação quer em análise numérica, nomeadamente na resolução de problemas de optimização, problemas de valores próprios, sistemas de equações não lineares, aproximações de funções, equações diferenciais ordinárias e parciais, equações integrais, etc. Para resolver um sistema de equações lineares recorre-se a dois tipos de métodos: *directos* e *iterativos*. Os primeiros são de preferência utilizados na resolução de sistemas de ordem moderada e matriz densa enquanto que os métodos iterativos são mais indicados para sistemas de ordem elevada e matriz esparsa.

Neste capítulo começa-se por estudar com um certo pormenor o método de Gauss que é o principal método directo. Faz-se a seguir referência a algumas variantes deste método. Depois de introduzir algumas noções sobre normas de matrizes, completa-se o estudo dos métodos directos com uma análise de erros pondo em relevo algumas dificuldades levantadas por estes métodos. Numa segunda parte do capítulo faz-se o estudo de métodos iterativos para sistemas lineares e não lineares.

3.1 Eliminação de Gauss

A eliminação de Gauss é o nome por que é habitualmente conhecido o método de resolução de sistemas de equações lineares por sucessiva eliminação das incógnitas. A eliminação de uma incógnita numa equação é conseguida adicionando a esta, outra equação previamente multiplicada por um factor adequado. A eliminação de incógnitas no sistema $\mathbf{Ax} = \mathbf{b}$ é efectuada de forma a obter-se um sistema equivalente, $\mathbf{Ux} = \mathbf{g}$, no qual $\mathbf{U}$ é uma matriz triangular superior. Este último sistema pode ser facilmente resolvido por um processo de *substituição ascendente*. Designemos o sistema original por $\mathbf{A}^{(1)}\mathbf{x} = \mathbf{b}^{(1)}$, em que

$$\mathbf{A}^{(1)} = \left[a_{ij}^{(1)}\right] \qquad 1 \leq i, j \leq n \qquad \mathbf{x} = [x_1, \ldots, x_n] \qquad \mathbf{b}^{(1)} = \left[b_1^{(1)}, \ldots, b_n^{(1)}\right]$$

onde n é a ordem do sistema. Assim, os várias passos do método de Gauss são os seguintes:

Método de Gauss

Passo 1: Suponhamos $a_{11}^{(1)} \neq 0$. Subtraimos à equação i ($i = 2, \ldots, n$) a 1ª equação multiplicada pelo factor m_{i1}, escolhido de modo a eliminar a 1ª incógnita x_1 nas equações 2 até n. O novo sistema é definido por

$$a_{ij}^{(2)} = a_{ij}^{(1)} - m_{i1} a_{1j}^{(1)} \qquad i = 2, \ldots, n \quad j = 1, \ldots, n \tag{3.1}$$

$$b_i^{(2)} = b_i^{(1)} - m_{i1} b_1^{(1)} \qquad i = 2, \ldots, n$$

Para que $a_{i1}^{(2)} = a_{i1}^{(1)} - m_{i1} a_{11}^{(1)}$, $i = 2, \ldots, n$, sejam nulos, basta escolher os factores m_{i1}, $i = 2, \ldots, n$, tais que

$$m_{i1} = a_{i1}^{(1)} / a_{11}^{(1)} \qquad i = 2, \ldots, n \tag{3.2}$$

A primeira linha de $\mathbf{A}$ e a primeira linha de $\mathbf{b}$ não são alteradas e a primeira coluna de $\mathbf{A}^{(1)}$, por baixo da diagonal principal, passa a ser zero na nova matriz $\mathbf{A}^{(2)}$. Obtém-se assim o sistema $\mathbf{A}^{(2)}\mathbf{x} = \mathbf{b}^{(2)}$, tal que

$$\begin{bmatrix} a_{11}^{(1)} & a_{12}^{(1)} & \cdots & a_{1n}^{(1)} \\ 0 & a_{22}^{(2)} & \cdots & a_{2n}^{(2)} \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ 0 & a_{n2}^{(2)} & \cdots & a_{nn}^{(2)} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{bmatrix} = \begin{bmatrix} b_1^{(1)} \\ b_2^{(2)} \\ \cdot \\ \cdot \\ \cdot \\ b_n^{(2)} \end{bmatrix}$$

Continuando a eliminação de incógnitas, temos em geral o seguinte:

Passo k: Seja $1 \leq k \leq n-1$. Suponhamos que $\mathbf{A}^{(\mathbf{k})}\mathbf{x} = \mathbf{b}^{(\mathbf{k})}$ tenha sido construído, com $x_1, x_2, \ldots, x_{k-1}$ eliminados nos sucessivos passos $1, 2, \ldots, k-1$; e $\mathbf{A}^{(k)}$ tem a forma:

$$\mathbf{A}^{(k)} = \begin{bmatrix} a_{11}^{(1)} & a_{12}^{(1)} & \cdot & \cdot & \cdot & \cdot & a_{1n}^{(1)} \\ 0 & a_{22}^{(2)} & & & & & a_{2n}^{(2)} \\ \cdot & & \cdot & & & & \cdot \\ \cdot & & & \cdot & & & \cdot \\ 0 & \cdot & \cdot & 0 & a_{kk}^{(k)} & \cdots & a_{kn}^{(k)} \\ \cdot & & & \cdot & \cdot & & \cdot \\ \cdot & & & \cdot & \cdot & & \cdot \\ 0 & \cdot & \cdot & 0 & a_{nk}^{(k)} & \cdots & a_{nn}^{(k)} \end{bmatrix}$$

Suponhamos $a_{kk}^{(k)} \neq 0$, em que $a_{kk}^{(k)}$ é designado por elemento pivot. Definamos os factores

$$m_{ik} = a_{ik}^{(k)} / a_{kk}^{(k)} \qquad i = k+1, \ldots, n \tag{3.3}$$

que são usados na eliminação de x_k nas equações $k+1$ até n. Definamos

$$a_{ij}^{(k+1)} = a_{ij}^{(k)} - m_{ik} a_{kj}^{(k)} \qquad i, j = k+1, \ldots, n \tag{3.4}$$

$$b_i^{(k+1)} = b_i^{(k)} - m_{ik} b_k^{(k)} \qquad i = k+1, \ldots, n \tag{3.5}$$

As primeiras k linhas de $\mathbf{A}^{(k)}$ e $\mathbf{b}^{(k)}$ não são alteradas e em $\mathbf{A}^{(k+1)}$ são introduzidos zeros na coluna k por baixo da diagonal principal. Repetindo este procedimento, ao fim de $n-1$ passos, obtemos $\mathbf{A}^{(n)}\mathbf{x} = \mathbf{b}^{(n)}$, tal que

$$\begin{bmatrix} a_{11}^{(1)} & \cdot & \cdot & \cdot & \cdot & a_{1n}^{(1)} \\ 0 & a_{22}^{(2)} & & & & a_{2n}^{(2)} \\ \cdot & \cdot & \cdot & & & \cdot \\ \cdot & & \cdot & \cdot & & \cdot \\ \cdot & & & \cdot & \cdot & \cdot \\ 0 & \cdot & \cdot & \cdot & 0 & a_{nn}^{(n)} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{bmatrix} = \begin{bmatrix} b_1^{(1)} \\ b_2^{(2)} \\ \cdot \\ \cdot \\ \cdot \\ b_n^{(n)} \end{bmatrix}$$

Resolução do sistema triangular: Por conveniência, façamos $\mathbf{U} = \mathbf{A}^{(n)}$ e $\mathbf{g} = \mathbf{b}^{(n)}$. O sistema $\mathbf{Ux} = \mathbf{g}$ é triangular superior, e é fácilmente resolúvel por substituição ascendente:

$$\begin{aligned} x_n &= g_n/u_{nn} \\ x_k &= (g_k - \textstyle\sum_{j=k+1}^{n} u_{kj}x_j)/u_{kk} \qquad k = n-1, n-2, \ldots, 1 \end{aligned} \tag{3.6}$$

■

Exemplo 3.1 *Utilizando o método de Gauss resolver o sistema*

$$\begin{aligned} 2x_1 + x_2 + 3x_3 &= 1 \\ 3x_1 + 2x_2 + 4x_3 &= 2 \\ 2x_1 + 2x_2 + x_3 &= 3 \end{aligned}$$

Para simplificar a notação, observando que as incógnitas x_1, x_2, x_3 só entram na substituição ascendente, vamos representar o sistema com a matriz aumentada:

$$\left[\mathbf{A}^{(1)}|\mathbf{b}^{(1)}\right] = [\mathbf{A}|\mathbf{b}] = \left[\begin{array}{ccc|c} 2 & 1 & 3 & 1 \\ 3 & 2 & 4 & 2 \\ 2 & 2 & 1 & 3 \end{array}\right]$$

Depois de obtermos os multiplicadores $m_{21} = 3/2$ e $m_{31} = 2/2 = 1$, temos que

$$\left[\mathbf{A}^{(2)}|\mathbf{b}^{(2)}\right] = \left[\begin{array}{ccc|c} 2 & 1 & 3 & 1 \\ 3-(3/2)2 & 2-(3/2)1 & 4-(3/2)3 & 2-(3/2)1 \\ 2-(1)2 & 2-(1)1 & 1-(1)3 & 3-(1)1 \end{array}\right] =$$

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

Então, $m_{32} = 1/(1/2) = 2$ e portanto

$$[\mathbf{U}|\mathbf{g}] = \left[\mathbf{A}^{(3)}|\mathbf{b}^{(3)}\right] = \left[\begin{array}{ccc|c} 2 & 1 & 3 & 1 \\ 0 & 1/2 & -1/2 & 1/2 \\ 0 & 1-(2)1/2 & -2-(2)(-1/2) & 2-(2)1/2 \end{array}\right] =$$

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

Resolvendo $\mathbf{Ux} = \mathbf{g}$ obtemos

$$x_3 = [1]/(-1) = -1$$

$$x_2 = [1/2 - (-1/2)(-1)]/(1/2) = 0$$

$$x_1 = [1 - 1(0) - 3(-1)]/2 = 2$$

■

É conveniente guardar os multiplicadores m_{ij}, definidos por (3.3), visto que frequentemente queremos resolver $\mathbf{Ax} = \mathbf{b}$ com o mesmo $\mathbf{A}$ mas com um diferente vector $\mathbf{b}$. O vector $\mathbf{g} = \mathbf{b}^{(n)}$ obtém-se aplicando sucessivamente para $k = 1, 2, \ldots, n-1$ as relações (3.5), onde estão presentes os multiplicadores m_{ij} e, como se depreende daquelas relações, $\mathbf{g}$ depende do vector $\mathbf{b}$. Pelo contrário, a matriz $\mathbf{U} = \mathbf{A}^{(n)}$ não depende de $\mathbf{b}$, como se pode verificar através das relações (3.4). Assim, se pretendermos resolver vários sistemas $\mathbf{Ax} = \mathbf{b}$ com a mesma matriz $\mathbf{A}$, basta-nos determinar uma única vez a matriz $\mathbf{U}$ a partir das relações (3.4). Para cada $\mathbf{b}$ determinamos, a partir das relações (3.5), o respectivo $\mathbf{g}$ e em seguida resolvemos o sistema $\mathbf{Ux} = \mathbf{g}$, utilizando as relações (3.6). Das relações (3.5), é fácil de ver que os vectores $\mathbf{b}$ e $\mathbf{g}$ estão relacionados pela equação matricial $\mathbf{Lg} = \mathbf{b}$, em que

$$\mathbf{L} = \begin{bmatrix} 1 & 0 & 0 & \cdot & \cdot & 0 \\ m_{21} & 1 & 0 & \cdot & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & & \cdot \\ \cdot & & \cdot & \cdot & \cdot & \cdot \\ \cdot & & & \cdot & 1 & 0 \\ m_{n1} & \cdot & \cdot & \cdot & m_{n,n-1} & 1 \end{bmatrix} \tag{3.7}$$

Então, somos conduzidos ao seguinte:

Teorema 3.1 *Seja* $\mathbf{A}$ *uma matriz tal que* $a_{kk}^{(k)} \neq 0$, $k = 1, 2, \ldots, n$. *Então, ela admite uma única factorização*

$$\mathbf{A} = \mathbf{LU}$$

em que $\mathbf{L}$ *é uma matriz triangular inferior de diagonal principal unitária e* $\mathbf{U}$ *é uma matriz triangular superior. A matriz* $\mathbf{U}$ *é a que se obtém utilizando a eliminação de Gauss e a matriz* $\mathbf{L}$ *é a dada por (3.7), constituida pelos coeficientes* m_{ij} *que intervêm na eliminação de Gauss.*

Demonstração: Começamos por relacionar um elemento a_{ij} de $\mathbf{A}$ com os elementos l_{ij} e u_{ij} das matrizes $\mathbf{L}$ e $\mathbf{U}$. Assim, temos que

$$a_{ij} = (\mathbf{LU})_{ij} = [l_{i1}, \ldots, l_{i,i-1}, 1, 0, \ldots, 0] \begin{bmatrix} u_{1j} \\ . \\ . \\ . \\ u_{jj} \\ 0 \\ . \\ . \\ . \\ 0 \end{bmatrix}$$

Para $i \leq j$,

$$a_{ij} = l_{i1}u_{1j} + l_{i2}u_{2j} + \ldots + l_{i,i-1}u_{i-1,j} + u_{ij}$$

e portanto[1]

$$u_{ij} = a_{ij} - \sum_{r=1}^{i-1} l_{ir}u_{rj} \qquad i \leq j \tag{3.8}$$

Para $i > j$,

$$a_{ij} = l_{i1}u_{1j} + l_{i2}u_{2j} + \ldots + l_{i,j-1}u_{j-1,j} + l_{ij}u_{jj}$$

e portanto

$$l_{ij} = (a_{ij} - \sum_{r=1}^{j-1} l_{ir}u_{rj})/u_{jj} \qquad i > j \tag{3.9}$$

admitindo que $u_{jj} \neq 0$. Veremos adiante, que os $a_{ij}^{(i)}$, presentes na matriz $\mathbf{A}^{(n)}$, e os m_{ij} podem obter-se, respectivamente, a partir de (3.8) e (3.9), se identificarmos $a_{ij}^{(i)}$ com u_{ij} e m_{ij} com l_{ij}. Assim de (3.8) e (3.9) fica provada a existência e unicidade da factorização $\mathbf{LU}$, visto que a condição $u_{jj} \neq 0$ em (3.9) é satisfeita pela hipótese feita que

$$a_{kk}^{(k)} \neq 0 \qquad k = 1, 2, \ldots, n$$

De (3.4), temos

$$a_{ij}^{(r)} = a_{ij}^{(r-1)} - m_{i,r-1}a_{r-1,j}^{(r-1)} \qquad r = 2, \ldots, k \leq i, j$$

que utilizado $k-1$ vezes conduz a

$$a_{ij}^{(k)} = a_{ij}^{(1)} - \sum_{r=1}^{k-1} m_{ir}a_{rj}^{(r)} \qquad k \leq i, j \tag{3.10}$$

Para $k = i$, esta relação é formalmente idêntica a (3.8). Se substituirmos (3.10) em (3.3) e fizermos $k = j$ obtemos

$$m_{ij} = (a_{ij}^{(1)} - \sum_{r=1}^{j-1} m_{ir}a_{rj}^{(r)})/a_{jj}^{(j)} \qquad j < i$$

[1]Quando o índice superior do somatório é menor do que o inferior convenciona-se que o somatório é nulo.

que é formalmente idêntica a (3.9). ■

Deste teorema, concluimos que a resolução do sistema $\mathbf{Ax} = \mathbf{b}$ pelo método de Gauss é equivalente a obtenção da factorização $\mathbf{A} = \mathbf{LU}$ e na resolução dos sistemas triangulares $\mathbf{Lg} = \mathbf{b}$ e $\mathbf{Ux} = \mathbf{g}$. Com efeito

$$\mathbf{Ax} = \mathbf{LUx} = \mathbf{L}(\mathbf{Ux}) = \mathbf{Lg} = \mathbf{b}$$

em que se faz $\mathbf{Ux} = \mathbf{g}$. Portanto para resolver $\mathbf{Ax} = \mathbf{b}$ basta obter por eliminação de Gauss as matrizes $\mathbf{L}$ e $\mathbf{U}$, resolver $\mathbf{Lg} = \mathbf{b}$ obtendo o vector $\mathbf{g}$ e finalmente resolver $\mathbf{Ux} = \mathbf{g}$. Se pretendermos determinar a solução $\mathbf{x}'$ do sistema $\mathbf{Ax}' = \mathbf{b}'$ em que $\mathbf{b}'$ é um vector diferente do de $\mathbf{b}$, então basta resolver $\mathbf{Lg}' = \mathbf{b}'$ obtendo o vector $\mathbf{g}'$ e resolver $\mathbf{Ux}' = \mathbf{g}'$, sem ter que efectuar novamente a factorização de $\mathbf{A}$ nas matrizes $\mathbf{L}$ e $\mathbf{U}$, que como veremos mais tarde é a fase do método de Gauss mais dispendiosa em quantidade de cálculos.

Para além desta vantagem, a factorização $\mathbf{LU}$ permite-nos de uma forma muito simples calcular $|\mathbf{A}|$, determinante de $\mathbf{A}$. Notamos que $|\mathbf{A}| = |\mathbf{L}||\mathbf{U}|$. Visto que $\mathbf{L}$ e $\mathbf{U}$ são triangulares os seus determinantes são o produto dos seus elementos diagonais. Assim $|\mathbf{L}| = 1$ e portanto

$$|\mathbf{A}| = |\mathbf{U}| = u_{11}u_{22}\dots u_{nn} = a_{11}^{(1)}a_{22}^{(2)}\dots a_{nn}^{(n)} \tag{3.11}$$

No computador, quando não é necessário guardar em memória a matriz $\mathbf{A}$, os elementos $a_{ij}^{(k+1)}$ das matrizes $\mathbf{A}^{(k+1)}$ são guardados no lugar dos $a_{ij}^{(k)}$. Como os $a_{ik}^{(k+1)}$, $i > k$ são nulos podemos guardar no seu lugar os m_{ik}. Assim, depois do passo $k-1$ da eliminação de Gauss, a matriz guardada no computador é a seguinte

$$\tilde{\mathbf{A}}^{(k)} = \begin{bmatrix} a_{11}^{(1)} & a_{12}^{(1)} & \cdot & \cdot & \cdot & \cdot & a_{1n}^{(1)} \\ m_{21} & a_{22}^{(2)} & & & & & a_{2n}^{(2)} \\ \cdot & & \cdot & & & & \cdot \\ \cdot & & & \cdot & & & \cdot \\ m_{k1} & \cdot & \cdot & m_{k,k-1} & a_{kk}^{(k)} & \cdots & a_{kn}^{(k)} \\ \cdot & & & \cdot & \cdot & & \cdot \\ \cdot & & & \cdot & \cdot & & \cdot \\ m_{n1} & \cdot & \cdot & m_{n,k-1} & a_{nk}^{(k)} & \cdots & a_{nn}^{(k)} \end{bmatrix} \tag{3.12}$$

Portanto, depois de termos efectuados os $n-1$ passos, e atendendo ao Teorema 3.1, obtemos

$$\tilde{\mathbf{A}}^{(n)} = (\mathbf{L} - \mathbf{I}) + \mathbf{U} = \begin{bmatrix} u_{11} & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & u_{1n} \\ l_{21} & u_{22} & & & & & & \cdot \\ \cdot & \cdot & \cdot & & & & & \cdot \\ \cdot & & \cdot & \cdot & & & & \cdot \\ l_{i1} & & & l_{i,i-1} & u_{ii} & & & u_{in} \\ \cdot & & & & \cdot & \cdot & & \cdot \\ \cdot & & & & & \cdot & \cdot & \cdot \\ l_{n1} & \cdot & \cdot & \cdot & \cdot & \cdots & l_{n,n-1} & u_{nn} \end{bmatrix}$$

Exemplo 3.2 *Utilizando o método de Gauss, entendido como um método de factorização, resolver o sistema*

$$\begin{aligned} 2x_1 + x_2 + 3x_3 &= 1 \\ 3x_1 + 2x_2 + 4x_3 &= 2 \\ 2x_1 + 2x_2 + x_3 &= 3 \end{aligned}$$

considerado no Exemplo 3.1. Determinar a partir da factorização o determinante da matriz do sistema considerado.

$$\tilde{\mathbf{A}}^{(1)} = \mathbf{A} = \begin{bmatrix} 2 & 1 & 3 \\ 3 & 2 & 4 \\ 2 & 2 & 1 \end{bmatrix}$$

Temos sucessivamente

$$\tilde{\mathbf{A}}^{(2)} = \left[\begin{array}{c|cc} 2 & 1 & 3 \\ \hline 3/2 & 1/2 & -1/2 \\ 1 & 1 & -2 \end{array}\right]$$

e

$$\mathbf{L} - \mathbf{I} + \mathbf{U} = \tilde{\mathbf{A}}^{(3)} = \left[\begin{array}{c|c|c} 2 & 1 & 3 \\ \hline 3/2 & 1/2 & -1/2 \\ \hline 1 & 2 & -1 \end{array}\right]$$

Portanto

$$\mathbf{L} = \begin{bmatrix} 1 & 0 & 0 \\ 3/2 & 1 & 0 \\ 1 & 2 & 1 \end{bmatrix} \qquad \mathbf{U} = \begin{bmatrix} 2 & 1 & 3 \\ 0 & 1/2 & -1/2 \\ 0 & 0 & -1 \end{bmatrix}$$

Pode-se verificar por multiplicação que $\mathbf{LU} = \mathbf{A}$. O sistema $\mathbf{Lg} = \mathbf{b}$ é

$$\begin{bmatrix} 1 & 0 & 0 \\ 3/2 & 1 & 0 \\ 1 & 2 & 1 \end{bmatrix} \begin{bmatrix} g_1 \\ g_2 \\ g_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$$

Resolvendo-o por substituição descendente

$$\begin{aligned} g_1 &= b_1/l_{11} && = 1/1 = 1 \\ g_2 &= (b_2 - l_{21}g_1)/l_{22} && = (2 - 3/2)/1 = 1/2 \\ g_3 &= (b_3 - l_{31}g_1 - l_{32}g_2)/l_{33} && = (3 - 1 - 2/2)/1 = 1 \end{aligned}$$

obtemos o vector $\mathbf{g} = [1, 1/2, 1]^T$. O sistema $\mathbf{Ux} = \mathbf{g}$ é

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

Resolvendo-o por substituição ascendente

$$\begin{aligned} x_3 &= g_3/u_{33} && = 1/(-1) = -1 \\ x_2 &= (g_2 - u_{23}x_3)/u_{22} && = (1/2 - 1/2)/(1/2) = 0 \\ x_1 &= (g_1 - u_{12}x_2 - u_{13}x_3)/u_{11} && = (1 - 0 + 3)/2 = 2 \end{aligned}$$

obtemos a solução $\mathbf{x} = [2, 0, -1]^T$. Utilizando (3.11) temos que $|\mathbf{A}| = |\mathbf{U}| =$
$= 2(1/2)(-1) = -1$. ■

Número de operações

Para analizar o número de operações necessárias na resolução de $\mathbf{Ax} = \mathbf{b}$ utilizando o método de Gauss, vamos considerar separadamente: (1) a construção de $\mathbf{L}$ e $\mathbf{U}$, (2) a resolução de $\mathbf{Lg} = \mathbf{b}$ e (3) a resolução de $\mathbf{Ux} = \mathbf{g}$.

1. *Cálculo de* $\mathbf{LU}$. No passo 1 da eliminação, $n-1$ divisões são utilizadas para calcular os factores $m_{i1},\ i = 2, \ldots, n$, dados por (3.2). Depois, $(n-1))^2$ multiplicações e $(n-1)^2$ adições são efectuadas para construir os novos elementos $a_{ij}^{(2)}$, dados por (3.1).[2] No passo k da eliminação, das relações (3.3) e (3.4), depreende-se que são necessárias $n-k$ divisões, $(n-k)^2$ multiplicações e $(n-k)^2$ adições. Assim, ao fim dos $n-1$ passos temos um total de

$$\sum_{k=1}^{n-1}(n-k) = \sum_{k=1}^{n-1} k = \frac{(n-1)n}{2}$$

divisões,

$$\sum_{k=1}^{n-1}(n-k)^2 = \sum_{k=1}^{n-1} k^2 = \frac{(n-1)n(2n-1)}{6}$$

multiplicações e $(n-1)n(2n-1)/6$ adições. Por conveniência de notação, seja $MD(.)$ e $AS(.)$ o número de multiplicações e divisões, e o número de adições e subtrações, respectivamente, para o cálculo da quantidade entre parênteses. Temos então

$$MD(\mathbf{LU}) = \frac{n(n^2-1)}{3} \approx \frac{n^3}{3}$$

$$AS(\mathbf{LU}) = \frac{n(n-1)(2n-1)}{6} \approx \frac{n^3}{3}$$

onde as aproximações são válidas quando n for elevado.

2. *Determinação de* $\mathbf{g} = \mathbf{b}^{(n)}$.

$$MD(\mathbf{g}) = \sum_{k=1}^{n-1} k = \frac{(n-1)n}{2}$$

$$AS(\mathbf{g}) = \sum_{k=1}^{n-1} k = \frac{(n-1)n}{2}$$

3. *Determinação de* $\mathbf{x}$.

$$MD(\mathbf{x}) = (\sum_{k=1}^{n-1} k) + n = \frac{n(n+1)}{2}$$

$$AS(\mathbf{x}) = \sum_{k=1}^{n-1} k = \frac{(n-1)n}{2}$$

[2] Os $a_{i1}^{(2)}, i = 2, \ldots, n$ não são calculados, visto que por construção são nulos.

Portanto, para resolver o sistema $\mathbf{Ax} = \mathbf{b}$, pelo método de Gauss, o número de operações a efectuar é:

$$MD(\mathbf{LU}, \mathbf{g}, \mathbf{x}) = \frac{n^3}{3} + n^2 - \frac{n}{3} \approx \frac{n^3}{3}$$

$$AS(\mathbf{LU}, \mathbf{g}, \mathbf{x}) = \frac{n(n-1)(2n+5)}{6} \approx \frac{n^3}{3}$$

O número de adições é praticamente o mesmo do que o número de multiplicações e divisões, por isso e para simplificar, a partir de agora só nos referiremos ao número de operações correspondentes a multiplicações e divisões.

Da contagem, feita atrás, do número de operações, podemos concluir que o maior custo na resolução de $\mathbf{Ax} = \mathbf{b}$ é na factorização $\mathbf{LU}$ ($\approx n^3/3$ operações). Depois de obtidas as matrizes $\mathbf{L}$ e $\mathbf{U}$ basta efectuar mais n^2 operações (resolvendo $\mathbf{Lg} = \mathbf{b}$ e $\mathbf{Ux} = \mathbf{g}$) para obter a solução $\mathbf{x}$. Assim uma vez obtida a factorização $\mathbf{LU}$, é pouco dispendioso resolver $\mathbf{Ax} = \mathbf{b}$ para outras escolhas do vector $\mathbf{b}$.

Notamos que a eliminação de Gauss é muito mais económica no número de operações do que *a regra de Cramer*. Se os determinantes na regra de Cramer são calculados usando desenvolvimentos por menores, então o número de operações é $(n+1)!$. Para $n = 10$, a eliminação de Gauss utiliza 430 operações enquanto que a regra de Cramer utiliza 39 916 800 operações!

Apesar do sistema $\mathbf{Ax} = \mathbf{b}$ se poder escrever na forma $\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$, nunca se resolve um sistema utilizando a matriz inversa $\mathbf{A}^{-1}$, por envolver um maior número de operações. Pode-se determinar $\mathbf{A}^{-1}$ resolvendo a equação $\mathbf{AX} = \mathbf{I}$, em que $\mathbf{A}^{-1} = \mathbf{X}$. Podemos escrever $\mathbf{X}$ e $\mathbf{I}$ em termos das suas colunas

$$\mathbf{X} = \left[\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}\right] \qquad \mathbf{I} = \left[\mathbf{e}^{(1)}, \ldots, \mathbf{e}^{(n)}\right]$$

em que os vectores $\mathbf{e}^{(1)}, \ldots, \mathbf{e}^{(n)}$ constituem a base canónica de $\mathbb{R}^n$. Então as colunas de $\mathbf{A}^{-1}$ obtém-se resolvendo os n sistemas

$$\mathbf{Ax}^{(1)} = \mathbf{e}^{(1)}, \ldots, \mathbf{Ax}^{(n)} = \mathbf{e}^{(n)}$$

que têm todos a mesma matriz $\mathbf{A}$. Assim tem-se

$$MD(\mathbf{A}^{-1}) = \frac{4}{3}n^3 - \frac{n}{3} \approx \frac{4}{3}n^3$$

Portanto o cálculo de $\mathbf{A}^{-1}$ é 4 vezes mais dispendioso do que a resolução de $\mathbf{Ax} = \mathbf{b}$ para um determinado vector $\mathbf{b}$. No entanto, com algumas modificações é possível reduzir o número de operações a

$$MD(\mathbf{A}^{-1}) = n^3$$

que é ainda o triplo do número de operações efectuadas na resolução de um único sistema. Notemos ainda, que conhecido $\mathbf{L}$ e $\mathbf{U}$ ou $\mathbf{A}^{-1}$, a determinação de $\mathbf{x}$ para um certo $\mathbf{b}$ requer n^2 multiplicações e divisões, quer resolvamos os sistemas $\mathbf{Lg} = \mathbf{b}$ e $\mathbf{Ux} = \mathbf{g}$ quer calculemos $\mathbf{A}^{-1}\mathbf{b}$. Portanto não há vantagem em guardar $\mathbf{A}^{-1}$ em vez de $\mathbf{L}$ e $\mathbf{U}$ quando tencionamos resolver sucessivamente o sistema $\mathbf{Ax} = \mathbf{b}$ com diferentes vectores $\mathbf{b}$.

3.2 Pesquisa de pivot

Em cada passo da eliminação de Gauss, nós supusemos que o elemento pivot $a_{kk}^{(k)}$ era diferente de zero. Se para um dado passo k tivermos $a_{kk}^{(k)} = 0$, então efectua-se troca de linhas de modo a que o novo $a_{kk}^{(k)}$ seja diferente de zero. Se isto não for possível é porque a matriz $\mathbf{A}$ é singular.

Não basta que o elemento pivot seja diferente de zero. Com efeito, se em valores absolutos o elemento pivot $a_{kk}^{(k)}$ for pequeno em relação aos elementos $a_{ij}^{(k)}$, $i, j \geq k$ da matriz $\mathbf{A}^{(k)}$, podemos obter um resultado afectado de um erro apreciável devido aos erros de arredondamento introduzidos pelo computador. Para minorar o efeito desses erros, recorre-se a vários tipos de *pesquisa de pivot*. Começamos por considerar as pesquisa *parcial* e *total* de pivot.

Definição 3.1 *Pesquisa parcial de pivot: No passo k da eliminação de Gauss, considere-se*

$$c_k = \max_{k \leq i \leq n} |a_{ik}^{(k)}| \tag{3.13}$$

valor máximo dos valores absolutos dos elementos $a_{ik}^{(k)}$ da coluna k localizados abaixo e sobre a diagonal principal. Seja r o menor índice das linhas, com $r \geq k$, para o qual o máximo c_k é atingido. Se $r > k$, então troca-se as linhas r e k na matriz $\tilde{\mathbf{A}}^{(k)}$, definida por (3.12), e no vector $\mathbf{b}$; e prossegue-se com o passo k da eliminação de Gauss. Esta estratégia implica que

$$|m_{ik}| \leq 1 \qquad i = k+1, \ldots, n$$

Definição 3.2 *Pesquisa total de pivot: No passo k da eliminação de Gauss, considere-se*

$$d_k = \max_{k \leq i,j \leq n} |a_{ij}^{(k)}|$$

Sejam r e s índices tais que $|a_{rs}^{(k)}| = d_k$, então troca-se as linhas r e k em $\tilde{\mathbf{A}}^{(k)}$ e $\mathbf{b}$ e troca-se as colunas s e k em $\tilde{\mathbf{A}}^{(k)}$ e $\mathbf{x}^T$; e prossegue-se com o passo k da eliminação de Gauss.

A pesquisa total de pivot é em geral a estratégia que permite reduzir mais os erros de arredondamento. No entanto, a sua utilização no computador conduz a um acréscimo de tempo de computação muito maior que quando da utilização da pesquisa parcial de pivot. A pesquisa parcial de pivot, com algumas modificações que estudaremos mais tarde, conduz quase sempre a resultados com uma precisão praticamente igual à dos resultados obtidos utilizando a pesquisa total de pivot. Por estas razões, as estratégias mais utilizadas nas aplicações são variantes da pesquisa parcial de pivot.

Exemplo 3.3 *Comparar a resolução do sistema*

$$0.0001456x_1 + 123.4x_2 = 124.1$$

$$2.885x_1 + 2.877x_2 = 3.874$$

utilizando o método de Gauss, (1) sem pesquisa de pivot, (2) com pesquisa parcial de pivot e (3) com pesquisa total de pivot, usando o computador em precisão simples. A solução exacta do sistema, com 8 algarismos significativos, é

$$x_1 = 0.33992411 \qquad x_2 = 1.0056722$$

1. *Resolução sem pesquisa de pivot.*

Tem-se sucessivamente:

$$m_{21} = 2.885/0.0001456 = 19814.561$$

$$a_{22}^{(2)} = 2.877 - (19814.561)(123.4) = -2445113.8$$

$$b_2^{(2)} = 3.874 - (19814.561)(124.1) = -2458983.3$$

$$x_2 = -2458983.3/(-2445113.8) = 1.0056723$$

$$x_1 = [124.1 - (123.4)(1.0056723)]/0.0001456 = 0.20959875$$

Portanto, x_2 foi calculado com 7 algarismos significativos mas o valor obtido para x_1 não tem significado.

2. *Resolução com pesquisa parcial de pivot.*

Trocando a ordem das equações o sistema passa a escrever-se na forma:

$$2.885x_1 + 2.877x_2 = 3.874$$

$$0.0001456x_1 + 123.4x_2 = 124.1$$

Então, tem-se sucessivamente:

$$m_{21} = 0.0001456/2.885 = 0.000050467937$$

$$a_{22}^{(2)} = 123.4 - (0.000050467937)(2.877) = 123.39986$$

$$b_2^{(2)} = 124.1 - (0.000050467937)(3.874) = 124.09980$$

$$x_2 = 124.09980/123.39986 = 1.0056722$$

$$x_1 = [3.874 - (2.877)(1.0056722)]/2.885 = 0.33992407$$

Portanto, x_1 e x_2 foram calculados respectivamente com 7 e 8 algarismos significativos, o que constitui uma melhoria substancial em relação aos resultados anteriores.

3. *Resolução com pesquisa total de pivot.*

Trocando a ordem das incógnitas o sistema passa a escrever-se na forma:

$$123.4x_1 + 0.0001456x_2 = 124.1$$

$$2.877x_1 + 2.885x_2 = 3.874$$

em que o elemento pivot 123.4 é o maior dos coeficientes $a_{ij}^{(1)}$. Então, tem-se sucessivamente:

$$m_{21} = 2.8770001/123.4 = 0.023314426$$

$$a_{22}^{(2)} = 2.8850000 - (0.023314426)(0.0001456) = 2.8849967$$

$$b_2^{(2)} = 3.874 - (0.023314426)(124.1) = 0.98067999$$

$$x_2 = 0.98067999/2.8849967 = 0.33992413$$

$$x_1 = [124.1 - (0.0001456)(0.33992413)]/123.4 = 1.0056722$$

Restabelecendo a ordem inicial para as incógnitas tem-se

$$x_1 = 0.33992413 \qquad x_2 = 1.0056722$$

Portanto, x_1 e x_2 foram calculados respectivamente com 7 e 8 algarismos significativos, o que para este caso não constitui uma melhoria em relação aos resultados obtidos utilizando pesquisa parcial de pivot. ■

Nem sempre a pesquisa parcial de pivot reduz os erros de arredondamento. Para ver isto, consideremos novamente o exemplo anterior em que se multiplicou a 1ª equação por 100000. Temos então o sistema

$$14.56x_1 + 12340000x_2 = 12410000$$

$$2.885x_1 + 2.877x_2 = 3.874$$

Como o elemento pivot $a_{11}^{(1)} = 14.56$ é maior do que $a_{21} = 2.885$, não há troca de linha e somos conduzidos aos mesmos resultados (desastrosos) do que no exemplo anterior. A pesquisa parcial de pivot não resultou porque estamos agora em presença de uma matriz desequilibrada, *i.e.*, uma matriz com elementos de ordem de grandeza muito diferente. A forma mais simples de contornar esta dificuldade é obrigar a que os elementos não excedam 1 em valor absoluto e que haja em cada linha pelo menos um elemento de módulo unitário. Para isto basta calcular o maior elemento em valor absoluto de cada linha

$$s_i = \max_{1 \le j \le n} |a_{ij}| \qquad i = 1, \ldots, n \tag{3.14}$$

e dividir a respectiva linha por este valor

$$\hat{a}_{ij} = a_{ij}/s_i \qquad i, j = 1, \ldots, n$$

Em vez de construir a matriz $\hat{\mathbf{A}} = [\hat{a}_{ij}]$ e em seguida recorrer à pesquisa parcial de pivot podemos aplicar directamente à matriz $\mathbf{A}$ *a pesquisa parcial de pivot com escala ou de equilibragem implícita*, que conduz exactamente as mesmas trocas de linhas com a vantagem de evitar a introdução de erros de arredondamento provocados pelo cálculo dos $\hat{a}_{ij}$.

Definição 3.3 *Pesquisa parcial de pivot com escala ou de equilibragem implícita: Determina-se previamente os s_i, $i = 1, \ldots, n$ definidos por (3.14). No passo k da eliminação de Gauss, considere-se*

$$c_k = \max_{k \le i \le n} \frac{|a_{ik}^{(k)}|}{s_i}$$

em vez do c_k definido por (3.13) utilizado na pesquisa parcial de pivot (Definição 3.1). Procede-se em seguida tal como na Definição 3.1 da pesquisa parcial de pivot.

As sucessivas trocas de linhas, provenientes da pesquisa de pivot, efectuadas ao longo da eliminação de Gauss são equivalentes a trocarmos essas linhas na matriz $\mathbf{A}$ antes da eliminação. Estas trocas podem ser representadas pela multiplicação de $\mathbf{A}$ por uma *matriz de permutação* $\mathbf{P}$, obtendo-se $\mathbf{PA}$. Assim o sistema $\mathbf{Ax} = \mathbf{b}$ passa a ser $\mathbf{PAx} = \mathbf{Pb}$, e depois da factorização $\mathbf{LUx} = \mathbf{Pb}$ em que

$$\mathbf{LU} = \mathbf{PA}$$

Vemos assim, que na eliminação de Gauss, a medida que se troca linhas de $\mathbf{A^{(k)}}$, temos que trocar também a ordem dos m_{ij} e dos b_i dessas linhas. A troca dos b_i pode ser feita utilizando um vector auxiliar $\mathbf{p}$ que vai registando as sucessivas trocas de linhas (é usual fazer $\mathbf{p} = [1, 2, \ldots, n]^T$ antes de iniciar a eliminação de Gauss). Este registo de trocas de linhas é utilizado na determinação correcta do determinante de $\mathbf{A}$.

Exemplo 3.4 *Utilizando o método de Gauss com pesquisa parcial de pivot com escala, resolver o sistema*

$$\begin{aligned} 2x_1 + x_2 + 3x_3 &= 1 \\ 3x_1 + 2x_2 + 4x_3 &= 2 \\ 2x_1 + 2x_2 + x_3 &= 3 \end{aligned}$$

considerado nos Exemplos 3.1 e 3.2. Determinar a partir da factorização o determinante da matriz do sistema considerado.

Considere-se o vector $\mathbf{s}$ cujas componentes são à partida

$$s_i = \max_{1 \le j \le n} |a_{ij}| \qquad i = 1, \ldots, n$$

No nosso caso temos que

$$\mathbf{p} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}, \quad \mathbf{s} = \begin{bmatrix} 3 \\ 4 \\ 2 \end{bmatrix}, \quad \mathbf{A^{(1)}} = \begin{bmatrix} 2 & 1 & 3 \\ 3 & 2 & 4 \\ \mathbf{2} & 2 & 1 \end{bmatrix}, \quad \begin{array}{l} a_{11}^{(1)}/s_1 = 2/3 \\ a_{21}^{(1)}/s_2 = 3/4 \\ a_{31}^{(1)}/s_3 = \mathbf{2/2} \end{array} \Rightarrow$$

$$\textit{troca de linha} \Rightarrow \begin{bmatrix} 3 \\ 2 \\ 1 \end{bmatrix}, \quad \begin{bmatrix} 2 \\ 4 \\ 3 \end{bmatrix}, \quad \begin{bmatrix} \mathbf{2} & 2 & 1 \\ 3 & 2 & 4 \\ 2 & 1 & 3 \end{bmatrix} \Rightarrow$$

$$\textit{passo } 1 \Rightarrow \begin{bmatrix} 3 \\ 2 \\ 1 \end{bmatrix}, \quad \begin{bmatrix} 2 \\ 4 \\ 3 \end{bmatrix}, \quad \left[\begin{array}{c|cc} 2 & 2 & 1 \\ \hline 3/2 & -1 & 5/2 \\ 1 & \mathbf{-1} & 2 \end{array}\right], \quad \begin{array}{l} a_{22}^{(2)}/s_2 = -1/4 \\ a_{32}^{(2)}/s_3 = \mathbf{-1/3} \end{array} \Rightarrow$$

$$\textit{troca de linha} \Rightarrow \begin{bmatrix} 3 \\ 1 \\ 2 \end{bmatrix} \quad , \quad \begin{bmatrix} 2 \\ 3 \\ 4 \end{bmatrix} \quad , \quad \left[\begin{array}{c|cc} 2 & 2 & 1 \\ \hline 1 & \mathbf{-1} & 2 \\ 3/2 & -1 & 5/2 \end{array}\right] \Rightarrow$$

$$\textit{passo } 2 \Rightarrow \begin{bmatrix} 3 \\ 1 \\ 2 \end{bmatrix} \quad , \quad \begin{bmatrix} 2 \\ 3 \\ 4 \end{bmatrix} \quad , \quad \left[\begin{array}{c|cc} 2 & 2 & 1 \\ \hline 1 & \underline{-1} & 2 \\ 3/2 & 1 & 1/2 \end{array}\right]$$

Temos então que

$$\mathbf{LUx} = \begin{bmatrix} b_3 \\ b_1 \\ b_2 \end{bmatrix} \qquad \mathbf{L} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 3/2 & 1 & 1 \end{bmatrix} \qquad \mathbf{U} = \begin{bmatrix} 2 & 2 & 1 \\ 0 & -1 & 2 \\ 0 & 0 & 1/2 \end{bmatrix}$$

$$\mathbf{Lg} = [3, 1, 2]^T \Rightarrow \mathbf{g} = [3, -2, -1/2]^T \quad ; \quad \mathbf{Ux} = \mathbf{g} \Rightarrow \mathbf{x} = [2, 0, -1]^T$$

O determinante de $\mathbf{A}$ é dado por

$$(-1)^{\textit{número de trocas}} u_{11}u_{22}\cdots u_{nn} = \; = (-1)^2(2)(-1)(1/2) = -1$$

■

3.3 Variantes da eliminação de Gauss

Há muitas variantes do método de Gauss. Algumas consistem em reescrever a eliminação de Gauss numa forma mais compacta, muitas vezes com a finalidade de utilizar técnicas para reduzir os erros de cálculo. Outras variantes são modificações ou simplificações, baseadas nas propriedades especiais de uma certa classe de matrizes, p.e., matrizes simétricas definidas positivas. Vamos considerar algumas destas variantes.

3.3.1 Métodos compactos

Vimos na demonstração do Teorema 3.1 que os elementos das matrizes $\mathbf{L}$ e $\mathbf{U}$ relacionavam-se entre eles e com os elementos da matriz $\mathbf{A}$ pelas relações (3.8) e (3.9). Podemos então, obter cada u_{ij} e cada l_{ij} num único passo efectuando o cálculo do somatório presente em (3.8) ou (3.9). Esta forma de proceder os cálculos para obter a factorização $\mathbf{LU}$ é conhecida por *método de Doolittle*. A ordem pela qual são calculados os u_{ij} e l_{ij} é condicionada pela relações (3.8) e (3.9); todos os u_{ij} e l_{ij} presentes no 2º membro destas relações deverão estar já previamente calculados.

Na pesquisa do elemento pivot intervém as quantidades $l_{ik}u_{kk}$ e não directamente os l_{ik}. Por esta razão, quando se pretende efectuar a eliminação de Gauss de uma forma compacta, é mais eficiente escolher para factorização de $\mathbf{A}$ o produto de uma matriz $\mathbf{L}$ triangular inferior, de diagonal principal não necessariamente unitária, por uma matriz $\mathbf{U}$ triangular superior de diagonal principal unitária ($u_{kk} = 1$). Somos conduzidos ao *método de Crout*, que passamos a descrever em pormenor.

Começamos por notar que $l_{ij} = 0,\ j > i$ e $u_{ij} = 0,\ i > j$, visto que $\mathbf{L}$ e $\mathbf{U}$ são matrizes triangulares respectivamente inferior e superior. Consideremos, para o método de Crout, a seguinte sequência de cálculos (outras variantes são possíveis): determinação sucessiva de uma coluna de $\mathbf{L}$ seguida de uma linha de $\mathbf{U}$.

Primeira coluna de L e primeira linha de U. Do que se acabou de dizer resulta imediatamente que

$$a_{i1} = \sum_{m=1}^{n} l_{im} u_{m1} = l_{i1} u_{11} \qquad i = 1, \ldots, n$$

$$a_{1j} = \sum_{m=1}^{n} l_{1m} u_{mj} = l_{11} u_{1j} \qquad j = 2, \ldots, n$$

pelo que, recordando que $\mathbf{U}$ tem diagonal unitária e portanto $u_{11} = 1$, obtemos as relações

$$l_{i1} = a_{i1} \qquad i = 1, \ldots, n \qquad \textit{primeira coluna de } \mathbf{L}$$

$$u_{1j} = a_{1j}/l_{11} \qquad j = 2, \ldots, n \qquad \textit{primeira linha de } \mathbf{U}$$

com o pressuposto que l_{11} seja diferente de zero.

Segunda coluna de L e segunda linha de U. Por outro lado também é verdade que

$$a_{i2} = \sum_{m=1}^{n} l_{im} u_{m2} = l_{i1} u_{12} + l_{i2} u_{22} \qquad i = 2, \ldots, n$$

$$a_{2j} = \sum_{m=1}^{n} l_{2m} u_{mj} = l_{21} u_{1j} + l_{22} u_{2j} \qquad j = 3, \ldots, n$$

Daqui tira-se que

$$l_{i2} = a_{i2} - l_{i1} u_{12} \qquad i = 2, \ldots, n \qquad \textit{segunda coluna de } \mathbf{L}$$

$$u_{2j} = (a_{2j} - l_{21} u_{1j})/l_{22} \qquad j = 3, \ldots, n \qquad \textit{segunda linha de } \mathbf{U}$$

com o pressuposto que l_{22} seja diferente de zero.

Coluna k de L e linha k de U. Não é dificil concluir que é possível deste modo obter todos os elementos das matrizes $\mathbf{L}$ e $\mathbf{U}$ e que as expressões gerais que os determinam são

$$l_{ik} = a_{ik} - \sum_{r=1}^{k-1} l_{ir} u_{rk} \qquad i = k, \ldots, n \qquad \textit{coluna } k \textit{ de } \mathbf{L}$$

$$u_{kj} = (a_{kj} - \sum_{r=1}^{k-1} l_{kr} u_{rj})/l_{kk} \qquad j = k+1, \ldots, n \qquad \textit{linha } k \textit{ de } \mathbf{U}$$

com o pressuposto que l_{kk} seja diferente de zero (comparar estas expressões com (3.8) e (3.9)).

A factorização pelo método de Doolittle ou pelo método de Crout conduz rigorosamente ao mesmo número de operações do que pelo método de Gauss. Assim pode-se perguntar se existe alguma vantagem na utilização dos métodos compactos. A resposta é afirmativa, visto que ao calcular num único passo um determinado elemento

de **L** ou **U** podemos efectuar este cálculo utilizando a precisão dupla do computador e armazenar em seguida esse elemento em precisão simples. Este procedimento aumenta a precisão dos resultados sem aumentar a ocupação da memória do computador.

Tal como no método de Gauss é possível a pesquisa de pivot parcial, sem ou com escala. No método de Crout basta substituir no cálculo dos c_k, referidos na Definição 3.1 ou 3.3, os $a_{ik}^{(k)}$ por l_{ik}. Vamos ilustrar com um exemplo.

Exemplo 3.5 *Resolver o sistema*

$$\mathbf{Ax} = \mathbf{b} \qquad \mathbf{A} = \begin{bmatrix} 2 & 1 & 3 \\ 3 & 2 & 4 \\ 2 & 2 & 1 \end{bmatrix} \qquad \mathbf{b} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$$

considerado em exemplos anteriores, utilizando o método de Crout, (1) sem pesquisa de pivot, (2) com pesquisa parcial de pivot e (3) com pesquisa parcial de pivot com escala.

1. *Crout sem pesquisa de pivot.*

$$\mathbf{A} = \left[\begin{array}{c|cc} \mathbf{2} & 1 & 3 \\ 3 & 2 & 4 \\ 2 & 2 & 1 \end{array}\right] \Rightarrow 1^a \textit{ linha de } \mathbf{U} \Rightarrow \left[\begin{array}{c|cc} \mathbf{2} & \underline{1/\mathbf{2}} & \underline{3/\mathbf{2}} \\ 3 & 2 & 4 \\ 2 & 2 & 1 \end{array}\right] \Rightarrow$$

$$2^a \textit{ coluna de } \mathbf{L} \Rightarrow \left[\begin{array}{cc|c} 2 & 1/2 & \underline{3/2} \\ 3 & 2-3/2 & 4 \\ 2 & 2-2/2 & 1 \end{array}\right] \Rightarrow$$

$$2^a \textit{ linha de } \mathbf{U} \Rightarrow \left[\begin{array}{cc|c} 2 & 1/2 & 3/2 \\ 3 & \mathbf{1/2} & \underline{(4-9/2)/(\mathbf{1/2})} \\ 2 & 1 & 1 \end{array}\right] \Rightarrow$$

$$3^a \textit{ coluna de } \mathbf{L} \Rightarrow \begin{bmatrix} 2 & 1/2 & 3/2 \\ 3 & 1/2 & -1 \\ 2 & 1 & 1-2(3/2)-1(-1) \end{bmatrix} = \begin{bmatrix} 2 & 1/2 & 3/2 \\ 3 & 1/2 & -1 \\ 2 & 1 & -1 \end{bmatrix}$$

Temos então que

$$\mathbf{A} = \mathbf{LU} \qquad \mathbf{L} = \begin{bmatrix} 2 & 0 & 0 \\ 3 & 1/2 & 0 \\ 2 & 1 & -1 \end{bmatrix} \qquad \mathbf{U} = \begin{bmatrix} 1 & 1/2 & 3/2 \\ 0 & 1 & -1 \\ 0 & 0 & 1 \end{bmatrix}$$

Atendendo que $\mathbf{LUx} = \mathbf{b} \Rightarrow \mathbf{Lg} = \mathbf{b}$ e $\mathbf{Ux} = \mathbf{g}$ temos

$$\begin{aligned} g_1 &= b_1/l_{11} && = 1/2 \\ g_2 &= (b_2 - l_{21}g_1)/l_{22} && = (2-3/2)/(1/2) = 1 \\ g_3 &= (b_3 - l_{31}g_1 - l_{32}g_2)/l_{33} && = (3-2/2-1)/(-1) = -1 \end{aligned}$$

$$\begin{aligned} x_3 &= g_3 && = -1 \\ x_2 &= g_2 - u_{23}x_3 && = 1-(-1)(-1) = 0 \\ x_1 &= g_1 - u_{12}x_2 - u_{13}x_3 && = 1/2 - 1/2(0) - (3/2)(-1) = 2 \end{aligned}$$

Portanto a solução do sistema dado é $\mathbf{x} = [2, 0, -1]^T$.

2. *Crout com pesquisa parcial de pivot.*

Consideremos o vector pivot $\mathbf{p}$, que vai registar as trocas de linhas efectuadas ao longo do processo de factorização, inicialmente dado por $\mathbf{p} = [1, 2, \ldots, n]^T$.

$$\mathbf{p} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} \quad , \quad \mathbf{A} = \begin{bmatrix} 2 & 1 & 3 \\ \mathbf{3} & 2 & 4 \\ 2 & 2 & 1 \end{bmatrix} \Rightarrow \textit{troca de linha} \Rightarrow \begin{bmatrix} 2 \\ 1 \\ 3 \end{bmatrix} \quad , \quad \left[\begin{array}{c|cc} \mathbf{3} & 2 & 4 \\ 2 & 1 & 3 \\ 2 & 2 & 1 \end{array}\right] \Rightarrow$$

$$1^a \textit{ linha de } \mathbf{U} \Rightarrow \begin{bmatrix} 2 \\ 1 \\ 3 \end{bmatrix} \quad , \quad \left[\begin{array}{c|cc} \mathbf{3} & \underline{2/\mathbf{3}} & \underline{4/\mathbf{3}} \\ 2 & 1 & 3 \\ 2 & 2 & 1 \end{array}\right] \Rightarrow$$

$$2^a \textit{ coluna de } \mathbf{L} \Rightarrow \begin{bmatrix} 2 \\ 1 \\ 3 \end{bmatrix} \quad , \quad \left[\begin{array}{cc|c} 3 & 2/3 & \underline{4/3} \\ 2 & 1-4/3 & 3 \\ 2 & 2-4/3 & 1 \end{array}\right] = \left[\begin{array}{cc|c} 3 & 2/3 & \underline{4/3} \\ 2 & -1/3 & 3 \\ 2 & \mathbf{2/3} & 1 \end{array}\right] \Rightarrow$$

$$\textit{troca de linha} \Rightarrow \begin{bmatrix} 2 \\ 3 \\ 1 \end{bmatrix} \quad , \quad \left[\begin{array}{cc|c} 3 & 2/3 & \underline{4/3} \\ 2 & \mathbf{2/3} & 1 \\ 2 & -1/3 & 3 \end{array}\right] \Rightarrow$$

$$2^a \textit{ linha de } \mathbf{U} \Rightarrow \begin{bmatrix} 2 \\ 3 \\ 1 \end{bmatrix} \quad , \quad \left[\begin{array}{cc|c} 3 & 2/3 & 4/3 \\ 2 & \mathbf{2/3} & \underline{(1-8/3)/(\mathbf{2/3})} \\ 2 & -1/3 & 3 \end{array}\right] \Rightarrow$$

$$3^a \textit{ coluna de } \mathbf{L} \Rightarrow \begin{bmatrix} 2 \\ 3 \\ 1 \end{bmatrix} \quad , \quad \begin{bmatrix} 3 & 2/3 & 4/3 \\ 2 & 2/3 & -5/2 \\ 2 & -1/3 & 3-8/3+5/6 \end{bmatrix} = \begin{bmatrix} 3 & 2/3 & 4/3 \\ 2 & 2/3 & -5/2 \\ 2 & -1/3 & -1/2 \end{bmatrix}$$

Temos então que

$$\mathbf{LUx} = \begin{bmatrix} b_2 \\ b_3 \\ b_1 \end{bmatrix} \qquad \mathbf{L} = \begin{bmatrix} 3 & 0 & 0 \\ 2 & 2/3 & 0 \\ 2 & -1/3 & -1/2 \end{bmatrix} \qquad \mathbf{U} = \begin{bmatrix} 1 & 2/3 & 4/3 \\ 0 & 1 & -5/2 \\ 0 & 0 & 1 \end{bmatrix}$$

$$\mathbf{Lg} = [2, 3, 1]^T \Rightarrow \mathbf{g} = [2/3, 5/2, -1]^T \quad ; \quad \mathbf{Ux} = \mathbf{g} \Rightarrow \mathbf{x} = [2, 0, -1]^T$$

3. *Crout com pesquisa parcial de pivot com escala ou de equilibragem implícita.*

Considere-se o vector $\mathbf{s}$ cujas componentes são à partida

$$s_i = \max_{1 \le j \le n} |a_{ij}| \quad , \quad i = 1, \ldots, n$$

No nosso caso temos que

$$\mathbf{p} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} \quad , \quad \mathbf{s} = \begin{bmatrix} 3 \\ 4 \\ 2 \end{bmatrix} \quad , \quad \mathbf{A} = \begin{bmatrix} 2 & 1 & 3 \\ 3 & 2 & 4 \\ \mathbf{2} & 2 & 1 \end{bmatrix} \quad , \quad \begin{array}{l} l_{11}/s_1 = 2/3 \\ l_{21}/s_2 = 3/4 \\ l_{31}/s_3 = \mathbf{2/2} \end{array} \Rightarrow$$

$$\textit{troca de linha} \Rightarrow \begin{bmatrix} 3 \\ 2 \\ 1 \end{bmatrix} , \begin{bmatrix} 2 \\ 4 \\ 3 \end{bmatrix} , \left[\begin{array}{c|cc} \mathbf{2} & 2 & 1 \\ 3 & 2 & 4 \\ 2 & 1 & 3 \end{array}\right] \Rightarrow$$

$$1^a \textit{ linha de } \mathbf{U} \Rightarrow \begin{bmatrix} 3 \\ 2 \\ 1 \end{bmatrix} , \begin{bmatrix} 2 \\ 4 \\ 3 \end{bmatrix} , \left[\begin{array}{c|cc} \mathbf{2} & 2/\mathbf{2} & 1/\mathbf{2} \\ 3 & 2 & 4 \\ 2 & 1 & 3 \end{array}\right] \Rightarrow$$

$$2^a \textit{ coluna de } \mathbf{L} \Rightarrow \begin{bmatrix} 3 \\ 2 \\ 1 \end{bmatrix} , \begin{bmatrix} 2 \\ 4 \\ 3 \end{bmatrix} , \left[\begin{array}{cc|c} \mathbf{2} & 2/\mathbf{2} & 1/\mathbf{2} \\ 3 & 2-3 & 4 \\ 2 & 1-2 & 3 \end{array}\right] = \left[\begin{array}{cc|c} 2 & 1 & 1/2 \\ 3 & -1 & 4 \\ 2 & -\mathbf{1} & 3 \end{array}\right] ,$$

$$\begin{array}{l} l_{22}/s_2 = -1/4 \\ l_{32}/s_3 = -\mathbf{1/3} \end{array} \Rightarrow \textit{troca de linha} \Rightarrow \begin{bmatrix} 3 \\ 1 \\ 2 \end{bmatrix} , \begin{bmatrix} 2 \\ 3 \\ 4 \end{bmatrix} , \left[\begin{array}{cc|c} 2 & 1 & 1/2 \\ 2 & -\mathbf{1} & 3 \\ 3 & -1 & 4 \end{array}\right] \Rightarrow$$

$$2^a \textit{ linha de } \mathbf{U} \Rightarrow \begin{bmatrix} 3 \\ 1 \\ 2 \end{bmatrix} , \begin{bmatrix} 2 \\ 3 \\ 4 \end{bmatrix} , \left[\begin{array}{cc|c} 2 & 1 & 1/2 \\ 2 & -\mathbf{1} & (3-2/2)/(-\mathbf{1}) \\ 3 & -1 & 4 \end{array}\right] \Rightarrow$$

$$3^a \textit{ coluna de } \mathbf{L} \Rightarrow \begin{bmatrix} 3 \\ 1 \\ 2 \end{bmatrix} , \begin{bmatrix} 2 \\ 3 \\ 4 \end{bmatrix} , \begin{bmatrix} 2 & 1 & 1/2 \\ 2 & -1 & -2 \\ 3 & -1 & 4-3/2-2 \end{bmatrix} = \begin{bmatrix} 2 & 1 & 1/2 \\ 2 & -1 & -2 \\ 3 & -1 & 1/2 \end{bmatrix}$$

Temos então que

$$\mathbf{LUx} = \begin{bmatrix} b_3 \\ b_1 \\ b_2 \end{bmatrix} \qquad \mathbf{L} = \begin{bmatrix} 2 & 0 & 0 \\ 2 & -1 & 0 \\ 3 & -1 & 1/2 \end{bmatrix} \qquad \mathbf{U} = \begin{bmatrix} 1 & 1 & 1/2 \\ 0 & 1 & -2 \\ 0 & 0 & 1 \end{bmatrix}$$

$$\mathbf{Lg} = [3,1,2]^T \Rightarrow \mathbf{g} = [3/2, 2, -1]^T \quad ; \quad \mathbf{Ux} = \mathbf{g} \Rightarrow \mathbf{x} = [2, 0, -1]^T$$

■

3.3.2 Método de Cholesky

Seja $\mathbf{A}$ uma matriz simétrica definida positiva de ordem n. A matriz $\mathbf{A}$ é definida positiva se

$$\mathbf{x}^T\mathbf{Ax} = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}x_ix_j > 0$$

para qualquer $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{x} \neq \mathbf{o}$. Para uma matriz deste tipo, há uma factorização muito conveniente, em que não é necessário pesquisa de pivot. É o *método de Cholesky*, que baseia-se no facto de existir a factorização

$$\mathbf{A} = \mathbf{LL}^T$$

em que $\mathbf{L}$ é uma matriz triangular inferior ($\mathbf{L}^T$ é triangular superior). O método requere somente o armazamento de $n(n+1)/2$ elementos para obter a factorização, em vez de n^2; e o número de operações é cerca de $\frac{1}{6}n^3$ em vez das usuais $\frac{1}{3}n^3$.

A obtenção dos l_{ij} é semelhante à encontrada nos métodos de Doolittle e Crout. Começamos por construir $\mathbf{L}$ multiplicando a primeira linha de $\mathbf{L}$ pela primeira coluna de $\mathbf{U} = \mathbf{L}^T$, obtendo

$$l_{11}^2 = a_{11}$$

Como $\mathbf{A}$ é definida positiva, $a_{11} > 0$, e portanto

$$l_{11} = \sqrt{a_{11}}$$

Multiplicando a primeira linha de $\mathbf{L}$ pelas restantes colunas de $\mathbf{L}^T$, obtemos os restantes elementos da 1ª coluna de $\mathbf{L}$

$$l_{i1} = a_{i1}/l_{11} \qquad i = 2, \ldots, n$$

Em geral temos para a k coluna de $\mathbf{L}$

$$l_{kk} = \sqrt{a_{kk} - \sum_{r=1}^{k-1} l_{kr}^2}$$

$$l_{ik} = \frac{a_{ik} - \sum_{r=1}^{k-1} l_{ir} l_{kr}}{l_{kk}} \qquad i = k+1, \ldots, n$$

Exemplo 3.6 *Utilizar o método de Cholesky para factorizar a matriz de Hilbert de ordem 3,*

$$\mathbf{A} = \begin{bmatrix} 1 & \frac{1}{2} & \frac{1}{3} \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{4} \\ \frac{1}{3} & \frac{1}{4} & \frac{1}{5} \end{bmatrix}$$

$$\begin{bmatrix} 1 & 0 & 0 \\ \frac{1}{2} & \frac{1}{3} & 0 \\ \frac{1}{3} & \frac{1}{4} & \frac{1}{5} \end{bmatrix} \Rightarrow 1^a \textit{ coluna de } \mathbf{L} \Rightarrow \left[\begin{array}{c|cc} \sqrt{1} & 0 & 0 \\ \frac{1}{2\sqrt{1}} & \frac{1}{3} & 0 \\ \frac{1}{3\sqrt{1}} & \frac{1}{4} & \frac{1}{5} \end{array}\right] \Rightarrow$$

$$2^a \textit{ coluna de } \mathbf{L} \Rightarrow \left[\begin{array}{ccc|c} 1 & 0 & & 0 \\ \frac{1}{2} & \sqrt{\frac{1}{3} - (\frac{1}{2})^2} = \frac{1}{2\sqrt{3}} & & 0 \\ \frac{1}{3} & (\frac{1}{4} - \frac{1}{3}\frac{1}{2})2\sqrt{3} = \frac{1}{2\sqrt{3}} & & \frac{1}{5} \end{array}\right] \Rightarrow$$

$$3^a \textit{ coluna de } \mathbf{L} \Rightarrow \begin{bmatrix} 1 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2\sqrt{3}} & 0 \\ \frac{1}{3} & \frac{1}{2\sqrt{3}} & \sqrt{\frac{1}{5} - (\frac{1}{3})^2 - (\frac{1}{2\sqrt{3}})^2} \end{bmatrix}$$

Temos portanto que

$$\mathbf{L} = \begin{bmatrix} 1 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2\sqrt{3}} & 0 \\ \frac{1}{3} & \frac{1}{2\sqrt{3}} & \frac{1}{6\sqrt{5}} \end{bmatrix}$$

■

Pode-se provar que para existir a factorização de Cholesky

$$\mathbf{A} = \mathbf{L}\mathbf{L}^T$$

com $\mathbf{A}$ e $\mathbf{L}$ matrizes reais é necessário e suficiente que $\mathbf{A}$ seja uma matriz simétrica definida positiva. Se $\mathbf{A}$ não for definida positiva pode eventualmente existir a factorização mas $\mathbf{L}$ não será real.

3.4 Análise de erros

Nem sempre a pesquisa de pivot consegue evitar que os resultados obtidos venham afectados de erros grosseiros. Consideremos, p. ex., o sitema

$$0.112x_1 + 0.101001x_2 = 0.112201$$

$$0.111x_1 + 0.100099x_2 = 0.111199$$

cuja solução exacta é

$$x_1 = 0.1 \qquad x_2 = 1$$

Resolvendo este sistema no computador em precisão simples obtemos o resultado:

$$x_1 = 0.13220689 \qquad x_2 = 0.96428573$$

Este sistema é muito mal condicionado visto que a sua solução vem completamente alterada quando se altera ligeiramente os parâmetros do sistema. Assim, p. ex., o sistema

$$0.112x_1 + 0.101001x_2 = 0.112201$$

$$0.111x_1 + 0.100099x_2 = 0.1112$$

que provem do anterior por adicionamento de 10^{-6} ao segundo membro da segunda equação, tem por solução

$$x_1 = 4.4913478 \qquad x_2 = -3.8695652$$

Enquanto que num sistema bem condicionado a pesquisa de pivot permite atenuar fortemente os erros de arredondamento e consequentemente obter uma solução com precisão suficiente, como vimos no Exemplo 3.3, num sistema mal condicionado a pesquisa de pivot só por si não permite geralmente obter uma solução com precisão suficiente. Veremos mais tarde como se pode obter uma solução com melhor precisão quando um sistema é mal condicionado. Antes, porém, vamos estudar a influência que tem na solução de um sistema um erro cometido nos parâmetros deste sistema. Para poder medir vectores e matrizes e os respectivos erros vamos ter que recorrer a normas. Assim apresentamos a seguir as principais propriedades de normas matriciais.

3.4.1 Normas de matrizes

Iremos considerar só normas de matrizes quadradas e reais induzidas por normas vectoriais. Assim, temos a seguinte:

Definição 3.4 *Seja* $\|.\|$ *uma norma vectorial, define-se uma norma da matriz* $\mathbf{A}$ *por*

$$\|\mathbf{A}\| = \max_{\mathbf{x}\neq\mathbf{o}} \frac{\|\mathbf{A}\mathbf{x}\|}{\|\mathbf{x}\|}$$

e designa-se $\|\mathbf{A}\|$ *por norma matricial induzida pela norma vectorial* $\|.\|$.

Notamos que usamos a mesma notação, $\|.\|$, para designar normas vectoriais ou matriciais. Pode-se mostrar que $\|\mathbf{A}\|$ é uma norma e que para duas matrizes da mesma ordem tem-se

$$\|\mathbf{A}\mathbf{B}\| \leq \|\mathbf{A}\|.\|\mathbf{B}\|$$

Da definição da norma induzida vê-se também que

$$\|\mathbf{A}\mathbf{x}\| \leq \|\mathbf{A}\|.\|\mathbf{x}\| \tag{3.15}$$

onde é verificada a igualdade pelo menos para um vector $\mathbf{y}$.

Vejamos alguns exemplos de normas induzidas de matrizes $n \times n$:
(a) A norma matricial induzida pela norma do máximo

$$\|\mathbf{x}\|_\infty = \max_{1\leq j\leq n} |x_j|$$

é o máximo dos somatórios dos módulos dos elementos de cada *linha*

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

Para provar isto basta mostrar que (3.15) é verificada para todos os vectores $\mathbf{x} \in \mathbb{R}^n$ e que existe um vector $\mathbf{y}$ tal que

$$\|\mathbf{A}\mathbf{y}\|_\infty = \|\mathbf{A}\|_\infty \|\mathbf{y}\|_\infty$$

Para um vector arbitrário $\mathbf{x}$ temos que

$$\begin{aligned}
\|\mathbf{A}\mathbf{x}\|_\infty &= \max_{1\leq i\leq n} |\textstyle\sum_{j=1}^{n} a_{ij}x_j| \\
&\leq \max_{1\leq i\leq n}\{(\max_{1\leq j\leq n} |x_j|) \textstyle\sum_{j=1}^{n} |a_{ij}|\} \\
&\leq \|\mathbf{x}\|_\infty \max_{1\leq i\leq n} \textstyle\sum_{j=1}^{n} |a_{ij}|
\end{aligned}$$

e portanto (3.15) é verificada. Seja k o índice de uma linha tal que

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

e seja $\mathbf{y}$ um vector cujas componentes y_j sejam 1 ou -1 de modo a que

$$a_{kj}y_j = |a_{kj}| \qquad j = 1, \ldots, n$$

Então $\|\mathbf{y}\|_\infty = 1$ e

$$\begin{aligned}\|\mathbf{A}\mathbf{y}\|_\infty &= \max_{1\le i\le n} |\textstyle\sum_{j=1}^n a_{ij}y_j| \\ &\ge |\textstyle\sum_{j=1}^n a_{kj}y_j| \\ &= \textstyle\sum_{j=1}^n |a_{kj}| = \|\mathbf{y}\|_\infty \max_{1\le i\le n} \sum_{j=1}^n |a_{ij}|\end{aligned}$$

Esta desigualdade reduz de facto a uma igualdade atendendo a (3.15), *i.e.*,

$$\|\mathbf{A}\mathbf{y}\|_\infty = \|\mathbf{A}\|_\infty \|\mathbf{y}\|_\infty$$

como pretendíamos.
(b) De forma análoga pode mostrar-se que a norma matricial induzida pela norma

$$\|\mathbf{x}\|_1 = \sum_{j=1}^n |x_j|$$

é o máximo dos somatórios dos módulos dos elementos de cada *coluna*

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

(c) Para o caso da norma matricial induzida pela norma Euclidiana

$$\|\mathbf{x}\|_2 = (\sum_{j=1}^n (x_j)^2)^{1/2}$$

precisamos da noção de raio espectral.

Definição 3.5 *O raio espectral de uma matriz* $\mathbf{A}$ *é definido por*

$$\rho(\mathbf{A}) = \max_{1\le i\le n} |\lambda_i(\mathbf{A})|$$

onde $\lambda_i(\mathbf{A})$ *designa o i-ésimo valor próprio de* $\mathbf{A}$.

Então, pode mostrar-se que

$$\|\mathbf{A}\|_2 = (\rho(\mathbf{A}^T\mathbf{A}))^{1/2}$$

Se $\mathbf{A}$ é simétrica, *i.e.*, $\mathbf{A}^T = \mathbf{A}$, então $\rho(\mathbf{A}^T\mathbf{A}) = \rho^2(\mathbf{A})$ e portanto *o raio espectral é igual a norma matricial Euclideana para matrizes simétricas.* Para uma matriz quadrada qualquer, em geral o raio espectral não é igual a norma $\|.\|_2$, mas tem-se o seguinte

Teorema 3.2 *(i) Para qualquer norma $\|.\|$ e qualquer matriz quadrada $\mathbf{A}$, tem-se*

$$\rho(\mathbf{A}) \leq \|\mathbf{A}\|$$

(ii) Por outro lado, dado uma matriz $\mathbf{A}$, para qualquer $\epsilon > 0$ existe sempre uma norma $\|.\|$ tal que

$$\|\mathbf{A}\| \leq \rho(\mathbf{A}) + \epsilon$$

■

Deste teorema concluimos que *o raio espectral de uma matriz é o ínfimo do conjunto das normas desta matriz.*

Exemplo 3.7 *Determinar as normas $\|\mathbf{A}\|_\infty$, $\|\mathbf{A}\|_1$, $\|\mathbf{A}\|_2$ e o raio espectral $\rho(\mathbf{A})$ da matriz*

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

Temos que

$$\|\mathbf{A}\|_\infty = \max(|1| + |-2|, |-3| + |4|) = \max(3, 7) = 7$$

$$\|\mathbf{A}\|_1 = \max(|1| + |-3|, |-2| + |4|) = \max(4, 6) = 6$$

Como

$$\mathbf{A}^T\mathbf{A} = \begin{bmatrix} 10 & -14 \\ -14 & 20 \end{bmatrix}$$

temos que

$$\rho(\mathbf{A}^T\mathbf{A}) = \max(15 - \sqrt{221}, 15 + \sqrt{221})$$

e portanto

$$\|\mathbf{A}\|_2 = \sqrt{15 + \sqrt{221}} \approx 5.465$$

Temos para o raio espectral

$$\rho(\mathbf{A}) = \max((5 - \sqrt{33})/2, (5 + \sqrt{33})/2) \approx 5.372$$

e por conseguinte $\mathbf{A}$ verifica o Teorema 3.2. ■

3.4.2 Número de condição de uma matriz

Depois deste breve resumo sobre normas, vamos então examinar a influência de uma pertubação nos parâmetros de um sistema $\mathbf{Ax} = \mathbf{b}$ na sua solução. Começamos por considerar o caso mais simples em que se perturba o vector $\mathbf{b}$. Assim consideremos a solução $\tilde{\mathbf{x}}$ do sistema perturbado

$$\mathbf{A}\tilde{\mathbf{x}} = \tilde{\mathbf{b}} = \mathbf{b} - \Delta\mathbf{b}$$

Seja

$$\Delta\mathbf{x} = \mathbf{x} - \tilde{\mathbf{x}}$$

o erro. Subtraindo $\mathbf{Ax} = \mathbf{b}$ ao sistema perturbado obtemos (supondo que $\mathbf{A}$ é invertível)

$$\mathbf{A}\Delta\mathbf{x} = \Delta\mathbf{b} \qquad \Delta\mathbf{x} = \mathbf{A}^{-1}\Delta\mathbf{b} \tag{3.16}$$

Aplicando normas em (3.16) e utilizando (3.15) temos as seguintes majorações

$$\|\mathbf{A}\|^{-1}\|\Delta\mathbf{b}\| \leq \|\Delta\mathbf{x}\| \leq \|\mathbf{A}^{-1}\|\|\Delta\mathbf{b}\| \tag{3.17}$$

Utilizando (3.17), podemos minorar e majorar o erro relativo $\|\Delta\mathbf{x}\|/\|\mathbf{x}\|$. Assim dividindo por $\|\mathbf{x}\| \neq 0$ obtemos

$$\frac{\|\mathbf{A}\|^{-1}\|\Delta\mathbf{b}\|}{\|\mathbf{x}\|} \leq \frac{\|\Delta\mathbf{x}\|}{\|\mathbf{x}\|} \leq \frac{\|\mathbf{A}^{-1}\|\|\Delta\mathbf{b}\|}{\|\mathbf{x}\|} \tag{3.18}$$

Atendendo a (3.15) e que $\mathbf{Ax} = \mathbf{b}$ e $\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$ podemos escrever sucessivamente (comparar com (3.17))

$$\|\mathbf{A}\|^{-1}\|\mathbf{b}\| \leq \|\mathbf{x}\| \leq \|\mathbf{A}^{-1}\|\|\mathbf{b}\|$$

e

$$\frac{1}{\|\mathbf{A}^{-1}\|\|\mathbf{b}\|} \leq \frac{1}{\|\mathbf{x}\|} \leq \frac{\|\mathbf{A}\|}{\|\mathbf{b}\|}$$

Substituindo este último resultado em (3.18) obtemos

$$\frac{1}{\|\mathbf{A}\|\|\mathbf{A}^{-1}\|} \cdot \frac{\|\Delta\mathbf{b}\|}{\|\mathbf{b}\|} \leq \frac{\|\Delta\mathbf{x}\|}{\|\mathbf{x}\|} \leq \|\mathbf{A}\|\|\mathbf{A}^{-1}\| \cdot \frac{\|\Delta\mathbf{b}\|}{\|\mathbf{b}\|} \tag{3.19}$$

Vamos designar por *número de condição* da matriz $\mathbf{A}$ a quantidade

$$cond(\mathbf{A}) = \|\mathbf{A}\|\|\mathbf{A}^{-1}\| \tag{3.20}$$

Assim podemos escrever (3.19) na forma

$$\frac{1}{cond(\mathbf{A})} \leq \frac{\|\Delta\mathbf{x}\|/\|\mathbf{x}\|}{\|\Delta\mathbf{b}\|/\|\mathbf{b}\|} \leq cond(\mathbf{A})$$

em que qualquer das igualdades é verificada pelo menos para um par de vectores $\mathbf{b}$ e $\Delta\mathbf{b}$. Vemos que se $cond(\mathbf{A})$ for grande, então para uma pequena perturbação relativa no vector $\mathbf{b}$ podemos ter uma grande perturbação relativa na solução $\mathbf{x}$ e que por outro lado para uma grande perturbação relativa no vector $\mathbf{b}$ podemos ter uma pequena perturbação relativa na solução $\mathbf{x}$.

O número de condição definido por (3.20) varia com a norma escolhida, mas é sempre superior a um. De facto

$$1 = \|\mathbf{I}\| = \|\mathbf{A}\mathbf{A}^{-1}\| \leq \|\mathbf{A}\|\|\mathbf{A}^{-1}\| = cond(\mathbf{A})$$

Podemos definir um número de condição em termos dos raios espectrais de $\mathbf{A}$ e $\mathbf{A}^{-1}$ e independente da norma escolhida. Assim consideremos o número de condição

$$cond(\mathbf{A})_\rho = \rho(\mathbf{A})\rho(\mathbf{A}^{-1})$$

Do Teorema 3.2 temos que

$$cond(\mathbf{A})_\rho \leq cond(\mathbf{A}) \tag{3.21}$$

para qualquer norma. Visto que os valores próprios de $\mathbf{A}^{-1}$ são os inversos dos de $\mathbf{A}$, temos que

$$cond(\mathbf{A})_\rho = \frac{\max_{1\leq i\leq n} |\lambda_i(\mathbf{A})|}{\min_{1\leq i\leq n} |\lambda_i(\mathbf{A})|}$$

e portanto

$$cond(\mathbf{A})_\rho \geq 1$$

Na determinação de um número de condição de uma matriz $\mathbf{A}$ intervem o cálculo da matriz inversa ou dos valores próprios de $\mathbf{A}$. Como qualquer destes cálculos são geralmente mais complicados do que a resolução do sistema $\mathbf{Ax} = \mathbf{b}$ é usual obter uma estimativa do número de condição sem recorrer a estes cálculos. Um processo para obter esta estimativa é o descrito de seguida. Consideremos o sistema $\mathbf{Ay} = \mathbf{c}$. Procedendo de uma forma equivalente à utilizada na obtenção de (3.17) obtemos as seguintes desigualdades

$$\frac{1}{\|\mathbf{A}\|} \leq \frac{\|\mathbf{y}\|}{\|\mathbf{c}\|} \leq \|\mathbf{A}^{-1}\|$$

Se escolhermos aleatoriamente p vectores $\mathbf{c}_k$ e determinarmos as soluções $\mathbf{y}_k$ dos sistemas $\mathbf{Ay}_k = \mathbf{c}_k$ com $k = 1, 2, \ldots, p$, então

$$cond_p(\mathbf{A}) = \|\mathbf{A}\| \max_{1\leq k\leq p} \frac{\|\mathbf{y}_k\|}{\|\mathbf{c}_k\|} \leq cond(\mathbf{A})$$

constitui geralmente uma boa aproximação de $cond(\mathbf{A})$ mesmo para p relativamente pequeno (p.ex. $p = 6$). Uma vez obtida a factorização de $\mathbf{A}$, a determinação dos p vectores $\mathbf{y}_k$ só necessita de pn^2 operações, o que constitui pouco trabalho computacional comparado com a resolução completa do sistema $\mathbf{Ax} = \mathbf{b}$.

Exemplo 3.8 *Determinar os números de condição $cond(\mathbf{A})_\infty$, $cond(\mathbf{A})_1$, $cond(\mathbf{A})_2$ e $cond(\mathbf{A})_\rho$ da matriz*

$$\mathbf{A} = \begin{bmatrix} 7 & 10 \\ 5 & 7 \end{bmatrix}$$

Determinar as estimativas $cond_p(\mathbf{A})_\infty$ e $cond_p(\mathbf{A})_1$ de $cond(\mathbf{A})_\infty$ e $cond(\mathbf{A})_1$ utilizando $p = 6$ vectores $\mathbf{y}_k$ obtidos aleatoriamente. Verificar que com esta matriz o sistema $\mathbf{Ax} = \mathbf{b}$, com $\mathbf{b} = [1, 0.7]^T$ é mal condicionado.

Começamos por notar que

$$\mathbf{A}^{-1} = \begin{bmatrix} -7 & 10 \\ 5 & -7 \end{bmatrix}$$

Então

$$cond(\mathbf{A})_\infty = \|\mathbf{A}\|_\infty \|\mathbf{A}^{-1}\|_\infty = (17)(17) = 289$$

e

$$cond(\mathbf{A})_1 = \|\mathbf{A}\|_1 \|\mathbf{A}^{-1}\|_1 = (17)(17) = 289$$

Como

$$cond(\mathbf{A})_2 = \|\mathbf{A}\|_2 \|\mathbf{A}^{-1}\|_2 = [\rho(\mathbf{A}^T\mathbf{A})\rho((\mathbf{A}^T\mathbf{A})^{-1})]^{1/2} = (cond(\mathbf{A}^T\mathbf{A})_\rho)^{1/2}$$

e os valores próprios de

$$\mathbf{A}^T\mathbf{A} = \begin{bmatrix} 74 & 105 \\ 105 & 149 \end{bmatrix}$$

são $(223 \pm \sqrt{223^2 - 4})/2$ temos que

$$cond(\mathbf{A})_2 \approx 223$$

Como os valores próprios de $\mathbf{A}$ são $7 \pm \sqrt{50}$ então

$$cond(\mathbf{A})_\rho = \frac{\sqrt{50} + 7}{\sqrt{50} - 7} \approx 198$$

Vemos assim que (3.21) é verificada. Utilizando sucessivamente os vectores

$$\begin{array}{ll} c_1 = [-0.564067, -0.404787]^T & c_2 = [+0.647882, -0.469623]^T \\ c_3 = [-0.648677, +0.334547]^T & c_4 = [-0.154625, +0.061334]^T \\ c_5 = [-0.240373, +0.005740]^T & c_6 = [+0.197823, -0.864721]^T \end{array}$$

em que as componentes foram obtidas aleatoriamente com valores no intervalo $[-1, 1]$, e resolvendo os respectivos sistemas $\mathbf{A}\mathbf{y}_k = \mathbf{c}_k$, $k = 1, \ldots, 6$ obtemos para os quocientes

$$\|\mathbf{y}_k\|_\infty / \|\mathbf{c}_k\|_\infty$$

e

$$\|\mathbf{y}_k\|_1 / \|\mathbf{c}_k\|_1$$

respectivamente os seguintes valores

$$0.1762 \quad \mathbf{14.2486} \quad 12.1574 \quad 10.9666 \quad 7.2388 \quad 11.6014$$

e

$$0.1162 \quad 14.1012 \quad 13.7013 \quad 13.4200 \quad 12.1166 \quad \mathbf{16.0691}$$

Assim obtemos para estimativas das normas $\|\mathbf{A}^{-1}\|_\infty$ e $\|\mathbf{A}^{-1}\|_1$ respectivamente 14.2486 e 16.0691. Portanto

$$cond_p(\mathbf{A})_\infty = 17 \times 14.2486 \approx 242$$

e

$$cond_p(\mathbf{A})_1 = 17 \times 16.0691 \approx 273$$

O sistema $\mathbf{A}\mathbf{x} = \mathbf{b}$ em causa é

$$7x_1 + 10x_2 = 1$$

$$5x_1 + 7x_2 = 0.7$$

cuja solução é

$$x_1 = 0 \qquad x_2 = 0.1$$

Para vermos que ele é mal condicionado basta considerar uma pequena perturbação no segundo membro. Assim, p.ex., o sistema

$$7\tilde{x}_1 + 10\tilde{x}_2 = 1.01$$

$$5\tilde{x}_1 + 7\tilde{x}_2 = 0.69$$

tem a solução

$$\tilde{x}_1 = -0.17 \qquad \tilde{x}_2 = 0.22$$

Podemos também verificar que para dois segundo membros bastante diferentes podemos obter soluções quase iguais. Assim a $\mathbf{b} = [20, 20]^T$ corresponde $\mathbf{x} = [60, -40]^T$ e a $\mathbf{b} = [3, 8]^T$ corresponde $\mathbf{x} = [59, -41]^T$. ■

Até agora só vimos a influência de uma perturbação no 2º membro $\mathbf{b}$ do sistema $\mathbf{Ax} = \mathbf{b}$ sobre a sua solução $\mathbf{x}$. Vamos agora considerar o caso geral em que a matriz $\mathbf{A}$ também é perturbada:

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

com

$$\tilde{\mathbf{A}} = \mathbf{A} - \Delta\mathbf{A} \qquad \tilde{\mathbf{b}} = \mathbf{b} - \Delta\mathbf{b}$$

Vamos supor que $\mathbf{A}$ é invertível e que a perturbação na matriz é suficientemente pequena de tal modo que

$$\|\Delta\mathbf{A}\|/\|\mathbf{A}\| < \frac{1}{cond(\mathbf{A})}$$

Com esta condição demonstra-se que

$$\frac{\|\Delta\mathbf{x}\|}{\|\mathbf{x}\|} \leq \frac{cond(\mathbf{A})}{1 - cond(\mathbf{A})\dfrac{\|\Delta\mathbf{A}\|}{\|\mathbf{A}\|}} \left(\frac{\|\Delta\mathbf{A}\|}{\|\mathbf{A}\|} + \frac{\|\Delta\mathbf{b}\|}{\|\mathbf{b}\|} \right) \tag{3.22}$$

Vemos assim que quanto maior for $cond(\mathbf{A})$ maior será o 2º membro da majoração (3.22).

3.4.3 Método da correcção residual

Mesmo para um sistema $\mathbf{Ax} = \mathbf{b}$ bem condicionado, pode acontecer que a solução calculada $\tilde{\mathbf{x}}$ seja afectada de um erro

$$\mathbf{e} = \mathbf{x} - \tilde{\mathbf{x}}$$

relativamente grande, devido aos erros de arredondamento introduzidos principalmente na fase da factorização $\mathbf{LU}$ de $\mathbf{A}$. Portanto o resíduo

$$\mathbf{r} = \mathbf{b} - \mathbf{A}\tilde{\mathbf{x}} \tag{3.23}$$

não é zero. Como

$$\mathbf{r} = \mathbf{b} - \mathbf{A}(\mathbf{x} - \mathbf{e}) = \mathbf{Ae}$$

vemos que $\mathbf{e}$ é a solução do sistema

$$\mathbf{Ae} = \mathbf{r} \tag{3.24}$$

Assim, teoricamente, para obter a solução exacta $\mathbf{x}$ basta efectuar a soma

$$\mathbf{x} = \tilde{\mathbf{x}} + \mathbf{e}$$

em que a *correcção* $\mathbf{e}$ é a solução do sistema (3.41).

Na prática, se não foi possível obter a solução exacta de $\mathbf{Ax} = \mathbf{b}$, normalmente também não é possível obter a solução exacta de $\mathbf{Ae} = \mathbf{r}$, mas sim uma solução aproximada $\tilde{\mathbf{e}}$. Assim, ao adicionar $\tilde{\mathbf{e}}$ a $\tilde{\mathbf{x}}$ não obtemos a solução exacta de $\mathbf{Ax} = \mathbf{b}$ mas uma aproximação de $\mathbf{x}$ que geralmente será melhor do que a aproximação $\tilde{\mathbf{x}}$. Com esta nova aproximação de $\mathbf{x}$, podemos determinar um novo resíduo, calcular uma nova correcção e obter mais uma aproximação de $\mathbf{x}$. Podemos continuar este processo iterativo até obter uma aproximação suficientemente boa de $\mathbf{x}$. Este processo de corrigir sucessivamente a solução aproximada é conhecido por *método da correcção residual* ou *do refinamento iterativo*.

Nas sucessivas correcções temos que resolver o sistema (3.24) em que o 2º membro $\mathbf{r}$ é diferente em cada correcção mas a matriz $\mathbf{A}$ é sempre a mesma. Assim, tendo obtido a factorização $\mathbf{LU}$ de $\mathbf{A}$, quando da obtenção de $\tilde{\mathbf{x}}$, basta resolver dois sistemas triangulares, $\mathbf{Lg} = \mathbf{r}$ e $\mathbf{Ue} = \mathbf{g}$, para obter uma nova correcção, o que torna este processo computacionalmente económico. Dado que a obtenção da factorização $\mathbf{LU}$ é afectada pelos erros de arredondamento, só é possível obter aproximações $\tilde{\mathbf{L}}$ e $\tilde{\mathbf{U}}$ de $\mathbf{L}$ e $\mathbf{U}$. Assim, admitindo que a influência dos erros de arredondamento é desprezável na fase de resolução dos sistemas triangulares em face da factorização, a solução aproximada $\tilde{\mathbf{e}}$ calculada para o sistema (3.24) é solução de

$$\tilde{\mathbf{L}}\tilde{\mathbf{U}}\tilde{\mathbf{e}} = \mathbf{Pr} \tag{3.25}$$

em que $\mathbf{P}$ é a matriz de permutação relacionada com as eventuais trocas de linhas.

Baseado nas relações (3.23) e (3.25) e nas considerações feitas atrás, podemos então apresentar um algoritmo para o método da correcção residual.

Algoritmo 3.1 *(Método da correcção residual)*

INICIALIZACÃO:
$\mathbf{x}^{(0)} = \tilde{\mathbf{x}}$ em que $\tilde{\mathbf{x}}$ é a solução de $\tilde{\mathbf{L}}\tilde{\mathbf{U}}\tilde{\mathbf{x}} = \mathbf{Pb}$

CICLO: PARA $k = 0, 1, \ldots$ FAZER

$\mathbf{r}^{(k)} = \mathbf{b} - \mathbf{Ax}^{(k)}$

$\tilde{\mathbf{L}}\tilde{\mathbf{U}}\mathbf{e}^{(k)} = \mathbf{Pr}^{(k)}$

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

SE $\|\mathbf{e}^{(k)}\| \leq \epsilon$ ENTÃO tomar $\mathbf{x}^{(k+1)}$ como solução de $\mathbf{Ax} = \mathbf{b}$, SAÍDA

FIM DO CICLO.

Notamos que para utilizar a correcção residual, não se pode destruir a matriz $\mathbf{A}$ quando da factorização $\mathbf{LU}$ o que implica o dobro de ocupação da memória. Por outro lado, dado que $\mathbf{r}^{(k)} \to \mathbf{o}$ quando $k \to \infty$ o cálculo de $\mathbf{Ax}^{(k)}$ deverá ser efectuado com uma precisão tal que o erro cometido neste cálculo seja desprezável em face de $\mathbf{r}^{(k)}$. Assim na prática, se for utilizada a precisão simples para armazenar os elementos de $\mathbf{A}$, $\mathbf{L}$ e $\mathbf{U}$, o cálculo de $\mathbf{Ax}^{(k)}$ deverá ser efectuado em precisão dupla.

Exemplo 3.9 *Resolver o sistema* $\mathbf{Ax} = \mathbf{b}$ *em que*

$$\mathbf{A} = \begin{bmatrix} 3.5 & 2.5 & 1.5 \\ 3 & 3 & 3.1 \\ 2.5 & 1.5 & 0.5 \end{bmatrix} \qquad \mathbf{b} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$$

utilizando o método da correcção residual. Considerar uma máquina que opera em vírgula flutuante, na base 10, com arredondamento simétrico e com 4 ou 8 algarismos na mantissa consoante se utilizar a precisão simples ou dupla.

As contas são efectuadas em precisão dupla e os resultados guardados em precisão simples. Assim utilizando o método de Crout obtém-se a factorização $\tilde{\mathbf{L}}\tilde{\mathbf{U}}$ em que

$$\tilde{\mathbf{L}} = \begin{bmatrix} 3.5 & 0 & 0 \\ 3 & 0.8571 & 0 \\ 2.5 & -0.2858 & 0.03354 \end{bmatrix} \qquad \tilde{\mathbf{U}} = \begin{bmatrix} 1 & 0.7143 & 0.4286 \\ 0 & 1 & 2.117 \\ 0 & 0 & 1 \end{bmatrix}$$

Resolvendo sucessivamente $\tilde{\mathbf{L}}\mathbf{g} = \mathbf{b}$ e $\tilde{\mathbf{U}}\tilde{\mathbf{x}} = \mathbf{g}$ obtemos

$$\mathbf{g} = [0.2857, 1.333, 79.51]^T$$

e

$$\mathbf{x}^{(0)} = \tilde{\mathbf{x}} = [85.50, -167.0, 79.51]^T$$

Calculando em seguida

$$\mathbf{r}^{(0)} = \mathbf{b} - \mathbf{A}\mathbf{x}^{(0)}$$

obtemos

$$\mathbf{r}^{(0)} = [-0.015, 0.019, -0.005]^T$$

Resolvendo sucessivamente $\tilde{\mathbf{L}}\mathbf{g} = \mathbf{r}^{(0)}$ e $\tilde{\mathbf{U}}\mathbf{e}^{(0)} = \mathbf{g}$, obtemos

$$\mathbf{g} = [-0.004286, 0.03717, 0.4871]^T$$

e

$$\mathbf{e}^{(0)} = [0.4970, -0.9940, 0.4871]^T$$

Atendendo que $\mathbf{x}^{(1)} = \mathbf{x}^{(0)} + \mathbf{e}^{(0)}$ temos que

$$\mathbf{x}^{(1)} = [86, -168, 80]^T$$

que é a solução exacta do sistema. Notamos que, se tivéssemos calculado $\mathbf{A}\mathbf{x}^{(0)}$ utilizando somente precisão simples, obteríamos para o resíduo o resultado

$$\mathbf{r}^{(0)} = [-0.1, 0, -0.06]^T$$

o que conduziria ao vector

$$\mathbf{x}^{(1)} = [86.70, -169.4, 80.70]^T$$

que constitui uma aproximação da solução de $\mathbf{Ax} = \mathbf{b}$ pior do que a aproximação inicial $\mathbf{x}^{(0)} = \tilde{\mathbf{x}}$. ■

3.5 Métodos iterativos

Muitos sistemas lineares são demasiadamente grandes para serem resolvidos por métodos directos. Para estes sistemas, os métodos iterativos são geralmente os únicos que permitem obter as respectivas soluções, e frequentemente estes métodos são mais rápidos do que os directos.

Para resolver iterativamente $\mathbf{Ax} = \mathbf{b}$ começamos por decompor $\mathbf{A}$ na soma

$$\mathbf{A} = \mathbf{M} + \mathbf{N}$$

e escrever o sistema na forma

$$\mathbf{Mx} = \mathbf{b} - \mathbf{Nx}$$

A matriz $\mathbf{M}$ é escolhida de tal modo que seja invertível e o sistema $\mathbf{Mz} = \mathbf{f}$ seja de fácil resolução para qualquer $\mathbf{f}$. Por exemplo, $\mathbf{M}$ pode ser diagonal, triangular ou tridiagonal. O método iterativo é definido por

$$\mathbf{Mx}^{(k+1)} = \mathbf{b} - \mathbf{Nx}^{(k)} \qquad k = 0, 1, \ldots \tag{3.26}$$

sendo $\mathbf{x}^{(0)}$ um vector dado. Consideraremos a seguir os métodos iterativos de *Jacobi* e de *Gauss-Seidel* que correspondem a determinadas escolhas da matriz $\mathbf{M}$. Antes porém, vamos ver que o método da *correcção residual*, estudado na secção anterior, pode considerar-se como um método iterativo definido por (3.26). Para isso, vamos dar outra forma a relação (3.26).

Como $\mathbf{N} = \mathbf{A} - \mathbf{M}$ temos que

$$\mathbf{Mx}^{(k+1)} = \mathbf{b} - \mathbf{Ax}^{(k)} + \mathbf{Mx}^{(k)} \qquad k = 0, 1, \ldots$$

e portanto podemos escrever

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \mathbf{M}^{-1}(\mathbf{b} - \mathbf{Ax}^{(k)}) \qquad k = 0, 1, \ldots \tag{3.27}$$

Vemos assim que tomando

$$\mathbf{M} = \mathbf{P}^{-1}\tilde{\mathbf{L}}\tilde{\mathbf{U}}$$

o método iterativo definido por (3.26) ou (3.27) reduz-se ao método da correcção residual.

3.5.1 Métodos de Jacobi e de Gauss-Seidel

Sendo a_{ij} , $i = 1, \ldots, n$, $j = 1, \ldots, n$ os elementos de $\mathbf{A}$, designemos por

$$\mathbf{L} = \begin{bmatrix} 0 & \cdot & \cdot & \cdot & \cdot & 0 \\ a_{21} & 0 & \cdot & \cdot & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ a_{n-1,1} & \cdot & \cdot & a_{n-1,n-2} & 0 & 0 \\ a_{n1} & \cdot & \cdot & \cdot & a_{n,n-1} & 0 \end{bmatrix}$$

$$\mathbf{D} = \begin{bmatrix} a_{11} & 0 & \cdot & \cdot & \cdot & 0 \\ 0 & a_{22} & 0 & \cdot & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & 0 & a_{n-1,n-1} & 0 \\ 0 & \cdot & \cdot & \cdot & 0 & a_{nn} \end{bmatrix}$$

e

$$\mathbf{U} = \begin{bmatrix} 0 & a_{12} & \cdot & \cdot & \cdot & a_{1n} \\ 0 & 0 & a_{23} & \cdot & \cdot & a_{2n} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & 0 & a_{n-1,n} \\ 0 & \cdot & \cdot & \cdot & \cdot & 0 \end{bmatrix}$$

as matrizes estritamente triangular inferior, diagonal e estritamente triangular superior que constituem a matriz $\mathbf{A}$. Portanto temos

$$\mathbf{A} = \mathbf{L} + \mathbf{D} + \mathbf{U}$$

3.5.1.1 Método de Jacobi (substituições simultáneas)

O método de Jacobi consiste em fazer

$$\mathbf{M} = \mathbf{D} \text{ e } \mathbf{N} = \mathbf{L} + \mathbf{U}$$

Para este método, temos portanto

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

ou ainda

$$a_{ii}x_i^{(k+1)} = b_i - \sum_{j=1, j\neq i}^{n} a_{ij}x_j^{(k)} \qquad i = 1, 2, \ldots, n \quad k = 0, 1, \ldots$$

onde $\mathbf{x}^{(k)} = [x_1^{(k)}, \ldots, x_n^{(k)}]^T$. Admitindo que todos os $a_{ii} \neq 0$ (o que é sempre possível por meio de convenientes permutações de linhas e de colunas desde que a matriz $\mathbf{A}$ seja regular) temos o seguinte método iterativo

$$x_i^{(k+1)} = (b_i - \sum_{j=1, j\neq i}^{n} a_{ij}x_j^{(k)})/a_{ii} \qquad i = 1, 2, \ldots, n \quad k = 0, 1, \ldots$$

Exemplo 3.10 *Aplicar o método de Jacobi à resolução do sistema* $\mathbf{A}\mathbf{x} = \mathbf{b}$ *com*

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & -1 \\ -3/4 & 2 & 0 \\ -1 & -1/4 & 2 \end{bmatrix} \qquad \mathbf{b} = \begin{bmatrix} 2 \\ 1 \\ 1 \end{bmatrix} \qquad \mathbf{x}^{(0)} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

Começamos por evidenciar as sucessivas incógnitas ficando o sistema na forma

$$\begin{aligned} x_1 &= 2 - 0x_2 + x_3 \\ x_2 &= (1 + 3/4x_1 - 0x_3)/2 \\ x_3 &= (1 + x_1 + 1/4x_2)/2 \end{aligned}$$

O método iterativo de Jacobi é, neste caso, dado por

$$\begin{aligned} x_1^{(k+1)} &= 2 - 0x_2^{(k)} + x_3^{(k)} \\ x_2^{(k+1)} &= (1 + 3/4x_1^{(k)} - 0x_3^{(k)})/2 \qquad k = 0, 1, \ldots \\ x_3^{(k+1)} &= (1 + x_1^{(k)} + 1/4x_2^{(k)})/2 \end{aligned}$$

Com a escolha $x_1^{(0)} = 0, x_2^{(0)} = 0, x_3^{(0)} = 0$, obtemos para as 2 primeiras iterações:

$$\begin{aligned} x_1^{(1)} &= 2 + 0 = 2 & x_1^{(2)} &= 2 + 1/2 = 5/2 \\ x_2^{(1)} &= (1 + 0)/2 = 1/2 & x_2^{(2)} &= (1 + 3/2)/2 = 5/4 \\ x_3^{(1)} &= (1 + 0 + 0)/2 = 1/2 & x_3^{(2)} &= (1 + 2 + 1/8)/2 = 25/16 \end{aligned}$$

A solução exacta do sistema é $\mathbf{x} = [164/29, 76/29, 106/29]^T$. Denominando o vector erro na iteração k por

$$\mathbf{e}^{(k)} = \mathbf{x} - \mathbf{x}^{(k)}$$

e notando que a norma do máximo $\|.\|_\infty$ é definida por

$$\|\mathbf{x}\|_\infty = \max_{1 \le i \le n} |x_i|$$

obtemos

$$\|\mathbf{e}^{(0)}\|_\infty = \max(164/29, 76/29, 106/29) = 164/29 = 5.655172$$

$$\|\mathbf{e}^{(1)}\|_\infty = 164/29 - 2 = 3.655172$$

e

$$\|\mathbf{e}^{(2)}\|_\infty = 164/29 - 5/2 = 3.155172$$

Na Tabela 3.1 apresentamos os resultados obtidos efectuando 8 iteracões. Destes resultados, vemos que para este sistema o método de Jacobi converge, se bem que lentamente. ■

3.5.1.2 Método de Gauss-Seidel (substituições sucessivas)

O método de Gauss-Seidel consiste em fazer

$$\mathbf{M} = \mathbf{L} + \mathbf{D} \text{ e } \mathbf{N} = \mathbf{U}$$

Para este método, temos portanto

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

Tabela 3.1: Exemplo do método de Jacobi

k	$x_1^{(k)}$	$x_2^{(k)}$	$x_3^{(k)}$	$\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\|_\infty$	$\|\mathbf{e}^{(k)}\|_\infty$	$\|\mathbf{e}^{(k)}\|_\infty / \|\mathbf{e}^{(k-1)}\|_\infty$
0	0.000000	0.000000	0.000000		5.655172	
1	2.000000	0.500000	0.500000	2.000000	3.655172	0.646341
2	2.500000	1.250000	1.562500	1.062500	3.155172	0.863208
3	3.562500	1.437500	1.906250	1.062500	2.092672	0.663251
4	3.906250	1.835938	2.460938	0.554688	1.748922	0.835736
5	4.460938	1.964844	2.682617	0.554688	1.194235	0.682841
6	4.682617	2.172852	2.976074	0.293457	0.972555	0.814375
7	4.976074	2.255981	3.112915	0.293457	0.679098	0.698262
8	5.112915	2.366028	3.270035	0.157120	0.542257	0.798496
∞	5.655172	2.620690	3.655172	0	0	$0.750000 = \rho(\mathbf{J})$

Passando para o segundo membro $\mathbf{L}\mathbf{x}^{(k+1)}$, fica

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

Admitindo novamente que todos os $a_{ii} \neq 0$, temos

$$x_i^{(k+1)} = (b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij} x_j^{(k)})/a_{ii} \qquad i = 1, 2, \ldots, n \quad k = 0, 1, \ldots \tag{3.28}$$

Notamos que as componentes $x_1^{(k+1)}, \ldots, x_{i-1}^{(k+1)}$ do vector $\mathbf{x}^{(k+1)}$ presentes no segundo membro de (3.28) já foram previamente determinadas quando se pretende calcular a componente $x_i^{(k+1)}$. Assim no método de Gauss-Seidel no cálculo da componente $x_i^{(k+1)}$ utilizamos os valores aproximados das componentes de $\mathbf{x}$ mais actualizados, enquanto que no método de Jacobi utilizamos somente as componentes de $\mathbf{x}^{(k)}$ enquanto não forem calculadas todas as componentes de $\mathbf{x}^{(k+1)}$. Por esta razão é usual denominar o método de *Jacobi* por método das *substituições simultáneas* enquanto que o de *Gauss-Seidel* por método das *substituições sucessivas*. Como no método de Gauss-Seidel aproveitamos logo o valor da última componente calculada de $\mathbf{x}^{(k+1)}$ no cálculo da componente seguinte, normalmente este método converge mais rapidamente do que o de Jacobi, no caso dos 2 convergirem. Podiamos, também, chegar a esta conclusão se notássemos que no método de Gauss-Seidel o sistema a resolver $\mathbf{M}\mathbf{x}^{(k+1)} = \mathbf{b} - \mathbf{N}\mathbf{x}^{(k)}$ tem uma matriz $\mathbf{M}$ que difere menos da matriz $\mathbf{A}$ do sistema a resolver do que no caso do método de Jacobi. Notamos também que no método de Gauss-Seidel, depois do cálculo de $x_i^{(k+1)}$, basta armazenar na memória do computador o vector $[x_1^{(k+1)}, \ldots, x_{i-1}^{(k+1)}, x_i^{(k+1)}, x_{i+1}^{(k)}, \ldots, x_n^{(k)}]^T$ enquanto que no método de Jacobi é necessário armazenar simultaneamente os vectores $\mathbf{x}^{(k)}$ e $\mathbf{x}^{(k+1)}$.

Exemplo 3.11 *Aplicar o método de Gauss-Seidel à resolução do sistema* $\mathbf{A}\mathbf{x} = \mathbf{b}$ *com*

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & -1 \\ -3/4 & 2 & 0 \\ -1 & -1/4 & 2 \end{bmatrix} \qquad \mathbf{b} = \begin{bmatrix} 2 \\ 1 \\ 1 \end{bmatrix} \qquad \mathbf{x}^{(0)} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

Tabela 3.2: Exemplo do método de Gauss-Seidel

k	$x_1^{(k)}$	$x_2^{(k)}$	$x_3^{(k)}$	$\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\|_\infty$	$\|\mathbf{e}^{(k)}\|_\infty$	$\|\mathbf{e}^{(k)}\|_\infty / \|\mathbf{e}^{(k-1)}\|_\infty$
0	0.000000	0.000000	0.000000		5.655172	
1	2.000000	1.250000	1.65625	2.000000	3.655172	0.646341
2	3.656250	1.871094	2.562012	1.656250	1.998922	0.546875
3	4.562012	2.210754	3.057350	0.905762	1.093161	0.546875
4	5.057350	2.396506	3.328238	0.495338	0.597822	0.546875
5	5.328238	2.498089	3.476380	0.270888	0.326934	0.546875
6	5.476380	2.553643	3.557396	0.148142	0.178792	0.546875
7	5.557395	2.584023	3.601701	0.081015	0.097777	0.546875
8	5.601701	2.600638	3.625930	0.044306	0.053472	0.546875
∞	5.655172	2.620690	3.655172	0	0	$0.546875 = \rho(\mathbf{S})$

considerados anteriormente.

Tal como no método de Jacobi, começamos por evidenciar as sucessivas incógnitas ficando o sistema na forma

$$\begin{aligned} x_1 &= 2 - 0x_2 + x_3 \\ x_2 &= (1 + 3/4x_1 - 0x_3)/2 \\ x_3 &= (1 + x_1 + 1/4x_2)/2 \end{aligned}$$

O método iterativo de Gauss-Seidel é neste caso dado por

$$\begin{aligned} x_1^{(k+1)} &= 2 - 0x_2^{(k)} + x_3^{(k)} \\ x_2^{(k+1)} &= (1 + 3/4x_1^{(k+1)} - 0x_3^{(k)})/2 \qquad k = 0, 1, \ldots \\ x_3^{(k+1)} &= (1 + x_1^{(k+1)} + 1/4x_2^{(k+1)})/2 \end{aligned}$$

Tal como vimos, no segundo membro da segunda equação está presente $x_1^{(k+1)}$ calculado na equação anterior e no segundo membro da terceira equação estão presentes $x_1^{(k+1)}$ e $x_2^{(k+1)}$ calculados nas equações anteriores. Com a escolha

$$x_1^{(0)} = 0 \qquad x_2^{(0)} = 0 \qquad x_3^{(0)} = 0$$

obtemos para as 2 primeiras iterações:

$$\begin{aligned} x_1^{(1)} &= 2 + 0 = 2 & x_1^{(2)} &= 2 + 53/32 = 117/32 \\ x_2^{(1)} &= (1 + 3/2)/2 = 5/4 & x_2^{(2)} &= (1 + 351/128)/2 = 479/256 \\ x_3^{(1)} &= (1 + 2 + 5/16)/2 = 53/32 & x_3^{(2)} &= (1 + 117/32 + 479/1024)/2 = 5247/2048 \end{aligned}$$

Utilizando as informações dadas no problema anterior apresentamos na Tabela 3.2 os resultados obtidos efectuando 8 iteracões. Destes resultados, vemos que para este sistema o método de Gauss-Seidel converge mais rapidamente que o método de Jacobi.

■

3.5.2 Critérios de convergência dos métodos iterativos

Para fazer um estudo da convergência dos métodos iterativos vamos necessitar de utilizar novamente normas de matrizes. Além dos conceitos introduzidos atrás, necessitamos da noção de *matriz convergente*. É toda a matriz para a qual

$$\lim_{k\to\infty} \mathbf{A}^k = \mathbf{O}$$

($\mathbf{O}$ designa a matriz $n \times n$ de elementos nulos). Podemos, então, enunciar o seguinte

Teorema 3.3 *(i)* $\mathbf{A}$ *é convergente sse* $\rho(\mathbf{A}) < 1$, *i.e., sse todos os valores próprios de* $\mathbf{A}$ *são menores do que um em valor absoluto. (ii)* $\mathbf{A}$ *é convergente sse para alguma norma se verifica que* $\|\mathbf{A}\| < 1$. ■

A segunda parte deste teorema (que é consequência do Teorema 3.2, apresentado atrás, e da primeira parte do Teorema 3.3, agora apresentado) fornece-nos um critério para verificar se uma matriz é convergente, que em geral é de aplicação mais fácil do que verificar se o raio espectral é inferior a unidade. Nomeadamente, o cálculo das normas $\|\mathbf{A}\|_\infty$ ou $\|\mathbf{A}\|_1$ é mais fácil do que o cálculo do raio espectral $\rho(\mathbf{A})$ e por conseguinte se verificarmos que uma dessas normas é inferior a unidade concluimos imediatamente que $\mathbf{A}$ é convergente.

3.5.2.1 Convergência dos métodos iterativos

Consideremos novamente a forma geral dos métodos iterativos dada por (3.26)

$$\mathbf{M}\mathbf{x}^{(k+1)} = \mathbf{b} - \mathbf{N}\mathbf{x}^{(k)} \qquad k = 0, 1, \ldots$$

ou por (3.27)

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \mathbf{M}^{-1}(\mathbf{b} - \mathbf{A}\mathbf{x}^{(k)}) \qquad k = 0, 1, \ldots$$

Destas relações e definindo a matriz

$$\mathbf{C} = -\mathbf{M}^{-1}\mathbf{N} = \mathbf{I} - \mathbf{M}^{-1}\mathbf{A} \tag{3.29}$$

podemos escrever

$$\mathbf{x}^{(k+1)} = \mathbf{C}\mathbf{x}^{(k)} + \mathbf{M}^{-1}\mathbf{b} \qquad k = 0, 1, \ldots \tag{3.30}$$

Como a solução $\mathbf{x}$ de $\mathbf{A}\mathbf{x} = \mathbf{b}$ satisfaz a equação

$$\mathbf{x} = \mathbf{C}\mathbf{x} + \mathbf{M}^{-1}\mathbf{b}$$

se subtrairmos a esta a equação anterior obtemos

$$\mathbf{e}^{(k+1)} = \mathbf{C}\mathbf{e}^{(k)} \qquad k = 0, 1, \ldots \tag{3.31}$$

sendo

$$\mathbf{e}^{(k)} = \mathbf{x} - \mathbf{x}^{(k)}$$

o vector erro. Aplicando sucessivamente esta equação com $k = 0, 1, \ldots$ obtemos

$$\mathbf{e}^{(k)} = \mathbf{C}^k\mathbf{e}^{(0)} \qquad k = 1, 2, \ldots \tag{3.32}$$

onde $\mathbf{e}^{(0)}$ é o erro inicial. Então, é evidente que uma *condição suficiente de convergência do método iterativo*, *i.e.*, que

$$\lim_{k\to\infty} \mathbf{e}^{(k)} = \mathbf{o}$$

é que $\mathbf{C}$ *seja convergente*, *i.e.*, que

$$\lim_{k\to\infty} \mathbf{C}^k = \mathbf{O}$$

e isto também é *necessário se o método for convergente para qualquer* $\mathbf{e}^{(0)}$. Atendendo ao Teorema 3.3, vemos assim que, um método iterativo é convergente se os valores próprios de $\mathbf{C}$ tiverem todos módulos inferior a um.

Teorema 3.4 *Qualquer que seja o vector* $\mathbf{x}^{(0)}$*, o método iterativo dado por (3.30) converge (i) sse* $\rho(\mathbf{C}) < 1$ *(ii) sse existir uma norma tal que* $\|\mathbf{C}\| < 1$. ■

Como é sabido, duas matrizes semelhantes têm os mesmos valores próprios. Portanto dado uma matriz não singular $\mathbf{S}$ tem-se que $\rho(\mathbf{SCS}^{-1}) = \rho(\mathbf{C})$. Sendo assim e como corolário do Teorema 3.4 temos o seguinte critério de convergência, que é importante na prática:

Teorema 3.5 *Para que o método iterativo dado por (3.30) converge, qualquer que seja o vector* $\mathbf{x}^{(0)}$ *basta que se verifique uma das seguintes condições:*

1. $\|\mathbf{C}\|_\infty < 1$
2. $\|\mathbf{C}\|_1 < 1$
3. $\|\mathbf{SCS}^{-1}\|_\infty < 1$
4. $\|\mathbf{SCS}^{-1}\|_1 < 1$

em que $\mathbf{S}$ *é uma matriz não singular.* ■

De (3.32) temos que

$$\|\mathbf{e}^{(k)}\| \leq \|\mathbf{C}\|^k \|\mathbf{e}^{(0)}\|$$

e portanto quanto menor for $\|\mathbf{C}\|$ será, em princípio, mais rápida a convergência. Notando que $\mathbf{C}$ é definida por (3.29), vemos que $\|\mathbf{C}\|$ será tanto mais pequena quanto mais $\mathbf{M}$ se aproximar de $\mathbf{A}$. Mais uma vez se explica porquê que normalmente o método de Gauss-Seidel ($\mathbf{M} = \mathbf{L} + \mathbf{D}$) converge mais rapidamente do que o de Jacobi ($\mathbf{M} = \mathbf{D}$). Também se explica a rápida convergência normalmente encontrada no método da correcção residual visto que para este método $\mathbf{M} = \mathbf{P}^{-1}\tilde{\mathbf{L}}\tilde{\mathbf{U}}$ constitui normalmente uma excelente aproximação de $\mathbf{A}$.

De (3.31) concluimos que

$$\|\mathbf{e}^{(k+1)}\|_\infty \leq c\|\mathbf{e}^{(k)}\|_\infty$$

onde $c = \|\mathbf{C}\|_\infty$. Podemos então escrever

$$\begin{aligned}\|\mathbf{x} - \mathbf{x}^{(k+1)}\|_\infty &\leq c\|\mathbf{x} - \mathbf{x}^{(k)}\|_\infty = c\|\mathbf{x} - \mathbf{x}^{(k+1)} + \mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\|_\infty \\ &\leq c\|\mathbf{x} - \mathbf{x}^{(k+1)}\|_\infty + c\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\|_\infty\end{aligned}$$

Se $c < 1$ obtemos a seguinte majoração para o erro

$$\|\mathbf{x} - \mathbf{x}^{(k+1)}\|_\infty \leq \frac{c}{1-c}\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\|_\infty$$

Se $c \leq 1/2$ então impor como critério de paragem das iterações

$$\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\|_\infty \leq \epsilon$$

i.e.,

$$|x_i^{(k+1)} - x_i^{(k)}| \leq \epsilon \qquad i = 1, \ldots, n \tag{3.33}$$

garante-nos que $\|\mathbf{x} - \mathbf{x}^{(k+1)}\|_\infty \leq \epsilon$ e portanto que todas as componentes do vector $\mathbf{x}$ são calculadas com um erro absoluto inferior a ϵ. Quando $c \approx 1$ a convergência é lenta e o erro $\|\mathbf{x} - \mathbf{x}^{(k+1)}\|_\infty$ pode ser muito superior a $\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\|_\infty$. Para ter uma indicação do valor de c podemos calcular o quociente

$$c_{k+1} = \frac{\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\|_\infty}{\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\|_\infty}$$

ou com maior segurança o máximo dos valores de c_{k+1} para várias iterações sucessivas. Esta estimativa de c baseia-se no facto de $c_{k+1} \leq c$. Com efeito, utilizando (3.31) podemos mostrar que

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

e portanto aplicando normas obtemos $c_{k+1} \leq c$.

3.5.2.2 Convergência dos métodos de Jacobi e Gauss-Seidel

O último critério de convergência acima referido (Teorema 3.5) é facilmente aplicável no método de Jacobi. Com efeito, para este método temos, como vimos, $\mathbf{M} = \mathbf{D}$ e portanto a matriz $\mathbf{C} = -\mathbf{M}^{-1}\mathbf{N}$, que para o método de Jacobi passamos a designar por $\mathbf{J}$, é dada por

$$\mathbf{J} = -\mathbf{D}^{-1}(\mathbf{A} - \mathbf{D}) = -\mathbf{D}^{-1}(\mathbf{L} + \mathbf{U})$$

ou ainda por

$$\mathbf{J} = \begin{bmatrix} 0 & \frac{-a_{12}}{a_{11}} & . & . & . & \frac{-a_{1n}}{a_{11}} \\ \frac{-a_{21}}{a_{22}} & 0 & \frac{-a_{23}}{a_{22}} & . & . & \frac{-a_{2n}}{a_{22}} \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ \frac{-a_{n-1,1}}{a_{n-1,n-1}} & . & . & \frac{-a_{n-1,n-2}}{a_{n-1,n-1}} & 0 & \frac{-a_{n-1,n}}{a_{n-1,n-1}} \\ \frac{-a_{n1}}{a_{n,n}} & . & . & . & \frac{-a_{n,n-1}}{a_{n,n}} & 0 \end{bmatrix}$$

Vemos imediatamente que, para o caso do método de Jacobi, a condição $\|\mathbf{C}\|_\infty < 1$ reduz-se a matriz $\mathbf{A}$ ser de diagonal estritamente dominante por *linhas*, *i.e.*,

$$\sum_{j=1, j\neq i}^{n} |a_{ij}| < |a_{ii}| \qquad i = 1, \ldots, n$$

e a condição $\|\mathbf{DCD}^{-1}\|_1 < 1$ reduz-se a matriz $\mathbf{A}$ ser de diagonal estritamente dominante por *colunas*, *i.e.*,

$$\sum_{i=1, i\neq j}^{n} |a_{ij}| < |a_{jj}| \qquad j = 1, \ldots, n$$

Assim podemos enunciar o seguinte:

Teorema 3.6 *Se* $\mathbf{A}$ *for de diagonal estritamente dominante por linhas ou por colunas, o método de Jacobi converge, independentemente do vector inicial* $\mathbf{x}^{(0)}$ *escolhido.* ■

De notar que qualquer matriz de diagonal estritamente dominante é sempre invertível.

Para considerar outro critério de convergência vamos introduzir a seguinte definição.

Definição 3.6 *Uma matriz* $\mathbf{A}$ *é irredutível se não for possível encontrar uma matriz de permutação* $\mathbf{P}$ *tal que* $\mathbf{P}^T\mathbf{AP}$ *tenha a forma*

$$\mathbf{P}^T\mathbf{AP} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{O} & \mathbf{A}_{22} \end{bmatrix} \tag{3.34}$$

onde $\mathbf{A}_{11}$ *é uma matriz* $p \times p$, $\mathbf{A}_{22}$ *uma matriz* $q \times q$ *com* $p + q = n$ *e* $\mathbf{O}$ *uma matriz* $q \times p$ *com elementos todos nulos.*

Um exemplo de uma matriz irredutível é a considerada nos Exemplos 3.10 e 3.11

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & -1 \\ -3/4 & 2 & 0 \\ -1 & -1/4 & 2 \end{bmatrix} \tag{3.35}$$

Com efeito se considerarmos a matriz de permutação

$$\mathbf{P} = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}$$

obtemos

$$\mathbf{P}^T\mathbf{AP} = \begin{bmatrix} 2 & -1/4 & -1 \\ 0 & 2 & -3/4 \\ -1 & 0 & 1 \end{bmatrix}$$

que não tem a forma do segundo membro de (3.34). À mesma conclusão chegaríamos se tivéssemos utilizado qualquer outra matriz de permutação. Para verificar a irredutibilidade de uma matriz podemos recorrer ao seguinte critério: Uma matriz $\mathbf{A}$

é irredutível sse, dados quaisquer índices $i = 1, \ldots, n$ e $j = 1, \ldots, n$, existe uma sequência de coeficientes de $\mathbf{A}$ não nulos da forma $\{a_{ir}, a_{rs}, \ldots, a_{tu}, a_{uj}\}$. Assim para a matriz definida por (3.35) temos as seguintes sequências, todas constituidas por elementos não nulos:

$$\begin{array}{ccc} \{a_{11}\} & \{a_{13}, a_{32}\} & \{a_{13}\} \\ \{a_{21}\} & \{a_{22}\} & \{a_{21}, a_{13}\} \\ \{a_{31}\} & \{a_{32}\} & \{a_{33}\} \end{array}$$

Deste critério podemos concluir que se todos os elementos de $\mathbf{A}$ adjacentes à diagonal principal forem diferentes de zero, *i.e.*, se $a_{i,i+1} \neq 0$ e $a_{i+1,i} \neq 0$ para $i = 1, \ldots, n-1$, então $\mathbf{A}$ é irredutível. Este tipo de matriz aparece frequentemente na resolução de equações diferenciais. Se existir um k tal que todos os elementos da linha e da coluna de ordem k forem diferentes de zero, então podemos também concluir que a matriz é irredutível para este caso. Vamos ainda introduzir a seguinte:

Definição 3.7 *(i) Uma matriz* $\mathbf{A}$ *diz-se de diagonal dominante por linhas se*

$$\sum_{j=1, j\neq i}^{n} |a_{ij}| \leq |a_{ii}| \qquad i = 1, \ldots, n$$

e se existe pelo menos um k *tal que*

$$\sum_{j=1, j\neq k}^{n} |a_{kj}| < |a_{kk}|$$

(ii) Uma matriz $\mathbf{A}$ *diz-se de diagonal dominante por colunas se*

$$\sum_{i=1, i\neq j}^{n} |a_{ij}| \leq |a_{jj}| \qquad j = 1, \ldots, n$$

e se existe pelo menos um k *tal que*

$$\sum_{i=1, i\neq k}^{n} |a_{ik}| < |a_{kk}|$$

De notar que qualquer matriz de diagonal dominante é sempre invertível. Podemos agora enunciar o seguinte:

Teorema 3.7 *Se* $\mathbf{A}$ *for irredutível e de diagonal dominante por linhas ou por colunas, o método de Jacobi converge, independentemente do vector inicial* $\mathbf{x}^{(0)}$ *escolhido.* ■

Visto que a matriz dada por (3.35) é de diagonal dominante por linhas, e como vimos é irredutível, podemos neste caso garantir a convergência do método de Jacobi. Notamos que esta matriz não é de diagonal *estritamente* dominante por linhas nem é de diagonal dominante por *colunas*.

Como vimos no estudo da convergência do método de Jacobi, o Teorema 3.6 é consequência imediata do Teorema 3.5 onde se supõe que $\|\mathbf{C}\|_\infty < 1$ ou $\|\mathbf{C}\|_1 < 1$

sendo $\mathbf{C}$ a matriz definida por (3.29). No caso do método de Gauss-Seidel esta matriz, que passamos a designar por $\mathbf{S}$, reduz-se a

$$\mathbf{S} = -(\mathbf{L} + \mathbf{D})^{-1}\mathbf{U} \tag{3.36}$$

e portanto já não é imediato a aplicação do referido critério. No entanto, é possível demonstrar para o método de Gauss-Seidel um teorema perfeitamente análogo ao do Teorema 3.6.

Teorema 3.8 *Se* $\mathbf{A}$ *for de diagonal estritamente dominante por linhas ou por colunas, o método de Gauss-Seidel converge, independentemente do vector inicial* $\mathbf{x}^{(0)}$ *escolhido.*
■

Também é válido o seguinte:

Teorema 3.9 *Se* $\mathbf{A}$ *for irredutível e de diagonal dominante por linhas ou por colunas, o método de Gauss-Seidel converge, independentemente do vector inicial* $\mathbf{x}^{(0)}$ *escolhido.*
■

Em geral, da convergência de um dos dois métodos não se pode inferir da convergência do outro; e se os dois convergir não se pode, com toda a generalidade, concluir que é o de Gauss-Seidel que converge mais rapidamente. No entanto, existe um caso particular importante para o qual é possível relacionar a rapidez dos dois métodos. Passamos a designar por *matriz não negativa (não positiva)* uma matriz de elementos não negativos (não positivos).

Teorema 3.10 *Se a matriz* $\mathbf{J} = -\mathbf{D}^{-1}(\mathbf{L} + \mathbf{U})$ *é não negativa, então para* $\mathbf{J}$ *e* $\mathbf{S} = -(\mathbf{L} + \mathbf{D})^{-1}\mathbf{U}$ *existe uma das seguintes relações:*

1. $\rho(\mathbf{S}) = \rho(\mathbf{J}) = 0$
2. $0 < \rho(\mathbf{S}) < \rho(\mathbf{J}) < 1$
3. $\rho(\mathbf{S}) = \rho(\mathbf{J}) = 1$
4. $\rho(\mathbf{S}) > \rho(\mathbf{J}) > 1$

■

A condição $\mathbf{J} \geq 0$ é satisfeita em particular se a matriz $\mathbf{A}$ tem os elementos da diagonal principal todos positivos e os restantes elementos não positivos: $a_{ii} > 0$, $a_{ij} \leq 0$, $i \neq j$ (como p.ex. a matriz $\mathbf{A}$ considerada nos Exemplos 3.10 e 3.11). Do Teorema 3.10, concluimos que para estas matrizes se um dos métodos converge, então o outro também converge e o método de Gauss-Seidel converge mais rapidamente do que o de Jacobi, visto que a rapidez de convergência está relacionada com o raio espectral.

No caso da matriz considerada nos Exemplos 3.10 e 3.11, dado por (3.35), correpondem os raios espectrais $\rho(\mathbf{J}) = 0.75$ e $\rho(\mathbf{S}) = 0.546875$ o que está conforme com o Teorema 3.10 e com os resultados numéricos obtidos anteriormente (ver Tabelas 3.1 e 3.2).

3.5.3 Método de relaxacção SOR

Acontece com frequência nas aplicações que, para uma dada matriz $\mathbf{A}$, seja $\rho(\mathbf{S}) \approx 1$ e portanto o método de Gauss-Seidel converge muito lentamente quando for $\rho(\mathbf{S}) < 1$. Nestes casos é usual modificar esse método introduzindo um *parâmetro de aceleração* ω. Assim em vez de fazer $\mathbf{M} = \mathbf{L} + \mathbf{D}$, passa-se a tomar por matriz $\mathbf{M}$ a matriz

$$\mathbf{M}(\omega) = \frac{1}{\omega}(\omega\mathbf{L} + \mathbf{D}) = \mathbf{L} + \frac{1}{\omega}\mathbf{D} \tag{3.37}$$

e por conseguinte a matriz $\mathbf{N}$ passa a ser

$$\mathbf{N}(\omega) = \frac{1}{\omega}((\omega - 1)\mathbf{D} + \omega\mathbf{U}) = (1 - \frac{1}{\omega})\mathbf{D} + \mathbf{U} \tag{3.38}$$

Somos assim conduzido a um *método de relaxacção* conhecido na literatura anglo-saxónica por SOR ("successive overrelaxation"):

$$(\omega\mathbf{L} + \mathbf{D})\mathbf{x}^{(k+1)} = \omega\mathbf{b} - ((\omega - 1)\mathbf{D} + \omega\mathbf{U})\mathbf{x}^{(k)} \qquad k = 0, 1, \ldots$$

Passando para o segundo membro $\omega\mathbf{L}\mathbf{x}^{(k+1)}$, fica

$$\mathbf{D}\mathbf{x}^{(k+1)} = \omega(\mathbf{b} - \mathbf{L}\mathbf{x}^{(k+1)} - \mathbf{U}\mathbf{x}^{(k)}) + (1 - \omega)\mathbf{D}\mathbf{x}^{(k)} \qquad k = 0, 1, \ldots$$

Admitindo todos os $a_{ii} \neq 0$, podemos multiplicar a equação por $\mathbf{D}^{-1}$ obtendo

$$\mathbf{x}^{(k+1)} = \omega\mathbf{z}^{(k+1)} + (1 - \omega)\mathbf{x}^{(k)} \qquad k = 0, 1, \ldots$$

com

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

Assim cada componente $x_i^{(k+1)}$ do vector $\mathbf{x}^{(k+1)}$ é obtida a partir das seguintes relações:

$$z_i^{(k+1)} = (b_i - \textstyle\sum_{j=1}^{i-1} a_{ij}x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij}x_j^{(k)})/a_{ii}$$

$$x_i^{(k+1)} = \omega z_i^{(k+1)} + (1 - \omega)x_i^{(k)} \qquad i = 1, 2, \ldots, n \quad k = 0, 1, \ldots$$

Vemos que quando $\omega = 1$ este método reduz-se ao método de Gauss-Seidel. Consoante seja $\omega > 1$ ou $\omega < 1$ trata-se de sobrerelaxacção (caso mais frequente) ou subrelaxacção.

Exemplo 3.12 *Aplicar o método de Gauss-Seidel com relaxacção (SOR) à resolução do sistema* $\mathbf{A}\mathbf{x} = \mathbf{b}$ *com*

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & -1 \\ -3/4 & 2 & 0 \\ -1 & -1/4 & 2 \end{bmatrix} \qquad \mathbf{b} = \begin{bmatrix} 2 \\ 1 \\ 1 \end{bmatrix} \qquad \mathbf{x}^{(0)} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

considerado anteriormente. Tomar $\omega = 1.2$

Tal como nos métodos de Jacobi e Gauss-Seidel, começamos por evidenciar as sucessivas incógnitas ficando o sistema na forma

$$\begin{aligned} x_1 &= 2 - 0x_2 + x_3 \\ x_2 &= (1 + 3/4x_1 - 0x_3)/2 \\ x_3 &= (1 + x_1 + 1/4x_2)/2 \end{aligned}$$

Temos então para este caso

$$\begin{aligned} z_1^{(k+1)} &= 2 - 0x_2^{(k)} + x_3^{(k)} \\ x_1^{(k+1)} &= \omega z_1^{(k+1)} + (1-\omega)x_1^{(k)} \\ z_2^{(k+1)} &= (1 + 3/4x_1^{(k+1)} - 0x_3^{(k)})/2 \\ x_2^{(k+1)} &= \omega z_2^{(k+1)} + (1-\omega)x_2^{(k)} \\ z_3^{(k+1)} &= (1 + x_1^{(k+1)} + 1/4x_2^{(k+1)})/2 \\ x_3^{(k+1)} &= \omega z_3^{(k+1)} + (1-\omega)x_3^{(k)} \end{aligned} \qquad k = 0, 1, \ldots$$

Com a escolha $x_1^{(0)} = 0, x_2^{(0)} = 0, x_3^{(0)} = 0$, obtemos para as 2 primeiras iterações:

$$\begin{aligned} z_1^{(1)} &= 2 + 0 = 2 \\ x_1^{(1)} &= 1.2 \times 2 - 0.2 \times 0 = 2.4 \\ z_2^{(1)} &= (1 + 3/4 \times 2.4)/2 = 1.4 \\ x_2^{(1)} &= 1.2 \times 1.4 - 0.2 \times 0 = 1.68 \\ z_3^{(1)} &= (1 + 2.4 + 1/4 \times 1.68)/2 = 1.91 \\ x_3^{(1)} &= 1.2 \times 1.91 - 0.2 \times 0 = 2.292 \end{aligned}$$

$$\begin{aligned} z_1^{(2)} &= 2 + 2.292 = 4.292 \\ x_1^{(2)} &= 1.2 \times 4.292 - 0.2 \times 2.4 = 4.6704 \\ z_2^{(2)} &= (1 + 3/4 \times 4.6704)/2 = 2.2514 \\ x_2^{(2)} &= 1.2 \times 2.2514 - 0.2 \times 1.68 = 2.36568 \\ z_3^{(2)} &= (1 + 4.6704 + 1/4 \times 2.36568)/2 = 3.13091 \\ x_3^{(2)} &= 1.2 \times 3.13091 - 0.2 \times 2.292 = 3.298692 \end{aligned}$$

Apresentamos na Tabela 3.3 os resultados obtidos efectuando 8 iteracões. Comparando estes resultados com os da Tabela 3.2, vemos que para este sistema o método de Gauss-Seidel converge mais rapidamente com relaxacção, com $\omega = 1.2$, do que sem relaxacção ($\omega = 1$). ■

Tal como para os outros métodos iterativos a convergência está relacionada com o raio espectral da matriz $\mathbf{C}$ definida por (3.29). Atendendo a (3.37) e (3.38), vemos que no caso do método de relaxacção esta matriz é

$$\mathbf{H}(\omega) = -(\omega\mathbf{L} + \mathbf{D})^{-1}((\omega - 1)\mathbf{D} + \omega\mathbf{U}) = \mathbf{I} - \omega(\omega\mathbf{L} + \mathbf{D})^{-1}\mathbf{A}$$

Tabela 3.3: Exemplo do método de Gauss-Seidel com relaxacção

k	$x_1^{(k)}$	$x_2^{(k)}$	$x_3^{(k)}$	$\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\|_\infty$	$\|\mathbf{e}^{(k)}\|_\infty$	$\|\mathbf{e}^{(k)}\|_\infty / \|\mathbf{e}^{(k-1)}\|_\infty$
0	0.000000	0.000000	0.000000		5.655172	
1	2.400000	1.680000	2.292000	2.400000	3.255172	0.575610
2	4.670400	2.365680	3.298692	2.270400	0.984772	0.302525
3	5.424350	2.567822	3.580045	0.753950	0.230822	0.234391
4	5.611184	2.611468	3.642422	0.186834	0.043988	0.190573
5	5.648669	2.619607	3.653658	0.037485	0.006503	0.147839
6	5.654656	2.620674	3.655163	0.005987	0.000516	0.079390
7	5.655264	2.620734	3.655236	0.000608	0.000092	0.178301
8	5.655231	2.620707	3.655197	0.000039	0.000058	0.631553
∞	5.655172	2.620690	3.655172	0	0	

Quando $\omega = 1$ esta matriz reduz-se a matriz $\mathbf{S}$ definida por (3.36).

Como a convergência do método de relaxação depende de $\rho(\mathbf{H}(\omega))$ (raio espectral da matriz $\mathbf{H}(\omega)$) vamos apresentar alguns teoremas relacionados com este raio espectral. O 1º diz-nos que o método de relaxação só pode convergir para valores do parâmetro de relaxação com $0 < \omega < 2$.

Teorema 3.11 *Para qualquer matriz quadrada* $\mathbf{A}$ *tem-se*

$$\rho(\mathbf{H}(\omega)) \geq |\omega - 1|$$

para todo o ω *real. Portanto*

$$\rho(\mathbf{H}(\omega)) \geq 1$$

para $\omega \notin (0, 2)$. ■

Teorema 3.12 *Se a matriz* $\mathbf{A}$ *é irredutível e a matriz* $\mathbf{J} = -\mathbf{D}^{-1}(\mathbf{L} + \mathbf{U})$ *é não negativa, e se o método de Jacobi converge, i.e.,* $\rho(\mathbf{J}) < 1$*, então* $\rho(\mathbf{H}(\omega))$ *é uma função estritamente decrescente no intervalo* $0 < \omega \leq \omega_0$ *em que* $1 \leq \omega_0 < 2$. ■

Isto significa que nestas condições somente a sobrerelaxação ($\omega > 1$) pode dar mais rápida convergência do que o método de Gauss-Seidel ($\omega = 1$).

Teorema 3.13 *Para matrizes* $\mathbf{A}$ *simétricas definidas positivas tem-se*

$$\rho(\mathbf{H}(\omega)) < 1$$

para qualquer $0 < \omega < 2$*, i.e., o método de relaxação para estas matrizes converge sse* $0 < \omega < 2$*. Em particular o método de Gauss-Seidel converge para matrizes simétricas definidas positivas.* ■

Vejamos agora como se comporta a função $\rho(\mathbf{H}(\omega))$ para o caso da matriz considerada no Exemplo 3.12. Representamos esta função na Figura 3.1. Desta figura

Figura 3.1: Raio espectral em função do parâmetro de relaxacção

Figura 3.2: Convergência do método de relaxacção

vemos que, para este exemplo, $\rho(\mathbf{H}(\omega))$ atinge um mínimo aproximadamente em $\omega = 1.2$ (valor escolhido no Exemplo 3.12, para o qual corresponde $\rho(\mathbf{H}(1.2)) = 0.210917$).

Como é difícil determinar o raio espectral, para estimar na prática o valor óptimo do parâmetro de aceleração ω costuma-se comparar o número de iterações para atingir uma certa precisão. Assim utilizando o critério dado por (3.33), *i.e.*, impondo que as diferenças entre as componentes das 2 últimas iteradas sejam em módulo todas inferior a um certo valor ϵ, determina-se o valor de ω a que corresponde o menor número de iterações. Este procedimento para determinar o ω óptimo só é compensador do ponto de vista computacional quando tivermos de resolver várias vezes sistemas com a mesma matriz $\mathbf{A}$ mas com diferentes 2º membros $\mathbf{b}$ (este tipo de situação ocorre frequentemente, p.ex. na resolução de equações diferenciais).

Para o caso da matriz considerada no Exemplo 3.12 e escolhendo em (3.33) a tolerância $\epsilon = 10^{-3}$ obtem-se os resultados representados na Figura 3.2. Desta figura confirmamos que o valor óptimo do parâmetro de aceleração é aproximadamente $\omega = 1.2$.

3.6 Sistemas de equações não lineares

Consideremos um sistema de n equações não lineares a n incógnitas:

$$\begin{array}{l} f_1(x_1, x_2, \ldots, x_n) = 0 \\ f_2(x_1, x_2, \ldots, x_n) = 0 \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ f_n(x_1, x_2, \ldots, x_n) = 0 \end{array}$$

Definindo os vectores $\mathbf{x} = [x_1, x_2, \ldots, x_n]^T$ e $\mathbf{f} = [f_1, f_2, \ldots, f_n]^T$ podemos escrever o sistema na forma matricial

$$\mathbf{f}(\mathbf{x}) = \mathbf{o} \tag{3.39}$$

Tal como para o caso de uma equação não linear a uma incógnita podemos escrever (3.39) na forma

$$\mathbf{x} = \mathbf{g}(\mathbf{x}) \tag{3.40}$$

em que $\mathbf{g} = [g_1, g_2, \ldots, g_n]^T$ é construido de modo que as equações (3.39) e (3.40) têm as mesmas soluções. Por exemplo, podemos escrever o vector $\mathbf{g}(\mathbf{x})$ na forma

$$\mathbf{g}(\mathbf{x}) = \mathbf{x} - \mathbf{A}(\mathbf{x})\mathbf{f}(\mathbf{x}) \tag{3.41}$$

onde $\mathbf{A}(\mathbf{x})$ é uma matriz quadrada de ordem n não singular com componentes $a_{ij}(\mathbf{x})$.

Escolhido a forma da matriz $\mathbf{A}(\mathbf{x})$, somos conduzidos à um método iterativo para a resolução de (3.39), bastando para isso calcular sucessivamente

$$\mathbf{x}^{(k+1)} = \mathbf{g}(\mathbf{x}^{(k)}) \quad k = 0, 1, \ldots \tag{3.42}$$

em que $\mathbf{x}^{(0)}$ é um vector escolhido previamente. A convergência de um método iterativo pode estudar-se utilizando a seguinte variante do teorema do ponto fixo.

Teorema 3.14 *Seja* $\mathbf{z}$ *uma solução de (3.40). Suponhamos que as componentes* $g_i(\mathbf{x})$ *possuem primeiras derivadas parciais contínuas e satisfazem a condição*

$$\max_{1\le i\le n}\sum_{j=1}^{n}|\partial g_i(\mathbf{x})/\partial x_j| \le c < 1 \tag{3.43}$$

para todo o $\mathbf{x}$ *em*

$$B_\epsilon(\mathbf{z}) = \{\mathbf{x} : \|\mathbf{x}-\mathbf{z}\|_\infty \le \epsilon\}$$

Então: Para qualquer $\mathbf{x}^{(0)} \in B_\epsilon(\mathbf{z})$*, todas as iteradas* $\mathbf{x}^{(k)}$ *de (3.42) também pertencem a* $B_\epsilon(\mathbf{z})$ *e convergem para a solução* $\mathbf{z}$ *de (3.40) a qual é única em* $B_\epsilon(\mathbf{z})$*.*

Demonstração: Para quaisquer $\mathbf{x}, \mathbf{y} \in B_\epsilon(\mathbf{z})$ da fórmula de Taylor temos

$$\mathbf{g}(\mathbf{x}) - \mathbf{g}(\mathbf{y}) = \mathbf{G}(\xi)(\mathbf{x}-\mathbf{y})$$

em que $\mathbf{G}$ é a matriz Jacobiana de $\mathbf{g}$

$$\mathbf{G} = \begin{bmatrix} \partial g_1/\partial x_1 & \partial g_1/\partial x_2 & \cdots & \partial g_1/\partial x_n \\ \partial g_2/\partial x_1 & \partial g_2/\partial x_2 & \cdots & \partial g_2/\partial x_n \\ \cdots & \cdots & \cdots & \cdots \\ \partial g_n/\partial x_1 & \partial g_n/\partial x_2 & \cdots & \partial g_n/\partial x_n \end{bmatrix}$$

calculada num ponto intermédio

$$\xi = \mathbf{x} - \theta(\mathbf{x}-\mathbf{y}) \qquad 0 < \theta < 1$$

entre $\mathbf{x}$ e $\mathbf{y}$. Aplicando a norma do máximo temos

$$\|\mathbf{g}(\mathbf{x}) - \mathbf{g}(\mathbf{y})\|_\infty \le \|\mathbf{G}\|_\infty \|\mathbf{x}-\mathbf{y}\|_\infty$$

e atendendo a condição (3.43) temos a seguinte majoração

$$\|\mathbf{g}(\mathbf{x}) - \mathbf{g}(\mathbf{y})\|_\infty \le c\,\|\mathbf{x}-\mathbf{y}\|_\infty$$

Se $\mathbf{x}^{(0)} \in B_\epsilon(\mathbf{z})$ então utilizando a majoração anterior temos

$$\begin{aligned} \|\mathbf{z}-\mathbf{x}^{(1)}\|_\infty &= \|\mathbf{g}(\mathbf{z}) - \mathbf{g}(\mathbf{x}^{(0)})\|_\infty \\ &\le c\,\|\mathbf{z}-\mathbf{x}^{(0)}\|_\infty \\ &\le c\,\epsilon < \epsilon \end{aligned}$$

e portanto $\mathbf{x}^{(1)}$ pertence também a $B_\epsilon(\mathbf{z})$. Por indução temos

$$\begin{aligned} \|\mathbf{z}-\mathbf{x}^{(k)}\|_\infty &= \|\mathbf{g}(\mathbf{z}) - \mathbf{g}(\mathbf{x}^{(k-1)})\|_\infty \\ &\le c\,\|\mathbf{z}-\mathbf{x}^{(k-1)}\|_\infty \\ &\le c^k\,\|\mathbf{z}-\mathbf{x}^{(0)}\|_\infty \\ &\le c^k\,\epsilon < \epsilon \end{aligned}$$

e portanto todos os $\mathbf{x}^{(k)}$ pertencem a $B_\epsilon(\mathbf{z})$ e a convergência para $\mathbf{z}$ é garantida visto que $c < 1$. A solução $\mathbf{z}$ é única visto que admitindo outra solução $\mathbf{y}$ temos

$$\|\mathbf{z}-\mathbf{y}\|_\infty = \|\mathbf{g}(\mathbf{z}) - \mathbf{g}(\mathbf{y})\|_\infty \le c\,\|\mathbf{z}-\mathbf{y}\|_\infty$$

o que só é possível se for $\mathbf{y} = \mathbf{z}$. ■

Se $\mathbf{g}$ for linear então o método iterativo tem a forma dada por (3.30) em que $\mathbf{G}$ reduz-se a matriz $\mathbf{C}$ dada por (3.29). Vemos que para este caso particular a condição de convergência (3.43) corresponde a impor $\|\mathbf{C}\|_\infty < 1$, condição já anteriormente utilizada no estudo dos sistemas de equações lineares.

Se a função $\mathbf{g}(\mathbf{x})$ é tal que para uma solução $\mathbf{z}$

$$\partial g_i(\mathbf{z})/\partial x_j = 0 \quad , \quad i, j = 1, 2, \ldots .n \tag{3.44}$$

e estas derivadas são contínuas perto de $\mathbf{z}$, então (3.43) é satisfeita para algum $\epsilon > 0$. É o caso, quando se escolha em (3.41)

$$\mathbf{A}(\mathbf{x}) = \mathbf{J}^{-1}(\mathbf{x})$$

sendo $\mathbf{J}$ a matriz Jacobiana de $\mathbf{f}$

$$\mathbf{J} = \begin{bmatrix} \partial f_1/\partial x_1 & \partial f_1/\partial x_2 & \cdots & \partial f_1/\partial x_n \\ \partial f_2/\partial x_1 & \partial f_2/\partial x_2 & \cdots & \partial f_2/\partial x_n \\ \cdots & \cdots & \cdots & \cdots \\ \partial f_n/\partial x_1 & \partial f_n/\partial x_2 & \cdots & \partial f_n/\partial x_n \end{bmatrix} \tag{3.45}$$

Com efeito, neste caso a função iteradora $\mathbf{g}(\mathbf{x})$ reduz-se a

$$\mathbf{g}(\mathbf{x}) = \mathbf{x} - \mathbf{J}^{-1}(\mathbf{x})\mathbf{f}(\mathbf{x})$$

e pode mostrar-se que (3.44) é satisfeita desde que, para algum $\epsilon > 0$, $\mathbf{J}$ seja invertível em $B_\epsilon(\mathbf{z})$ e $\mathbf{f} \in C^2(B_\epsilon(\mathbf{z}))$, *i.e.*, $\mathbf{f}$ seja 2 vezes continuamente diferenciável em $B_\epsilon(\mathbf{z})$. O método iterativo da forma (3.42), com a função iteradora dada por (3.45), é a *generalização do método de Newton* e, com as condições anteriores, converge quadraticamente, *i.e.*, existe um número M tal que

$$\|\mathbf{z} - \mathbf{x}^{(k+1)}\|_\infty \leq M \, \|\mathbf{z} - \mathbf{x}^{(k)}\|_\infty^2 \quad k = 0, 1, \ldots$$

Na prática, o método de Newton

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \mathbf{J}^{-1}(\mathbf{x}^{(k)}) \, \mathbf{f}(\mathbf{x}^{(k)}) \quad k = 0, 1, \ldots$$

obtem-se resolvendo o sistema

$$\mathbf{J}(\mathbf{x}^{(k)})\Delta\mathbf{x}^{(k)} = -\mathbf{f}(\mathbf{x}^{(k)}) \tag{3.46}$$

e fazendo

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \Delta\mathbf{x}^{(k)} \tag{3.47}$$

A razão disso é de evitar a inversão da matriz $\mathbf{J}(\mathbf{x}^{(k)})$, operação como se sabe mais trabalhosa do que a resolução do sistema dado por (3.46).

Exemplo 3.13 *Aplicar o método de Newton à resolução do sistema não linear*

$$\begin{aligned} x_1^2 + x_2 x_3 &= 1 \\ 2x_1 x_2 + x_3 &= 6 \\ x_1 x_3 + 2x_2^2 &= 1 \end{aligned}$$

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

Começamos por determinar $\mathbf{J}(\mathbf{x})$ e $\mathbf{f}(\mathbf{x})$.

$$\mathbf{J}(\mathbf{x}) = \begin{bmatrix} 2x_1 & x_3 & x_2 \\ 2x_2 & 2x_1 & 1 \\ x_3 & 4x_2 & x_1 \end{bmatrix} \qquad \mathbf{f}(\mathbf{x}) = \begin{bmatrix} x_1^2 + x_2x_3 - 1 \\ 2x_1x_2 + x_3 - 6 \\ x_1x_3 + 2x_2^2 - 1 \end{bmatrix} \tag{3.48}$$

1ª iteração:Fazendo $\mathbf{x} = \mathbf{x}^{(0)} = [1, 0, 1]^T$, (3.48) reduz-se a

$$\mathbf{J}(\mathbf{x}^{(0)}) = \begin{bmatrix} 2 & 1 & 0 \\ 0 & 2 & 1 \\ 1 & 0 & 1 \end{bmatrix} \qquad \mathbf{f}(\mathbf{x}^{(0)}) = \begin{bmatrix} 0 \\ -5 \\ 0 \end{bmatrix}$$

e o sistema dado por (3.46), com $k = 0$, reduz-se, neste caso, a

$$\begin{bmatrix} 2 & 1 & 0 \\ 0 & 2 & 1 \\ 1 & 0 & 1 \end{bmatrix} \Delta\mathbf{x}^{(0)} = \begin{bmatrix} 0 \\ 5 \\ 0 \end{bmatrix}$$

A sua solução é

$$\Delta\mathbf{x}^{(0)} = \begin{bmatrix} -1 \\ 2 \\ 1 \end{bmatrix}$$

Utilizando (3.47) com $k = 0$, temos portanto

$$\mathbf{x}^{(1)} = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix} + \begin{bmatrix} -1 \\ 2 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 2 \\ 2 \end{bmatrix}$$

2ª iteração: Substituindo em (3.48) $\mathbf{x}$ por $\mathbf{x}^{(1)}$, o sistema dado por (3.46) é agora

$$\begin{bmatrix} 0 & 2 & 2 \\ 4 & 0 & 1 \\ 2 & 8 & 0 \end{bmatrix} \Delta\mathbf{x}^{(1)} = \begin{bmatrix} -3 \\ 4 \\ -7 \end{bmatrix}$$

cuja solução é

$$\Delta\mathbf{x}^{(1)} = \begin{bmatrix} 37/34 \\ -39/34 \\ -12/34 \end{bmatrix}$$

Utilizando novamente (3.47) agora com $k = 1$, temos

$$\mathbf{x}^{(2)} = \begin{bmatrix} 0 \\ 2 \\ 2 \end{bmatrix} + \begin{bmatrix} 37/34 \\ -39/34 \\ -12/34 \end{bmatrix} = \begin{bmatrix} 37/34 \\ 29/34 \\ 56/34 \end{bmatrix}$$

Na Tabela 3.4 apresentamos os resultados obtidos efectuando 10 iterações. Destes resultados, vemos que, para este sistema e com o vector inicial dado, só a partir da 7ª iteração é que o método começa a convergir para a solução e a convergência só começa a ser quadrática a partir da 9ª iteração. ■

Notamos que na resolução de sistemas não lineares, o método de Newton leva-nos a resolver em cada iteração um sistema linear cuja matriz $\mathbf{J}(\mathbf{x}^{(k)})$ é diferente da utilizada na iteração anterior. Várias modificações do método de Newton têm sido criadas, com a finalidade de evitar a resolução de sistemas lineares com matrizes diferentes em todas as iterações.

Tabela 3.4: Exemplo do método de Newton

k	$x_1^{(k)}$	$x_2^{(k)}$	$x_3^{(k)}$	$\lVert\mathbf{x}^{(k)}-\mathbf{x}^{(k-1)}\rVert_\infty$	$\lVert\mathbf{e}^{(k)}\rVert_\infty$	$\lVert\mathbf{e}^{(k)}\rVert_\infty/\lVert\mathbf{e}^{(k-1)}\rVert_\infty$
0	1.000000	0.000000	1.000000		4.947856	
1	0.000000	2.000000	2.000000	2.000000	3.947856	0.797892
2	1.088235	0.852941	1.647059	1.147059	4.300797	1.089401
3	−5.850712	−5.297865	29.36768	27.72062	23.41982	5.445460
4	−2.445027	−2.268177	15.54482	13.82285	9.596966	0.409780
5	−0.960780	−0.920138	8.233546	7.311275	2.285690	0.238168
6	1.043993	0.814904	11.25522	3.021672	5.307361	2.321995
7	0.578240	0.475886	5.765444	5.489774	0.419142	0.078974
8	0.152830	0.163623	6.215665	0.450222	0.267809	0.638946
9	0.158825	0.163879	5.947947	0.267718	0.000273	0.001019
10	0.159098	0.163872	5.947856	0.000273	0.000000	0.000000
∞	+0.159098	+0.163872	5.947856	0	0	0

3.7 Problemas

1. Considere a matriz

$$\mathbf{A} = \begin{bmatrix} 3 & -4 & 6 \\ 4 & -2 & 5 \\ 2 & 2 & 1 \end{bmatrix}$$

   (a) Efectue a factorização de $\mathbf{A}$ utilizando o método de Gauss com pivot equilibrado.

   (b) Utilizando a factorização obtida na alínea anterior, resolva $\mathbf{Ax} = \mathbf{b}$ com $\mathbf{b} = [7/6, 1, 0]^T$.

2. Qual o efeito da utilização de pivots nos métodos de eliminação?

3. Qual a solução do sistema $\mathbf{Ax} = \mathbf{b}$, sabendo que $\mathbf{A} = \mathbf{EF}$, $\mathbf{Ec} = \mathbf{b}$, $\mathbf{Fd} = \mathbf{c}$?

4. Na resolução do sistema $\mathbf{Ax} = \mathbf{b}$ pelo método de Crout chegou-se, numa determinada fase da factorização, à configuração da matriz auxiliar $\mathbf{C}$ que a seguir se indica.

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 2 & 1 & 1 \\ 0 & 2 & 0 & 4 & 2 \\ 2 & 0 & 8 & 6 & 10 \\ 2 & 1 & 6 & 8 & 13 \\ 1 & 2 & 2 & 5 & 4 \end{bmatrix} \qquad \mathbf{C} = \begin{bmatrix} 1 & 0 & 2 & 1 & 1 \\ 0 & 2 & 0 & 2 & 1 \\ 2 & 0 & 4 & 1 & 2 \\ 2 & 1 & 2 & p & q \\ 1 & 2 & 0 & r & s \end{bmatrix} \qquad \mathbf{b} = \begin{bmatrix} 0 \\ 2 \\ 0 \\ 1 \\ 2 \end{bmatrix}$$

   (a) Determine p, q, r e s e calcule a solução do sistema.

   (b) Na resolução do sistema anterior teria interesse a utilização do método iterativo de Gauss-Seidel? Porquê?

5. Dos métodos utilizados na resolução de sistemas indique um para o qual é fundamental obter previamente a factorização da matriz. Justifique a sua resposta indicando sucintamente o que aconteceria se para este método não utilizasse a factorização.

6. Dada a matriz
$$\mathbf{A} = \begin{bmatrix} 2 & -4 & -6 \\ 0 & 1 & 2 \\ 1 & 0 & 2 \end{bmatrix}$$
pretende-se obter a sua inversa $\mathbf{A}^{-1}$.

   (a) Calcule a 1ª coluna de $\mathbf{A}^{-1}$, utilizando o método de factorização de Crout.

   (b) Obtenha uma aproximação da 2ª coluna de $\mathbf{A}^{-1}$ utilizando o método de Gauss-Seidel, considerando como aproximação inicial o vector $[0, 1, 1]^T$ e efectuando apenas uma iteração. Caso tivesse utilizado o método de Jacobi, qual seria a aproximação obtida.

7. Considere o seguinte sistema.
$$\begin{bmatrix} 2 & 0 & 0 \\ 2 & -1 & 0 \\ 0 & 1 & -2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} -4 \\ 1 \\ 2 \end{bmatrix}$$
Partindo de $\mathbf{x}^{(0)} = [3, -1, 8]^T$ obtenha $\mathbf{x}^{(1)}$ efectuando uma iteração utilizando o método de Jacobi. Com este vector $\mathbf{x}^{(1)}$ efectue uma nova iteração utilizando agora o método de Gauss-Seidel.

8. Na resolução do sistema $\mathbf{Ax} = \mathbf{b}$, os métodos iterativos podem escrever-se na forma
$$\mathbf{x}^{(k+1)} = \mathbf{C}\mathbf{x}^{(k)} + \mathbf{M}^{-1}\mathbf{b}$$
com
$$\mathbf{C} = -\mathbf{MN} \qquad \mathbf{A} = \mathbf{M} + \mathbf{N}$$

   (a) Mostre que
$$\|\mathbf{x} - \mathbf{x}^{(k+1)}\|_\infty \leq \|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\|_\infty \frac{\beta}{1-\beta}$$
   em que $\beta = \|\mathbf{C}\|_\infty$ é a norma de $\mathbf{C}$ induzida pela norma do máximo.

   (b) i. Qual a condição a impor aos coeficientes a_{ij} de $\mathbf{A}$ de forma a que
$$\|\mathbf{x} - \mathbf{x}^{(k+1)}\|_\infty \leq \|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\|_\infty$$
   no método iterativo de Jacobi?

   ii. Determine $\beta = \|\mathbf{C}\|_\infty$ para o método de Jacobi quando a matriz é:
$$\mathbf{A} = \begin{bmatrix} 3 & -1 & 0 \\ 0 & 3 & -1 \\ -1 & 1 & 3 \end{bmatrix}$$

9. Considere o seguinte sistema:

$$\begin{bmatrix} 4 & -2 & 1 \\ -2 & 10 & -1/2 \\ 1 & -1/2 & 5/4 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

   (a) Haveria vantagem em utilizar na resolução deste sistema o método de Gauss-Seidel? Justifique o seu raciocínio. Determine, por este método, o valor da 1ª iterada $\mathbf{x}^{(1)}$ tomando $\mathbf{x}^{(0)} = [0.5, 3.5, 15.5]^T$.

   (b) Através de um método de factorização para resolução de sistemas de equações lineares que, para este caso, seja o mais aconselhável, resolva o sistema.

10. Verificando-se num método iterativo o critério de paragem

$$\max_{1 \le i \le n} |x_i^{(k+1)} - x_i^{(k)}| \le 5 \times 10^{-2}$$

será que

$$\max_{1 \le i \le n} |x_i^{(k+1)} - x_i| \le 5 \times 10^{-2}$$

em que $\mathbf{x} = [x_1, \ldots, x_n]^T$ é a solução do sistema a resolver? Justifique.

11. Em cálculo numérico utilizam-se muitas vezes os métodos iterativos em detrimento dos métodos directos. Qual é o objectivo?

12. Considere o sistema:

$$\begin{bmatrix} 5 & 0 & 10 \\ 0 & 10 & 0 \\ 10 & 0 & 34 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} -97 \\ 514 \\ -380 \end{bmatrix}$$

   (a) Que método de factorização deve empregar na resolução do sistema? Justifique. Efectue essa factorização e indique como prosseguiria os cálculos para obter o vector solução.

   (b) Aplicando o método de Gauss-Seidel à resolução do sistema, o que pode afirmar quanto à convergência? Recorrendo à forma não matricial, determine a 1ª aproximação do vector solução, considerando como aproximação inicial o vector nulo.

13. Considere o seguinte sistema:

$$\begin{bmatrix} 2 & 0 & 0 & 0 \\ 2 & -1 & 0 & 0 \\ 2 & 1 & -2 & 0 \\ 0 & 1 & 0 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} -4 \\ 1 \\ 2 \\ 1 \end{bmatrix}$$

Partindo de $\mathbf{x}^{(0)} = [1, -1, 8, 7]^T$ obtenha $\mathbf{x}^{(1)}$ efectuando uma iteração utilizando o método de Jacobi. Com este vector $\mathbf{x}^{(1)}$ efectue uma nova iteração utilizando agora o método de Gauss-Seidel.

14. Dado o sistema de equações não lineares:

$$\begin{aligned} -x^2 + y^2 + z^2 &= -1 \\ 2xy + 2y^2 + z^2 + 3z &= 3 \\ 2xz - yz &= -2 \end{aligned}$$

(a) Aplicando o método de Newton mostre que a matriz jacobiana do sistema, considerando $[x_0, y_0, z_0]^T = [2, 1, -1]^T$, é:

$$\mathbf{W} = \begin{bmatrix} -4 & 2 & -2 \\ 2 & 8 & 1 \\ -2 & 1 & 3 \end{bmatrix}$$

(b) Resolva, pelo método de Crout, o sistema que lhe permite calcular a nova aproximação do vector solução do sistema não linear.

(c) Recorrendo à forma não matricial do método de Gauss-Seidel, determine a 1ª aproximação do vector solução de $\mathbf{Wx} = [1, 0, 0]^T$, considerando como aproximação inicial o vector nulo.

15. Mostre que, se aplicar o método de Newton (método para resolver sistemas não lineares) à resolução de um sistema de equações lineares, obtém a solução exacta logo na primeira iteração, seja qual for a aproximação inicial (e desprezando erros de arredondamento). Porque não usar, então, este método para resolver sistemas lineares?

16. Na resolução, pelo método de Newton, do seguinte sistema de equações

$$\begin{aligned} 2x - xz &= a \\ 2yz + y - 1 &= b \\ 3x + z^2 &= c \end{aligned}$$

ao tomar para solução inicial $[x_0, y_0, z_0]^T = [0, 2, 1]^T$ obteve-se depois de uma iteração $[x_1, y_1, z_1]^T = [1, 2, 3]^T$. Determine a, b, c.

17. Dado o sistema

$$\begin{aligned} x + a_1 y &= b_1 \\ a_2 x + y &= b_2 \end{aligned}$$

mostre que o método iterativo de Gauss-Seidel é convergente sse $|m| < 1$ com $m = a_1 a_2$ e que

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

em que x é a solução do sistema e $x^{(k)}$ é o valor aproximado na iteração k.

18. Dado o sistema

$$\begin{aligned} x_1 + a_1 x_2 &= b_1 \\ a_2 x_1 + x_2 &= b_2 \end{aligned}$$

mostre que o método iterativo de Jacobi é convergente sse $|m| < 1$ com $m = a_1 a_2$ e que

$$\max_{i=1,2} |x_i - x_i^{(k+1)}| \leq \frac{\alpha}{1-\alpha} \max_{i=1,2} |x_i^{(k+1)} - x_i^{(k)}|$$

em que $\alpha = \max_{i=1,2} |a_i|$, $[x_1, x_2]^T$ é a solução do sistema e $[x_1^{(k)}, x_2^{(k)}]^T$ é o valor aproximado na iteração k.

19. Dado o sistema de equações não lineares:

$$\begin{aligned} &e^{2x} + 3\sin y + \cos z = 1 \\ &\cos x + e^{2y} + \sin z = 2 \\ &e^{2x} + z = 2 \end{aligned}$$

pretende-se aplicar o método de Newton, considerando para vector inicial $[x_0, y_0, z_0]^T = [0, 0, 0]^T$.

(a) Factorize a matriz Jacobiana utilizando o método de Gauss com pivot equilibrado.

(b) Utilizando a factorização obtida na alínea anterior, obtenha a 1ª iterada $[x_1, y_1, z_1]^T$.

20. Seja $\mathbf{B}$ uma matriz $n \times n$ triangular superior com todos os elementos da sua diagonal principal nulos; então $\mathbf{B}^n = \mathbf{O}$. Utilize este resultado para mostrar que o método de Gauss-Seidel converge num número finito p de iterações quando a matriz $\mathbf{A}$ do sistema for triangular superior (e não singular). Qual é o valor de p?

Capítulo 4

Interpolação Polinomial

4.1 Introdução

Este capítulo e o seguinte são dedicados a aproximação de funções, *i.e.*, dado uma função f obter uma função g, em geral mais simples, que aproxima f segundo um certo critério. Ao enunciar um problema de aproximação consideram-se dois aspectos:

1. Definição da função a aproximar:

   (a) É conhecida uma relação que defina a função no seu domínio (expressão analítica, desenvolvimento em série, fórmula integral, equação diferencial ou integral, etc.)

   (b) São só conhecidos valores da função, habitualmente afectados de erros, num conjunto numerável de pontos.

2. Finalidade da aproximação:

   Resolução de equações não lineares, diferenciais e integrais; derivação ou integração da função; cálculo, traçado do gráfico ou obtenção de uma expressão analítica da função, etc.

Na resolução de um dado problema de aproximação consideram-se ainda mais dois aspectos.

3. Escolha da classe de funções aproximadoras:

   Polinómios de um determinado grau, splines, funções trignométricas, racionais (quocientes de polinómios) ou exponenciais, etc.

4. Escolha do critério de aproximação:

   Interpolação ou minimização de uma norma ou semi-norma do erro (habitualmente as normas do máximo, Euclidiana e L_1).

Definidas as escolhas (3) e (4), a aproximação consiste em obter um elemento da classe de funções aproximadoras que satisfaça ao critério de aproximação.

Neste capítulo, iremos estudar a interpolação polinomial. A interpolação consiste em escolher dentro de uma classe de funções a função g (interpoladora) cujo gráfico, definido por $y = g(x)$, passa por um dado conjunto de pontos (x_i, f_i), $i = 0, 1 \ldots, n$. O critério de aproximação consiste, portanto, em impor $g(x_i) = f(x_i)$, $i = 0, 1, \ldots, n)$. Devido às suas propriedades algébricas e analíticas, os polinómios são as funções mais frequentemente utilizadas na aproximação e em particular na interpolação. A interpolação polinomial é utilizada preferencialmente na resolução de equações não lineares, diferenciais e integrais, na derivação e integração numérica e na interpolação de tabelas; constituindo a base de muitos métodos numéricos.

Dado que a interpolação polinomial fornece, frequentemente, uma aproximação grosseira da função, tem vindo a ser substituída pela interpolação por spline e pela interpolação racional. O interesse pela interpolação por spline tem vindo a crescer nos últimos anos. É utilizada, por exemplo, no traçado de gráficos, na derivação e, tal como a interpolação racional, no cálculo computacional de funções.

Como já referimos, a interpolação tem várias aplicações importantes. A mais antiga consiste em interpolar uma tabela, *i.e.*, definida uma função $f(x)$ por meio de uma tabela, obter um valor aproximado de $f(x)$ correspondente a um argumento x não tabelado ($x \neq x_i$, $i = 0, 1, \ldots, n$). Hoje em dia a utilização de tabelas caiu um pouco em desuso visto que podemos, quase sempre, determinar directamente o valor de uma função com o auxílio do computador.

Neste capítulo, são deduzidas as fórmulas de Lagrange e de Newton. Porque a interpolação é uma aproximação, dá-se especial relevo ao erro cometido. Assim, faz-se referência ao exemplo de Runge e apresenta-se a interpolação utilizando para nós os zeros de polinómios de Chebyshev o que permite obter uma aproximação que quase minimiza o erro no sentido da norma do máximo.

4.2 Fórmula interpoladora de Lagrange

Consideremos f uma função real definida num intervalo $I = [a, b]$ e $x_0, \ldots, x_n$ $n + 1$ pontos distintos de I. Escrevemos

$$f_i = f(x_i) \qquad i = 0, 1, \ldots, n$$

Pretendemos construir um polinómio $p(x)$ que interpola $f(x)$ nos pontos (x_i, f_i), $i = 0, 1, \ldots, n$, *i.e.*,

$$p(x_i) = f_i \qquad i = 0, 1, \cdots, n \tag{4.1}$$

Designaremos por nós de interpolação as abcissas x_i, $i = 0, 1, \ldots, n$. Seja

$$p(x) = a_0 + a_1 x + \ldots + a_m x^m$$

o polinómio de grau m que satisfaz a (4.1). Então $a_0, a_1, \ldots, a_m$ podem obter-se resolvendo o seguinte sistema

$$\begin{array}{l} a_0 + a_1 x_0 + \cdots + a_m x_0^m = f_0 \\ a_0 + a_1 x_1 + \cdots + a_m x_1^m = f_1 \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ a_0 + a_1 x_n + \cdots + a_m x_n^m = f_n \end{array}$$

de $n+1$ equações lineares a $m+1$ incógnitas. Para podermos garantir a existência e unicidade da solução deste sistema façamos $m = n$. O sistema pode escrever-se na forma matricial

$$X\ a\ =\ f \tag{4.2}$$

com

$$X = \begin{bmatrix} 1\ x_0 \cdots x_0^n \\ 1\ x_1 \cdots x_1^n \\ \cdots\ \ \cdots\cdots \\ 1\ x_n \cdots x_n^n \end{bmatrix} \qquad \begin{array}{l} a{=}[a_0, a_1, \ldots, a_n]^T \\ f{=}[f_0, f_1, \ldots, f_n]^T \end{array}$$

A matriz X é conhecida por *matriz de Vandermonde*. Sabe-se que, para x_i distintos, X é uma matriz não singular e portanto (4.2) tem uma só solução a. Teoricamente o problema da interpolação polinomial fica resolvido com a resolução do sistema (4.2). Todavia existem fórmulas, que iremos apresentar neste capítulo que permitem resolver este mesmo problema com menor trabalho computacional.

Como existem várias fórmulas para obter um polinómio interpolador p, de grau igual ou inferior a n (eventualmente $a_n = 0$), é importante mostrar que do conjunto dos polinómios de grau inferior ou igual a n, só um é que pode interpolar uma dada função f em $n+1$ pontos. Sejam p e q dois polinómios interpoladores de grau quanto muito igual a n, tais que

$$p(x_i) = q(x_i) = f_i \qquad i = 0, 1, \ldots, n \tag{4.3}$$

Seja $r(x) = p(x) - q(x)$. Então r é um polinómio de grau menor ou igual a n e de (4.3) temos que

$$r(x_i) = 0 \qquad i = 0, 1, \ldots, n$$

Visto que r tem $n+1$ zeros e $grau(r) \leq n$ temos que $r(x) \equiv 0$. Logo $q(x) \equiv p(x)$, o que prova a unicidade do polinómio interpolador.

Apresentamos a seguir uma fórmula que prova a existência do polinómio interpolador e ao mesmo tempo permite a interpolação deste. Começamos por considerar o caso particular em que f é uma função tal que para um certo i $(0 \leq i \leq n)$

$$f_i = 1$$

e

$$f_j = 0 \ \text{ para } \ \mathrm{j} \neq \mathrm{i}$$

O polinómio interpolador da função f, que designaremos por $l_i(x)$, tem n zeros $x_0, x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n$. Logo $l_i(x)$ é um polinómio de grau igual a n que se pode escrever na forma

$$l_i(x) = C(x - x_0)(x - x_1) \cdots (x - x_{i-1})(x - x_{i+1}) \cdots (x - x_n)$$

em que C é uma constante. A condição $l_i(x_i) = f_i = 1$ implica que

$$C = [(x_i - x_0)(x_i - x_1) \cdots (x_i - x_{i-1})(x_i - x_{i+1}) \cdots (x_i - x_n)]^{-1}$$

Temos então

$$l_i(x) = \prod_{j=0, j\neq i}^{n} \left(\frac{x - x_j}{x_i - x_j}\right) \qquad 0 \leq i \leq n$$

Este polinómio l_i, de grau n, denomina-se por *polinómio de Lagrange*, e foi construido de modo a verificar-se

$$l_i(x_j) = \begin{cases} 1 & j = i \\ 0 & j \neq i \end{cases} \tag{4.4}$$

Podemos agora considerar o caso geral da interpolação polinomial caracterizada por (4.1). O polinómio interpolador p_n pode exprimir-se directamente em termos dos polinómios $l_0, l_1, \ldots, l_n$ por meio da *fórmula interpoladora de Lagrange*:

$$p_n(x) = \sum_{i=0}^{n} f_i \, l_i(x)$$

É fácil de ver que $p_n(x)$ satisfaz a (4.1) tendo em conta as propriedades (4.4) dos polinómios $l_i(x)$ e que $p_n(x)$ é um polinómio de grau menor ou igual a n por ser a soma de polinómios $l_i(x)$ de grau igual a n. A fórmula de Lagrange mostra a existência do polinómio interpolador o que nos permite enunciar o teorema seguinte.

Teorema 4.1 *Dado $n+1$ abcissas distintas $x_0, x_1, \ldots, x_n$ e as respectivas $n+1$ ordenadas $f_0, f_1, \ldots, f_n$, de todos os polinómios de grau menor ou igual a n existe um único polinómio $p_n(x)$ que interpola nos pontos (x_i, f_i), $i = 0, 1, \ldots, n$.* ■

Consideremos os casos mais simples da fórmula de Lagrange. Começamos com $n = 1$; neste caso são dados dois pontos $(x_0, f_0), (x_1, f_1)$. Então

$$l_0(x) = \frac{x - x_1}{x_0 - x_1} \qquad l_1(x) = \frac{x - x_0}{x_1 - x_0}$$

e

$$p_1(x) = f_0 l_0(x) + f_1 l_1(x) = f_0 \frac{x - x_1}{x_0 - x_1} + f_1 \frac{x - x_0}{x_1 - x_0}$$

ou ainda

$$p_1(x) = f_0 + \frac{f_1 - f_0}{x_1 - x_0}(x - x_0) \tag{4.5}$$

Como se vê, este é o caso da interpolação linear. Nota-se que $p_1(x_0) = f_0$, $p_1(x_0) = f_1$. Com $n = 2$ obtemos a interpolação quadrática.

$$p_2(x) = f_0 \frac{(x - x_1)\,(x - x_2)}{(x_0 - x_1)\,(x_0 - x_2)} + f_1 \frac{(x - x_0)\,(x - x_2)}{(x_1 - x_0)\,(x_1 - x_2)} + f_2 \frac{(x - x_0)\,(x - x_1)}{(x_2 - x_0)\,(x_2 - x_1)}$$

Por exemplo o polinómio de grau menor ou igual a 2 que passa pelos três pontos $(0, 2)$, $(-1, 3)$ e $(2, 6)$ é

$$\begin{aligned} p_2(x) &= 2\,\frac{(x+1)\,(x-2)}{(0+1)\,(0-2)} + 3\,\frac{(x-0)\,(x-2)}{(-1-0)\,(-1-2)} + 6\,\frac{(x-0)\,(x+1)}{(2-0)\,(2+1)} \\ &= x^2 + 2 \end{aligned}$$

Se substituirmos o ponto $(0,2)$ por $(0,4)$ obtemos $p_2(x) = x+4$ que é um polinómio de grau um porque, neste caso particular, os pontos $(0,4)$, $(-1,3)$ e $(2,6)$ são colineares.

Para reduzir ao mínimo o número de operações a efectuar vamos escrever a fórmula de Lagrange na seguinte forma

$$p_n(x) = \omega_{n+1}(x) \sum_{i=0}^{n} \frac{f_i}{(x - x_i)\, \omega'_{n+1}(x_i)}$$

em que

$$\omega_{n+1}(x) = \prod_{j=0}^{n} (x - x_j) \tag{4.6}$$

e portanto

$$\omega'_{n+1}(x_i) = \prod_{j=0, j\neq i}^{n} (x_i - x_j)$$

Verifica-se que depois do cálculo dos $f_i/\omega'_{n+1}(x_i)$, que dependem só dos pontos dados (x_i, f_i), ainda é necessário efectuar $2n+2$ multiplicações ou divisões e $2n+1$ adições, para calcular o polinómio interpolador num dado valor de x.

Com a fórmula de Newton, que apresentamos na secção seguinte, para cada valor de x só é necessário efectuar n multiplicações e $2n$ adições. Não bastando isto, a fórmula de Lagrange é computacionalmente ineficiente para passar de um polinómio interpolador para outro de maior grau, *i.e.*, aumentar o número de pontos de interpolação (veremos mais tarde que a comparação entre resultados com polinómios interpoladores de graus diferentes permite-nos geralmente decidir qual o grau a utilizar).

4.3 Fórmula de Newton. Diferenças divididas

Em face do que acabamos de expr, pretendemos obter o polinómio interpolador $p_n(x)$ utilizando o polinómio $p_{n-1}(x)$ que interpola $f(x)$ em $x_0, x_1, \ldots, x_{n-1}$. Escrevemos assim

$$p_n(x) = p_{n-1}(x) + c_n(x) \tag{4.7}$$

Como $grau(p_n) \leq n$ e $grau(p_{n-1}) \leq n-1$ então a função c_n é um polinómio de grau menor ou igual a n. Por outro lado como

$$c_n(x_i) = p_n(x_i) - p_{n-1}(x_i) = f_i - f_i = 0 \qquad i = 0, 1, \ldots, n-1$$

este polinómio $c_n(x)$ tem n zeros $x_0, x_1, \ldots, x_{n-1}$. Logo

$$c_n(x) = A_n\, \omega_n(x) = A_n(x - x_0)\,(x - x_1) \cdots (x - x_{n-1}) \tag{4.8}$$

A_n obtém-se atendendo que $p_n(x_n) = f(x_n)$. De (4.7) e (4.8) tem-se

$$A_n = \frac{f(x_n) - p_{n-1}(x_n)}{(x_n - x_0) \cdots (x_n - x_{n-1})}$$

Este coeficiente A_n só depende de $x_0, x_1, \ldots, x_n$ e dos respectivos valores de $f(x)$. Por isso é habitual representá-lo por $f[x_0, x_1, \ldots, x_n]$ e denominá-lo *diferença dividida de*

ordem n visto que, como veremos a seguir (ver (4.12)), é possível exprimi-lo sob a forma de um quociente de duas diferenças.

De (4.7) e (4.8) obtemos

$$p_n(x) = p_{n-1}(x) + f[x_0, x_1, \ldots, x_n]\,(x - x_0)\cdots(x - x_{n-1}) \tag{4.9}$$

Se em (4.9) fizermos sucessivamente $n = 1, 2, \ldots$ e notarmos que $p_0(x) = f(x_0)$ obtemos

$$\begin{aligned}
&p_1(x) = f(x_0) + f[x_0, x_1]\,(x - x_0)\\
&p_2(x) = f(x_0) + f[x_0, x_1]\,(x - x_0) + f[x_0, x_1, x_2]\,(x - x_0)(x - x_1)\\
&\cdots\cdots\cdots\cdots\cdots\cdots\\
&p_n(x) = f(x_0) + \sum_{i=1}^{n} f[x_0, \ldots, x_i]\,(x - x_0)\cdots(x - x_{i-1})
\end{aligned} \tag{4.10}$$

A fórmula (4.10) é conhecida por *fórmula interpoladora de Newton* com diferenças divididas e, como já referimos atrás, é do ponto de vista computacional bastante melhor do que a fórmula de Lagrange.

Atendendo a (4.9) ou (4.10), podemos fazer a seguinte:

Definição 4.1 *A diferença dividida* $f[x_0, \ldots, x_n]$ *é o coeficiente do termo em* x^n *do polinómio* $p_n(x)$ *que interpola* $f(x)$ *em* $x_0, \ldots, x_n$.

Utilizando esta definição vamos obter uma fórmula simples para calcular esta diferença dividida. Seja $f[x_1, \ldots, x_{n-1}, x_n]$ e $f[x_0, x_1, \ldots, x_{n-1}]$ os coeficientes dos termos x^{n-1} dos polinómios $q_{n-1}(x)$ e $p_{n-1}(x)$ que interpolam $f(x)$ respectivamente em $x_1, \ldots, x_{n-1}, x_n$ e $x_0, x_1, \ldots, x_{n-1}$. Então

$$p(x) = \frac{x - x_0}{x_n - x_0}\, q_{n-1}(x) - \frac{x - x_n}{x_n - x_0}\, p_{n-1}(x) \tag{4.11}$$

é um polinómio de grau menor ou igual a n. Como

$$q_{n-1}(x_i) = f(x_i) \qquad i = 1, \ldots, n$$

e

$$p_{n-1}(x_i) = f(x_i) \qquad i = 0, \ldots, n-1$$

é fácil verificar que

$$p(x_i) = f(x_i) \qquad i = 0, \ldots, n$$

Logo $p(x) = p_n(x)$ e portanto $f[x_0, x_1, \ldots, x_n]$ é o coeficiente do termo em x^n de $p(x)$. Atendendo a (4.11) e notando que $f[x_1, \ldots, x_{n-1}, x_n]$ e $f[x_0, x_1, \ldots, x_{n-1}]$ são os coeficientes do termo em x^n respectivamente dos polinómios $(x - x_0)q_{n-1}(x)$ e $(x - x_n)p_{n-1}(x)$ temos a seguinte fórmula importante

$$f[x_0, \ldots, x_n] = \frac{f[x_1, \ldots, x_{n-1}, \mathbf{x_n}] - f[\mathbf{x_0}, x_1, \ldots, x_{n-1}]}{\mathbf{x_n} - \mathbf{x_0}} \tag{4.12}$$

Se em (4.10) fizermos $n = 1$ (interpolação linear) e compararmos com (4.5) vemos imediatamente que $f[x_0] = f(x_0)$ e que a diferença dividida de 1ª ordem é dada por

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

ou genéricamente

$$f[x_i, x_j] = f[x_j, x_i] = \frac{f(x_j) - f(x_i)}{x_j - x_i} \tag{4.13}$$

A partir de diferenças de 1^a ordem obtêm-se diferenças de 2^a ordem que por sua vez permitem obter diferenças de 3^a ordem e assim por diante, utilizando (4.12) ou genericamente

$$f[x_i, \ldots, x_{i+k}] = \frac{f[x_{i+1}, \ldots, x_{i+k-1}, \mathbf{x_{i+k}}] - f[\mathbf{x_i}, x_{i+1}, \ldots, x_{i+k-1}]}{\mathbf{x_{i+k}} - \mathbf{x_i}} \tag{4.14}$$

A diferença dividida de ordem k $f[x_i, \ldots, x_{i+k}]$ obtém-se a partir de duas diferenças divididas de ordem $k-1$ $f[x_{i+1}, \ldots, x_{i+k}]$ e $f[x_i, \ldots, x_{i+k-1}]$ que têm $k-1$ argumentos comuns $x_{i+1}, \ldots, x_{i+k-1}$. Como $f[x_i, \ldots, x_{i+k}]$ é o coeficiente do termo em x^k do polinómio $p_k(x)$ que interpola $f(x)$ em $x_i, \ldots, x_{i+k}$ pode-se permutar a ordem dos argumentos de $f[x_i, \ldots, x_{i+k}]$ e por conseguinte exprimir $f[x_i, \ldots, x_{i+k}]$ utilizando duas quaisquer diferenças divididas de ordem $k-1$ cujos argumentos são escolhidos entre os $x_i, \ldots, x_{i+k}$. Assim, para a diferença dividida de 2ª ordem $f[x_1, x_2, x_3]$ temos, p.ex., que

$$\begin{aligned} f[x_1, x_2, x_3] &= f[x_2, x_3, x_1] = f[x_3, x_1, x_2] = \frac{f[x_2, x_3] - f[x_1, x_2]}{x_3 - x_1} \\ &= \frac{f[x_2, x_3] - f[x_1, x_3]}{x_2 - x_1} = \frac{f[x_1, x_2] - f[x_1, x_3]}{x_2 - x_3} \end{aligned} \tag{4.15}$$

É usual apresentar as diferenças divididas sob a forma de uma tabela triangular como a que se segue. Nesta tabela só estão presentes diferenças com argumentos cujos índices são consecutivos. Não apareça diferenças tais como

$$f[x_0, x_2]\ ,\quad f[x_1, x_3]\ ,\quad f[x_0, x_2, x_4]\ ,\quad f[x_0, x_2, x_3, x_4]$$

Para obter por exemplo $f[x_1, x_2, x_3]$ dado por (4.15) a partir da tabela considera-se as duas diagonais

$$f[x_1, x_2, x_3]\ ,\quad f[x_1, x_2]\ ,\quad f(x_1)$$

e

$$f[x_1, x_2, x_3]\ ,\quad f[x_2, x_3]\ ,\quad f(x_3)$$

indicadas na tabela a ponteado; o numerador de $f[x_1, x_2, x_3]$ é a diferença dos termos adjacentes de $f[x_1, x_2, x_3]$ nas diagonais, *i.e.*, $f[x_1, x_2]$ e $f[x_2, x_3]$; o denominador de $f[x_1, x_2, x_3]$ é a diferença dos argumentos x_1 e x_3 dos últimos termos das diagonais, *i.e.*, $f[x_1]$ e $f[x_3]$.

Exemplo 4.1 *Ilustrar para $f(x) = \sqrt{x}$, com $x_0 = 1$, $x_i = x_0 + ih$, $i = 1, 2, 3, 4$, $h = 0.05$, a utilização da fórmula interpoladora de Newton (4.10). Obter aproximações de f(x), para $x = \overline{x}$ com $\overline{x} = 0$, $\overline{x} = 1.13$ e $\overline{x} = 2.2$, utilizando sucessivamente polinómios interpoladores de grau 1, 2, 3 e 4.*

Tabela 4.1: Tabela de diferenças divididas

x_i		$f(x_i)$	$f[x_i, x_{i+1}]$	$f[\ ,\ ,\ ]$	$f[\ ,\ ,\ ,\ ]$	$f[\ ,\ ,\ ,\ ,\ ]$
x_0		$f(x_0)$				
			$f[x_0, x_1]$			
x_1	$\cdots$	$f(x_1)$		$f[x_0, x_1, x_2]$		
			$f[x_1, x_2]$		$f[x_0, x_1, x_2, x_3]$	
x_2		$f(x_2)$		$f[x_1, x_2, x_3]$		$f[x_0, x_1, x_2, x_3, x_4]$
			$f[x_2, x_3]$		$f[x_1, x_2, x_3, x_4]$	
x_3	$\cdots$	$f(x_3)$		$f[x_2, x_3, x_4]$		
			$f[x_3, x_4]$			
x_4		$f(x_4)$				

Tabela 4.2: Diferenças divididas de $f(x) = \sqrt{x}$

x_i	$f(x_i)$	$f[x_i, x_{i+1}]$	$f[\ ,\ ,\ ]$	$f[\ ,\ ,\ ,\ ]$	$f[\ ,\ ,\ ,\ ,\ ]$
1.00	**1.000000**				
		0.49390			
1.05	1.024695		**−0.1162**		
		0.48228		**0.052000**	
1.10	1.048809		−0.1084		**−0.033333**
		0.47144		0.045333	
1.15	1.072381		−0.1016		
		0.46128			
1.20	1.095445				

Primeiro construimos a tabela de diferenças divididas para $f(x) = \sqrt{x}$, apresentada na Tabela 4.2. A diagonal superior é constituida pelas diferenças divididas

$$f[1.00] = 1 \qquad f[1.00, 1.05] = 0.4939 \qquad f[1.00, 1.05, 1.10] = -0.1162$$

$$f[1.00, 1.05, 1.10, 1.15] = 0.052 \qquad f[1.00, 1.05, 1.10, 1.15, 1.20] = -0.033333$$

Estas diferenças foram obtidas utilizando (4.12); p.ex.,

$$\begin{aligned} f[1.00, 1.05, 1.10, 1.15] &= \frac{f[1.05, 1.10, 1.15] - f[1.00, 1.05, 1.10]}{1.15 - 1.00} \\ &= \frac{-0.1084 + 0.1162}{0.15} = 0.052 \end{aligned}$$

Se substituirmos estas diferenças na fórmula de Newton e fizermos $x_0 = 1$, $x_1 = 1.05$, $x_2 = 1.10$, $x_3 = 1.15$, $x_4 = 1.20$ obtemos sucessivamente os polinómios interpoladores $p_1(x), \ldots, p_4(x)$. O polinómio $p_4(x)$ que interpola todos os pontos da Tabela 4.2 é dado por

$$\begin{aligned} p_4(x) = \; & 1 + 0.4939(x-1) - 0.1162(x-1)(x-1.05) + \\ & +0.052(x-1)(x-1.05)(x-1.1) - \\ & -0.033333(x-1)(x-1.05)(x-1.1)(x-1.15) \end{aligned}$$

Pretendemos com os sucessivos polinómios interpoladores obter aproximações de $f(x)$ num ponto $x = \overline{x}$. Consideremos primeiramente $\overline{x} = 1.13$. Utilizando a função ω_{n+1} definida por (4.6) temos

$$p_1(\overline{x}) = p_0(\overline{x}) + 0.4939(\overline{x} - 1) = 1 + 0.064207 = \underline{1.06}4207$$

$$p_2(\overline{x}) = p_1(\overline{x}) - 0.1162\, \omega_1(\overline{x})(\overline{x} - 1.05) = p_1(1.13) - 0.0012085 = \underline{1.0629}99$$

$$p_3(\overline{x}) = p_2(\overline{x}) + 0.052\, \omega_2(\overline{x})(\overline{x} - 1.1) = p_2(1.13) + 0.0000162 = \underline{1.063015}$$

$$p_4(\overline{x}) = p_3(\overline{x}) - 0.033333\, \omega_3(\overline{x})(\overline{x} - 1.15) = p_3(1.13) + 0.0000002 = \underline{1.063015}$$

Notando que $f(\overline{x}) = 1.063015$ vemos que a interpolação linear conduz neste caso a uma aproximação com 3 algarismos significativos e que o polinómio interpolador de grau 3 conduz a uma aproximação com 7 algarismos significativos (mesmo número de algarismos significativos do que os valores tabelados da função). Além dos resultados obtidos para $\overline{x} = 1.13$ apresentamos na Tabela 4.3 resultados para $\overline{x} = 0$ e $\overline{x} = 2.2$ em que $\overline{x}$ é exterior ao intervalo $[1.0, 1.2]$. ■

Nos casos em que $\overline{x}$ é exterior a $\text{int}[x_0, x_1, \ldots, x_n]$, menor intervalo que contém os nós de interpolação, estamos a efectuar uma extrapolação. Em geral, como se depreende da Tabela 4.3, a extrapolação conduz a resultados que podem estar totalmente desprovidos de precisão. Mesmo para a interpolação propriamente dita pode-se apresentar exemplos em que o erro é elevado. Por isso vamos obter expressões para determinar o erro.

4.4 Erro de interpolação

Continuamos a denominar $p_n(x)$ o polinómio de grau menor ou igual a n que interpola $f(x)$ nos $n+1$ pontos distintos (x_i, f_i), $i = 0, \ldots, n$. O erro de interpolação cometido em x é dado por

$$e_n(x) = f(x) - p_n(x)$$

Tabela 4.3: Interpolação e extrapolação de $f(x) = \sqrt{x}$

$\overline{x} = 0.00$		$f(\overline{x}) = 0.000000$		
n	$p_n(\overline{x})$	$f(\overline{x}) - p_n(\overline{x})$	$p_n(\overline{x}) - p_{n-1}(\overline{x})$	λ_n
0	1.000000	−1.000000		
1	0.506100	−0.506100	−0.493900	+1.0247
2	0.384090	−0.384090	−0.122010	+3.1480
3	0.324030	−0.324030	−0.060060	+5.3951
4	0.279755	−0.279755	−0.044275	+6.3186

$\overline{x} = 1.13$		$f(\overline{x}) = 1.063015$		
n	$p_n(\overline{x})$	$f(\overline{x}) - p_n(\overline{x})$	$p_n(\overline{x}) - p_{n-1}(\overline{x})$	λ_n
0	$\underline{1.}000000$	+0.063015		
1	$\underline{1.06}4207$	−0.001192	+0.064207	−0.0186
2	$\underline{1.0629}99$	+0.000016	−0.001208	−0.0136
3	$\underline{1.063015}$	$< 0.5 \times 10^{-6}$	+0.000016	+0.0158
4	$\underline{1.063015}$	$< 0.5 \times 10^{-6}$	$< 0.5 \times 10^{-6}$	+0.2308

$\overline{x} = 2.20$		$f(\overline{x}) = 1.483240$		
n	$p_n(\overline{x})$	$f(\overline{x}) - p_n(\overline{x})$	$p_n(\overline{x}) - p_{n-1}(\overline{x})$	λ_n
0	$\underline{1}.000000$	+0.483240		
1	$\underline{1}.592680$	−0.109440	+0.592680	−0.1847
2	$\underline{1}.432324$	+0.050916	−0.160356	−0.3175
3	$\underline{1.5}11260$	−0.028020	+0.078936	−0.3550
4	$\underline{1.4}58130$	+0.025110	−0.053130	−0.4726

Consideremos $\overline{x}$ um nó qualquer diferente de $x_0, \ldots, x_n$ e $q_{n+1}(x)$ o polinómio de grau menor ou igual a $n+1$ que interpola $f(x)$ em $x_0, \ldots, x_n, \overline{x}$; logo $q_{n+1}(\overline{x}) = f(\overline{x})$. Utilizando (4.9) temos

$$q_{n+1}(x) = p_n(x) + f[x_0, \ldots, x_n, \overline{x}] \quad \prod_{i=0}^{n}(x - x_i)$$

Logo

$$f(\overline{x}) = q_{n+1}(\overline{x}) = p_n(\overline{x}) + f[x_0, \ldots, x_n, \overline{x}] \quad \prod_{i=0}^{n}(\overline{x} - x_i)$$

Então para qualquer $\overline{x} \neq x_0, \ldots, x_n$ temos para o *erro de interpolação*

$$e_n(\overline{x}) = f[x_0, \ldots, x_n, \overline{x}] \quad \prod_{i=0}^{n}(\overline{x} - x_i) \tag{4.16}$$

Vemos que o erro de interpolação tem uma forma semelhante ao termo

$$p_{n+1}(\overline{x}) - p_n(\overline{x}) = f[x_0, \ldots, x_n, x_{n+1}] \quad \prod_{i=0}^{n}(\overline{x} - x_i)$$

que se incluíria na fórmula de Newton se tomassemos em conta mais um ponto de interpolação (x_{n+1}, f_{n+1}). Por conseguinte, sempre que $f[x_0, \ldots, x_n, \overline{x}]$ for da ordem de grandeza de $f[x_0, \ldots, x_n, x_{n+1}]$ então

$$p_{n+1}(\overline{x}) - p_n(\overline{x})$$

é da ordem de grandeza de

$$e_n(\overline{x}) = f(\overline{x}) - p_n(\overline{x})$$

Pondo

$$f[x_0, \ldots, x_n, \overline{x}] = (1 + \lambda_{n+1}) f[x_0, \ldots, x_n, x_{n+1}]$$

então

$$e_n(\overline{x}) = (1 + \lambda_{n+1})[p_{n+1}(\overline{x}) - p_n(\overline{x})]$$

e, como fácilmente se depreende,

$$e_{n+1}(\overline{x}) = f(\overline{x}) - p_{n+1}(\overline{x}) = \lambda_{n+1}[p_{n+1}(\overline{x}) - p_n(\overline{x})]$$

Logo impor que

$$|p_{n+1}(\overline{x}) - p_n(\overline{x})| \leq \epsilon \tag{4.17}$$

garante que

$$|e_{n+1}(\overline{x})| \leq \epsilon$$

se $|\lambda_{n+1}| \leq 1$, caso em que

$$0 \leq f[x_0, \ldots, x_n, \overline{x}]/f[x_0, \ldots, x_n, x_{n+1}] \leq 2$$

Exemplo 4.2 *Com os dados do Exemplo 4.1 relacionar o erro e_n com $p_n - p_{n-1}$.*

Vamos utilizar os resultados da Tabela 4.3. Assim, para $\overline{x} = 1.13$, temos

$$p_1(\overline{x}) - p_0(\overline{x}) = 1.064207 - 1 = 0.064207$$

Como

$$e_0(\overline{x}) = 1.063015 - 1 = 0.063015$$

e

$$e_1(\overline{x}) = 1.063015 - 1.064207 = -0.001192$$

vemos que para este caso

$$\mid f(\overline{x}) - p_0(\overline{x}) \mid \approx \mid p_1(\overline{x}) - p_0(\overline{x}) \mid$$

e

$$\mid f(\overline{x}) - p_1(\overline{x}) \mid \ll \mid p_1(\overline{x}) - p_0(\overline{x}) \mid$$

e que

$$\lambda_1 = -0.001192/0.064207 = -0.0186$$

Na Tabela 4.3 apresentamos os vários valores de $p_n(\overline{x}) - p_{n-1}(\overline{x})$ e de λ_n para $\overline{x} = 0$, 1.13 e 2.2. Vemos que normalmente $\mid \lambda_n \mid \leq 1$ (o caso desfavorável $\overline{x} = 0$ é um exemplo de extrapolação que, como já referimos atrás, conduz a resultados com grande falta de precisão). Portanto, escolhida uma tolerância ϵ para o erro de interpolação, basta aumentar o número de pontos $n+1$ até que seja satisfeita a condição (4.17), o que normalmente garante que o erro absoluto de interpolação seja inferior a ϵ. ■

Vamos agora ver que a diferença dividida de ordem k $f[x_0, \ldots, x_k]$ pode relacionar-se com a derivada de ordem k de f; a relação obtida será depois utilizada em (4.16) para estimar o erro $e_n(\overline{x})$.

Teorema 4.2 *Sejam $k+1$ nós distintos $x_0, \ldots, x_k$ e $[a,b] \equiv \text{int}[x_0, \ldots, x_k]$. Seja f definida em $[a,b]$ e $f \in C^k(a,b)$. Então existe um $\xi \in (a,b)$ tal que*

$$f[x_0, \ldots, x_k] = \frac{f^{(k)}(\xi)}{k!} \tag{4.18}$$

Demonstração: Para $k = 1$, o teorema reduz-se ao teorema de Lagrange ($f[x_0, x_1] = [f(x_1) - f(x_0)]/(x_1 - x_0)$). No caso geral, nota-se que $e_k(x) = f(x) - p_k(x)$ tem pelo menos $k+1$ zeros distintos $x_0, \ldots, x_k$ em $[a,b]$. Como f e portanto e_k são diferenciáveis até a ordem k em (a,b), então segue-se do teorema de Rolle que e_k' tem pelo menos k zeros em (a,b), logo e_k'' tem pelo menos $k-1$ zeros em (a,b), prosseguindo assim conclui-se finalmente que $e_k^{(k)}$ tem pelo menos um zero ξ em (a,b). Então

$$e_k^{(k)}(\xi) = f^{(k)}(\xi) - p_k^{(k)}(\xi) = 0 \tag{4.19}$$

Por outro lado, $f[x_0, \ldots, x_k]$ é por definição o coeficiente do termo em x^k do polinómio interpolador $p_k(x)$. Logo, para qualquer x,

$$p_k^{(k)}(x) = f[x_0, \ldots, x_k]\ k! \tag{4.20}$$

De (4.19) e (4.20) tira-se (4.18) como pretendiamos. ■ Atendendo a (4.18), podemos escrever

$$\min_{x \in (a,b)} \frac{f^{(k)}(x)}{k!} \leq f[x_0, \ldots, x_k] \leq \max_{x \in (a,b)} \frac{f^{(k)}(x)}{k!} \tag{4.21}$$

Exemplo 4.3 *Para $f[1.00, 1.05, 1.10]$ dado na Tabela 4.2 de diferenças divididas de $f(x) = \sqrt{x}$ verifique (4.18) e (4.21).*

Usando (4.18) temos que

$$f[1.00, 1.05, 1.10] = f''(\xi)/2! = -0.125(\xi)^{-3/2}$$

com $1.0 < \xi < 1.1$. Dado que $f[1.00, 1.05, 1.10] = -0.1162$ é fácil de ver neste caso que $\xi = 1.049871$. Pode-se ainda ver neste caso que (4.21) reduz-se a

$$-0.125(1)^{-3/2} < -0.1162 < -0.125(1.1)^{-3/2} = -0.108348$$

■

Voltando agora à expressão (4.16) do erro de interpolação e utilizando o Teorema 4.2 podemos enunciar o seguinte teorema.

Teorema 4.3 *Seja f definida em $[a, b]$ e $f \in C^{n+1}(a, b)$. Seja $p_n(x)$ o polinómio de grau menor ou igual a n que interpola $f(x)$ nos $n+1$ nós distintos $x_0, \ldots, x_n$ em $[a, b]$. Então para todo o $x \in [a, b]$ existe um $\xi = \xi(x) \in (c, d)$, com $[c, d] \equiv \text{int}[x_0, \ldots, x_n, x]$, tal que*

$$e_n(x) = f(x) - p_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!} \prod_{i=0}^{n} (x - x_i) \tag{4.22}$$

■

No caso da interpolação linear, $n = 1$, (4.22) reduz-se a

$$e_1(x) = \frac{f''(\xi)}{2} (x - x_0)(x - x_1)$$

Se x é um ponto entre x_0 e x_1, então $x_0 < \xi < x_1$ (supõe-se sem perda de generalidade que $x_0 < x_1$). Supondo que $| f''(x) | \leq M_2$ para $x_0 < x < x_1$, então,

$$| e_1(x) | \leq \frac{M_2}{2} \max_{x_0 \leq t \leq x_1} | (t - x_0)(t - x_1) | \qquad x_0 \leq x \leq x_1$$

Designando $x_1 - x_0$ por h, o máximo verifica-se para $t = (x_0 + x_1)/2$ e é $h^2/4$. Logo para qualquer $x \in [x_0, x_1]$ tem-se

$$| e_1(x) | \leq \frac{M_2}{8} h^2$$

No caso da extrapolação, em que $x \notin [x_0, x_1]$, tem-se

$$| (x - x_0)(x - x_1) | > h^2/4$$

excepto se

$$x_0 - \frac{\sqrt{2} - 1}{2} h \leq x < x_0$$

ou

$$x_1 \leq x < x_1 + \frac{\sqrt{2} - 1}{2} h$$

Por isso a extrapolação conduz geralmente a resultados com pouca precisão.

No exemplo $f(x) = \sqrt{x}$, considerado anteriormente, os resultados apresentados na Tabela 4.3 mostram-nos a falta de precisão para os casos de extrapolação $x = 0$ e $x = 2.2$. A maior imprecisão para $x = 0$ do que para $x = 2.2$ deve-se ao facto de $f^{(n+1)}(\xi)$ em (4.22) tomar valores absolutos mais elevados quando $x = 0$ ($f^{(n+1)}(x)$ não é limitada em $x = 0$).

No caso geral $n \geq 1$, para facilitar o estudo do erro dado por (4.22), vamos considerar os nós igualmente espaçados, *i.e.*, $x_{i+1} - x_i = h$. Temos então que $x_i = x_0 + ih$. Introduzimos a mudança de variável

$$x = x_0 + \theta h \quad \text{ou} \quad \theta = \frac{x - x_0}{h} \tag{4.23}$$

Assim $x - x_i = (\theta - i)h$ e para $\theta = i$ temos $x = x_i$. Vemos que (4.22) reduz-se a

$$e_n(x) = f^{(n+1)}(\xi) h^{n+1} C_{n+1}(\theta) \tag{4.24}$$

em que a função binomial $C_k(\theta)$ é definida por

$$C_k(\theta) = \begin{cases} 1 & k = 0 \\ \dfrac{\theta(\theta-1)\cdots(\theta-k+1)}{k!} & k > 0 \end{cases} \tag{4.25}$$

Nota-se que $C_k(\theta)$ é precisamente um coeficiente binomial $\begin{pmatrix} \theta \\ k \end{pmatrix}$ quando θ é inteiro.

Atendendo a (4.24) vemos que o erro de interpolação está relacionado com a função binomial. Por isso vamos apresentar algumas propriedades desta função. Da definição (4.25) vê-se imediatamente que $C_{n+1}(\theta)$ é um polinómio em θ de grau $n+1$ cujos zeros são $\theta = 0, 1, \ldots, n$. Considerando dois pontos $\theta = n/2 - t$ e $\theta = n/2 + t$, equidistantes do ponto $n/2$ (ponto médio dos zeros), obtém-se a seguinte relação de simetria

$$C_{n+1}(\frac{n}{2} - t) = (-1)^{n+1} C_{n+1}(\frac{n}{2} + t)$$

i.e., $C_{n+1}(\theta)$ é simétrico ou antisimétrico em relação a $\theta = n/2$ conforme n for ímpar ou par. Em cada intervalo $(i, i+1)$, com $0 \leq i \leq n-1$, $|C_{n+1}(\theta)|$ possue um único máximo. Tem-se ainda que para $\theta \neq 0, 1, \ldots, n$

$$\begin{aligned} |C_{n+1}(\theta)| &< |C_{n+1}(\theta - 1)| \qquad \theta \leq \frac{n}{2} \\ |C_{n+1}(\theta)| &< |C_{n+1}(\theta + 1)| \qquad \theta \geq \frac{n}{2} \end{aligned} \tag{4.26}$$

o que implica que os sucessivos máximos de $|C_{n+1}(\theta)|$ tomem valores cada vez maiores a medida que θ se aproxima de um dos extremos de $[0, n]$. Estas propriedades de $C_{n+1}(\theta)$ são ilustradas na Figura 4.1 onde se representa $C_6(\theta)$, $C_7(\theta)$ e $C_{11}(\theta)$.

De (4.26) e da Figura 4.1 depreende-se que $|C_{n+1}(\theta)|$ tem, no intervalo $[0, n]$, dois máximos absolutos para valores de θ pertencentes respectivamente aos intervalos $[0, 1/2]$ e $[n - 1/2, n]$. Designando por

$$M_e(n+1) = \max_{0 \leq \theta \leq n} h^{n+1} |C_{n+1}(\theta)|$$

Figura 4.1: Funções binomial C_6, C_7 e C_{11} associada ao erro de interpolação nos casos de 6, 7 e 11 nós equidistantes

estes máximos[1] multiplicados pelo factor h^{n+1} presente em (4.24), temos que

$$h^{n+1}|C_{n+1}(\frac{1}{2})| = \frac{h^{n+1}(2n)!}{2^{2n+1}(n!)^2(n+1)} \leq M_e(n+1) \leq \frac{1}{2}\lim_{\theta\to 0} h^{n+1}|C_{n+1}(\theta)/\theta| = \frac{h^{n+1}}{2(n+1)} \tag{4.27}$$

Pode-se mostrar que $M_e(n+1)$ é assintótico a

$$\frac{h^{n+1}}{e(n+1)\ln(n+1)} \tag{4.28}$$

quando n tende para infinito. Por outro lado, para a interpolação perto do centro do intervalo $[0, n]$, temos que

$$M_e^*(n+1) = \max_{\theta\in[\frac{n}{2}-\frac{1}{2},\frac{n}{2}+\frac{1}{2}]} h^{n+1}|C_{n+1}(\theta)| = \frac{h^{n+1}n!}{2^{2n+1}[\frac{n}{2}]![\frac{n+1}{2}]!}$$

em que [.] designa a parte inteira de um número real. Pode-se mostrar que $M_e^*(n+1)$ é assintótico a

$$\frac{h^{n+1}}{2^{n+1}\sqrt{\pi n/2}} \tag{4.29}$$

quando n tende para infinito. De (4.28) e (4.29) concluimos que $M_e(n+1)/M_e^*(n+1)$ é assintótico a

$$\frac{\sqrt{\pi/2}\ 2^{n+1}}{e\sqrt{n+1}\ \ln(n+1)}$$

quando n tende para infinito. Deste último resultado podemos concluir que, p.ex., para $n = 20$

$$M_e(n+1) \approx 69305\ M_e^*(n+1)$$

Vemos assim que os máximos absolutos de $|C_{n+1}(\theta)|$, localizados perto dos extremos do intervalo $[0, n]$, tomem valores cada vez maiores relativamente aos restantes máximos a medida que n for maior.

Perante as propriedades de $C_{n+1}(\theta)$, podemos concluir que para reduzir o erro de interpolação os nós $x_0, \ldots, x_n$ deverão ser escolhidos de modo que o ponto de interpolação $\overline{x}$ esteja o mais perto possível do ponto médio de int$[x_0, \ldots, x_n]$. Na Tabela 4.2, referente ao Exemplo 4.1 apresentado atrás, ao pretendermos obter aproximações de $f(\overline{x})$ com $\overline{x} = 1.13$ utilizando os sucessivos polinómios interpoladores deveríamos ter utilizado sucessivamente os nós $x_0 = 1.15$, $x_1 = 1.1$, $x_2 = 1.2$, $x_3 = 1.05$, $x_4 = 1.0$ (deveria ser $x_4 = 1.25$ se a tabela contivesse este valor).

Do que foi dito até agora neste capítulo, podemos então apresentar o seguinte algoritmo, baseado na fórmula interpoladora de Newton (4.10), nas diferenças divididas (4.14), no critério de paragem (4.17) e na propriedade (4.26) da função binomial.

Algoritmo 4.1 *Interp* $(x, f, n, \overline{x}, p, k, \epsilon)$

[1] Os máximos são atingidos para valores de θ que são assintóticos a $1/\ln(n+1)$ e $n - 1/\ln(n+1)$ quando n tende para infinito.

NOTA: x e f são vectores de componentes x_i e $f(x_i)$, $i = 0, \ldots, n$, respectivamente; p é o valor do polinómio interpolador de grau k na abcissa $\overline{x}$.

ORDENACÃO DOS PONTOS: Ordenar (x_i, f_i) de modo a que $|x_{i+1} - \overline{x}| \geq |x_i - \overline{x}|$, $i = 0, \ldots, n-1$.

INICIALIZACÃO:

$d_0 = f(x_0),\ pa = f(x_0),\ \omega = \overline{x} - x_0.$

CICLO: PARA $k = 1, \ldots, n$ FAZER

$d_k = f(x_k).$

CICLO: PARA $j = 1, \ldots, k$ FAZER

fazer $d_{k-j} = (d_{k-j+1} - d_{k-j})/(x_k - x_{k-j}).$

FIM DO CICLO j.

$p = pa + \omega d_0.$

SE $|p - pa| \leq \epsilon$ ENTÃO, SAÍDA

$pa = p$

fazer $\omega = \omega.(\overline{x} - x_k)$

FIM DO CICLO k

Nota-se que a tabela de diferenças divididas é construida no ciclo j à medida que é necessária para o cálculo do polinómio interpolador. Para o exemplo da Tabela 4.2 com $\overline{x} = 1.13$ representa-se na Tabela 4.4 esquematicamente a sequência da construção da tabela de diferenças divididas.

Vamos agora ver, com alguns exemplos, que o erro de interpolação nem sempre tende para zero quando o grau do polinómio interpolador tende para infinito.

Consideremos a função $f(x) = \cos x$ tabelada para $x_i = ih,\ i = 0, \ldots, n$. Para $h = 2\pi$, é imediato que qualquer que seja o número $n+1$ de nós x_i utilizados, o polinómio interpolador é sempre igual a um. No entanto se h for suficientemente pequeno é de esperar, em face da expressão do erro (4.24), que, para um dado $x > 0$, $\lim_{n\to\infty} e_n(x) = 0$. Com efeito, utilizando (4.27) ou (4.28), pode provar-se, não só para $f(x) = \cos x$, mas para qualquer outra função indefinidamente derivável com todas as suas derivadas limitadas em $(0, \infty)$, que $\lim_{n\to\infty} e_n(x) = 0$ se for $h \leq 1$.

Consideremos a aproximação de uma dada função $f(x)$ num intervalo $[a, b]$ usando polinómios interpoladores. Escolhemos nós igualmente espaçados $x_i = a + ih,\ i = 0, \ldots, n,\ h = (b-a)/n$. Será que, para $a \leq x \leq b$, $\lim_{n\to\infty} e_n(x) = 0$? Para o caso $f(x) = e^x$ em $[0, 1]$ a resposta é sim. Mas para os casos $f(x) = \sqrt{x}$ em $[0, 1]$ ou $f(x) = |x|$ em $[0, 1]$ a resposta é não. Nestes últimos casos f não é indefinidamente derivável em $[a, b]$. No entanto pode-se apresentar exemplos de funções indefinidamente deriváveis tais que $\lim_{n\to\infty} e_n(x) \neq 0$. Um exemplo clássico, devido a *Runge*, é

$$f(x) = (1 + 25x^2)^{-1}, x \in [-1, 1]$$

Tabela 4.4: Construção da tabela de diferenças divididas

	x_i	$f(x_i)$	$f[x_i, x_{i+1}]$	$f[\ ,\ ,\ ]$	$f[\ ,\ ,\ ,\ ]$	$f[\ ,\ ,\ ,\ ,\ ]$
	$x_4 = 1.00$	d_4				
			d_3			
	$x_3 = 1.05$	d_3		d_2		
			d_2		d_1	
	$x_1 = 1.10$	d_1		d_1		d_0
$\overline{x} = 1.13$			d_0		d_0	
	$x_0 = 1.15$	d_0		d_0		
			d_1			
	$x_2 = 1.20$	d_2				

Neste caso para qualquer $|x| > 0.728$,

$$\sup_{n \in \mathbb{N}} |f(x) - p_n(x)| = \infty$$

Nas Figuras 4.2 e 4.3 são ilustradas os 2 últimos casos referidos.

Destes exemplos podemos concluir que nem sempre é vantajoso utilizar polinómios interpoladores de grau elevado. Assim um programa baseado no Algoritmo 4.1, apresentado atrás, deverá ser protegido contra a eventualidade de $|p_{k+1}(\overline{x}) - p_k(\overline{x})|$ começar a crescer quando k cresce, visto que nesta eventualidade normalmente $|e_{k+1}(\overline{x})| > |e_k(\overline{x})|$.

4.5 Nós igualmente espaçados

A fórmula interpoladora de Newton pode simplificar-se quando os nós x_i estiverem igualmente espaçados. Seja $x_i = x_0 + ih$, $f_i = f(x_i)$ com i inteiro. Definimos *diferenças* (não divididas) *progressivas* por

$$\Delta f_i = f_{i+1} - f_i \tag{4.30}$$

A partir destas diferenças de 1ª ordem podemos definir diferenças de 2ª ordem

$$\Delta^2 f_i = \Delta(\Delta f_i) = \Delta f_{i+1} - \Delta f_i = f_{i+2} - 2f_{i+1} + f_i \tag{4.31}$$

Figura 4.2: Interpolação de $|x|$ com 11 e 21 nós equidistantes

Figura 4.3: Interpolação de $(1 + 25x^2)^{-1}$ com 11 e 21 nós equidistantes

Em geral defina-se diferença progressiva de ordem k por

$$\Delta^k f_i = \begin{cases} f_i & k = 0 \\ \Delta^{k-1} f_{i+1} - \Delta^{k-1} f_i & k \geq 1 \end{cases} \tag{4.32}$$

Por indução prova-se que

$$\Delta^k f_i = \sum_{j=0}^{k} (-1)^j \begin{pmatrix} k \\ j \end{pmatrix} f_{i+k-j} \tag{4.33}$$

em que

$$\begin{pmatrix} k \\ j \end{pmatrix} = \frac{k!}{j!(k-j)!}$$

é um coeficiente binomial (para $k = 1$ e $k = 2$, (4.33) reduz-se a (4.30) e (4.31) respectivamente). Também por indução pode-se provar que

$$f_{i+k} = \sum_{j=0}^{k} \begin{pmatrix} k \\ j \end{pmatrix} \Delta^j f_i \tag{4.34}$$

Tal como as diferenças divididas podemos apresentar as diferenças progressivas sob a forma de uma tabela triangular como a Tabela 4.5.

Tabela 4.5: Tabela de diferenças progressivas

x_i	f_i	Δf_i	$\Delta^2 f_i$	$\Delta^3 f_i$	$\Delta^4 f_i$
x_0	f_0				
		Δf_0			
x_1	f_1		$\Delta^2 f_0$		
		Δf_1		$\Delta^3 f_0$	
x_2	f_2		$\Delta^2 f_1$		$\Delta^4 f_0$
		Δf_2		$\Delta^3 f_1$	
x_3	f_3		$\Delta^2 f_2$		
		Δf_3			
x_4	f_4				

Visto que pretendemos simplificar a fórmula de Newton vamos relacionar as diferenças divididas com as diferenças progressivas. Utilizando (4.13) e (4.30) temos para diferenças de 1ª ordem que

$$f[x_i, x_{i+1}] = \frac{f_{i+1} - f_i}{x_{i+1} - x_i} = \frac{1}{h} \Delta f_i$$

Para as diferenças de ordem $k \geq 0$ vamos mostrar, por indução, que

$$f[x_i, \ldots, x_{i+k}] = \frac{1}{k! h^k} \Delta^k f_i \tag{4.35}$$

Como, por definição, tem-se

$$f[x_i] = f(x_i) = f_i = \Delta^o f_i$$

(4.35) é válido para $k = 0$. Admitindo que (4.35) é válido para $k = m \geq 0$ e utilizando (4.14) e (4.32) temos que

$$\begin{aligned} f[x_i, \ldots, x_{i+m+1}] &= \frac{f[x_{i+1}, \ldots, x_{i+m+1}] - f[x_i, \ldots, x_{i+m}]}{x_{i+m+1} - x_i} \\ &= \frac{\Delta^m f_{i+1}/(m!h^m) - \Delta^m f_i/(m!h^m)}{(m+1)h} \\ &= \frac{\Delta^{m+1} f_i}{(m+1)!h^{m+1}} \end{aligned}$$

o que prova que (4.35) também é válido para $k = m + 1$. Aproveitamos para notar, atendendo a (4.18) e (4.35), que

$$\Delta^k f_i = h^k f^{(k)}(\xi) \qquad x_i < \xi < x_{i+k} \tag{4.36}$$

Este resultado (4.36) é utilizado na derivação numérica.

Substituindo (4.35) na fórmula de Newton (4.10) temos, para o polinómio p_n de grau menor ou igual que n que interpola f em $x_0, \ldots, x_n$, que

$$p_n(x) = \sum_{i=0}^{n} \frac{1}{i!h^i} \Delta^i f_0 \prod_{j=0}^{i-1} (x - x_j)$$

Utilizando a mudança de variável (4.23), $\theta = (x - x_0)/h$, e a função binomial $C_i(\theta)$, definida atrás em (4.25), obtemos a *fórmula interpoladora de Newton progressiva*

$$\begin{aligned} p_n(x_0 + \theta h) &= f_0 + \theta \Delta f_0 + \frac{\theta(\theta-1)}{2!} \Delta^2 f_0 + \cdots + \frac{\theta(\theta-1)\cdots(\theta-n+1)}{n!} \Delta^n f_0 \\ &= \sum_{i=0}^{n} C_i(\theta) \Delta^i f_0 \end{aligned} \tag{4.37}$$

Se $\theta = k$ com k inteiro, $0 \leq k \leq n$, então (4.37) reduz-se a

$$p_n(x_0 + kh) = f(x_k) = \sum_{i=0}^{n} \binom{k}{i} \Delta^i f_0 = \sum_{i=0}^{k} \binom{k}{i} \Delta^i f_0$$

que coincida com (4.34).

Exemplo 4.4 *Para os mesmos dados do Exemplo 4.1 ilustrar a utilização da fórmula de Newton progressiva para $f(x) = \sqrt{x}$.*

As diferenças não divididas de $\sqrt{x}$ são dadas na Tabela 4.6[2] Escolhendo $x_0 = 1$ então temos, p.ex.,

$$\Delta^2 f_1 = \Delta f_2 - \Delta f_1 = 0.023572 - 0.024114 = -0.000542$$

[2]Nas tabelas de diferenças não divididas é habitual representar as diferenças sem os zeros não significativos.

Tabela 4.6: Diferenças progressivas de $f(x) = \sqrt{x}$

x_i	f_i	Δf_i	$\Delta^2 f_i$	$\Delta^3 f_i$	$\Delta^4 f_i$
1.00	1.000000				
		24695			
1.05	1.024695		−581		
		24114		39	
1.10	1.048809		−542		−5
		23572		34	
1.15	1.072381		−508		
		23064			
1.20	1.095445				

resultado que poderíamos obter utilizando (4.35). De facto da Tabela 4.2 temos

$$f[x_1, x_2, x_3] = -0.1084$$

e como

$$h = 1.05 - 1.00 = 0.05$$

obtem-se

$$\Delta^2 f_1 = 2! h^2 f[x_1, x_2, x_3] = -0.000542$$

Suponhamos que, utilizando a Tabela 4.6, queremos obter $p_n(1.13)$ para $n = 1, 2, 3, 4$. Para minimizar o erro de interpolação vamos escolher $x_0, \ldots, x_n$ os nós mais perto de 1.13. Assim, notando que

$$\theta = (1.13 - x_0)/0.05$$

, teremos sucessivamente:

$$\begin{array}{llllll}
n = 1 & x_0 = 1.10 & x_1 = 1.15 & & & \theta = 0.6 \\
n = 2 & x_0 = 1.10 & x_1 = 1.15 & x_2 = 1.20 & & \theta = 0.6 \\
n = 3 & x_0 = 1.05 & x_1 = 1.10 & x_2 = 1.15 & x_3 = 1.20 & \theta = 1.6 \\
n = 4 & x_0 = 1.00 & x_1 = 1.05 & x_2 = 1.10 & x_3 = 1.15 \quad x_4 = 1.20 & \theta = 2.6
\end{array}$$

Então, utilizando (4.37), teremos

$$\begin{aligned}
p_1(1.13) &= 1.048809 + 0.6(0.023572) = \underline{1.062}952 \\
p_2(1.13) &= p_1(1.13) + \frac{0.6(0.6-1)}{2}(-0.000508) = \underline{1.06301}3 \\
p_3(1.13) &= 1.024695 + 1.6(0.024114) + \frac{1.6(0.6)}{2}(-0.000542) + \\
&\quad \frac{1.6(0.6)(-0.4)}{6}(0.000034) = \underline{1.063015} \\
p_4(1.13) &= 1. + 2.6(0.024695) + \frac{2.6(1.6)}{2}(-0.000581) + \frac{2.6(1.6)(0.6)}{6}(0.000039) + \\
&\quad \frac{2.6(1.6)(0.6)(-0.4)}{24}(-0.000005) = \underline{1.063015}
\end{aligned}$$

Vemos que os valores obtidos para $p_1(1.13)$ e $p_2(1.13)$ constituem melhores aproximações de $\sqrt{1.13}$ do que os obtidos anteriormente na Tabela 4.3, visto termos escolhidos os nós de interpolação mais perto de 1.13. ■

Vamos ainda introduzir outra fórmula interpolador com diferenças não divididas. Definimos *diferenças regressivas* por

$$\begin{aligned} \nabla f_i &= f_i - f_{i-1} \\ \nabla^{k+1} f_i &= \nabla^k f_i - \nabla^k f_{i-1} \qquad k \geq 1 \end{aligned}$$

Tal como para as anteriores diferenças, apresentamos na Tabela 4.7 a respectiva tabela triangular.

Tabela 4.7: Tabela de diferenças regressivas

x_i	f_i	∇f_i	$\nabla^2 f_i$	$\nabla^3 f_i$	$\nabla^4 f_i$
x_{-4}	f_{-4}				
		∇f_{-3}			
x_{-3}	f_{-3}		$\nabla^2 f_{-2}$		
		∇f_{-2}		$\nabla^3 f_{-1}$	
x_{-2}	f_{-2}		$\nabla^2 f_{-1}$		$\nabla^4 f_0$
		∇f_{-1}		$\nabla^3 f_0$	
x_{-1}	f_{-1}		$\nabla^2 f_0$		
		∇f_0			
x_0	f_0				

Pode provar-se fácilmente por indução que para $k \geq 0$

$$\nabla^k f_i = \Delta^k f_{i-k}$$

Logo, existem para as diferenças regressivas resultados análogos aos obtidos para as diferenças progressivas. Nomeadamente, a (4.35) corresponde

$$\nabla^k f_i = k!\, h^k\, f[x_{i-k}, \ldots, x_i] \tag{4.38}$$

Substituindo na fórmula de Newton (4.10) $x_0, x_1, \ldots, x_n$ por $x_0, x_{-1}, \ldots, x_{-n}$, utilizando (4.38) e a mudança de variável (4.23) obtemos a *fórmula interpoladora de Newton regressiva.*

$$\begin{aligned} p_n(x_0 + \theta h) &= f_0 + \theta \nabla f_0 + \frac{\theta(\theta+1)}{2!} \nabla^2 f_0 + \cdots + \frac{\theta(\theta+1)\cdots(\theta+n-1)}{n!} \nabla^n f_0 \\ &= \sum_{i=0}^{n} C_i(\theta + i - 1) \nabla^i f_0 \end{aligned} \tag{4.39}$$

Exemplo 4.5 *Ilustrar a utilização da fórmula de Newton regressiva (4.39), considerando a extrapolação de $\sqrt{x}$ com $\overline{x} = 2.2$, a partir da Tabela 4.6.*

Neste caso $x_0 = 1.2$, $\theta = 20$, logo

$$\begin{aligned}
p_1(2.2) &= 1.095445 + 20(0.023064) = \underline{1}.556725 \\
p_2(2.2) &= p_1(2.2) + \frac{20(20+1)}{2}(-0.000508) = \underline{1.4}50045 \\
p_3(2.2) &= p_2(2.2) + \frac{20(21)(20+2)}{6}(0.000034) = \underline{1.5}02405 \\
p_4(2.2) &= p_3(2.2) + \frac{20(21)(22)(20+3)}{24}(-0.000005) = \underline{1.4}58130
\end{aligned}$$

Como era de esperar só $p_4(2.2)$ coincide com o valor já anteriormente obtido na Tabela 4.3. ■

4.6 Nós de Chebyshev

Da expressão (4.22) do erro de interpolação vemos que este é influenciado por dois termos (o polinómio ω_{n+1} é o definido por (4.6)):

$$f^{(n+1)}(\xi) \text{ e } \omega_{n+1}(x)/(n+1)!$$

O último termo é independente da função interpolada e já vimos algumas das suas propriedades quando os nós x_i estão igualmente espaçados, caso em que se reduz a $h^{n+1}C_{n+1}(\theta)$ (ver (4.24)). Neste caso $|\omega_{n+1}(x)/(n+1)!|$ atinge os seus maiores valores na proximidade dos extremos do intervalo de interpolação $[a, b]$. É de esperar que com uma distribuição de nós $x_0, \ldots, x_n$ mais densa na proximidade dos extremos de $[a, b]$ do que no seu centro, $|\omega_{n+1}(x)/(n+1)!|$ atinge um valor máximo mais pequeno do que com uma distribuição de nós igualmente espaçados. Designando por

$$M(n+1) = \max_{a \le x \le b} |\omega_{n+1}(x)/(n+1)!|$$

mostraremos mais tarde, que para o intervalo $[-1, 1]$ ($a = -1, b = 1$), a distribuição de $n+1$ nós para a qual corresponde o menor $M(n+1)$ é

$$t_i = \cos\frac{(2i+1)\pi}{2(n+1)} \qquad i = 0, \ldots, n \tag{4.40}$$

Estes nós são os zeros de um polinómio T_{n+1} de grau $n+1$ conhecido na literatura por polinómio de *Chebyshev*.

Defina-se polinómio de Chebyshev de grau n como sendo

$$T_n(x) = \cos(n \arccos x) \tag{4.41}$$

Vejamos que T_n é de facto um polinómio de grau n. Para $n = 0$ e $n = 1$ (4.41) reduz-se a

$$T_0(x) = 1 \qquad T_1(x) = x$$

Fazendo a mudança de variável $\cos\theta = x$, obtemos

$$T_n(\cos\theta) = \cos n\theta$$

portanto podemos escrever

$$T_{n-1}(\cos\theta) = \cos(n-1)\theta = \cos n\theta\cos\theta + \sin n\theta\sin\theta$$
$$T_{n+1}(\cos\theta) = \cos(n+1)\theta = \cos n\theta\cos\theta - \sin n\theta\sin\theta$$

Somando estas duas expressões, obtemos a fórmula de recorrência

$$T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x) \tag{4.42}$$

Fazendo sucessivamente $n = 1, 2, 3, \ldots$ nesta fórmula obtemos

$$\begin{aligned} T_2(x) &= 2x^2 - 1 \\ T_3(x) &= 4x^3 - 3x \\ T_4(x) &= 8x^4 - 8x^2 + 1 \\ &\cdots \end{aligned}$$

Em geral, se T_{n-1} e T_n forem polinómios respectivamente de grau $n-1$ e n, então da fórmula de recorrência (4.42) concluimos que T_{n+1} é um polinómio de grau $n+1$. Como T_n é um polinómio de grau n, para $n = 0$ e $n = 1$, fica demonstrado por indução que, para qualquer n natural, T_n é um polinómio de grau n.

Por indução também se demonstra que o coeficiente do termo em x^n de $T_n(x)$ é 2^{n-1} para $n \geq 1$. Assim, podemos escrever

$$2^{-n}T_{n+1}(x) = \tau_{n+1}(x) = (x - t_0)(x - t_1)\cdots(x - t_n) \tag{4.43}$$

em que $t_0, t_1, \ldots, t_n$ são os zeros do polinómio de Chebyshev T_{n+1}. É fácil de ver que estes zeros são os dados por (4.40) e correspondem aos zeros de $\cos(n+1)\theta$ no intervalo $[0, \pi]$. Portanto todos os zeros de um polinómio de Chebyshev localizam-se no intervalo $[-1, 1]$. Por outro lado, os extremos (máximos e mínimos) de T_{n+1} no intervalo $[-1, 1]$ localizam-se nos pontos

$$t'_i = \cos\frac{i\pi}{n+1} \qquad i = 0, 1, \ldots, n+1$$

e correspondem aos extremos de $\cos(n+1)\theta$ no intervalo $[0, \pi]$. Portanto

$$T_{n+1}(t'_i) = (-1)^i \qquad i = 0, 1, \ldots, n+1 \tag{4.44}$$

Assim, temos que

$$\max_{-1\leq x\leq 1} |T_{n+1}(x)| = 1$$

Para melhor compreensão das propriedades dos polinómios de Chebyshev representamos na Figura 4.4 alguns destes.

Exemplo 4.6 *Considerar a aproximação da função $exp(x)$ no intervalo $[-1, 1]$ por um polinómio p_3 de grau 3 utilizando (1) nós igualmente espaçados e (2) os zeros do polinómio de Chebyshev T_4.*

Figura 4.4: Polinómios de Chebyshev T_1, T_2, T_3, T_4, T_5 e T_6

No primeiro caso, os nós de interpolação são

$$x_0 = -1 \qquad x_1 = -0.3333333 \qquad x_2 = 0.3333333 \qquad x_3 = 1$$

Depois de determinarmos o respectivo polinómio interpolador p_3, podemos obter

$$\max_{x\in[-1,1]} |\exp(x) - p_3(x)| = 0.00998$$

No caso dos nós de interpolação serem

$$\begin{aligned} t_0 &= \cos\frac{\pi}{8} = 0.9238795 \\ t_1 &= \cos\frac{3\pi}{8} = 0.3826834 \\ t_2 &= \cos\frac{5\pi}{8} = -0.3826834 \\ t_3 &= \cos\frac{7\pi}{8} = -0.9238795 \end{aligned}$$

podemos obter

$$\max_{x\in[-1,1]} |\exp(x) - c_3(x)| = 0.00666$$

em que c_3 é o polinómio interpolador. No conjunto $\mathcal{P}_3$ dos polinómios de grau menor ou igual a 3 existe um polinómio q_3 que melhor aproxima a função $exp(x)$, designado por polinómio *minimax*, e podemos obter

$$\min_{p_3\in\mathcal{P}_3} \max_{x\in[-1,1]} |\exp(x) - p_3(x)| = |\exp(x) - q_3| = 0.00553$$

Dos resultados apresentados, concluímos que ao escolher para nós de interpolação os zeros de T_4 obtemos um polinómio c_3 que se aproxima bastante do polinómio minimax q_3. ■

Se no intervalo $[-1, 1]$ utilizarmos para nós de interpolação os zeros de T_{n+1} então, atendendo a (4.43) e (4.44), temos

$$\tau_{n+1}(t'_i) = \frac{(-1)^i}{2^n} \qquad i = 0, 1, \ldots, n+1 \tag{4.45}$$

e portanto o máximo de $|\omega_{n+1}(x)/(n+1)!|$ é neste caso

$$M_c(n+1) = \max_{-1\le x\le 1} |\tau_{n+1}(x)/(n+1)!| = \frac{1}{2^n(n+1)!}$$

Vejamos que $M_c(n+1)$ é o mínimo dos $M(n+1)$, *i.e.*, que temos o seguinte:

Teorema 4.4 *Seja $\tau_{n+1} = 2^{-n}T_{n+1}$ e $\omega_{n+1}(x)$ qualquer polinómio de grau $n+1$ da forma $(x - x_0)\cdots(x - x_n)$. Então*

$$\max_{-1\le x\le 1} |\tau_{n+1}(x)| = \frac{1}{2^n} \le \max_{-1\le x\le 1} |\omega_{n+1}(x)| \tag{4.46}$$

Demonstração: Suponhamos que, ao contrário do que o teorema afirma, existe um polinómio ω_{n+1} tal que

$$\max_{-1\leq x\leq 1} |\omega_{n+1}(x)| < \frac{1}{2^n} \tag{4.47}$$

e consideremos o polinómio $r = \tau_{n+1} - \omega_{n+1}$. Notando que τ_{n+1} e ω_{n+1} são polinómios de grau $n+1$ com o coeficiente de x^{n+1} unitário, é óbvio que $grau(r) \leq n$. Por outro lado, atendendo a (4.45), temos

$$r(t_i') = (-1)^i 2^{-n} - \omega_{n+1}(t_i') \qquad i = 0, 1, \ldots, n+1 \tag{4.48}$$

Como se admitiu que ω_{n+1} satisfaz a relação (4.47), podemos concluir de (4.48) que o polinómio r, de grau quanto muito igual a n, assume nos $n+2$ pontos t_i' alternadamente sinais diferentes e portanto r tem pelo menos $n+1$ zeros; o que implica que $r \equiv 0$. Então $\omega_{n+1} = \tau_{n+1}$ e portanto (4.47) não pode ser satisfeita. Fica assim provada a validade de (4.46). ■

No Exemplo 4.6, ao utilizar para nós de interpolação os zeros de T_4, vimos que o polinómio interpolador c_3 assim obtido é quase o polinómio minimax q_3. Este facto pode ser explicado pelo Teorema 4.4. Atendendo a (4.46), é fácil de ver que c_3 seria o polinómio minimax q_3 se $f^{(4)}(x) = exp(x)$ fosse constante no intervalo $[-1, 1]$, visto que neste caso o erro de interpolação seria proporcional a $\tau_{n+1}(x)$.

Quando o intervalo for $[a, b]$, em vez de $[-1, 1]$, basta efectuar a mudança de variável

$$x = \frac{a+b}{2} + \frac{b-a}{2}t \qquad t \in [-1, 1]$$

para transformar o intervalo $[-1, 1]$ no intervalo $[a, b]$. Assim se no intervalo $[a, b]$ utilizarmos para nós de interpolação

$$x_i = \frac{a+b}{2} + \frac{b-a}{2}\cos\frac{(2i+1)\pi}{2(n+1)} \qquad i = 0, \ldots, n$$

zeros de

$$T_{n+1}((x - \frac{a+b}{2})/(\frac{b-a}{2}))$$

então, o máximo de $|\omega_{n+1}(x)/(n+1)!|$ é neste caso

$$M_c(n+1) = \max_{a\leq x\leq b} |\omega_{n+1}(x)/(n+1)!| = \frac{(b-a)^{n+1}}{2^{2n+1}(n+1)!}$$

Pode-se mostrar que $M_c(n+1)$ é assintótico a

$$\frac{(\frac{b-a}{n})^{n+1}\, e^n}{2^{2n+1}\sqrt{n+1}\,\sqrt{2\pi}} \tag{4.49}$$

quando n tende para infinito. Vimos anteriormente, que para nós de interpolação igualmente espaçados, o comportamento assintótico de $M_e(n+1)$ era dado por (4.28). Substituindo em (4.28) h por $(b-a)/n$ e atendendo a (4.49), vemos que

$$M_e(n+1)/M_c(n+1)$$

é assintótico a

$$\sqrt{\frac{\pi}{2}}\ \frac{(\frac{4}{e})^{n+1}}{\sqrt{n+1}\ \ln(n+1)} \tag{4.50}$$

quando n tende para infinito. De igual modo atendendo a (4.29), vemos que

$$M_c(n+1)/M_e^*(n+1)$$

é assintótico a

$$\frac{1}{e}\ (\frac{e}{2})^{n+1} \tag{4.51}$$

quando n tende para infinito. Concluímos que, para n elevado, apesar de

$$M_c(n+1) \gg M_e^*(n+1)$$

temos

$$M_c(n+1) \ll M_e(n+1)$$

Assim utilizando (4.50) e (4.51), temos, p.ex., para $n = 20$

$$M_e(n+1) \approx 299.58\ M_c(n+1)$$
$$M_c(n+1) \approx 231.34\ M_e^*(n+1)$$

Na Figura 4.5 representamos no intervalo $[-1, 1]$ a função $\omega_{n+1}(x)/(n+1)!$ respectivamente para os casos dos nós igualmente espaçados e dos nós coincidentes com os zeros de T_{n+1}, para $n = 5$ e $n = 6$.

Referimos anteriormente 2 casos para o qual a sucessão $\{p_n\}$ de polinómios interpoladores não converge para a função $f(x)$ quando n tende para infinito no caso dos nós de interpolação estarem igualmente espaçados no intervalo $[-1, 1]$. Os casos foram para as funções $f(x) = |x|$ e $f(x) = (1+25x^2)^{-1}$ (exemplo de Runge). Se os nós de interpolação forem escolhidos de modo a serem os zeros do polinómio de Chebyshev de grau $n+1$, então a sucessão $\{c_n\}$ de polinómios interpoladores já converge para a função f. Isto acontece sempre que a função f for de classe $C^2([a, b])$. Nas Figuras 4.6 e 4.7 são ilustradas os 2 casos referidos.

4.7 Interpolação inversa

Somos frequentemente confrontados com o seguinte problema: dada uma função f para a qual se conhecem os pares de valores (x_i, y_i), $i = 0, 1, \ldots, n$, com $y_i = f(x_i)$, e dado um valor $\overline{y}$, pretende-se obter o correspondente valor $\overline{x}$ tal que $\overline{y} = f(\overline{x})$. Este problema é resolvido, ainda que em geral aproximadamente, pela interpolação inversa. Esta consiste em tratar x como variável dependente e $y = f(x)$ como variável independente e determinar o valor $p_n(\overline{y})$ a partir do polinómio $p_n(y)$ que interpola a função inversa $x = g(y) = f^{-1}(y)$. Estamos a admitir que a função inversa existe, *i.e.*, que a correspondência $y \to g(y)$ é unívoca, e portanto $f'(x) \neq 0$ para todo o x pertencente ao menor intervalo que contenha os x_i, $i = 0, 1, \ldots, n$, e o valor pretendido $\overline{x} = g(\overline{y})$.

Figura 4.5: Polinómios $\omega_6(x)/6!$ e $\omega_7(x)/7!$

Figura 4.6: Interpolação de $|x|$ com 11 e 21 nós coincidentes com os zeros de T_{n+1}

Figura 4.7: Interpolação de $(1 + 25x^2)^{-1}$ com 11 e 21 nós coincidentes com os zeros de T_{n+1}

O polinómio interpolador $p_n(y)$ pode obter-se, por exemplo, utilizando a fórmula interpoladora de Newton

$$p_n(y) = x_0 + \sum_{i=1}^{n} x[y_0, \ldots, y_i](y - y_0) \cdots (y - y_{i-1})$$

Atendendo a (4.22), o erro de interpolação é dado por

$$\overline{x} - p_n(\overline{y}) = \frac{g^{(n+1)}(\eta)}{(n+1)!} \quad (\overline{y} - y_i)$$

onde $\eta \in \text{int}[y_0, \ldots, y_n, \overline{y}]$. As derivadas de g podem exprimir-se em termos das derivadas de f, bastando para isso diferenciar a identidade

$$g[f(x)] = x$$

Obtem-se, por exemplo, as seguintes relações

$$g'(y) = [f'(x)]^{-1} \qquad g''(y) = -f''(x)[f'(x)]^{-3}$$

$$g'''(y) = [3(f''(x))^2 - f'(x)f'''(x)][f'(x)]^{-5}$$

Pode verificar-se, por indução, que para $g^{(k)}(y)$ com $k \geq 1$ a relação pretendida tem o factor $[f'(x)]^{-2k+1}$. Assim, se $\mid f'(x) \mid$ for pequeno para algum $x \in \text{int}[x_0, \ldots, x_n, \overline{x}]$, é de esperar que quanto maior for n pior será a interpolação.

Exemplo 4.7 *Considerar a função de Bessel*

$$J_0(x) = \frac{1}{\pi} \int_0^{\pi} \cos(x \cos \theta) d\theta$$

para a qual apresentamos na Tabela 4.8 alguns dos seus valores, tirados de Abramowitz e Stegun "Handbook of mathemetical functions". Determinar o valor $\overline{x}$ tal que $J_0(\overline{x}) = -0.35$.

Começamos por trocar a ordem das colunas x e J_0 na Tabela 4.8 e construimos uma tabela de diferenças divididas $x[y_i, \ldots, y_{i+k}]$ com $y_i = J_0(x_i)$. Notamos que na construção desta tabela não podem intervir os últimos dois pares de valores da Tabela 4.8 visto $J_0(x)$ ter um mínimo entre $x = 3.8$ e $x = 3.9$ ($J_0'(3.83171) = 0$). Na Tabela 4.9 apresenta-se as sucessivas aproximações $p_n(y)$ de $\overline{x} = 3.32650$. ■

Vemos que neste caso para $n \geq 3$ o erro da interpolação inversa vai aumentando com n. Isto era de esperar visto a derivada de J_0 tomar valores pequenos nos últimos pontos da tabela. Se para um dado x determinassemos $J_0(x)$ por interpolação directa bastaria tomar $n = 4$ para que o erro de interpolação fosse inferior a 5×10^{-9}. Sendo assim se pretendermos obter um valor suficientemente aproximado de $\overline{x}$, tal que $f(\overline{x}) = J_0(\overline{x}) + 0.35 = 0$, é preferível utilizar um dos métodos numéricos utilizados na resolução de equações não lineares, calculando f a partir da interpolação directa de J_0 (neste exemplo bastaria resolver a equação $p_4(x) + 0.35 = 0$ em que p_4 é o polinómio interpolador de grau 4 de J_0).

Tabela 4.8: Valores da função de Bessel $J_0(x)$

x	$J_0(x)$	x	$J_0(x)$
3.1	-0.29206435	3.6	-0.39176898
3.2	-0.32018817	3.7	-0.39923020
3.3	-0.34429626	3.8	-0.40255641
3.4	-0.36429560	3.9	-0.40182601
3.5	-0.38012774	4.0	-0.39714981

Tabela 4.9: Resultados da interpolação inversa

n	y_n	x_n	$x[y_0, \ldots, y_n]$	$p_n(y)$	$p_n(y) - p_{n-1}(y)$
0	-0.34429626	3.3	3.300000×10^0	$\underline{3.3}0000$	
1	-0.36429560	3.4	-5.000165×10^0	$\underline{3.32}852$	2.9×10^{-2}
2	-0.32018817	3.2	1.932055×10^1	$\underline{3.326}94$	-1.6×10^{-3}
3	-0.38012774	3.5	-2.904548×10^2	$\underline{3.326}24$	-7.1×10^{-4}
4	-0.39176898	3.6	9.489871×10^3	$\underline{3.326}93$	7.0×10^{-4}
5	-0.39923020	3.7	-7.717985×10^5	$\underline{3.32}457$	2.4×10^{-3}
6	-0.40255641	3.8	2.634572×10^8	$\underline{3.3}6425$	4.0×10^{-2}

4.8 Problemas

1. Considere da função $f(x) = (9x - 3)/(x - 2)$ o conjunto de pontos:

x	-3	-2	-1	0	1
$f(x)$	6	5.25	4	1.5	-6

   (a) Considerando apenas os argumentos -2, -1 e 0 determine a expressão do polinómio interpolador por aplicação da fórmula de Lagrange.

   (b) Se calcular o valor do polinómio interpolador para $x = 2$ obtém -7.25. Como interpreta que se obtenha o valor anterior sendo $f(x)$ descontínua no ponto $x = 2$.

2. Considere uma função $f : \mathbb{R} \to \mathbb{R}$ para a qual só se conhece os valores tabelados:

x	-2	-1	0	1
$f(x)$	12	4	-2	y

Sabendo que f é um polinómio de 2 grau:

(a) Indique qual o valor de y e calcule o coeficiente do termo de maior grau do polinómio interpolador.

(b) Considerando os 3 primeiros pontos, calcule o polinómio interpolador de Lagrange. O resultado seria o mesmo, caso considerasse os 4 pontos; porquê?

3. Dada a seguinte função tabelada:

x	0	2	4	6
$f(x)$	-5	1	25	55

(a) Utilize a fórmula interpoladora de Newton regressiva para determinar $f(5)$.

(b) Por inversão da tabela e utilizando somente os 3 primeiros pontos desta, determine o zero da função f.

4. Considere a seguinte tabela:

x	-3	-1	1	3	5	7	9
$f(x)$	-21	-4	-0.5	1	1.5	0.5	-1

(a) Escolhendo os pontos que em princípio minimizam o erro de interpolação, utilize a fórmula de Newton regressiva para obter uma aproximação quadrática de $f(4.5)$.

(b) Em face deste resultado qual o conjunto de 3 pontos é que escolheria para determinar, por interpolação inversa, um valor aproximado do zero de f localizado no intervalo $(1, 3)$. Justifique.

5. Demonstre que se $x_0, x_1, \ldots, x_n$, forem $(n + 1)$ nós distintos, tais que

$$x_j = x_0 + jh \qquad j = 0, 1, \ldots, n$$

e $f_0, f_1, \ldots, f_n$, os valores correspondentes da função f nesses nós, se obtem, para todo o par (i, j) tal que $0 \leq i < j \leq n$ e com $k = j - i$:

$$f[x_i, x_{i+1}, \ldots, x_j] = \Delta^k f_i/(k!h^k)$$

6. Considere a função tabelada

x	-2	0	2	4
$f(x)$	-17	5	-5	y

que se sabe representar o polinómio $p_3(x) = x^3 + a_2x^2 + a_1x + 5$.

(a) Que relação existe entre o polinómio interpolador nos 3 primeiros pontos e o polinómio p_3?

(b) Qual o valor da diferença f[-2,0,2,4]?

(c) Se $\Delta^2 f(0) = 16$, qual o valor de $f(4)$?

7. Considere a seguinte tabela:

x	1	3	5
$f(x)$	-1.5	1	2

(a) Obtenha o valor interpolado para $x = 4$, utilizando a fórmula de Newton regressiva.

(b) Determine por inversão da tabela o zero de f.

8. Considere a tabela

x	-3	-1	1	3
$f(x)$	-63	-1	1	β

representativa do polinómio $p_3(x) = 2.5x^3 + a_2x^2 + a_1x + a_0$. Calcule o valor de $\Delta^3 f(-3)$ e depois o valor de β.

9. Considere a função $f(x) = x^7 - 3x^4 + 1$ e o conjunto discreto de pontos

x	-2	-1	0	1	2
$f(x)$	-175	-3	1	-1	81

(a) Considerando os 3 primeiros pontos e utilizando a fórmula de Lagrange, obtenha o polinómio interpolador (não é necessário exprimi-lo numa soma de potências de x). Como justifica que o valor desse polinómio para $x = 2$ não seja igual a $f(2)$?

(b) Calcule por interpolação inversa a primeira raiz da função em $(-2, 2)$. Qual o conjunto de pontos que deve considerar? Justifique.

10. Considere a seguinte tabela da função de erro (*erf*).

x	$erf(x)$
0.0	0.000000
0.2	0.222703
0.4	0.428392
0.6	0.603856
0.8	0.742101
1.0	0.842701
1.2	0.910314
1.4	0.952285
1.6	0.976348
1.8	0.989091
2.0	0.995322

(a) Escolhendo os pontos que em princípio minimizam o erro de interpolação, utilize a fórmula de Newton progressiva para obter uma aproximação quadrática de *erf*(0.67).

(b) Sabe-se que

$$|\frac{d^{m+1}erf(x)}{dx^{m+1}}| \leq m^3$$

para $1 \leq m \leq 5$ e $0 \leq x \leq 2$.

i. Obtenha um majorante para o erro de interpolação cometido no cálculo anterior.

ii. Diga qual o espaçamento que a tabela deveria ter para que, para $\forall x \in [0, 2]$, o erro de interpolação linear seja sempre inferior a 10^{-4}.

11. Considere uma tabela de valores da função e^x. Diga qual o espaçamento h que a tabela deve ter para que, para $\forall x \in [0, 1]$, o erro de interpolação linear seja sempre inferior a 10^{-6}.

12. Considere a seguinte tabela

x	-2	-1	0	1	2
$f(x)$	6	2	0	α	0

(a) Utiliza a fórmula de Lagrange para obter o polinómio interpolador p_4. Qual o valor de α de modo a que p_4 se reduz a um polinómio de grau não superior a 3.

(b) Para $\alpha = 1.5$, calcule por interpolação inversa uma aproximação de $x \in (-2, 0)$ tal que $f(x) = 3$.

13. Verifique que a fórmula interpoladora de Lagrange, para abcissas igualmente espaçadas, se reduz a

$$p_n(x_0 + h\theta) = \frac{(-1)^n}{n!} \prod_{j=0}^{n} (\theta - j) \sum_{i=0}^{n} (-1)^i \begin{pmatrix} n \\ i \end{pmatrix} \frac{f_i}{\theta - i}$$

14. Considere a interpolação linear numa tabela de valores de $e^x, 0 \leq x \leq 2$, com $h = .01$. Majore o erro de interpolação. Supondo que se pretendesse um erro inferior a 10^{-5}, qual o valor que escolheria para h.

15. Por interpolação inversa, usando a Tabela 4.2, determine o valor $\overline{x}$ tal que $\sqrt{\overline{x}} = 1.01$. Para $n = 2$ majore o erro de interpolação e compare esta majoração com o erro atendendo a que $\overline{x} = 1.0201$.

Capítulo 5

Aproximação segundo os mínimos quadrados

5.1 Introdução

A aproximação de uma função f por uma *função aproximadora* g pode considerar-se em 2 situações distintas:

1. *Caso contínuo*: a função f é definida num certo intervalo $[a, b]$.
2. *Caso discreto*: só se conhecem $n+1$ pares de valores (x_i, f_i) com $i = 0, 1, \ldots, n$ onde f_i é o valor da função $f(x)$ para $x = x_i$ ou uma aproximação deste.

O *caso discreto* surge, p.ex., quando se pretende ajustar uma curva a um conjunto de pontos. Neste caso, f_i é geralmente só um valor aproximado de $f(x_i)$, o que acontece, p.ex., quando f_i é obtido experimentalmente ou provém de cálculos afectados por erros. O problema de aproximar uma função f no *caso contínuo* surge quando é necessário substituir esta função por outra que tenha determinadas características, p.ex., facilmente diferenciável, primitivável ou computável.

O *caso discreto* pode ser resolvido utilizando a interpolação polinomial que é das formas mais simples de aproximar uma função f. No entanto nem sempre é possível obter um polinómio interpolador p_n que constitua uma boa aproximação de f. Já vimos exemplos (nomeadamente o de Runge) em que quanto maior for o grau n do polinómio pior é a aproximação. Vimos também que os polinómios interpoladores não conduziam a bons resultados na *extrapolação*. Por outro lado, quando os valores conhecidos f_i são aproximações de $f(x_i)$, então a interpolação conduz geralmente a resultados desastrosos; o que facilmente se compreende visto que forçamos o polinómio interpolador a passar por pontos (x_i, f_i) que não pertencem à curva $f(x)$.

Dado um conjunto de pontos $(x_0, f_0), \ldots, (x_n, f_n)$, uma alternativa à interpolação polinomial consiste em obter um polinómio

$$p_m(x) = a_0 + a_1 x + \cdots + a_m x^m$$

de grau $m < n$, relativamente baixo, que segundo um certo critério se ajusta melhor àquele conjunto de pontos. Mais geralmente podemos considerar uma função

aproximadora

$$g(x) = F(x, a_0, a_1, \ldots, a_m)$$

que depende de certos parâmetros $a_0, a_1, \ldots, a_m$, como, p.ex., as funções:

$$g(x) = a_0 + \frac{a_1}{1+x} \qquad g(x) = a_0 + a_1 x^2$$

$$g(x) = \frac{a_0}{1+a_1 x} \qquad g(x) = a_0 + x^{a_1}$$

As 2 primeiras são exemplos de funções lineares nos parâmetros, enquanto que as últimas não são lineares nos parâmetros. Iremos considerar somente o caso em que $g(x)$ depende linearmente dos parâmetros. Assim temos

$$g(x) = \sum_{k=0}^{m} a_k \phi_k(x) \tag{5.1}$$

onde os $\{\phi_k\}$ são um conjunto de *funções de base* seleccionadas a priori e os $\{a_k\}$ são os *parâmetros* a determinar. Os $\{\phi_k(x)\}$ podem ser, p.ex., o conjunto dos monómios $\{x^k\}$ (caso do polinómio aproximador), o conjunto das funções trigonométricas $\{\sin kx\}$, o conjunto das funções exponenciais $\{\exp kx\}$, etc.

Os valores dos parâmetros $\{a_k\}$ são determinados de modo que os desvios

$$d_i = f_i - g_i \qquad i = 0, 1, \ldots, n$$

entre os dados f_i e os valores $g_i = g(x_i)$ da função aproximadora g sejam tão pequenos quanto possível, segundo um certo critério. Se impusermos como critério que

$$d_i = 0 \qquad i = 0, 1, \ldots, n$$

caimos no caso da *interpolação* com $m = n$. Designando respectivamente por d, f e g os vectores (de $\mathbb{R}^{n+1}$) de componentes d_i, f_i e g_i com $i = 0, \ldots, n$, e escolhendo em $\mathbb{R}^{n+1}$ uma norma $\| \ \|$ vectorial, temos o seguinte critério para obter os parâmetros $\{a_k\}$:

Seja f a função a aproximar, $\{\phi_k\}$ um dado conjunto de funções, $\| \ \|$ uma determinada norma e m um dado número natural; então dentro da classe de funções da forma

$$g = \sum_{k=0}^{m} a_k \phi_k$$

determina-se a que minimiza

$$\|d\| = \|f - g\|$$

Como exemplos de normas em $\mathbb{R}^{n+1}$ tem-se

$$\|f\|_p = (\sum_{i=0}^{n} |f_i|^p)^{1/p}, \quad 1 \le p < \infty$$

$$\|f\|_\infty = \lim_{p\to\infty} ||f||_p = \max_{0\le i\le n} |f_i|$$

Vamos escolher dentro das normas $|| \; ||_p$ a norma Euclidiana $\| \;\; \|_2$, por ser a que conduz à técnica mais simples e por outro lado porque no caso do ajustamento de uma função a um conjunto de dados experimentais cujos erros são normalmente distribuídos é a que conduz normalmente a uma melhor aproximação. Neste caso temos que minimizar

$$Q = \|d\|_2^2 = \sum_{i=0}^{n} [d_i]^2$$

Este critério é conhecido pelo dos *mínimos quadrados*. Como é sabido, a norma $\| \;\; \|_2$ é induzida de um produto interno e portanto podemos defini-la por

$$\|f\|_2 = (f, f)^{1/2}$$

em que $(\; , \;)$ é o produto interno

$$(f, g) = \sum_{i=0}^{n} f_i g_i \tag{5.2}$$

No *caso contínuo*, em que f está definida no intervalo $[a, b]$, para utilizar todos os valores de f no intervalo $[a, b]$ teremos que utilizar integrais em vez de somatórios. Tem-se então

$$\|f\|_p = (\int_a^b |f(x)|^p dx)^{1/p} \quad 1 \le p < \infty$$

$$\|f\|_\infty = \max_{x\in[a,b]} |f(x)|$$

Quando se pretende obter uma aproximação computacional da função f ou traçar o seu gráfico, escolhe-se, naturalmente, a norma do máximo (ou norma uniforme) $\| \; \|_\infty$ e tenta-se obter uma aproximação minimax da função. Dado que não é fácil obter esta aproximação, o que se faz é minimizar $|| \; ||_2$ que provém de um produto interno e que portanto é mais fácil de obter. Sendo assim, utilizamos o produto interno

$$(f, g) = \int_a^b f(x)g(x)dx \tag{5.3}$$

e a norma

$$\|f\|_2 = (\int_a^b [f(x)]^2 dx)^{1/2}$$

admitindo que f e g são funções reais de quadrado integrável em (a, b). Então o critério dos mínimos quadrados, para o *caso contínuo*, consiste em minimizar

$$Q = \|d\|_2^2 = \int_a^b [d(x)]^2 dx$$

em que a função

$$d(x) = f(x) - g(x)$$

é o desvio entre a função f e a função aproximadora g.

Por conseguinte, para obtermos uma aproximação segundo o critério dos mínimos quadrados, temos que minimizar

$$Q = \|f - g\|_2^2 = \|f - \sum_{k=0}^{m} a_k \phi_k\|_2^2 \tag{5.4}$$

que se reduz para o *caso contínuo* a

$$Q = \int_a^b [f(x) - g(x)]^2 dx \tag{5.5}$$

e para o *caso discreto* a

$$Q = \sum_{i=0}^{n} [f_i - g_i]^2 \tag{5.6}$$

5.2 Sistema de equações normais

Segundo o critério dos mínimos quadrados, a quantidade Q tem de ser minimizada, sendo a minimização sobre os parâmetros $a_0, a_1, \ldots, a_m$ da função aproximadora g. Com a finalidade de minimizar Q, consideremos Q como sendo função desses parâmetros $a_0, a_1, \ldots, a_m$, *i.e.*,

$$Q = Q(a_0, a_1, \ldots, a_m)$$

Então, para que Q atinja o seu mínimo, basta que se verifique a condição

$$\frac{\partial Q}{\partial a_j} = 0 \qquad j = 0, 1, \ldots, m$$

Notando que f é independente dos parâmetros a_j e atendendo a (5.5) e (5.6), esta condição reduz-se para o *caso contínuo* a

$$\frac{\partial}{\partial a_j} \int_a^b [f(x) - g(x)]^2 dx = 2 \int_a^b [f(x) - g(x)] \frac{\partial[-g(x)]}{\partial a_j} dx = 0 \qquad j = 0, 1, \ldots, m$$

e para o *caso discreto* a

$$\frac{\partial}{\partial a_j} \sum_{i=0}^{n} [f_i - g_i]^2 = 2 \sum_{i=0}^{n} [f_i - g_i] \frac{\partial[-g_i]}{\partial a_j} = 0 \qquad j = 0, 1, \ldots, m$$

Como estamos a considerar que $g(x)$ depende linearmente dos parâmetros, então de (5.1) temos

$$\frac{\partial g}{\partial a_j} = \phi_j$$

e portanto as condições anteriores reduzem-se respectivamente para o *caso contínuo*

$$\int_a^b [f(x) - g(x)] \phi_j(x) dx = 0 \qquad j = 0, 1, \ldots, m$$

e para o *caso discreto* a

$$\sum_{i=0}^{n}[f_i - g_i]\phi_j(x_i) = 0 \qquad j = 0, 1, \ldots, m$$

Estas condições podem escrever-se simplesmente na forma de um produto interno

$$(\phi_j, [f - g]) = 0 \qquad j = 0, 1, \ldots, m \tag{5.7}$$

ou ainda

$$(\phi_j, g) = (\phi_j, f) \qquad j = 0, 1, \ldots, m \tag{5.8}$$

em que no caso discreto ϕ_j é um vector de $\mathbb{R}^{n+1}$ de componentes $\phi_j(x_i)$ com $i = 0, \ldots, n$. Substituindo (5.1)

$$g(x) = \sum_{k=0}^{m} a_k \phi_k(x)$$

em (5.8) e atendendo à linearidade do produto interno, obtemos finalmente

$$\sum_{k=0}^{m} a_k(\phi_j, \phi_k) = (\phi_j, f) \qquad j = 0, 1, \ldots, m \tag{5.9}$$

que é um sistema de $m + 1$ equações lineares a $m + 1$ incógnitas $a_0, a_1, \ldots, a_m$. Este sistema (5.9) é conhecido por *sistema de equações normais*; que pode escrever-se na forma matricial

$$\begin{bmatrix} (\phi_0, \phi_0) & (\phi_0, \phi_1) & \cdots & (\phi_0, \phi_m) \\ (\phi_1, \phi_0) & (\phi_1, \phi_1) & \cdots & (\phi_1, \phi_m) \\ \cdots & \cdots & \cdots & \cdots \\ (\phi_m, \phi_0) & (\phi_m, \phi_1) & \cdots & (\phi_m, \phi_m) \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_m \end{bmatrix} = \begin{bmatrix} (\phi_0, f) \\ (\phi_1, f) \\ \vdots \\ (\phi_m, f) \end{bmatrix} \tag{5.10}$$

A designação de *normal* para este sistema deve-se à equivalência entre o sistema (5.10) e a condição (5.7) onde se exige que o desvio $f - g$ seja ortogonal a todos os vectores ϕ_j.

Teorema 5.1 *A matriz simétrica G presente em (5.10) é definida positiva sse os vectores ϕ_k forem linearmente independentes.*

Demonstração: Uma matriz G é definida positiva sse para qualquer vector não nulo $a = [a_0, a_1, \ldots, a_m]^T$ é verificada a desigualdade:

$$a^T G a > 0 \tag{5.11}$$

Notando que a componente de ordem j do vector Ga é

$$\sum_{k=0}^{m}(\phi_j, \phi_k)a_k = (\phi_j, \sum_{k=0}^{m} a_k \phi_k)$$

temos que

$$\begin{aligned} a^T G a &= \textstyle\sum_{j=0}^{m} a_j(\phi_j, \sum_{k=0}^{m} a_k \phi_k) \\ &= (\textstyle\sum_{j=0}^{m} a_j \phi_j, \sum_{k=0}^{m} a_k \phi_k) \\ &= \| \textstyle\sum_{k=0}^{m} a_k \phi_k \|_2^2 \geq 0 \end{aligned}$$

Como, para quaisquer $a_0, a_1, \ldots, a_m$ não simultaneamente todos nulos,

$$\|\sum_{k=0}^{m} a_k\phi_k\|_2 \neq 0$$

sse os ϕ_k forem linearmente independentes, então concluimos que a desigualdade (5.11) é verificada como pretendíamos. ■

O facto da matriz G do sistema normal (5.10) ser definida positiva garante-nos a existência de uma única solução a, para cada f não nulo, *i.e.*, a aproximação de uma função f segundo o critério dos mínimos quadrados, e para um determinado conjunto de funções de base ϕ_k, tem uma única solução g. Por outro lado, como a matriz é simétrica e definida positiva o método numérico aconselhável para resolver o sistema (5.10) é o de *Cholesky*.

A aproximação dos mínimos quadrados

$$g(x) = \sum_{k=0}^{m} a_k\phi_k(x)$$

que acabamos de obter pode enterpretar-se como a solução do seguinte problema de aproximação.

> Seja V um espaço euclidiano (espaço linear com produto interno) e S o subespaço de V de dimensão finita gerado por $\{\phi_0, \ldots, \phi_m\}$. Sendo $f \in V$ determinar dentro dos elementos de S o vector g mais próximo de f.

Como se sabe a solução (única) g deste problema é a projecção ortogonal de f sobre S, e portanto $f - g$ é ortogonal a S. É evidente que a condição (5.7) é necessária e suficiente para que $f - g$ seja ortogonal a S; a aproximação dos mínimos quadrados é a solução g deste problema de aproximação.

Se os ϕ_k forem linearmente independentes, então constituem uma base de S e a dimensão de S é $m + 1$. No caso discreto o espaço V é $\mathbb{R}^{n+1}$. No caso contínuo podemos tomar para V por exemplo o espaço das funções de quadrado integrável[1] em $[a, b]$ munido do produto interno (5.3).

Relembramos que o mínimo de Q definido por (5.4) é atingido quando as condições de ortogonalidade (5.7) forem satisfeitas. Destas condições temos que $(f - g, g) = 0$, condição que vamos utilizar para obter outra forma de Q. Temos que

$$Q = ||f - g||_2^2 = (f - g, f - g) = (f - g, f) - (f - g, g)$$

Da condição de ortogonalidade temos

$$Q = (f - g, f) = ||f||_2^2 - (g, f)$$

De (5.1), temos finalmente

$$Q = ||f||_2^2 - \sum_{k=0}^{m} a_k(\phi_k, f) \tag{5.12}$$

[1] Se $\int_a^b [f(x) - g(x)]^2\, dx = 0$ associa-se a f e a g o mesmo elemento de V

que, especialmente no caso contínuo, pode ser mais cómodo de utilizar para calcular Q do que

$$Q = ||f - g||_2^2 \quad g = \sum_{k=0}^{m} a_k \phi_k \tag{5.13}$$

5.2.1 Caso discreto

Neste caso os elementos do sistema de equações normais (5.10) são

$$(\phi_j, \phi_k) = \sum_{i=0}^{n} \phi_j(x_i)\phi_k(x_i) \qquad k = 0, 1, \ldots, m \quad j = 0, 1, \ldots, m$$

e

$$(\phi_j, f) = \sum_{i=0}^{n} \phi_j(x_i) f_i \qquad j = 0, 1, \ldots, m$$

Podemos escrever a matriz G do sistema normal como o produto

$$G = X^T X$$

onde

$$X = \begin{pmatrix} \phi_0(x_0) & \phi_1(x_0) & \cdots & \phi_m(x_0) \\ \phi_0(x_1) & \phi_1(x_1) & \cdots & \phi_m(x_1) \\ \cdots & \cdots & \cdots & \cdots \\ \phi_0(x_n) & \phi_1(x_n) & \cdots & \phi_m(x_n) \end{pmatrix}$$

é uma matriz $(n+1) \times (m+1)$ cujas colunas são os vectores ϕ_j. Denominando por a o vector de componentes $a_0, \ldots, a_m$ as equacões normais (5.10) podem ser expressas por

$$X^T X a = X^T f \tag{5.14}$$

em que $Xa = g$. Escrevendo (5.14) na forma $X^T(f - g) = 0$ obtem-se a condição (5.7) de ortogonalidade. Vemos de (5.14) que a solução dos mínimos quadrados discreto pode considerar-se como a solução dos mínimos quadrados do sistema sobredimensionado $(n > m)$ $Xa = f$, i.e., aquela que minimiza $||Xa - f||_2$.

O caso particular mais importante é quando a função aproximadora (ou de ajustamento) é um *polinómio*. Neste caso

$$\phi_j(x) = x^j$$

e portanto o sistema normal é da forma:

$$\begin{bmatrix} n+1 & \sum_{i=0}^n x_i & \cdots & \sum_{i=0}^n x_i^m \\ \sum_{i=0}^n x_i & \sum_{i=0}^n x_i^2 & \cdots & \sum_{i=0}^n x_i^{m+1} \\ \cdots & \cdots & \cdots & \cdots \\ \sum_{i=0}^n x_i^m & \sum_{i=0}^n x_i^{m+1} & \cdots & \sum_{i=0}^n x_i^{2m} \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_m \end{bmatrix} = \begin{bmatrix} \sum_{i=0}^n f_i \\ \sum_{i=0}^n x_i f_i \\ \vdots \\ \sum_{i=0}^n x_i^m f_i \end{bmatrix} \tag{5.15}$$

e a matriz X reduz-se a matriz de Vandermonde

$$X = \begin{pmatrix} 1 & x_0 & \cdots & x_0^m \\ 1 & x_1 & \cdots & x_1^m \\ \cdots & \cdots & \cdots & \cdots \\ 1 & x_n & \cdots & x_n^m \end{pmatrix}$$

Se o polinómio de ajustamento for linear então a aproximação é designada correntemente por *regressão linear*. Neste caso particular, $m = 1$, as equações normais têm a seguinte forma

$$a_0(n+1) + a_1 \sum_{i=0}^{n} x_i = \sum_{i=0}^{n} f_i$$

$$a_0 \sum_{i=0}^{n} x_i + a_1 \sum_{i=0}^{n} x_i^2 = \sum_{i=0}^{n} x_i f_i$$

Resolvendo este sistema, podemos exprimir a_0 e a_1 na forma:

$$a_1 = \frac{\sum_{i=0}^{n} f_i(x_i - \overline{x})}{\sum_{i=0}^{n}(x_i - \overline{x})^2} \qquad a_0 = \overline{f} - a_1\overline{x} \tag{5.16}$$

onde

$$\overline{x} = \frac{1}{n+1}\sum_{i=0}^{n} x_i \qquad \overline{f} = \frac{1}{n+1}\sum_{i=0}^{n} f_i \tag{5.17}$$

Assim no caso da regressão linear o polinómio de ajustamento tem a forma:

$$p_1(x) = \overline{f} + a_1(x - \overline{x}) \tag{5.18}$$

com $\overline{f}$, a_1 e $\overline{x}$ dados por (5.16) e (5.17).

Exemplo 5.1 *Seja $n = 4$, em que os pontos dados são definidos por:*

$$x_0 = -2 \quad x_1 = -1 \quad x_2 = 0 \quad x_3 = 1 \quad x_4 = 2$$

$$f_0 = 0 \quad f_1 = 1 \quad f_2 = 2 \quad f_3 = 1 \quad f_4 = 0$$

Obter um polinómio que se ajusta a estes pontos, segundo o critério dos mínimos quadrados.

Um gráfico destes pontos sugere a utilização de uma aproximação quadrática ($m = 2$). Para construir o sistema de equações normais (5.15) calculamos

$$\textstyle\sum_{i=0}^{4} x_i^0 = 5 \quad \sum_{i=0}^{4} x_i^1 = 0 \quad \sum_{i=0}^{4} x_i^2 = 10 \quad \sum_{i=0}^{4} x_i^3 = 0 \quad \sum_{i=0}^{4} x_i^4 = 34$$

$$\textstyle\sum_{i=0}^{4} x_i^0 f_i = 4 \quad \sum_{i=0}^{4} x_i^1 f_i = 0 \quad \sum_{i=0}^{4} x_i^2 f_i = 2$$

o que implica que o sistema seja o seguinte:

$$\begin{aligned} &5\,a_0 + 10\,a_2 = 4 \\ &10\,a_1 = 0 \\ &10\,a_0 + 34\,a_2 = 2 \end{aligned} \tag{5.19}$$

A solução deste sistema é

$$a_0 = 58/35 \qquad a_1 = 0 \qquad a_2 = -3/7$$

e o correspondente polinómio mínimos quadrados é

$$p_2(x) = -3/7x^2 + 58/35$$

Notamos que

$$\begin{array}{ll} g_2 = 1.65714286 & f_2 = 2 \\ g_1 = g_3 = 1.22857143 & f_1 = f_3 = 1 \\ g_0 = g_4 = -0.05714286 & f_0 = f_4 = 0 \end{array}$$

onde $g_i = p_2(x_i)$. Portanto fazendo uso de (5.6) obtemos para este exemplo

$$Q = 8/35 = 0.2285714$$

■

Em referência à aproximação por mínimos quadrados, um problema importante é a escolha do m, o grau do polinómio aproximador. Infelizmente, não existe uma teoria geral que nos resolva o problema. Às vezes uma inspecção visual do gráfico dos dados pode ser útil. Ralston e Rabinowitz dão um método de selecionar m que consiste em aumentar m até que

$$\frac{Q}{n-m} = \frac{\sum_{i=0}^{n}[f_i - p_m(x_i)]^2}{n-m}$$

deixa de aumentar.

Nem sempre o ajustamento polinomial é o mais adequado. Pode acontecer que dado um conjunto de pontos saibamos qual o tipo de função que se ajusta melhor a este.

Exemplo 5.2 *Para o conjunto de pontos considerado no Exemplo 5.1 obter a função do tipo*

$$a_0 \cos x + a_1$$

que se lhe ajusta melhor, segundo o critério dos mínimos quadrados.

A função aproximadora é

$$g(x) = a_0\phi_0(x) + a_1\phi_1(x)$$

com

$$\phi_0(x) = \cos x \qquad \phi_1(x) = 1$$

Assim, o sistema de equações normais é, para este caso:

$$\begin{bmatrix} \sum_{i=0}^{4} \cos^2 x_i & \sum_{i=0}^{4} \cos x_i \\ \sum_{i=0}^{4} \cos x_i & 5 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \end{bmatrix} = \begin{bmatrix} \sum_{i=0}^{4} \cos x_i f_i \\ \sum_{i=0}^{4} f_i \end{bmatrix}$$

Atendendo a que

$$\sum_{i=0}^{4} \cos x_i = 1.24831094 \qquad \sum_{i=0}^{4} \cos^2 x_i = 1.93020954$$

e

$$\sum_{i=0}^{4} f_i = 4 \qquad \sum_{i=0}^{4} \cos x_i f_i = 3.08060461$$

e resolvendo o sistema obtemos a seguinte aproximação

$$g(x) = 1.28630648 \ \cos x \ + \ 0.47885579$$

Notamos que

$$\begin{array}{ll} g_2 = 1.76516227 & f_2 = 2 \\ g_1 = g_3 = 1.17385015 & f_1 = f_3 = 1 \\ g_0 = g_4 = -0.05643658 & f_0 = f_4 = 0 \end{array}$$

Portanto temos para este exemplo (comparar com o resultado obtido no Exemplo 5.1)

$$Q = 0.1219667$$

■

5.2.2 Caso contínuo

Neste caso os elementos do sistema normal (5.10) são

$$(\phi_j, \phi_k) = \int_a^b \phi_j(x)\phi_k(x)\, dx \qquad k = 0, 1, \ldots, m \quad j = 0, 1, \ldots, m$$

e

$$(\phi_j, f) = \int_a^b \phi_j(x) f(x)\, dx \qquad j = 0, 1, \ldots, m$$

Para o caso particular

$$\phi_j(x) = x^j$$

com $a = 0$ e $b = 1$ estes elementos reduzem-se a

$$(\phi_j, \phi_k) = \int_0^1 x^{j+k}\, dx = \frac{1}{j+k+1} \qquad k = 0, 1, \ldots, m \quad j = 0, 1, \ldots, m$$

e

$$(\phi_j, f) = \int_0^1 x^j f(x)\, dx \qquad j = 0, 1, \ldots, m$$

e portanto o sistema de equações normais (5.10) toma a forma

$$\begin{bmatrix} 1 & \frac{1}{2} & \cdots & \frac{1}{m+1} \\ \frac{1}{2} & \frac{1}{3} & \cdots & \frac{1}{m+2} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{1}{m+1} & \frac{1}{m+2} & \cdots & \frac{1}{2m+1} \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_m \end{bmatrix} = \begin{bmatrix} \int_0^1 f(x)\, dx \\ \int_0^1 x f(x)\, dx \\ \vdots \\ \int_0^1 x^m f(x)\, dx \end{bmatrix} \tag{5.20}$$

A matriz H deste sistema é conhecida por matriz de *Hilbert de ordem* $m + 1$.

Exemplo 5.3 *Obter a aproximação quadrática de* $f(x) = \exp(x)$ *no intervalo* $[0, 1]$, *segundo o critério dos mínimos quadrados.*

Neste caso o sistema (5.20) reduz-se a

$$\begin{bmatrix} 1 & \frac{1}{2} & \frac{1}{3} \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{4} \\ \frac{1}{3} & \frac{1}{4} & \frac{1}{5} \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} \int_0^1 \exp(x)\, dx \\ \int_0^1 x \exp(x)\, dx \\ \int_0^1 x^2 \exp(x)\, dx \end{bmatrix} = \begin{bmatrix} e-1 \\ 1 \\ e-2 \end{bmatrix}$$

cuja solução é

$$\begin{aligned} a_0 &= 3(13e - 35) = 1.01299131 \\ a_1 &= 12(49 - 18e) = 0.85112505 \\ a_2 &= 30(7e - 19) = 0.83918398 \end{aligned}$$

Portanto o polinómio de ajustamento é

$$p_2(x) = 1.01299131 + 0.85112505x + 0.83918398x^2$$

Nota-se que este polinómio constitui um aproximação relativamente boa de $f(x) = \exp(x)$ em [0,1]. Por exemplo:

$$\begin{aligned} p_2(0.0) &= 1.01299131 & \exp(0.0) &= 1.00000000 \\ p_2(0.5) &= 1.64834983 & \exp(0.5) &= 1.64872127 \\ p_2(1.0) &= 2.70330034 & \exp(1.0) &= 2.71828183 \end{aligned}$$

■

5.2.3 Mau condicionamento do sistema normal

Quando o grau do polinómio aproximador, segundo o critério dos mínimos quadrados, toma valores moderados ou grandes somos geralmente conduzidos à resolução de um sistema de equações normais mal condicionado. Consequentemente o polinómio obtido numericamente difere substancialmente do polinómio que satisfaz o critério dos mínimos quadrados.

Para ilustrar o mau condicionamento que afecta a maioria das aproximações que utilizam o critério dos mínimos quadrados, consideremos o caso em que as abcissas dos pontos estão igualmente espaçadas no intervalo $[0, 1]$. Temos então,

$$x_i = i/n \qquad i = 0, 1, \ldots, n$$

Se pretendermos obter um ajustamento polinomial, então os elementos da matriz G do sistema de equações normais (5.15) são

$$(\phi_j, \phi_k) = \sum_{i=0}^{n} x_i^{j+k} = (n+1) \sum_{i=0}^{n} (\frac{i}{n})^{j+k} (\frac{1}{n+1}) \approx (n+1) \int_0^1 x^{j+k} dx = \frac{n+1}{j+k+1}$$

Utilizamos o facto de

$$\sum_{i=0}^{n} (\frac{i}{n})^{j+k} (\frac{1}{n+1})$$

ser uma soma de Riemann que aproxima o integral

$$\int_0^1 x^{j+k} dx$$

Como vimos este integral é um elemento da matriz do sistema (5.20), referido atrás quando consideramos a aproximação polinomial no caso contínuo no intervalo [0,1]. Portanto, concluimos que $G \approx (n+1)H$, onde H é a matriz de Hilbert de ordem $m+1$

$$H = \begin{bmatrix} 1 & \frac{1}{2} & \cdots & \frac{1}{m+1} \\ \frac{1}{2} & \frac{1}{3} & \cdots & \frac{1}{m+2} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{1}{m+1} & \frac{1}{m+2} & \cdots & \frac{1}{2m+1} \end{bmatrix}$$

Mesmo para m pequeno a matriz de Hilbert é muito mal condicionada. É de notar que na obtenção de um polinómio segundo o critério dos minímos quadrados o mau condicionamento depende só dos valores dos argumentos x_i que intervêm na matriz, e não dos valores de f_i.

Vemos assim que apesar das fórmulas para construir os polinómios mínimos quadrados serem fáceis de obter, elas tendem a ser numericamente instáveis. Mé - todos com funções ortogonais, discutidos a seguir, fornecem um remédio para estas dificuldades numéricas.

5.3 Problemas

1. Determine a parábola $g(x) = ax^2 + b$ que se ajusta melhor (no sentido dos mínimos quadrados) aos dados:

x	-1	0	1
$f(x)$	3.1	0.9	2.9

2. Considere a seguinte tabela:

x	1	3	5
$f(x)$	-1.5	1	2

Determine a função $g(x)$ da forma $a/(x+2) + b$ que se ajusta aos pontos da tabela utilizando o critério dos mínimos quadrados.

3. Considere a seguinte tabela da função de erro (*erf*).

x	$erf(x)$
0.8	0.742101
1.0	0.842701
1.2	0.910314

Verifique que a função $g(x)$ da forma $(a/x) + bx$ que melhor se ajusta aos pontos da tabela, no sentido dos mínimos quadrados, é caracterizada por $a = 0.19677$, $b = 0.62935$.

4. Considere a seguinte tabela:

x	-3	-1	1
$f(x)$	-21	-4	-0.5

Determine a função $g(x)$ da forma $a/(x+4)+b$ que se ajusta aos pontos da tabela utilizando o critério dos mínimos quadrados.

5. Considere a função $f(x) = x^7 - 3x^4 + 1$ e o conjunto discreto de pontos

x	-2	-1	0	1	2
$f(x)$	-175	-3	1	-1	81

(a) Em relação ao conjunto discreto de pontos, pretende-se calcular a função p_2 de ajustamento quadrática. Mostra que o sistema de equações normais a que é conduzido por aplicação do método dos mínimos quadrados é:

$$\begin{bmatrix} 5 & 0 & 10 \\ 0 & 10 & 0 \\ 10 & 0 & 34 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} -97 \\ 514 \\ -380 \end{bmatrix}$$

(b) Qual o sistema de equações normais a que é conduzido por ajustamento quadrático à função f, no intervalo $[-2, 2]$?

6. Considere a tabela

x	-2	-1	0	1	2
$f(x)$	1	2	-1	0	1

Determine a função $g(x)$ da forma $a\sin(\frac{\pi}{2}x)+b\cos(\frac{\pi}{2}x)$ que se ajusta aos pontos da tabela utilizando o critério dos mínimos quadrados.

7. Considere a seguinte tabela

x	-2	-1	0	1	2
$f(x)$	6	2	0	-1	0

Determine a função $g(x)$ da forma ax^2+bx que se ajusta aos pontos da tabela, utilizando o critério dos mínimos quadrados.

8. Considere a tabela

x	-5	-3	-1	1	3	5
$f(x)$	1	0	-1	2	1	2

Determine a função $g(x)$ da forma $a/x+bx^2$ que se ajusta aos pontos da tabela com abcissas -3, -1, 1 e 3 utilizando o critério dos mínimos quadrados.

9. Considere a tabela

x	-3	-1	1
$f(x)$	-63	-1	1

Determine a função $g(x)$ da forma $a/(x+4)+bx^2$ que se ajusta aos pontos da tabela utilizando o critério dos mínimos quadrados.

Capítulo 6

Integração numérica

6.1 Introdução

O cálculo de integrais aparece constantemente na resolução dos mais variados problemas. Na maioria dos casos, não podem ser determinados explicitamente por fórmulas simples. Somos conduzidos, então, à resolução numérica. Os principais métodos de integração baseiam-se na substituição da função por polinómios interpoladores e na respectiva integração destes. Neste capítulo deduzem-se as principais regras de integração numérica. Consideram-se sucessivamente fórmulas com pontos (1) sem restrições (obtidos p. ex. experimentalmente), (2) igualmente espaçados (fórmulas de Newton-Cotes), e (3) coincidentes com os zeros de polinómios ortogonais (fórmulas de Gauss), deduzindo-se o erro de truncatura associado a algumas destas fórmulas. Realçam-se algumas dificuldades que surgem na utilização das fórmulas obtidas e tenta-se contorná-las através do uso de regras compostas, incluíndo as adaptativas.

Consideremos o integral definido

$$I(f) = \int_a^b f(x)\, dx \tag{6.1}$$

com $[a, b]$ finito. Neste capítulo, pretendemos desenvolver alguns métodos numéricos para calcular (6.1). Os métodos que estudaremos, consistem em utilizar *fórmulas de integração*, também designadas por *fórmulas de quadratura*, com a forma geral:

$$I_n(f) = \sum_{i=0}^{n} A_i f(x_i) \tag{6.2}$$

em que $I_n(f)$ constitui uma aproximação de $I(f)$. Na generalidade dos casos, os pontos de integração $x_0, x_1, \ldots, x_n$, para os quais é calculada a função integranda, estão localizados no intervalo $[a, b]$. Os pesos $A_0, A_1, \ldots, A_n$ são determinados, independentemente da função a integrar, segundo um certo critério de modo a que $I_n(f)$ constitui em geral uma boa aproximação de $I(f)$.

O critério mais simples é impor que a fórmula de integração $I_n(f)$ conduz ao valor exacto do integral $I(f)$ quando f for um polinómio q_n de grau inferior ou igual

a n, *i.e.*,

$$I_n(q_n) = I(q_n)$$

Quando uma fórmula de integração satisfaz a esta condição e existe um polinómio q_{n+1} de grau $n+1$ tal que

$$I_n(q_{n+1}) \neq I(q_{n+1})$$

diz-se que a *fórmula é de grau* n.

Dado o conjunto de pontos $x_0, x_1, \ldots, x_n$, para obter uma fórmula de grau não inferior a n, começa-se por considerar o polinómio interpolador p_n de grau n que interpola a função f nestes pontos. Como vimos, p_n pode obter-se utilizando a fórmula interpoladora de Lagrange

$$p_n(x) = \sum_{i=0}^{n} f(x_i) l_i(x) \tag{6.3}$$

em que

$$l_i(x) = \prod_{j=0, j\neq i}^{n} (x - x_j) / \prod_{j=0, j\neq i}^{n} (x_i - x_j)$$

são os polinómios de Lagrange. Como

$$p(x_i) = f(x_i) \qquad i = 0, \ldots, n$$

temos que

$$I_n(f) = I_n(p_n)$$

Além disso, como pretendemos que seja

$$I_n(p_n) = I(p_n)$$

vemos que é necessário que

$$I_n(f) = I(p_n)$$

Então para obter uma fórmula de integração de f, basta integrar exactamente o polinómio interpolador (6.3) entre a e b obtendo

$$I_n(f) = \int_a^b \sum_{i=0}^{n} f(x_i) l_i(x)\, dx = \sum_{i=0}^{n} f(x_i) \int_a^b l_i(x)\, dx$$

Somos assim conduzidos a uma fórmula de integração do tipo dado por (6.2) em que os pesos são dados por

$$A_i = \int_a^b l_i(x)\, dx \tag{6.4}$$

É obvio que esta fórmula assim obtida é de grau não inferior a n, visto que se f for um polinómio q_n de grau inferior ou igual a n, o seu polinómio interpolador p_n coincide com ele próprio e temos portanto que

$$I_n(q_n) = I(q_n)$$

como pretendíamos.

Exemplo 6.1 *Obter uma fórmula de integração numérica para o integral* $\int_{-1}^{1} f(x)\,dx$ *que utiliza* $f(-2)$, $f(0)$ *e* $f(2)$, *recorrendo aos polinómios de Lagrange.*

Começamos por calcular

$$\int_{-1}^{1}(x-x_r)(x-x_s)\,dx = \int_{-1}^{1}[x^2-(x_r+x_s)x+x_rx_s]\,dx = 2(1/3+x_rx_s)$$

Utilizando este resultado em (6.4) obtemos

$$A_0 = \int_{-1}^{1}\frac{(x-0)(x-2)}{(-2-0)(-2-2)}\,dx = \frac{2(1/3)}{8} = 1/12$$

$$A_1 = \int_{-1}^{1}\frac{(x+2)(x-2)}{(0+2)(0-2)}\,dx = \frac{2(1/3-4)}{-4} = 11/6$$

$$A_2 = \int_{-1}^{1}\frac{(x+2)(x-0)}{(2+2)(2-0)}\,dx = \frac{2(1/3)}{8} = 1/12$$

Substituindo estes valores em (6.2), obtemos a fórmula pretendida

$$I_2(f) = \frac{1}{12}f(-2)+\frac{11}{6}f(0)+\frac{1}{12}f(2)$$

Notamos que esta fórmula é de grau 3. Com efeito,

$$I_2(x^3) = I(x^3) = 0$$

e por construção a fórmula é pelo menos de grau 2. Como um polinómio de grau 3 pode sempre exprimir-se na combinação linear de x^3 e de um polinómio de grau 2 concluimos que a fórmula é de grau 3 visto que

$$I_2(x^4) \neq I(x^4)$$

■

6.1.1 Método dos coeficientes indeterminados

Vamos apresentar um método alternativo para determinar os pesos para uma fórmula de grau não inferior a n. Este método é conhecido por *método dos coeficientes indeterminados* e baseia-se no facto de uma fórmula de grau n ser exacta para qualquer polinómio de grau $k \leq n$, em particular para

$$x^k \qquad k = 0, 1, \ldots, n$$

Como I e I_n são lineares e qualquer polinómio pode exprimir-se numa combinação linear de x^k, concluimos que para que a fórmula (6.2) seja de grau não inferior a n é *necessário* e *suficiente* que os pesos satisfaçam ao sistema:

$$\sum_{i=0}^{n} A_i x_i^k = \int_a^b x^k\,dx \quad , \qquad k = 0, 1, \ldots, n \tag{6.5}$$

Podemos escrever este sistema na forma matricial:

$$\begin{bmatrix} 1 & 1 & 1 & \cdot & \cdot & \cdot & 1 \\ x_0 & x_1 & x_2 & \cdot & \cdot & \cdot & x_n \\ x_0^2 & x_1^2 & x_2^2 & \cdot & \cdot & \cdot & x_n^2 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ x_0^n & x_1^n & x_2^n & \cdot & \cdot & \cdot & x_n^n \end{bmatrix} \begin{bmatrix} A_0 \\ A_1 \\ A_2 \\ \cdot \\ \cdot \\ A_n \end{bmatrix} = \begin{bmatrix} (b-a)/1 \\ (b^2-a^2)/2 \\ (b^3-a^3)/3 \\ \cdot \\ \cdot \\ (b^{n+1}-a^{n+1})/(n+1) \end{bmatrix} \tag{6.6}$$

Como a matriz deste sistema é invertível (desde que todos os x_i sejam distintos) o vector solução $(A_0, A_1, \ldots, A_n)^T$ existe e é único.

Exemplo 6.2 *Obter a fórmula de integração numérica para o integral $\int_{-1}^{1} f(x)\, dx$ que utiliza $f(-2)$, $f(0)$ e $f(2)$, recorrendo ao método dos coeficientes indeterminados.*

O sistema a resolver é:

$$\begin{bmatrix} 1 & 1 & 1 \\ -2 & 0 & 2 \\ (-2)^2 & 0 & (2)^2 \end{bmatrix} \begin{bmatrix} A_0 \\ A_1 \\ A_2 \end{bmatrix} = \begin{bmatrix} 1-(-1) \\ \left[(1)^2-(-1)^2\right]/2 \\ \left[(1)^3-(-1)^3\right]/3 \end{bmatrix} = \begin{bmatrix} 2 \\ 0 \\ 2/3 \end{bmatrix}$$

Cuja solução é

$$\begin{bmatrix} A_0 \\ A_1 \\ A_2 \end{bmatrix} = \begin{bmatrix} 1/12 \\ 11/6 \\ 1/12 \end{bmatrix}$$

que, obviamente, coincide com a obtida no Exemplo 6.1. ■

6.2 Fórmulas de Newton-Cotes

As fórmulas de integração de grau não inferior a n com pontos $x_i = x_0 + ih$, $i = 0, 1, \ldots, n$ igualmente espaçados são conhecidas por *fórmulas de Newton-Cotes*. Quando os pontos x_0 e x_n coincidirem respectivamente com a e b, extremos do intervalo de integração, então as fórmulas dizem-se *fechadas*; no caso dos pontos x_i localizarem-se todos no intervalo aberto (a, b) então elas são designadas por *abertas*. As duas fórmulas de Newton-Cotes fechadas mais populares são as que correspondem a tomar $n = 1$ e $n = 2$ e são conhecidas respectivamente por *regra dos trapézios* e por *regra de Simpson*, que passamos a estudar de seguida.

6.2.1 Regra dos trapézios

Para obter a regra dos trapézios basta utilizar (6.2) e (6.4) ou (6.6) com $n = 1$, $x_0 = a$ e $x_1 = b$. Assim atendendo que

$$A_0 = \int_a^b \frac{x-b}{a-b}\, dx = \frac{\frac{1}{2}x^2 - bx|_a^b}{a-b} = \frac{b-a}{2}$$

e

$$A_1 = \int_a^b \frac{x-a}{b-a}\, dx = \frac{\frac{1}{2}x^2 - ax|_a^b}{b-a} = \frac{b-a}{2}$$

temos

$$I_1(f) = (\frac{b-a}{2})[f(a) + f(b)] \tag{6.7}$$

Imediatamente, vemos que (6.7) corresponde a área do trapézio representado na Figura 6.1.

Figura 6.1: Regra dos trapézios

Para obter uma expressão do erro de integração vamos utilizar a fórmula do erro de interpolação

$$f(x) - p_1(x) = f[a, b, x](x-a)(x-b)$$

em que $p_1(x)$ é o polinómio que interpola $f(x)$ nos pontos de integração a e b. Temos então para o erro de integração

$$E_1(f) = I(f) - I_1(f) = I(f) - I(p_1) = \int_a^b f[a, b, x](x-a)(x-b)\, dx$$

Notando que $(x-a)(x-b)$ não muda de sinal no intervalo de integração $[a, b]$, podemos utilizar o teorema do valor intermédio para integrais. Supondo que
$f \in C^2([a, b])$, temos assim

$$\begin{aligned} E_1(f) &= f[a, b, \eta]\int_a^b (x-a)(x-b)\, dx \qquad \eta \in (a, b) \\ &= \left[\frac{1}{2}f''(\xi)\right]\left[-\frac{1}{6}(b-a)^3\right] \qquad \xi \in (a, b) \end{aligned}$$

em que na última igualdade se atendeu a relação entre diferenças divididas e derivadas. Então

$$E_1(f) = -\frac{(b-a)^3}{12} f''(\xi) \qquad \xi \in (a, b)$$

Se $b-a$ não for suficientemente pequeno, a regra dos trapézios (6.7) conduz a uma aproximação grosseira do integral $I(f)$. Por isso, passamos a subdividir o intervalo $[a,b]$ em n subintervalos $[x_{i-1},x_i]$, $i=1,\ldots,n$ e por conseguinte exprimir $I(f)$ numa soma de integrais

$$I(f)=\int_a^b f(x)\,dx=\sum_{i=1}^n\int_{x_{i-1}}^{x_i} f(x)\,dx$$

Se aplicarmos (6.7) a cada um deste integrais e designarmos $h_i=x_i-x_{i-1}$ obtemos

$$I_n(f)=\sum_{i=1}^n(\frac{h_i}{2})[f(x_{i-1})+f(x_i)]$$

e

$$E_n(f)=I(f)-I_n(f)=-\sum_{i=1}^n\frac{h_i^3}{12}f''(\xi_i)\qquad \xi_i\in(x_{i-1},x_i)$$

Se escolhermos os pontos x_i igualmente espaçados, temos que $h_i=h$ e $x_i=a+hi$ sendo $h=(b-a)/n$ o espaçamento. Obtemos assim a *regra dos trapézios composta:*

$$I_n(f)=h\left[\frac{f(a)+f(b)}{2}+\sum_{i=1}^{n-1}f(x_i)\right]$$

que designaremos simplesmente por regra dos trapézios. O erro de integração é

$$E_n(f)=-\sum_{i=1}^n\frac{h^3}{12}f''(\xi_i)=-\frac{h^3n}{12}\left[\frac{1}{n}\sum_{i=1}^n f''(\xi_i)\right]$$

Para o termo entre parênteses temos

$$\min_{a\le x\le b}f''(x)\le S=\frac{1}{n}\sum_{i=1}^n f''(\xi_i)\le\max_{a\le x\le b}f''(x)$$

Visto que, por hipótese, $f''(x)$ é contínua em $[a,b]$, existe um $\xi\in(a,b)$ tal que $f''(\xi)=S$. Portanto podemos escrever

$$E_n(f)=-\frac{(b-a)h^2}{12}f''(\xi)\qquad \xi\in(a,b) \tag{6.8}$$

Exemplo 6.3 *Calcular*

$$I=\int_0^\pi e^x\cos x\,dx \tag{6.9}$$

utilizando a regra dos trapézios. O valor correcto é $I=-12.07034631639$.

Designando por f a função integranda, temos para $n=2$ que $h=\pi/2$ e portanto

$$I_2=h[\frac{f(0)+f(\pi)}{2}+f(\pi/2)]=-17.389259$$

Tabela 6.1: Ilustração da regra dos trapézios

n	I_n	E_n	$E_{n/2}/E_n$
2	-17.389259	$5.32E-0$	
4	-13.336023	$1.27E-0$	4.20
8	-12.382162	$3.12E-1$	4.06
16	-12.148004	$7.77E-2$	4.02
32	-12.089742	$1.94E-2$	4.00
64	-12.075194	$4.85E-3$	4.00

Para $n = 4$ temos $h = \pi/4$ e

$$\begin{aligned} I_4 &= h[\frac{f(0)+f(\pi)}{2} + f(\pi/4) + f(\pi/2) + f(3\pi/4)] \\ &= \frac{1}{2}I_2 + h[f(\pi/4) + f(3\pi/4)] = -13.336023 \end{aligned}$$

Para $n = 8$ temos $h = \pi/8$ e

$$I_8 = \frac{1}{2}I_4 + h[f(\pi/8) + f(3\pi/8) + f(5\pi/8) + f(7\pi/8)] = -12.382162$$

Prosseguindo desta forma podemos obter os restantes valores de I_n representados na Tabela 6.1. Desta tabela concluimos que o erro decresce aproximadamente de um factor 4 quando o número n de subintervalos é duplicado. Este facto era previsível pelo aparecimento do factor $h^2 = (b-a)^2/n^2$ na forma do erro (6.8). Este exemplo evidencia a necessidade do cálculo da função num número relativamente grande de pontos para obter uma precisão suficiente. Apresentaremos neste capítulo outras regras de integração, para as quais geralmente o número de cálculos é mais reduzido do que para a regra dos trapézios. ■

6.2.2 Regra de Simpson

Para obter a regra de Simpson utilizamos, tal como para a regra dos trapézios, (6.2) e (6.4) ou (6.6) agora com $n = 2$, $x_0 = a$, $x_1 = c = a + h$ e $x_2 = b$ em que $h = (b-a)/2$. Temos então a regra de Simpson

$$I_2(f) = \frac{h}{3}\left[f(a) + 4f(a+h) + f(b)\right]$$

A partir do erro de interpolação obtemos o erro de integração

$$\begin{aligned} E_2(f) &= I(f) - I_2(f) \\ &= \int_a^b f[a,b,c,x](x-a)(x-c)(x-b)\,dx \end{aligned}$$

Como o polinómio $(x-a)(x-c)(x-b)$ muda de sinal em $x = c$ não podemos utilizar directamente o teorema do valor intermédio para integrais. Para podermos utilizar

este teorema vamos recorrer primeiro à fórmula interpoladora de Newton. Assim considerando um novo ponto de interpolação arbitrário d, podemos escrever

$$\begin{aligned} E_2(f) &= \int_a^b f[a,b,c,d](x-a)(x-c)(x-b)\,dx \\ &+ \int_a^b f[a,b,c,d,x](x-a)(x-c)(x-b)(x-d)\,dx \end{aligned}$$

Notando que $f[a,b,c,d]$ é independente da variável de integração x e que

$$\int_a^b (x-a)(x-c)(x-b)\,dx = 0 \tag{6.10}$$

obtemos

$$E_2(f) = \int_a^b f[a,b,c,d,x](x-a)(x-c)(x-b)(x-d)\,dx$$

Para aplicar o teorema do valor intermédio para integrais basta fazer $d = c$. Supondo que $f \in C^4([a,b])$, temos então[1]

$$\begin{aligned} E_2(f) &= \int_a^b f[a,b,c,c,x](x-a)(x-c)^2(x-b)\,dx \\ &= f[a,b,c,c,\eta]\int_a^b (x-a)(x-c)^2(x-b)\,dx \qquad \eta \in (a,b) \\ &= \frac{f^{(4)}(\xi)}{4!}\left[-\frac{4}{15}h^5\right] \qquad \xi \in (a,b) \end{aligned}$$

Portanto

$$E_2(f) = -\frac{h^5}{90} f^{(4)}(\xi) \qquad \xi \in (a,b) \tag{6.11}$$

De (6.11) vemos que $E_2(f) = 0$ se f for um polinómio de grau menor ou igual a 3 e portanto a regra de Simpson é de grau 3, apesar de ter sido deduzida a partir de um polinómio interpolador de grau 2. Este facto deve-se a (6.10).

Podemos novamente construir uma regra composta. Para $n \geq 2$ e par, definimos $h = (b-a)/n$, $x_i = a + ih$, $f_i = f(x_i)$, $i = 0, \ldots, n$. Então

$$I(f) = \int_a^b f(x)\,dx = \sum_{i=1}^{n/2} \int_{x_{2i-2}}^{x_{2i}} f(x)\,dx$$

$$I_n(f) = \sum_{i=1}^{n/2} \frac{h}{3}(f_{2i-2} + 4f_{2i-1} + f_{2i})$$

$$E_n(f) = I(f) - I_n(f) = -\sum_{i=1}^{n/2} \frac{h^5}{90} f^{(4)}(\xi_i) \qquad \xi_i \in (x_{2i-2}, x_{2i})$$

[1] A diferença dividida $f[a,b,c,c,x]$ pode-se definir como sendo o $\lim_{d\to c} f[a,b,c,d,x]$.

Simplificando, obtemos a *regra de Simpson composta*:

$$\begin{aligned} I_n(f) &= \frac{h}{3}[f_0 + 4f_1 + 2f_2 + 4f_3 + 2f_4 + \cdots + 2f_{n-2} + 4f_{n-1} + f_n] \\ &= \frac{h}{3}\left[(f_0 + f_n) + 4\sum_{i=1}^{n/2} f_{2i-1} + 2\sum_{i=1}^{n/2-1} f_{2i}\right] \end{aligned}$$

Para o erro, tal como para a regra dos trapézios, temos

$$E_n(f) = -\frac{h^5(n/2)}{90}\left[\frac{2}{n}\sum_{i=1}^{n/2} f^{(4)}(\xi_i)\right]$$

$$E_n(f) = -\frac{h^4(b-a)}{180} f^{(4)}(\xi) \qquad \xi \in (a,b) \tag{6.12}$$

Exemplo 6.4 *Utilizar a regra de Simpson para calcular o integral (6.9)*

$$I = \int_0^\pi e^x \cos x \, dx$$

considerado no Exemplo 6.3.

Temos para $n = 2$ que $h = \pi/2$ e portanto

$$I_2 = \frac{h}{3}[f(0) + f(\pi) + 4f(\pi/2)] = \frac{h}{3}[2P_2 + 4Q_2] = -11.592840$$

com

$$P_2 = \frac{f(0) + f(\pi)}{2} \qquad Q_2 = f(\pi/2)$$

Para $n = 4$ temos $h = \pi/4$ e

$$\begin{aligned} I_4 &= \frac{h}{3}[f(0) + f(\pi) + 2f(\pi/2) + 4f(\pi/4) + 4f(3\pi/4)] \\ &= \frac{h}{3}[2P_4 + 4Q_4] = -11.984944 \end{aligned}$$

com

$$P_4 = P_2 + Q_2 \qquad Q_4 = f(\pi/4) + f(3\pi/4)$$

Para $n = 8$ temos $h = \pi/8$ e

$$I_8 = \frac{h}{3}[2P_8 + 4Q_8] = -12.064209$$

com

$$P_8 = P_4 + Q_4 \qquad Q_8 = f(\pi/8) + f(3\pi/8) + f(5\pi/8) + f(7\pi/8)$$

Prosseguindo desta forma podemos obter os restantes valores de I_n representados na Tabela 6.2. Desta tabela concluimos que o erro decresce aproximadamente de um factor 16 quando o número n de subintervalos é duplicado. Este facto era previsível pelo aparecimento do factor $h^4 = (b-a)^4/n^4$ na forma do erro (6.12). Comparando os resultados com os apresentados na Tabela 6.1, vemos que a regra de Simpson é computacionalmente melhor do que a dos trapézios. ■

Tabela 6.2: Ilustração da regra de Simpson

n	I_n	E_n	$E_{n/2}/E_n$
2	-11.592840	$-4.78E-1$	
4	-11.984944	$-8.54E-2$	5.6
8	-12.064209	$-6.14E-3$	14.9
16	-12.069951	$-3.95E-4$	15.5
32	-12.070321	$-2.49E-5$	15.9
64	-12.070345	$-1.56E-6$	16.0

6.2.3 Fórmulas de graus superiores

Acabamos de estudar as 2 fórmulas fechadas de Newton-Cotes mais simples. Em geral para obter fórmulas fechadas de Newton-Cotes basta utilizar (6.2) e (6.4) ou (6.6) com $x_i = a + ih$, $i = 0, \ldots, n$ em que $h = (b-a)/n$. Como vimos, com $n = 1$ e $n = 2$ obtem-se respectivamente a regra dos trapézios e a regra de Simpson. Para $n = 3$ obtemos a *regra dos três oitavos:*

$$I_3(f) = \frac{3h}{8}\left[f(a) + 3f(a+h) + 3f(b-h) + f(b)\right]$$

e o respectivo erro

$$E_3(f) = -\frac{3h^5}{80} f^{(4)}(\xi) \qquad \xi \in (a,b)$$

e para $n = 4$ obtemos a seguinte fórmula de integração

$$I_4(f) = \frac{2h}{45}\left[7f(a) + 32f(a+h) + 12f(\frac{a+b}{2}) + 32f(b-h) + 7f(b)\right]$$

e o respectivo erro

$$E_4(f) = -\frac{8h^7}{945} f^{(6)}(\xi) \qquad \xi \in (a,b)$$

Para as fórmulas compostas, definindo $h = (b-a)/n$, $x_i = a + ih$, $f_i = f(x_i)$, $i = 0, \ldots, n$, temos respectivamente

$$I_n(f) = \frac{3h}{8}\left[(f_0 + f_n) + 3\sum_{i=1}^{n/3} f_{3i-2} + 3\sum_{i=1}^{n/3} f_{3i-1} + 2\sum_{i=1}^{n/3-1} f_{3i}\right]$$

$$E_n(f) = -\frac{h^4(b-a)}{80} f^{(4)}(\xi) \qquad \xi \in (a,b)$$

e

$$I_n(f) = \frac{2h}{45}\left[7(f_0 + f_n) + 32\sum_{i=1}^{n/4} f_{4i-3} + 12\sum_{i=1}^{n/4} f_{4i-2} + 32\sum_{i=1}^{n/4} f_{4i-1} + 14\sum_{i=1}^{n/4-1} f_{4i}\right]$$

$$E_n(f) = -\frac{2h^6(b-a)}{945} f^{(6)}(\xi) \qquad \xi \in (a,b)$$

Vemos que a fórmula de integração obtida fazendo $n = 3$ é só de grau $n = 3$ enquanto que a fórmula de integração obtida fazendo $n = 4$ é de grau $n + 1 = 5$. Pode-se demonstrar que a fórmula de Newton-Cotes para n ímpar (caso da regra dos trapézios) é de grau n enquanto que para n par (caso da regra de Simpson) é de grau $n + 1$.

Dado um intervalo $[a, b]$ e o conjunto de pontos $x_i = a + ih$, $i = 0, \ldots, n$, igualmente espaçados, vimos que existe funções indefinidamente diferenciáveis para as quais a sucessão p_n de polinómios que interpolem naqueles pontos uma destas funções não converge para a função. É de esperar que algo de semelhante acontece com as sucessivas fórmulas de Newton-Cotes $I_n(f)$ baseadas em polinómios interpoladores p_n. De facto utilizando as sucessivas fórmulas $I_n(f)$ para calcular

$$I = \int_{-4}^{4} \frac{1}{1 + x^2}\, dx$$

verifique-se que I_n não converge para I quando n tende para infinito. Por essa razão não se utiliza fórmulas de Newton-Cotes de grau elevado, preferindo-se recorrer às fórmulas compostas correspondentes às fórmulas de grau mais baixo.

Das fórmulas *abertas* de Newton-Cotes, só vamos apresentar a *regra do ponto médio* que é baseada no polinómio de grau 0 que interpola a função integranda a meio do intervalo. Temos assim

$$I_1 = (b - a) f\left(\frac{a + b}{2}\right) \tag{6.13}$$

$$E_1 = \frac{(b - a)^3}{24} f''(\xi) \qquad \xi \in (a, b) \tag{6.14}$$

Para a fórmula composta, fazendo $h = (b - a)/n$, temos[2]

$$I_n(f) = h \sum_{i=1}^{n} f(a + (i - 1/2)h)$$

$$E_n(f) = \frac{h^2(b - a)}{24} f''(\xi) \qquad \xi \in (a, b)$$

6.3 Fórmulas de Gauss

Para obter as fórmulas de Newton-Cotes impusemos pontos de integração igualmente espaçados e em seguida, utilizando (6.4) ou (6.6), fomos obter os pesos. Suponhamos agora que, não escolhendo a priori os pontos de integração, pretendemos obter os pontos x_i e os pesos A_i de modo que a fórmula de integração

$$I_n(f) = \sum_{i=1}^{n} A_i f(x_i) \tag{6.15}$$

seja do maior grau possível. Uma fórmula assim obtida é designada por fórmula de Gauss.

[2] O número de pontos de integração passa a ser designado por n e não por $n + 1$ como até agora.

Iremos considerar só integrais com intervalo de integração finito $[a, b]$. Como neste caso, um integral é sempre redutível, por meio de uma mudança de variável, a

$$\int_a^b g(t)\,dt = (\frac{b-a}{2})\int_{-1}^1 g(\frac{a+b+(b-a)x}{2})\,dx = \int_{-1}^1 f(x)\,dx \tag{6.16}$$

só vamos obter fórmulas de integração de Gauss para calcular integrais da forma

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

Para que a fórmula $I_n(f)$ seja de grau m, o erro

$$E_n(f) = \int_{-1}^1 f(x)\,dx - \sum_{i=1}^n A_i f(x_i)$$

tem de ser nulo para $f(x) = x^k$, $k = 0, 1, \ldots, m$. Temos assim o seguinte sistema não linear

$$\sum_{i=1}^n A_i x_i^k = \int_{-1}^1 x^k\,dx = \frac{1-(-1)^{k+1}}{k+1} \qquad k = 0, 1, \ldots, m \tag{6.17}$$

Como temos $2n$ incógnitas ($x_1, \ldots, x_n$ e $A_1, \ldots, A_n$) o sistema deverá ter $2n$ equações e portanto $m = 2n - 1$. Mais tarde (Teorema 6.1), demonstraremos que uma fórmula de integração da forma (6.15) não pode ser de grau superior a $2n - 1$.

Para $n = 1$ o sistema (6.17) reduz-se a

$$\begin{aligned} A_1 &= 2 \\ A_1 x_1 &= 0 \end{aligned}$$

cuja solução é $x_1 = 0$ e $A_1 = 2$. Portanto a fórmula de integração (6.15) reduz-se neste caso a

$$I_1(f) = 2f(0) \tag{6.18}$$

que é a regra do ponto médio (6.13). Esta regra deve ser de grau $2 - 1 = 1$ o que condiz com a expressão do erro (6.14).

Para $n = 2$ o sistema (6.17) reduz-se a

$$\begin{aligned} A_1 + A_2 &= 2 \\ A_1 x_1 + A_2 x_2 &= 0 \\ A_1 x_1^2 + A_2 x_2^2 &= 2/3 \\ A_1 x_1^3 + A_2 x_2^3 &= 0 \end{aligned}$$

cuja solução é

$$x_2 = -x_1 = \sqrt{3}/3 \qquad A_1 = A_2 = 1$$

Portanto a fórmula de integração (6.15) reduz-se neste caso a

$$I_2(f) = f(-\sqrt{3}/3) + f(\sqrt{3}/3) \tag{6.19}$$

que é de grau $2 \times 2 - 1 = 3$. A regra de Simpson que é também de grau 3 utiliza 3 pontos de integração. A fórmula de Gauss com 3 pontos de integração é de grau 5.

Para obter as fórmulas de Gauss com mais pontos temos que resolver sistemas não lineares da forma (6.17). Como a resolução destes sistemas não é fácil, vamos utilizar outra via para determinar os pontos de integração x_i. O teorema seguinte estabelece as condições a que devem satisfazer os x_i para que a fórmula de integração (6.15) seja de grau $2n - 1$.

Teorema 6.1 *A fórmula de integração (6.15) é de grau $2n - 1$ sse os pontos x_i são escolhidos de modo a que o polinómio*

$$\psi_n(x) = (x - x_1) \cdots (x - x_n)$$

satisfaça a relação

$$I(\psi_n q_{n-1}) = \int_{-1}^{1} \psi_n(x) q_{n-1}(x)\, dx = 0 \tag{6.20}$$

para qualquer polinómio $q_{n-1}(x)$ de grau igual ou inferior a $n - 1$ e os pesos A_i são determinados de modo a que a fórmula (6.15) seja de grau não inferior a $n - 1$, i.e., que os A_i satisfaçam (6.4) e o sistema

$$\sum_{i=1}^{n} A_i x_i^k = \int_{-1}^{1} x^k\, dx = \frac{1 + (-1)^k}{k + 1} \qquad k = 0, 1, \ldots, n - 1 \tag{6.21}$$

(que é equivalente a (6.5) com uma indexação diferente).

Demonstração: **(1)** Começamos por mostrar que, qualquer que seja a escolha dos x_i, a fórmula (6.15) não pode ser de grau superior a $2n - 1$, *i.e.*, existe sempre um polinómio q_{2n} de grau $2n$ para o qual

$$I(q_{2n}) \neq I_n(q_{2n})$$

De facto

$$I_n(\psi_n^2) = 0$$

visto

$$\psi_n(x_i) = 0 \qquad i = 0, \ldots, n$$

mas

$$I(\psi_n^2) = \int_{-1}^{1} [\psi_n(x)]^2\, dx > 0$$

Logo

$$I(\psi_n^2) \neq I_n(\psi_n^2)$$

em que ψ_n^2 é um polinómio de grau $2n$. **(2)** Vejamos agora que (6.20) é necessário para que a fórmula (6.15) seja de grau $2n - 1$. Se (6.15) é de grau $2n - 1$, então

$$I(\psi_n q_{n-1}) = I_n(\psi_n q_{n-1})$$

donde

$$\int_{-1}^{1} \psi_n(x) q_{n-1}(x)\, dx = \sum_{i=1}^{n} A_i \psi_n(x_i) q_{n-1}(x_i) = 0$$

(3) Para concluir a demonstração, vejamos que (6.20) garante que a fórmula (6.15) seja de grau $2n-1$. Consideremos um polinómio p_{2n-1} de grau igual ou inferior a $2n-1$ e seja p_{n-1} o seu polinómio interpolador nos pontos $x_1, \ldots, x_n$. Então

$$p_{2n-1}(x) = p_{n-1}(x) + \psi_n(x) q_{n-1}(x)$$

em que q_{n-1} é um polinómio de grau igual ou inferior a $n-1$. Temos então

$$I(p_{2n-1}) = I(p_{n-1}) + I(\psi_n q_{n-1}) = I(p_{n-1}) \tag{6.22}$$

em que a última igualdade é verificada quando for satisfeita a condição (6.20). Por outro lado temos

$$I_n(p_{2n-1}) = I_n(p_{n-1}) = I(p_{n-1}) \tag{6.23}$$

em que a última igualdade é garantida pela construção dos pesos A_i. De (6.22) e (6.23), concluímos como pretendíamos que

$$I_n(p_{2n-1}) = I(p_{2n-1})$$

■

Do que acabamos de demonstrar vemos que os pontos de integração da fórmula de Gauss de grau $2n-1$ devem ser os zeros de um polinómio ψ_n de grau n que satisfaça (6.20). Vamos então estudar as propriedades dos polinómios ψ_n, ou melhor, de uns polinómios P_n, múltiplos desses, conhecidos na literatura por *polinómios de Legendre*.

6.3.1 Polinómios de Legendre

Os *polinómios de Legendre* P_n podem ser definidos pela fórmula de *Rodriguez:*

$$P_n(x) = \frac{1}{2^n\, n!} \frac{d^n}{dx^n}(x^2-1)^n \tag{6.24}$$

Dado que $(x^2-1)^n$ é um polinómio de grau $2n$ em que o coeficiente de x^{2n} é unitário, concluimos, de (6.24), que P_n é um polinómio de grau n cujo coeficiente α_n de x^n é

$$\alpha_n = \frac{2n(2n-1)\cdots(2n-n+1)}{2^n\, n!}$$

ou ainda

$$\alpha_n = \frac{(2n)!}{2^n\, (n!)^2} = \frac{1.3\ldots(2n-1)}{n!} \tag{6.25}$$

Notando que para $k = 1, \ldots, n$

$$\frac{d^{n-k}}{dx^{n-k}}(x^2-1)^n$$

anula-se em $x = \pm 1$, utilizando a definição (6.24) e efectuando n integrações por parte obtemos

$$\int_{-1}^{1} P_n(x) q_{n-1}(x)\, dx = \frac{(-1)^n}{2^n\, n!} \int_{-1}^{1} (x^2-1)^n \frac{d^n}{dx^n} q_{n-1}(x)\, dx$$

Se $q_{n-1}(x)$ for um polinómio de grau igual ou inferior a $n-1$ então da igualdade anterior concluimos que

$$\int_{-1}^{1} P_n(x) q_{n-1}(x)\, dx = 0 \tag{6.26}$$

que é equivalente a condição (6.20) do Teorema 6.1. Logo os zeros do polinómio de Legendre P_n de grau n são os pontos de integração da fórmula de Gauss designada mais propriamente por *Gauss-Legendre*. De (6.26) concluimos que para 2 polinómios de Legendre P_n e P_m existe a *relação de ortogonalidade*

$$\int_{-1}^{1} P_n(x) P_m(x)\, dx = 0 \qquad m \neq n \tag{6.27}$$

Assim os *polinómios de Legendre* constituem um conjunto de *funções ortogonais*. Para $m = n$, temos atendendo a (6.26)

$$\begin{aligned} \gamma_n &= \int_{-1}^{1} P_n^2(x)\, dx \\ &= \int_{-1}^{1} P_n(x) \alpha_n x^n\, dx \end{aligned} \tag{6.28}$$

Atendendo a (6.24) e integrando por partes n vezes o último integral obtido temos

$$\gamma_n = \frac{\alpha_n}{2^n} \int_{-1}^{1} (1 - x^2)^n\, dx$$

Integrando por partes $n+1$ vezes este último integral e substituindo α_n pelo seu valor dado por (6.25), obtemos finalmente

$$\gamma_n = \int_{-1}^{1} P_n^2(x)\, dx = \frac{2}{2n+1} \tag{6.29}$$

Teorema 6.2 *Os polinómios de Legendre P_n podem obter-se utilizando a fórmula de recorrência*

$$\begin{aligned} &P_0(x) = 1 \qquad P_1(x) = x \\ &P_{n+1}(x) = \frac{1}{n+1}[(2n+1) x P_n(x) - n P_{n-1}(x)] \qquad n = 1, \ldots \end{aligned} \tag{6.30}$$

Demonstração: Atendendo que α_n, dado por (6.25), é o coeficiente de x^n de $P_n(x)$, vemos que

$$Q_n(x) = P_{n+1}(x) - \frac{\alpha_{n+1}}{\alpha_n} x P_n(x) \tag{6.31}$$

é um polinómio de grau igual ou inferior a n. Como os polinómios P_i são de grau i podemos exprimir o polinómio $Q_n(x)$ como uma combinação linear de polinómios de Legendre:

$$Q_n(x) = \sum_{i=0}^{n} a_i P_i(x) \tag{6.32}$$

e substituir na igualdade anterior (6.31). Multiplicando a igualdade assim obtida por $P_m(x)$, integrando entre -1 e 1 e atendendo a ortogonalidade dos polinómios de Legendre, obtemos

$$\begin{aligned}\int_{-1}^{1} P_{n+1}(x)P_m(x)\,dx - \frac{\alpha_{n+1}}{\alpha_n}\int_{-1}^{1} xP_n(x)P_m(x)\,dx &= \int_{-1}^{1}\sum_{i=0}^{n} a_iP_i(x)P_m(x)\,dx \\ &= a_m\int_{-1}^{1} P_m^2(x)\,dx = a_m\gamma_m\end{aligned}$$

em que o valor de γ_m é dado por (6.29). Se fizermos sucessivamente $m = 0, \ldots, n$ e atendermos a (6.26), obtemos

$$\begin{aligned}&a_m = 0 \qquad m = 0, \ldots, n-2 \\ &a_{n-1}\gamma_{n-1} = -\frac{\alpha_{n+1}}{\alpha_n}\int_{-1}^{1} xP_n(x)P_{n-1}(x)\,dx \\ &a_n\gamma_n = -\frac{\alpha_{n+1}}{\alpha_n}\int_{-1}^{1} xP_n^2(x)\,dx = 0\end{aligned}$$

Vemos que em (6.32), a excepção de a_{n-1}, todos os coeficientes a_i são nulos. Atendendo a (6.27) e (6.28) temos ainda

$$a_{n-1}\gamma_{n-1} = -\frac{\alpha_{n+1}}{\alpha_n}\int_{-1}^{1} P_n(x)\alpha_{n-1}x^n\,dx = -\frac{\alpha_{n-1}\alpha_{n+1}}{\alpha_n^2}\gamma_n$$

Utilizando este resultado em (6.32) e substituindo em (6.31), obtemos

$$\begin{aligned}P_{n+1}(x) - \frac{\alpha_{n+1}}{\alpha_n}xP_n(x) &= a_{n-1}P_{n-1}(x) \\ &= -\frac{\alpha_{n-1}\alpha_{n+1}}{\alpha_n^2}\frac{\gamma_n}{\gamma_{n-1}}P_{n-1}(x)\end{aligned}$$

Basta atender agora a (6.25) (6.29) para obter a fórmula de recorrência (6.30), como pretendíamos. ■

Utilizando esta fórmula de recorrência podemos obter alguns dos polinómios de Legendre de grau mais baixo:

$$\begin{aligned}P_0(x) &= 1 \\ P_1(x) &= x \\ P_2(x) &= \frac{1}{2}(3x^2 - 1) \\ P_3(x) &= \frac{1}{2}(5x^3 - 3x) \\ P_4(x) &= \frac{1}{8}(35x^4 - 30x^2 + 3)\end{aligned}$$

6.3.2 Fórmulas de Gauss-Legendre. Erro de integração

Como vimos os zeros $x_1, \ldots, x_n$ do polinómio de Legendre P_n são os pontos de integração da fórmula de Gauss-Legendre (6.15) de grau $2n - 1$. Os pesos A_i são determinados ou utilizando (6.4) ou resolvendo o sistema (6.21).

Exemplo 6.5 *Obter sucessivamente as fórmulas de Gauss-Legendre com 1,2,3 e 4 pontos.*

$n = 1 : x_1 = 0$ é o zero de P_1. Utilizando (6.21) com $k = 0$, temos $A_1 = 2$. Obtemos portanto (6.18).

$n = 2 : x_1 = 1/\sqrt{3}$ e $x_2 = -1/\sqrt{3}$ são os zeros de P_2. Resolvendo (6.21) temos $A_1 = A_2 = 1$. Obtemos portanto (6.19).

$n = 3 : x_1 = 0$, $x_2 = -\sqrt{3/5}$ e $x_3 = \sqrt{3/5}$ são os zeros de P_3. Resolvendo o sistema

$$\begin{bmatrix} 1 & 1 & 1 \\ 0 & -\sqrt{3/5} & \sqrt{3/5} \\ 0 & 3/5 & 3/5 \end{bmatrix} \begin{bmatrix} A_1 \\ A_2 \\ A_3 \end{bmatrix} = \begin{bmatrix} 2 \\ 0 \\ 2/3 \end{bmatrix}$$

obtemos $A_1 = 8/9$ e $A_2 = A_3 = 5/9$. A fórmula é por conseguinte

$$I_3(f) = 8/9\ f(0) + 5/9\ f(-\sqrt{3/5}) + 5/9\ f(\sqrt{3/5})$$

$n = 4 : x_1 = -x_3 = \sqrt{(3 - 2\sqrt{6/5})/7}$ e $x_2 = -x_4 = \sqrt{(3 + 2\sqrt{6/5})/7}$ são os zeros de P_4. Resolvendo (6.21) temos

$$A_1 = A_3 = 1/2 + (1/6)\sqrt{5/6}$$

e

$$A_2 = A_4 = 1/2 - (1/6)\sqrt{5/6}$$

A fórmula é por conseguinte

$$\begin{aligned} I_4(f) = \ & 0.6521451549[f(-0.3399810436) + f(0.3399810436)] + \\ & +0.3478548451[f(-0.8611363116) + f(0.8611363116)] \end{aligned}$$

■

Na Tabela 6.3 apresentamos os pontos de integração x_i e os pesos A_i das fórmulas de Gauss-Legendre de ordem mais baixa.

Vamos agora obter uma expressão para o erro de integração da fórmula de Gauss-Legendre (6.15). A fórmula $I_n(f)$ é construida de modo a que

$$I_n(p_{n-1}) = I(p_{n-1})$$

em que p_{n-1} é o polinómio que interpola f nos pontos de integração $x_1, \ldots, x_n$ (zeros do polinómio de Legendre P_n). Sendo assim o erro de integração obtém-se integrando o erro de interpolação:

$$\begin{aligned} E_n(f) &= I(f) - I_n(f) \\ &= \int_{-1}^{1} [f(x) - p_{n-1}(x)]\ dx \\ &= \int_{-1}^{1} f[x_1, \ldots, x_n, x](x - x_1) \cdots (x - x_n)\ dx \end{aligned}$$

Tabela 6.3: Pontos e pesos das fórmulas de Gauss-Legendre

n	x_i	A_i
2	$\pm.5773502692$	1.0000000000
3	$\pm.7745966692$	.5555555556
	0.0	.8888888889
4	$\pm.8611363116$	.3478548451
	$\pm.3399810436$	.6521451549
5	$\pm.9061798459$	.2369268851
	$\pm.5384693101$	.4786286705
	0.0	.5688888889
6	$\pm.9324695142$	.1713244924
	$\pm.6612093865$	.3607615730
	$\pm.2386191861$	.4679139346
7	$\pm.9491079123$	.1294849662
	$\pm.7415311856$	.2797053915
	$\pm.4058451514$	.3818300505
	0.0	.4179591837
8	$\pm.9602898565$	.1012285363
	$\pm.7966664774$	.2223810345
	$\pm.5255324099$	.3137066459
	$\pm.1834346425$	.3626837834

Como o polinómio

$$\psi_n(x) = (x - x_1) \cdots (x - x_n)$$

muda de sinal $n + 1$ vezes no intervalo de integração $[-1, 1]$ não podemos utilizar directamente o teorema do valor intermédio para integrais. Para podermos utilizar este teorema vamos recorrer primeiro à fórmula interpoladora de Newton. Assim considerando n novos pontos de interpolação arbitrários $y_n, \ldots, y_n$, podemos escrever

$$\begin{aligned} E_n(f) = \; & f[x_1, \ldots, x_n, y_1] \int_{-1}^{1} \psi_n(x)\, dx + f[x_1, \ldots, x_n, y_1, y_2] \int_{-1}^{1} \psi_n(x)(x - y_1)\, dx + \\ & + \cdots + f[x_1, \ldots, x_n, y_1, \ldots, y_n] \int_{-1}^{1} \psi_n(x)(x - y_1) \cdots (x - y_{n-1})\, dx + \\ & + \int_{-1}^{1} f[x_1, \ldots, x_n, y_1, \ldots, y_n, x] \psi_n(x)(x - y_1) \cdots (x - y_n)\, dx \end{aligned}$$

Atendendo a que o polinómio ψ_n satisfaz a (6.20) temos que

$$E_n(f) = \int_{-1}^{1} f[x_1, \ldots, x_n, y_1, \ldots, y_n, x] \psi_n(x)(x - y_1) \cdots (x - y_n)\, dx$$

Para aplicar o teorema do valor intermédio para integrais basta fazer $y_i = x_i$, $i = 1, \ldots, n$. Supondo que $f \in C^{2n}([-1,1])$, temos então

$$\begin{aligned} E_n(f) &= \int_{-1}^{1} f[x_1, \ldots, x_n, x_1, \ldots, x_n, x]\psi_n(x)\psi_n(x)\, dx \\ &= f[x_1, \ldots, x_n, x_1, \ldots, x_n, \eta]\int_{-1}^{1} \psi_n^2(x)\, dx \qquad \eta \in (-1,1) \\ &= \frac{f^{(2n)}(\xi)}{(2n)!}\int_{-1}^{1} \psi_n^2(x)\, dx \qquad \xi \in (-1,1) \end{aligned}$$

Notando que

$$\psi_n(x) = \frac{1}{\alpha_n} P_n(x)$$

em que α_n é dado por (6.29) e P_n é o polinómio de Legendre de grau n e atendendo a (6.29) temos sucessivamente

$$E_n(f) = \frac{f^{(2n)}(\xi)}{(2n)!\alpha_n^2}\int_{-1}^{1} P_n^2(x)\, dx$$

e

$$E_n(f) = \frac{f^{(2n)}(\xi)}{(2n)!}\, e_n \qquad \xi \in (-1,1)$$

$$e_n = \frac{2^{2n+1}\,(n!)^4}{(2n+1)\,[(2n)!]^2}$$

Desta expressão concluimos, mais uma vez, que a fórmula de integração de Gauss-Legendre com n pontos é de ordem $2n-1$. Com efeito se f for um polinómio de grau igual ou inferior a $2n-1$ então $E_n(f) = 0$ e se $f(x) = x^{2n}$ então $E_n(f) = e_n \neq 0$.

Se utilizarmos a fórmula assintótica de Stirling:

$$n! \approx e^{-n} n^n \sqrt{2\pi n}$$

que permite estimar $n!$ quando for $n \gg 1$, então obtemos

$$e_n \approx \frac{\pi}{4^n}$$

quando for $n \gg 1$. Definindo

$$M_k = \max_{-1 \le x \le 1} \frac{|f^{(k)}(x)|}{k!} \qquad k \ge 0$$

obtemos a seguinte majoração para o erro de integração quando n for grande:

$$|E_n(f)| \le \frac{\pi}{4^n} \cdot M_{2n}$$

Para a maioria das funções temos que $\sup_{k \ge 0} M_k < \infty$ (normalmente $\lim_{k\to\infty} M_k = 0$) e portanto para estas funções o erro de integração tende exponencialmente para zero quando $n \to \infty$.

Pode-se demonstrar que para qualquer função f, integrável em $[-1,1]$, $I_n(f) \to I(f)$ quando $n \to \infty$, sendo $I_n(f)$ a fórmula de integração de Gauss-Legendre com n pontos. Lembramos que isto não é verdade quando $I_n(f)$ é a fórmula de integração de Newton-Cotes baseada no polinómio interpolador p_n.

Exemplo 6.6 *Utilizar as fórmulas de Gauss-Legendre para calcular o integral (6.9)*

$$I = \int_0^{\pi} e^x \cos x \, dx$$

considerado nos Exemplos 6.3 e 6.4.

Antes de utilizarmos as fórmulas de Gauss-Legendre, temos que reduzir I a um integral com o intervalo de integração $[-1, 1]$, utilizando (6.16):

$$I = \int_0^{\pi} e^t \cos t \, dt = \int_{-1}^{1} (\pi/2) \, e^{\pi(x+1)/2} \cos(\pi(x+1)/2) \, dx$$

Apresentamos na Tabela 6.4 os valores aproximados I_n do integral I e os respectivos erros $E_n = I - I_n$. Desta tabela concluimos que o erro decresce muito rapidamente

Tabela 6.4: Ilustração das fórmulas de Gauss-Legendre

n	I_n	E_n
2	-12.33621046570	$2.66E-1$
3	-12.12742045017	$5.71E-2$
4	-12.07018949029	$-1.57E-4$
5	-12.07032853589	$-1.78E-5$
6	-12.07034633110	$1.47E-8$
7	-12.07034631753	$1.14E-9$
8	-12.07034631639	$< 5.0E-12$

quando o número n de pontos de integração aumenta. Comparando os resultados com os apresentados nas Tabelas 6.1 e 6.2, vemos que as fórmulas de Gauss-Legendre são computacionalmente melhores do que as regras dos trapézios e de Simpson. ■

6.4 Integração adaptativa

É frequente que uma função integranda apresenta um comportamento muito diferente ao longo do intervalo de integração. Por exemplo, para o integral

$$I = \int_0^1 \sqrt{x} \, dx \tag{6.33}$$

a função integranda $\sqrt{x}$ tem um declive infinito em $x = 0$ mas é bem comportada para pontos perto de 1.

Os métodos até agora apresentados, utilizam um conjunto de pontos de integração distribuidos de uma forma praticamente uniforme ao longo do intervalo de integração. Quando uma função é mal comportada na vizinhança de um ponto s no intervalo $[a, b]$, é necessário utilizar muitos pontos de integração perto de s. Isto implica, com uma distribuição uniforme, utilizar desnecessariamente muitos pontos de

integração no resto do intervalo de integração $[a, b]$. Para obviar a este inconveniente, na *integração adaptativa* a distribuição dos pontos de integração depende do comportamento da função integranda, sendo maior a densidade destes pontos onde a função tiver pior comportamento.

Para explicar o conceito de integração adaptativa vamos utilizar a regra de Simpson. Designamos por I_{ad} o valor aproximado de um integral

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

utilizando a integração adaptativa. Pretendemos que seja

$$|I - I_{ad}| \leq \epsilon \tag{6.34}$$

em que ϵ é a tolerância para o erro cometido. Numa dada fase da integração adaptativa o intervalo de integração está dividido em sub-intervalos de comprimentos não necessariamente todos iguais. Seja $[a_i, b_i]$ um destes sub-intervalos. Vamos utilizar a regra de Simpson para calcular o integral:

$$I_i = \int_{a_i}^{b_i} f(x)\, dx$$

Assim designando $h_i = (b_i - a_i)/2$ e utilizando a regra de Simpson com 3 e 5 pontos temos

$$I_{i2} = \frac{h_i}{3}[f(a_i) + 4f(a_i + h_i) + f(b_i)] \tag{6.35}$$

e

$$I_{i4} = \frac{h_i/2}{3}[f(a_i) + 4f(a_i + h_i/2) + 2f(a_i + h_i) + 4f(a_i + 3h_i/2) + f(b_i)] \tag{6.36}$$

Atendendo a (6.12) temos que

$$I_i - I_{i2} = -\frac{h_i^4(b_i - a_i)}{180} f^{(4)}(\xi_{i2}) \qquad \xi_{i2} \in (a_i, b_i)$$

e

$$I_i - I_{i4} = -\frac{(h_i/2)^4(b_i - a_i)}{180} f^{(4)}(\xi_{i4}) \qquad \xi_{i4} \in (a_i, b_i)$$

Se admitirmos que

$$f^{(4)}(\xi_{i4}) \approx f^{(4)}(\xi_{i2})$$

então

$$I_i - I_{i4} \approx \frac{1}{16}(I_i - I_{i2})$$

Desta aproximação tiramos a seguinte estimativa para o erro da aproximação I_{i4}

$$E_{i4} = I_i - I_{i4} \approx \frac{1}{15}(I_{i4} - I_{i2}) \tag{6.37}$$

Como pretendemos que (6.34) seja verificada, temos que impor

$$|I_i - I_{i4}| \leq \epsilon_i = \frac{b_i - a_i}{b - a}\epsilon$$

Como não conhecemos o valor de I_i, temos que fazer uso da aproximação (6.37) e portanto se

$$\frac{1}{15}|I_{i4} - I_{i2}| \leq \epsilon_i \tag{6.38}$$

aceitamos I_{i4} como aproximação suficientemente boa de I_i. Se (6.38) não é verificada então o intervalo $[a_i, b_i]$ é dividido ao meio e para cada um dos 2 novos subintervalos repete-se o processo atrás. O valor aproximado I_{ad} de I é obtido somando todos os I_{i4} que satisfazem a condição (6.38).

Exemplo 6.7 *Utilizando a regra de Simpson adaptativa determinar uma aproximação I_{ad} do integral (6.33)*

$$I = \int_0^1 \sqrt{x}\, dx$$

com um erro inferior a $\epsilon = 0.5E - 3$.

Começamos por dividir o intervalo de integração $[0, 1]$ em 2 sub-intervalos $[0, 0.5]$ e $[0.5, 1]$. Aplicamos (6.35) e (6.36) ao intervalo $[0.5, 1]$. Neste caso

$$h_i = (1 - 0.5)/2 = 0.25$$

$$\epsilon_i = \frac{0.5}{1} 0.0005 = 0.00025$$

e portanto

$$I_{i2} = \frac{1}{12}[\sqrt{0.5} + 4\sqrt{0.75} + \sqrt{1}] = 0.43093403$$

$$I_{i4} = \frac{1}{24}[\sqrt{0.5} + 4\sqrt{0.625} + 2\sqrt{0.75} + 4\sqrt{0.875} + \sqrt{1}] = 0.43096219$$

Como

$$\frac{1}{15}(I_{i4} - I_{i2}) = 0.0018775E - 3 < 0.25E - 3$$

a condição (6.38) é satisfeita. Portanto o valor I_{i4} vai contribuir para o valor aproximado I_{ad} do integral I. Notamos que o erro cometido no cálculo do integral de $\sqrt{x}$ no intervalo $[0.5, 1]$ é

$$E_{i4} = I_i - I_{i4} = 0.43096441 - 0.43096219 = 0.00222E - 3$$

Vemos que a sua estimativa $\frac{1}{15}(I_{i4} - I_{i2})$ é da mesma ordem de grandeza. Aplicamos agora (6.35) e (6.36) ao intervalo $[0, 0.5]$. Temos novamente $h_i = 0.25$ e portanto

$$I_{i2} = \frac{1}{12}[\sqrt{0} + 4\sqrt{0.25} + \sqrt{0.5}] = 0.22559223$$

$$I_{i4} = \frac{1}{24}[\sqrt{0} + 4\sqrt{0.125} + 2\sqrt{0.25} + 4\sqrt{0.375} + \sqrt{0.5}] = 0.23211709$$

Como

$$\frac{1}{15}(I_{i4} - I_{i2}) = 0.43499E - 3 > 0.25E - 3$$

a condição (6.38) não é satisfeita. Portanto é necessário subdividir o intervalo $[0, 0.5]$ em dois: $[0, 0.25]$ e $[0.25, 0.5]$; e para cada um destes novos intervalos proceder como

Tabela 6.5: Ilustração da regra de Simpson adaptativa

$[a_i, b_i]$	I_{i4}	$E_{i4} = I_i - I_{i4}$	$(I_{i4} - I_{i2})/15$	ϵ_i
$[0, 0.125]$	$.02901464$	$.44814E-3$	$.05437E-3$	$.0625E-3$
$[0.125, 0.25]$	$.05387027$	$.00028E-3$	$.00023E-3$	$.0625E-3$
$[0.25, 0.5]$	$.15236814$	$.00079E-3$	$.00066E-3$	$.1250E-3$
$[0.5, 1]$	$.43096219$	$.00222E-3$	$.00188E-3$	$.2500E-3$
$[a, b]$	I_{ad}	$I - I_{ad}$	$\sum(I_{i4} - I_{i2})/15$	ϵ
$[0, 1]$	$.66621524$	$.45142E-3$	$.05714E-3$	$.5000E-3$

anteriormente. Neste exemplo, a aproximação I_{i4} satisfaz a condição (6.38) no intervalo $[0.25, 0.5]$ mas não no intervalo $[0, 0.25]$. Assim temos que novamente subdividir este último intervalo em dois: $[0, 0.125]$ e $[0.125, 0.25]$; para os quais já é satisfeita a condição (6.38). Os vários passos da integração adaptativa estão resumidos na Tabela 6.5 em que a última linha corresponde a soma dos valores apresentados nas linhas anteriores. Vemos que o erro total cometido $I - I_{ad} = 0.00045142$ é menor do que a tolerância $\epsilon = 0.0005$. Se, para este exemplo, não tivessemos aplicado a integração adaptativa à regra de Simpson, em vez de 17 pontos, teríamos de utilizar 33 pontos de integração. ■

Podemos concluir que sempre que a função integranda tenha um comportamento que varia muito no intervalo de integração devemos utilizar uma integração adaptativa, visto que com este procedimento diminuímos substancialmente o número de cálculos a efectuar.

6.5 Problemas

1. Considere a tabela

x	-3	-1	1
$f(x)$	-63	-1	1

representativa do polinómio $p_3(x) = 2.5x^3 + a_2x^2 + a_1x + a_0$

(a) Os correspondentes polinómios de Lagrange são:

$$l_0(x) = \frac{1}{8}(x^2 - 1) \qquad l_1(x) = \frac{-1}{4}(x^2 + 2x - 3) \qquad l_2(x) = \frac{1}{8}(x^2 + 4x + 3)$$

Verifique que é correcta a expressão de $l_2(x)$.

(b) Os pesos que intervêm na fórmula de quadratura destinada ao cálculo numérico de $I = \int_{-3}^{1} f(x)\, dx$ são $A_0 = A_2 = 2/3$ e $A_1 = 8/3$. Verifique o valor de A_0, e calcule numericamente I.

(c) Mostre que o último valor obtido em 1b é o valor exacto de I.

2. Considere a seguinte tabela da função de erro (*erf*).

x	$erf(x)$
0.0	0.000000
0.2	0.222703
0.4	0.428392
0.6	0.603856
0.8	0.742101
1.0	0.842701

   (a) Utilize a regra de Simpson para obter uma aproximação de $\int_0^{0.8} erf(x)\, dx$.

   (b) Sabendo que
$$\left|\frac{d^{m+1} erf(x)}{dx^{m+1}}\right| \leq m^3$$
   para $1 \leq m \leq 5$ e $0 \leq x \leq 2$: Obtenha um majorante para o erro de integração cometido no cálculo anterior.

3. Considere a seguinte tabela:

x	-3	-1	1	3	5
$f(x)$	-21	-4	-0.5	1	1.5

   Pretende-se determinar numericamente $I = \int_0^2 f(x)\, dx$ utilizando os valores $f(-1)$, $f(1)$ e $f(3)$.

   (a) Resolva este problema pelo método dos coeficientes indeterminados, utilizando o método de Crout para efectuar a factorização da matriz que surge nesta resolução.

   (b) Supondo que f é um polinómio de grau não superior a m, determine qual o maior valor que m pode tomar de forma a que o resultado obtido seja igual ao valor I do integral (despreza-se os erros de arredondamento). Justifique a sua resposta.

4. Considere uma função $f : \mathbb{R} \to \mathbb{R}$ para a qual só se conhece os valores tabelados:

x	-2	-1	0
$f(x)$	12	4	-2

   Pretende-se calcular $I = \int_{-2}^0 f(x)\, dx$.

   (a) Obtenha o valor do integral pela regra dos trapézios e pela regra de Simpson.

   (b) O resultado obtido pela regra dos trapézios pode ser exacto se f tiver um certo comportamento. Qual este comportamento? Faça um esboço gráfico para este caso.

   (c) Supondo que f é um polinómio de grau 3, determina o erro cometido no cálculo de I pela regra dos trapézios, utilizando exclusivamente os resultados obtidos em 4a. Justifique.

5. Considere da função $f(x) = (9x - 3)/(x - 2)$ o conjunto de pontos:

x	-3	-2	-1	0	1
$f(x)$	6	5.25	4	1.5	-6

   (a) i. Considerando apenas os 3 primeiros pontos escreva o sistema de equações lineares a que é conduzido por aplicação do método dos coeficientes indeterminados ao cálculo do integral entre -1 e 0.
   ii. Resolva o sistema anterior pelo método de Crout.
   iii. O que pode afirmar quanto à convergência do método de Gauss-Seidel quando aplicado ao sistema obtido em 5(a)i? Utilizando como aproximação inicial o vector nulo, efectue uma iteração usando este método (na forma não matricial) na resolução do sistema.

   (b) Considere o integral $I = \int_{-3}^{1} f(x)\, dx$.
   i. Calcule I com todos os pontos tabelados aplicando as regras dos trapézios e de Simpson.
   ii. Pretende-se agora calcular I aplicando a fórmula de Gauss-Legendre com 4 pontos. Obtenha a expressão numérica final que permita, substituindo valores, obter imediatamente uma aproximação do integral da função f.

6. Pretende-se obter a fórmula de integração

$$A_0 f(-2) + A_1 f(0) + A_2 f(2)$$

de modo a que ela seja pelo menos de grau 2 para o integral $\int_{-1}^{1} f(x)\, dx$.

   (a) Resolva este problema utilizando o método de Crout para efectuar a factorização da matriz que surge nesta resolução.
   (b) Mostre que a fórmula obtida é de grau 3 .

7. Considere o integral $I = \int_{2}^{3} f(x)\, dx$ com $f(x) = x^3$.

   (a) Determine I utilizando a regra de Simpson com 5 pontos.
   (b) Determine o mesmo integral utilizando a quadratura de Gauss-Legendre (considere apenas 2 pontos de integração).
   (c) Compare os resultados obtidos em 7a e 7b, comentando-os. Obteria o mesmo resultado com a regra de Simpson se tivesse utilizado somente 3 pontos? Porquê?

8. Demonstre que na regra de integração do ponto médio se tem:

$$\int_{x_0-h/2}^{x_0+h/2} f(x)\, dx = hf(x_0) + E \qquad E = \frac{h^3 f''(\xi)}{24} \qquad \xi \in (x_0 - \frac{h}{2}, x_0 + \frac{h}{2})$$

9. Determine, utilizando sucessivamente polinómios de Lagrange e o método dos coeficientes indeterminados, , os pesos da fórmula aberta

$$I_3(f) = A_1 f(0) + A_2 f(1/2) + A_3 f(-1/2)$$

de modo a que ela seja pelo menos de grau 2 para o integral $I = \int_{-1}^{1} f(x)\, dx$. Mostre que a fórmula obtida é de grau 3. Utilize-a para estimar o valor de $\int_0^4 t^{-1/2}\, dt$.

10. As regras dos Trapézios e de Simpson podem ser deduzidas utilizando polinómios interpoladores. Quais os graus destes polinómios?

11. Para as regras dos Trapézios e de Simpson compostas existirá alguma restrição:

    (a) No número de sub-intervalos.

    (b) No espaçamento dos pontos de integração?

12. O comprimento de um arco de curva da função $y = f(x)$, entre $x = a$ e $x = b$ pode ser determinado através da expressão

$$s = \int_a^b \sqrt{1 + (dy/dx)^2}\, dx$$

    (a) Determine o comprimento da curva $y^2 = x^3$ entre $x = 0$ e $x = 4$, utilizando a regra de Simpson com 5 pontos.

    (b) O resultado obtido em 12a é exacto? Porquê? Obtenha um majorante do erro cometido.

13. Sendo $f(x) = \sqrt[3]{x}/(x^2 + \ln x)$, pretende-se calcular $I = \int_1^3 f(x)\, dx$ através da fórmula de Gauss-Legendre considerando 3 pontos. Obtenha a expressão numérica final que permita, substituindo valores, obter imediatamente uma aproximação de I.

14. Verifique que os pesos da fórmula de Gauss-Legendre com 3 pontos são:

$$A_1 = 5/9 \quad A_2 = 8/9 \quad A_3 = 5/9$$

    (a) Utilizando polinómios de Lagrange.

    (b) Utilizando o método dos coeficientes indeterminados.

15. Pretende-se calcular $I = \int_{-2}^{2}(x^7 - 3x^4 + 1)\, dx$. Para aplicar nas melhores condições a quadratura de Gauss-Legendre, qual o número de pontos que deve considerar? Justifique.

16. Pretende-se calcular numericamente $I = \int_{-1}^{1} f(x)\, dx$ pelo método dos coeficientes indeterminados utilizando da função f somente os valores $f(x_0)$, $f(x_1)$ e $f(x_2)$ em que x_0, x_1 e x_2 são os abaixo indicados.

    (a) Justificando a sua escolha, diga que tipo de função deverá ser f por forma a obtermos o valor exacto de I (despreza-se erros de arredondamento) quando:

        i. $x_0 = -1$, $x_1 = 0$, $x_2 = 1$.

        ii. $x_0 = 0$, $x_1 = -1$, $x_2 = 2$.

        iii. $x_0 = -0.7745967$, $x_1 = 0$, $x_2 = 0.7745967$ com $P_3(x_i) = 0$ $i = 0, 1, 2$ em que P_3 é o polinómio de Legendre do 3 grau.

    (b) Para o caso 16(a)ii determine a fórmula de integração a utilizar.

17. Sejam $I = \int_a^b f(x)\,dx$ o valor de um integral e I_{2n} e I_n os valores obtidos para esse integral pelo método de Simpson com $2n$ e n sub-intervalos respectivamente. Mostre que sob certas condições (e diga quais) se obtém o erro

$$E_{2n} = I - I_{2n}$$

pela fórmula

$$E_{2n} = (I_{2n} - I_n)/15$$

18. Considere o integral $I = \int_0^\infty g(x)\,dx$ em que $g(x)$ pode ser escrito na forma $g(x) = \exp(-x)f(x)$.

    (a) Determine os pesos A_0 e A_1 da regra de integração

    $$I_1 = A_0 f(x_0) + A_1 f(x_1)$$

    com $x_0 = 2 - \sqrt{2}$ e $x_1 = 2 + \sqrt{2}$, de forma a que ela seja pelo menos de grau 1.

    (b) Mostre que, a regra assim obtida é de grau 3. Atenda a que

    $$\int_0^\infty x^k \exp(-x)\,dx = k!$$

19. Justificando, obtenha um critério de paragem para a regra dos trapézios com integração adaptativa.

20. Pretende-se obter a fórmula de integração

$$A_0 f(0) + A_1[f(x_1) + f(-x_1)]$$

de modo a que ela seja pelo menos de grau 2 para o integral $\int_{-1}^1 f(x)\,dx$.

    (a) Exprime A_0 e A_1 em função de x_1.

    (b) Mostre que a fórmula obtida é pelo menos de grau 3 e determine x_1 de modo a que ela seja de grau 5.

    (c) Determine x_1 de modo a que $A_0 = A_1$.

21. A partir da tabela

x	-5	-3	-1	1	3	5
$f(x)$	1	0	-1	2	1	2

pretende-se calcular o integral $I = \int_{-3}^5 f(x)\,dx$. Obtenha o valor deste integral utilizando a regra dos trapézios e a regra de Simpson.

22. Deduza a fórmula de integração

$$\int_0^h \sqrt{x} f(x)\,dx \approx A_1 f(h/4) + A_2 f(h/2) + A_3 f(3h/4)$$

que é exacta quando f for um polinómio de grau ≤ 2.

23. Determine o grau da seguinte fórmula

$$\int_{-1}^1 f(x)\,dx \approx \frac{2}{3}[f(-\sqrt{2}/2) + f(0) + f(\sqrt{2}/2)]$$

Capítulo 7

Equações diferenciais

7.1 Introdução

A maioria dos problemas de aplicação passam pela resolução de equações diferenciais. Neste capítulo apresentam-se alguns dos principais métodos de resolução numérica de equações diferenciais ordinárias com condições iniciais. Não se incluiem, porém, dada a sua complexidade, os métodos de extrapolação . Devido à falta de preparação dos alunos, começa-se por um breve estudo teórico das equações diferenciais. Apresentam-se, de seguida, vários métodos de resolução numérica de uma equação de primeira ordem. O primeiro método apresentado é o de Euler. Por ser o mais simples, é o estudado com maior profundidade, o que facilita a abordagem aos restantes métodos. Seguem-se os métodos de Runge-Kutta, baseados na fórmula de Taylor, e os métodos multipasso, que são baseados na interpolação polinomial.

Um grande número de problemas em engenharia e outros ramos da ciência podem formular-se utilizando equações diferenciais (ED). Um dos mais simples exemplos de ED é

$$\frac{dy}{dx} = \lambda y \tag{7.1}$$

cuja solução geral é

$$y(x) = C \exp(\lambda y)$$

com C arbitrário. Se impusermos à solução desta ED a condição inicial

$$y(x_0) = y_0 \tag{7.2}$$

i.e., se a função y passar a tomar o valor inicial y_0 quando a variável $x = x_0$, então

$$C = y_0 \exp(-\lambda x_0)$$

deixa de ser arbitrário. Portanto o problema de valor inicial definido por (7.1) e (7.2) tem por única solução

$$y(x) = y_0 \exp[\lambda(x - x_0)]$$

Este problema pode considerar-se como um caso particular do seguinte

$$y' = f(x, y) \qquad y(x_0) = y_0 \tag{7.3}$$

em que f é uma função contínua num certo dominínio D do plano xy e (x_0, y_0) é um ponto de D. Admitamos que $f'_y = \partial f/\partial y$ também é contínua em D o que garante a existência de uma única solução para (7.3).

Se a função f se puder exprimir na forma

$$f(x, y) = a_0(x)\, y(x) + g(x)$$

então a ED diz-se linear. Como caso especial de ED lineares consideremos

$$y'(x) = \lambda y(x) + g(x) \qquad y(x_0) = y_0 \tag{7.4}$$

cuja solução é

$$y(x) = y_0 \exp[\lambda(x - x_0)] + \int_{x_0}^{x} \exp[\lambda(x - t)] g(t)\, dt \tag{7.5}$$

Neste caso para resolver numericamente (7.4) basta integrar numericamente o último termo de (7.5) por um dos métodos já estudados anteriormente.

Como exemplo de uma ED não linear podemos considerar

$$y'(x) = \frac{1}{1 + x^2} - 2\, y^2(x) \qquad y(0) = 0$$

cuja solução é

$$y(x) = \frac{x}{1 + x^2}$$

Vejamos outro exemplo de uma ED não linear

$$y'(x) = \frac{y^4(x)}{3x^2} \qquad y(-1) = -1 \tag{7.6}$$

cuja solução é

$$y(x) = x^{1/3} \qquad x \leq 0$$

Para $x \geq 0$ a ED (7.6) admite uma infinidade de soluções da forma

$$y(x) = \frac{x^{1/3}}{(1 + Cx)^{1/3}} \qquad x \geq 0$$

Este facto deve-se a que

$$f'_y = 4y^3(x)/(3x^2) = 4/(3x)$$

não é contínua em $x = 0$. Na Figura 7.1 representa-se algumas soluções de (7.6).

Em geral não é possível resolver uma ED por via analítica ou utilizando os métodos de integração directa. Assim somos levados a estudar métodos numéricos de resolução de ED. É o que iremos fazer neste capítulo onde apresentaremos os principais métodos numéricos para resolver o problema de valor inicial (7.3) para ED de 1ª ordem.

Em todos estes métodos iremos determinar valores aproximados da solução de (7.3) para um conjunto finito de pontos $x_0 < x_1 < \cdots < x_n < \cdots < x_N$. À *solução*

Figura 7.1: Soluções de (7.6)

exacta $y(x_n)$ corresponde o *valor aproximado* y_n que se obtem usando alguns dos valores $y_{n-1}, y_{n-2}, \ldots$ obtidos em passos anteriores. Conforme y_n é obtido utilizando y_{n-1} ou utilizando mais do que um valor aproximado da solução o método designa-se por unipasso ou por multipasso. Dentro dos *métodos unipasso* podemos incluir os métodos baseados nas séries de Taylor nomeadamente o *método de Euler* e os *métodos de Runge-Kutta*. Dentro dos *métodos multipasso* destacam-se os *métodos preditor-corrector* nomeadamente os *métodos de Adams*.

7.2 Método de Euler

Consideremos o problema de valor inicial

$$y' = f(x, y) \qquad y(x_0) = y_0 \tag{7.7}$$

Pretende-se obter uma aproximação de $y(x)$ para $x_0 \leq x \leq b$. Consideremos o conjunto de pontos $x_n = x_0 + n\,h,\ n = 0, 1, \ldots, N$ com $h = (b - x_0)/N$ e $x_N = b$. Recorrendo ao teorema de Taylor obtemos

$$y(x_{n+1}) = y(x_n) + hf(x_n, y(x_n)) + \frac{h^2}{2}y''(\xi_n) \qquad \xi_n \in (x_n, x_{n+1}) \tag{7.8}$$

atendendo a que $h = x_{n+1} - x_n$ e $f(x_n, y(x_n)) = y'(x_n)$ visto y satisfazer à ED (7.7). Fazendo $n = 0$ e desprezando o último termo em (7.8) obtemos

$$y_1 = y_0 + h\ f(x_0, y_0)$$

Considerando y_1 uma aproximação de $y(x_1)$, substituindo $y(x_1)$ por y_1 em (7.8), com $n = 1$, e procedendo como anteriormente obtemos para aproximação de $y(x_2)$

$$y_2 = y_1 + h\ f(x_1, y_1)$$

Deste modo podemos gerar uma aproximação numérica da solução de (7.7) nos pontos $x_n,\ n = 1, \ldots, N$ fazendo

$$y_{n+1} = y_n + h\ f(x_n, y_n) \qquad n = 0, 1, \ldots, N-1 \tag{7.9}$$

Este método de obter as aproximações y_n de $y(x_n)$ é conhecido pelo *método de Euler*.

Exemplo 7.1 *Utilizar o método de Euler para obter, no intervalo* $[0, 1]$*, uma solução numérica do seguinte problema de valor inicial*

$$y' = y \qquad y(0) = 1 \tag{7.10}$$

A solução exacta é $y(x) = \exp(x)$.

Começamos, p.ex., por escolher $h = 0.25$. Notando que neste caso $f(x, y) = y$ teremos

$$\begin{aligned}
y(0.25) &= 1.284025 \simeq y_1 = 1.000000 + 0.25(1.000000) = 1.250000\\
y(0.50) &= 1.648721 \simeq y_2 = 1.250000 + 0.25(1.250000) = 1.562500\\
y(0.75) &= 2.117000 \simeq y_3 = 1.562500 + 0.25(1.562500) = 1.953125\\
y(1.00) &= 2.718282 \simeq y_4 = 1.953125 + 0.25(1.953125) = 2.441406
\end{aligned}$$

Na Figura 7.2 representamos a solução exacta e as aproximações. Vemos que o ponto

Figura 7.2: Ilustração do método de Euler

(x_1, y_1) encontra-se sobre a tangente à curva $y(x)$ no ponto (x_0, y_0). Para obter uma melhor aproximação devemos escolher h mais pequeno. Por exemplo com $h = 0.1$ e $h = 0.05$ indicamos na Tabela 7.1 o erro cometido para alguns pontos x_n. ■

Nota-se neste exemplo que para um valor fixo de x o erro decresce practicamente de metade quando h é dividido por 2. Mais tarde veremos que o erro no

Tabela 7.1: Resultados referentes ao Exemplos 7.1 e 7.2

		$y(x_n) - y_n$		$h(e/2)\,(e^{x_n} - 1)$	
x_n	$y(x_n)$	$h = 0.1$	$h = 0.05$	$h = 0.1$	$h = 0.05$
0.2	1.2214	0.0114	0.0059	0.0301	0.0150
0.4	1.4918	0.0277	0.0144	0.0668	0.0334
0.6	1.8221	0.0506	0.0263	0.1117	0.0559
0.8	2.2255	0.0820	0.0427	0.1666	0.0833
1.0	2.7183	0.1245	0.0650	0.2336	0.1168

método de Euler é majorado por uma quantidade proporcional a h. Diz-se que o erro é de $O(h)$ (ordem de h) e que o método de Euler é um método de ordem 1. Sendo assim, para obter uma boa aproximação de (7.7) utilizando o método de Euler necessitamos de escolher um h pequeno o que obriga a ter que utilizar (7.9) N vezes com $N = (b - x_0)/h$ grande. Nas próximas secções vamos apresentar métodos de ordem superior que, em geral, permitem obter uma melhor precisão . Antes porém, vamos fazer um estudo do erro e da convergência do método de Euler.

7.2.1 Majoração do erro e convergência do método de Euler

Pretendemos majorar o *erro global*

$$e_n = y(x_n) - y_n$$

do método de Euler. Para isso subtraímos (7.9) de (7.8) obtendo

$$e_{n+1} = e_n + h[f(x_n, y(x_n)) - f(x_n, y_n)] + \frac{h^2}{2} y''(\xi_n) \qquad \xi \in (x_n, x_{n+1}) \tag{7.11}$$

O termo

$$T_n = \frac{h^2}{2} y''(\xi_n) \qquad \xi \in (x_n, x_{n+1}) \tag{7.12}$$

corresponde a termos truncado em (7.8) o desenvolvimento em série de Taylor e portanto denomina-se *erro de truncatura*. Também é conhecido por *erro de discretização local*, por corresponder ao erro que se cometeria ao passar de x_n para x_{n+1} se fosse $y_n = y(x_n)$. Usando o teorema do valor intermédio temos que

$$\begin{aligned} f(x_n, y(x_n)) - f(x_n, y_n) &= f'_y(x_n, \overline{y}_n)\,(y(x_n) - y_n) \\ &= f'_y(x_n, \overline{y}_n) e_n \end{aligned} \tag{7.13}$$

onde $\overline{y}_n \in \text{int}(y_n, y(x_n))$. Substituindo (7.12) e (7.13) em (7.11) temos que

$$e_{n+1} - e_n = h f'_y(x_n, \overline{y}_n) e_n + T_n \tag{7.14}$$

Como admitimos a continuidade de f'_y, existe uma constante $K > 0$ tal que para $x_0 \leq x \leq b$

$$|f'_y(x, y)| \leq K \tag{7.15}$$

Vamos supor que y'' é limitado, *i.e.*,

$$|y''(x)| \leq Y_2 \qquad x_0 \leq x \leq b$$

em que Y_2 é uma constante positiva. Então

$$|e_{n+1}| \leq (1 + hK)\, |e_n| \; + \frac{h^2}{2}\, Y_2 \tag{7.16}$$

Aplicando (7.16) recursivamente obtemos

$$|e_n| \leq \; (1 + hK)^n |e_0| \; + \; [1 + (1 + hK) + \cdots + (1 + hK)^{n-1}] \frac{h^2}{2}\, Y_2$$

ou ainda

$$|e_n| \leq \; (1 + hK)^n |e_0| \; + \; \frac{(1 + hK)^n - 1}{K} \; \frac{h}{2} Y_2$$

Notando que $e^x \geq 1 + x$ para todo o x real temos então que

$$\begin{aligned} |e_n| &\leq (e^{nhK} - 1) \frac{hY_2}{2K} + \; |e_0|\, e^{nhK} \\ &\leq h\, \frac{Y_2}{2}\, \frac{e^{(x_n - x_0)K} - 1}{K} \; + \; |e_0|\, e^{(x_n - x_0)K} \end{aligned} \tag{7.17}$$

atendendo a que $nh = x_n - x_0$. Como $e_0 = 0$ vemos que o erro global $e_n = O(h)$, logo tende para zero quando $h \to 0$ (admitindo a não existência de erros de arredondamento).

Note-se que o erro local T_n dado por (7.12) é $O(h^2)$ enquanto o erro global e_n é só de $O(h)$. Uma explicação para este facto é a seguinte: enquanto que T_n contribui para a variação do erro $e_{n+1} - e_n$ (ver (7.14)) quando se passa do ponto (x_n, y_n) para o ponto (x_{n+1}, y_{n+1}) o erro e_n é a acumulação dos n erros cometidos nas sucessivas passagens desde (x_0, y_0) até (x_n, y_n).

Exemplo 7.2 *Utilizar (7.17) no problema de valor inicial dado por (7.10), já considerado no Exemplo 7.1.*

Notando que $\partial f / \partial y = 1$, $y'' = \exp(x)$ temos $K = 1$ e para $0 \leq x \leq 1$, $Y_2 = e$. Assim (7.17) reduz-se neste caso a

$$|e^{x_n} - y_n| \leq h(e/2)\, (e^{x_n} - 1) \tag{7.18}$$

Apresentamos na Tabela 7.1 alguns valores tomados pelo 2º membro da desigualdade (7.18). Comparando estes valores com os já obtidos anteriormente para o 1º membro de (7.18), vemos que para este caso o majorante é aproximadamente o dobro do erro. ■

Na obtenção da majoração (7.17) desprezamos os erros de arredondamento. Denominando ϵ_n o *erro de arredondamento* associado ao cálculo de y_{n+1} dado por (7.9) e $\tilde{y}_{n+1}$ o valor numérico resultante temos

$$\tilde{y}_{n+1} = \tilde{y}_n + hf(x_n, \tilde{y}_n) + \epsilon_n \qquad n = 0, 1, \ldots, N - 1$$

Denominando $\tilde{e}_n = y(x_n) - \tilde{y}_n$ o erro global e procedendo de uma forma análoga a quando da majoração (7.17) obtemos

$$|\tilde{e}_n| \leq |y_0 - \tilde{y}_0|\, e^{(x_n - x_0)K} + \left(\frac{hY_2}{2} + \frac{\epsilon}{h}\right) \frac{e^{(x_n - x_0)K} - 1}{K}$$

em que $\epsilon = \max_{0 \leq n \leq N-1} |\epsilon_n|$. Vemos que quando $h \to 0$ o termo devido aos erros de arredondamento tende para infinito. Logo o erro poderá atingir um mínimo para um certo valor de h. Para conseguirmos uma maior precisão dos resultados devemos reduzir a influência dos erros de arredondamento utilizando uma aritmética de dupla precisão ou então utilizar um método cujo erro de método seja menor do que o de Euler.

7.3 Métodos de Runge-Kutta

O método de Euler foi obtido utilizando os dois primeiros termos do desenvolvimento em série de Taylor da solução da ED (7.7). Se utilizarmos mais termos do desenvolvimento, vamos obter métodos de ordem superior a 1. Assim recorrendo ao teorema de Taylor temos que

$$\begin{aligned} y(x_{n+1}) = \; & y(x_n) + hy'(x_n) + \frac{h^2}{2} y''(x_n) + \cdots + \frac{h^r}{r!} y^{(r)}(x_n) + \\ & + \frac{h^{r+1}}{(r+1)!} y^{(r+1)}(\xi_n) \qquad \xi_n \in (x_n, x_{n+1}) \end{aligned} \tag{7.19}$$

Desprezando em (7.19) o último termo e fazendo sucessivamente $n = 0, 1, \ldots, N-1$ obtemos um método de ordem r para resolver (7.7). O método de Euler corresponde a fazer $r = 1$ e atender a que $y' = f$.

Para $r = 2$ necessitamos de relacionar y'' com f. Temos que

$$y''(x) = \frac{d}{dx} f(x, y) \;=\; f'_x(x, y(x)) \;+\; f'_y(x, y(x)) y'(x)$$

O método numérico para $r = 2$ reduz-se então a

$$y_{n+1} = y_n + hf_n + \frac{h^2}{2}[f'_x(x_n, y_n) + f'_y(x_n, y_n) f_n] \tag{7.20}$$

em que $f_n = f(x_n, y_n)$. Este método tem o grave inconveniente de necessitar o cálculo de derivadas. Para evitar derivações de f recorre-se aos métodos de *Runge-Kutta* (RK).

Os métodos de RK de ordem 2 têm a seguinte forma geral

$$y_{n+1} = y_n + h(c_1 V_1 + c_2 V_2) \tag{7.21}$$

em que

$$V_1 = f_n = f(x_n, y_n) \tag{7.22}$$

$$V_2 = f(x_n + a_2 h, y_n + b_2 h V_1) \tag{7.23}$$

As constantes c_1, c_2, a_2, b_2 são determinadas de modo a que a diferença entre os $2^{\underline{os}}$ membros de (7.20) e (7.21) seja $O(h^3)$ quando $h \to 0$. Com esta finalidade vamos aplicar à (7.23) o teorema de Taylor para funções de duas variáveis.

$$V_2 = f_n + a_2 h f'_x(x_n, y_n) + b_2 h V_1 f'_y(x_n, y_n) + O(h^2) \tag{7.24}$$

Substituindo (7.24) em (7.21) temos

$$y_{n+1} = y_n + h(c_1 + c_2) f_n + h^2[c_2 a_2 f'_x(x_n, y_n) + c_2 b_2 f'_y(x_n, y_n) f_n] + O(h^3) \tag{7.25}$$

Comparando (7.25) com (7.20) concluímos que

$$c_1 + c_2 = 1 \qquad c_2 a_2 = c_2 b_2 = \frac{1}{2} \tag{7.26}$$

para que (7.21) seja um método de RK de ordem 2. Resolvendo o sistema (7.26) em ordem a a_2 e substituindo em (7.21), obtemos a expressão geral dos *métodos de RK de ordem 2*

$$y_{n+1} = y_n + h(1 - \frac{1}{2a_2}) f_n + \frac{h}{2a_2} f(x_n + a_2 h\,,\, y_n + a_2 h f_n)$$

Os valores mais utilizados para a_2 são 2/3, 1/2 e 1 a que correspondem os seguintes métodos de RK de ordem 2:

$$y_{n+1} = y_n + \frac{h}{4} f_n + \frac{3h}{4} f(x_n + \frac{2}{3}h, y_n + \frac{2}{3}h f_n) \tag{7.27}$$

$$y_{n+1} = y_n + h f(x_n + \frac{1}{2}h, y_n + \frac{1}{2}h f_n) \tag{7.28}$$

$$y_{n+1} = y_n + h\frac{1}{2}[f_n + f(x_n + h, y_n + h f_n)] \tag{7.29}$$

O método (7.27) é o método de RK de ordem 2 que conduz geralmente a um erro mais pequeno e reduz-se a um método de ordem 3 quando $f(x, y)$ for independente de y. Quando $f(x, y)$ é uma função de x o método (7.29) corresponde a *regra dos trapézios* e o método (7.28) corresponde a *regra do ponto médio*. Nota-se que $y_n + a_2 h f_n$ é o valor aproximado de $y(x_n + a_2 h)$ quando se utiliza o método de Euler. Assim, podemos considerar $f(x_n + a_2 h, y_n + a_2 h f_n)$ como uma aproximação de $f(x, y)$ para $x = x_n + a_2 h$. Vemos que o método de RK de ordem 2 utiliza aproximações de $f(x, y)$ em dois pontos x_n e $x_n + a_2 h$ enquanto que o método de Euler só utiliza em x_n. Podemos assim considerar que os métodos (7.28) e (7.29) são modificações do de Euler. O método (7.28), designado por método de Euler modificado, calcula $f(x, y)$ no ponto médio entre x_n e x_{n+1} e o método (7.29), designado por método de Heun, calcula a média dos valores que $f(x, y)$ toma em x_n e x_{n+1}.

Exemplo 7.3 *Aplicar os métodos de ordem 2 de RK (7.27), (7.28) e (7.29) e o baseado na série de Taylor (7.20) ao seguinte problema de valor inicial*

$$y' = -y^2 \qquad y(0) = 1$$

cuja solução é $y(x) = (1 + x)^{-1}$.

Apresentamos na Tabela 7.2 o erro cometido com os vários métodos referidos atrás incluindo o método de Euler (7.9). Da tabela vê-se que os erros para os métodos de ordem 2 são bastante mais pequenos do que para o método de ordem 1 (Euler). Para este exemplo é o método (7.29) que dá melhores resultados e comparando este método para $h = 0.1$ e $h = 0.05$ vemos que quando h é dividido por 2 o erro vem aproximadamente dividido por 4 o que confirma numericamente o facto do erro ser de $O(h^2)$. ■

Tabela 7.2: $y(x_n) - y_n$

x_n	SOL. EX. $y(x_n)$	TAYL. (7.20) $h = 0.1$	RK (7.27) $h = 0.1$	RK (7.28) $h = 0.1$
0.2	0.833333	−0.001392	−0.000883	−0.001010
0.4	0.714286	−0.001737	−0.001104	−0.001262
0.6	0.625000	−0.001732	−0.001104	−0.001261
0.8	0.555556	−0.001612	−0.001029	−0.001174
1.0	0.500000	−0.001461	−0.000934	−0.001066

	RK (7.29)		EULER (7.9)	
x_n	$h = 0.1$	$h = 0.05$	$h = 0.1$	$h = 0.05$
0.2	−0.000629	−0.000151	0.014333	0.006717
0.4	−0.000789	−0.000190	0.018901	0.008991
0.6	−0.000790	−0.000190	0.019836	0.009531
0.8	−0.000738	−0.000178	0.019338	0.009357
1.0	−0.000671	−0.000162	0.018287	0.008895

Pode-se obter métodos de RK de ordem superior a 2 por uma via análoga à utilizada na obtenção dos métodos de ordem 2. Um exemplo de um *método de RK de ordem* 4 é dado por

$$\begin{aligned} &y_{n+1} = y_n + \frac{h}{6}(V_1 + 2V_2 + 2V_3 + V_4) \\ &V_1 = f(x_n, y_n) \qquad V_2 = f(x_n + \tfrac{1}{2}h, y_n + \tfrac{1}{2}hV_1) \\ &V_3 = f(x_n + \tfrac{1}{2}h, y_n + \tfrac{1}{2}hV_2) \qquad V_4 = f(x_n + h, y_n + hV_3) \end{aligned} \tag{7.30}$$

Por ser o método de RK mais utilizado é denominado simplesmente *método de Runge-Kutta*. O método de RK (7.30) reduz-se à *regra de Simpson* quando $f(x, y)$ não depende de y.

Do que acabamos de ver, podemos concluir que quanto maior for a ordem do métodos de RK maior será o número de vezes em que é calculado $f(x, y)$ num passo. Para obviar a este inconveniente apresentaremos a seguir outra classe de métodos.

7.4 Métodos multipasso

7.4.1 Métodos de Adams-Bashforth

Os métodos apresentados até agora, métodos baseados nas séries de Taylor e métodos de Runge-Kutta, são métodos unipasso. Para obter uma aproximação da solução do problema de valor inicial (7.7) em x_{n+1} estes métodos só necessitam do conhecimento de uma aproximação da solução de (7.7) em x_n. Pelo contrário os métodos multipasso necessitam do conhecimento de uma aproximação da solução em vários pontos.

Com a finalidade de obter uma classe de métodos multipasso consideremos a ED (7.7) e integremos entre x_n e x_{n+1} obtendo

$$y(x_{n+1}) = y(x_n) + \int_{x_n}^{x_{n+1}} f(x, y(x))\, dx \tag{7.31}$$

Seja P_{r-1} o polinómio de grau $r-1$ que interpola f nos r pontos $x_n, x_{n-1}, \ldots, x_{n-r+1}$. Recorrendo à fórmula de Newton regressiva temos

$$\begin{aligned} P_{r-1}(x) = \quad & f_n + \theta \nabla f_n + \frac{\theta(\theta+1)}{2!}\nabla^2 f_n + \cdots + \\ & + \frac{\theta(\theta+1)\cdots(\theta+r-2)}{(r-1)!}\nabla^{r-1} f_n \end{aligned} \tag{7.32}$$

em que $f_n = f(x_n, y(x_n))$, $\nabla f_n = f_n - f_{n-1}$ e $\theta = (x - x_n)/h$. O erro de interpolação é dado por

$$\begin{aligned} E_{r-1}(x) &= f(x, y(x)) - P_{r-1}(x) \\ &= \frac{\theta(\theta+1)\cdots(\theta+r-1)}{r!} h^r y^{(r+1)}(\zeta_x) \qquad \zeta_x \in (x_{n-r+1}, x_{n+1}) \end{aligned} \tag{7.33}$$

com $x \in (x_n, x_{n+1})$. Substituindo em (7.31) f por P_{r-1} obtemos

$$y_{n+1} = y_n + h\,(f_n + \sum_{j=1}^{r-1} \gamma_j \nabla^j f_n) \qquad n \geq r-1 \tag{7.34}$$

em que

$$\gamma_j = \frac{1}{j!}\int_0^1 \theta(\theta+1)\cdots(\theta+j-1)\, d\theta$$

e $f_n = f(x_n, y_n)$. O método (7.34) é conhecido por *Adams-Bashforth* e atendendo a (7.33) tem por erro de truncatura

$$T_n = \gamma_r\, h^{r+1} y^{(r+1)}(\xi_n) \qquad \xi_n \in (x_{n-r+1}, x_{n+1})$$

Dado que $T_n = 0(h^{r+1})$ concluimos que o método (7.34) é de ordem r. Apresentamos a seguir os métodos de Adams-Bashforth de ordem mais baixa.

$$\begin{aligned} & r = 1 \\ & y_{n+1} = y_n + h f_n \qquad n \geq 0 \\ & T_n = \tfrac{1}{2} h^2 y''(\xi_n) \end{aligned} \tag{7.35}$$

$$
\begin{aligned}
&r = 2 \\
&y_{n+1} = y_n + \frac{h}{2}(3f_n - f_{n-1}) \qquad n \geq 1 \\
&T_n = \frac{5}{12}h^3 y'''(\xi_n)
\end{aligned} \tag{7.36}
$$

$$
\begin{aligned}
&r = 3 \\
&y_{n+1} = y_n + \frac{h}{12}(23f_n - 16f_{n-1} + 5f_{n-2}) \qquad n \geq 2 \\
&T_n = \frac{3}{8}h^4 y^{(4)}(\xi_n)
\end{aligned} \tag{7.37}
$$

$$
\begin{aligned}
&r = 4 \\
&y_{n+1} = y_n + \frac{h}{24}(55f_n - 59f_{n-1} + 37f_{n-2} - 9f_{n-3}) \qquad n \geq 3 \\
&T_n = \frac{251}{720}h^5 y^{(5)}(\xi_n)
\end{aligned} \tag{7.38}
$$

Notamos que o método de ordem 1 não é mais do que o método de Euler. Os métodos de Adams-Bashforth são métodos de r passos que fazem intervir não só y_n mas também $y_{n-1}, \ldots, y_{n-r+1}$. Por isso $y_1, \ldots, y_{r-1}$ têm de ser determinados por outros métodos; os métodos multipasso não são auto-suficientes. Têm, contudo, a vantagem sobre os métodos de Runge-Kutta de, em cada passo, só se calcular um novo valor f_n da função f.

7.4.2 Métodos de Adams-Moulton

Consideremos novamente (7.31) mas agora construimos o polinómio Q_{r-1} que interpola f nos r pontos $x_{n+1}, x_n, \ldots, x_{n-r+2}$. Temos então

$$
\begin{aligned}
Q_{r-1}(x) = \; & f_{n+1} + (\theta - 1)\nabla f_{n+1} + \frac{(\theta-1)\theta}{2!}\nabla^2 f_{n+1} + \cdots + \\
& + \frac{(\theta-1)\theta\cdots(\theta+r-3)}{(r-1)!}\nabla^{r-1} f_{n+1}
\end{aligned}
$$

O erro de interpolação é agora dado por

$$
\begin{aligned}
E_{r-1}(x) &= f(x, y(x)) - Q_{r-1}(x) \\
&= \frac{(\theta-1)\theta\cdots(\theta+r-2)}{r!}h^r y^{(r+1)}(\zeta_x) \qquad \zeta_x \in (x_{n-r+2}, x_{n+1})
\end{aligned} \tag{7.39}
$$

Substituindo em (7.31) f por Q_{r-1} obtemos

$$
y_{n+1} = y_n + h(f_{n+1} + \sum_{j=1}^{r-1} \delta_j \nabla^j f_{n+1}) \qquad n \geq r-2 \tag{7.40}
$$

em que

$$
\delta_j = \frac{1}{j!}\int_0^1 (\theta-1)\theta\cdots(\theta+j-2)\,d\theta
$$

O método (7.40) é conhecido por *Adams-Moulton* e atendendo a (7.39) tem por erro de truncatura

$$T_n = \delta_r h^{r+1} y^{(r+1)}(\xi_n) \qquad \xi_n \in (x_{n-r+2}, x_{n+1})$$

Apresentamos a seguir os métodos de Adams-Moulton de ordem r mais baixa.

$$\begin{aligned} &r = 1 \\ &y_{n+1} = y_n + hf_{n+1} \qquad n \geq 0 \\ &T_n = -\frac{1}{2}h^2 y''(\xi_n) \end{aligned} \tag{7.41}$$

$$\begin{aligned} &r = 2 \\ &y_{n+1} = y_n + \frac{h}{2}(f_{n+1} + f_n) \qquad n \geq 0 \\ &T_n = -\frac{1}{12}h^3 y'''(\xi_n) \end{aligned} \tag{7.42}$$

$$\begin{aligned} &r = 3 \\ &y_{n+1} = y_n + \frac{h}{12}(5f_{n+1} + 8f_n - f_{n-1}) \qquad n \geq 1 \\ &T_n = -\frac{1}{24}h^4 y^{(4)}(\xi_n) \end{aligned} \tag{7.43}$$

$$\begin{aligned} &r = 4 \\ &y_{n+1} = y_n + \frac{h}{24}(9f_{n+1} + 19f_n - 5f_{n-1} + f_{n-2}) \qquad n \geq 2 \\ &T_n = -\frac{19}{720}h^5 y^{(5)}(\xi_n) \end{aligned} \tag{7.44}$$

Quando $f(x, y)$ é uma função só de x o método (7.42) corresponde à regra dos trapézios. Comparando o erro de truncatura destes métodos com os métodos de Adams-Bashforth vemos que para métodos da mesma ordem é o de Adams-Moulton que fornece geralmente melhores resultados.

Nos métodos de Adams-Moulton intervêm o cálculo de $f_{n+1} = f(x_{n+1}, y_{n+1})$ e por conseguinte não é, em geral, possível explicitar y_{n+1}. Por essa razão estes métodos denominam-se *implícitos*. Num método implícito é necesssário, em geral, resolver uma equação não linear cuja solução é y_{n+1}. Vamos utilizar o método iterativo linear do ponto fixo em que a aproximação inicial de y_{n+1} é dada por um método *explícito* como por exemplo um dos métodos de Adams-Bashforth.

7.4.3 Métodos preditor-corrector

Consideremos o método implícito de ordem 2 dado por (7.42). Supondo que temos uma aproximação inicial $y_{n+1}^{(0)}$ de y_{n+1}, então podemos corrigir este valor utilizando (7.42). Assim temos

$$y_{n+1}^{(1)} = y_n + \frac{h}{2}\left[f(x_n, y_n) + f(x_{n+1}, y_{n+1}^{(0)})\right]$$

Continuando o processo iterativo temos que

$$y_{n+1}^{(i+1)} = y_n + \frac{h}{2}\left[f(x_n, y_n) + f(x_{n+1}, y_{n+1}^{(i)})\right] \qquad i = 0, 1, \ldots \tag{7.45}$$

Para determinar as condições para as quais o processo iterativo converge subtraímos (7.45) de (7.42).

$$y_{n+1} - y_{n+1}^{(i+1)} = \frac{h}{2}\left[f(x_{n+1}, y_{n+1}) - f(x_{n+1}, y_{n+1}^{(i)})\right]$$

Utilizando o teorema do valor intermédio e a majoração (7.15) ($|\partial f/\partial y| \leq K$) obtemos

$$|y_{n+1} - y_{n+1}^{(i+1)}| \leq \frac{hK}{2}\,|y_{n+1} - y_{n+1}^{(i)}|$$

Vemos que escolhendo $h < 2/K$ o método iterativo converge. Supondo que $h \leq 1/K$ então

$$|y_{n+1} - y_{n+1}^{(i+1)}| \leq |y_{n+1}^{(i+1)} - y_{n+1}^{(i)}|$$

Logo podemos escolher para critério de paragem

$$|y_n^{(i+1)} - y_{n+1}^{(i)}| \leq \epsilon$$

sendo ϵ um valor previamente escolhido.

Normalmente determina-se a aproximação inicial $y_{n+1}^{(0)}$ utilizando um método explícito da mesma ordem que a do método implícito. Assim no caso considerado utilizamos a fórmula de Adams-Bashforth (7.36)

$$y_{n+1}^{(0)} = y_n + \frac{h}{2}\left[3f(x_n, y_n) - f(x_{n-1}, y_{n-1})\right] \tag{7.46}$$

Esta fórmula permite *prever* aproximadamente o valor de y_{n+1}. Por outro lado utilizando (7.45) uma ou mais vezes vai-se *corrigindo* o valor de y_{n+1}. Por isso (7.46) conjuntamente com (7.45) constitui um exemplo de um *método preditor-corrector*.

Consideremos o par preditor-corrector de ordem 4, dado pelas fórmulas (7.38) e (7.44). Normalmente h é escolhido de modo que seja só necessário uma iteração. Temos então que

$$\begin{aligned} y_{n+1}^{(0)} &= y_n + \frac{h}{24}[55f(x_n, y_n) - 59f(x_{n-1}, y_{n-1}) + 37f(x_{n-2}, y_{n-2}) - 9f(x_{n-3}, y_{n-3})] \\ y_{n+1} &= y_n + \frac{h}{24}[9f(x_{n+1}, y_{n+1}^{(0)}) + 19f(x_n, y_n) - 5f(x_{n-1}, y_{n-1}) + f(x_{n-2}, y_{n-2})] \end{aligned} \tag{7.47}$$

Vemos que só é necessário determinar 2 novos valores de f: $f(x_n, y_n)$ e $f(x_{n+1}, y_{n+1}^{(0)})$. No método RK tinhamos que determinar 4 valores de f. Além disso a partir de $y_{n+1} - y_{n+1}^{(0)}$ é possível obter uma estimativa do erro $y(x_{n+1}) - y_{n+1}$. Por outro lado para inicializar o preditor é necessário determinar por outros métodos, o de RK por exemplo, os valores de y_1, y_2 e y_3 (y_0 é dado).

7.4.4 Outros métodos multipasso

Para obtermos outros métodos multipasso podemos, tal como fizemos no caso dos métodos de Adams, integrar a ED (7.7), mas agora, entre x_{n-p} e x_{n+1} com $p \geq 0$ inteiro. Voltando a substituir f pelo polinómio $P_{r-1,k}$ que interpola f nos pontos $x_{n+k}, x_{n-1+k}, \ldots, x_{n-r+1+k}$ com $k = 0$ ou $k = 1$ conforme se pretende obter um método explícito ou implícito, temos então

$$y_{n+1} = y_{n-p} + \int_{x_{n-p}}^{x_{n+1}} P_{r-1,k}(x)\, dx \tag{7.48}$$

Aos casos $p = 0$, $k = 0$ e $p = 0$, $k = 1$ correspondem os já considerados métodos de Adams-Bashforth e Adams-Moulton.

Consideremos agora o caso $k = 1$, $p = r - 2$. Este caso corresponde às *fórmulas fechadas de Newton-Cotes* quando $f(x, y)$ é uma função de x. Assim, $r = 2$ (7.48) reduz-se a (7.42) a que corresponde a *regra dos trapézios.* Com $r = 3$ temos

$$\begin{aligned} y_{n+1} &= y_{n-1} + \tfrac{h}{3}\,(f_{n+1} + 4f_n + f_{n-1}) \qquad n \geq 1 \\ T_n &= -\tfrac{1}{90}\, h^5 y^{(5)}(\xi_n) \end{aligned} \tag{7.49}$$

a que corresponde a *regra de Simpson.* . Com $r = 4$ temos

$$\begin{aligned} y_{n+1} &= y_{n-2} + \tfrac{3}{8} h\,(f_{n+1} + 3f_n + 3f_{n-1} + f_{n-2}) \qquad n \geq 2 \\ T_n &= -\tfrac{3}{80}\, h^5 y^{(5)}(\xi_n) \end{aligned}$$

a que corresponde a *regra dos* 3/8. Nota-se que neste caso o método é de ordem r para r par e de ordem $r + 1$ para r ímpar.

Consideremos o caso $k = 0$, $p = r$. Este caso corresponde às *fórmulas abertas de Newton-Cotes* quando $f(x, y)$ é uma função só de x. Assim, com $r = 1$ (7.48) reduz-se a

$$\begin{aligned} y_{n+1} &= y_{n-1} + 2hf_n \qquad n \geq 1 \\ T_n &= \tfrac{1}{3}\, h^3 y'''(\xi_n) \end{aligned} \tag{7.50}$$

a que corresponde a *regra do ponto médio.* Com $r = 2$ temos

$$\begin{aligned} y_{n+1} &= y_{n-2} + \tfrac{3}{2}\, h(f_n + f_{n-1}) \qquad n \geq 2 \\ T_n &= \tfrac{3}{4}\, h^3 y'''(\xi_n) \end{aligned}$$

e com $r = 3$ temos

$$\begin{aligned} y_{n+1} &= y_{n-3} + \tfrac{4}{3}\, h(2f_n - f_{n-1} + 2f_{n-2}) \qquad n \geq 3 \\ T_n &= \tfrac{14}{45}\, h^5 y^{(5)}(\xi_n) \end{aligned} \tag{7.51}$$

que é conhecida por *fórmula de Milne.* Nota-se que neste caso o método é de ordem r para r par e de ordem $r + 1$ para r ímpar.

O par preditor-corrector dado pelas fórmulas (7.51) e (7.49) é conhecido por Milne. As fórmulas são ambas de ordem 4. Comparando os respectivos erros de truncatura com os do par Adams (7.38) e (7.44) de ordem 4 somos levados a concluir que as fórmulas de Milne conduzem a melhores resultados do que as de Adams. Esta conclusão, no entanto, é falsa. Apesar do erro de truncatura de (7.49) ser geralmente menos de metade do que a de (7.44) verifica-se normalmente que para x suficientemente grande o erro global é maior quando se utiliza a fórmula de Milne. Iremos constatar este facto para um exemplo considerado a seguir.

7.5 Comparação dos vários métodos

Com a finalidade de comparar os vários métodos apresentados, consideremos o seguinte :

Exemplo 7.4 *Resolver pelos vários métodos apresentados o problema de valor inicial*

$$y' = -y \qquad y(0) = 1$$

Apresentamos na Tabela 7.3 o erro cometido ao utilizar os seguintes métodos de ordem 4: Runge-Kutta (7.30), preditor de Adams-Bashforth (7.38), Preditor-corrector de Adams (7.47), preditor de Milne (7.51) e preditor-corrector de Milne (7.51), (7.49) (com uma única iteração). Dos resultados podemos tirar as seguintes conclusões. O método RK é o que conduz a resultados mais precisos. As fórmulas preditor sem a utilização de um corrector conduzem a resultados piores (no caso do preditor de Milne os resultados chegam a ser desastrosos). A partir de $x \geq 6$ vê-se, da tabela, que o preditor-corrector de Milne é pior do que o de Adams e que o erro no método de Milne para $x \geq 14$ é da ordem de grandeza da própria solução $\exp(-x)$. Notamos ainda que quando h é dividido por 2 os erros vêm divididos por um número superior a 2^4 o que confirma que os métodos são de ordem 4. ■

Deste exemplo e do que vimos atrás podemos tirar as seguintes conclusões: Os métodos de *Runge-Kutta* são auto-iniciáveis e de precisão comparável com os preditor-corrector da mesma ordem (frequentemente são mais precisos). Todavia estes métodos requerem geralmente que se calcule um maior número de vezes a função $f(x, y)$ da ED do que nos métodos preditor-corrector e por conseguinte só são aconselhados quando $f(x, y)$ for uma função simples de computar. Os métodos *preditor-corrector* têm características complementares das apontadas para os métodos RK. Além disso para certos métodos, p.ex., o método de Milne, pode-se verificar uma acumulação de erros que torna o método inútil na determinação da solução para valores de x grandes. Por outro lado, para os métodos preditor-corrector é mais fácil construir um algoritmo que permita controlar o erro de truncatura em cada passo.

Tabela 7.3: $y(x_n) - y_n$

	SOL. EXACTA	RUNGE – KUTTA	
x_n	$y(x_n)$	$h = 0.1$	$h = 0.05$
2	$1.353E-1$	$-2.3E-7$	$-4.8E-9$
4	$1.832E-2$	$-6.4E-8$	$-2.1E-9$
6	$2.479E-3$	$-1.3E-8$	$-3.6E-10$
8	$3.355E-4$	$-2.3E-9$	$-6.2E-11$
10	$4.540E-5$	$-4.0E-10$	$-1.0E-11$
12	$6.144E-6$	$-6.4E-11$	$-1.7E-12$
14	$8.315E-7$	$-1.0E-11$	$-2.7E-13$
16	$1.125E-7$	$-1.6E-12$	$-4.1E-14$

	PRED. A. – BASHF.		PRED. – COR. ADAMS	
x_n	$h = 0.1$	$h = 0.05$	$h = 0.1$	$h = 0.05$
2	$-9.4E-6$	$-5.9E-7$	$1.0E-6$	$5.5E-8$
4	$-2.8E-6$	$-1.7E-7$	$3.1E-7$	$1.6E-8$
6	$-5.8E-7$	$-3.4E-8$	$6.4E-8$	$3.2E-9$
8	$-1.1E-7$	$-6.2E-9$	$1.2E-8$	$5.8E-10$
10	$-1.8E-8$	$-1.1E-9$	$2.0E-9$	$9.8E-11$
12	$-2.9E-9$	$-1.7E-10$	$3.3E-10$	$1.6E-11$
14	$-4.6E-10$	$-2.7E-11$	$5.2E-11$	$2.5E-12$
16	$-7.2E-11$	$-4.2E-12$	$8.1E-12$	$3.9E-13$

	PRED. MILNE		PRED. – COR. MILNE	
x_n	$h = 0.1$	$h = 0.05$	$h = 0.1$	$h = 0.05$
2	$-9.2E-6$	$-4.1E-7$	$3.6E-7$	$1.6E-8$
4	$-1.9E-4$	$-7.7E-6$	$2.0E-7$	$7.4E-9$
6	$-5.1E-3$	$-2.1E-4$	$2.1E-7$	$7.2E-9$
8	$-1.4E-1$	$-5.9E-3$	$3.1E-7$	$1.2E-8$
10	$-3.8E+0$	$-1.7E-1$	$4.9E-7$	$2.0E-8$
12	$-1.1E+2$	$-4.6E+0$	$8.0E-7$	$3.6E-8$
14	$-2.9E+3$	$-1.3E+2$	$1.3E-6$	$6.4E-8$
16	$-7.9E+4$	$-3.6E+3$	$2.1E-6$	$1.1E-7$

7.6 Problemas

1. Mostre que o método de Euler falha ao aproximar a solução $y = (\frac{2}{3}x)^{3/2}, x \geq 0$ do problema $y' = y^{1/3}$, $y(0) = 0$. Explique porquê.

2. Considere o problema de valor inicial $y' = -2y^2$, $0 \leq x \leq 1$, $y(0) = 1$. Obtenha uma majoração do erro para $x = 1$ em termos do passo h quando se utiliza o método de Euler (recorra a (7.17), sabendo que $y \neq 0$).

3. Usando séries de Taylor obtenha o método de ordem 3 para resolver $y' = x - y^2$, $y(0) = 0$.

4. Mostre que o método de RK

$$\begin{aligned} y_{n+1} &= y_n + \frac{h}{6}\,(V_1 + 4V_2 + V_3) \\ V_1 &= f(x_n, y_n) \\ V_2 &= f(x_n + h/2, y_n + h/2\,V_1) \\ V_3 &= f(x_n + h, y_n + 2hV_2 - hV_1) \end{aligned}$$

   é de ordem 3.

5. Para o problema $y' = x - y^2,\ y(0) = 0$ exprima a aproximação y_1 em termos de h ao utilizar os métodos referidos nos problemas 3 e 4.

6. Utilize o par preditor-corrector de Adams de ordem 2 com $h = 0.1$ para obter uma aproximação $y(0.2)$ da solução de $y' = x + y,\ y(0) = 0$ (efectue só uma iteração). Utilize o método de RK de ordem 2 (7.29) para obter uma aproximação de $y(0.1)$. Compare os valores obtidos pelo preditor e corrector.

7. Deduza o preditor (7.51) do método de Milne.

8. Mostre que o erro de truncatura do método explícito do ponto médio (7.50) é

$$T_n = \frac{1}{3}\,h^3 y'''(\xi_n) \qquad \xi_n \in (x_{n-1}, x_{n+1})$$

9. Considere o método de RK de 3 ordem:

$$\begin{aligned} y_{n+1} &= y_n + \tfrac{1}{4}hV_1 + \tfrac{3}{4}hV_3 \\ V_1 &= f(x_n, y_n) \\ V_2 &= f(x_n + \tfrac{1}{3}h, y_n + \tfrac{1}{3}hV_1) \\ V_3 &= f(x_n + \tfrac{2}{3}h, y_n + \tfrac{2}{3}hV_2) \end{aligned}$$

   Mostre que ele coincide com o método de Taylor da mesma ordem para a equação $y' = x + y$.

10. Considere o problema inicial $y' = xy$, $y(0) = 1$. Para um espaçamento genérico h obtenha uma aproximação de $y(2h)$ utilizando o método preditor corrector de Adams de 2 ordem (utilizando só uma iteração). Recorre a fórmula de Euler se necessário.

11. Considere os métodos de Runge-Kutta de 2 ordem no caso de f só ser função de x. Determine a_2 (ver formulário) de modo a que corresponda para este caso um método de 3 ordem.

12. Considere o problema inicial $y' = x + y$, $y(0) = 0$. Para um espaçamento genérico h obtenha uma aproximação de $y(3h)$ utilizando o método de Adams-Bashforth de 3 ordem. Recorra à fórmula de Euler se necessário.

13. Considere o problema inicial $y = (1/y') - x$, $y(0) = 1$. Para um espaçamento genérico h obtenha uma aproximação de $y(2h)$ utilizando o método preditor corrector de Adams de 2 ordem (utilizando só uma iteração). Recorre a fórmula de Euler se necessário.

Bibliografia

Bibliografia básica

1. ATKINSON, K.E., "An Introduction to Numerical Analysis", John Wiley & Sons, New York, 1978.

2. CONTE, S.D. & BOOR, C. de, "Elementary Numerical Analysis", Mc Graw-Hill, London, 1980.

3. RICE, J.R., "Numerical Methods, Software and Analysis", Mc Graw-Hill, London, 1983.

Bibliografia geral

4. BLUM, E.K., "Numerical Analysis and Computation Theory and Practice", Addison-Wesley, 1972.

5. HENRICI, P., "Elements of Numerical Analysis", Wiley, 1964.

6. ISAACSON, P. & KELLER, H.B., "Analysis of Numerical Methods", Wiley, 1966.

7. KINCAID, D. & CHENEY, W., "Numerical Mathematics and computing", Monterey: Brooks, 1985.

8. PRESS, W.H. & FLANNERY, B.P. & TEUKOLSKY, S.A. & VETTERLING, W.T., "Numerical recipes", Cambridge Univ. Press, 1986.

9. RALSTON, A. & RABINOWITZ, P., "A First Course in Numerical Analysis", Mc Graw-Hill, London, 1978.

10. SIBONY, M. & MARDON, J.C., "Analyse Numérique", Hermann, Paris, 1982.

11. STARK, P.A., "Introduction to Numerical Methods", Macmillan, 1970.

12. STOER, J. & BULIRSCH, R., "Introduction to Numerical Analysis", Springer-Verlag, New York, 1980.

13. VANDERGRAFT, J.S., "Introduction to Numerical Computations", Academic Press, 1983.

14. VETTERLING, W.T. & TEUKOLSKY, S.A. & PRESS, W.H. & FLANNERY, B.P., "Numerical recipes example book (FORTRAN)", Cambridge Univ. Press, 1985.

Bibliografia especializada

15. DAVIS, P.J. & RABINOWITZ, P., "Methods of Numerical Integration", Academic Press, 1984.

16. POWELL, M.J.D., "Approximation Theory and Methods", Cambridge University Press, 1981.

17. STROUD, A.H., "Numerical Quadrature and Solution of Ordinary Differential Equation", Springer-Verlag, 1974.

Resolução dos problemas

Teoria dos erros

1. (a) Bem condicionado.
   (b) $w = x/(2+\sqrt{x+4})$.

2. (a) $\delta_{z_1} = \dfrac{-k}{\sqrt{k^2-c}}\delta_k - \dfrac{1}{2}\dfrac{c}{\sqrt{k^2-c}(\sqrt{k^2-c}-k)}\delta_c + \dfrac{1}{2}\dfrac{k^2}{\sqrt{k^2-c}(\sqrt{k^2-c}-k)}\delta_{ar_1} + \dfrac{1}{2}\dfrac{\sqrt{k^2-c}}{\sqrt{k^2-c}-k}\delta_{ar_2} + \dfrac{\sqrt{k^2-c}}{\sqrt{k^2-c}-k}\delta_{ar_3} + \delta_{ar_4}$.
   (b) $z_1 = -\dfrac{c}{\sqrt{k^2-c}+k}$.

3. (a) $\delta_{w_1} = \delta_f + [(x_2-x_1)\delta_{ar_1} - (x_1-x_0)\delta_{ar_2}]/w + \delta_{ar_3}$.
   (b) w_1.

4. (a) $\delta_z = \dfrac{x(1-\cos x)}{z}\delta_x + \dfrac{-\sin x}{z}\delta_{ar_1} + \delta_{ar_2}$.
   (b) Bem condicionado mas instável.
   $\epsilon_z \approx 0.03$.

5. $\delta_z = \dfrac{-y^2}{1-y^2}\delta_y - \dfrac{1}{2}\dfrac{y^2}{1-y^2}\delta_{ar_1} + \dfrac{1}{2}\delta_{ar_2} + \delta_{ar_3}$.

6. $\delta_P = \dfrac{a(\sin a+\cos a)}{\sin a+\cos a+1}\delta_a + \delta_h + \dfrac{\sin a}{\sin a+\cos a+1}\delta_{ar_1} + \dfrac{\cos a}{\sin a+\cos a+1}\delta_{ar_2} + \delta_{ar_3} + \delta_{ar_4}$.

7. $\delta_z = \delta_f + \delta_{ar_1} + \dfrac{-x^2}{x^2+\log x}\delta_{ar_2} + \dfrac{-\log x}{x^2+\log x}\delta_{ar_3} - \delta_{ar_4} + \delta_{ar_5}$.
   $\delta_f = \dfrac{-5x^2+\log x-3}{3(x^2+\log x)}\delta_x$.

8. $\delta_c = \dfrac{a^2}{a^2+b^2}\delta_a + \dfrac{b^2}{a^2+b^2}\delta_b + \beta\cot\beta\cdot\delta_\beta + \dfrac{1}{2}\dfrac{a^2}{a^2+b^2}\delta_{ar_1} + \dfrac{1}{2}\dfrac{b^2}{a^2+b^2}\delta_{ar_2} + \dfrac{1}{2}\delta_{ar_3} + \delta_{ar_4} + \delta_{ar_5} + \delta_{ar_6}$.

9. $\delta_{x_1} = \dfrac{1}{2}\dfrac{\sqrt{a}}{\sqrt{a}-1}\delta_a + \dfrac{\sqrt{a}}{\sqrt{a}-1}\delta_{ar_1} + \delta_{ar_2}$ com $a = 1.0012345$.
   Um algarismo significativo.
   $x_1 = \dfrac{0.0012345}{\sqrt{1.0012345}+1}$.

11. (a) $\delta_{x_{m+1}} = \dfrac{x_m^2 - c}{x_m^2 + c}\delta_{x_m} + \dfrac{c}{x_m^2 + c}\delta_c + \dfrac{c}{x_m^2 + c}\delta_{ar_1} + \delta_{ar_2} + \delta_{ar_3}$.

    (b) Seis algarismos significativos.

12. (a) $\delta_f = -\dfrac{x}{\sqrt{x^2 - 1}}\delta_x$.
    Bem condicionado.
    Instável.

    (b) Um algarismo significativo.
    $z = -\dfrac{1}{x + \sqrt{x^2 + 1}}$.

13. (a) $\delta_{w_2} = \delta_f + \dfrac{x}{x - y}\delta_{ar_1} + \dfrac{-y}{x - y}\delta_{ar_2} + \delta_{ar_3}$.

14. (a) $\delta_{w_3} = \delta_f + \delta_{ar_1} + \dfrac{x}{x - y}\delta_{ar_2} + \dfrac{-y}{x - y}\delta_{ar_3} + \delta_{ar_4}$.

Equações não lineares

1. (a) i. $g(x) = \sqrt[3]{0.6x}$.
    ii. $K = g'(0.7) = 0.357$, $|z - x_0| < 0.3$; 9 iterações.
    iii. $x_0 = 1$.
    iv. $a_2 = 0.7772138$.

    (b) $x_0 = -1$, $x_1 = -0.83333$.

2. (a) 1 raiz em (-1,0), 1 raiz em (1,2), 2 raízes complexas conjugadas.

    (b) $x_0 = -1$, $x_1 = -0.8$.

    (c) $x_1 = 1.21896$.

3. (a) $x_0 \neq \sqrt{a}$, $x_{2k+1} = a/x_0$, $x_{2k} = x_0$.

    (b) $x_0 = 1$, $x_1 = 2$, $x_2 = 1$, $a_2 = 1.5$.
    $x_0 = 1.5$, $x_1 = 4/3$, $x_2 = 1.5$, $a_2 = 17/12 = 1.41(6)$.

4. (a) $z = 2$.

5. 2 raízes > 0 e 1 raiz < 0 ou 2 complexas conj. e 1 raiz < 0.

7. $x_0 = 2$, $x_1 = 1.444445$.

8. (a) $x_1 = 0.5 - 0.3125/4 = 0.421875$.

    (b) Falsa pos. $x_4 = 0.369241$; secante $x_4 = 0.428657$.

9. (a) 2 raízes.
    $[0.7, 0.8]$.

    (b) $x_2 = -1.630717$. $x_{-1} = -1$, $x_0 = -2$.

10. 7.11287×10^{-5}.

11. (a) 2 ou 0 raízes em (-1,0), 2 ou 0 raízes em (0,1).
    (b) $x_1 = 0.3435$.
    1 alg. sign.
    (c) $x_0 = 0.9$, $x_1 = 0.865387$.

15. $A = (1 + \log 3)/3$, sendo $z = 1/\sqrt[3]{3}$.

16. Acelerar a convergência de métodos iterativos de convergência linear.

17. (a) 1 raiz em (0,0.8) e 1 raiz em (0.8,1).
    (b) $x_1 = 0.339981$.
    É garantida a convergência.
    (c) $x_1 = 0.861064$.

21. 0.00243.

22. 17/12.
    $x_{-1} = 0$, $x_0 = 2$.

24. (b) $x_0 = 2$. $x_1 = 1.4$.
    (c) $A \in [-1/11, -1/40]$.

Sistemas de equações

1. (a)

$$\mathbf{L} = \begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 3/2 & 7/6 & 1 \end{bmatrix} \quad \mathbf{U} = \begin{bmatrix} 2 & 2 & 1 \\ 0 & -6 & 3 \\ 0 & 0 & 1 \end{bmatrix} \quad \mathbf{p} = \begin{bmatrix} 3 \\ 2 \\ 1 \end{bmatrix}$$

   (b) $\mathbf{x} = [1/6, -1/6, 0]^T$.

3. **d**

4. (a) $p = 2$, $r = 0$, $q = 3$, $s = 1$. $\mathbf{x} = [0, 1, 0, 0, 0]^T$.

5. Método da correcção residual.

6. (a) $[1, 1, -1/2]^T$.
   (b) G.S.: $\mathbf{x}^{(1)} = [5, -1, -5/2]$. J.: $\mathbf{x}^{(1)} = [5, -1, 0]$.

7. $\mathbf{x}^{(1)} = [-2, 5, -3/2]$. $\mathbf{x}^{(2)} = [-2, -5, -7/2]$.

8. (b) i. $\forall i = 1, \ldots, n$, $\sum_{j=1, j\neq i}^{n} |a_{ij}/a_{ii}| \leq 1/2$.
   ii. $\beta = 2/3$.

9. (a) Não.
   $\mathbf{x}^{(1)} = [-2.125, 0.35, 2.64]^T$.

(b) Choleski:

$$\mathbf{L} = \begin{bmatrix} 2 & 0 & 0 \\ -1 & 3 & 0 \\ 1/2 & 0 & 1 \end{bmatrix} \quad \mathbf{g} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \quad \mathbf{x} = \begin{bmatrix} -1/4 \\ 0 \\ 1 \end{bmatrix}$$

13. $\mathbf{x}^{(1)} = [-2, 1, -1/2, 1]^T$. $\mathbf{x}^{(2)} = [-2, -5, -11/2, 3]^T$.

14. (b)

$$\mathbf{L} = \begin{bmatrix} -4 & 0 & 0 \\ 2 & 9 & 0 \\ -2 & 0 & 4 \end{bmatrix} \quad \mathbf{U} = \begin{bmatrix} 1 & -1/2 & 1/2 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad \mathbf{g} = \begin{bmatrix} -1/4 \\ -1/18 \\ 1/8 \end{bmatrix} \quad \Delta\mathbf{x} = \begin{bmatrix} -49/144 \\ -1/18 \\ 1/8 \end{bmatrix} \quad \mathbf{x}^{(1)} =$$

(c) $\mathbf{x}^{(1)} = [-1/4, 1/16, -3/16]^T$.

16. $a = 1$, $b = 13$, $c = 8$.

19. (a)

$$\mathbf{J}(\mathbf{x}^{(0)}) = \begin{bmatrix} 2 & 3 & 0 \\ 0 & 2 & 1 \\ 2 & 0 & 1 \end{bmatrix} \quad \mathbf{L} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 3/2 & 1 \end{bmatrix} \quad \mathbf{U} = \begin{bmatrix} 2 & 0 & 1 \\ 0 & 2 & 1 \\ 0 & 0 & -5/2 \end{bmatrix} \quad \mathbf{p} = \begin{bmatrix} 3 \\ 2 \\ 1 \end{bmatrix}$$

(b) $\mathbf{x}^{(1)} = \Delta\mathbf{x}^{(0)} = [-1/5, -1/5, 2/5]^T$.

Interpolação polinomial

1. (a) $p_2(x) = 2.625(x+1)x - 4(x+2)x + 0.75(x+2)(x+1)$.

2. (a) $y = -6$.
   $a_2 = 1$.
   (b) $p_2(x) = 6(x+1)x - 4(x+2)x - (x+2)(x+1)$.
   Sim.

3. (a) $p_3(5) = 40$.
   (b) 1.7083332.

4. (a) $p_2(4.5) = 1.515625$.
   (b) (-1,-4), (1,-0.5) e (3,1).

6. (a) Polinómios diferentes.
   (b) 1.
   (c) 1.

7. (a) $p_2(4) = 1.6875$.
   (b) $q_2(0) = 2.114286$.

8. $\Delta^3 f(-3) = 3!\ 2^3\ 2.5 = 120.$
   $\beta = 63.$

9. (a) $p_2(x) = -87.5(x+1)x + 3(x+2)x + 0.5(x+2)(x+1).$
   $p_2(2) = -495 \neq f(2).$
   (b) $q_2(0) = -0.254162.$

10. (a) $p_2(0.67) = 0.6564754.$
    (b) i. $|e_2(0.67)| \leq \frac{2^3}{3!}|(0.67 - -0.4)(0.67 - 0.6)(0.67 - 0.8)| = 3.277 \times 10^{-3}.$
    ii. $\frac{h^2}{8} < 10^{-4} \Rightarrow h < 2.82843 \times 10^{-2}.$

11. $\frac{e\ h^2}{8} < 10^{-6} \Rightarrow h < 1.715528 \times 10^{-3}.$

12. (a) $p_4(x) = \frac{1}{4}(x^2-1)x(x-2) - \frac{1}{3}(x^2-4)x(x-1) - \frac{\alpha}{6}(x^2-4)x(x+1).$
    $\alpha = -1/2.$
    (b) $q_2(3) = -1.375.$

14. $|e_1(x)| \leq \frac{e^2\ 0.01^2}{8} = 9.2364 \times 10^{-5}.$
    $\frac{e^2\ h^2}{8} < 10^{-5} \Rightarrow h < 3.3 \times 10^{-3}.$

Aproximação segundo os mínimos quadrados

1. $g(x) = 2.1x^2 + 0.9.$

2. $g(x) = -18.44146/(x+2) + 4.65665.$

4. $g(x) = -25.584675/(x+4) + 4.5766112.$

5. (b)

$$\begin{bmatrix} 4 & 0 & 16/3 \\ 0 & 16/3 & 0 \\ 16/3 & 0 & 64/5 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} -172/5 \\ 1024/9 \\ -2192/21 \end{bmatrix}$$

6. $-\sin(\frac{\pi}{2}x) - \cos(\frac{\pi}{2}x).$

7. $g(x) = 25/34x^2 - 3/2x.$

8. $g(x) = 1.534749/x + 0.060113x^2.$

Integração numérica

1. (b) $I_2(f) = -44.$
   (c) $I(q_2) = I_2(q_2),\ I(x^3) = I_2(x^3) = -20 \Rightarrow I(q_3) = I_2(q_3).$ Logo a fórmula é pelo menos de grau 3.

2. (a) $I_4(erf) = 0.3270081$.
   (b) $|E_4(erf)| \leq \frac{0.2^4\ 0.8}{180}\ 3^3 = 1.92 \times 10^{-4}$.
3. (a) $I_2(f) = \frac{1}{12}f(-1) + \frac{11}{6}f(1) + \frac{1}{12}f(3) = -\frac{7}{6}$.
   (b) $m = 3$.
4. (a) RT=9; RS=26/3.
   (c) $\frac{26}{3} - 9 = -\frac{1}{3}$.
5. (b) i. RT: 10.75; RS: 11.(6)
6. (a) $A_0 = A_2 = 1/12$, $A_1 = 11/6$.
7. (a) RS: 16.25.
   (b) RS: 16.25.
9. $A_1 = -2/3$, $A_2 = A_3 = 4/3$.
   3.2634582.
10. R.T. pol. de grau 1.
    R.S. pol. de grau 2.
11. (a) Nenhuma; par.
    (b) Nenhuma; uniforme.
12. (a) R.S.: 18.13288.
    (b) $E_4(g) < \frac{1}{3}(\frac{9}{2})^4(4)^{-\frac{7}{2}} = 1.067887$.
13. $I_3(f) = 5/9f(2 - \sqrt{3/5}) + 8/9f(2) + 5/9f(2 + \sqrt{3/5})$.
15. 4 pontos.
16. (a) i. Polinómio de grau ≤ 3.
        ii. Polinómio de grau ≤ 2.
        iii. Polinómio de grau ≤ 5.
    (b) $I_2(f) = 5/3f(0) + 2/9f(-1) + 1/9f(2)$.
18. (a) $A_0 = (2 + \sqrt{2})/4$, $A_1 = (2 - \sqrt{2})/4$.
19. $I - I_{2n} \approx (I_{2n} - I_n)/3$. $|I_{i2} - I_{i1}| \leq 3\epsilon_i = 3\frac{b_i - a_i}{b - a}\epsilon$.
20. (a) $A_1 = \frac{1}{3x_1^2}$; $A_0 = 2 - 2A_1$.
    (b) $x_1 = \sqrt{3/5}$.
    (c) $x_1 = \sqrt{2}/2$.
21. R.T.: 6; R.S.: 4.
23. Grau 3.