1 Representação de Números e Erros

1.1 Sistemas de Vírgula Flutuante

Para podermos fazer cálculos é necessário antes de mais escolhermos um sistema para representar os números com que trabalhamos. Supondo que vamos trabalhar com números reais, os sistemas habitualmente utilizados para os representar são os **sistemas de vírgula (ponto) flutuante.** Assim, vamos começar por definir matematicamente estes sistemas.

Seja β um número natural, diferente de 1, a que chamaremos **base** do sistema. A base é o número de dígitos diferentes que usamos para representar os números. A base mais corrente é a decimal, em que se usam dez dígitos (os algarismos); quando usamos esta base, temos $\beta = 10$. No entanto, se atentarmos à forma como os números são representados internamente nos computadores e outros sistemas de cálculo, verificaremos que a base aí utilizada é a binária, ou seja, $\beta = 2$, já que, por razões técnicas, é conveniente trabalhar apenas com dois símbolos diferentes: 0 e 1. Nesse caso, cada símbolo representado designa-se por **bit.** Uma vez escolhida a base, qualquer elemento x do sistema de vírgula flutuante representa-se na forma

$$x = \sigma \times 0.a_1a_2a_3...a_n \times \beta^t \tag{1.1}$$

onde σ representa o sinal ($\sigma = \pm 1$), a_i são dígitos da base considerada e t é um número inteiro. A sucessão de dígitos $a_1a_2a_3...a_n$, onde $a_1 \neq 0$, designa-se *mantissa.* Assim, além da base, qualquer sistema de vírgula flutuante caracteriza-se pelo **comprimento da mantissa,** isto é, o número n de dígitos que a compõem. Finalmente, um sistema de vírgula flutuante caracteriza-se pelos limites inferior e superior do expoente t, que representaremos respectivamente por t_1e t_2. Chegamos assim à seguinte definição.

Definição 1. Sistema de vírgula flutuante com base β e n dígitos na mantissa:

$$VF(\beta, n, t_1, t_2) = \{x \in I\!R : x = \sigma \times 0.a_1a_2a_3...a_n \times \beta^t, \sigma = \pm 1, a_1 \neq 0, t \in [t_1, t_2]\} \cup \{0\}.$$

De acordo com esta definição, como é natural, o número 0 pertence a qualquer sistema VF , embora formalmente ele não possa ser representado na forma (1.1), já que o primeiro dígito da mantissa, por definição, é diferente de zero.

Exemplo 1.1. Calculadora em que os números são representados na base decimal, com 12 dígitos na mantissa e com o expoente entre -99 e 99. Neste caso, o sistema utilizado é $VF(10, 12, -99, 99)$.

Exemplo 1.2. Computador em que os números são representados na base binária, sendo reservados 56 bits para a mantissa e 8 bits para o expoente. Dos 8 bits do expoente 7 são reservados ao seu valor absoluto e um ao sinal, pelo que o seu valor pode variar entre $-2^7+1 = -127$ e $2^7-1 = 127$. Logo, o sistema considerado é $VF(2, 56, -127, 127)$. Note-se que, devido à condição $a_1 \neq 0$, no caso do sistema binário obtém-se $a_1 = 1$, qualquer que seja o número representado. Isto faz com que o primeiro dígito da mantissa seja supérfluo. Na prática, esse dígito é usado para representar o sinal da mantissa.

Propriedades dos sistemas de vírgula flutuante:

1. Qualquer sistema VF é finito. Determinemos o número de elementos positivos do sistema $VF(\beta, n, t_1, t_2)$. O número de mantissas diferentes é $\beta^{n-1}(\beta - 1)$ (o primeiro dígito da mantissa não pode ser 0, por definição). O número de expoentes diferentes é $t_2 - t_1 + 1$. Logo, o número de elementos do sistema $VF(\beta, n, t_1, t_2)$ é $\beta^{n-1}(\beta - 1)(\ t_2 - t_1 + 1)$ (no caso do exemplo 1.1, temos $9 \times 199 \times 10^{11} \approx 1.8 \times 10^{14}$ elementos, enquanto que no exemplo 1.2 o número de elementos é $255 \times 2^{55} \approx 0.92 \times 10^{19}$).

2. Em qualquer sistema de vírgula flutuante existe um elemento **máximo**, cujo valor é $M = (1 - \beta^{-n})\beta^{t_2}$(no caso do exemplo 1.1, temos $M = (1 - 10^{-12})10^{99} \approx 10^{99}$, enquanto que para o exemplo 1.2 temos $M = (1 - 2^{-155})2^{127} \approx 1,70 \times 10^{38}$).

3. Em qualquer sistema de vírgula flutuante existe um elemento com o **mínimo valor absoluto**, igual a $m = \beta^{-1}\beta^{t_1} = \beta^{t_1-1}$(no caso do exemplo 1.1, temos $m = 10^{-100}$, enquanto que para o exemplo 1.2 temos $m = 2^{-128} \approx 2,9 \times 10^{-39}$).

1.2 Representação de números em sistemas de vírgula flutuante

Visto que qualquer sistema VF é finito, ele contém apenas uma pequena parte dos números reais. Quando um número real não pertence ao sistema VF considerado, para o representar nesse sistema é necessário fazer uma certa aproximação, chamada **arredondamento.** Denotemos $fl(x)$ a representação do real x no sistema VF considerado. Se $x \in VF(\beta, n, t_1, t_2)$ então $fl(x) = x$ (diz-se que x tem representação exacta nesse sistema). Caso contrário, isto é, se $x \notin VF(\beta, n, t_1, t_2)$, há que escolher $fl(x)$ e essa escolha pode ser feita de diferentes maneiras. Para melhor compreender este processo, suponhamos que $x = \sigma \times 0.a_1a_2a_3...a_na_{n+1}... \times \beta^t$. Note-se que

qualquer número real pode ser representado nesta forma, sendo que a mantissa, regra geral, é uma dízima infinita. Segundo a forma mais simples de arredondamento, o **arredondamento por corte,** escolhe-se:

$$fl(x) = \sigma \times 0.a_1a_2a_3...a_n \times \beta^t.$$

Outra forma mais sofisticada de determinar $fl(x)$ consiste em defini-lo pela fórmula

$$fl(x) = \begin{cases} \sigma \times 0.a_1a_2a_3...a_n \times \beta^t, \ se \ a_{n+1} < \beta/2; \\ \sigma \times (0.a_1a_2a_3...a_n + \beta^{-n}) \times \beta^t, \ se \ a_{n+1} \geq \beta/2. \end{cases}$$

Esta forma de aproximação chama-se **arredondamento simétrico.** Este tipo de arredondamento envolve um erro igual ao do arredondamento por corte, no caso de $a_{n+1} < \beta/2$, ou menor, no caso em que $a_{n+1} \geq \beta/2$.

Note-se que, ao considerar um certo sistema VF, nem todos os números reais podem ser representados nele, mesmo com arredondamento. Os números x, tais que $|x| > M$ ou $|x| < m$, não têm qualquer representação no sistema, pelo que ao tentar representá-los ocorrem situações de erro. No primeiro caso, essas situações deignam-se por *overflow,* enquanto no segundo caso são referidas como *underflow* .

Exercício 1.1 Para cada um dos seguintes números reais obtenha (caso seja possível) a sua representação no sistema $VF(10, 3, -99, 99)$,utilizando arredondamento simétrico:

a) $x = 10$
b) $x = 0.001235$
c) $x = 1001$
d) $x = 1/3$
d) $x = 10^{100}$
e) $x = 10^{-101}$

Resposta:

x	**fl(x)**
100	$0,100 \times 10^3$
$0,001235$	$0,124 \times 10^{-2}$
-1001	$-0,100 \times 10^4$
1/3	$0,333$
10^{100}	não tem representação (overflow)
10^{-101}	não tem representação (underflow)

1.3 Erros de arredondamento

Quando se aproxima um número real x pela sua representação em vírgula flutuante, $fl(x)$, comete-se um erro geralmente designado por *erro de arredondamento:*

$$e_{ar} = fl(x) - x.$$

Outras grandezas relacionadas são o *erro de arredondamento absoluto*

$$|e_{ar}| = |x - fl(x)|$$

e o *erro de arredondamento relativo:*

$$|\delta_{ar}| = \frac{|x - fl(x)|}{|x|}.$$

Para caracterizar a precisão com que os números reais são aproximados num sistema VF utiliza-se o conceito de *unidade de arredondamento.* A unidade de arredondamento do sistema VF é um número real u, tal que

$$|\delta_{ar}| \leq u, \qquad \forall x \in I\!R, m \leq |x| \leq M.$$

O valor de u depende, evidentemente, dos parâmetros do sistema VF considerado, mais precisamente, de n e β. Logicamente, para o mesmo valor de β, a unidade de arredondamento será tanto mais pequena quanto maior for n, isto é, quanto mais dígitos utilizarmos para representar os números tanto menor será o erro de arredondamento relativo. Para analisarmos esta questão em pormenor, consideremos um número real x arbitrário e representemo-lo na forma $x = \sigma \times 0.a_1a_2a_3...a_na_{n+1}... \times \beta^t$. Para simplificar, comecemos por considerar o caso do arredondamento por corte. Como já vimos, neste caso $fl(x) = \sigma \times 0.a_1a_2a_3...a_n \times \beta^t$. Por conseguinte, o erro de arredondamento absoluto é

$$|e_{ar}| = |x - fl(x)| = 0,00...0a_{n+1}... \times \beta^t < \beta^{t-n}.$$

No que diz respeito ao erro de arredondamento relativo, temos

$$|\delta_{ar}| = \frac{|x - fl(x)|}{|x|} < \frac{\beta^{t-n}}{\beta^{t-1}} = \beta^{1-n}.$$

Por conseguinte, qualquer que seja x,tal que $m \leq |x| \leq M$,verifica-se

$$|\delta_{ar}| < \beta^{1-n}.$$

Logo, podemos afirmar que, no caso do arredondamento por corte, a unidade de arredondamento é

$$u = \beta^{1-n}.$$

Raciocínios semelhantes levam-nos à conclusão que, no caso do arredondamento simétrico, a unidade de arredondamento é

$$u = \frac{1}{2}\beta^{1-n}.$$

Por exemplo, no caso do sistema $VF(10, 12, -99, 99)$, e assumindo que o arredondamento é simétrico, temos $u = 0,5 \times 10^{-11}$.

Exercício 1.2. Para cada um dos números referidos no exercício 1.3, considerando de novo o sistema $VF(10, 3, -99, 99)$, determine o erro de arredondamento absoluto e o erro de arredondamento relativo (nos casos em que existe representação). Compare este último com a unidade de arredondamento do sistema.

Resposta:

x	$fl(x)$	$\lvert e_{ar}\rvert$	$\lvert\delta_{ar}\rvert$
100	$0,100 \times 10^3$	0	0
$0,001235$	$0,124 \times 10^{-2}$	$0,5 \times 10^{-5}$	$0,004$
-1001	-$0,100 \times 10^4$	1	$0,001$
$1/3$	$0,333$	$0,33 \times 10^{-3}$	$0,001$

A unidade de arredondamento, neste caso, é $0,5 \times 10^{-2} = 0,005$, pelo que todos os números considerados têm erro de arredondamento relativo inferior a este valor, como seria de esperar.

1.4 Propagação dos Erros

Sejam $\widetilde{x}$ e $\widetilde{y}$ valores aproximados dos números reais x e y, respectivamente. Denotaremos por e_x e $|\delta_x|$ o erro e o erro relativo de $\widetilde{x}$, repectivamente:

$$e_x = \widetilde{x} - x, \ |\delta_x| = \left|\frac{\widetilde{x} - x}{x}\right|.$$

De modo análogo se definem o erro e o erro relativo de $\widetilde{y}$. Suponhamos que $\widetilde{x}$ e $\widetilde{y}$ são dados de um cálculo que pretendemos efectuar. O nosso objectivo é determinar qual o efeito dos erros destes dados no resultado. Para começar, consideremos o caso das operações aritméticas.

Adição/Subtracção

Representemos por $e_{x\pm y}$o erro de $\widetilde{x} \pm \widetilde{y}$. Note-se que

$$\widetilde{x} \pm \widetilde{y} = (x + e_x) \pm (y + e_y) = (x \pm y) + (e_x \pm e_y).$$

Por conseguinte, para o erro de $\widetilde{x} \pm \widetilde{y}$ temos

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

e, para o erro absoluto,

$$|e_{x\pm y}| \leq |e_x| + |e_y|.$$

Quanto ao erro relativo, podemos escrever

$$|\delta_{x\pm y}| = \frac{|e_x \pm e_y|}{|x \pm y|} \leq \frac{|x\delta x| + |y\delta y|}{|x \pm y|}. \tag{1.2}$$

Daqui resulta que, se $x \pm y$ for próximo de zero, então o erro relativo de $x \pm y$ pode ser muito maior que o dos dados. Voltaremos a este assunto mais tarde.

Multiplicação

No caso da multiplicação, temos

$$\widetilde{x}.\widetilde{y} = (x + e_x).(y + e_y) = (x.y) + ye_x + xe_y + e_xe_y.$$

Admitindo que e_x e e_y são grandezas pequenas, o seu produto pode ser desprezado na expressão anterior, pelo que obtemos

$$e_{x.y} = \widetilde{x}.\widetilde{y} - x.y \approx ye_x + xe_y.$$

Logo, para o erro relativo do produto, resulta

$$|\delta_{x.y}| = \frac{|e_{x.y}|}{|x.y|} \approx \frac{|ye_x + xe_y|}{|x.y|} \leq |\delta x| + |\delta y|. \tag{1.3}$$

Divisão

Para deduzir a fórmula do erro do quociente, suponhamos que os valores de e_x e e_y são desprezáveis em comparação com x e y, respectivamente. Podemos então fazer a seguinte aproximação:

$$\frac{\widetilde{x}}{\widetilde{y}} = (x+e_x)\frac{1}{y}\frac{1}{1+\frac{e_y}{y}} \approx (x+e_x)\frac{1}{y}\left(1-\frac{e_y}{y}\right) = \frac{x}{y} + \frac{ye_x - xe_y}{y^2}.$$

Daqui resulta que

$$e_{x/y} = \frac{\widetilde{x}}{\widetilde{y}} - \frac{x}{y} \approx \frac{ye_x - xe_y}{y^2}.$$

Quanto ao erro relativo do quociente, obtém-se

$$|\delta_{x/y}| = |e_{x/y}|\frac{|y|}{|x|} \approx \frac{|ye_x - xe_y|}{y^2}\frac{|y|}{|x|} \leq \frac{|e_x|}{|x|} + \frac{|e_y|}{|y|} = |\delta x| + |\delta y|. \tag{1.4}$$

Os cálculos que acabámos de efectuar mostram que, no caso da multiplicação e da divisão, o erro relativo dos resultados é da mesma ordem que o erro relativo dos dados, ou seja, destas operações não resulta uma perda de precisão. Já no caso da adição e da subtracção, como vimos, tal perda de precisão pode ocorrer. Esse fenómeno designa-se cancelamento subtractivo e é ilustrado pelo exercício que se segue.

Exercício 1.3. Considere os números $x = \pi$ e $y = 2199/700$.

1. Determine $\tilde{x}$ e $\tilde{y}$ com 4 dígitos na mantissa, usando arredondamento simétrico. Obtenha ainda $\tilde{x} - \tilde{y}$.

   Solução.

   $$\begin{array}{ll} x = 0.3141592\cdots \times 10; & \widetilde{x} = fl(x) = 0.3142\cdots \times 10; \\ y = 0.3141428\cdots \times 10; & \widetilde{y} = fl(y) = 0.3141\cdots \times 10. \end{array}$$

   Logo, $\tilde{x} - \tilde{y} = 0.1 \times 10^{-2}$.

2. Calcule os erros absolutos e relativos de $\tilde{x}$ e $\tilde{y}$. Comente.

   Solução.

Dado	Erro absoluto	Erro relativo
x	$\lvert e_x\rvert = \lvert x - \widetilde{x}\rvert = 0.41 \times 10^{-3}$	$\lvert\delta x\rvert = \lvert e_x\rvert/\lvert x\rvert = 0.131 \times 10^{-3}$
y	$\lvert e_y\rvert = \lvert y - \widetilde{y}\rvert = 0.43 \times 10^{-3}$	$\lvert\delta y\rvert = \lvert e_y\rvert/\lvert y\rvert = 0.137 \times 10^{-3}$.

   Como era de esperar, os erros de arredondamento relativos dos dados são inferiores à unidade de arredondamento do sistema que, neste caso, é $u = 0.5 \times 10^{1-4} = 0.5 \times 10^{-3}$.

3. Represente os números x e y em ponto flutuante, mas com 6 algarismos na mantissa. Com base nestas novas aproximações, calcule de novo $\tilde{x} - \tilde{y}$ e comente.

   Solução. Neste caso temos:

$$\begin{aligned} x &= 0.3141592\cdots \times 10; \qquad & \widetilde{x} &= fl(x) = 0.314159\cdots \times 10; \\ y &= 0.3141428\cdots \times 10; \qquad & \widetilde{y} &= fl(y) = 0.314143\cdots \times 10. \end{aligned}$$

   Logo, $\tilde{x} - \tilde{y} = 0.16 \times 10^{-3}$, o que é muito diferente do valor obtido na alínea 1. Isto sugere que, na alínea 1, houve uma perda de precisão, resultante de cancelamento subtractivo.

4. Tomando como valor exacto da diferença o resultado da alínea anterior, determine o erro relativo do valor de $\tilde{x}-\tilde{y}$, obtido na alínea 1. Se usasse a estimativa (1.2) para o erro relativo da diferença, chegaria à mesma conclusão?

   Solução. Comparando os resultados da alínea 1 e da alínea 3, temos

$$|\delta_{x-y}| = \frac{|e_{x-y}|}{|x-y|} \approx \frac{0.001 - 0.00016}{0.00016} = 5.25.$$

   Vemos que o erro relativo do resultado da alínea 1 é muito superior à unidade, o que significa uma perda total de precisão. Por outro lado, se usássemos a estimativa (1.2), teríamos

$$|\delta_{x-y}| \leq \frac{|x\delta x| + |y\delta y|}{|x-y|} \approx \frac{0.00084}{0.00016} = 5.25,$$

   ou seja, neste caso verifica-se a igualdade na relação (1.2).

1.5 Estabilidade de algoritmos

Quando se efectua um cálculo, geralmente efectua-se por passos. Assim, o erro cometido em cada passo acumula-se com os erros dos passos anteriores. Em resultado, o erro do resultado final pode ser muito maior do que o erro de cada passo isolado.

O conjunto dos passos que levam à obtenção de um dado resultado designa-se *algoritmo*. O mesmo resultado pode ser obtido, em princípio, através de algoritmos diferentes. No entanto, os erros propagam-se de forma diferente em cada algoritmo. Por isso, os resultados que se obtêm, para o mesmo problema, através de algoritmos diferentes podem divergir significativamente. Surge assim a definição de *estabilidade numérica*.

Definição. Um algoritmo diz-se estável (ou numericamente estável) para um certo conjunto de dados se, a pequenos valores dos erros relativos de arredondamento dos dados e da unidade de arredondamento do sistema corresponderem pequenos valores do erro relativo do resultado.

O exercício que se segue ilustra o conceito de estabilidade numérica.

Exercício 1.3 Considere a função real de variável real

$$f(x) = \frac{1 - \cos x}{x^2}. \tag{1.5}$$

1. Supondo que utiliza um sistema de vírgula flutuante com 10 dígitos na mantissa e arredondamento simétrico, calcule $f(10^{-6})$ aplicando a fórmula (1.5).

   Solução. A fórmula (1.5) pode ser aplicada em 3 passos. Sendo $x = 10^{-6}$, temos

$$\begin{aligned} z_1 &= cosx = 1; \\ z_2 &= 1 - z_1 = 0; \\ z_3 &= \tilde{f}(x) = \tfrac{z_2}{x^2} = 0. \end{aligned}$$

2. Obtenha uma aproximação de $f(10^{-6})$, utilizando o desenvolvimento de f em série de Taylor, em torno de $x = 0$.

   Solução. Como é sabido, para valores de x próximos de zero, a função $cos(x)$ admite o seguinte desenvolvimento em série de Taylor:

$$cosx = 1 - \frac{x^2}{2} + \frac{x^4}{4!} + O(x^6),$$

   Daqui obtém-se facilmente que

$$f(x) = \frac{1 - cos\ x}{x^2} = \frac{1}{2} - \frac{x^2}{4!} + O(x^4). \tag{1.6}$$

   Utilizando a fórmula (1.6), num sistema VF com 10 dígitos, obtém-se $f(10^{-6}) = 0.5000000000$.

3. Sabendo que $1 - \cos x = 2\ \sin^2(x/2)$, calcule $f(10^{-6})$ utilizando uma nova fórmula para f.

   Solução. Transformando a fórmula (1.5) obtém-se

$$f(x) = \frac{1 - cos^2 x}{x^2} = \frac{2}{x^2}\, sin^2(x/2). \tag{1.7}$$

À semelhança do que fizemos na alínea 1, apliquemos a fórmula (1.7) em 5 passos:

$$\begin{gathered} w_1 = x/2 = 0.5 \times 10^{-6}; \\ w_2 = sen(w_1) = 0.5 \times 10^{-6} \\ w_3 = w_2^2 = 0.25 \times 10^{-12}; \\ w_4 = w_3/x^2 = 0.25; \\ w_5 = \bar{f}(x) = 2w_4 = 0.5. \end{gathered}$$

4. Compare os valores obtidos nas alíneas anteriores e classifique os algoritmos quanto à estabilidade.

   Solução. Verifica-se que o valor obtido na alínea 3 é uma boa aproximação de $f(10^{-6})$, já que coincide com o valor dado pela série de Taylor (em todos os 10 dígitos). Pelo contrário, o valor obtido pelo algoritmo da alínea 1 dava uma má aproximação (nem um único dígito correcto). Este facto deve-se apenas aos erros de arredondamento cometidos, já que as fórmulas usadas são equivalentes e, se trabalhássemos com valores exactos, obteríamos em ambos os casos o mesmo resultado. Esta situação pode interpretar-se do seguinte modo: para valores de x próximos de 0 o algoritmo considerado em 1) é instável, enquanto o algoritmo considerado em 3) é estável.

2 Métodos Numéricos para Equações Não Lineares

2.1 Localização de raizes de uma função

Seja f uma função real, definida num certo intervalo $[a, b]$. O ponto $z \in [a, b]$ diz-se um zero ou uma raiz de f se e só se $f(z) = 0$. No caso de $f'(z) \neq 0, z$ diz-se um zero simples. Se $f'(z) = 0$, z diz-se um zero múltiplo. Mais precisamente, se $f \in C^k(z)$ e se $f'(z) = f''(z) = ... = f^{(k-1)}(z) = 0, f^{(k)}(z) \neq 0$,então z diz-se um zero de multiplicidade k da função f.

Exemplo 2.1. Seja **f** um polinómio de grau n. Então, de acordo com o teorema fundamental da álgebra, f tem, no total, n raizes em C (somando as multiplicidades). Em particular, se tivermos $f(x) = x^2 + 2x + 1, f$ tem uma raiz de multiplicidade dois (raiz dupla) em $z = -1$. De facto, verifica-se $f(-1) = 0$; visto que $f'(x) = 2x + 2$, temos$f'(-1) = 0$; além disso, como $f''(x) = 2$, $f''(-1) \neq 0$. Se considerarmos o polínómio de terceiro grau $f(x) = x^3 - x = x(x-1)(x+1)$, verificamos facilmente que ele tem três

raizes simples: $z_1 = -1, z_2 = 0, z_3 = 1$. Quanto ao polinómio $f(x) = x^3 + 1$, tem apenas uma raiz real ($z_1 = -1$) e duas raizes complexas ($z_{2,3} = \frac{1 \pm \sqrt{3}i}{2}$).

De um modo geral, a determinação das raizes de um polinómio de grau n (ou zeros de uma equação algébrica) é um problema muito complexo que ocupou os matemáticos de várias civilizações. Desde o princípio do século XX sabe-se, graças a Abel, que não existem fórmulas resolventes para equações algébricas de grau superior a 4. Mais precisamente, não é possível, através de uma fórmula geral, exprimir todas as raizes de uma equação algébrica de grau superior a 4 através dos seus coeficientes.

Este exemplo ilusta a importância dos métodos numéricos para a resolução de equações. Até no caso de equações relativamente simples, como as equações algébricas, é impossível clacular as suas raizes unicamente através de fórmulas analíticas. Por outro lado, mesmo nos casos em que existem fórmulas resolventes, estas são por vezes tão complexas que se torna mais eficiente determinar as raizes a partir de um método numérico. Tal é o caso de algumas equações algébricas de terceiro e quarto grau, por exemplo. Naturalmente, isso pressupõe que se escolha um método adequado à equação considerada.

Para tratar o problema do cálculo numérico de todas as raizes da função f é necessário, em primeiro lugar, localizar as raizes, isto é, determinar, para cada raiz, um intervalo que a contenha e que não contenha nenhuma outra. Com esse objectivo, recordemos dois teoremas da análise matemática.

Teorema 2.1 (teorema de Bolzano). Se f for contínua em $[a, b]$ e se $f(a)f(b) < 0$, então f tem, pelo menos, uma raiz em$]a, b[$.

Teorema 2.2 (corolário do teorema de Rolle). Se f for contínua em $[a, b]$ e continuamente diferenciável em$]a, b[$ e se $f'(x) \neq 0$ em $]a, b[$ então f tem, no máximo, uma raiz em$]a, b[$.

Combinando estes dois teoremas e alguns outros resultados da análise, é possível, em muitas situações, localizar as raizes reais de uma equação.

Exemplo 2.2. Com base nos teoremas 2.1 e 2.2, determinar o número de raizes reais da equação

$$e^x - x^2 - 2x = 0.5$$

e determinar um intervalo que contenha cada uma delas (mas não contenha as restantes).

Este problema é equivalente a determinar os zeros da função de variável real $f(x) = e^x - x^2 - 2x - 0.5$. Esta função é evidentemente contínua em $\mathbb{R}$,

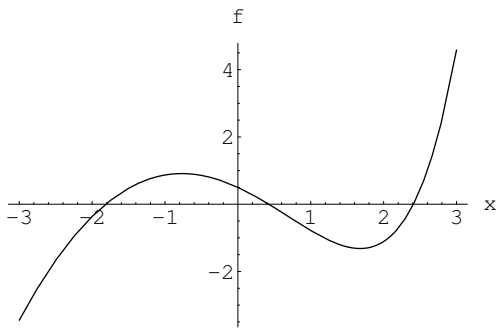

Figure 1: Gráfico relativo ao Exemplo 2.2

assim como todas as suas derivadas de qualquer ordem (ver o seu gráfico na fig.1).

Para facilitar a análise do problema, comecemos por calcular uma tabela de valores de f e de f'.

Tabela 2.1

x	-3	-2	-1	0	1	2	3
$f(x)$	-3.45	-0.365	0.868	0.5	-0.782	-1.11	4.59
$f'(x)$	4.05	2.14	0.368	-1	-1.28	1.39	12.1

Observando a tabela 2.1, verifica-se imediatamente que o teorema 2.1 é aplicável a f nos intervalos $[-2,-1]$,$[0,1]$e $[2,3]$. Daqui se conclui que a equação considerada tem, pelo menos, 3 raizes: $z_1 \in [-2,-1], z_2 \in [0,1]$ e $z_3 \in [2,3]$. Pela aplicação do teorema 2.2 podemos concluir também que, em cada um destes intervalos, a função f tem exactamente 1 raiz. Para isso, consideremos a primeira e a segunda derivada de f : $f'(x) = e^x - 2x - 2, f''(x) = e^x - 2$.

Em relação à segunda derivada, verifica-se facilmente que ela é positiva para $x > \ln 2$ e negativa para $x < \ln 2$. Temos $f''(\ln 2) = 0, f'''(\ln 2) = 2$, pelo que f' tem em $\ln 2$ um ponto de mínimo. Assim, no intervalo $[-2,-1], f'$ é decrescente. Recorrendo de novo à tabela 2.1, verifica-se que f' é sempre positiva neste intervalo, pelo que, pelo teorema 2.2, podemos concluir que f tem uma única raiz z_1 no intervalo $[-2,-1]$. Do mesmo modo, podemos observar que a função f' é crescente em $[2,3]$ e, de acordo com a tabela, toma sempre valores positivos neste intervalo. Aplicando o teorema 2.2 neste intervalo, constata-se que f tem nele uma única raiz z_3. Para aplicar o teorema 2.2 no intervalo [0,1], comecemos por recordar que a função f' tem

um ponto de mínimo em $x = \ln 2$, que pertence a este intervalo. Note-se que $f'(\ln 2) = -1.38 < 0$; uma vez que, de acordo com a tabela, $f'(0)$ e $f'(1)$ também são negativos, podemos concluir que f' é negativa em todo o intervalo [0,1]. Logo, o teorema 2.2 é aplicável neste intervalo e a função tem nele uma única raiz z_2. Resta esclarecer uma questão: será que a função f tem alguma raiz além das que acabámos de localizar? Para responder a esta pergunta, recordemos que a segunda derivada de f tem uma única raiz real $(\ln 2)$. Daqui, pelo teorema de Rolle, somos levados a concluir que a primeira derivada de f tem, no máximo, duas raizes reais. Finalmente, aplicando o teorema de Rolle a f',conclui-se que f tem, no máximo, 3 raizes reais. Como já vimos que existem, pelo menos, 3 raizes (z_1, z_2 e z_3), concluímos que estas são as únicas raizes de f.

2.2 Método da Bissecção

Um dos métodos mais simples para o cálculo aproximado de raizes é o método da bissecção. Para se poder aplicar este método basta conhecer um intervalo que contenha uma única raiz da função considerada (a qual é, por condição, contínua). A ideia do método é construir uma sucessão de intervalos encaixados $[a, b] \supset [a_1, b_1] \supset ... \supset [a_k, b_k]$ tais que a)cada intervalo tem o comprimento igual a metade do intervalo anterior; b)$f(a_i)f(b_i) < 0, i = 1, 2, ..., k$. Desta última condição, pelo teorema 2.1, resulta que a raiz é um ponto comum a todos os intervalos da sucessão. Assim, construindo um número suficientemente grande de intervalos, é possível aproximar a raiz com a precisão que se pretender. Vejamos em pormenor o algoritmo deste método.

1° Passo. Dado um intervalo $[a, b]$ e uma função f tais que $f(a)f(b) < 0$, determina-se o ponto médio desse intervalo: $x_1 = \frac{a+b}{2}$. Se, por feliz coincidência, se verificar $f(x_1) = 0$, x_1 é a raiz procurada Suponhamos que $f(x_1) \neq 0$,então verifica-se $f(x_1)f(a) < 0$ ou $f(x_1)f(a) > 0$. No primeiro caso, podemos afirmar que $z \in [a, x_1]$; no segundo caso, $z \in [x_1, b]$. Então o intervalo $[a_1, b_1]$ pode ser definido do seguinte modo:

se $f(x_1)f(a) < 0$, então $a_1 = a; b_1 = x_1$; senão, $a_1 = x_1$; $b_1 = b$.

Em qualquer dos casos, o novo intervalo $[a_1, b_1]$ satisfaz $f(a_1)f(b_1) < 0$.

2° Passo. Repetem-se as acções do primeiro passo, substituindo o intervalo $[a, b]$ por $[a_1, b_1]$ e representando por x_2 o ponto médio deste intervalo.Deste passo, resulta o intervalo $[a_2, b_2]$.

Generalizando, no **k-ésimo** passo (iteração) realizam-se as seguintes operações.

Determina-se o ponto médio do intervalo anterior: $x_k = \frac{a_{k-1}+b_{k-1}}{2}$;

Se $f(x_k)f(a_{k-1}) < 0$, então $a_k = a_{k-1}; b_k = x_k$; senão, $a_k = x_k; b_k = b_{k-1}$. Do **k-ésimo** passo obtém-se o intervalo $[a_k, b_k]$.

O processo é interrompido quando for satisfeita a condição de paragem: $b_k - a_k < \varepsilon$,onde ε é uma tolerância estabelecida de acordo com a precisão que se pretende obter.

Note-se que o comprimento do k-ésimo intervalo, por construção, é $b_k - a_k = \frac{b-a}{2^k}$, pelo que tende para zero, quando k tende para infinito. Logo, qualquer que seja ε, a condição de paragem é satisfeita ao fim de um certo número de passos, que depende do comprimento do intervalo inicial e de ε. Mais precisamente, temos

$$\frac{b-a}{2^k} < \varepsilon \Leftrightarrow \frac{b-a}{\varepsilon} < 2^k \Longleftrightarrow k > \log_2(\frac{b-a}{\varepsilon}).$$

Assim, o número de passos do método da bissecção que é necessário realizar até satisfazer a condição de paragem é o mínimo inteiro k, tal que $k > \log_2(\frac{b-a}{\varepsilon})$.

Se tomarmos como k-ésima aproximação da raiz z o valor de x_k podemos afirmar que o erro absoluto de x_k satisfaz

$$|z - x_k| < \frac{b_{k-1} - a_{k-1}}{2} = \frac{b-a}{2^k}.$$

É costume nos métodos computacionais representar o erro da k-ésima aproximação da raiz por e_k; usando esta notação, podemos afirmar que, no método da bissecção é válida a estimativa do erro:

$$|e_k| < \frac{b-a}{2^k}.$$

Exercício 2.1. **a)**Recorrendo ao teorema 2.1, justifique que a raiz cúbica de 2 pertence ao intervalo $[1.2, 1.3]$.

b) Baseando-se na alínea anterior, efectue três iterações (passos) do método da bissecção com o objectivo de calcular um valor aproximado da raiz cúbica de 2.

c) Quantas iterações teria que efectuar se pretendesse determinar $\sqrt[3]{2}$ com um erro absoluto inferior a 0.001?

Resposta. Comecemos por observar que determinar a raiz cúbica de 2 equivale a resolver a equação $f(x) = x^3 - 2 = 0$.

a) Temos que $f(1.2) = 1.2^3 - 2 = -0.272 < 0$ e $f(1.3) = 1.3^3 - 2 = 0.197 > 0$. Uma vez que a função f é contínua, pelo teorema 2.1 concluímos que a raiz procurada está no intervalo $[1.2, 1.3]$.

b) Comecemos com o intervalo $[a, b] = [1.2, 1.3]$. Então $x_1 = \frac{a+b}{2} = 1.25$. Verifica-se que $f(1.25) = -0.047 < 0$, de onde resulta que $f(1.25)f(1.2) > 0$. Logo, o intervalo a considerar na iteração seguinte é $[a_1, b_1] = [1.25, 1.3]$. Por conseguinte, $x_2 = \frac{a_1+b_1}{2} = 1.275$. Neste caso, $f(1.275) = 0.0727 > 0$, de onde resulta que $f(1.275)f(1.25) < 0$. Assim, o intervalo a considerar na terceira iteração é $[a_2, b_2] = [1.25, 1.275]$. Finalmente, $x_3 = \frac{a_2+b_2}{2} = 1.2625$. Neste ponto, temos $f(1.2625) = 0.012 > 0$, pelo que o intervalo a considerar na iteração seguinte será $[a_3, b_3] = [1.25, 1.2625]$

c) O comprimento do intervalo inicial é $b - a = 0.1$. Logo, para atingir uma precisão de $\varepsilon = 0.001$, o número de iterações é dado por $\log_2(\frac{b-a}{\varepsilon}) = \log_2(\frac{0.1}{0.001}) = 6.64$. Ou seja, a precisão será atingida ao fim de 7 iterações.

2.3 Método do Ponto Fixo

2.3.1 Definição e exemplos de pontos fixos

Definição. Seja g uma função real, definida num certo intervalo $[a, b] \subset I\!R$. O número $z \in [a, b]$ diz-se um ponto fixo de g se $g(z) = z$.

Exemplo 2.3

a) Seja $g(x) = \alpha x + \beta$, $\alpha \neq 1, \alpha, \beta \in I\!R$.

O ponto fixo de g é o que satisfaz $\alpha z + \beta = z$, ou seja $z = \frac{\beta}{1-\alpha}$. Por exemplo, se for $\alpha = 2, \beta = -3$, obtém-se $z = 3$ (fig.2).

b) Seja $g(x) = x^2+1$. Neste caso, a equação dos pontos fixos é $z^2+1 = z$. Por conseguinte, temos $z = \frac{1}{2} \pm \sqrt{\frac{1}{2}^2 - 1}$, ou seja, não existem pontos fixos reais (fig. 3).

c) $g(x) = x^2$. A equação dos pontos fixos, neste caso, é $z^2 = z$. Logo, existem dois pontos fixos: $z_1 = 0, z_2 = 1$ (fig.4).

d) $g(x) = \cos x$. Embora não seja possível determinar analiticamente o ponto fixo desta função, é fácil verificar que ela tem um ponto fixo (único) no intervalo $[0, 1]$. Com efeito, se definirmos $f(x) = \cos x - x$, verifica-se que $f(0) = 1, f(1) = \cos 1 - 1 < 0$. Logo, sendo a função f contínua, pelo teorema 2.1, ela tem, pelo menos, um zero z em $]0, 1[$. Nesse ponto z verifica-se $\cos z = z$, pelo que z é um ponto fixo de g. Por outro lado, f é uma função continuamente diferenciável e a sua derivada, $f'(x) = -senx - 1$, é negativa em $[0, 1]$. Logo, pelo Teorema 2.2, f tem uma única raiz neste intervalo, que é também o único ponto fixo de g (fig.5).

Dada uma função g, determinar os seus pontos fixos equivale a calcular as raizes da equação $g(x) - x = 0$, ou, por outras palavras, calcular os zeros da função $f(x) = g(x) - x$. Inversamente, se for dada uma equação $f(x) = 0$,

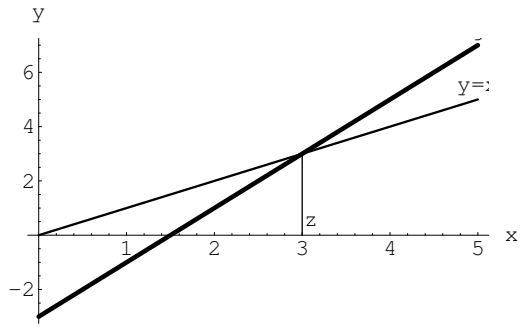

Figure 2: Exemplo 2.3a

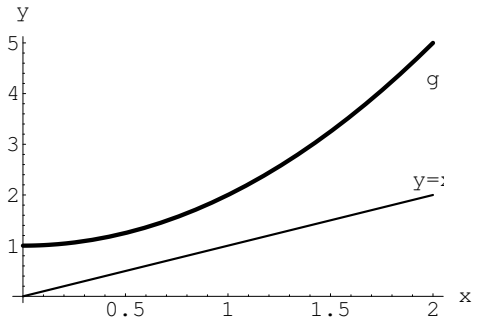

Figure 3: Exemplo 2.3b

calcular as raizes dessa equação equivale a determinar os pontos fixos de outra função.

Exemplo 2.4. Consideremos de novo a equação $e^x - x^2 - 2x = 0.5$ (exemplo 2.2). Esta equação pode ser reescrita de várias formas, todas elas equivalentes:

$$\frac{e^x - x^2 - 0.5}{2} = x; \tag{2.1}$$

$$\sqrt{e^x - 2x - 0.5} = x; \tag{2.2}$$

$$\ln(x^2 + 2x + 0.5) = x. \tag{2.3}$$

No caso da equação (2.1), as raizes da equação inicial são vistas como os pontos fixos da função $g_1(x) = \frac{e^x - x^2 - 0.5}{2}$. Em relação à equação (2.2), ela remete-nos para os pontos fixos de $g_2(x) = \sqrt{e^x - 2x - 0.5}$. Note-se

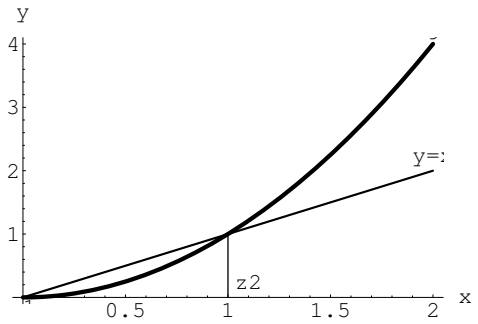

Figure 4: Exemplo 2.3c

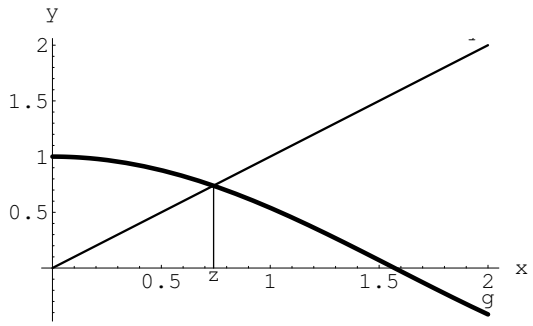

Figure 5: Exemplo 2.3d

que, neste caso, as equações só são equivalentes para valores positivos de x (a função g_2 não pode tomar valores negativos). Em particular, a raiz z_1, negativa, não é ponto fixo de g_2. Quanto à equação (2.3), diz-nos que as raizes da equação inicial são pontos fixos da função $g_3(x) = \ln(x^2+2x+0.5)$. Neste caso, a equivalência também não é válida para qualquer valor de x, já que o domínio da função g_3 só inclui os valores de x,para os quais $x^2+2x+0.5 > 0$. Das raizes da equação inicial apenas z_2 e z_3 satisfazem esta condição. Logo, z_2 e z_3 são também pontos fixos de g_3, enquanto z_1 não o é.

O Exemplo 2.4 mostra-nos que as raizes de uma equação dada podem ser tratadas como pontos fixos de diferentes funções. Veremos que esse facto pode ser utilizado na escolha de métodos numéricos para o cálculo aproximado destas raizes.

2.3.2 Sucessões numéricas geradas por funções

Dada uma função real g,com domínio num certo intervalo $[a, b]$, e um número x_0, tal que $x_0 \in [a, b]$, é possível gerar uma sucessão de números reais $\{x_n\}$ do seguinte modo:

$$x_{k+1} = g(x_k), \qquad k = 0, 1, ... \tag{2.4}$$

Se a imagem do intervalo [a,b] estiver contida no próprio intervalo, então a relação (2.4) permite-nos definir uma sucessão infinita de elementos do conjunto considerado. Neste caso, chamaremos a g a função iteradora e aos termos x_k da sucessão as *iteradas.* Veremos como as sucessões geradas desse modo podem ser utilizadas para aproximar as raizes de uma equação dada.

Exemplo 2.5. Seja $g(x) = x^2$. O domínio desta função é R e a imagem do intervalo [0,1] por esta função é o próprio intervalo.Se tomarmos $x_0 = 0$, a função g gera uma sucessão constante:$\{0, 0, 0, ..\}$. Se considerarmos $0 < x_0 < 1$,então a sucessão gerada é $\{x_0, x_0^2, x_0^4, ...\}$ e converge para 0, que é um dos pontos fixos de g. Se tomarmos $x_0 = 1$, a sucessão das iteradas é de novo constante: $\{1, 1, 1, ..\}$(1 também é um ponto fixo de g). Se tomarmos $x_0 > 1$, a sucessão vai ser divergente (tende para infinito).

O exemplo 2.5 sugere-nos que, quando a sucessão gerada por uma função g converge, então o seu limite é um ponto fixo da função g. De facto, pode provar-se que assim é.

Teorema 2.3. Seja $\{x_n\}$ uma sucessão gerada pela função g que converge para um certo limite z. Se g for contínua em z, então z é ponto fixo de g.

Demonstração . Uma vez que $z = \lim_{n\to\infty} x_n$, temos

$$z = \lim_{n\to\infty} x_{n+1} = \lim_{n\to\infty} g(x_n).$$

Da continuidade de g em z resulta que $\lim_{n\to\infty} g(x_n) = g(\lim_{n\to\infty} x_n) = g(z)$. Obtemos assim que $z = g(z)$,como se pretendia demonstrar.

Exercício 2.2. Considere a sucessão gerada pela função $g(x) = sen(x)$, com $x_0 = 1$. Prove que esta sucessão converge. Qual é o seu limite?

Resposta. Para provar que a sucessão converge, basta provar que é monótona e limitada. Note-se que, sendo $0 < x < 1$, temos $0 < sen(x) < x$. Daqui resulta que 1) todos os termos da sucessão considerada pertencem ao intervalo [0,1]; 2) a sucessão é motótona decrescente, já que $x_{k+1} = sen(x_k) < x_k$.Por conseguinte a sucessão é monótona e limitada, logo é convergente. De acordo com o teorema 2.3, a sucessão considerada, sendo

convergente, deve convergir para um ponto fixo da função iteradora. O único ponto fixo da função $g(x) = sen(x)$ é $z = 0$, logo é para esse ponto que a sucessão converge.

2.3.3 Teorema do Ponto Fixo

O teorema 2.3 diz-nos que uma sucessão gerada por uma função iteradora g, a convergir, converge para um ponto fixo daquela função. Fica por responder a questão: em que condições essa sucessão converge? A resposta a esta questão é dada por um teorema fundamental da Análise Matemática, o *teorema do ponto fixo.* Embora este teorema possa ser formulado num contexto mais vasto, por agora, limitar-os-emos a uma versão simplificada, em que g é uma função de uma variável real.

Teorema 2.4 (teorema do ponto fixo). Seja g uma função de variável real e $[a, b]$ um intervalo tais que:

1. $g([a, b]) \subset [a, b]$;
2. g é continuamente diferenciável em $[a, b]$;
3. $\max_{x\in[a,b]} |g'(x)| = L < 1$;

Então são válidas as seguintes afirmações:

1. g tem um único ponto fixo z em $[a, b]$;
2. se $x_0 \in [a, b]$, a sucessão gerada pela função g converge para o ponto fixo z.

Demonstração. 1. Para demonstrar a existência de, pelo menos, um ponto fixo, defina-se a função $h(x) = g(x) - x$. Esta função é obviamente contínua em $[a, b]$; se tivermos $g(a) = a$ (ou $g(b) = b$), teremos que a (ou b, respectivamente) é ponto fixo de g. Caso contrário, de acordo com a condição 1), a função h satisfaz $h(a) = g(a) - a > 0, h(b) = g(b) - b < 0$; então, pelo teorema de Bolzano, existe, pelo menos, um ponto $z \in [a, b]$, tal que $h(z) = 0$, ou seja, $g(z) = z$; logo, z é ponto fixo de g.

Para demonstrar a unicidade, suponhamos que existem em $[a, b]$ dois pontos fixos distintos z_1,z_2. Por definição, temos $g(z_1) = z_1, g(z_2) = z_2$. Logo, $|g(z_1) - g(z_2)| = |z_1 - z_2|$. Por outro lado, usando o Teorema de Lagrange e a condição 3, temos

$$|g(z_1) - g(z_2)| \leq \max_{x\in[a,b]} |g'(x)||z_1 - z_2| = L|z_1 - z_2|.$$

Obtemos assim que

$$|z_1 - z_2| \leq L|z_1 - z_2|,$$

ou seja,

$$|z_1 - z_2|(1 - L) \leq 0. \tag{2.5}$$

Mas, de acordo com a condição 3, $L < 1$; logo, da desigualdade (2.5) resulta que $|z_1 - z_2| = 0$, o que contradiz a hipótese de z_1 e z_2 serem distintos. Desta contradição conclui-se a unicidade do ponto fixo.

2. Para demonstrar a segunda afirmação seja x_0 um ponto arbitrário de $[a, b]$. Pela condição 1 do teorema, temos que $x_1 = g(x_0)$ também pertence ao intervalo $[a, b]$. Do mesmo modo se conclui que todos os elementos da sucessão, gerada pela função g, pertencem àquele intervalo. Vamos agora provar que esta sucessão converge para o ponto fixo z. Pela condição 3 do teorema, temos

$$|x_n - z| = |g(x_{n-1}) - g(z)| \leq L|x_{n-1} - z|. \tag{2.6}$$

Aplicando n vezes a desigualdade (2.6), conclui-se que

$$|x_n - z| \leq L^n|x_0 - z|. \tag{2.7}$$

Como, por condição, $L < 1$, da desigualdade (2.7) resulta que $|x_n - z| \to 0$, quando $n \to \infty$ (qualquer que seja $x_0 \in [a, b]$). Fica assim demonstrada a segunda afirmação do teorema.

O teorema do ponto fixo não só nos garante a existência de um único ponto fixo z da função g num dado intervalo, como nos indica um método para obter aproximações desse ponto. Na realidade, se tomarmos qualquer ponto inicial x_0 dentro do intervalo e construirmos a sucessão, gerada pela função g, de acordo com o teorema do ponto fixo essa sucessão converge para z. O método baseado nesta construção chama-se **método do ponto fixo**. Este método permite-nos, dada uma função iteradora g e um intervalo $[a, b]$ (tal que sejam satisfeitas as condições do teorema 2.4), obter uma aproximação tão precisa quanto quisermos do ponto fixo de g em $[a, b]$. O algoritmo é extremamente simples:

1. Escolher um ponto $x_0 \in [a, b]$;
2. calcular cada nova iterada pela fórmula $x_n = g(x_{n-1})$, $n = 1, 2, ...$;
3. Parar quando se obtiver uma aproximação aceitável (sobre os critérios de paragem falaremos mais abaixo).

2.3.4 Estimativa do erro do método do ponto fixo

Para efeitos práticos, interessa-nos não só saber as condições em que um método converge, mas também estimar o erro das aproximações obtidas. No caso do método do ponto fixo, a resposta a esta questão é dada pelo seguinte teorema.

Teorema 2.5. Nas condições do Teorema 2.4, são válidas as seguintes estimativas do erro:

$$|x_n - z| \leq L^n |x_0 - z| \tag{2.8}$$

(estimativa apriori),

$$|x_n - z| \leq \frac{L}{1-L}|x_n - x_{n-1}|, \qquad n \geq 1 \tag{2.9}$$

(estimativa aposteriori), onde x_{n-1}, x_n, são duas iteradas consecutivas do método do ponto fixo e $L = \max_{x\in[a,b]} |g'(x)|$.

Demonstração. A fórmula (2.8) já foi provada durante a demonstração do teorema do ponto fixo (ver (2.7)). Quanto à fórmula (2.9), pode ser provada do seguinte modo. Comecemos por observar que

$$|x_{n-1} - z| = |z - x_{n-1}| \leq |z - x_n| + |x_n - x_{n-1}|. \tag{2.10}$$

Por outro lado, visto que, de acordo com (2.6),

$$|x_n - z| \leq L|x_{n-1} - z|,$$

da fórmula (2.10) resulta que

$$|x_{n-1} - z|(1 - L) \leq |x_n - x_{n-1}|. \tag{2.11}$$

Observando que $1-L > 0$ (pela condição 3 do teorema 2.4), podem dividir-se por este valor ambos os membros da desigualdade (2.11), obtendo-se assim

$$|z - x_{n-1}| \leq \frac{1}{1-L}|x_n - x_{n-1}|. \tag{2.12}$$

Finalmente, das desigualdades (2.12) e (2.6) obtém-se a estimativa (2.9).

Exercício 2.3. Considere a equação $cos(x) = 2x$.

1. Com base no teorema do ponto fixo, mostre que esta equação tem uma única raiz no intervalo $[0.4, 0.5]$ e que o método do ponto fixo converge para essa raiz.

   Resolução. Comecemos por observar que qualquer raiz da equação dada é um ponto fixo de $g(x) = \frac{cosx}{2}$. Mostremos agora que a função g satisfaz, no dado intervalo, as condições do teorema do ponto fixo.

   (a) Para mostrar que $g([0.4, 0.5]) \subset [0.4, 0.5]$ comecemos por calcular as imagens dos extremos do intervalo: $g[0.4] = cos(0.4)/2 = 0.46053 \in [0.4, 0.5]$; $g(0.5) = cos(0.5)/2 = 0.43879 \in [0.4, 0.5]$. Por outro lado, a função g é decrescente em $[0.4, 0.5]$ ($g'(x) = -senx/2$ é negativa naquele intervalo). Daqui se conclui que $g([0.4, 0.5]) \subset [0.4, 0.5]$.

   (b) g é continuamente diferenciável em $I\!R$ e, em particular, no intervalo considerado.

   (c) $L = \max_{x\in[0.4,0.5]} |g'(x)| = \max_{x\in[0.4,0.5]} \frac{|senx|}{2} = \frac{sen0.5}{2} = 0.2397 < 1$.

   Concluimos assim que todas as condições do torema do ponto fixo estão satisfeitas, pelo que o método do ponto fixo com a função iteradora $g(x) = cosx/2$ converge para o ponto fixo.

2. Tomando como aproximação inicial $x_0 = 0.4$, calcule as duas primeiras iteradas do método.

   Resolução. Tomando como aproximação inicial $x_0 = 0.4$, as duas primeiras aproximações iniciais são $x_1 = g(x_0) = 0.46053; x_2 = g(x_1) = 0.44791$.

3. Obtenha uma estimativa do erro da aproximação x_2, calculada na alínea anterior.

   Resolução.Usando a fórmula (2.9), obtém-se

   $$|z - x_2| \leq \frac{L}{1-L}|x_2 - x_1| = \frac{0.2397}{1-0.2397}|0.44791 - 0.46053| = 0.00397.$$

4. Nas condições da alínea anterior, quantas iterações é necessário efectuar para garantir que o erro absoluto da aproximação obtida é inferior a 0.001?

   Resolução. Para responder a esta questão, é necessário aplicar a estimativa *a priori* (2.8). De acordo com esta estimativa, temos

   $$|x_n - z| \leq L^n|x_0 - z| \leq 0.2397^n|0.5 - 0.4| = 0.1 \times 0.2397^n, \qquad n \geq 1.$$

Logo, para garantir que o erro absoluto da n-ésima iterada é inferior a um certo ϵ, basta escolher n de tal modo que $0.2397^n < 10\epsilon$. Desta inequação, resulta que $n > \log_{0.2397}(10\epsilon)$. No caso de $\epsilon = 0.001$, obtém-se $n > 3.22$, de onde se conclui que o erro satisfaz a condição exigida ao fim de 4 iterações.

2.3.5 Classificação dos pontos fixos

De acordo com o teorema do ponto fixo, a convergência das sucessões geradas por uma certa função g num intervalo $[a, b]$ depende do comportamento da sua derivada g' nesse intervalo. Isto leva-nos a classifiar os pontos fixos z de uma função g de acordo com o valor de $g'(z)$. Neste parágrafo iremos assumir que g é continuamente diferenciável (ou seja, g' é contínua), pelo menos, numa vizinhança de cada ponto fixo de g.

Assim, um ponto fixo z de uma função g diz-se **atractor** se $|g'(z)| < 1$. De facto, se $|g'(z)| < 1$ e g' é contínua em z, então existe uma vizinhança $V_\epsilon(z) = (z - \epsilon, z + \epsilon)$ tal que $\max_{x \in V_\epsilon(z)} |g'(z)| = L < 1$. Por outro lado, se $x \in V_\epsilon(z)$, temos $|g(x) - g(z)| \leq L|x - z| < |x - z| < \epsilon$, ou seja, $g(x)$ também pertence a $V_\epsilon(z)$. Log, se o intervalo $[a, b]$ estiver contido em $V_\epsilon(z)$, nesse intervalo a função g satisfaz as condições do teorema do ponto fixo. Concluimos portanto que, **se z for um ponto fixo atractor, então existe uma vizinhança $V_\epsilon(z)$ tal que, se $x_0 \in V_\epsilon(z)$ então a sucessão gerada por g converge para z.**

Por outro lado, se z é um ponto fixo de g e $|g'(z)| > 1$, então z diz-se um **repulsor**. Nesse caso, é fácil verificar que nenhuma sucessão gerada pela função g converge para z (excepto a sucessão constante $\{z, z, \dots\}$). De facto, se z é um repulsor, então existe uma vizinhança $V_\epsilon(z) = (z - \epsilon, z + \epsilon)$ tal que $|g'(z)| > 1, \forall x \in V_\epsilon(z)$. Assim, seja x_k um termo de uma sucessão gerada pela função g e suponhamos que $x_k \in V_\epsilon(z), x_k \neq z$. Então temos $|x_{k+1} - z| = |g(x_k) - g(z)| \geq \min_{x \in V_\epsilon(z)} |g'(x)||x_k - z| > |x_k - z|$. Logo, x_{k+1} está mais distante de z do que x_k. Um raciocínio semelhante, aplicado aos termos seguintes da sucessão, mostra-nos que esta não pode convergir para z.

Finalmente, se tivermos $|g'(z)| = 1$, então o ponto fixo z diz-se **neutro**. Neste caso, existem sucessões geradas pela função g que convergem para z e outras que não convergem (mesmo que x_0 esteja próximo do ponto fixo z).

Exemplo 2.6. Consideremos a função $g(x) = kx(1 - x)$, onde $k > 0$. Esta função é conhecida como "função logística" e tem diversas aplicações em ecologia matemática. Para determinarmos os pontos fixos desta função,

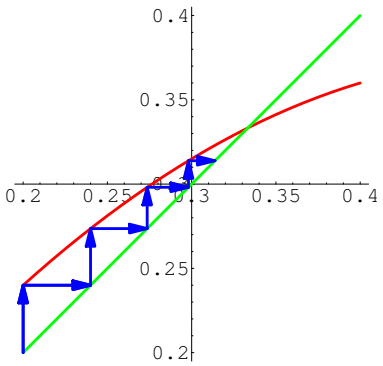

Figure 6: Iterações da função $g(x) = 1.5x(1-x)$, com $x_0 = 0.2$

.

para um certo valor de k, resolva-se a equação

$$kz(1-z) = z. \tag{2.13}$$

É fácil verificar que esta equação tem 2 raizes: $z_1 = 0$ e $z_2 = 1 - 1/k$. Consideremos o caso de $k = 1.5$. Nesse caso, os dois pontos fixos de g são $z_1 = 0$ e $z_2 = 1/3$. Para os classificarmos, observemos que $g'(x) = 1.5 - 3x$. Logo $g'(0) = 1.5$ e $g'(1/3) = 1.5 - 1 = 0.5$, ou seja, z_1 é **repulsor** e z_2 é **atractor**. Isto significa que a) nenhuma sucessão gerada pela função g poderá convergir para 0 (excepto a sucessão constante, igual a 0); b) se x_0 for suficientemente próximo de 1/3, a sucessão gerada por g converge para $z_2 = 1/3$. Mais precisamente pode provar-se que, se $0 < x_0 < 1$, então a sucessão $\{x_k\}$ converge para z_2. As figuras 6 e 7 ilustram esta afirmação.

Exemplo 2.7. Seja $g(x) = x^2 + x$. Esta função tem um ponto fixo (único) em $z = 0$. Visto que $g'(z) = 2z + 1 = 1$, este ponto fixo é **neutro**. Vejamos qual é o comportamento das sucessões geradas por esta função.

Seja $x_0 = 0.12$. Então, temos que $x_1 = x_0^2 + x_0 = 0.1344$; $x_2 = x_1^2 + x_1 = 0.152463$, etc. É fácil verificar que, neste caso, a sucessão é crescente e tende para $+\infty$.

Mas se escolhermos como iterada inicial $x_0 = -0.12$, já teremos $x_1 = x_0^2 + x_0 = -0.1056$;$x_2 = x_1^2 + x_1 = -0.0945$. Neste caso, a sucessão é crescente e converge para o ponto fixo $z = 0$. As figuras 8 e 9 ilustram este exemplo.

2.3.6 Monotonia das iteradas do método do ponto fixo

Suponhamos que z é um ponto fixo atractor da função g. Como vimos no parágrafo anterior, isto verifica-se sempre que $|g'(z)| < 1$, isto é, $-1 <$

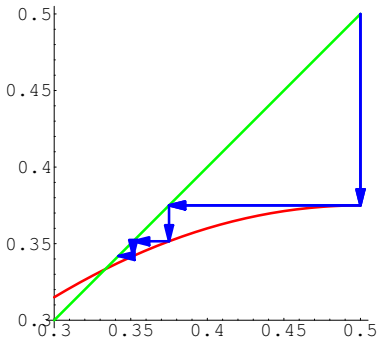

Figure 7: Iterações da função $g(x) = 1.5x(1-x)$, com $x_0 = 0.5$

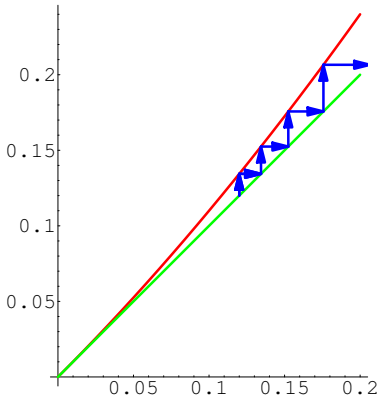

Figure 8: Iterações da função $g(x) = x^2 + x$, com $x_0 = 0.12$

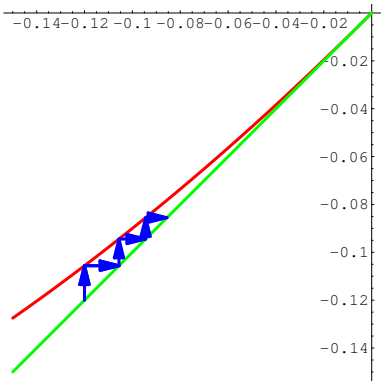

Figure 9: Iterações da função $g(x) = x^2 + x$, com $x_0 = -0.12$

$g'(z) < 1$. Como vimos, neste caso, qualquer sucessão gerada pela função g, com x_0 suficientemente próximo de z, converge para z. Neste parágrafo, vamos investigar em que condições essa sucessão é monótona (crescente ou decrescente). Tal como antes, admitimos que g é continuamente diferenciável numa vizinhança de z.

Caso 1. Suponhamos que $0 < g'(z) < 1$. Então, da continuidade da derivada de g, resulta que existe uma vizinhança $V_\epsilon(z) = (z - \epsilon, z + \epsilon)$, tal que, se $x \in V_\epsilon(z)$, então $0 < g'(x) < 1$. Suponhamos que x_k é um termo de uma sucessão, gerada pela função g, tal que $x_k \in V_\epsilon(z)$. Para sermos mais específicos admitamos que $z < x_k < z + \epsilon$. Nesse caso, uma vez que $x_{k+1} = g(x_k)$, aplicando o teorema de Lagrange, existe um ponto ξ_k, $z \leq \xi_k \leq x_k$, tal que

$$x_{k+1} - z = g(x_k) - g(z) = g'(\xi_k)(x_k - z). \tag{2.14}$$

Por construção, temos $x_k - z > 0$ e $g'(\xi_k) > 0$. Logo, $x_{k+1} > z$. Concluimos que, se $x_k > z$, então também $x_{k+1} > z$. Por outro lado, uma vez que z é um ponto atractor ($g'(\xi_k) < 1$), x_{k+1} deve estar mais próximo de z do que x_k, de onde se conclui que $x_{k+1} < x_k$. Como o mesmo raciocínio se aplica a todas as iteradas subsequentes, podemos concluir que, neste caso, a sucessão $\{x_n\}$ é **decrescente** (pelo menos, a partir da ordem k). Esta situação é ilustrada pelo gráfico da fig. 7. Se tivermos $x_k < z$, o mesmo raciocínio leva-nos a conlusão que $x_{k+1} > x_k$. Nesse caso, a sucessão das iteradas será **crescente**. Esta situação é a que está representada na fig. 6 e na fig. 9. Em qualquer das duas situações, as sucessões são **monótonas**.

Caso 2. Suponhamos agora que $-1 < g'(z) < 0$. Então, da continuidade da derivada de g, resulta que existe uma vizinhança $V_\epsilon(z) = (z - \epsilon, z + \epsilon)$,

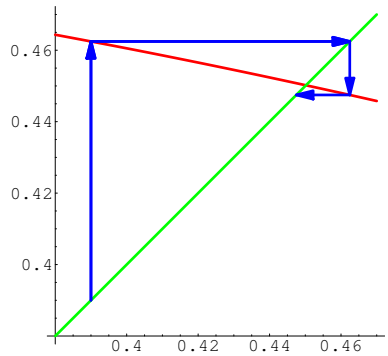

Figure 10: Iterações da função $g(x) = \frac{cosx}{2}$, com $x_0 = 0.39$

.

tal que, se $x \in V_\epsilon(z)$, então $-1 < g'(x) < 0$. Admitindo que x_k pertence a essa vizinhança, a igualdade (2.14) é aplicável. Neste caso, admitindo que $x_k > z$, dessa igualdade resulta que $x_{k+1} < z$ (uma vez que $g'(\xi_k) < 0$). Se aplicarmos o mesmo raciocínio às iteradas seguintes, concluímos que $x_{k+2} > z$, $x_{k+3} < z$, etc. Se, pelo contrário, tivermos $x_k < z$, então $x_{k+1} > z$, $x_{k+2} < z$, etc. Ou seja, neste caso, as iteradas vão ser alternadamente maiores ou menores que z (uma sucessão deste tipo diz-se **alternada**). Uma propriedade interessante das sucessões alternadas é que o limite da sucessão está sempre entre dois termos consecutivos, isto é $x_k < z < x_{k+1}$ ou $x_{k+1} < z < x_k$. Isto permite-nos obter um majorante do erro absoluto de x_{k+1} além daqueles que foram obtidos em 2.3.4:

$$|x_{k+1} - z| < |x_{k+1} - x_k|. \tag{2.15}$$

A sucessão das iteradas do exercício 2.3, em que $g'(z) < 0$, é um exemplo de uma sucessão alternada. Na fig. 10 estão representados graficamente alguns termos desta sucessão.

2.3.7 Divergência do método do ponto fixo

O estudo dos pontos repulsores no parágrafo 2.3.5 permite-nos formular o seguinte critério de divergência do método do ponto fixo.

Teorema 2.6 Se g for continuamente diferenciável em $[a, b]$ e se $|g'(x)| > 1, \forall x \in [a, b]$, então nenhuma sucessão gerada pela função g pode convergir para qualquer ponto do intervalo $[a, b]$.

Demonstração. De acordo com as condições do teorema e com a classificação dos pontos fixos, se a função g tiver algum ponto fixo em $[a, b]$

esse ponto fixo é repulsor. Por outro lado, se uma sucessão gerada pela função g convergir, ela converge para um ponto fixo de g (Teorema 2.3). Da conjugação destes dois factos resulta a afirmação do teorema.

2.3.8 Ordem de convergência

Um dos conceitos fundamentais da teoria dos métodos iterativos é o conceito de **ordem de convergência**. Este conceito permite-nos comparar a rapidez com que diferentes métodos convergem e escolher, em cada caso, o método mais eficiente. Nas definições que se seguem representaremos por $\{x_n\}$,$n \in N$, um sucessão convergente de números reais e por z o seu limite.

Definição. Diz-se que uma sucessão $\{x_n\}$ tem **convergência, pelo menos, linear** (pelo menos,de ordem 1) se existir uma constante $k_\infty < 1$ tal que

$$k_\infty = \lim_{n\to\infty} \frac{|z - x_{n+1}|}{|z - x_n|};$$

este limite designa-se **coeficiente assimptótico de convergência**.

Sempre que $k_\infty \neq 0$, a ordem de convergência é exactamente 1. Se $k_\infty = 0$, a ordem de convergência é superior a 1, caso que estudaremos mais adiante.

Note-se que o coeficiente assimptótico de convergência nos permite comparar, quanto à rapidez de convergência, métodos diferentes que tenham convergência linear. Com efeito, quanto mais pequeno (mais próximo de 0) for k_∞ mais rápida será a convergência.

Exemplo 2.8. Consideremos a sucessão $\{x_n\}$, tal que $x_{n+1} = x_n/a$, $a > 1$, com $x_0 \in I\!R$. É fácil verificar que esta sucessão converge para $z = 0$, qualquer que seja $x_0 \in I\!R$, já que este é o único ponto fixo da função iteradora $g(x) = x/a$. Além disso, este ponto é atractor, já que $g'(x) = 1/a < 1$, $\forall x \in I\!R$. Verifiquemos que esta sucessão tem convergência linear. Para isso, calculemos o quociente:

$$k_\infty = \lim_{n\to\infty} \frac{|z - x_{n+1}|}{|z - x_n|} = \lim_{n\to\infty} \frac{|x_{n+1}|}{|x_n|} = \frac{1}{a} < 1. \tag{2.16}$$

Concluímos assim que temos **convergência linear**. Além disso, o coeficiente assimptótico de convergência é $k_\infty = \frac{1}{a}$. Assim, a convergência será tanto mais rápida quanto maior for a.

Analisemos agora os casos em que a ordem de convergência é superior a 1 (convergência supralinear).

Definição. Diz-se que uma sucessão $\{x_n\}$ tem **convergência, pelo menos, de ordem** $p > 1$ se existir uma constante k_∞ tal que

$$k_\infty = \lim_{n\to\infty} \frac{|z - x_{n+1}|}{|z - x_n|^p};$$

esta constante designa-se **coeficiente assimptótico de convergência**. Se $k_\infty \neq 0$, então a ordem de convergência é exactamente p.

Exemplo 2.9. Consideremos a sucessão $\{x_n\}$, tal que $x_{n+1} = bx_n^\alpha$, onde $b \neq 0$, $\alpha > 1$, com $|x_0| < |b|^{\frac{-1}{\alpha-1}}$. É fácil verificar que esta sucessão converge para $z = 0$, se x_0 satisfizer a condição indicada. De facto, o ponto $z = 0$ é um ponto fixo da função iteradora $g(x) = bx^\alpha$. Além disso, este ponto é atractor, já que $g'(0) = 0$. Por outro lado, se $|x_0| < |b|^{\frac{-1}{\alpha-1}}$, então $|x_1| < |x_0|$ e, de um modo geral, teremos que $|x_{n+1}| < |x_n|$, $\forall n$, pelo que a sucessão é decrescente em módulo. Verifiquemos qual é a ordem de convergência desta sucessão. Para isso calculemos o limite:

$$\lim_{n\to\infty} \frac{|z - x_{n+1}|}{|z - x_n|^p} = \lim_{n\to\infty} \frac{|x_{n+1}|}{|x_n|^p} = \lim_{n\to\infty} \frac{|bx_n^\alpha|}{|x_n|^p}. \tag{2.17}$$

Para que este limite seja finito, devemos ter $p = \alpha$. Neste caso, verifica-se $k_\infty = |b|$. Logo, a ordem de convergência é α e o coeficiente assimptótico de convergência é $|b|$.

A ordem de convergência do método do ponto fixo depende das propriedades da função iteradora g. O teorema que se segue diz-nos quais as que condições que a função g deve satisfazer para garantir que o método do ponto fixo tenha ordem de convergência $p \geq 1$.

Teorema 2.6 (ordem de convergência do método do ponto fixo). Suponhamos que g satisfaz as condições do teorema do ponto fixo em $[a, b]$ e que, além disso, $g \in C^p[a, b]$, com $p \geq 1$. Seja ainda $g'(z) = g''(z) = ... = g^{(p-1)}(z) = 0$, $g^{(p)}(z) \neq 0$. Então

1. g tem um único ponto fixo z em $[a, b]$;
2. se $x_0 \in [a, b]$, então a sucessão gerada pela função g converge para z com ordem p;
3. o coeficente assimptótico de convergência é $k_\infty = \frac{|g^{(p)}(z)|}{p!}$.

Demonstração. A primeira afirmação resulta do teorema 2.4 (teorema do ponto fixo). Resta-nos provar os pontos 2 e 3. Com esse fim, escreva-se o desenvolvimento de g em série de Taylor, em torno de z:

$$g(x) = g(z) + g'(z)(x - z) + \frac{g''(z)}{2}(x - z)^2 + \cdots + \frac{g^{(p)}(\xi)}{p!}(x - z)^p, \tag{2.18}$$

onde $\xi \in int(z, x)$. Em particular, se escrevermos a fórmula (2.18) com $x = x_m$, atendendo às condições do teorema, obtém-se

$$g(x_m) = g(z) + \frac{g^{(p)}(\xi_m)}{p!}(x_m - z)^p, \tag{2.19}$$

onde $\xi_m \in int(z, x_m)$. Da fórmula (2.19) resulta imediatamente que

$$x_{m+1} - z = \frac{g^{(p)}(\xi_m)}{p!}(x_m - z)^p. \tag{2.20}$$

Dividindo ambos os membros de (2.20) por $(x_m - z)^p$ e tomando o módulo, obtém-se

$$\frac{|x_{m+1} - z|}{|x_m - z|^p} = \frac{|g^{(p)}(\xi_m)|}{p!}. \tag{2.21}$$

Calculando o limite quando $m \to \infty$ em ambos os membros de (2.21), obtém-se

$$\lim_{m\to\infty} \frac{|x_{m+1} - z|}{|x_m - z|^p} = \frac{|g^{(p)}(z)|}{p!}. \tag{2.22}$$

Da igualdade (2.22) resulta imediatamente que a sucessão $\{x_m\}$ tem ordem de convergência p e que $k_\infty = \frac{|g^{(p)}(z)|}{p!}$.

Observe-se que, como caso particular do Teorema 2.6, com $p = 1$, se obtém a seguinte afirmação: se g satisfizer as condições do teorema do ponto fixo em $[a, b]$ e se $g'(z) \neq 0$, então, qualquer que seja $x_0 \in [a, b]$, a sucessão gerada pela função g converge linearmente para z e o coeficiente assimptótico de convergência é $k_\infty = |g'(z)|$.

Exercício 2.10 Considere a função iteradora $g(x) = \frac{1}{2}\left(x + \frac{1}{x}\right)$.

1. Mostre que os pontos fixos de g são $z_1 = 1$,$z_2 = -1$.

   Resolução. Da equação $g(z) = \frac{1}{2}\left(z + \frac{1}{z}\right) = z$, multiplicando ambos os membros por $2z$, obtém-se $z^2 + 1 = 2z^2$. daqui resulta que $z^2 = 1$, pelo que os pontos fixos de g são $z_1 = 1$,$z_2 = -1$.

2. Mostre que estes pontos são atractores.

   Resolução. Visto que $g'(x) = \frac{1}{2} - \frac{1}{2x^2}$, obtém-se $g'(1) = g'(-1) = 0$. Logo, estes pontos são atractores.

3. Seja $x_0 \in [1, 2]$. Mostre que a sucessão gerada pela função g converge para $z_1 = 1$, determine a ordem e o coeficiente assimptótico de convergência.

Resolução. Em primeiro lugar, mostremos que a função g satisfaz as condições do teorema do ponto fixo em $[1,2]$. Já sabemos que $g'(x) = \frac{1}{2} - \frac{1}{2x^2}$, logo g é continuamente diferenciável em $[1,2]$. Além disso, é fácil verificar que $g'(x) \geq 0, \forall x \in [1,2]$. Logo, g é crescente em $[1,2]$. Então, para mostrar que $g([1,2]) \subset [1,2]$, basta-nos verificar que $g(1) = 1 \in [1,2]$ e $g(2) = 5/4 \in [1,2]$. Por outro lado, temos $\max_{x\in[1,2]} |g'(x)| = |g(2)| = \frac{3}{8} < 1$.

Para determinar a ordem de convergência e o coeficiente assimptótico de convergência da sucessão considerada, vamos aplicar o teorema 2.6. Para isso, analisemos as derivadas de g. Já sabemos que $g'(1) = 0$. Quanto à segunda derivada, temos $g''(x) = \frac{1}{x^3}$. Logo,g'' é contínua em $[1,2]$ e $g''(1) = 1 \neq 0$. Daqui resulta que o teorema 2.6 é aplicável com $p = 2$ e a ordem de convergência é 2. Quanto ao coeficiente assimptótico de convergência, temos $k_\infty = \frac{|g''(1)|}{2} = \frac{1}{2}$.

2.4 Método de Newton

No parágrafo anterior vimos que o método do ponto fixo tem um vasto domínio de aplicação e permite, na maioria dos casos, obter boas aproximações de raizes de equações. No entanto, vimos também que, em geral, aquele método garante apenas primeira ordem de convergência (ordens superiores só se obtêm, de acordo com o teorema 2.6, se algumas derivadas da função iteradora se anularem no ponto fixo).

O método de Newton, que vamos estudar neste prarágrafo, tem a importante vantagem de proporcionar, em geral, convergência de segunda ordem (quadrática). É um dos métodos mais frequentemente utilizados, já que combina a rapidez de convergência com a simplicidade do processo iterativo. Mais adiante, veremos que o método de Newton pode ser encarado como um caso particular do método do ponto fixo. Por agora, vamos introduzi-lo através da sua interpretação geométrica.

2.4.1 Interpretação geométrica do método de Newton

Seja f uma função continuamente diferenciável num certo intervalo $[a,b]$. Suponhamos que nesse intervalo a função f tem uma única raiz e que a sua derivada não se anula ($f'(x) \neq 0, \forall x \in [a,b]$).

Sendo x_0 um ponto arbitrário de $[a,b]$, podemos traçar a tangente ao gráfico de f que passa pelo ponto $(x_0, f(x_0))$ (ver fig. 11). Sendo $f'(x_0) \neq 0$, por condição, essa recta intersecta o eixo das abcissas num certo ponto $(x_1, 0)$. Para determinar x_1, comecemos por escrever a equação da tangente

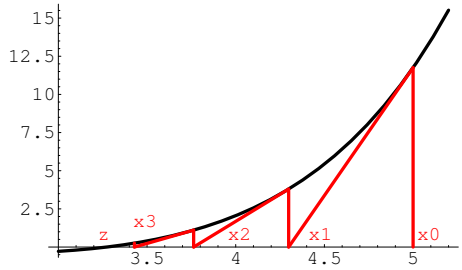

Figure 11: Interpretação geométrica do método de Newton

.

ao gráfico de f:

$$y - f(x_0) = f'(x_0)(x - x_0). \tag{2.23}$$

Igualando y a 0 na equação (2.23), obtém-se a abcissa x_1 procurada:

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}. \tag{2.24}$$

As iteradas seguintes obtêm-se do mesmo modo. Mais precisamente, para determinar x_2, traça-se a tangente ao gráfico de f que passa pelo ponto $(x_1, f(x_1))$ e procura-se o ponto onde essa recta intersecta o eixo das abcissas; e assim sucessivamente. Obtém-se deste modo uma sucessão de pontos $\{x_k\}, k = 0, 1, 2, \ldots,$ que podem ser calculados pela fórmula de recorrência

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}. \tag{2.25}$$

A interpretação geométrica, representada na Fig.11, sugere-nos que esta sucessão converge para a raiz z da equação considerada. Nos parágrafos seguintes, vamos demonstrar analiticamente que de facto assim é.

2.4.2 Estimativa do erro do método de Newton

Em primeiro lugar, vamos deduzir uma fórmula que nos permite majorar o erro de cada iterada do método de Newton, admitindo que é conhecido um majorante do erro da iterada anterior. Vamos supor que a função f satisfaz no intervalo $[a, b]$ as condições formuladas no parágrafo anterior (f é continuamente diferenciável em $[a, b]$ e a sua derivada não se anula neste intervalo). Além disso, vamos supor que a segunda derivada de f também é contínua neste intervalo. Seja $\{x_n\}, n = 0, 1, 2, \ldots$ a sucessão das iteradas

do método de Newton (que se consideram pertencentes ao intervalo $[a, b]$). Se desenvolvermos f em série de Taylor, em torno de x_k, obtém-se

$$f(x) = f(x_k) + (x - x_k)f'(x_k) + \frac{(x - x_k)^2}{2} f''(\xi_k), \tag{2.26}$$

onde $\xi_k \in int(x_k, x)$. Escrevendo o desenvolvimento (2.26) com $x = z$, obtém-se

$$f(z) = f(x_k) + (z - x_k)f'(x_k) + \frac{(z - x_k)^2}{2} f''(\xi_k) = 0, \tag{2.27}$$

onde $\xi_k \in int(x_k, z)$. Uma vez que, por condição, $f'(x_k) \neq 0$, podemos dividir ambos os membros de (2.27) por $f'(x_k)$, obtendo assim

$$\frac{f(x_k)}{f'(x_k)} + (z - x_k) + \frac{(z - x_k)^2}{2f'(x_k)} f''(\xi_k) = 0. \tag{2.28}$$

Atendendo à fórmula iteradora do método de Newton (2.25), da equação (2.28) resulta que

$$z - x_{k+1} = -\frac{(z - x_k)^2}{2f'(x_k)} f''(\xi_k). \tag{2.29}$$

A igualdade (2.29) dá-nos a relação que procurávamos entre o erro de $x_{k+1}(e_{k+1})$ e o erro de $x_k(e_k)$. No segundo membro desta desigualdade aparece o valor $f''(\xi_k)$, o qual não podemos calcular exactamente, já que sabemos apenas que ξ_k é um ponto situado entre x_k e z. Por isso, para podermos majorar o erro absoluto de $x_k(|e_k|)$, precisamos de majorar o módulo da segunda derivada de f (que se supõe contínua). Introduzindo a notação

$$M = \max_{x \in [a,b]} |f''(x)| \tag{2.30}$$

da igualdade (2.29) obtem-se a seguinte relação:

$$|e_{k+1}| \leq |e_k|^2 \frac{M}{2|f'(x_k)|}. \tag{2.31}$$

Saliente-se que na desigualdade (2.31) $|e_{k+1}|$ é comparado com o quadrado de $|e_k|$, o que indica um rápido decrescimento do erro. Seja

$$\mu = \min_{x \in [a,b]} |f'(x)|. \tag{2.32}$$

Então, a desigualdade (2.31) pode ser reforçada, substituindo $|f'(x_k)|$ por μ:

$$|e_{k+1}| \leq |e_k|^2 \frac{M}{2\mu}. \tag{2.33}$$

Nesta última desigualdade o segundo membro não depende de k. Na prática, usam-se frequentemente as fórmulas (2.31) e (2.33)para obter uma estimativa de $|e_{k+1}|$.

Exemplo 2.10. Consideremos a equação $f(x) = cosx - 2x = 0$, que já foi analisada no Exercício 2.3. Vamos obter aproximações da raiz desta equação, situada no intervalo $[0.4, 0.5]$, aplicando o método de Newton, e majorar o erro dessas aproximações. Sendo $x_0 = 0.4$, da fórmula (2.25) obtém-se que $x_1 = 0.45066547$ e $x_2 = 0.45018365$. Vamos agora obter majorantes de $|e_1|$ e $|e_2|$. Em primeiro lugar, note-se que $|e_0| \leq 0.5 - 0.4 = 0.1$. Para podermos aplicar a desigualdade (2.31) é necessário majorar $|f''(x)|$ e minorar $|f'(x)|$. Temos $f'(x) = -\sin x - 2$ e $f''(x) = -\cos x$, logo

$$\mu = \min_{x\in[0.4,0.5]} |f'(x)| = \min_{x\in[0.4,0.5]} |2 + \sin x| = 2 + \sin 0.4 = 2.389;$$

$$M = \max_{x\in[0.4,0.5]} |f''(x)| = \max_{x\in[0.4,0.5]} |\cos x| = \cos 0.4 = 0.921.$$

Por conseguinte, da desigualdade (2.33) resulta a seguinte majoração para o erro absoluto de x_1:

$$|e_1| \leq \frac{M}{2\mu}|e_0|^2 \leq \frac{0.921}{2 \times 2.389} 0.01 = 0.001927.$$

Em relação ao erro de x_2, obtem-se, do mesmo modo:

$$|e_2| \leq \frac{M}{2\mu}|e_1|^2 \leq \frac{0.921}{2 \times 2.389} 0.001927 = 0.696 \times 10^{-7}.$$

Vemos assim que, com duas iteradas apenas, conseguiu-se obter um resultado de alta precisão. Para terminar este exemplo apresentamos na Tabela 2.2 uma comparação dos resultados obtidos para esta equação pelos métodos de Newton e do ponto fixo.

k	x_k(Ponto Fixo)	$\|e_k\|$	x_k(Newton)	$\|e_k\|$
0	0.4	0.0501	0.4	0.0501
1	0.46053	0.0105	0.45066547	0.48×10^{-3}
2	0.44791	0.0022	0.45018365	0.4×10^{-7}

Tabela 2.2. Resultados numéricos do método de Newton e do método do ponto fixo (Exemplo 2.10).

2.4.3 Condições suficientes de convergência do método de Newton

Até agora, analisámos o erro do método de Newton, partindo do princípio que a aproximação inicial é tal que as iteradas convergem para a raiz procurada. No entanto, nem sempre é fácil prever, para uma dada aproximação inicial, se o método de Newton vai convergir, e para que raiz ele vai convergir (se a equação tiver várias raizes). Neste parágrafo, vamos enunciar um conjunto de condições que, uma vez satisfeitas, garantem que, se a aproximação inicial x_0 do método de Newton, pertencer a um certo intervalo, então o método converge para a raiz da equação que se encontra nesse intervalo.

Teorema 2.7 Seja f uma função de variável real que satisfaz as seguintes condições:

1. f f é contínua em $[a, b]$;$f(a)f(b) < 0$.
2. $f \in C^1([a, b])$,$f'(x) \neq 0$ em [a,b].
3. $f \in C^2([a, b])$,$f''(x) \geq 0$ ou $f''(x) \leq 0$ em $[a, b]$.
4. $\frac{|f(a)|}{|f'(a)|} < b - a$, $\frac{|f(b)|}{|f'(b)|} < b - a$.

Então, qualquer que seja a aproximação inicial $x_0 \in [a, b]$, o método de Newton converge para a única raiz z de f em $[a, b]$.

Nalgumas situações tem interesse também a seguinte variante deste teorema:

Teorema 2.8 Suponhamos que f satisfaz as primeiras 3 condições do teorema 2.7 . Então, se a aproximação inicial x_0 for tal que $f(x_0)f''(x) \geq 0, \forall x \in [a, b]$, o método de Newton converge para a única raiz z de f em $[a, b]$ e a sucessão das iteradas é **monótona**.

Não iremos fazer a demonstração completa destes teoremas, mas apenas investigar o significado e a razão de ser de cada uma das condições. As primeiras condições, como sabemos pelos Teoremas 2.1 e 2.2, garantem que a função considerada tem um único zero em $[a, b]$. Além disso, a segunda condição é essencial para o método de Newton, pois se ela não se verificar (isto é, se a derivada de f se anular nalgum ponto de $[a, b]$), o método de Newton pode não ser aplicável. Quanto à terceira condição, ela significa que no domínio considerado a segunda derivada de f não muda de sinal ou, por outras palavras, a função não tem pontos de inflexão no intervalo considerado. Para entendermos a razão de ser desta condição, analisemos o seguinte exemplo.

Exemplo 2.11. Consideremos $f(x) = x^3 - x$, no intervalo $[-0.5, 05]$. Neste intervalo, a função é continuamente diferenciável, com $f'(x) = 3x^2 - 1$. Além disso, f tem sinais opostos nos extremos do intervalo ($f(-0.5) = 3/8$, $f(0.5) = -3/8$) e f' não se anula (é sempre negativa). Por conseguinte, as duas primeiras condições do teorema 2.8 estão satisfeitas no intervalo $[-0.5, 0.5]$. Em relação à terceira condição, temos $f''(x) = 6x$, logo, $f''(x)$ muda de sinal em $x = 0$, pelo que esta condição não é satisfeita. Vejamos agora através de um exemplo que a convergência do método de Newton não está garantida para qualquer aproximação inicial no intervalo $[-0.5, 0.5]$. Seja $x_0 = 1/\sqrt{5}$. Embora este ponto pertença ao intervalo considerado, verifica-se imediatamente que as iteradas do método de Newton, neste caso, formam uma sucessão divergente: $x_1 = -1/\sqrt{5}, x_2 = 1/\sqrt{5}, x_3 = -1/\sqrt{5}, \ldots$.

Vejamos agora um exemplo que ilustra a importância da condição 4.

Exemplo 2.12. Seja $f(x) = \ln x$, que tem uma única raiz em $z = 1$. Se considerarmos, por exemplo, o intervalo $[0.5, 3]$ vemos que neste intervalo estão satisfeitas as primeiras 3 condições dos teoremas 2.7 e 2.8:

1. $f(0.5)f(3) < 0$;
2. $f'(x) = 1/x \neq 0, \forall x \in [0.5, 3]$;
3. $f''(x) = -1/x^2 < 0, \forall x \in [0.5, 3]$.

No entanto, a convergência do método de Newton não está assegurada para qualquer aproximação inicial neste intervalo. Se tomarmos, por exemplo, $x_0 = 3$, temos $x_1 = 3 - 3\,ln(3) < 0$, pelo que o método de Newton não pode ser aplicado ($f(x)$ não está definida para $x < 0$).

Neste caso, é fácil ver que falha a condição 4 do teorema 2.7. Com efeito, temos

$$\frac{|f(3)|}{|f'(3)|} = 3ln(3) > 3 - 0.5 = 2.5.$$

Para que o método de Newton convergisse para a raiz procurada, neste caso, deveríamos escolher, por exemplo, $x_0 = 0.5$. Nesse caso, a convergência resultaria do Teorema 2.8, já que se verifica $f(0.5)f''(x) > 0, \forall x \in [0.5, 3]$.

Sobre o significado geométrico da condição 4 podemos dizer o seguinte: se ela se verificar, então, tomando $x_0 = a$, a iterada x_1 satisfaz $|x_1 - a| = \frac{|f(a)|}{|f'(a)|} < |b - a|$ (ou seja, a distância de x_1 a a é menor que o comprimento do intervalo $[a, b]$). Logo, x_1 pertence a esse intervalo. Continuando este raciocínio, pode mostrar-se que todas as iterdas seguintes continuam a pertencer ao intervalo $[a, b]$. Se começarmos o processo iterativo a partir de

$x_0 = b$ e utilizarmos a condição $\frac{|f(b)|}{|f'(b)|} < |b-a|$, um raciocínio semelhante leva-nos à mesma conlusão, isto é, a condição 4 garante que **se** $x_0 \in [a,b]$**, todas as iteradas do método de Newton se mantêm dentro desse intervalo**.

2.4.4 Ordem de convergência do método de Newton

O método de Newton também pode ser encarado como um caso particular do método do ponto fixo. Essa abordagem tem a vantagem de nos permitir analisar a convergência do método de Newton com base nos resultados teóricos sobre o método do ponto fixo.

Consideremos a equação $f(x) = 0$ e suponhamos que ela tem uma única raiz num certo intervalo $[a,b]$. Admitamos ainda que $f \in C^1([a,b])$ e que $f'(x) \neq 0, \forall x \in [a,b]$. Então a equação considerada é equivalente a

$$x - \frac{f(x)}{f'(x)} = x. \tag{2.34}$$

Se definirmos a função iteradora

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

podemos dizer que a equação (2.34) é a equação dos pontos fixos de g. Logo, as raizes de f, que também são pontos fixos de g, podem ser aproximadas pelo processo iterativo

$$x_{k+1} = g(x_k) = x_k - \frac{f(x_k)}{f'(x_k)}. \tag{2.35}$$

Verificamos portanto que este método é idêntico ao método Newton, aplicado à função $f(x)$. Logo, para determinar a ordem de convergência do método de Newton, basta determinar, com base no teorema 2.6, a ordem de convergência da sucessão gerada por esta função iteradora. Para isso, comecemos por calcular as suas derivadas. Temos

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

Daqui, tomando em consideração que $f(z) = 0$ e $f'(z) \neq 0$, resulta que $g'(z) = 0$. Quanto à segunda derivada de g, temos

$$g''(x) = \frac{\left(f'(x)f''(x) + f(x)f'''(x)\right) f'(x)^2 - f(x)f''(x)(f'(x)^2)'}{f'(x)^4}.$$

Logo,

$$g''(z) = \frac{f''(z)}{f'(z)}.$$

Podemos assim concluir o seguinte:

a) Se $f''(z) \neq 0$, então $g''(z) \neq 0$ (uma vez que, por condição, $f'(z) \neq 0$). Nesse caso, de acordo com o teorema 2.6, o método de Newton (ou seja, o método do ponto fixo com a função iteradora $g(x) = x - \frac{f(x)}{f'(x)}$) tem ordem de convergência 2 (convergência quadrática). Além disso, o coeficiente assimptótico de convergência é dado por

$$k_\infty = \frac{|f''(z)|}{2|f'(z)|}$$

.

b) Se $f''(z) = 0$, então $g''(z) = 0$ e o método de Newton tem ordem de convergência, pelo menos, 3 (para saber qual a ordem concreta é necessário analisar as derivadas de ordem superior de g.

Exemplo 2.13. Consideremos a equação $f(x) = x^3 - x = 0$. Uma das suas raizes é $z = 0$. Se aplicarmos o método de Newton ao cálculo aproximado desta raiz, isso equivale a utilizar o método do ponto fixo com a função iteradora

$$g(x) = x - \frac{f(x)}{f'(x)} = \frac{2x^3}{3x^2 - 1}.$$

Analisemos a ordem do método neste caso. Para isso comecemos por verificar que $f'(0) = -1 \neq 0$ e $f''(0) = 0$. Então, de acordo com a análise que acabámos de realizar, o método deve ter ordem, pelo menos, 3. Vejamos qual é, de facto, a ordem de convergência. Sabemos que $g''(0) = \frac{f''(0)}{f'(0)} = 0$. Para determinar $g'''(0)$, observemos que a função g admite, em torno de $z = 0$, um desenvolvimento em série de Taylor da forma

$$g(x) = \frac{-2x^3}{1 - 3x^2} = -2x^3 + O(x^5),$$

de onde se conclui que $g'''(x) = -12 + O(x^2)$(quando $x \to 0$), pelo que $g'''(0) = -12$. Temos, portanto, convergência de ordem 3 . O coeficente assimptótico de convergência, de acordo com o Teorema 2.6, é

$$k_\infty = \frac{|g'''(0)|}{3!} = 2.$$

2.5 Método da Secante

Terminaremos este capítulo com a análise do método da secante. Tal como no caso do método de Newton, a fórmula iteradora deste método pode ser deduzida a partir da sua interpretação geométrica.

2.5.1 Interpretação geométrica do método da secante

Seja f uma função de variável real, contínua num certo intervalo $[a, b]$, e suponhamos que f tem nesse intervalo um único zero z. Para aplicar o método da secante, escolhem-se dois números, x_0 e x_1, no intervalo $[a, b]$, e considera-se a recta que passa pelos pontos $(x_0, f(x_0))$ e $(x_1, f(x_1))$ (secante ao gráfico de f). A equação dessa recta é

$$y - f(x_1) = \frac{f(x_1) - f(x_0)}{x_1 - x_0}(x - x_0). \tag{2.36}$$

Depois, determina-se o ponto onde esta recta intersecta o eixo das abcissas. Esse ponto existe, desde que $f(x_0) \neq f(x_1)$, condição que consideramos satisfeita. Desigando por x_2 a abcissa desse ponto, obtém-se a seguinte equação para x_2:

$$x_2 = x_1 - \frac{x_1 - x_0}{f(x_1) - f(x_0)} f(x_1). \tag{2.37}$$

Deste modo, a nova aproximação da raiz x_2 fica definida a partir de x_0 e x_1. A fórmula que nos permite determinar cada aproximação x_{k+1} a partir das duas anteriores (x_k e x_{k-1}) é análoga a (2.37):

$$x_{k+1} = x_k - \frac{x_k - x_{k-1}}{f(x_k) - f(x_{k-1})} f(x_k), \qquad k = 1, 2, \ldots \tag{2.38}$$

Um exemplo de aplicação do método da secante está representado na fig.12.

2.5.2 Estimativa do erro do método da secante

No caso do método de Newton, vimos que o erro de cada iterada pode ser estimado a partir do erro da iterada anterior e das propriedades da função f. No caso do método da secante, é de realçar uma diferença fundamental: é que cada iterada depende das duas iteradas anteriores, e não apenas da última. Neste caso, diz-se que temos um método iterativo a dois passos. Sendo assim, é natural que o erro de cada iterada do método da secante possa ser determinado a partir dos erros das duas últimas iteradas. Suponhamos

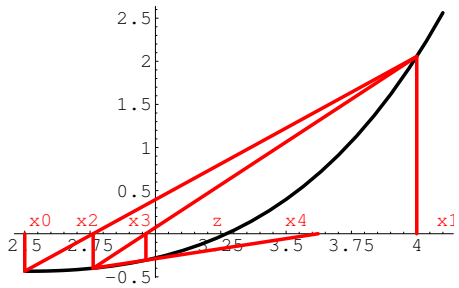

Figure 12: Interpretação geométrica do método da secante.

então que x_{m-1} e x_m são duas iteradas consecutivas do método da secante. A iterada seguinte, x_{m+1}, pode ser determinada através da fórmula (2.38). Representemos os erros de x_{m-1} e x_m por e_{m-1} e e_m, respectivamente; isto é, $e_{m-1} = z - x_{m-1}$ e $e_m = z - x_m$. Além disso, suponhamos que a função f é duas vezes continuamente diferenciável num intervalo $I \subset [a, b]$ que contém x_{m-1}, x_m, x_{m+1} e z e que f' não se anula em I. Então, usando resultados da teoria da interpolação (ver, por exemplo, [2]), pode mostrar-se que e_{m+1} (erro de x_{m+1}) satisfaz a desigualdade:

$$e_{m+1} = -\frac{f''(\xi_m)}{2f'(\eta_m)} e_m e_{m-1}, \tag{2.39}$$

onde ξ_m e η_m representam pontos que pertencem ao intervalo I acima referido. Note-se que a fórmula (2.39) é muito parecida com a fórmula (2.29) para o erro do método de Newton. A diferença é que, como seria de esperar, neste caso, o erro da nova iterada do método da secante é avaliado a partir do produto dos erros das duas últimas iteradas, enquanto que no método de Newton o erro da nova iterada é avaliado a partir do quadrado do erro da iterada anterior.

Na prática, para usar a fórmula (2.39), à semelhança do que fizemos no caso do método de Newton, convém majorar em I o módulo da segunda derivada de f e minorar o módulo da sua segunda derivada. Para simplificar, suponhamos que $I = [a, b]$ e seja

$$M = \max_{x \in [a,b]} |f''(x)|,$$

$$\mu = \min_{x \in [a,b]} |f'(x)|.$$

Então, da fórmula (2.39) resulta imediatamente a seguinte majoração para

o erro absoluto do método da secante:

$$|e_{m+1}| \leq \frac{M}{2\mu}|e_m||e_{m-1}|. \tag{2.40}$$

Normalmente, os erros absolutos das suas iteradas iniciais, $|e_0|$ e $|e_1|$, são majorados pelo comprimento do intervalo $[a, b]$, isto é, são evidentes as desigualdades $|e_0| < |b-a|$ e $|e_1| < |b-a|$. A partir daí, os erros das sucessivas iteradas são majorados por recorrência, isto é: o erro $|e_2|$ majora-se a partir dos erros $|e_0|$ e $|e_1|$; o erro $|e_3|$ majora-se a partir dos erros $|e_1|$ e $|e_2|$; e assim sucessivamente.

Exemplo 2.14. Consideremos mais uma vez a equação $f(x) = \cos(x) - 2x = 0$, que tem uma raiz no intervalo $[0.4, 0.5]$. Para aproximar essa raiz pretende-se usar o método da secante.

1. Tomando como aproximações iniciais os pontos $x_0 = 0.5$ e $x_1 = 0.4$, calcule as iteradas x_2 e x_3 pelo método da secante.

Aplicando a fórmula (2.38), temos

$$x_2 = x_1 - \frac{x_1 - x_0}{f(x_1) - f(x_0)} f(x_1) = 0.449721; \tag{2.41}$$

$$x_3 = x_2 - \frac{x_2 - x_1}{f(x_2) - f(x_1)} f(x_2) = 0.450188. \tag{2.42}$$

2. Obtenha majorantes do erro absoluto de x_0, x_1, x_2 e x_3.

O caminho mais fácil seria majorar $|e_0|$ e $|e_1|$ pelo comprimento do intervalo: $|b - a| = |0.5 - 0.4| = 0.1$. O majorante pode, no entanto, ser um pouco melhorado, se tivermos em conta o sinal de f em cada um dos pontos x_i calculados. Para tal, observemos a tabela:

i	x_i	$f(x_i)$
0	0.5	-0.122
1	0.4	0.121
2	0.449721	0.0011
3	0.450188	-0.00001

Da observação da tabela conclui-se que os pontos x_1 e x_2 se encontram à esquerda da raiz z (onde f é positiva), enquanto x_0 e x_3 se encontram à direita (onde f é negativa). Sendo assim, para os erros de x_0 e x_1 obtêm-se os seguintes majorantes:

$$|e_0| = |z - x_0| \leq |x_2 - x_0| = |0.449721 - 0.5| = 0.050258,$$

$$|e_1| = |z - x_1| \leq |x_3 - x_1| = |0.450188 - 0.4| = 0.050188.$$

Recordemos do exemplo 2.10 que neste caso se tem $M = 0.921$, $\mu = 2.389$. Daqui, pela estimativa (2.40), obtém-se

$$|e_2| \leq \frac{M}{2\mu}|e_1||e_0| \leq 0.193 \times 0.050188 \times 0.050258 = 0.4868 \times 10^{-3}, \quad (2.43)$$

$$|e_3| \leq \frac{M}{2\mu}|e_2||e_1| \leq 0.193 \times 0.4868 \times 10^{-3} \times 0.050188 = 0.4715 \times 10^{-5}. \quad (2.44)$$

Vemos assim que, ao fim de duas iterações, o método da secante nos proporciona uma aproximação com um erro da ordem de 10^{-5}. No caso de método de Newton, com o mesmo número de iterações, obtém-se um erro da ordem de 10^{-7} (ver exemplo 2.10). Assim, este exemplo sugere que o método de Newton converge mais rapidamente que o da secante. Por outro lado, no mesmo exemplo, a precisão que se consegue obter com duas iteradas do método do ponto fixo é da ordem de 10^{-2}. Logo, a julgar por este exemplo, é de esperar que a ordem de convergência do método da secante esteja entre a ordem do método do ponto fixo no caso geral(1) e a do método de Newton (2). Esta conjectura é confirmada pelos resultados que vamos expor na próxima secção.

2.5.3 Convergência do método da secante

Com base na estimativa do erro que foi deduzida no parágrafo anterior pode provar-se o seguinte teorema sobre a convergência do método da secante (ver demonstração em [2]).

Teorema 2.9 Seja f uma função duas vezes continuamente diferenciável numa vizinhança de z e tal que $f'(z) \neq 0$. Então, se os valores iniciais x_0 e x_1 forem suficientemente próximos de z, a sucessão $\{x_m\}$ gerada pelo método da secante converge para z.

O exemplo numérico que foi analisado na secção anterior sugere que o método da secante é mais rápido que o método do ponto fixo (que , no caso geral, tem ordem 1), mas menos rápido que o de Newton, que tem convergência quadrática. Assim, se $\{x_m\}$ for uma sucessão gerada pelo método da secante, existe um número real p, tal que $1 < p < 2$, para o qual se verifica

$$\lim_{m\to\infty} \frac{|z - x_{m+1}|}{|z - x_m|^p} = K_\infty, \quad (2.45)$$

onde K_∞ é uma constante positiva. (De acordo com a designação introduzida na secção 2.3.8, esta constante designa-se coeficiente assimptótico de

convergência.) Mais precisamente, pode provar-se (ver detalhes em [1]) que, no caso do método da secante, temos

$$p = \frac{1+\sqrt{5}}{2} \approx 1.6,$$

isto é, a ordem de convergência é dada pelo chamado *número de ouro* (sobre a importância dessse número e as suas numerosas aplicações ver, por exemplo, [3]).

O Teorema 2.9, acima enunciado, tem a desvantagem de não ser facilmente aplicável. Na realidade, o que significa a condição "se x_0 e x_1 forem suficientemente próximos de z?" Para a prática, são bastante mais úteis proposições, do tipo dos Teoremas 2.7 e 2.8, que nos dêm condições suficientes para a convergência do método da secante, desde que as aproximações iniciais pertençam a um dado intervalo. Passamos a enunciar tais teoremas.

Teorema 2.10. Se forem satisfeitas as condições do teorema 2.7, então o método da secante converge para a raiz z de f em $[a, b]$, quaisquer que sejam as aproximações iniciais x_0,x_1, pertencentes a $[a, b]$.

Teorema 2.11. Se forem satisfeitas as primeiras 3 condições do teorema 2.7, e se as aproximações iniciais satisfizerem $f(x_0)f''(x) \geq 0$,$f(x_1)f''(x) \geq 0$, $\forall x \in [a, b]$, então o método da secante converge para a raiz z de f em $[a, b]$.

3 Métodos Numéricos para Sistemas de Equações

3.1 Normas Matriciais

Neste capítulo trataremos de métodos computacionais para a resolução de sistemas de equações (lineares e não lineares). Para a análise do erro destes métodos, necessitaremos frequentemente de recorrer a normas vectoriais e matriciais, pelo que começaremos por fazer uma breve introdução sobre este tema.

Seja E um espaço linear. Tipicamente, teremos $E = I\!R^n$ (vectores reais de dimensão n) ou $E = I\!R^{n\times n}$ (matrizes reais de n linhas e n colunas).

Definição 3.1 . Uma aplicação ϕ de E em $I\!R_0^+$ diz-se uma norma se satisfizer as seguintes condições:

1. $\phi(x) \geq 0$, $\forall x \in E$, sendo $\phi(x) = 0$, se e só se $x = 0$.

2. $\phi(\lambda x) = |\lambda|\phi(x)$, $\forall x \in E$, $\lambda \in I\!R$.

3. $\phi(x+y) \leq \phi(x) + \phi(y), \forall x, y \in E$.

Começaremos por rever alguns exemplos de normas em $I\!R^n$. Como habitualmente, representaremos qualquer elemento de $I\!R^n$ por $x = (x_1, x_2, \ldots, x_n)$, onde $x_i \in I\!R$.

1. Norma do máximo. $\phi(x) = \|x\|_\infty = \max_{i=1,\ldots,n} |x_i|$.

2. Norma 1. $\phi(x) = \|x\|_1 = \sum_{i=1}^n |x_i|$.

3. Norma euclidiana. $\phi(x) = \|x\|_2 = \sqrt{\sum_i^n |x_i|^2}$.

4. Norma p. $\phi(x) = \|x\|_p = (\sum_i^n |x_i|^p)^{\frac{1}{p}}$, $p \geq 1$.

Note-se que a norma 1 e a norma euclidiana são casos particulares das normas p, respectivamente, com $p = 1$ e $p = 2$. A norma do máximo é um caso limite, quando $p \to \infty$. Pode provar-se que todos os exemplos indicados definem normas, isto é, satisfazem as 3 condições acima formuladas.

Passamos agora a considerar o caso de $E = I\!R^{n\times n}$. Neste caso, os elementos de E são matrizes de n linhas e n colunas, por exemplo:

$$A = \begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ a_{n1} & a_{n2} & \ldots & a_{nn} \end{bmatrix}.$$

Represente-se por $\|.\|_v$ uma norma qualquer em $I\!R_n$. Então podemos definir uma norma $\|.\|_M$ em E pela seguinte igualdade:

$$\|A\|_M = sup_{x\in I\!R^n, x\neq 0} \frac{\|A\,x\|_v}{\|x\|_v}. \tag{3.1}$$

Neste caso, diz-se que a norma $\|.\|_M$ é a **norma matricial,induzida pela norma vectorial** $\|.\|_v$.

Esta forma de construir normas matriciais permite-nos associar uma norma matricial a cada uma das nomas vectoriais introduzidas. As normas matriciais introduzidas deste modo gozam de duas propriedades muito úteis:

1. A norma $\|.\|_M$ é compatível com a norma $\|.\|_v$, isto é,

$$\|A\,x\|_v \leq \|A\|_M\,\|x\|_v, \quad \forall x \in I\!R^n, \quad \forall A \in I\!R^{n\times n}. \tag{3.2}$$

Esta propriedade é uma consequência imediata da fórmula (3.1).

2. A norma $\|.\|_M$ é regular, isto é,

$$\|A\,B\|_M \leq \|A\|_M\,\|B\|_M, \quad \forall A, B \in I\!R^{n\times n}. \tag{3.3}$$

São as seguintes as normas matriciais, induzidas pelas normas vectoriais mais correntes.

1. A norma matricial induzida pela **norma do máximo** chama-se **norma por linha** e é definida pela seguinte fórmula:

$$\|A\|_\infty = \max_{i=1,\ldots,n} \sum_{j=1}^{n} |a_{ij}|. \tag{3.4}$$

2. A norma matricial induzida pela **norma 1** chama-se **norma por coluna** e é definida pela seguinte fórmula:

$$\|A\|_1 = \max_{j=1,\ldots,n} \sum_{i=1}^{n} |a_{ij}|. \tag{3.5}$$

3. A norma matricial induzida pela **norma euclidiana** é definida pela seguinte fórmula:

$$\|A\|_2 = \sqrt{\rho(A^T A)}. \tag{3.6}$$

Na fórmula (3.6) o símbolo $\rho(A)$ representa o **raio espectral** de A, que se define como o máximo dos módulos dos valores próprios de A:

$$\rho(A) = \max_{i=1,\ldots,n} |\lambda_i|. \tag{3.7}$$

Note-se que, se A for uma matriz simétrica, isto é, se $A^T = A$, então verifica-se

$$\|A\|_2 = \sqrt{\rho(A^T A)} = \sqrt{\rho(A^2)} = \rho(A), \tag{3.8}$$

isto é, o raio espectral de A coincide com a sua norma euclidiana.

Exemplo 3.1. Consideremos a matriz

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

Neste caso, as normas de A são

$$\begin{aligned} \|A\|_\infty &= \max(6, 8, 6) = 8; \\ \|A\|_1 &= \max(5, 5, 10) = 10; \\ \|A\|_2 &= \sqrt{36.3} \approx 6.02; \end{aligned} \tag{3.9}$$

Para calcular $\|A\|_2$, foram necessários os seguintes cálculos auxiliares:

$$B = A^T A = \begin{bmatrix} 9 & 3 & 4 \\ 3 & 11 & 6 \\ 4 & 6 & 34 \end{bmatrix}. \tag{3.10}$$

Os valores próprios de B são $\lambda_1 = 6.8$,$\lambda_2 = 10.9$, $\lambda_3 = 36.3$.

Para terminar este exemplo, calculemos o raio espectral de A. Os valores próprios de A são $\tilde{\lambda}_1 = 3.69$,$\tilde{\lambda}_{2,3} = 2.15 \pm i3.07$. Logo,$|\tilde{\lambda}_2| = |\tilde{\lambda}_3| = 3.75$. Por conseguinte, $\rho(A) = \max(3.69, 3.75) = 3.75$.

Note-se que, no exemplo 3.1, qualquer das normas de A é maior que o raio espectral da matriz. Tal não acontece por acaso. Para terminar este parágrafo, vamos verificar uma relação existente entre as normas e o raio espectral das matrizes.

Teorema 3.1. Seja $A \in I\!R^{n\times n}$. Então, qualquer que seja a norma matricial p, associada a uma certa norma vectorial em $I\!R^n$, verifica-se

$$r_\sigma(A) \leq \|A\|_p. \tag{3.11}$$

Demonstração. Seja x um vector próprio de A, associado ao valor próprio λ, tal que $|\lambda| = r_\sigma(A)$. Então

$$\|A\,x\|_p = \|\lambda\,x\|_p = |\lambda|\,\|x\|_p. \tag{3.12}$$

Então podemos afirmar que

$$\|A\|_p = \sup_{x\in I\!R^n, x\neq 0} \frac{\|A\,x\|_p}{\|x\|_p} \geq |\lambda|, \tag{3.13}$$

de onde resulta a afirmação do teorema.

3.2 Condicionamento de Sistemas Lineares

Como vimos no capítulo 1, um dos aspectos mais importantes a ter em consideração quando se quando se analisam métodos numéricos para a solução de um determinado problema é a sensibilidade desses métodos em relação a pequenos erros nos dados. Se for dado um certo sistema linear

$$A\,x = b$$

os dados são o segundo membro do sistema ($b \in I\!R^n$) e a matriz ($A \in I\!R^{n\times n}$), que se supõe não singular. Vamos, portanto, analisar até que ponto um

pequeno erro, em termos relativos, do vector b ou da matriz A, pode afectar a solução do sistema. Para isso, representemos por $\tilde{A}$ uma perturbação da matriz A.

$$\tilde{A} \approx A. \tag{3.14}$$

Analogamente, representemos por $\tilde{b}$ uma perturbação do segundo membro do sistema:

$$\tilde{b} \approx b. \tag{3.15}$$

Se substituirmos A por $\tilde{A}$ e b por $\tilde{b}$ no sistema inicial, obteremos um novo sistema, cuja solução representaremos por $\tilde{x}$.

Como vimos no capítulo 1, para medir a precisão dos resultados numéricos é essencial o conceito de erro relativo. Vamos, portanto, designar por erro relativo de um vector $\tilde{x}$ (numa certa norma vectorial p) o quociente

$$\|\delta x\|_p = \frac{\|\tilde{x} - x\|_p}{\|x\|_p}; \tag{3.16}$$

analogamente, designaremos por erro relativo de uma matriz $\tilde{A}$ (na norma matricial p)o quociente

$$\|\delta A\|_p = \frac{\|\tilde{A} - A\|_p}{\|A\|_p}. \tag{3.17}$$

O nosso objectivo é, escolhendo uma certa norma vectorial e a norma matricial associada, estimar o erro relativo $\|\delta x\|_p$ em função dos erros relativos $\|\delta b\|_p$ e $\|\delta A\|_p$.

Definição 3.2. Um sistema linear diz-se *bem condicionado* se e só se a pequenos erros relativos do segundo membro e da matriz correspondem pequenos erros relativos da solução.

Para analisarmos o problema do condicionamento, comecemos por considerar o caso mais simples em que $\|\delta A\|_p = 0$. Nesse caso, temos

$$A\,\tilde{x} = \tilde{b} \tag{3.18}$$

De (3.18) obtém-se

$$\tilde{x} - x \;=\; A^{-1}\,(\tilde{b} - b) \tag{3.19}$$

Por conseguinte, qualquer que seja a norma vectorial escolhida, é válida a seguinte estimativa para o erro absoluto de $\tilde{x}$:

$$\|\tilde{x} - x\|_p \;\leq\; \|A^{-1}\|_p\,\|\tilde{b} - b\|_p. \tag{3.20}$$

Por outro lado, da igualdade $A\,x = b$ obtém-se

$$\|b\|_p \leq \|A\|_p \, \|x\|_p, \tag{3.21}$$

logo

$$\frac{1}{\|x\|_p} \leq \frac{\|A\|_p}{\|b\|_p}. \tag{3.22}$$

Multiplicando cada um dos membros de 3.20 pelo membro correspondente de 3.22, obtém-se

$$\frac{\|\tilde{x} - x\|_p}{\|x\|_p} \leq \|A\|_p \, \|A^{-1}\|_p \frac{\|\tilde{b} - b\|_p}{\|b\|_p}. \tag{3.23}$$

Obtivemos assim a estimativa que procurávamos para o erro relativo da solução, em função do erro relativo do segundo membro. A desigualdade (3.23) leva-nos à definição de número de condição de uma matriz.

Definição 3.3. Seja A uma matriz invertível. Chama-se *número de condição de A* (na norma p) o seguinte valor

$$cond_p(A) = \|A\|_p \, \|A^{-1}\|_p. \tag{3.24}$$

Como resulta da desigualdade 3.23, o número de condição da matriz A dá -nos a relação entre o erro relativo da solução de um sistema linear e o erro relativo do seu segundo membro. Assim, **se o número de condição de A for alto, daí resulta que pequenos erros relativos do segundo membro podem provocar erros muito significativos na solução, o que significa, por definição, que o sistema é mal condicionado**.

Note-se, entretanto, que o número de condição de uma matriz é sempre maior ou igual a 1, desde que consideremos normas matriciais induzidas (ver problema 3-b no fim desta secção). Por conseguinte, um sistema bem condicionado é aquele que tem o número de condição não muito maior que 1.

Também se usa a definição do número de condição através do raio espectral, sendo nesse caso dado pela expressão

$$cond_*(A) = r_\sigma(A) \, r_\sigma(A^{-1}). \tag{3.25}$$

De acordo com o teorema 3.1, podemos escrever

$$cond_*(A) \leq cond_p(A), \tag{3.26}$$

qualquer que seja a norma matricial p considerada ($p \geq 1$). Daqui resulta que, se o número de condição $cond_*(A)$ for alto, todos os números de condição da matriz são altos, pelo que o sistema é mal condicionado. No entanto, pode acontecer que o sistema seja mal condicionado, mesmo que o número de condição $cond_*(A)$ seja pequeno (ver problema 1 desta secção).

Atendendo a que os valores pró prios de A^{-1} são os inversos dos valores pró prios de A, o número de condição $cond_*(A)$ pode escrever-se sob a forma

$$cond_*(A) \;=\; \frac{\max_{\lambda_i \in \sigma(A)} |\lambda_i|}{\min_{\lambda_i \in \sigma(A)} |\lambda_i|}. \tag{3.27}$$

No caso de a matriz A ser simétrica, então, como foi observado, a sua norma euclidiana coincide com o raio espectral, pelo que podemos escrever:

$$cond_2(A) \;=\; cond_*(A). \tag{3.28}$$

Consideremos agora o caso mais geral em que o sistema linear pode estar afectado de erros, não só no segundo membro, mas também na própria matriz. Sobre este caso, é conhecido o seguinte teorema.

Teorema 3.2. Consideremos o sistema linear $A\,x = b$, onde A é uma matriz invertível. Sejam δA e δb definidos pelas igualdades (3.17) e (3.16), respectivamente, e suponhamos que

$$\|A - \tilde{A}\|_p \;\leq\; \frac{1}{\|A^{-1}\|_p}. \tag{3.29}$$

Então verifica-se a seguinte desigualdade:

$$\|\delta x\|_p \;\leq\; \frac{cond(A)}{1 \,-\, cond_p(A)\,\|\delta A\|_p}\;\left\{\|\delta A\|_p \,+\, \|\delta b\|_p\right\}. \tag{3.30}$$

Observação. Note-se que a desigualdade 3.23, obtida anteriormente, é um caso particular de 3.30, que se obtém fazendo $\|\delta A\|_p = 0$. A demonstração do teorema 3.2 encontra-se, por exemplo, em [1].

A desigualdade (3.30) confirma a nossa conclusão de que os sistemas lineares com números de condição altos são mal condicionados. O exemplo que

se segue mostra como os problemas de mau condicionamento podem surgir mesmo em sistemas de pequenas dimensões, com matrizes aparentemente "bem comportadas".

Exemplo 3.2 Consideremos o sistema linear $A\,x = b$, onde

$$A = \begin{bmatrix} 10 & 7 & 8 & 7 \\ 7 & 5 & 6 & 5 \\ 8 & 6 & 10 & 9 \\ 7 & 5 & 9 & 10 \end{bmatrix}, \quad b = \begin{bmatrix} 32 \\ 23 \\ 33 \\ 31 \end{bmatrix}. \tag{3.31}$$

Verifica-se imediatamente, por substituição, que a solução deste sistema é $x = (1, 1, 1, 1)^T$. Sabe-se que a matriz A é não singular. Além disso, ela é, evidentemente, simétrica e a sua norma, por linhas ou por colunas, é

$$\|A\|_\infty = \|A\|_1 = max(32, 23, 33, 31) = 33.$$

Entretanto, se substituirmos o vector b por um vector $\tilde{b}$, tal que

$$\tilde{b} = (32.1, 22.9, 33.1, 30.9)^T,$$

a solução do sistema passa a ser

$$\tilde{x} = (9.2, -12.6, 4.5, -1.1)^T,$$

o que é totalmente diferente da solução do sistema inicial. Por outras palavras, uma perturbação do segundo membro, tal que

$$\|\delta b\|_\infty = \frac{0.1}{33} \approx 0,3\%$$

leva-nos a uma nova solução, cuja norma é cerca de 13 vezes maior que a da solução original.

Observemos ainda o que acontece se a matriz A sofrer uma ligeira perturbação das suas componentes, sendo substituída por

$$\tilde{A} = \begin{bmatrix} 10 & 7 & 8.1 & 7.2 \\ 7.08 & 5.04 & 6 & 5 \\ 8 & 5.98 & 9.89 & 9 \\ 6.99 & 5 & 9 & 9.98 \end{bmatrix}, \tag{3.32}$$

mantendo-se o segundo membro inalterado. Neste caso, a solução do sistema passa a ser

$$\tilde{x} = (-81, 137, -34, 22)^T.$$

Neste caso, a diferença em relação à solução inicial é ainda mais acentuada. Entretanto, a norma da perturbação é relativamente pequena:

$$\|A - \tilde{A}\|_\infty \;=\; max(0.3, 0.12, 0.13, 0.03) = 0.3,$$

de onde resulta

$$\|\delta A\|_\infty \;=\; \frac{0.3}{33} \approx 0,9\%.$$

Vejamos como interpretar estes factos, com base na teoria do condicionamento que expusemos nesta secção. Para isso, precisamos de conhecer a inversa de A :

$$A^{-1} \;=\; \begin{bmatrix} 25 & -41 & 10 & -6 \\ -41 & 68 & -17 & 10 \\ 10 & -17 & 5 & -3 \\ -6 & 10 & -3 & 2 \end{bmatrix}. \tag{3.33}$$

Podemos imediatamente constatar que

$$\|A^{-1}\|_\infty \;=\; max(82, 136, 35, 21) = 136.$$

Assim, o número de condição de A, segundo a norma ∞ (que coincide com a norma 1 neste caso), tem o valor

$$cond_\infty(A) \;=\; cond_1(A) \;=\; 33 \times 136 = 4488.$$

Conhecendo o valor do número de condição, já não nos surpreende o facto de as pequenas perturbações que introduzimos no segundo membro e na matriz terem alterado completamente a solução. Com efeito, a estimativa (3.23), aplicada a este caso, diz-nos que

$$\|\delta x\| \;\leq\; 4488 \times 0,3\% \;=\; 1346\%,$$

o que explica inteiramente os resultados obtidos, no que diz respeito à perturbação do segundo membro do sistema. No caso das alterações produzidas na matriz A, não se pode aplicar a estimativa 3.30, uma vez que, para a perturbação considerada, não se verifica a condição

$$\|A - \tilde{A}\|_\infty \;\leq\; \frac{1}{\|A^{-1}\|_\infty}.$$

No entanto, dado o alto valor do número de condição, seria de esperar, de qualquer modo, que a solução do sistema ficasse muito alterada, tal como se verificou.

3.2.1 Problemas sobre Condicionamento

1. Seja A uma matriz quadrada, de dimensão n, com a forma

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

   (a) Calcule A^{-1}.

   (b) Determine os números de condição $cond_1(A)$ e $cond_\infty(A)$.

   (c) Sejam b_1 e b_2 dois vectores de $I\!R^n$ tais que

$$\|\delta b\|_\infty = \frac{\|b_1 - b_2\|_\infty}{\|b_1\|_\infty} \leq 10^{-5}.$$

   Sejam x_1 e x_2, respectivamente, as soluções dos sitemas $A\,x = b_1$ e $A\,x = b_2$. Determine um majorante de

$$\|\delta x\|_\infty = \frac{\|x_1 - x_2\|_\infty}{\|x_1\|_\infty}$$

   no caso de $n = 20$. Comente.

2. Seja A a matriz real

$$A = \begin{bmatrix} 1 & 0 & 1 \\ 1 & -1 & 0 \\ a & 0 & 3 \end{bmatrix},$$

   onde $a \in I\!R$. Suponhamos que, ao resolver o sistema $A\,x = b$, com um certo valor de a, se obteve a solução $\tilde{x} = (1, 1, 1)$. Supondo que o valor de a está afectado de um certo erro, de valor absoluto não superior a ϵ, determine um majorante de $\|\Delta x\|_\infty$, onde Δx é a diferença entre a solução obtida e a que se obteria se fosse conhecido o valor exacto de a.

3.3 Métodos Directos

Para resolver um sistema linear, podemos optar por dois caminhos diferentes:

1. ou procuramos a solução por um método de aproximações sucessivas, utilizando um *método iterativo*;

2. ou optamos por reduzir o sistema a uma forma mais simples, de modo a obter a solução exacta através de simples substituições. Nesse caso, dizemos que estamos a aplicar um *método directo*.

Neste parágrafo, falaremos de métodos directos. Quando se utiliza um método directo, sabe-se que o erro do método é nulo, visto que o método conduz teoricamente à solução exacta do problema. Isto, porém não quer dizer que a solução que se obtém, de facto, seja exacta, visto que, como sabemos, quando se efectuam cálculos numéricos são inevitáveis os erros de arredondamento.

3.3.1 Método da Eliminação de Gauss

Um dos métodos mais simples e mais conhecidos para a solução de sistemas lineares é o *método da eliminação de Gauss*. A ideia básica deste método consiste em reduzir o sistema dado, $A\,x = b$, a um sistema equivalente, $U\,x = b'$, onde U é uma matriz triangular superior. Este último sistema pode ser resolvido imediatamente, por substituição, começando pela última equação. Assim, podemos dizer que a resolução de um sistema pelo método de Gauss se divide em três etapas:

1. Redução da matriz A à forma triangular superior;
2. Transformação do segundo membro do sistema;
3. Resolução do sistema com a matriz triangular superior.

Vejamos com mais pormenor em que consiste cada uma destas etapas e avaliemos, em termo de número de operações aritméticas, o volume dos cálculos correspondentes.

1.Redução da matriz A à forma triangular superior. Suponhamos que a matriz dada tem a forma

$$A = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix}. \tag{3.34}$$

Admitindo que $a_{11} \neq 0$, esta matriz pode ser transformada, somando a cada linha (a começar pela $2^{\underline{a}}$), um múltiplo da $1^{\underline{a}}$. Assim, obtém-se uma

nova matriz $A^{(1)}$, com a forma:

$$A^{(1)} = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ 0 & a_{22}^{(1)} & \dots & a_{2n}^{(1)} \\ \dots & \dots & \dots & \dots \\ 0 & a_{n2}^{(1)} & \dots & a_{nn}^{(1)}. \end{bmatrix}. \tag{3.35}$$

As componentes de $A^{(1)}$ obtêm-se através da fórmula:

$$a_{ij}^{(1)} = a_{ij} - m_{i1}\, a_{1j}, \tag{3.36}$$

onde

$$m_{i1} = \frac{a_{i1}}{a_{11}}. \tag{3.37}$$

Continuando este processo, podemos a seguir transformar a matriz $A^{(1)}$ numa nova matriz $A^{(2)}$, que será triangular superior, até à $2^{\underline{a}}$ coluna. Generalizando, vamos efectuar uma sucessão de transformações, em que cada transformada $A^{(k)}$ tem a forma

$$A^{(k)} = \begin{bmatrix} a_{11} & a_{12} & \dots & \dots & \dots & a_{1n} \\ 0 & a_{22}^{(1)} & \dots & \dots & \dots & a_{2n}^{(1)} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & 0 & a_{kk}^{(k-1)} & \dots & a_{kn}^{(k-1)} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & 0 & a_{nk}^{(k)} & \dots & a_{nn}^{(k)} \end{bmatrix}. \tag{3.38}$$

As componentes de $A^{(k)}$ obtêm-se a partir das de $A^{(k-1)}$ através da fórmula:

$$a_{ij}^{(k)} = a_{ij}^{(k-1)} - m_{ik}\, a_{kj}^{(k-1)}, \quad i = k+1, \dots, n; \;\; j = k+1, \dots, n; \tag{3.39}$$

onde

$$m_{ik} = \frac{a_{ik}^{(k-1)}}{a_{kk}^{(k-1)}} \tag{3.40}$$

(neste caso, pressupõe-se que $a_{kk}^{(k-1)} \neq 0$). Ao fim de $n-1$ transformações obtém-se

$$A^{(n-1)} = \begin{bmatrix} a_{11} & a_{12} & \dots & \dots & \dots & a_{1n} \\ 0 & a_{22}^{(1)} & \dots & \dots & \dots & a_{2n}^{(1)} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & 0 & a_{kk}^{(k-1)} & \dots & a_{kn}^{(k-1)} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & \dots & \dots & 0 & a_{nn}^{(n-1)} \end{bmatrix}. \tag{3.41}$$

No caso de algum dos coeficientes $a_{kk}^{(k-1)}$ ser igual a 0, para se poder aplicar a eliminação de Gauss torna-se necessário alterar a ordem das linhas. Esse caso será visto em detalhe mais adiante, durante a resolução do Exemplo 3.3. Note-se que, se a matriz A for não singular, existe sempre uma permutação das suas linhas, depois da qual ela pode ser reduzida à forma (3.41), com todos os elementos da diagonal principal diferentes de 0.

2. Transformação do segundo membro. O segundo membro do sistema é sujeito a transformações análogas às que se operam sobre a matriz, de modo a garantir a equivalência do sistema resultante ao inicial. Assim, a transformação do vector b também se realiza em $n-1$ passos, sendo a primeira transformada, $b^{(1)}$, obtida segundo a fórmula:

$$b_i^{(1)} = b_i - m_{i1}\, b_1, \quad i = 2, 3, \ldots, n. \tag{3.42}$$

Analogamente, a k-ésima transformada do segundo membro é obtida pela fórmula:

$$b_i^{(k)} = b_i - m_{ik}\, b_k^{(k-1)}, \quad i = k+1, \ldots, n. \tag{3.43}$$

Os coeficientes m_{ik} são dados pelas fórmulas (3.37) e (3.40).

3. Resolução do sistema triangular superior. Depois de reduzido o sistema inicial à forma triangular superior, com a matriz (3.41), a sua solução obtém-se facilmente, mediante um processo de substituições sucessivas:

$$\begin{array}{rcl} x_n & = & \dfrac{b_n^{(n-1)}}{a_{nn}^{(n-1)}}; \\ x_{n-1} & = & \dfrac{b_{n-1}^{(n-2)} - a_{n-1,n}^{(n-2)}\, x_n}{a_{n-1,n-1}^{(n-2)}}; \\ \ldots & \ldots & \ldots \\ x_1 & = & \dfrac{b_1 - \sum_{i=2}^{n} a_{1,i}\, x_i}{a_{1,1}}. \end{array} \tag{3.44}$$

Vejamos agora qual o número de operações aritméticas necessárias para efectuar cada uma das etapas que acabámos de descrever.

1. Redução da matriz à forma triangular superior. O número de operações necessárias para a transformação da matriz A está relacionado com o número de vezes que são aplicadas as fórmulas (3.36) e (3.39).

No 1º passo, a fórmula (3.36) é aplicada $(n-1)^2$ vezes. Isto implica que se realizem $(n-1)^2$ multiplicações e outras tantas adições (ou subtracções). Entretanto, para calcular os quocientes da fórmula (3.37), efectuam-se $n-1$ divisões. Todas estas operações continuam a efectuar-se nos passos seguintes da transformação, mas em menor número, de acordo com a quantidade de componentes que são alteradas em cada passo. Em geral, no k-ésimo passo efectuam-se $(n-k)^2$ multiplicações e outras tantas adições (ou subtracções), assim como $n-k$ divisões. Assim, o número total de multiplicações efectuadas durante a transformação da matriz, igual ao número de adições (ou subtracções), é

$$M(n) \;=\; AS(n) \;=\; \sum_{k=1}^{n-1}(n-k)^2 \;=\; \frac{n(n-1)(2n-1)}{6}, \tag{3.45}$$

Quanto às divisões, o número total é

$$D(n) \;=\; \sum_{k=1}^{n-1}(n-k) \;=\; \frac{n(n-1)}{2}. \tag{3.46}$$

Assim, o número total de operações efectuadas durante a transformação da matriz é, em termos assimptóticos, para valores altos de n,

$$TO(n) \;= M(n) + AS(n) + D(n) \approx \frac{2n^3}{3} + O(n^2). \tag{3.47}$$

2. Transformação do segundo membro. Quando transformamos o vector b, aplicamos a fórmula (3.43) às suas componentes. Esta fórmula é aplicada $n-k$ vezes durante o k-ésimo passo da transformação, o que implica $n-k$ multiplicações e outras tantas adições (ou subtracções). Assim, o número total de multiplicações efectuadas durante a transformação do segundo membro, igual ao número de adições (ou subtracções), é

$$M(n) \;=\; AS(n) \;=\; \sum_{k=1}^{n-1}(n-k) \;=\; \frac{n(n-1)}{2}, \tag{3.48}$$

Por conseguinte, o número total de operações exigidas por esta etapa dos cálculos é, em termos assimptóticos,

$$TO(n) \;=\; M(n) \;+\; AS(n) \;\approx\; n^2. \tag{3.49}$$

3. Resolução do sistema triangular. Para resolver o sistema triangular, efectuamos a série de substituições (3.44). Como resulta destas fórmulas, o número total de multiplicações para resolver o sistema é $n(n-1)/2$, igual ao número total de adições (ou subtracções). Quanto ao número de divisões, é igual a n.

Assim, o número total de operações efectuadas para resolver o sistema triangular é, em termos assimptóticos,

$$TO(n) \;=\; M(n) \;+\; AS(n) \;+\; D(n) \;\approx\; n^2. \tag{3.50}$$

Exemplo 3.3. Consideremos o sistema linear $A\,x = b$, onde

$$A = \begin{bmatrix} 2 & 1 & 3 \\ -2 & -1 & 1 \\ 2 & 4 & 2 \end{bmatrix}, b = \begin{bmatrix} 5 \\ -1 \\ 4 \end{bmatrix}. \tag{3.51}$$

Pretende-se reslolver este sistema pelo método da eliminação de Gauss. Comecemos por reduzir A à forma triangular superior. O primeiro passo para isso, como vimos, consiste em transformar a matriz A na matriz $A^{(1)}$. Neste caso, usando as fórmulas (3.36) e (3.37), obtém-se:

$$\begin{aligned} a_{22}^{(1)} &= a_{22} \;-\; m_{21}\, a_{12} = 0, \\ a_{23}^{(1)} &= a_{23} \;-\; m_{21}\, a_{13} = 4, \end{aligned} \tag{3.52}$$

onde

$$m_{21} \;=\; \frac{a_{21}}{a_{11}} = -1. \tag{3.53}$$

Verifica-se que o novo elemento da diagonal principal, $a_{22}^{(1)}$, é nulo. Como sabemos, neste caso não é possível aplicar o método da eliminação de Gauss directamente à matriz A. Vamos então proceder a uma troca de linhas; mais precisamente, troquemos a segunda linha com a terceira. Obtém-se assim o novo sistema $A'\,x = b'$, onde

$$A' = \begin{bmatrix} 2 & 1 & 3 \\ 2 & 4 & 2 \\ -2 & -1 & 1 \end{bmatrix}, b' = \begin{bmatrix} 5 \\ 4 \\ -1 \end{bmatrix}. \tag{3.54}$$

Aplicando o método da eliminação de Gauss a este sistema, usemos de novo as fórmulas (3.36) e (3.37), que nos dão

$$\begin{aligned} a_{22}^{(1)'} &= a'_{22} \;-\; m'_{21}\, a_{12} = 4 - 1 = 3 \\ a_{23}^{(1)'} &= a'_{23} \;-\; m'_{21}\, a_{13} = 2 - 3 = -1, \\ a_{32}^{(1)'} &= a'_{32} \;-\; m'_{31}\, a_{12} = -1 + 1 = 0, \\ a_{33}^{(1)'} &= a'_{33} \;-\; m'_{31}\, a_{13} = 1 + 3 = 4, \end{aligned} \tag{3.55}$$

onde

$$m'_{21} = \frac{a'_{21}}{a_{11}} = 1, \tag{3.56}$$

$$m'_{31} = \frac{a'_{31}}{a_{11}} = -1. \tag{3.57}$$

Obtemos assim a matriz transformada da forma

$$A' = \begin{bmatrix} 2 & 1 & 3 \\ 0 & 3 & -1 \\ 0 & 0 & 4 \end{bmatrix}. \tag{3.58}$$

Como esta matriz é triangular superior, não é necessário transformá-la mais.

A segunda etapa da aplicação do método da eliminação de Gauss consiste em transformar o segundo membro do sistema, isto é, o vector b'. Para isso, utilizamos a fórmula (3.42), que neste caso nos dá

$$\begin{aligned} b_2^{(1)'} &= b'_2 - m'_{21}\, b'_1 = 4 - 5 = -1; \\ b_3^{(1)'} &= b'_3 - m'_{31}\, b'_1 = -1 + 5 = 4. \end{aligned} \tag{3.59}$$

Obtemos assim o vector transformado $b^{(1)'} = (5, -1, 4)^T$.

Por último, resta-nos resolver o sistema triangular superior $A^{(1)'}\, x = b^{(1)'}$. Para isso, usamos a substituição ascendente, isto é, começamos por determinar x_3 a partir da última equação, para depois determinar x_2 da segunda, e x_1 da primeira. Usando as fórmulas (3.44), obtém-se

$$\begin{aligned} x_3 &= \frac{b_3^{(1)'}}{a_{33}^{(1)}} = 1; \\ x_2 &= \frac{b_2^{(2)} - a_{23}^{(1)}\, x_3}{a_{22}^{(1)}} = \tfrac{-1+1}{2} = 0; \\ x_1 &= \frac{b_1 - a_{13}\, x_3 - -a_{12}\, x_2}{a_{1,1}} = \tfrac{5-3}{2} = 1. \end{aligned} \tag{3.60}$$

Assim, obtemos a solução do sistema: $x = (1, 0, 1)^T$.

3.3.2 Influência dos erros de arredondamento

Até agora, ao falarmos do método de Gauss, não entrámos em consideração com com os erros cometidos durante os cálculos. Na secção 3.2 já vimos que pequenos erros nos dados iniciais do sistema podem afectar muito a solução, se a matriz for mal condicionada. Além dos erros dos dados iniciais,

que são inevitáveis, há que ter em conta também o erro computacional, resultante dos arredondamentos efectuados durante os cálculos. Um dos inconvenientes do método de Gauss, assim como de outros métodos directos, de que falaremos adiante, consiste em que esses erros têm frequentemente tendência para se propagar durante os cálculos, de tal modo que podem adquirir um peso muito grande na solução, mesmo que o sistema seja bem condicionado. No entanto, o efeito destes erros pode ser muito atenuado, se durante os cálculos for usada uma estratégia adequada, a chamada *estratégia de pivot*. Vamos ver em que consiste esta estratégia.

Ao falarmos da transformação da matriz A, vimos que, para que ela se pudesse efectuar, é necessário que todos os elementos da diagonal principal da matriz triangular superior U sejam diferentes de 0. Estes elementos são representados por $a_{kk}^{(k-1)}$ e são chamados geralmente os *pivots*, dada a sua importância para a aplicação do método de Gauss [1]. Vimos também que, no caso de um dos pivots ser nulo, se podia mesmo assim aplicar o método, desde que se efectuasse uma troca de linhas na matriz. Se o pivot não for nulo, mas próximo de 0, o método continua a ser teoricamente aplicável, mesmo sem trocas de linhas. Só que, ao ficarmos com um denominador muito pequeno no segundo membro de (3.40), cria-se uma situação em que os erros de arredondamento podem propagar-se de uma forma desastrosa. A estratégia de pivot tem por objectivo evitar que isto aconteça. Assim, em cada passo da transformação da matriz, verifica-se o pivot e, caso se considere conveniente, efectua-se uma troca de linhas que substitui o pivot por outra componente maior. A estratégia de pivot tem diversas variantes, das quais falaremos aqui de duas: a *pesquisa parcial* e a *pesquisa total* de pivot.

Pesquisa parcial de pivot. Neste caso, em cada passo da transformação da matriz, é inspeccionada a coluna k da matriz $A^{(k-1)}$, mais precisamente, as componentes dessa coluna que se situam da diagonal principal para baixo. Seja

$$c_k = \max_{k \leq i \leq n} \left| a_{ik}^{(k-1)} \right| . \tag{3.61}$$

Se o máximo no segundo membro de (3.61) for atingido para $i = k$, isso significa que o actual pivot é, em módulo, a maior componente daquela coluna. Nesse caso, continuam-se os cálculos normalmente. Se o máximo for atingido para um certo $i \neq k$, então troca-se de lugar a linha k com a

[1] em francês, *pivot* tem o significado de *base, apoio*

linha i e só depois se prosseguem os cálculos. Evidentemente, ao fazer esssa troca, também se efectua uma permutação correspondente nas componentes do vector b.

Pesquisa total de pivot. De acordo com esta estratégia, mais radical, é inspeccionada não só a coluna k da matriz $A^{(k-1)}$, mas também todas as colunas subsequentes. Seja

$$c_k = \max_{k \leq i,j \leq n} \left| a_{ij}^{(k-1)} \right| . \tag{3.62}$$

Sejam i' e j', respectivamente, os valores dos índices i e j para os quais é atingido o máximo no segundo membro de (3.62). Se i' não coincidir com k, a linha i' troca de lugar com a linha k. Se, além disso, j' não coincidir com k, então a coluna j' também vai trocar de lugar com a coluna k (o que corresponde a uma troca de lugar entre as incógnitas $x_{j'}$ e x_k).

Comparando as duas variantes acima mencionadas da pesquisa de pivot, vê-se que a pesquisa total é bastante mais dispendiosa do que a parcial, uma vez que exige um número de comparações muito maior. Entretanto, a prática do cálculo numérico tem demonstrado que, na grande maioria dos casos reais, a pesquisa parcial conduz a resultados praticamente tão bons como os da total. Isto explica que, hoje em dia, a pesquisa parcial seja mais frequentemente escolhida quando se elaboram algoritmos baseados no método de Gauss.

O exemplo que se segue mostra até que ponto os erros de arredondamento podem influir na solução de um sistema linear, quando é aplicado o método da eliminação de Gauss. Vamos ver também como a pesquisa parcial de pivot pode contribuir para melhorar esta situação.

Exemplo 3.4. Consideremos o sistema linear $A\,x = b$, onde

$$A = \begin{bmatrix} 10^{-6} & 0 & 1 \\ 1 & 10^{-6} & 2 \\ 1 & 2 & -1 \end{bmatrix}, \quad b = \begin{bmatrix} 1 \\ 3 \\ 2 \end{bmatrix}. \tag{3.63}$$

Se procurarmos resolvê -lo, utilizando o método da eliminação de Gauss,

chegamos ao sistema equivalente $U\,x\,=\,b'$,onde

$$U\,=\,\begin{bmatrix} 10^{-6} & 0 & 1 \\ 0 & 10^{-6} & 2-10^{6} \\ 0 & 0 & 2\times 10^{12}-5\times 10^{6}-1 \end{bmatrix},\;\; b'\,=\,\begin{bmatrix} 1 \\ 3-10^{6} \\ 2\times 10^{12}-7\times 10^{6}+2 \end{bmatrix}. \tag{3.64}$$

Na realidade, a matriz U e o vector b' que vamos obter não são exactamente os da fórmula (3.64), devido aos erros de arredondamento. Suponhamos que os cálculos são efectuados num computador em que os números são representados no sistema decimal, com *seis dígitos* na mantissa. Então , em vez de U e b', teremos a seguinte matriz e o seguinte vector:

$$\tilde{U}\,=\,\begin{bmatrix} 1.00000\times 10^{-6} & 0 & 1.00000 \\ 0 & 1.00000\times 10^{-6} & -0.999998\times 10^{6} \\ 0 & 0 & 1.99999\times 10^{12} \end{bmatrix},\;\; \tilde{b}\,=\,\begin{bmatrix} 1 \\ -0.999997\times 10^{6} \\ 1.99999\times 10^{12} \end{bmatrix}. \tag{3.65}$$

Assim, ao resolvermos o sistema (3.65) por substituição ascendente, obtemos sucessivamente:

$$\begin{aligned} \tilde{x_3} &= \frac{1.99999\times 10^{12}}{1.99999\times 10^{12}} &&= 1.00000; \\ \tilde{x_2} &= \frac{-0.999997\times 10^{6}+0.999998\times 10^{6}\;\tilde{x_3}}{1.00000\times 10^{-6}} &&= 1.00000\times 10^{6}; \\ *[0.5cm]\tilde{x_1} &= \frac{1.00000-1.00000\;\tilde{x_3}}{1.00000\times 10^{-6}} &&= 0. \end{aligned} \tag{3.66}$$

Substituindo os valores assim obtidos no sistema dado, verifica-se que eles estão longe de o satisfazer, o que indica que este resultado apresenta um erro relativo muito grande. Este erro, no entanto, não tem a ver com o condicionamento do sistema, visto que o número de condição de A tem o valor

$$cond_{\infty}(A)\,=\,\|A\|_{\infty}\|A^{-1}\|_{\infty}\approx 3\times 4=12;$$

logo, o sistema não se pode considerar mal condicionado. Nestas condições, temos razões para pensar que o mau resultado obtido resulta da instabilidade numérica do método, a qual, como vimos, pode ser contrariada através da pesquisa de pivot. Vejamos então que resultado obteríamos, se aplicássemos a pesquisa parcial de pivot. Neste caso, começaríamos por trocar a primeira linha de A com a segunda, visto que $a_{21}\,>\,a_{11}$. Depois da primeira trans-

formação, obteríamos a matriz $A^{(1)}$, com a seguinte forma:

$$A^{(1)} = \begin{bmatrix} 1 & 10^{-6} & 2 \\ 0 & -10^{-12} & 1-2\times 10^{-6} \\ 0 & 2-10^{-6} & -3 \end{bmatrix}. \tag{3.67}$$

Neste caso, a pesquisa de pivot leva-nos trocar a segunda linha com a terceira, visto que $a_{32} > a_{22}$. Depois de efectuar esta troca, realiza-se a segunda transformação da matriz, que nos leva ao sistema $A^{(2)}x = b^{(2)}$. Se os cálculos forem realizados com a precisão acima referida, obtemos a seguinte matriz e o seguinte vector:

$$A^{(2)} = \begin{bmatrix} 1.00000 & 1.00000\times 10^{-6} & 2.00000 \\ 0 & 2.00000 & -3.00000 \\ 0 & 0 & 9.99998\times 10^{-1} \end{bmatrix}, \quad b^{(2)} = \begin{bmatrix} 3.00000 \\ -1.00000 \\ 9.99997\times 10^{-1} \end{bmatrix}. \tag{3.68}$$

Resolvendo o sistema (3.68), obtém-se a seguinte solução:

$$\begin{aligned} x_3 &= \frac{9.99997\times 10^{-1}}{1.00000} &&= 9.99999\times 10^{-1}; \\ *[0.5cm]x_2 &= \frac{-1.00000+3.00000\,x_3}{2.00000} &&= 1.00000; \\ x_1 &= 3.00000 - 2.00000\,x_3 - 1.00000\times 10^{-6}\,x_2 &&= 9.99999\times 10^{-1}. \end{aligned} \tag{3.69}$$

Esta solução é bastante diferente da que obtivemos, quando não foi utilizada a pesquisa de pivot. Se substituirmos estes valores no sistema (3.63), veremos que a nova solução está correcta, dentro dos limites da precisão utilizada. Este exemplo mostra-nos como a pesquisa de pivot pode desempenhar um papel essencial na eliminação dos problemas da instabilidade numérica quando se resolvem sistemas lineares pelo método da eliminação de Gauss.

3.3.3 Métodos de Factorização

Neste parágrafo, vamos tratar de métodos directos que se baseiam na factorização da matriz do sistema linear.

Definição 3.4.Chama-se *factorização LU* de uma matriz A à sua representação sob a forma do produto de de duas matrizes:

$$A = LU,$$

onde L é triangular inferior e U é triangular superior.

Se for conhecida a factorização LU de uma matriz A, então o sistema linear $A\,x = b$ reduz-se a dois sistemas lineares com matrizes triangulares:

$$\begin{aligned} L\,g &= b, \\ U\,x &= g, \end{aligned}$$

onde g é um vector auxiliar. Além da solução de sistemas lineares, a factorização LU tem outras aplicações , como por exemplo o cálculo de determinantes. Com efeito, o determinante de A é igual ao produto dos determinantes de L e U, os quais se calculam imediatamente, dado estas matrizes serem triangulares:

$$\begin{aligned} det\,L &= l_{11}\,l_{22}\ldots l_{nn}, \\ det\,U &= u_{11}\,u_{22}\ldots u_{nn}. \end{aligned}$$

Note-se que, para calcularmos o determinante a partir da definição , teríamos de somar, para uma matriz de ordem n, $n!$ parcelas, cada uma das quais é um produto de n componentes da matriz A. Tal cálculo significaria, só para uma matriz de ordem 10, efectuar mais de 30 milhões de multiplicações ! Compreende-se portanto que tal maneira de calcular um determinante não é aplicável na prática. Em compensação, se utilizarmos a factorização, o mesmo determinante pode ser calculado apenas com algumas centenas de operações aritméticas.

Uma vantagem dos métodos de factorização consiste em que, uma vez factorizada uma matriz, se pretendermos resolver vários sistemas diferentes com essa matriz, basta resolver os sistemas triangulares correspondentes (as matrizes L e U só precisam de ser determinadas uma vez). Isto é muito importante, dado que, como vamos ver, nos métodos de factorização a determinação das matrizes L e U é precisamente a etapa mais dispendiosa, em termos de número de operações .

O problema da factorização LU de uma matriz $A \in L^n$, não singular, admite sempre múltiplas soluções. Com efeito, podemos abordar o problema como um sistema de n equações :

$$a_{ij} = \sum_{k=1}^{n} l_{ik}\,u_{kj}; \quad i = 1,\ldots,n;\ j = 1,\ldots,n; \tag{3.70}$$

onde l_{ik} e u_{kj} são as incógnitas e representam as componentes das matrizes L e U, respectivamente. Uma vez que cada uma destas matrizes tem $\frac{n(n+1)}{2}$ componentes não nulas, o número total de incógnitas do sistema é $n(n+1)$, superior, portanto, ao número de equações. Isto explica que o sistema seja

indeterminado, isto é, que ele admita muitas soluções diferentes. A cada uma dessas soluções corresponde uma certa factorização , que se caracteriza por um conjunto de condições suplementares. Vamos analisar alguns tipos particulares de factorização, com maior interesse prático.

3.3.4 Factorização de Doolittle

Este tipo de factorização caracteriza-se pelo conjunto de condições :

$$l_{ii} \;=\; 1; \;\; i = 1, \ldots, n. \tag{3.71}$$

Vamos mostrar como, a partir destas condiçoes, se podem deduzir as fórmulas para as componentes das matrizes L e U, que ficam assim completamente determinadas.

Seja a_{ij} uma componente da matriz A, com $i \leq j$. Então , atendendo à forma das matrizes L e U, bem como à condição 3.71, podemos escrever:

$$a_{ij} \;=\; \sum_{k=1}^{i} l_{ik}\, u_{kj} \;=\; \sum_{k=1}^{i-1} l_{ik}\, u_{kj} \;+\; u_{ij}; \;\; i = 1, \ldots, n; \; j = i, \ldots, n. \tag{3.72}$$

Da fórmula 3.72 resulta imediatamente que

$$u_{ij} \;=\; a_{ij} \;-\; \sum_{k=1}^{i-1} l_{ik}\, u_{kj}. \tag{3.73}$$

A fim de deduzir uma fórmula análoga para a matriz L, consideremos o caso de uma componente a_{ij}, com $i > j$. Neste caso, em vez da fórmula 3.72, temos

$$a_{ij} \;=\; \sum_{k=1}^{j} l_{ik}\, u_{kj} \;=\; \sum_{k=1}^{j-1} l_{ik}\, u_{kj} \;+\; l_{ij}\, u_{jj}, \;\; i = 1, \ldots, n; \; j = i, \ldots, n. \tag{3.74}$$

Daqui resulta

$$l_{ij} \;=\; \frac{a_{ij} \;-\; \sum_{k=1}^{j-1} l_{ik}\, u_{kj}}{u_{jj}}. \tag{3.75}$$

Utilizando as fórmulas 3.73 e 3.75, podem calcular-se todas as componentes das matrizes L e U. Para isso, basta que todas as componentes da diagonal principal de U sejam diferentes de 0. Se, durante processo de cálculo, se obtiver alguma dessas componentes igual a 0, então , tal como acontece no método da eliminação de Gauss, deve-se proceder a alterações na matriz U.

Neste caso, o mais prático é alterar a ordem das colunas de U, mantendo a matriz L sem alteração . Isto corresponde a trocar a ordem das colunas de A, ou seja, a trocar a ordem das incógnitas. Neste caso, ao calcular o determinante de A com base na factorização , deve-se entrar em conta com as permutações efectuadas. Assim,

$$det\, A \;=\; (-1)^{Nt}\, det\, L\; det\, U, \tag{3.76}$$

onde Nt é o número de trocas de colunas efectuadas. A troca de colunas de L também pode ser aplicada para atenuar os problemas de instabilidade numérica que podem ocorrer durante a factorização . Para isso, usa-se a mesma estratégia da pesquisa parcial de pivot que acima descrevemos, para o método de Gauss.

É interessante notar que o método da eliminação de Gauss é idêntico ao método de Doolittle, podendo, neste sentido, ser considerado também um método de factorização . Para verificar isso, recordemos que no método da eliminação de Gauss se obtém uma matriz triangular superior U, dada pela fórmula 3.41. Além disso, durante o cálculo da matriz U são utilizados os coeficientes m_{ik}, $k = 1, \ldots, n;\; i = k+1, \ldots, n$, definidos pela fórmula 3.40. Se dispusermos estes coeficientes numa matriz de ordem n, preenchermos a sua diagonal principal com 1 e as restantes posições com 0, obtemos a seguinte matriz triangular inferior:

$$L \;=\; \begin{bmatrix} 1 & 0 & \ldots & 0 \\ m_{21} & 1 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ \ldots & m_{n-1,n-2} & 1 & 0 \\ \ldots & m_{n,n-2} & m_{n,n-1} & 1 \end{bmatrix}. \tag{3.77}$$

Teorema 3.3. As matrizes L e U, dadas, respectivamente, pelas fórmulas 3.77 e 3.41,formam uma factorização LU da matriz A, idêntica à factorização de Doolittle.

Demonstração . Vamos demonstrar as igualdades

$$a_{ij}^{(i-1)} \;=\; u_{ij},\;\; i = 1, \ldots, n;\; j = i, \ldots, n; \tag{3.78}$$

$$m_{ij} \;=\; l_{ij},\; j = 1, \ldots, n;\; i = j, \ldots, n. \tag{3.79}$$

Para isso, basta comparar as fórmulas do método de Gauss com as da factorização de Doolittle. Em primeiro lugar, sabemos que

$$a_{1j} = u_{1j}, \;\; j = 1, \ldots, n; \;\; m_{i1} = \frac{a_{i1}}{a_{11}}, \; i = 2, \ldots, n. \tag{3.80}$$

Vamos agora supor, por indução , que a igualdade 3.78 se verifica para as linhas da matriz U, com índice $k = 1, \ldots, i-1$, e que a igualdade 3.79 se verifica para todas colunas de L, com índice $k = 1, \ldots, j-1$. Verifiquemos que, nesse caso, as mesmas identidades se mantêm válidas para a i-ésima linha de U e para a j-ésima coluna de L. De facto, de acordo com a fórmula 3.39 do método de Gauss, temos

$$a_{ij}^{(k)} = a_{ij}^{(k-1)} - m_{ik}\, a_{kj}^{(k-1)}, \;\; (k = 1, \ldots, n-1), \tag{3.81}$$

onde se subentende que $a_{ij}^{(0)} = a_{ij}$, $i = 1, \ldots, n$; $j = 1, \ldots, n$. Aplicando a fórmula 3.81, sucessivamente, com $k = 1, \ldots, i-1$ obtém-se:

$$\begin{array}{ccccccc}
a_{ij}^{(1)} & = & a_{ij} - m_{i1}\, a_{1j}; & & & & \\
a_{ij}^{(2)} & = & a_{ij}^{(1)} - m_{i2}\, a_{2j}^{(1)} & = & a_{ij} - m_{i1}\, a_{1j} - m_{i2}\, a_{2j}^{(1)}; & & \\
\ldots & \ldots & \ldots & \ldots & \ldots & & \\
a_{ij}^{(i-1)} & = & a_{ij}^{(i-2)} - m_{i,i-1}\, a_{i-1,j}^{(i-2)} & = & a_{ij} - \sum_{k=1}^{i-1} m_{ik}\, a_{kj}^{(k-1)}. & &
\end{array} \tag{3.82}$$

Se, de acordo com a hipótese da indução , substituirmos os coeficientes $m_{i,k}$ e $a_{k,j}^{(k-1)}$, no segundo membro de 3.82, por l_{ik} e u_{kj}, obtemos a fórmula 3.73, de onde se conclui que $a_{ij}^{(i-1)} = u_{ij}$, com $j = i, \ldots, n$.

Considerando agora a j-ésima coluna de L, as suas componentes, de acordo com 3.40, têm a forma

$$m_{ij} = \frac{a_{ij}^{(j-1)}}{a_{jj}^{(j-1)}}; \; i = j, \ldots, n. \tag{3.83}$$

Do mesmo modo como deduzimos a fórmula 3.82, podemos mostrar que

$$a_{ij}^{(j-1)} = a_{ij} - \sum_{k=1}^{j-1} m_{ik}\, a_{kj}^{(k-1)}. \tag{3.84}$$

Se, no segundo membro da fórmula 3.83, substituirmos o numerador de acordo com 3.84, obtemos

$$m_{ij} = \frac{a_{ij} - \sum_{k=1}^{j-1} m_{ik}\, a_{kj}^{(k-1)}}{a_{jj}^{(j-1)}}; \; i = j, \ldots, n. \tag{3.85}$$

Mas, de acordo com a hipótese da indução , podemos substituir, no segundo membro de 3.85, $a_{kj}^{(k-1)}$ por u_{kj},$k = 1, \ldots, j$ e m_{ik} por l_{ik},$k = 1 \ldots, i$. Então o segundo membro de 3.85 fica igual ao segundo membro de 3.75, de onde se conclui que $m_{ij} = l_{ij}$, para todas as componentes da j-ésima coluna da matriz L. Fica assim provada, por indução , a afirmação do teorema. $\square$

Do teorema que acabamos de demonstrar resulta que os métodos de Gauss e de Doolittle são idênticos, no sentido em que, dado um certo sistema linear, para o resolver segundo cada um desses métodos, efectuam-se exactamente as mesmas operações aritméticas. Em particular, as três etapas que distinguimos no método de Gauss coincidem com as etapas do método de Doolittle (ou de qualquer outro método de factorização):

1. factorização da matriz A (redução da matriz à forma triangular);
2. resolução do sistema $L\,g = b$ (transformação do segundo membro);
3. resolução do sistema $U\,x = g$ (resolução do sistema triangular superior).

Então , de acordo com o que dissemos em relação ao método de Gauss, podemos concluir que a etapa mais dispendiosa dos cálculos, quando se aplica o método de Doolittle, é a primeira, exigindo cerca de $2n^3/3$ operações aritméticas. As outras duas etapas exigem cerca de n^2 operações cada uma. Estes números também se aplicam à factorização de Crout, de que falaremos a seguir.

3.3.5 Factorização de Crout

Outro tipo, bastante comum, de factorização (a factorização de Crout) baseia-se nas condições

$$u_{ii} = 1, \quad i = 1, \ldots, n. \tag{3.86}$$

As fórmulas para as componentes das matrizes L e U da factorização de Crout deduzem-se da mesma maneira que no caso da factorização de Doolittle. Assim no caso de $i \geq j$, é válida a igualdade

$$a_{ij} = \sum_{k=1}^{j} l_{ik}\, u_{kj} = \sum_{k=1}^{j-1} l_{ik}\, u_{kj} + l_{ij}, \quad j = 1, \ldots, n; \; i = j, \ldots, n. \tag{3.87}$$

Daqui obtém-se imediatamente

$$l_{ij} = a_{ij} - \sum_{k=1}^{j-1} l_{ik}\, u_{kj}. \tag{3.88}$$

No que diz respeito à matriz L, partimos da igualdade

$$a_{ij} = \sum_{k=1}^{i} l_{ik}\, u_{kj} = \sum_{k=1}^{i-1} l_{ik}\, u_{kj} + l_{ii}\, u_{ij}; \;\; i = 1, \ldots, n; \; j = 1, \ldots, i. \tag{3.89}$$

Da igualdade 3.89 resulta

$$u_{ij} = \frac{a_{ij} - \sum_{k=1}^{i-1} l_{ik}\, u_{kj}}{l_{ii}}. \tag{3.90}$$

As fórmulas 3.88 e 3.90, aplicadas alternadamente (começando com a primeira coluna de L e acabando com a última linha de U) permitem-nos determinar completamente as matrizes L e U da factorização de Crout, desde que se verifique $l_{ii} \neq 0$, $i = 1, \ldots, n$. Se, durante o processo de factorização , se verificar que $l_{ii} = 0$, para um certo i, procede-se a uma troca de linhas na matriz L, mantendo U sem alteração . Esta troca é acompanhada por uma permutação análoga das linhas da matriz A e das entradas do segundo membro do sistema. Tal como no caso da factorização de Doolittle, estas permutações implicam uma troca de sinal no determinante, de acordo com a fórmula 3.76.

Também no caso da factorização de Crout é conveniente aplicar a pesquisa parcial de pivot, conduzindo a trocas de linhas quando os elementos diagonais l_{ii} forem pequenos em módulo.

Exemplo 3.5. Consideremos o sistema $A\,x = b$, onde

$$A = \begin{bmatrix} 2 & 1 & 3 \\ -2 & -1 & 1 \\ 2 & 4 & 2 \end{bmatrix}, \;\; b = \begin{bmatrix} 5 \\ -1 \\ 4 \end{bmatrix}. \tag{3.91}$$

Comecemos por factorizar A segundo o método de Doolittle. Como resulta da fórmula 3.73, neste caso a primeira linha de U é igual à primeira linha de A, ou seja,

$$u_{11} = 2, \;\; u_{12} = 1, \;\; u_{13} = 3. \tag{3.92}$$

Calculando os elementos da primeira coluna de L, de acordo com a fórmula 3.75, obtemos

$$l_{11} = 1, \;\; l_{21} = \frac{a_{21}}{u_{11}} = -1, \;\; l_{31} = \frac{a_{31}}{u_{11}} = 1. \tag{3.93}$$

Passemos ao cálculo da segunda linha de U. Temos

$$\begin{array}{rclcl} u_{22} & = & a_{22} - l_{21}\, u_{12} & = & 0; \\ u_{23} & = & a_{23} - l_{21}\, u_{13} & = & 4. \end{array} \tag{3.94}$$

Como sabemos, sendo $u_{22} = 0$, não é possível prosseguir os cálculos sem alterar a matriz A. Assim, e uma vez que $u_{23} \neq 0$, vamos trocar de lugar a segunda com a terceira coluna de U, fazendo simultaneamente a mesma troca em A. Sejam U' e A', respectivamente, as matrizes resultantes destas permutações. Então podemos escrever

$$\begin{array}{rcl} u'_{22} & = & u_{23}, \\ u'_{23} & = & u_{22}. \end{array} \tag{3.95}$$

Continuando o processo de factorização com as matrizes U' e A', obtém-se

$$\begin{array}{rcccl} l_{32} & = & \dfrac{a'_{32} - l_{31}\, u'_{12}}{u'_{22}} & = & \dfrac{a_{33} - l_{31}\, u_{13}}{u_{23}} = -\dfrac{1}{4}; \\ u'_{33} & = & a'_{33} - l_{31}\, u'_{13} - l_{32}\, u'_{23} & = & a_{32} - l_{31}\, u_{12} - l_{32}\, u_{22} = 3. \end{array} \tag{3.96}$$

Recapitulando, obtivemos a seguinte factorização de A:

$$L = \begin{bmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 1 & -\frac{1}{4} & 1 \end{bmatrix}, \;\; U' = \begin{bmatrix} 2 & 3 & 1 \\ 0 & 4 & 0 \\ 0 & 0 & 3 \end{bmatrix}. \tag{3.97}$$

Para obter a solução do sistema dado, comecemos por resolver o sistema com a matriz triangular inferior $L\,g = b$, de acordo com o método habitual:

$$\begin{array}{rclcrcl} g_1 & = & b_1 & = & 5; & & \\ -g_1 + g_2 & = & b_2 & \Leftrightarrow & g_2 & = & 4; \\ g_1 - g_2/4 + g_3 & = & b_3 & \Leftrightarrow & g_3 & = & 0. \end{array} \tag{3.98}$$

Ao resolver depois o sistema $U'\,x = g$, temos de ter em conta que a segunda coluna de U trocou de lugar com a terceira. Isto equivale a uma troca de lugares entre x_2 e x_3. Assim, temos

$$\begin{array}{rclcrcl} 3x_2 & = & g_3 & \Leftrightarrow & x_2 & = & 0; \\ 4x_3 & = & g_2 & \Leftrightarrow & x_3 & = & 1. \\ 2x_1 + 3x_3 + x_2 & = & g_1 & \Leftrightarrow & x_1 & = & 1. \end{array} \tag{3.99}$$

Se, em vez do método de Doolittle, quisermos aplicar a factorização de Crout, teremos de basear os cálculos nas fórmulas 3.88 e 3.90. Nesse caso, a primeira coluna de L fica igual à primeira coluna de A. Para a primeira linha de U, obtém-se

$$u_{11} = 1; \;\; u_{12} = \frac{a_{12}}{l_{11}} = \frac{1}{2}; \;\; u_{13} = \frac{a_{13}}{l_{11}} = \frac{3}{2}. \tag{3.100}$$

Na segunda coluna de L, obtemos:

$$\begin{array}{rclcl} l_{22} & = & a_{22} - l_{21}\, u_{12} & = & 0; \\ l_{32} & = & a_{32} - l_{31}\, u_{12} & = & 3. \end{array} \tag{3.101}$$

Uma vez que $l_{22} = 0$, torna-se necessário trocar a segunda com a terceira linha de L (e, consequentemente, de A). Feita esta permutação , obtemos

$$\begin{array}{rclcl} l'_{22} & = & l_{32} & = & 3; \\ l'_{32} & = & l_{22} & = & 0. \end{array} \tag{3.102}$$

Resta calcular as componentes da segunda linha de U e terceira coluna de L :

$$\begin{array}{rclcl} u_{23} & = & \dfrac{a'_{23} - l'_{21} u_{13}}{l'_{22}} & = & -\dfrac{1}{3}; \\ l'_{33} & = & a'_{33} - l'_{31} u_{13} - l'_{32} u_{23} & = & 4. \end{array} \tag{3.103}$$

Consequentemente, a factorização de Crout da matriz dada tem a forma:

$$L' = \begin{bmatrix} 2 & 0 & 0 \\ 2 & 3 & 0 \\ -2 & 0 & 4 \end{bmatrix}, \;\; U = \begin{bmatrix} 1 & \frac{1}{2} & \frac{3}{2} \\ 0 & 1 & -\frac{1}{3} \\ 0 & 0 & 1 \end{bmatrix}. \tag{3.104}$$

A partir de qualquer uma das factorizações de A obtidas, utilizando a fórmula 3.76, calcula-se facilmente o determinante de A

$$det\, A = det\, L' \, (-1)^1 = det\, U' \, (-1)^1 = -24.$$

Se quisermos resolver o sistema dado, com base na factorização de Crout, basta considerar o segundo membro $b' = (5, 4, -1)^T$ (uma vez que foi trocada a segunda com a terceira linha de U), após o que se resolvem os sistemas $L'\, g = b'$ e $U\, x = g$, utilizando, como sempre, a substuição descendente (para o primeiro sistema) e a substituição ascendente (para o segundo).

3.3.6 Factorização de Cholesky

Os dois tipos de factorização de que falámos acima existem para qualquer matriz não singular (ainda que seja necessário efectuar uma troca de linhas ou colunas). Quanto à *factorização de Cholesky*, de que vamos falar a seguir, só existe para matrizes simétricas e definidas positivas. Embora se trate de uma restrição muito forte, este tipo de factorização não deixa de ter interesse prático, visto que estas propriedades são satisfeitas pelas matrizes que surgem em muitos problemas de cálculo numérico, por exemplo, no método dos mínimos quadrados e em certos problemas de valores de fronteira para equações diferenciais. A grande vantagem desta factorização consiste em que só temos de calcular uma matriz, L, visto que uma matriz simétrica e definida positiva pode ser representada sob a forma:

$$A = L\,L^T. \tag{3.105}$$

Isto significa que o número de operações para resolver um sistema linear fica reduzido a cerca de metade, quando se compara o método de Cholesky com outros métodos de factorização ou com o método de Gauss. Para estudar melhor as propriedades da factorização de Cholesky, comecemos por demonstrar o seguinte teorema.

Teorema 3.4. Seja A uma matriz simétrica e definida positiva. Então existe uma matriz real, triangular inferior, L, tal que se verifica a factorização 3.105.

Demonstração . Provemos esta afirmação por indução em n (ordem da matriz). Para o caso de $n = 1$, a afirmação é trivial, visto que, nesse caso, A é um número positivo e L, a sua raiz quadrada. Suponhamos agora que a afirmação do teorema é verdadeira para qualquer matriz de ordem não superior a $n - 1$. Seja $\hat{A}$ o menor principal de A, de ordem $n - 1$. Verifiquemos que $\hat{A}$ também é uma matriz definida positiva. Na realidade, se A é definida positiva, então verifica-se

$$\sum_{i=1,j=1}^{n} a_{ij}\,x_i\,x_j \;>\; 0, \;\; \forall x \in \mathbf{R^n}, x \neq 0. \tag{3.106}$$

Em particular, se considerarmos $x = (x_1, x_2, \ldots, x_{n-1}, 0)$, onde x_i, $i = 1, \ldots, n-1$ são números reais arbitrários, de 3.106 resulta

$$\sum_{i=1,j=1}^{n-1} a_{ij}\,x_i\,x_j \;>\; 0; \tag{3.107}$$

o que significa que $\hat{A}$ também é definida positiva.

Vamos procurar uma matriz triangular inferior L, de ordem n, com a forma

$$L = \begin{bmatrix} \hat{L} & 0 \\ \gamma^T & z \end{bmatrix}, \tag{3.108}$$

onde $\hat{L}$ é uma matriz de ordem $n-1$, $\gamma \in \mathbf{R}^{n-1}$, $z \in \mathbf{R}$. Queremos provar que, para essa matriz L, se verifica

$$L\,L^T = \begin{bmatrix} \hat{L} & 0 \\ \gamma^T & z \end{bmatrix} \begin{bmatrix} \hat{L}^T & \gamma \\ 0 & z \end{bmatrix} = A = \begin{bmatrix} \hat{A} & c \\ c^T & d \end{bmatrix}, \tag{3.109}$$

onde $c \in \mathbf{R}^{n-1}$ e $d = a_{nn} \in \mathbf{R}$. Por hipótese da indução , existe uma matriz real, triangular inferior $\hat{L}$, tal que

$$\hat{A} = \hat{L}\,\hat{L}^T. \tag{3.110}$$

Resta provar que existe um vector $\gamma \in \mathbf{R}^{n-1}$ e um número real z, para os quais se verifica a igualdade 3.109. Quanto à existência de γ, ela resulta de o sistema $\hat{L}\,\gamma = c$ ter solução .De facto, a matriz $\hat{L}$ é não singular, uma vez que

$$(det\,\hat{L})^2 = det\,\hat{A} > 0$$

(visto que $\hat{A}$ é definida positiva).Mostremos agora que existe um número real z, para o qual a igualdade 3.109 é verdadeira. Para que a igualdade 3.109 seja verdadeira, deve verificar-se

$$det\,A = (\,det\,\hat{L})^2\,z^2 > 0. \tag{3.111}$$

Uma vez que $det\,A > 0$,a equação 3.111 tem duas raizes e o valor de z pode ser determinado, escolhendo a raiz positiva. Assim, o teorema fica demonstrado, por indução . $\square$

Observação . Note-se que o problema da factorização de Cholesky admite mais do que uma solução . Isso verifica-se pela equação 3.111, a qual admite duas raizes reais:

$$z_{1,2} = \pm\frac{\sqrt{det\,A}}{det\,\hat{L}} \tag{3.112}$$

Por convenção , escolhe-se sempre a raiz positiva desta equação , o que significa que todos os elementos da diagonal principal de L são positivos.

Vejamos agora, em termos práticos, como se pode calcular a matriz L da factorização de Cholesky. Seja a_{ij} uma componente de A, com $i \geq j$. Então, da igualdade 3.105 resulta

$$a_{ij} = \sum_{k=1}^{j} l_{ik}\, l_{jk} = \sum_{k=1}^{j-1} l_{ik}\, l_{jk} + l_{ij}\, l_{jj}, j = 1, \ldots, n;\ i = j, \ldots, n. \quad (3.113)$$

No caso de $i = j$, da igualdade 3.113 obtém-se a fórmula para os elementos da diagonal principal de L:

$$l_{ii} = \sqrt{a_{ii} - \sum_{k=1}^{i-1} l_{ik}^2};\ \ i = 1, \ldots, n. \quad (3.114)$$

De acordo com o teorema 3.2, todos os elementos da diagonal principal de L são reais, pelo que o segundo membro de 3.114 é sempre real. Uma vez calculado l_{jj}, podemos obter as restantes componentes da j-ésima coluna de L. Da fórmula 3.113 resulta:

$$l_{ij} = \frac{a_{ij} - \sum_{k=1}^{j-1} l_{ik}\, l_{jk}}{l_{jj}};\ \ i = j+1, \ldots, n. \quad (3.115)$$

Assim, usando as fórmulas 3.114 e 3.115, alternadamente, pode ser obtida a factorização de Cholesky da matriz A.

Exemplo 3.6. Consideremos a matriz de ordem n com a forma geral

$$A = \begin{bmatrix} 4 & 2 & 0 & \ldots & 0 \\ 2 & 5 & 2 & \ldots & 0 \\ 0 & 2 & 5 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & \ldots & 2 & 5 & 2 \\ 0 & \ldots & 0 & 2 & 5 \end{bmatrix}. \quad (3.116)$$

Trata-se de uma matriz simétrica tridiagonal, isto é

$$a_{ij} \neq 0 \Rightarrow |i - j| \leq 1.$$

Matrizes com estas características aparecem frequentemente nas aplicações. Vamos tentar obter a sua factorização de Cholesky. Como não é simples determinar, à partida, se a matriz dada é definida positiva, vamos tentar

utilizar as fórmulas 3.114 e 3.115 e verificar se elas são aplicáveis. Comecemos, como sempre, pela componente l_{11}. De acordo com 3.114, o seu valor é o seguinte:

$$l_{11} = \sqrt{a_{11}} = 2. \tag{3.117}$$

As restantes componentes desta coluna são dadas pela fórmula 3.115:

$$\begin{aligned} l_{21} &= \frac{a_{21}}{l_{11}} = 1; \\ l_{k1} &= \frac{a_{k1}}{l_{11}} = 0; \quad k = 3, \ldots, n. \end{aligned} \tag{3.118}$$

Vamos provar agora, por indução , que as restantes colunas da matrix L têm a mesma estrutura, isto é para a coluna j, verifica-se:

$$\begin{aligned} l_{jj} &= 2; \\ l_{j+1,j} &= 1; \\ l_{i,j} &= 0; \quad i = j+2, \ldots, n. \end{aligned} \tag{3.119}$$

Para a primeira coluna, as fórmulas 3.119 já estão provadas. Suponhamos agora que estas fórmulas são válidas para todas as colunas, até à $j-1$. Vejamos o que acontece com a coluna j. De acordo com a fórmula 3.114, podemos escrever

$$l_{jj} = \sqrt{a_{jj} - \sum_{k=1}^{j-1} l_{jk}^2} = \sqrt{a_{jj} - l_{j,j-1}^2} = 2. \tag{3.120}$$

Depois, aplicando a fórmula 3.115, obtemos

$$\begin{aligned} l_{j+1,j} &= \frac{a_{j+1,j}}{l_{jj}} = 1; \\ l_{i,j} &= 0; \qquad i = j+2, \ldots, n. \end{aligned} \tag{3.121}$$

Fica assim provado que a factorização de Cholesky da matriz dada é definida por uma matriz triangular inferior com a forma

$$L = \begin{bmatrix} 2 & 0 & 0 & \ldots & 0 \\ 1 & 2 & 0 & \ldots & 0 \\ 0 & 1 & 2 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & \ldots & 1 & 2 & 0 \\ 0 & \ldots & 0 & 1 & 2 \end{bmatrix}. \tag{3.122}$$

O determinante de A pode ser calculado com base nesta factorização :

$$det\, A \,=\, (det\, L)^2 \,=\, (l_{11}\, l_{22}\, \ldots\, l_{nn})^2 \,=\, (2^n)^2 \,=\, 4^n. \tag{3.123}$$

Uma vez que a fórmula 3.123 é válida para qualquer n, ela pode servir para calcularmos os menores principais da matriz A dada. Assim, temos

$$A_1 \,=\, 4,\; A_2 \,=\, 4^2\; \ldots, A_n \,=\, det\, A \,=\, 4^n.$$

Fica assim provado que todos os menores principais de A são positivos, de onde resulta que A é definida positiva.

3.4 Nétodos Iterativos para Sistemas Lineares

Neste parágrafo, vamos estudar alguns métodos iterativos, utilizados no cálculo aproximado de soluções de sistemas lineares. Antes disso, porém, vamos apresentar alguns conceitos gerais, relativos aos métodos iterativos, e que nos voltarão a ser úteis nos capítulos posteriores.

3.4.1 Noções Básicas sobre Métodos Iterativos

Em muitos problemas matemáticos, quando se revela impossível ou muito difícil calcular a solução exacta de uma certa equação , opta-se por obter um valor aproximado dessa solução . Esse valor aproximado é geralmente obtido por um *método de aproximações sucessivas*, ou *método iterativo*, em que cada nova aproximação é obtida a partir da anterior (ou das anteriores), através de um algoritmo conhecido, pretendendo-se deste modo tornar o erro de aproximação cada vez mais pequeno.

Definição 3.5. Seja E um espaço normado e X um subconjunto de E. Chama-se *método iterativo a p passos* em E (definido sobre X) a qualquer aplicação Ψ, que a cada sucessão de p elementos $(\xi_0, \ldots, \xi_{p-1}) \in X^p$, faz corresponder uma sucessão infinita $\{x^{(k)}\}_{k=0}^{\infty}$,
$x^{(k)} \in E$, com as seguintes propriedades:

1. Os primeiros p termos da sucessão $\{x^{(k)}\}_{k=0}^{\infty}$ são os dados:

$$x^{(i)} = \xi_i,\; i = 0, \ldots, p-1.$$

2. Os restantes elementos elementos de $\{x^{(k)}\}_{k=0}^{\infty}$ são obtidos a partir destes, de acordo com a fórmula

$$x^{(k+p)} = \phi(x^k, x^{k+1}, \ldots, x^{k+p-1}),$$

onde ϕ é uma função conhecida (chamada a *função iteradora*) com domínio em X^p e valores em E.

Naturalmente, na prática só se calcula um número finito de termos da sucessão $\{x^{(k)}\}_{k=0}^{\infty}$ (também chamados *iteradas*), tantos quantos necessários para alcançar a precisão pretendida. Por isso, a cada método iterativo estão geralmente associados *critérios de paragem*, isto é, regras que nos permitem verificar se uma dada iterada tem ou não a precisão exigida.

Definição 3.6. Dizemos que um método iterativo a p passos, definido sobre x, é convergente para um certo $x \in \mathbf{R^n}$, se, para quaisquer valores iniciais $(\xi_0, \ldots, \xi_{p-1}) \in A$ se verificar $x^{(k)} \to x$, quando $k \to \infty$ (segundo a norma considerada em $\mathbf{R^n}$).

Note-se que a convergência em espaços de dimensão finita não depende da norma considerada. Daí que, no caso dos métodos iterativos para sistemas lineares, que vamos estudar nos próximos parágrafos, a convergência numa certa norma é equivalente à convergência noutra norma qualquer.

Além da convergência, outra propriedade importante dos métodos iterativos é a sua *estabilidade*.

Definição 3.7. Um método iterativo Ψ a p passos, definido no conjunto X, diz-se *estável em* $A \subset X$, se existir uma constante $C > 0$, tal que

$$\max_{n \in N} \|x^{(n)} - y^{(n)}\| \leq C \max_{i=1,\ldots,n} \|\xi_i - \eta_i\| \; \forall \xi, \eta \in A^p, \tag{3.124}$$

onde $\{x^{(n)}\}$ e $\{y^{(n)}\}$ são , respectivamente, as sucessões geradas a partir de $\xi = (\xi_0, \xi_1, \ldots, \xi_{p-1})$ e $\eta = (\eta_0, \eta_1, \ldots, \eta_{p-1})$.

Para representar o erro da k-ésima iterada usaremos a notação $e^{(k)}$, ou seja, $e^{(k)} = x^{(k)} - x$. Nos próximos parágrafos, vamos analisar alguns métodos iterativos para o cálculo aproximado da solução do sistema linear

$$A\,x = b, \tag{3.125}$$

onde $A \in \mathbf{R}^{n \times n}$ e $b \in \mathbf{R^n}$. Suponhamos, como habitualmente, que a matriz A é não singular, pelo que o sistema (3.125) tem uma única solução .

Com o objectivo de construir um método iterativo, começamos por reduzir o sistema (3.125) à forma equivalente

$$x = G(x) = Cx + g, \tag{3.126}$$

onde C é uma certa matriz (a que chamaremos *matriz de iteração*), e g é um vector auxiliar ($g \in \mathbf{R^n}$). Uma vez escrito o sistema na forma (3.126), podemos dizer que a sua solução é um ponto fixo da função G (função definida em $\mathbf{R^n}$ e com valores no mesmo espaço). A ideia é determinar o ponto fixo de G por um método análogo ao método do ponto fixo, utilizado no capítulo anterior para aproximar os pontos fixos de funções de uma variável. Assim, dada uma certa aproximação inicial $x^{(0)} \in \mathbf{R^n}$,vamos construir uma sucessão de vectores através da seguinte fórmula de recorrência:

$$x^{(k+1)} = G(x^{(k)}) = Cx^{(k)} + g, \qquad k = 0, 1, \ldots \tag{3.127}$$

Naturalmente, esta transformação do sistema pode ser feita de muitas maneiras diferentes, dando origem a diferentes métodos iterativos. Vamos descrever dois deles.

3.4.2 Método de Jacobi

Para deduzir a fórmula do método de Jacobi, começamos por reescrever o sistema (3.125) do seguinte modo:

$$\begin{array}{c} x_1 = \frac{b_1 - a_{12}x_2 - a_{13}x_3 - \cdots - a_{1n}x_n}{a_{11}} \\ x_2 = \frac{b_2 - a_{21}x_1 - a_{23}x_3 - \cdots - a_{2n}x_n}{a_{22}} \\ \cdots \\ x_n = \frac{b_n - a_{n1}x_1 - a_{n2}x_2 - \cdots - a_{n,n-1}x_{n-1}}{a_{nn}} \end{array} . \tag{3.128}$$

Ficamos asim com o sistema escrito na forma (3.126), de tal modo que o sistema (3.128) é equivalente ao inicial. Para poder escrever o sistema nesta forma, basta que todos os elementos da diagonal principal da matriz A sejam diferentes de 0 ($a_{ii} \neq 0$,$i = 1, \ldots, n$). Se iterarmos agora a função

G correspondente ao sistema (3.128), obtém-se a seguinte fórmula iteradora:

$$\begin{aligned} x_1^{(k+1)} &= \frac{b_1 - a_{12}x_2^{(k)} - a_{13}x_3^{(k)} - \cdots - a_{1n}x_n^{(k)}}{a_{11}} \\ x_2^{(k+1)} &= \frac{b_2 - a_{21}x_1^{(k)} - a_{23}x_3^{(k)} - \cdots - a_{2n}x_n^{(k)}}{a_{22}} \qquad , \qquad k = 0, 1, 2, \ldots \\ &\ldots \\ x_n^{(k+1)} &= \frac{b_n - a_{n1}x_1^{(k)} - a_{n2}x_2^{(k)} - \cdots - a_{n,n-1}x_{n-1}^{(k)}}{a_{nn}} \end{aligned} \tag{3.129}$$

A fórmula (3.129) pode escrever-se na seguinte forma compacta:

$$x_i^{(k+1)} = \frac{b_i}{a_{ii}} - \frac{\sum_{j=1,\ldots,n, j\neq i}^{n} a_{ij}x_j^{(k)}}{a_{ii}}, \tag{3.130}$$

$i = 1, \ldots, n,\ k = 0, 1, 2, \ldots$

Exemplo 3.7. Consideremos o sistema $A\,x = b$, onde

$$A = \begin{bmatrix} 2 & 1 & 0 \\ -1 & 2 & 1 \\ 0 & -1 & 2 \end{bmatrix}, \quad b = \begin{bmatrix} 2 \\ 2 \\ 1 \end{bmatrix}. \tag{3.131}$$

1. Efectuar uma iteração do método de Jacobi, tomando como aproximação inicial $x^{(0)} = (0.5, 0.8, 1)$.

   Resolução:

$$\begin{aligned} x_1^{(1)} &= \frac{b_1 - a_{12}x_2^{(0)} - a_{13}x_3^{(0)}}{a_{11}} = \tfrac{1}{2}\,(2 - 0.8 - 0) = 0.6; \\ x_2^{(1)} &= \frac{b_2 - a_{21}x_1^{(0)} - a_{23}x_3^{(0)}}{a_{22}} = \tfrac{1}{2}(2 + 0.5 - 1) = 0.75; \\ x_3^{(1)} &= \frac{b_3 - a_{31}x_1^{(0)} - a_{32}x_2^{(0)}}{a_{33}} = \tfrac{1}{2}(1 - 0 + 0.8) = 0.9; \end{aligned} \tag{3.132}$$

2. Sabendo que a solução exacta do sistema é $x = (0.583, 0.833, 0.97)$, calcule $\|e^{(0)}\|_1$ e $\|e^{(1)}\|_1$.

   Resolução:

$$\begin{aligned} e^{(0)} &= x - x^{(0)} = (0.083, 0.033, -0.083) \quad \|e^{(0)}\|_1 = 0.199; \\ e^{(1)} &= x - x^{(1)} = (-0.0017, 0.083, 0.017) \quad \|e^{(1)}\|_1 = 0.117. \end{aligned} \tag{3.133}$$

   A norma do erro da primeira iterada é menor, o que significa que ela está mais próxima da solução exacta do que a aproximação inicial. No entanto, daqui nada podemos concluir sobre a convergência do método. Esta será analisada mais tarde, segundo os métodos que vamos estudar nos próximos parágrafos.

3.4.3 Método de Gauss-Seidel

O método de Gauss-Seidel é um dos métodos iterativos mais comuns para resolução de sistemas lineares. Para deduzirmos a sua forma iteradora, partimos de novo do sistema na forma (3.128).

Neste caso, porém, a fórmula iteradora tem o seguinte aspecto:

$$\begin{aligned} x_1^{(k+1)} &= \frac{b_1 - a_{12}x_2^{(k)} - a_{13}x_3^{(k)} - \cdots - a_{1n}x_n^{(k)}}{a_{11}} \\ x_2^{(k+1)} &= \frac{b_2 - a_{21}x_1^{(k+1)} - a_{23}x_3^{(k)} - \cdots - a_{2n}x_n^{(k)}}{a_{22}} \\ &\cdots \\ x_n^{(k+1)} &= \frac{b_n - a_{n1}x_1^{(k+1)} - a_{n2}x_2^{(k+1)} - \cdots - a_{n,n-1}x_{n-1}^{(k+1)}}{a_{nn}} \end{aligned} \quad , \qquad k = 0, 1, 2, \ldots \tag{3.134}$$

A diferença em relação ao método de Jacobi está em que, para determinar a componente $x_i^{(k+1)}$ da iterada $(k+1)$ (com $i > 1$) utilizamos as componentes $x_1^{(k+1)}, ..., x_{i-1}^{(k+1)}$ dessa mesma iterada, enquanto que no método de Jacobi as componentes de $x^{(k+1)}$ eram calculadas apenas a partir das componentes de $x^{(k)}$ (iterada anterior). A fórmula (3.134) pode ser escrita na seguinte forma compacta:

$$x_i^{(k+1)} = \frac{b_i}{a_{ii}} - \frac{\sum_{j=1}^{i-1} a_{i,j}x_j^{(k+1)} + \sum_{j=i+1}^{n} a_{i,j}x_j^{(k)}}{a_{ii}}, \tag{3.135}$$

$i = 1, \ldots, n$, $k = 0, 1, 2, \ldots$. Vejamos um exemplo de aplicação.

Exemplo 3.8. Consideremos de novo o sistema (3.131).

1. Efectuar uma iteração do método de Gauss-Seidel, tomando como aproximação inicial $x^{(0)} = (0.5, 0.8, 1)$.

   Resolução:

$$\begin{aligned} x_1^{(1)} &= \frac{b_1 - a_{12}x_2^{(0)} - a_{13}x_3^{(0)}}{a_{11}} = \tfrac{1}{2}(2 - 0.8 - 0) = 0.6; \\ x_2^{(1)} &= \frac{b_2 - a_{21}x_1^{(1)} - a_{23}x_3^{(0)}}{a_{22}} = \tfrac{1}{2}(2 + 0.6 - 1) = 0.8; \\ x_3^{(1)} &= \frac{b_3 - a_{31}x_1^{(1)} - a_{32}x_2^{(1)}}{a_{33}} = \tfrac{1}{2}(1 - 0 + 0.8) = 0.9; \end{aligned} \cdot \tag{3.136}$$

2. Sabendo que a solução exacta do sistema é $x = (0.583, 0.833, 0.97)$, calcule $\|e^{(0)}\|_1$ e $\|e^{(1)}\|_1$.

Resolução:

$$\begin{array}{ll} e^{(0)} = x - x^{(0)} = (0.083, 0.033, -0.083) & \|e^{(0)}\|_1 = 0.199; \\ e^{(1)} = x - x^{(1)} = (-0.0017, 0.033, 0.017) & \|e^{(1)}\|_1 = 0.067. \end{array} \tag{3.137}$$

Tal como acontecia no caso do método de Jacobi, também aqui a norma do erro diminui da aproximação inicial para a primeira iterada, o que significa que esta está mais próxima da solução exacta do que a aproximação inicial. Na caso do método de Gauss-Seidel, a diferença entre os dois erros é mais acentuada. Embora, como já foi dito, nada se possa concluir destes factos, eles estão de acordo com a análise que vamos fazer mais adiante sobre a convergência dos métodos iterativos.

3.4.4 Forma Matricial dos Métodos Iterativos

Para podermos escrever as fórmulas iteradoras dos métodos iterativos de um modo mais compacto e estudar a sua convergência, convém representá-los na forma matricial. Para isso, dada uma certa matriz A, começamos por definir as matrizes L, D, U, tais que

$$L = \begin{bmatrix} 0 & 0 & \dots & 0 \\ a_{21} & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & 0 \end{bmatrix}, D = \begin{bmatrix} a_{11} & 0 & \dots & 0 \\ 0 & a_{22} & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & a_{nn} \end{bmatrix}, U = \begin{bmatrix} 0 & a_{12} & \dots & a_{1n} \\ 0 & 0 & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 0 \end{bmatrix}. \tag{3.138}$$

Obviamente, verifica-se $A = L + D + U$. Nesta secção, vamos supor que todas as entradas diagonais da matriz A são diferentes de zero, ou seja, $a_{ii} \neq 0$, $i = 1, \dots, n$. Daqui resulta que a matriz D é invertível.

Método de Jacobi na Forma Matricial

Utilizando estas notações, vejamos como se pode escrever a fórmula iteradora do método de Jacobi na forma (3.127), identificando o vector g e a matriz de iteração correspondentes.

Começamos por escrever a fórmula (3.130) com o auxílio das matrizes L, D e U:

$$x^{(k+1)} = D^{-1}\left(b - Lx^{(k)} - Ux^{(k)}\right). \tag{3.139}$$

Esta equação também pode ser escrita na seguinte forma:

$$x^{(k+1)} = D^{-1}b - D^{-1}(L + U)x^{(k)}. \tag{3.140}$$

Comparando esta equação com a fórmula geral para os métodos iterativos (3.127), concluímos que, no caso do método de Jacobi, o vector auxiliar g e a matriz de iteração têm a forma

$$g_j = D^{-1}b, \qquad C_j = -D^{-1}(L+U). \tag{3.141}$$

Uma vez que a matriz D é diagonal e que todas as entradas da sua diagonal são diferentes de zero, a sua inversa pode ser determinada imediatamente:

$$D^{-1} = \begin{bmatrix} \frac{1}{a_{11}} & 0 & \dots & 0 \\ 0 & \frac{1}{a_{22}} & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \frac{1}{a_{nn}} \end{bmatrix}. \tag{3.142}$$

Por conseguinte, a matriz de iteração C_j tem a seguinte forma:

$$C_j = -D^{-1}(L+U) = \begin{bmatrix} 0 & -\frac{a_{12}}{a_{11}} & \dots & -\frac{a_{1n}}{a_{11}} \\ -\frac{a_{21}}{a_{22}} & 0 & \dots & -\frac{a_{2n}}{a_{22}} \\ \dots & \dots & \dots & \dots \\ -\frac{a_{n1}}{a_{nn}} & -\frac{a_{n2}}{a_{nn}} & \dots & 0 \end{bmatrix}. \tag{3.143}$$

Método de Gauss-Seidel na Forma Matricial

Vejamos agora como se pode escrever a fórmula iteradora do método de Gauss-Seidel na forma (3.127).

Com o auxílio das matrizes L, D e U, a fórmula (3.135) tem o seguinte aspecto:

$$x^{(k+1)} = D^{-1}\left(b - Lx^{(k+1)} - Ux^{(k)}\right). \tag{3.144}$$

Multiplicando ambos os membros de (3.144) à esquerda por D, obtém-se:

$$Dx^{(k+1)} = b - Lx^{(k+1)} - Ux^{(k)}. \tag{3.145}$$

Juntando no primeiro membro os termos que contêm $x^{(k+1)}$, temos a seguinte equação:

$$(L+D)x^{(k+1)} = b - Ux^{(k)}. \tag{3.146}$$

Uma vez que a matriz D, por condição, é invertível, $L+D$ também o é (o seu determinante é igual ao determinante de D). Assim, podemos escrever

$$x^{(k+1)} = (L+D)^{-1}b - (L+D)^{-1}Ux^{(k)}. \tag{3.147}$$

Finalmente, comparando a equação (3.147) com a fórmula geral para os métodos iterativos, concluímos que, no caso do método de Gauss-Seidel, o vector auxiliar g e a matriz de iteração têm a forma

$$g_{gs} = (L+D)^{-1}b, \qquad C_{gs} = -(L+D)^{-1}U. \tag{3.148}$$

Neste caso, não é possível encontrar uma forma explícita para a inversa de $(L+D)$. Tudo o que se pode dizer é que é uma matriz triangular inferior e que os seus elementos diagonais são os inversos dos elementos diagonais de A. Logo, também não é possível encontrar uma forma explícita para a matriz de iteração C_{gs}.

Exemplo 3.9. Determinemos os vectores g e as matrizes de iteração dos métodos de Jacobi e do método de Gauss-Seidel para o sistema do exemplo 3.7. No caso do método de Jacobi, o vector g tem a forma

$$g_j = D^{-1}b = (\frac{b_1}{a_{11}}, \frac{b_2}{a_{22}}, \frac{b_3}{a_{33}}). \tag{3.149}$$

A matriz C_J obtém-se da aplicação directa da fórmula (3.143):

$$C_J = -D^{-1}(L+U) = \begin{bmatrix} 0 & -\frac{1}{2} & 0 \\ \frac{1}{2} & 0 & -\frac{1}{2} \\ 0 & \frac{1}{2} & 0 \end{bmatrix}. \tag{3.150}$$

No caso do método de Gauss-Seidel, para poder determinar o vector g e a matriz de iteração começamos por calcular a inversa de $L+D$:

$$(L+D)^{-1} = \begin{bmatrix} \frac{1}{2} & 0 & 0 \\ \frac{1}{4} & \frac{1}{2} & 0 \\ \frac{1}{8} & \frac{1}{4} & \frac{1}{2} \end{bmatrix}. \tag{3.151}$$

Daqui, pelas fórmulas (3.148), resulta que

$$g_{gs} = (1, 1, \frac{5}{4}), \qquad C_{gs} = \begin{bmatrix} 0 & -\frac{1}{2} & 0 \\ 0 & -\frac{1}{4} & -\frac{1}{2} \\ 0 & -\frac{1}{8} & -\frac{1}{4} \end{bmatrix}. \tag{3.152}$$

3.4.5 Convergência dos Métodos Iterativos para Sistemas Lineares

Uma vez definido um método iterativo para a aproximação da solução de um sistema, é fundamental saber em que condições esse método gera uma sucessão que converge para a solução exacta. Nos teoremas que se seguem,

estabelecem-se condições sobre a matriz A que garantem a convergência dos métodos iterativos considerados.

Conforme resulta das fórmulas (3.126) e (3.127), os erros das iteradas, que representaremos por $e^{(k)}$, $k = 0, 1, \ldots$, satisfazem a igualdade

$$e^{(k+1)} = x^{(k+1)} - x = C(x^{(k)} - x); \; k = 0, 1, 2, \ldots \tag{3.153}$$

A igualdade 3.153 também pode ser escrita na forma

$$e^{(k+1)} = C\, e^{(k)}; \; k = 0, 1, 2, \ldots \tag{3.154}$$

onde C é a matriz de iteração do método considerado. No parágrafo anterior já foi analisada a forma das matrizes de iteração dos métodos de Jacobi e Gauss-Seidel.

Vejamos agora que propriedades deve apresentar a matriz C para garantir a convergência de um método iterativo. Em primeiro lugar, notemos que da igualdade 3.154 resulta imediatamente uma fórmula que exprime o erro de qualquer iterada através do erro da aproximação inicial:

$$e^{(k)} = C^k\, e^{(0)}; \; k = 0, 1, 2, \ldots \tag{3.155}$$

Definição 3.7 Uma matriz $C \in \mathbf{R}^{n \times n}$, que representa uma aplicação linear no espaço vectorial $\mathbf{R^n}$, diz-se *convergente* se e só se

$$\lim_{k \to \infty} C^k\, x = 0, \; \forall x \in \mathbf{R^n}. \tag{3.156}$$

Estamos agora em condições de formular o teorema que nos dá uma condição necessária e suficiente para a convergência dos métodos iterativos baseados na fórmula 3.127.

Teorema 3.5. Seja $\{x^{(k)}\}_{k=0}^{\infty}$ uma sucessão em $\mathbf{R^n}$, gerada pela fórmula 3.127, com uma certa matriz C. Então esta sucessão converge para a solução do sistema $Ax = b$, qualquer que seja a aproximação inicial x_0, se e só se a matriz C for convergente.

Demonstração .*Condição suficiente.* Seja C uma matriz convergente e seja $e^{(k)}$ o erro da k-ésima iterada. Então , de acordo com as fórmulas 3.155 e 3.156, temos

$$\lim_{k \to \infty} e^{(k)} = \lim_{k \to \infty} C^k\, e^{(0)} = 0, \tag{3.157}$$

qualquer que seja o vector $e^{(0)} \in \mathbf{R^n}$, independentemente da norma considerada. Isto significa que o método iterativo converge, qualquer que seja a aproximação inicial $x^{(0)} \in \mathbf{R^n}$.

Condição necessária. Suponhamos que C não é convergente. Então existe um vector $v \in \mathbf{R^n}$ tal que a sucessão $\{C^k\, v\}_{k=0}^{\infty}$ não converge para o vector nulo. Seja $x^{(0)} = x + v$, onde x é a solução exacta do sistema. Então , de acordo com 3.155, temos $e^{(k)} = C^k\, v$ e, de acordo com a definição de v, a sucessão $\{e^{(k)}\}_{k=0}^{\infty}$ não tende para o vector nulo. Isto, por sua vez, significa que o método iterativo não é convergente, no caso da aproximação inicial $x^{(0)} = x + v$. $\square$

Na prática, pode não ser fácil averiguar se a matriz C é ou não convergente. Vamos a seguir apresentar dois teoremas que nos ajudam a verificar esta propriedade.

Teorema 3.6. Seja $C \in \mathbf{R}^{n\times n}$. Se existir uma norma matricial M, associada a uma certa norma vectorial V, tal que

$$\|C\|_M < 1 \tag{3.158}$$

então a matriz C é convergente.

Demonstração . Seja x um vector arbitrário de $\mathbf{R^n}$. Então , de acordo com as propriedades das normas matriciais, temos

$$\|C^k\, x\|_V \leq \|C^k\|_M\, \|x\|_V \leq (\|C\|_M)^k\, \|x\|_V. \tag{3.159}$$

Da desigualdade 3.159 resulta imediatamente que, se $\|C\|_M < 1$, então

$$\lim_{k\to\infty} \|C^k\, x\|_V = 0, \tag{3.160}$$

o que significa, por definição , que a matriz C é convergente. $\square$

Teorema 3.7. Para que a matriz $C \in \mathbf{R}^{n\times n}$ seja convergente é necessário e suficiente que

$$r_\sigma(C) < 1. \tag{3.161}$$

Demonstração .*Condição suficiente.* Se tivermos $r_\sigma(C) = \rho < 1$, de acordo com o teorema 2.1.31 de [4], para qualquer $\epsilon > 0$, existe uma norma matricial $N(\epsilon)$ tal que

$$\|C\|_{N(\epsilon)} \leq \rho + \epsilon. \tag{3.162}$$

Se considerarmos $\epsilon = \frac{1-\rho}{2}$, obtemos

$$\|C\|_{N(\epsilon)} \leq \frac{\rho + 1}{2} < 1. \tag{3.163}$$

Da desigualdade 3.163 resulta, pelo teorema 3.5, que a matriz C é convergente.

Condição necessária. Suponhamos que a condição 3.161 não se verifica, isto é, que $r_\sigma(C) \geq 1$. Então existe, pelo menos, um valor próprio λ de C, tal que $|\lambda| = \rho \geq 1$. Seja v um vector próprio de C, associado ao valor próprio λ. Então , para qualquer norma vectorial, verifica-se

$$\|C^k\, v\| = \|\lambda^k\, v\| = |\lambda|^k\, \|v\|. \tag{3.164}$$

Da igualdade 3.164, visto que $|\lambda| = \rho \geq 1$, resulta que a sucessão $\{C^k\, v\}_{k=0}^{\infty}$ não converge para o vector nulo, pelo que a matriz C não é convergente.$\square$

Se aplicarmos os teoremas 3.5 e 3.6 aos métodos de Jacobi e Gauss-Seidel, podemos obter outros critérios de convergência para estes métodos. A fim de formularmos estes critérios numa forma mais concisa, vamos introduzir mais algumas definições .

Definição 3.8. Diz-se que a matriz $A \in \mathbf{R}^{n\times n}$ tem a *diagonal estritamente dominante por linhas*, se forem satisfeitas as condições :

$$\sum_{j=1, j\neq i}^{n} |a_{ij}| < |a_{ii}|,\ i = 1, \ldots, n. \tag{3.165}$$

Definição 3.9 . Diz-se que a matriz $A \in \mathbf{R}^{n\times n}$ tem a *diagonal estritamente dominante por colunas*, se forem satisfeitas as condições :

$$\sum_{i=1, i\neq j}^{n} |a_{ij}| < |a_{jj}|,\ j = 1, \ldots, n. \tag{3.166}$$

Do teorema 3.6 resulta o seguinte critério de convergência para o método de Jacobi.

Teorema 3.8. Se a matriz A tiver a diagonal estritamente dominante por linhas, então o método de Jacobi converge para a solução do sistema $A\,x = b$, qualquer que seja a aproximação inicial $x^{(0)} \in \mathbf{R^n}$.

Demonstração . Se a matriz A tiver a diagonal estritamente dominante por linhas, das desigualdades 3.165 resulta

$$\sum_{j=1, j\neq i}^{n} \frac{|a_{ij}|}{|a_{ii}|} < 1 \; i = 1, \ldots, n. \tag{3.167}$$

De acordo com a forma da matriz C_J, dada por (3.141), as desigualdades 3.167 implicam

$$\|C_J\|_\infty = \max_{i=1,\ldots,n} \sum_{j=1, j\neq i}^{n} \frac{|a_{ij}|}{|a_{ii}|} < 1, \tag{3.168}$$

De acordo com o teorema 3.6, a condição 3.168 garante que a matriz C_J é convergente. Isto, por sua vez, implica, de acordo com o teorema 3.5, que o método de Jacobi é convergente, qualquer que seja a aproximação inicial. □

Um critério análogo é válido no caso de a matriz A ter a diagonal estritamente dominante por colunas.

Teorema 3.9. Se a matriz A tiver a diagonal estritamente domi nante por colunas, então o método de Jacobi converge para a solução do sistema $A\,x = b$, qualquer que seja a aproximação inicial $x^{(0)} \in \mathbf{R^n}$.

Demonstração . Suponhamos que a matriz A satisfaz as condições 3.166. Seja D uma matriz diagonal com a forma $D = diag(a_{11}, \ldots, a_{nn})$. Então podemos definir uma norma matricial M pela igualdade

$$\|C\|_M = \|D\,C\,D^{-1}\|_1, \; \forall C \in L^n. \tag{3.169}$$

Das condições 3.166 obtem-se facilmente que

$$\|C_J\|_M = \|D\,C_J\,D^{-1}\|_1 < 1. \tag{3.170}$$

De acordo com o teoremas 3.5 e 3.6, da desigualdade 3.170 resulta que o método de Jacobi converge para a solução do sistema $A\,x = b$, qualquer que seja a aproximação inicial $x^{(0)} \in \mathbf{R^n}$. □.

Exemplo 3.10 Voltemos ao sistema do exemplo 3.7 . Recordemos que

a matriz A daquele sistema tem a forma

$$A = \begin{bmatrix} 2 & 1 & 0 \\ -1 & 2 & 1 \\ 0 & -1 & 2 \end{bmatrix}. \tag{3.171}$$

Verifica-se facilmente que esta matriz não tem a diagonal estritamente dominante por linhas, uma vez que, neste caso,

$$|a_{21}| + |a_{23}| = 2 = |a_{22}|.$$

Do mesmo modo se pode verificar que aquela matriz também não tem a diagonal estritamente dominante por colunas. Por conseguinte, os teoremas 3.8 e 3.9 não são , neste caso, aplicáveis. Vejamos se é possível aplicar directamente o teorema 3.7.

A matriz C_J, neste caso, tem a forma:

$$C_J = \begin{bmatrix} 0 & -1/2 & 0 \\ 1/2 & 0 & -1/2 \\ 0 & 1/2 & 0 \end{bmatrix}. \tag{3.172}$$

Os valores próprios desta matriz são as raizes da equação

$$\lambda^3 + \frac{\lambda}{2} = 0.$$

Determinando estas raizes, obtém-se

$$\lambda_1 = 0,\ \lambda_2 = \frac{i}{\sqrt{2}},\ \lambda_3 = -\frac{i}{\sqrt{2}}.$$

Por conseguinte, o raio espectral de C_J é

$$r_\sigma(C_J) = |\lambda_2| = \frac{1}{\sqrt{2}} < 1.$$

Logo, pelo teorema 3.7, podemos concluir que o método de Jacobi converge para a solução do sistema considerado, qualquer que seja a aproximação inicial.

Vamos agora averiguar as condições que garantem a convergência do método de Gauss-Seidel. Representemos por C_{gs} a matriz

$$C_{gs} = -(L + D)^{-1}U \tag{3.173}$$

De acordo com o teorema 3.5, o método de Gauss-Seidel converge, qualquer que seja a aproximação inicial, se e só se a matriz C_S for convergente. Para isso, de acordo com o teorema 3.7, é necessário e suficiente que o raio espectral daquela matriz seja menor que 1.

Por outro lado, pode mostrar-se que o método de Gauss-Seidel, tal como o de Jacobi, converge sempre que a matriz do sistema tem a diagonal estritamente dominante por linhas. Para isso, comecemos por introduzir as seguintes notações :

$$\alpha_i = \sum_{j=1}^{i-1} \left| \frac{a_{ij}}{a_{ii}} \right|, \ \beta_i = \sum_{j=i+1}^{n} \left| \frac{a_{ij}}{a_{ii}} \right|. \tag{3.174}$$

Sendo conhecidos α_i e β_i,$i = 1, \ldots, n$, define-se a grandeza η através da fórmula

$$\eta = \max_{i=1,\ldots,n} \left(\frac{\beta_i}{1 - \alpha_i} \right). \tag{3.175}$$

Note-se que, se a matriz A tiver a diagonal estritamente dominante, por linhas, então $\alpha_i + \beta_i < 1, i = 1, \ldots, n$, de onde resulta que $\eta < 1$.

Teorema 3.10. Seja A uma matriz com a diagonal estritamente dominante por linhas. Então o método de Gauss-Seidel converge, qualquer que seja a aproximação inicial e é válida a estimativa do erro

$$\|e^{(k)}\|_\infty \leq \eta^k \, \|e^{(0)}\|_\infty. \tag{3.176}$$

Demonstração . Da fórmula 3.135 deduz-se facilmente que o erro da k-ésima iterada do método de Gauss-Seidel satisfaz a igualdade

$$e_i^{(k+1)} = \frac{1}{a_{ii}} \left(- \sum_{j=1}^{i-1} a_{ij} \, e_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij} \, e_j^{(k)} \right), \ i = 1, 2, \ldots, n; \ k = 0, 1, \ldots. \tag{3.177}$$

Tomando o módulo de ambos os membros da igualdade 3.177 e entrando em conta com as definições das grandezas α_i e β_i, obtém-se

$$|e_i^{(k+1)}| \leq \alpha_i \, \|e^{(k+1)}\|_\infty + \beta_i \|e^{(k)}\|_\infty, \ i = 1, 2, \ldots, n; \ k = 0, 1, \ldots. \tag{3.178}$$

Seja m o índice para o qual se verifica $|e_m^{(k+1)}| = \|e^{(k+1)}\|_\infty$. Então , escrevendo a desigualdade 3.178, com $i = m$, obtém-se

$$\|e^{(k+1)}\|_\infty \leq \alpha_m \, \|e^{(k+1)}\|_\infty + \beta_m \|e^{(k)}\|_\infty, \ k = 0, 1, \ldots. \tag{3.179}$$

Daqui resulta

$$\|e^{(k+1)}\|_\infty(1 - \alpha_m) \leq \beta_m\|e^{(k)}\|_\infty, \; k = 0, 1, \ldots. \tag{3.180}$$

Visto que $\alpha_m < 1$, podemos dividir ambos os membros de 3.180 por $1 - \alpha_m$ e obter

$$\|e^{(k+1)}\|_\infty \leq \frac{\beta_m}{1 - \alpha_m}\|e^{(k)}\|_\infty \leq \eta\|e^{(k)}\|_\infty, \; k = 0, 1, \ldots. \tag{3.181}$$

Da desigualdade 3.181 resulta a estimativa do erro 3.176. Por outro lado, uma vez que a matriz tem a diagonal estritamente dominante, por linhas, $\eta < 1$. Logo, a desigualdade 3.176 implica que

$$\lim_{k\to\infty} \|e^{(k)}\|_\infty = 0,$$

o que garante a convergência do método de Gauss-Seidel, qualquer que seja a aproximação inicial . $\square$

Exemplo 3.11 Consideremos o mesmo sistema linear dos exemplos anteriores. Neste caso, iremos debruçar-nos sobre o convergência do método de Gauss-Seidel. Já vimos que a matriz deste sistema não tem a diagonal estritamente dominante por linhas. Por conseguinte, o teorema 3.10 não é, neste caso, aplicável. Vejamos se é possível aplicar directamente o teorema 3.7.

A matriz C_{gs}, de acordo com (3.151), tem a forma:

$$C_{gs} = \begin{bmatrix} 0 & -1/2 & 0 \\ 0 & -1/4 & -1/2 \\ 0 & -1/8 & -1/4 \end{bmatrix}. \tag{3.182}$$

Os valores próprios desta matriz são as raizes da equação

$$\lambda^3 + \frac{\lambda^2}{2} = 0.$$

Determinando estas raizes, obtém-se

$$\lambda_1 = \lambda_2 = 0, \quad \lambda_3 = -\frac{1}{2}.$$

Por conseguinte, o raio espectral de C_{gs} é

$$r_\sigma(C_{gs}) = |\lambda_3| = \frac{1}{2}.$$

Logo, pelo teorema 3.7, podemos concluir que o método de Gauss-Seidel converge para a solução do sistema considerado, qualquer que seja a aproximação inicial.

3.4.6 Rapidez de Convergência dos Métodos Iterativos. Análise do Erro

Nos parágrafos anteriores, estudámos as condições que garantem a convergência dos métodos iterativos de Jacobi e de Gauss-Seidel. Atendendo aos resultados já obtidos, vamos agora classificar estes métodos e compará-los quanto à rapidez de convergência. Considerando qualquer norma vectorial V e a norma matricial M, a ela associada, podemos afirmar que, para qualquer método iterativo que verifique a igualdade 3.154, se verifica

$$\|e^{(k+1)}\|_V \leq \|C\|_M \, \|e^{(k)}\|_V. \tag{3.183}$$

A rapidez de convergência depende, naturalmente, das propriedades da matriz C e da aproximação inicial escolhida. Nalguns casos especiais, pode acontecer que a solução exacta seja obtida com um número finito de iterações . No entanto, na maioria dos casos com interesse prático, verifica-se que a ordem de convergência dos métodos aqui analisados é precisamente 1, ou seja, *a convergência é linear*. Como sabemos, a rapidez de convergência de métodos da mesma ordem é caracterizada pelo factor assimpótico de convergência. Para avaliar esse factor, recorre-se frequentemente ao limite

$$c_1 = \lim_{k\to\infty} \frac{\|e^{(k+1)}\|_V}{\|e^{(k)}\|_V}. \tag{3.184}$$

A existência do limite c_1 depende das propriedades da matriz C e da norma V considerada. Além disso, para a mesma matriz C, o limite pode ter diferentes valores, conforme a aproximação inicial escolhida. Pode-se mostrar que, se a matriz C tiver um único valor próprio $\lambda \in \mathbf{R}$, tal que $|\lambda| = r_\sigma(C)$, então, para a maioria das aproximações iniciais, o limite c_1 existe e verifica-se $c_1 = r_\sigma(C)$. Logo, se o limite c_1 existir e o método iterativo vonvergir, então $0 < c_1 < 1$ e este valor pode ser tomado como o factor assimptótico de convergência. Isto significa que, para valores de c_1 próximos de 0, teremos convergência rápida, enquanto que para valores de c_1 próximos de 1 teremos convergência lenta (isto é, são necessárias muitas iterações para atingir uma dada precisão). Na prática, o valor de c_1 não pode ser obtido directamente da fórmula 3.184, uma vez que os valores $\|e^{(k+1)}\|_V$ e $\|e^{(k)}\|_V$ não são, em geral, conhecidos para nenhuma iterada. Por isso, recorre-se frequentemente à igualdade

$$x^{(k+1)} - x^{(k)} = -e^{(k+1)} + e^{(k)} = -C\,e^{(k)} + C\,e^{(k-1)} = C(x^{(k)} - x^{(k-1)}). \tag{3.185}$$

Desta igualdade depreende-se que a diferença entre iteradas sucessivas varia com k do mesmo modo que o erro $e^{(k)}$ (ambas estas grandezas satisfazem uma relação do tipo 3.154. Logo, se o limite 3.184 existir, também existe o limite

$$c_1' = \lim_{k\to\infty} \frac{\|x^{(k+1)} - x^{(k)}\|_V}{\|x^{(k)} - x^{(k-1)}\|_V}. \tag{3.186}$$

e os dois limites (c_1 e c_1') têm o mesmo valor, para a maioria das aproximações iniciais. A fórmula 3.186 pode ser utilizada na prática para avaliar c_1' e, logo, c_1. Com esse fim, calcula-se, para sucessivos valores de k, a razão

$$r^{(k)} = \frac{\|x^{(k+1)} - x^{(k)}\|_V}{\|x^{(k)} - x^{(k-1)}\|_V},$$

até que o seu valor estabilize. O número assim obtido é tomado como uma estimativa de c_1. Por outro lado, os valores de $r^{(k)}$ também podem ser utilizados para obter estimativas do erro $e^{(k)}$. Na realidade, se considerarmos um valor c_2 tal que $r^{(k)} \leq c_2$, $\forall k > k_0$ (aqui k_0 representa a ordem a partir da qual o valor de $r^{(k)}$ estabiliza), podemos esperar que, para $k > k_0$, se verifique

$$\|e^{(k+1)}\|_V = \|x^{(k+1)} - x\|_V \leq c_2 \|x^{(k)} - x\|_V. \tag{3.187}$$

Por outro lado, da desigualdade triangular, temos

$$\|x^{(k)} - x\|_V \leq \|x^{(k)} - x^{(k+1)}\|_V + \|x^{(k+1)} - x\|_V \tag{3.188}$$

Das desigualdades 3.188 e 3.187 resulta

$$\|x^{(k)} - x\|_V \leq \|x^{(k)} - x^{(k+1)}\|_V + c_2\|x^{(k)} - x\|_V, \tag{3.189}$$

de onde

$$\|x^{(k)} - x\|_V (1 - c_2) \leq \|x^{(k)} - x^{(k+1)}\|_V. \tag{3.190}$$

Uma vez que $c_2 < 1$, por construção , da desigualdade 3.190 obtém-se

$$\|e^{(k)}\|_V = \|x^{(k)} - x\|_V \leq \frac{\|x^{(k)} - x^{(k+1)}\|_V}{1 - c_2}. \tag{3.191}$$

De 3.191, utilizando 3.187, obtém-se também

$$\|e^{(k+1)}\|_V = \|x^{(k+1)} - x\|_V \leq \frac{c_2}{1 - c_2}\|x^{(k)} - x^{(k+1)}\|_V. \tag{3.192}$$

A desigualdade 3.192 permite-nos majorar o erro de uma dada iterada, bastando para isso conhecer a diferença entre as duas últimas iteradas e o valor de c_2.

k	$x_1^{(k)}$	$x_2^{(k)}$	$x_3^{(k)}$	$\|x^{(k+1)} - x^{(k)}\|_2$	$r^{(k)}$
1	0.6	0.75	0.9	0.15	
2	0.625	0.85	0.875	0.106066	0.7071064
3	0.575	0.875	0.925	0.07500	0.7071066
4	0.5625	0.825	0.9375	0.05303	0.7071069
5	0.5875	0.8125	0.9125	0.03750	0.7071068
6	0.59375	0.8375	0.90625	0.02652	0.7071083
7	0.58125	0.84375	0.91875	0.01875	0.7071075
8	0.578125	0.83125	0.921875	0.01326	0.7071061
9	0.584375	0.828125	0.915625	0.00938	0.7071068

Table 1: Resultados do método de Jacobi para o exemplo 3.12

k	$x_1^{(k)}$	$x_2^{(k)}$	$x_3^{(k)}$	$\|x^{(k+1)} - x^{(k)}\|_2$	$r^{(k)}$
1	0.6	0.8	0.9	0.141421	
2	0.6	0.85	0.925	0.055902	0.3952846
3	0.575	0.825	0.9125	0.037500	0.6708187
4	0.5875	0.8375	0.91875	0.018750	0.5
5	0.58125	0.83125	0.915625	0.009375	0.5

Table 2: Resultados do método de Gauss-Seidel para o exemplo 3.12

Exemplo 3.12 . Regressando, mais uma vez, ao sistema linear que foi analisado no exemplo 3.10, vamos efectuar uma análise do erro, para os métodos de Jacobi e de Gauss-Seidel. Com este fim, foi aplicado cada um destes métodos ao sistema considerado. Partindo da aproximação inicial $x^{(0)} = (0.5, 0.8, 1.0)$, foram efectuadas iterações até que fosse satisfeita a condição

$$\|x^{(k)} - x^{(k+1)}\|_2 \leq 0.01.$$

Em cada iteração foi avaliada a norma $\|x^{(k)} - x^{(k+1)}\|_2$ e, a partir da $2^{\underline{a}}$ iteração , a razão $r^{(k)}$ correspondente. Os resultados obtidos no caso do método de Jacobi são dados na tabela 3.1, enquanto os resultados obtidos para o método de Gauss-Seidel se encontram na tabela 3.2. Nestas tabelas verifica-se nitidamente que os valores de $r^{(k)}$ tendem para $c_1 = 0.7071$, no caso do método de Jacobi, e para $c_1 = 0.5$, no caso do método de Gauss-Seidel. Estes valores coincidem com os raios espectrais das matrizes C_J e C_{gs}, respectivamente (ver exemplos 3.10 e 3.11). Com base nestes valores, podemos obter estimativas do erro para cada um dos métodos.

Para o método de Jacobi, de acordo om a fórmula 3.191, considerando $c_2 = 0.70711$, temos

$$\|e^{(9)}\|_2 \leq \frac{c_2}{1 - c_2}\|x^{(9)} - x^{(8)}\|_2 \leq 0.0242.$$

No caso do método de Gauss-Seidel, tomando $c_2 = 0.5$, temos

$$\|e^{(5)}\|_2 \leq \frac{c_2}{1 - c_2}\|x^{(5)} - x^{(4)}\|_2 \leq 0.01.$$

No exemplo que acabámos de ver, constatámos que o método de Gauss-Seidel convergia mais rapidamente que o de Jacobi, o que resulta de o raio espectral da matriz C_{gs} ser inferior ao da matriz C_J. Vamos ver que este caso não é uma excepção , no que diz respeito à relação entre os dois métodos.

A fim de compararmos o método de Gauss-Seidel com o de Jacobi, quanto à rapidez de convergência, consideremos o caso em que a matriz A do sistema tem a diagonal estritamente dominante por linhas. Neste caso, de acordo com os teoremas 3.8 e 3.10, ambos os métodos convergem para a solução exacta, qualquer que seja a aproximação inicial escolhida. Além disso, para o método de Jacobi é válida a estimativa do erro

$$\|e^{(k)}\|_\infty \leq \mu^k \|e^{(0)}\|_\infty, \; k = 1, 2, \ldots \tag{3.193}$$

onde $\mu = \|C_J\|_\infty$. Recordando a forma da matriz C_J, dada por 3.143, e as definições das grandezas α_i e β_i, dadas por 3.174, podemos concluir que

$$\mu = \max_{i=1,\ldots,n} (\alpha_i + \beta_i). \tag{3.194}$$

Por outro lado, para o método de Gauss-Seidel, segundo o teorema 3.10, é válida a estimativa do erro

$$\|e^{(k)}\|_\infty \leq \eta^k \|e^{(0)}\|_\infty, \; k = 1, 2, \ldots \tag{3.195}$$

Para estabelecer uma relação entre a rapidez de convergência dos dois métodos, basta-nos portanto comparar o parâmetro μ da fórmula 3.193 com o parâmetro η da fórmula 3.195. Atendendo à definição de μ segundo a fórmula 3.194, temos

$$(\alpha_i + \beta_i) - \frac{\beta_i}{1 - \alpha_i} = \frac{\alpha_i(1 - \alpha_i - \beta_i)}{1 - \alpha_i} \geq \frac{\alpha_i(1 - \mu)}{1 - \alpha_i} \geq 0,$$

de onde resulta imediatamente

$$\mu \geq \eta, \tag{3.196}$$

quando a matriz A tem a diagonal estritamente dominante por linhas. Isto explica que na maioria dos exemplos práticos, em que entram tais matrizes, a convergência do método de Gauss-Seidel seja mais rápida que a do método de Jacobi.

Exemplo 3.13. Consideremos o sistema $A\,x = b$, onde A é uma matriz de ordem n com a forma geral

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

Neste caso, verifica-se imediatamente que a matriz A tem a diagonal estritamente dominante, por linhas, pelo que tanto o método de Gauss-Seidel como o de Jacobi convergem, qualquer que seja a aproximação inicial. Além disso, atendendo às fórmulas 3.174, temos

$$\alpha_i = \beta_i = 2/5, \;\; i = 1, 2, \dots, n.$$

Daqui, pelas fórmulas 3.194 e 3.175, obtém-se imediatamente

$$\mu = 4/5, \;\; \eta = 2/3.$$

Assim, neste exemplo verifica-se a desigualdade 3.196. Por conseguinte, é de esperar que, no caso deste sistema, o método de Gauss-Seidel convirja mais rapidamente que o de Jacobi.

Note-se, porém, que esta comparação entre os dois métodos só é válida para matrizes com a diagoal estritamente dominante por linhas. No caso geral, nem sempre o método de Gauss-Seidel é mais rápido que o de Jacobi, havendo mesmo casos particulares em que o segundo é convergente e o primeiro não .

Um aspecto importante a salientar é que os métodos iterativos para sistemas lineares, quando convergem para qualquer aproximação inicial, são *estáveis* (ver Definição 3.7). Ou seja, partindo de dois vectores iniciais próximos, ξ_0 e η_0 , obtêm-se sempre duas sucessões $\{x^{(n)}\}$ e $\{y^{(n)}\}$

igualmente próximas, convergindo para o mesmo vector x (solução exacta). Esta propriedade é de grande importância prática, uma vez que no cálculo numérico são inevitáveis os erros se arredondamento, os quais, como já vimos, podem propagar-se ao longo de uma sucessão de operações , conduzindo a erros muito grandes no resultado final. Esta situação verifica-se, por exemplo, na resolução de sistemas lineares por métodos directos, mesmo que eles sejam bem condicionados. Os métodos iterativos, desde que sejam aplicados a sistemas bem condicionados, são sempre estáveis, ou seja, quando se usam estes métodos não há perigo de os erros de arredondamento cometidos nos cálculos poderem resultar em erros significativos no resultado final. Isto representa, portanto, uma importante vantagem dos métodos iterativos sobre os directos, sobretudo quando se trata de resolver sistemas de grandes dimensões .

References

[1] K.E. Atkinson, *An Introduction to Numerical Analysis*, John Wiley & sons, New York, 1978.

[2] M. Carpentier,*Análise Numérica*, Secção de Folhas, IST, 1993.

[3] M. Graça, P. Lima, *Matemática Experimental*, ISTPress, 2007.

[4] P.Lima, *Métodos Numéricos da Álgebra*, disponível em www.math.ist.utl.pt/∼plima/LMAC/mna.pdf .