Métodos Numéricos da Álgebra

Pedro Trindade Lima

Março de 1997

Índice

1 Erros no Cálculo Numérico **3**

2 Métodos Numéricos da Álgebra Linear **8**

2.1 Complementos de Álgebra Linear 10
2.1.1 Valores Próprios e Formas Canónicas de Matrizes . . . 10
2.1.2 Normas e Convergência em Espaços Vectoriais 20
2.1.3 Normas Matriciais 26
2.1.4 Problemas 36
2.2 Condicionamento de Sistemas Lineares 38
2.3 Métodos Directos 45
2.3.1 Método da Eliminação de Gauss 45
2.3.2 Métodos de Factorização 54
2.3.3 Problemas 68
2.4 Métodos Iterativos para Sistemas Lineares 70
2.4.1 Noções Básicas sobre Métodos Iterativos 70
2.4.2 Método de Jacobi 76
2.4.3 Método de Gauss-Seidel 83
2.4.4 Rapidez de Convergência dos Métodos Iterativos. Análise do Erro 88
2.4.5 Optimização de Métodos Iterativos 95
2.4.6 Problemas 106
2.5 Métodos Numéricos para o Cálculo de Valores e Vectores Próprios 108
2.5.1 Localização de Valores Próprios 108
2.5.2 Condicionamento do Problema de Valores Próprios . . 113
2.5.3 Método das Potências 117
2.5.4 Aceleração de Convergência 123

3 Equações e Sistemas Não-Lineares **126**
3.1 Método do Ponto Fixo 128

Capítulo 1

Erros no Cálculo Numérico

Ao resolver um determinado problema matemático por um certo método, chegamos a resultados que, de um modo geral, não constituem a *solução exacta* do problema, mas apenas *valores aproximados* desta. Daí que um dos problemas fundamentais da Análise Numérica consiste em avaliar o erro dos resultados obtidos por um determinado método, ou seja, determinar o valor máximo que esse erro pode atingir, considerando os dados do problema e os meios utilizados para o resolver.

Antes de mais, convém fixar o significado de alguns termos que vamos utilizar frequentemente e as respectivas notações. Assim, se x for o valor exacto de uma certa grandeza e $\tilde{x}$ for um valor aproximado de x , chamaremos *erro absoluto de* $\tilde{x}$ e representaremos por e_x a diferença entre os dois valores, isto é,

$$e_x = x - \tilde{x}. \tag{1.1}$$

Note-se, no entanto, que o erro absoluto é um conceito que tem pouco significado quando se trata de avaliar a precisão de um valor aproximado. Por exemplo, se dissermos que o erro absoluto de um certo resultado não excede uma milésima , este resultado tanto pode ser considerado muito preciso (se estivermos a tratar com grandezas da ordem dos milhões), como pode não ter precisão nenhuma (se o resultado exacto for uma grandeza de ordem inferior às milésimas). Logo, a precisão dos resultados numéricos terá de ser avaliada de outro modo, por exemplo, através do *erro relativo*, que é dado pela razão entre o erro absoluto e o valor da grandeza considerada. Ou seja, se representarmos o erro relativo por δx, teremos

$$\delta x = \frac{e_x}{x}. \tag{1.2}$$

Uma vez que o erro relativo nunca pode ser superior à unidade (isso significaria que o erro da grandeza é superior a ela própria, em termos absolutos), é frequente representar-se este erro sob a forma de percentagem.

Agora que já fixámos algumas notações a propósito de erros, vamos analisar as razões que podem estar na origem do erro dum resultado numérico. Esta análise é fundamental, uma vez que , se não conhecermos a origem do erro, também não poderemos, em geral, determinar até que ponto esse erro afecta o resultado obtido nem saberemos o que fazer para atenuar o seu efeito.

Em primeiro lugar, quando se resolve numericamente um determinado problema prático, há que formulá-lo em termos matemáticos. Isto significa que determinadas leis que regem os fenómenos naturais são descritas através de equações que relacionam diversas grandezas. Ao fazê-lo, torna-se necessário, frequentemente, ignorar certos aspectos da realidade, cujo efeito se considera desprezável. Por exemplo, ao estudar o movimento de um corpo em queda livre, a pequena distância da superfície da terra, em certas condições, pode ser ignorada a força de resistência do ar e o movimento de rotação da Terra. Este processo conduz a simplificações significativas no modelo matemático que facilitam a obtenção de resultados numéricos. Além disso, as equações dos modelos matemáticos contêm frequentemente constantes (no exemplo acima referido, a constante g da gravidade) cujo valor é determinado experimentalmente, apenas com uma certa precisão. Os próprios dados do problema (por exemplo, a posição e a velocidade inicial do corpo) são obtidos, geralmente, por medições e, como tal, também estão sujeitos a erros. Todos estes factores de erro estão relacionados com *inexactidão na formulação matemática do problema* e os erros deles resultantes são conhecidos geralmente como *erros inerentes*. Sendo os erros inerentes inevitáveis (eles são inerentes à própria formulação matemática do problema), é importante saber até que ponto eles podem influir no resultado final, para podermos ajuizar se a formulação matemática do problema escolhida está razoavelmente de acordo com a realidade. Caso contrário, torna-se necessário procurar outro modelo matemático.

O segundo tipo de erros que vamos considerar está directamente relacionado com o método escolhido para resolver o problema matemático e

chegar a um resultado numérico. Na maioria dos casos, não são conhecidas fórmulas que permitam obter a solução exacta do problema matemático com um número finito de operações aritméticas. Por exemplo, as soluções de uma simples equação algébrica de grau superior a 4 não podem, de um modo geral, ser expressas através de uma fórmula resolvente. Noutros casos, embora existam fórmulas que exprimem os resultados exactos, a aplicação directa destas fórmulas exige um tal volume de cálculos que se torna, de facto, impraticável. Estes factos levam-nos, em geral, a abdicar de encontrar a *solução exacta* do problema e a procurar, em vez dela , uma *solução aproximada*, obtida, por exemplo, por um método de aproximações sucessivas. A diferença entre a solução exacta do problema matemático e a solução aproximada, obtida pelo método utilizado, é designada *erro do método*. Ao longo do nosso programa, iremos estudar métodos para resolver diversos problemas matemáticos e, para cada método, trataremos de analisar o erro a ele associado. Nisto consiste precisamente um dos objectivos centrais da Análise Numérica.

Finalmente, existe ainda um terceiro factor muito importante que está na origem dos erros dos resultados numéricos: a impossibilidade prática de representar com exactidão qualquer número real. Sabemos que a representação de números através de uma sucessão de dígitos (seja no sistema decimal ou noutro sistema qualquer) só permite representar exactamente um conjunto finito de números. Os números irracionais, por exemplo, não têm representação exacta em nenhum sistema. Logo, quando precisamos de efectuar um certo cálculo com números reais, temos, em geral, que os "arredondar", isto é , substituir cada número por outro próximo, que tem representação exacta no sistema de numeração escolhido. Os erros resultantes destas aproximações chamam-se *erros de arredondamento*. Na prática, para chegarmos à solução de um problema, temos de efectuar uma sucessão (por vezes, muito longa) de operações aritméticas, onde os erros de arredondamento podem ocorrer a cada passo. O efeito total dos erros de arredondamento constitui aquilo a que se chama o *erro computacional*. Logo, o erro computacional é a diferença entre o resultado que se obteria se fosse possível efectuar os cálculos sem cometer erros de arredondamento e o valor que se obtém na prática, ao efectuar os cálculos num certo sistema de representação numérica. É evidente, portanto, que o erro computacional depende essencialmente dos meios de cálculo utilizados, em particular, do sistema de representação de números. Este erro pode ser reduzido se aumentarmos o número de dígitos utilizados

para representar os números, diminuindo assim os erros de arredondamento. No entanto, convém lembrar que o erro computacional depende também da ordem pela qual os cálculos são efectuados (ou, por outras palavras, do *algoritmo* utilizado), pois é isso que determina a forma como os erros se vão propagar de passo para passo. Uma vez que o mesmo método numérico pode ser aplicado através de algoritmos diferentes, é essencial escolher aqueles algoritmos (ditos *estáveis*) que não permitem a acumulação dos erros, de modo a que o erro computacional não adquira um peso excessivo no resultado final.

No exemplo que se segue, tentaremos distinguir os diferentes tipos de erros cometidos na resolução de um problema concreto.

Exemplo. Suponhamos que, ao observar um ponto material em movimento, se dispõe de aparelhos que permitem determinar, em cada instante, a sua posição e a sua velocidade. Representemos por x, y, z as suas coordenadas e por v_x, v_y e v_z, as componentes do seu vector velocidade ao longo de cada eixo. O nosso objectivo é determinar, a cada instante, o espaço percorrido pelo ponto material, desde o início da contagem do tempo.

Comecemos por formular matematicamente o problema. Nas condições dadas, o espaço $s(t)$, percorrido até ao instante t, pode ser expresso através de um integral de linha, mais precisamente, do integral do módulo da velocidade ao longo do percurso :

$$s(t) = \int_{t_0}^{t} |\mathbf{v}(\xi)| d\xi, \tag{1.3}$$

onde

$$|\mathbf{v}(\xi)| = \sqrt{v_x(\xi)^2 + v_y(\xi)^2 + v_z(\xi)^2} \tag{1.4}$$

e t_0 representa o momento inicial.

Uma vez que, na prática, o integral presente na equação 1.3 não pode ser calculado exactamente, vamos optar por aproximar aquele integral pelo seguinte método. Admitindo que as medições podem ser efectuadas em sucessivos instantes $t_1, t_2, \ldots, t_n$, com pequenos intervalos entre si, vamos proceder como se a velocidade se mantivesse constante dentro de cada intervalo de tempo $[t_i, t_{i+1}]$. Então, o espaço percorrido durante esse intervalo de tempo é considerado igual a $|\mathbf{v}(t_i)|(t_{i+1} - t_i)$ e, quando $t = t_n$,o integral do segundo

membro da equação 1.3 pode ser aproximado pela seguinte soma :

$$s(t_n) \approx s_n = \sum_{i=1}^{n} |\mathbf{v}(t_i)|(t_{i+1} - t_i). \tag{1.5}$$

.

Vejamos como, neste caso, os diferentes tipos de erros afectariam o resultado final. Em primeiro lugar, mesmo que conseguíssemos calcular o valor exacto $s(t)$ do integral da equação 1.3, esse valor podia não ser exactamente o espaço percorrido pelo ponto, uma vez que, para o obtermos, teríamos de nos basear nos valores da velocidade , v_x, v_y e v_z, obtidos experimentalmente e, logo, sujeitos a erros. Assim, a diferença entre o espaço de facto percorrido pelo ponto material até ao instante t, e o valor $s(t)$, dado pela equação 1.3 , pode ser considerado como o *erro inerente* da solução deste problema.

Em segundo lugar, temos o erro cometido ao aproximar o integral da equação 1.3 pela soma da equação 1.5. A esse erro, ou seja, à diferença $s(t_n) - s_n$, chamaremos o *erro do método*.

Finalmente, não esqueçamos que, ao calcular a soma da equação 1.5 , vamos inevitavelmente cometer *erros de arredondamento*, quer ao efectuar as adições, quer ao extrair as raizes quadradas, pelo que obteríamos, de facto, não o valor numérico de s_n, mas um valor aproximado, que representaremos por $\tilde{s}_n$. A diferença $s_n - \tilde{s}_n$ constitui, neste caso, o *erro computacional*.

Capítulo 2

Métodos Numéricos da Álgebra Linear

Entre os problemas de Álgebra Linear que mais frequentemente se encontram nas aplicações da Matemática, encontram-se os sistemas de equações lineares e os problemas de valores e vectores próprios.

Os sistemas de equações lineares surgem nas mais diversas situações e as suas dimensões podem variar desde as unidades até aos milhões . Como alguns exemplos típicos de problemas que conduzem a sistemas de equações lineares, poderemos citar as aproximações pelo método dos mínimos quadrados; o método dos coeficientes indeterminados (utilizado no cálculo de integrais, na resolução de equações diferenciais, etc.); e a resolução numérica de problemas de valores de fronteira para equações diferenciais.

Não é de estranhar, portanto, a grande diversidade de métodos que ao longo da história da Matemática foram e continuam a ser criados para a resolução de sistemas lineares. A escolha do método a utilizar para resolver um determinado sistema depende, essencialmente, da sua dimensão e das suas propriedades, nomeadamente, do seu *condicionamento*, propriedade fundamental de que falaremos no parágrafo 2.2. Por exemplo, a regra de Cramer, um dos primeiros métodos conhecidos para a resolução deste problema, funciona bem para pequenos sistemas (3–4 equações) mas torna-se praticamente impossível de utilizar para sistemas de maior dimensão . (Basta lembrar que o cálculo de um determinante de ordem n envolve $n!(n-1)$ multiplicações). Já o método de Gauss, de que falaremos no parágrafo 2.3 , permite trabalhar com sistemas de maiores dimensões (da ordem das dezenas), mas continua,

como veremos, a ter grandes limitações . Os métodos iterativos, de que falaremos no parágrafo 2.4, fornecem, em muitos casos, uma boa alternativa para o cálculo de uma solução aproximada dum sistema linear.

No que diz respeito aos problemas de valores e vectores próprios, eles também surgem em muitas aplicações , em particular, nos modelos matemáticos da teoria das oscilações , da mecânica quântica e da hidrodinâmica. Os métodos numéricos para a resolução destes problemas têm algo em comum com os métodos para sistemas lineares, mas também têm aspectos específicos importantes, que os transformam num objecto de estudo independente. Também neste campo a diversidade dos métodos é muito grande e a escolha depende essencialmente de dois factores: a dimensão da matriz e o número de valores próprios que se pretende calcular. Falaremos sobre este assunto no parágafo 2.5 .

Hoje em dia, uma importante direcção da investigação na análise numérica consiste na elaboração e estudo de métodos eficientes que permitam resolver sistemas lineares gigantes , cujas dimensões podem atingir a ordem dos milhões . Este objectivo torna-se cada vez mais actual, à medida que cresce a necessidade de resolver com grande precisão as equações dos modelos matemáticos das ciências aplicadas. Problemas matemáticos que até há poucos anos eram impossíveis de resolver, devido ao volume de cálculos que envolvem, ficam hoje ao alcance dos investigadores, graças à utilização dos mais modernos meio de cálculo, em particular, dos computadores com processamento paralelo e vectorial. No entanto, não podemos esquecer que estes meios só poderão ser utilizados de forma eficiente se os métodos de cálculo evoluirem de modo a poderem tirar partido de todas as suas potencialidades. Daí que a criação de novos métodos numéricos da álgebra linear, a par do aperfeiçoamento dos métodos já conhecidos, seja um tema de grande actualidade.

2.1 Complementos de Álgebra Linear

Antes de entrarmos no estudo dos métodos numéricos para a resolução de sistemas lineares e de problemas de valores próprios, começaremos por rever alguns conceitos e teoremas de Álgebra Linear, de que necessitaremos para a análise que vamos efectuar.

2.1.1 Valores Próprios e Formas Canónicas de Matrizes

Seja A uma matriz quadrada de ordem n (real ou complexa). Representaremos por A^T a *transposta de* A, ou seja, a matriz cujas linhas são as colunas de A. Por conseguinte,

$$a_{ij}^T = a_{ji}.$$

Usaremos a notação A^* para representar a *transposta da conjugada de* A. Logo,

$$a_{ij}^* = \overline{a_{ji}}.$$

Definição 2.1.1. A matriz A diz-se *hermitiana* se for verdadeira a igualdade $A = A^*$.

No caso particular de A ser real, então dizer que A é hermitiana equivale a dizer que A é *simétrica*, isto é, que $A = A^T$.

Definição 2.1.2. Diz-se que a matriz A é *unitária* se se verificar

$$A^*A = AA^* = I,$$

onde I representa a matriz identidade.

Deste modo, se a matriz A for unitária, então $A^* = A^{-1}$, onde A^{-1} denota, como habitualmente, a inversa de A. As matrizes reais unitárias, para as quais $A^T = A^{-1}$, também se costumam chamar *ortogonais*.

Definição 2.1.3. O número $\lambda \in \mathbb{C}$ diz-se *valor próprio* da matriz A se existir um vector $x \in \mathbb{C}^n$, não nulo, tal que

$$Ax = \lambda x. \tag{2.1.1}$$

Nesse caso, diz-se que x é um *vector próprio de* A associado a λ.

Como é sabido do curso de Álgebra (ver, por exemplo [8]) a existência de soluções não nulas da equação 2.1.1 implica que

$$det(A - \lambda I) = 0. \tag{2.1.2}$$

A equação 2.1.2 é conhecida como a *equação característica* da matriz A. Desenvolvendo o determinante do primeiro membro da equação 2.1.2 por menores, verifica-se facilmente que este determinante constitui um polinómio de grau n, em relação a λ. Este polinómio é chamado o *polinómio característico* de A. Visto que os valores próprios de A são, por definição , as raizes do polinómio característico de A, podemos concluir imediatamente que *qualquer matriz A tem n valores próprios, entre reais e complexos (contando com as respectivas multiplicidades).* Naturalmente, um valor próprio diz-se *múltiplo* se corresponder a uma raiz *múltipla* do polinómio característico. No entanto, no caso dos valores próprios, é importante distinguir entre a *multiplicidade algébrica* e a *multiplicidade geométrica.*

Definição 2.1.4. Chama-se *multiplicidade algébrica* do valor próprio λ à sua multiplicidade, enquanto raiz do polinómio característico de A. A *multiplicidade geométrica* de λ é o número máximo de vectores próprios independentes, associados a λ.

Nota. Como resulta da definição 2.1.3, os vectores próprios associados a um dado valor próprio λ formam um espaço linear, chamado o *espaço próprio associado a λ*. A dimensão deste espaço não pode ser superior à multiplicidade de λ, como raiz do polinómio característico. Por outras palavras, *a multiplicidade geométrica de um valor próprio não pode ser superior à sua multiplicidade algébrica.* No entanto, as duas podem não ser iguais, como se pode verificar no exemplo que se segue.

Exemplo 2.1.5. Seja A a seguinte matriz

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

Neste caso, o polinómio característico de A tem a forma

$$det(A - \lambda I) = (1 - \lambda)^2 (2 - \lambda)^2. \tag{2.1.3}$$

De acordo com a definição 2.1.3 , os valores próprios de A são :

$$\lambda_{1,2} = 1, \ \lambda_{3,4} = 2, \tag{2.1.4}$$

ou seja, a matriz A tem dois valores próprios distintos, cada um deles com multiplicidade algébrica igual a 2. Se resolvermos o sistema linear

$$(A - 2I)\, x = 0, \tag{2.1.5}$$

verificamos que a sua solução geral pode ser escrita sob a forma

$$x = C_1 (0, 0, 1, 0) + C_2(0, 0, 0, 1), \tag{2.1.6}$$

onde C_1 e C_2 são constantes arbitrárias. Daqui podemos concluir que a matriz A tem, no máximo, dois vectores próprios independentes associados ao valor próprio $\lambda_{3,4} = 2$, o que significa que *a multiplicidade geométrica deste valor próprio é 2* (igual à multiplicidade algébrica). No caso do valor próprio $\lambda_{1,2} = 1$, verifica-se que a solução geral do sistema linear

$$(A - I)\, x = 0 \tag{2.1.7}$$

tem a forma

$$x = C\, (1, 0, 0, 0), \tag{2.1.8}$$

onde C é uma constante arbitrária, ou seja, não existe mais do que um vector próprio independente, associado ao valor próprio $\lambda_{1,2} = 1$. Neste caso, *a multiplicidade geométrica do valor próprio* $\lambda_{1,2} = 1$ *é igual a 1*, inferior portanto à sua multiplicidade algébrica.

Definição 2.1.6 . Sejam A e B duas matrizes quadradas da mesma ordem. Então A diz-se *semelhante* a B se existir uma matriz P, não singular, tal que

$$B = P^{-1}\,A\,P. \tag{2.1.9}$$

Note-se que, se A é semelhante a B, então da equação 2.1.9 resulta que

$$A = Q^{-1}\,B\,Q, \tag{2.1.10}$$

onde $Q = P^{-1}$, o que significa que B *é semelhante a* A. Trata-se, portanto, de uma relação simétrica, como seria de esperar. É útil referirmos aqui algumas propriedades das matrizes semelhantes.

1. Se A e B forem semelhantes, então têm a mesma equação característica.

   Demonstração . Sejam A e B duas matrizes semelhantes. Então existe P tal que

   $$\begin{aligned} det\,(B - \lambda\,I) &= det\,(P^{-1}\,A\,P - \lambda\,I) = \\ &= det\,(P^{-1}\,(A - \lambda\,I)\,P) = \\ &= det\,(P^{-1})\,det\,(A - \lambda\,I)\,det\,(P). \end{aligned} \tag{2.1.11}$$

   Uma vez que $det\,(P^{-1})\,det\,(P) = det\,(I) = 1$, de 2.1.11 resulta que A e B têm o mesmo polinómio característico e, por conseguinte, a mesma equação característica □.

2. Duas matrizes semelhantes A e B têm os mesmos valores próprios. Se x for um vector próprio de A, associado ao valor próprio λ, então $z = P^{-1}x$ é um vector próprio de B, associado também a λ.

   Demonstração . A primeira parte da afirmação é consequência imediata da propriedade 1. Para verificar a veracidade da segunda parte da afirmação , notemos que , se λ e x satisfazem a igualdade

   $$A\,x = \lambda\,x, \tag{2.1.12}$$

   então também se verifica

   $$P^{-1}\,A\,P\,(P^{-1}\,x) = \lambda\,P^{-1}\,x. \tag{2.1.13}$$

Se definirmos o vector z como $z = P^{-1}x$, e atendendo a que $B = P^{-1}AP$, da equação 2.1.13 resulta

$$B\,z = \lambda\,z, \tag{2.1.14}$$

o que significa que z é um vector próprio da matriz B , também associado ao valor próprio λ $\square$.

3. A propriedade 1 significa que o polinómio característico se mantém invariante, no caso de transformações de semelhança. Uma vez que o *traço* de uma matriz (ou seja, a soma dos elementos da sua diagonal principal) é igual ao coeficiente do termo em λ^{n-1} do polinómio característico, podemos concluir que *duas matrizes semelhantes têm o mesmo traço* . Do mesmo modo se pode provar que *duas matrizes semelhantes têm o mesmo determinante.*

Nos próximos parágrafos vamos utilizar algumas vezes a redução das matrizes a certas formas, ditas *canónicas.* Isto significa que vamos, através de transformações de semelhança, obter matrizes mais simples, que conservam as propriedades fundamentais das matrizes iniciais. A forma canónica a utilizar depende das propriedades da matriz inicial e dos objectivos que se pretende alcançar. Algumas das formas mais frequentemente utilizadas são a *forma normal de Schur*,a *forma diagonal* e a *forma de Jordan* , de que falaremos a seguir.

Teorema 2.1.7 (forma normal de Schur). Seja A uma matriz complexa de ordem n. Então existe uma matriz unitária U que reduz A à forma

$$T = U^*\,A\,U, \tag{2.1.15}$$

onde T é *triangular superior.*

Demonstração . Este teorema demonstra-se por indução em n. No caso de $n = 1$, a afirmação do teorema é trivial, com $U = [1]$. Admitamos agora que a afirmação do teorema é válida para qualquer matriz de ordem $n \leq k - 1$. Vmos mostrar que então ela também é válida para matrizes de ordem $n = k$.

Seja λ_1 um valor próprio de A, e $u^{(1)}$ um vector próprio, associado a λ_1, tal que $\|u^{(1)}\|_2 = 1$. Começando com $u^{(1)}$, constroi-se uma base ortonormal

em $\mathbb{C}^k$, que representaremos por $\{u^{(1)}, u^{(2)}, \ldots, u^{(k)}\}$. Seja P_1 uma matriz de ordem k, cujas colunas são constituídas pelos vectores desta base:

$$P_1 = [u^{(1)}, u^{(2)}, \ldots, u^{(k)}].$$

Por construção , as colunas de P_1 são ortogonais entre si. Logo, $P_1^* P_1 = I$, ou seja, $P_1^{-1} = P_1^*$. Definindo a matriz B_1 como

$$B_1 = P_1^* \, A \, P_1. \tag{2.1.16}$$

Vamos provar que B_1 tem a forma

$$B_1 = \begin{bmatrix} \lambda_1 & \alpha_2 & \ldots & \alpha_k \\ 0 & & & \\ 0 & & A_2 & \\ 0 & & & \end{bmatrix}, \tag{2.1.17}$$

onde A_2 é uma submatriz de ordem $k-1$ e $\alpha_2, \ldots, \alpha_k$ são certos números. Para provarmos isso, comecemos por verificar que

$$\begin{aligned} AP_1 &= A[u^{(1)}, u^{(2)}, \ldots, u^{(k)}] = \\ &= [Au^{(1)}, Au^{(2)}, \ldots, Au^{(k)}] = \\ &= [\lambda_1 u^{(1)}, v^{(2)}, \ldots, v^{(k)}], \end{aligned} \tag{2.1.18}$$

onde $v^{(j)} = Au^{(j)}, j = 2, \ldots, k$.Multiplicando ambos os membros de 2.1.18, à esquerda, por P_1^*, obtém-se

$$P_1^* \, A \, P_1 = B_1 = [\lambda_1 \, P_1^* \, u^{(1)}, \ldots, P_1^* \, v^{(k)}]. \tag{2.1.19}$$

Notemos ainda que $P_1^* u^{(1)} = e^{(1)} = (1, 0, \ldots, 0)$. Logo, de 2.1.19 resulta que

$$B_1 = [\lambda_1 \, e^{(1)}, w^{(2)}, \ldots, w^{(k)}], \tag{2.1.20}$$

onde $w^{(j)} = P_1^* \, v^{(j)}, j = 2, \ldots, k$, ficando assim provada a igualdade 2.1.17.

De acordo com a hipótese da indução , existe uma matriz unitária P_2, de ordem $k-1$, tal que reduz A_2 à forma

$$\hat{T} = \hat{P_2}^* \, A_2 \, \hat{P_2}, \tag{2.1.21}$$

onde $\hat{T}$ é triangular superior, de ordem $k-1$. Seja

$$P_2 = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & & & \\ \dots & & \hat{P}_2 & \\ 0 & & & \end{bmatrix}. \tag{2.1.22}$$

Então P_2 é unitária e verifica-se

$$P_2^*\, B_1\, P_2 = \begin{bmatrix} \lambda_1 & \gamma_2 & \dots & \gamma_k \\ 0 & & & \\ \dots & & \hat{P}_2{}^*\, A_2\, \hat{P}_2 & \\ 0 & & & \end{bmatrix} = \begin{bmatrix} \lambda_1 & \gamma_2 & \dots & \gamma_k \\ 0 & & & \\ \dots & & \hat{T} & \\ 0 & & & \end{bmatrix} = T, \tag{2.1.23}$$

onde T é triangular superior. Das igualdades 2.1.23 e 2.1.16 resulta

$$T = P_2^*\, B_1\, P_2 = P_2^*\, P_1^*\, A\, P_1\, P_2 = U^*\, A\, U, \tag{2.1.24}$$

onde $U = P_1\, P_2$. Resta constatar que U é unitária, (como produto de duas matrizes unitárias), ficando assim o teorema 2.1.7 demonstrado por indução $\square$.

Nota. Uma vez que a matriz T, que nos dá a forma normal de Schur, é triangular superior, o seu polinómio característico é

$$det(T - \lambda I) = (t_{11} - \lambda) \dots (t_{nn} - \lambda), \tag{2.1.25}$$

de onde se conclui que os valores próprios de T são os elementos da sua diagonal principal. Por conseguinte, e uma vez que as transformações de semelhança não alteram os valores próprios (como vimos acima), concluímos que os valores próprios de A são os elementos diagonais de T, isto é, $\lambda_i = t_{ii}$, i=1,...,k.

A redução de uma matriz hermitiana à forma diagonal, de que vamos falar a seguir, pode ser considerada como um caso particular da forma normal de Schur, com grande significado prático.

Teorema 2.1.8. Seja A uma matriz *hermitiana* de ordem n. Então A *tem n valores próprios reais* $\lambda_1, \lambda_2, \dots, \lambda_n$, não necessariamente diferentes

entre si, e *n vectores próprios associados* $u^{(1)}, u^{(2)}, \ldots, u^{(n)}$, que formam uma base ortonormal em $\mathbb{C}^n$. Além disso, existe uma matriz unitária U tal que

$$U^* A U = D, \tag{2.1.26}$$

onde D é uma matriz diagonal cujas entradas não nulas são os valores próprios de A :

$$D = diag(\lambda_1, \lambda_2, \ldots, \lambda_n) \tag{2.1.27}$$

Demonstração . Pelo teorema 2.1.7 , existe uma matriz unitária U, tal que

$$U^* A U = T, \tag{2.1.28}$$

onde T é triangular superior. Atendendo a que A é hermitiana, de 2.1.28 resulta que

$$T^* = U^* A^* U = T, \tag{2.1.29}$$

o que significa que T também é, neste caso, hermitiana. Se T é triangular superior e hermitiana, então T é diagonal. Além disso, de 2.1.29 resulta que $\overline{t_{ii}} = t_{ii}, i = 1, 2, \ldots, n$, o que significa que todos os elementos diagonais de T são reais. Logo, pelo teorema 2.1.7, concluímos que *todos os valores próprios de A são reais.* Sendo assim, temos

$$T = diag(\lambda_1, \lambda_2, \ldots, \lambda_n), \tag{2.1.30}$$

onde todos os λ_i são reais. Para demonstar a segunda parte do teorema, representemos a matriz unitária U sob a forma

$$U = [u^{(1)}, \ldots, u^{(n)}], \tag{2.1.31}$$

onde os vectores $u^{(i)}, i = 1, \ldots, n$ formam uma base ortonormal em $\mathbb{C}^n$. Falta mostrar que estes vectores são vectores próprios de A. Da igualdade $T = U^* A U$ resulta $UT = AU$. Logo

$$A\,[u^{(1)}, \ldots, u^{(n)}] = [u^{(1)}, \ldots, u^{(n)}] \begin{bmatrix} \lambda_1 & 0 & \ldots & 0 \\ 0 & \lambda_2 & \ldots & 0 \\ 0 & \ldots & \ldots & \ldots \\ 0 & \ldots & \ldots & \lambda_n \end{bmatrix}. \tag{2.1.32}$$

A igualdade 2.1.32 pode desdobrar-se em n igualdades:

$$A\,u^{(j)} = \lambda_j\,u^{(j)}, j = 1, \ldots, n \tag{2.1.33}$$

Estas igualdades significam que os vectores $u^{(j)}$ são os vectores próprios da matriz A, tal como se pretendia demonstrar $\square$.

Nota. Nas condições do teorema 2.1.8 , pode-se mostrar que, no caso particular de a matriz A ser *real e simétrica*, ela tem n vectores próprios *reais*, que podem ser escolhidos de modo a formarem *uma base ortonormal em* $\mathbf{R^n}$.

O teorema anterior diz-nos que as matrizes hermitianas, que constituem uma classe com grandes aplicações práticas, podem ser reduzidas a uma forma canónica extremamente simples: a forma diagonal. Em relação às matrizes não hermitianas, tudo o que sabemos, por enquanto, é que podem ser reduzidas à forma normal de Schur (triangular superior). Este resultado, como vamos ver, pode ser melhorado, através da *forma canónica de Jordan*, que está mais próxima da forma diagonal.

Definição 2.1.9. Chama-se *bloco de Jordan de dimensão n* a uma matriz da forma

$$J_n(\lambda) = \begin{bmatrix} \lambda & 1 & 0 & \dots & 0 \\ 0 & \lambda & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & 0 & \lambda & 1 \\ 0 & \dots & 0 & 0 & \lambda \end{bmatrix}. \tag{2.1.34}$$

Note-se que *qualquer bloco de Jordan de dimensão n tem um único valor próprio* λ*, de multiplicidade algébrica igual a n e multiplicidade geométrica igual a 1* . Ou seja, tem *um único vector próprio independente, associado ao valor próprio* λ. Em particular, um bloco de Jordan de dimensão 1 é constituído apenas pelo número λ.

Definição 2.1.10 Diz-se que uma matriz J se encontra na *forma canónica*

de jordan se ela tiver a seguinte estrutura

$$J = \begin{bmatrix} J_{n_1}(\lambda_1) & 0 & \dots & 0 \\ 0 & J_{n_2}(\lambda_2) & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & J_{n_k}(\lambda_k) \end{bmatrix}, \tag{2.1.35}$$

onde $J_{n_i}(\lambda_i)$ representa um bloco de Jordan de dimensão n_i, com valor próprio λ_i. Note-se que os valores próprios λ_i dos vários blocos não são necessariamente distintos. Por exemplo, a matriz A, considerada no **exemplo 2.1.5**, encontra-se na forma canónica de Jordan, pois a sua estrutura é a seguinte:

$$A = \begin{bmatrix} J_2(1) & 0 & 0 \\ 0 & J_1(2) & 0 \\ 0 & 0 & J_1(2) \end{bmatrix}, \tag{2.1.36}$$

ou seja, ela decompõe-se num bloco de Jordan, de dimensão 2, correspondente ao valor próprio $\lambda_{1,2} = 1$, e dois blocos de Jordan, de dimensão 1, correspondentes ao valor próprio $\lambda_{3,4} = 2$.

Uma *matriz diagonal* também se encontra na forma de Jordan, pois *é constituída por n blocos de Jordan de dimensão 1* . Sobre a redução das matrizes à forma de Jordan existe o seguinte teorema.

Teorema 2.1.11. Para qualquer matriz quadrada A (real ou complexa), de dimensão n, existe uma matriz não singular P tal que

$$P^{-1} A P = J, \tag{2.1.37}$$

onde J é uma matriz na forma canónica de Jordan. Os elementos λ_i, da diagonal principal de J, são os valores próprios de A e o número de blocos de Jordan em que se decompõe a matriz J é igual ao número de vectores próprios independentes de A. Finalmente, o número de vezes que cada valor próprio λ_i aparece na diagonal principal de J é igual à sua multiplicidade algébrica.

A demonstração do teorema 2.1.11 faz parte do curso de Álgebra linear, pelo que remetemos o leitor para qualquer manual desse curso, por exemplo, [8].

2.1.2 Normas e Convergência em Espaços Vectoriais

Em Análise Numérica, necessitamos frequentemente de medir os erros de grandezas aproximadas, o que torna indispensável o uso de conceitos como o de *norma*. Quando o resultado que estamos a considerar é simplesmente um número, a norma introduz-se de forma muito simples, pois é igual ao *módulo* (em $\mathbf{R}$ ou em $\mathbb{C}$, conforme o caso). No caso mais geral de um espaço vectorial a n dimensões ($\mathbf{R^n}$ ou $\mathbb{C}^\mathbf{n}$) são possíveis diferentes generalizações do conceito de norma, que nós vamos rever neste parágrafo.

Definição 2.1.12 . Seja E um espaço vectorial e N uma função de E em $\mathbf{R}$. Então N diz-se uma norma se verificar as seguintes propriedades:

1. $N(x) \geq 0, \forall x \in V;\ N(x) = 0,\ sse\, x = 0.$
2. $N(\alpha x) = |\alpha|\, N(x),\ \forall x \in E,\ \forall \alpha \in \mathbf{R}$.
3. $N(x+y) \leq N(x) + N(y),\ \forall x, y \in E$ (desigualdade triangular).

Utiliza-se frequentemente a notação $\|x\|_N = N(x)$, onde o índice N junto do sinal de norma serve para especificar a norma considerada (este índice pode ser omitido, quando não for necessário fazer essa especificação).

Na prática, utilizam-se muitos tipos diferentes de normas, de acordo com as grandezas a que se aplicam e os objectivos a alcançar. Seguem-se alguns dos exemplos mais frequentes, no caso do espaço considerado ser $\mathbf{R^n}$ ou $\mathbb{C}^\mathbf{n}$.

Exemplo 2.1.13. *Norma euclidiana.* A norma euclidiana, que traduz o conceito habitual de comprimento de um vector, tem a seguinte expressão

$$\|x\|_2 = \sqrt{\sum_{i=1}^{n} |x_i|^2} \tag{2.1.38}$$

Exemplo 2.1.14. *Norma do máximo.* A norma do máximo, também frequentemente utilizada, define-se do seguinte modo:

$$\|x\|_\infty = \max_{1\leq i\leq n} |x_i|. \tag{2.1.39}$$

Exemplo 2.1.15. *p-normas.* Mais um tipo de normas frequentemente utilizadas é dado pela expressão

$$\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p\right)^{1/p}, \tag{2.1.40}$$

onde $p > 1$.

Nota. A norma euclidiana é um caso particular de uma p-norma, com $p = 2$. A norma do máximo pode ser obtida como um caso limite das p-normas, quando $p \to \infty$, o que explica a notação utilizada.

As propriedades 1 e 2 verificam-se imediatamente para qualquer uma das normas aqui apresentadas. Quanto à propriedade 3, ela verifica-se imediatamente para a norma do máximo, mas requer algum esforço no caso das p-normas (ver problema 2 de §2.1.4).

Um espaço vectorial no qual está definida uma norma diz-se um **espaço normado**.

Num espaço normado é possível introduzir o conceito de *conjunto limitado*.

Definição 2.1.16. Seja E um espaço normado, a um elemento de E e r um número real positivo. Chama-se *bola de raio* r com centro em a, segundo a norma N, e representa-se por $B_a^N(r)$ o seguinte subconjunto de E :

$$B_a^N(r) = \{x \in E : \|x - a\|_N \leq r\}. \tag{2.1.41}$$

Definição 2.1.17. Um conjunto X, contido num espaço normado E, diz-se *limitado*, sse existir um número $r > 0$, tal que X está contido numa bola de raio r com centro no elemento nulo de E.

Note-se que o conceito de conjunto limitado está ligado a uma norma concreta, pelo que, de um modo geral, um conjunto pode ser limitado numa certa norma, e não o ser noutra. No entanto, no caso de espaços de dimensão finita ($\mathbf{R^n}$ ou $\mathbf{\mathbb{C}^n}$), se um conjunto for limitado segundo uma certa norma, também o é segundo outra norma qualquer. Os dois lemas que se seguem têm como objectivo a verificação deste facto.

Lema 2.1.18. Seja E um espaço de dimensão finita, munido com a *norma do máximo*, e seja $B_0^\infty(1)$ uma bola segundo esta norma. Então , qualquer que seja a norma $N(x)$ considerada, existe um número $r > 0$, tal que

$$N(x) \leq r, \ \forall\, x \ \in \ B_0^\infty(1). \tag{2.1.42}$$

Demonstração . Seja $e^{(1)}, \ldots, e^{(n)}$ uma base em E. Então x pode ser representado na forma

$$x = \sum_{i=1}^{n} x_i \ e^{(i)}. \tag{2.1.43}$$

Utilizando as propriedades das normas, podemos escrever

$$N(x) \leq \sum_{i=1}^{n} |x_i| \ N(e^{(i)}) \ \leq \ \max_{1\leq i\leq n} |x_i| \ \sum_{i=1}^{n} N(e^{(i)}). \tag{2.1.44}$$

Uma vez que $x \in B_0^\infty(1)$, sabemos que $\max_{1\leq i\leq n} |x_i| \leq 1$, pelo que a desigualdade 2.1.42 se verifica com $r = \sum_{i=1}^{n} N(e^{(i)})$, ficando assim o lema 2.1.18 demonstrado.□

Sendo dada uma norma no espaço vectorial E,é possível definir também o conceito de continuidade para as funções com domínio em E.

Definição 2.1.19. Seja f uma função de E em $\mathbf{R}$ e seja X um subconjunto de E, onde E é um espaço normado com a norma N. Então f diz-se *contínua em* X se e só se for satisfeita a condição :

$$\forall \epsilon > 0 \, \exists \delta > 0 : N(x_1 - x_2) < \delta \Rightarrow |f(x_1) - f(x_2)| < \epsilon, \ \forall x_1, x_2 \in X. \tag{2.1.45}$$

Lema 2.1.20. Seja $N(x)$ uma norma no espaço vectorial E. Então $N(x)$ *é uma função contínua de* E *em* $\mathbf{R}$.

Demonstração . Para verificar que a condição 2.1.45 é satisfeita pela função $N(x)$ em todo o espaço E, consideremos dois elementos arbitrários de E, x_1 e x_2, tais que $N(x_1 - x_2) < \epsilon$. Então , da desigualdade triangular resulta (ver problema 1 de §2.1.4):

$$|N(x_1) - N(x_2)| \leq N(x_1 - x_2) < \epsilon. \tag{2.1.46}$$

Por conseguinte, a condição 2.1.45 verifica-se em todo o espaço E , com $\delta = \epsilon$, ficando assim demonstrado o lema 2.1.20.□

Podemos agora demonstrar o seguinte teorema.

Teorema 2.1.21. Sejam N e M duas normas num espaço E de dimensão finita. Então existem constantes positivas C_1, C_2 , tais que

$$C_1\, M(x) \;\leq\; N(x) \;\leq\; C_2\, M(x),\ \forall x \in E. \tag{2.1.47}$$

Demonstração . Basta considerar o caso em que N é uma norma arbitrária e M é a norma do máximo. Consideremos em E uma esfera S de raio 1 (segundo a norma do máximo) :

$$S = \{x \in E : \|x\|_\infty = 1\}.$$

Uma vez que S é a fronteira de uma bola de raio 1 pela norma do máximo, de acordo com o lema 2.1.18, S é limitado segundo qualquer norma N. Por outro lado, S é um conjunto fechado. Logo, uma vez que E é um espaço de dimensão finita, S é um compacto. De acordo com o lema 2.1.20, a norma N define uma função contínua em E. Logo, segundo o teorema de Weierstrass, N tem um máximo e um mínimo em S, pelo que podemos escrever

$$C_1 \;\leq\; N(x) \;\leq\; C_2,\ \forall x \in S. \tag{2.1.48}$$

Além disso, podemos afirmar que $C_1 > 0$, uma vez que S não contém o elemento nulo de E. Notemos ainda que, se y for um elemento arbitrário de

E, existe um elemento $x \in S$ e um número positivo α , tais que $y = \alpha\, x$ e $\|y\|_\infty = \alpha$. Então , da desigualdade 2.1.48 resulta que

$$C_1 \|y\|_\infty \leq N(y) \leq C_2 \|y\|_\infty \forall y \in E. \tag{2.1.49}$$

Fica assim demonstrado o teorema 2.1.21. □

Diz-se que duas normas N e M num espaço E são equivalentes se existirem constantes positivas C_1 e C_2, tais que

$$C_1\, M(x) \leq N(x) \leq C_2\, M(x),\ \forall x \in E.$$

Assim, o teorema 2.1.21 significa que todas as normas num espaço de dimensão finita são equivalentes.

Exemplo 2.1.22. Determinemos as constantes C_1 e C_2 da desigualdade 2.1.47 no caso em que se compara a norma euclidiana com a norma do máximo em $\mathbf{R^n}$. Temos, por definição ,

$$\|x\|_2 = \sqrt{\sum_{i=1}^{n} |x_i|^2}.$$

Seja $x \in S$, onde S é uma esfera de raio 1 segundo a norma do máximo. Neste caso, verifica-se facilmente que

$$\max_{x \in S} \|x\|_2 = \|(1, 1, \ldots, 1)\|_2 = \sqrt{n}.$$

Por conseguinte, a segunda parte da desigualdade 2.1.47 verifica-se com $C_2 = \sqrt{n}$. Por outro lado, também é evidente que

$$\min_{x \in S} \|x\|_2 = \|(1, 0, \ldots, 0)\|_2 = 1.$$

Por conseguinte, a primeira parte da desigualdade 2.1.47 também é satisfeita, com $C_1 = 1$.□

Finalmente, vamos introduzir o conceito de convergência em espaços normados, extremamente importante para a Análise Numérica.

Definição 2.1.23. Seja $\left\{x^{(n)}\right\}_{n=1}^{\infty}$ uma sucessão de elementos de um espaço normado E. Diz-se que esta sucessão *converge para um certo elemento*

x *de* E , segundo a norma N (simbolicamente, escreve-se $x^{(n)} \xrightarrow{N} x$) se e só se

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

Do teorema 2.1.21 resulta que, se uma dada sucessão $\left\{x^{(n)}\right\}_{n=1}^{\infty}$ convergir para um certo x em E (espaço de dimensão finita) segundo uma certa norma N, então ela converge também para x segundo qualquer outra norma M. Na realidade, de acordo com a definição 2.1.23, se $x^{(n)} \xrightarrow{N} x$, então $\|x^{(n)} - x\|_N \to 0$. Mas, pelo teorema 2.1.21, qualquer que seja a norma M, existe uma constante C_2, tal que $M(y) \leq C_2\, N(y)$, $\forall y \in E$. Logo, podemos afirmar que

$$\lim_{n\to\infty} \|x^{(n)} - x\|_M \leq C_2 \lim_{n\to\infty} \|x^{(n)} - x\|_N = 0,$$

o que significa que a sucessão considerada também converge para x segundo a norma M. Esta observação permite-nos, enquanto estivermos a tratar de espaços de dimensão finita, falar de convergência sem nos referirmos a nenhuma norma em especial. Por isso, diremos simplesmente que *a sucessão* $\left\{x^{(n)}\right\}_{n=1}^{\infty}$ *converge (ou tende) para* x e escreveremos simbolicamente $x^{(n)} \to x$.

2.1.3 Normas Matriciais

Na análise de métodos numéricos que vamos efectuar a seguir, teremos de introduzir normas em espaços de matrizes. Sabemos que as matrizes quadradas de ordem n formam um espaço vectorial de dimensão n^2, topologicamente equivalente, portanto, ao espaço $\mathbb{C}^{n^2}$ ou $\mathbf{R}^{n^2}$, conforme se trate de matrizes de entradas reais ou complexas.

Se encararmos o problema deste ponto de vista, qualquer das normas que introduzimos em espaços vectoriais também serve como norma matricial. No entanto, nem todas as normas possíveis têm interesse prático. Vamos, portanto, começar por definir algumas condições suplementares que as normas matriciais devem satisfazer para nos permitirem alcançar os nossos objectivos. Visto que nos espaços de matrizes está definido o produto, operação que não existe em todos os espaços vectoriais, é importante que a norma matricial esteja, num certo sentido, de acordo com essa operação .

Definição 2.1.24. Seja M uma norma no espaço das matrizes quadradas de ordem n (que representaremos, daqui em diante, por L^n). A norma M diz-se *regular* se e só se for satisfeita a condição

$$\|A\,B\|_M \;\leq\; \|A\|_M\,\|B\|_M,\; \forall A, B \in L^n. \tag{2.1.50}$$

As matrizes quadradas são geralmente utilizadas para representar aplicações lineares do espaço $\mathbf{R^n}$ ou $\mathbb{C}^{\mathbf{n}}$ em si próprio. Por isso, é natural exigir que exista uma certa relação entre a norma matricial e a norma no espaço vectorial onde é definida a aplicação .

Definição 2.1.25. Seja M uma norma no espaço L^n e V uma norma no espaço vectorial E^n (que pode ser $\mathbf{R^n}$ ou $\mathbb{C}^{\mathbf{n}}$, conforme se trate de matrizes de entradas reais ou complexas). Diz-se que a norma M é *compatível* com a norma V se e só se for verdadeira a relação

$$\|A\,x\|_V \;\leq\; \|A\|_M\,\|x\|_V,\; \forall A \in L^n,\; \forall x \in E^n. \tag{2.1.51}$$

Todas as normas matriciais que vamos utilizar são regulares e compatíveis com as normas vectoriais conhecidas. Isto torna mais fácil obter estimativas de erro e resultados relativos a convergência.

Exemplo 2.1.26. Consideremos no espaço L^n a norma F, definida do seguinte modo:

$$F(A) = \left[\sum_{i,j=1}^{n} |a_{ij}|^2\right]^{1/2} \tag{2.1.52}$$

Esta norma é conhecida como a *norma de Frobenius*.

Da sua definição resulta que a norma de Frobenius é análoga à norma euclidiana, definida para os espaços vectoriais e, por conseguinte, satisfaz todas as condições da definição 2.1.12. Além disso, pode-se demonstrar que esta norma é *regular*. Sabendo que $\|x\|_2 \leq \|x\|_1$, para qualquer vector x,e utilizando a desigualdade de Cauchy-Schwarz, temos

$$\begin{aligned} F(A\,B) &= \left[\sum_{i,j=1}^{n} |\sum_{k=1}^{n} a_{ik}\,b_{kj}|^2\right]^{1/2} \leq \\ &\leq \left[\sum_{i,j=1}^{n}\sum_{k=1}^{n} |a_{ik}|^2 \sum_{k=1}^{n} |b_{kj}|^2\right]^{1/2} = \\ &= \left[\sum_{i,k=1}^{n} |a_{ik}|^2\right]^{1/2} \left[\sum_{k,j=1}^{n} |b_{kj}|^2\right]^{1/2} = \\ &= F(A)\,F(B) \end{aligned} \tag{2.1.53}$$

Podemos ainda provar que a norma de Frobenius é *compatível* com a norma euclidiana para vectores. Para isso, consideremos um vector $x \in E^n$ e uma matriz $A \in L^n$. Então , usando de novo a desigualdade de Cauchy-Schwarz, temos

$$\begin{aligned} \|A\,x\|_2 &= \left[\sum_{i=1}^{n} \left|\sum_{j=1}^{n} a_{ij}\,x_j\right|^2\right]^{1/2} \leq \\ &\leq \sum_{i=1}^{n} \left[\sum_{j=1}^{n} |a_{ij}|^2\right]^{1/2} \left[\sum_{j=1}^{n} |x_j|^2\right]^{1/2} = \\ &= F(A)\,\|x\|_2, \end{aligned} \tag{2.1.54}$$

tal como pretendíamos demonstrar.

Uma forma natural de introduzir num espaço de matrizes uma norma que satisfaz todas as condições impostas consiste em *induzir* neste espaço uma

norma matricial, *associada* à norma vectorial considerada. Vamos explicar a seguir o significado exacto desta expressão .

Definição 2.1.27. Diz-se que uma norma matricial M em L^n está *associada* a uma certa norma vectorial V em E^n se e só se ela for definida pela igualdade

$$\|A\|_M = \sup_{x \in E^n, x \neq 0} \frac{\|A\,x\|_V}{\|x\|_V} \tag{2.1.55}$$

Também se diz, neste caso, que a norma M é a norma *induzida* em L^n pela norma vectorial V.

Vamos provar que a igualdade 2.1.55 define, de facto, uma norma matricial, *regular* e *compatível com a norma* V .

Para começar, mostremos que o segundo membro de 2.1.55 é uma grandeza finita. Em primeiro lugar, note-se que a igualdade 2.1.55 pode ser escrita na forma

$$\|A\|_M = \sup_{x \in B_0^M(1)} \|A\,x\|_V, \tag{2.1.56}$$

onde $B_0^M(1)$ representa uma bola de raio unitário com centro na origem. Assim, sendo $B_0^M(1)$ um conjunto compacto, e uma vez que $\|A\,x\|_V$ é uma função contínua em S, pelo teorema de Weierstrass ela tem um máximo (finito) nesse conjunto.

Verifica-se facilmente que o segundo membro de 2.1.55 satisfaz as condições 1,2 e 3 da definição de norma. Além disso, da definição 2.1.27 resulta imediatamente que uma norma matricial associada a uma certa norma vectorial é compatível com essa norma. Com efeito, da igualdade 2.1.55 obtém-se

$$\|A\,x\|_V \;\leq\; \|A\|_M\,\|x\|_V \tag{2.1.57}$$

Por último, verifiquemos que uma norma associada é sempre regular. Sejam A e B duas matrizes de L^n e x, um vector de E^n. Então , de acordo com a desigualdade 2.1.57

$$\|(A\,B\,)x\|_V \;=\; \|A\,(B\,x)\|_V \;\leq\; \|A\|_M\,\|B\,x\|_V \;\leq\; \|A\|_M\,\|B\|_M\,\|x\|_V. \tag{2.1.58}$$

Dividindo ambos os membros de 2.1.58 por $\|x\|_V$, obtém-se

$$\frac{\|(A\,B\,)x\|_V}{\|x\|_V} \;\leq\; \|A\|_M\,\|B\|_M,\ \forall x \neq 0, \tag{2.1.59}$$

de onde resulta que $\|A\,B\|_M \leq \|A\|_M\,\|B\|_M$. Uma nova definição de norma associada, equivalente a 2.1.55, pode ser dada pela igualdade

$$\|A\|_M = sup_{x\in E^n,\,\|x\|_V\leq 1}\|A\,x\|_V. \tag{2.1.60}$$

Vamos agora introduzir as normas matriciais que se utilizam mais frequentemente, induzindo-as a partir das normas vectoriais, estudadas no parágrafo anterior. Como acabamos de demonstrar, qualquer norma induzida desse modo é, por construção , regular e compatível com a norma vectorial a que está associada.

1. consideremos a norma vectorial definida por

$$\|x\|_1 = \sum_{i=1}^{n}|x_j|. \tag{2.1.61}$$

Tentemos determinar a forma da norma matricial associada a 2.1.61. Temos

$$\begin{aligned}
\|A\,x\|_1 &= \sum_{i=1}^{n}\left|\sum_{j=1}^{n}a_{ij}\,x_j\right| \leq \\
&\leq \sum_{i=1}^{n}\sum_{j=1}^{n}(|a_{ij}|\,|x_j|) = \\
&= \sum_{j=1}^{n}\left(|x_j|\sum_{i=1}^{n}|a_{ij}|\right) \leq \\
&\leq \|x\|_1 \max_{j=1,\ldots,n}\sum_{i=1}^{n}|a_{ij}|.
\end{aligned} \tag{2.1.62}$$

Representemos por c a grandeza

$$c = \max_{j=1,\ldots,n}\sum_{i=1}^{n}|a_{ij}|. \tag{2.1.63}$$

Queremos provar que a grandeza c é a norma matricial associada à norma vectorial 2.1.61, ou seja, que

$$c = \sup_{x\in E^n, x\neq 0}\frac{\|A\,x\|_1}{\|x\|_1}. \tag{2.1.64}$$

Mas, de 2.1.62 obtém-se

$$\frac{\|A\,x\|_1}{\|x\|_1} \leq c,\ \forall x \in E^n. \tag{2.1.65}$$

Para mostrar que a igualdade 2.1.64 é verdadeira, resta provar que existe, pelo menos, um vector $x \in E^n$, para o qual a inequação 2.1.62 se converte em igualdade. Seja k o índice da coluna para a qual é atingido o máximo no segundo membro da igualdade 2.1.63. Então , se na desigualdade 2.1.62 substituirmos $x = e^{(k)} = (0, \ldots, 1, 0, \ldots, 0)$, obtém-se

$$\frac{\|A\,x\|_1}{\|x\|_1} = \sum_{i=1}^{n} |a_{ik}| = c. \tag{2.1.66}$$

Fica assim demonstrado que a grandeza c, definida pela igualdade 2.1.64, é uma norma matricial em L^n, associada à norma vectorial 2.1.61. Atendendo à fórmula de cálculo desta norma, ela é conhecida como *norma por coluna* e identificada através do índice 1. Assim, podemos escrever:

$$\|A\|_1 = \sup_{x \in E^n, x \neq 0} \frac{\|A\,x\|_1}{\|x\|_1}. \tag{2.1.67}$$

2. Consideremos agora no espaço vectorial E^n a norma do máximo:

$$\|x\|_\infty = \max_{1 \leq i \leq n} |x_i|. \tag{2.1.68}$$

Para qualquer matriz $A \in L^n$ temos

$$\|A\,x\|_\infty = \max_{1 \leq i \leq n} \left| \sum_{j=1}^{n} a_{ij}\,x_j \right|. \tag{2.1.69}$$

De 2.1.69 obtém-se

$$\|A\,x\|_\infty \leq \max_{1 \leq j \leq n} |x_j| \max_{1 \leq i \leq n} \sum_{j=1}^{n} |a_{ij}|. \tag{2.1.70}$$

Representemos por c a grandeza

$$c = \max_{1 \leq i \leq n} \sum_{j=1}^{n} |a_{ij}|. \tag{2.1.71}$$

Então a desigualdade 2.1.70 adquire a forma

$$\|A\,x\|_\infty \leq c\|x\|_\infty, \forall x \in E^n, \forall A \in L^n. \tag{2.1.72}$$

Verifiquemos agora que, para um certo vector $x \in E^n$, a inequação 2.1.72 se converte em igualdade. Seja k o índice da linha para a qual é atingido o máximo no segundo membro de 2.1.71. Consideremos um vector com a seguinte forma

$$x = (sign(a_{k1}), sign(a_{k2}), \ldots, sign(a_{kn})). \tag{2.1.73}$$

Introduzindo o vector $b = A\,x$ e entrando em conta com 2.1.71 e 2.1.73, obtém-se

$$\|A\,x\|_\infty = \max_{1\leq i\leq n} |b_i| = b_k = \sum_{j=1}^{n} |a_{kj}| = c, \tag{2.1.74}$$

tal como pretendíamos demonstrar. Fica assim demonstrado que a igualdade 2.1.71 define a norma matricial, associada à norma do máximo. Esta norma é geralmente conhecida como a *norma por linha* e assinalada com o índice ∞. Assim, podemos escrever

$$\|A\|_\infty = \max_{1\leq i\leq n} \sum_{j=1}^{n} |a_{ij}|. \tag{2.1.75}$$

3. Vamos agora definir a norma matricial, associada à norma euclidiana. Para isso, começaremos por definir raio espectral.

Definição 2.1.28. Seja $A \in L^n$ e sejam $\lambda_1, \lambda_2, \ldots, \lambda_n$ os valores próprios de A. Chama-se *raio espectral de* A e representa-se por $r_\sigma(A)$ o seguinte valor:

$$r_\sigma(A) = \max_{1\leq i\leq n} |\lambda_i|. \tag{2.1.76}$$

Proposição 2.1.29. A norma matricial, associada à norma euclidiana de vectores, tem a forma

$$\|A\|_2 = \sqrt{r_\sigma(A^*A)}. \tag{2.1.77}$$

Demonstração . Usaremos a notação $\langle u, v\rangle$ para representar o produto interno de dois vectores $u, v \in E^n$. Note-se que, qualquer que seja a matriz $A \in L^n$, a matriz $B = A^*A$ é *hermitiana* e, de acordo com o teorema 2.1.8, tem n valores próprios reais. Pode-se mostrar também que todos estes valores são não negativos (ver problema 9 de §2.1.4). Numerando-os por ordem decrescente, podemos escrever

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n \geq 0. \tag{2.1.78}$$

Além disso, também segundo o teorema 2.1.8, a matriz $B = A^*A$ tem n vectores próprios independentes $u^{(1)}, u^{(2)}, \ldots, u^{(n)}$, que podem ser escolhidos de modo a formar uma base ortonormal em E^n. Suponhamos que estes vectores estão numerados de tal modo que $u^{(i)}$ é um vector próprio associado a $\lambda_i, i = 1, \ldots, n$. Podemos, pois, decompor um vector arbitrário x na seguinte forma:

$$x = \sum_{j=1}^{n} \alpha_j \, u^{(j)} \tag{2.1.79}$$

e, utilizando esta decomposição , podemos escrever

$$A^*A\, x = \sum_{j=1}^{n} \alpha_j \, A^*A \, u^{(j)} \; = \; \sum_{j=1}^{n} \alpha_j \, \lambda_j \, u^{(j)}. \tag{2.1.80}$$

Por outro lado, com base na representação da norma euclidiana através do produto interno, temos

$$\|A\, x\|_2^2 \; = \; \langle A\, x, A\, x\rangle \; = \langle x, A^*A\, x\rangle. \tag{2.1.81}$$

Utilizando a decomposição 2.1.80 no segundo membro de 2.1.81 e com base na ortonormalidade dos vectores próprios de A^*A, podemos escrever:

$$\begin{aligned} \|A\, x\|_2^2 \;\; &= \;\; \langle \sum_{j=1}^{n} \alpha_j u^{(j)} \, , \sum_{i=1}^{n} \alpha_i \, \lambda_i \, u^{(i)} \rangle \; = \\ &= \;\; \sum_{i=1}^{n} \; |\alpha_i|^2 \, \lambda_i \; \leq \\ &\leq \;\; \lambda_1 \sum_{i=1}^{n} \; |\alpha_i|^2 \; = \\ &= \;\; \lambda_1 \, \|x\|_2^2. \end{aligned} \tag{2.1.82}$$

Extraindo a raiz quadrada de ambos os membros da igualdade 2.1.82, e notando que $\lambda_1 = r_\sigma(A^*A)$,temos

$$\|A\,x\|_2 \leq \sqrt{r_\sigma(A^*A)}\,\|x\|_2, \forall x \in E^n. \tag{2.1.83}$$

Para acabarmos de provar a igualdade 2.1.77, basta indicar um vector x, para o qual a inequação 2.1.83 se converta numa igualdade. Seja $x = u^{(1)}$. Então , por construção, $A\,x = \lambda_1\,x$. Logo

$$\|A\,x\|_2^2 = \lambda_1\langle x, x\rangle = \lambda_1\,\|x\|_2^2 = r_\sigma(A^*A)\|x\|_2^2. \tag{2.1.84}$$

Extraindo a raiz quadrada de ambos os membros de 2.1.84 obtém-se a igualdade pretendida. Fica assim demonstrada a proposição 2.1.29. □.

Para terminar este parágrafo, vamos estabelecer relações entre as normas e o raio espectral das matrizes.

Teorema 2.1.30. Seja $A \in L^n$. Então , qualquer que seja a norma matricial M, associada a uma certa norma vectorial em E^n, verifica-se

$$r_\sigma(A) \leq \|A\|_M. \tag{2.1.85}$$

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

$$\|A\,x\|_V = \|\lambda\,x\|_V = |\lambda|\,\|x\|_V. \tag{2.1.86}$$

Então podemos afirmar que

$$\|A\|_M = \sup_{x\in E^n, x\neq 0} \frac{\|A\,x\|_V}{\|x\|_V} \geq |\lambda|, \tag{2.1.87}$$

de onde resulta a afirmação do teorema. □

Teorema 2.1.31 . Para qualquer $\epsilon > 0$ e qualquer matriz $A \in L^n$, existe uma certa norma $N(\epsilon)$, associada a uma norma vectorial em E^n, tal que

$$\|A\|_{N(\epsilon)} \leq r_\sigma(A) + \epsilon. \tag{2.1.88}$$

Demonstração . De acordo com o teorema 2.1.11, para qualquer matriz A existe uma matriz não singular P, tal que $P^{-1}\,A\,P\,=\,J$, onde J é a forma canónica de Jordan de A . Seja D uma matriz diagonal com a forma

$$D = diag(1, \epsilon, \epsilon^2, \ldots, \epsilon^{n-1}). \tag{2.1.89}$$

Logo, existe uma matriz $\hat{J}$, semelhante a J e a A, do seguinte modo:

$$\hat{J}\,=\,D^{-1}\,J\,D = Q^{-1}\,A\,Q, \tag{2.1.90}$$

onde $Q\,=\,P\,D$. Então verifica-se que todas as entradas da matriz $\hat{J}$ são iguais às da matriz J , à excepção das entradas iguais a 1, situadas na supra-diagonal de J, que passam a ser iguais a ϵ, na supra-diagonal de $\hat{J}$. Por outras palavras, $\hat{J}$ tem a forma

$$\hat{J}\,=\,\begin{pmatrix} \lambda_1 & j_{12} & 0 & \ldots & 0 \\ 0 & \lambda_2 & j_{23} & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & \ldots & \ldots & \lambda_{n-1} & j_{n-1,n} \\ 0 & \ldots & \ldots & \ldots & \lambda_n \end{pmatrix}, \tag{2.1.91}$$

onde λ_i representa um valor próprio de A e $j_{i-1,i}\,=\,0$ ou $j_{i-1,i}\,=\,\epsilon$, $i\,=\,1,\ldots,n$. Por conseguinte,

$$\|\hat{J}\|_\infty\,\leq\,r_\sigma(A)+\epsilon. \tag{2.1.92}$$

Vamos agora definir uma norma vectorial $V(\epsilon)$ do seguinte modo:

$$\|x\|_{V(\epsilon)}\,=\,\|Q^{-1}\,x\|_\infty,\,\forall x\in E^n. \tag{2.1.93}$$

Seja $N(\epsilon)$ a norma matricial, associada a $V(\epsilon)$. Usando as igualdades 2.1.93 e 2.1.90, obtém-se

$$\begin{aligned} \|A\|_{N(\epsilon)} &= \max_{\|A\,x\|_{V(\epsilon)}=1} \|A\,x\|_{V(\epsilon)}\,= \\ &= \max_{\|Q^{-1}\,x\|_\infty=1} \|Q^{-1}\,A\,x\|_\infty\,= \\ &= \max_{\|y\|_\infty=1} \|Q^{-1}\,A\,Q\,y\|_\infty\,= \\ &= \max_{\|y\|_\infty=1} \|\hat{J}\,y\|_\infty = \|\hat{J}\|_\infty\,\leq \\ &\leq r_\sigma(A)+\epsilon. \end{aligned} \tag{2.1.94}$$

Fica assim demonstrado o teorema 2.1.31. □.

Dos teoremas 2.1.30 e 2.1.31 resulta que *o raio espectral de uma matriz é o ínfimo de todas as suas normas* (associadas a normas vectoriais). No caso das matrizes hermitianas, temos $r_\sigma(A) = \|A\|_2$ (ver problema 7 de §2.1.4), o que significa que, para esta matrizes, a norma euclidiana é a mínima de todas as normas associadas. No caso geral, pode não existir nenhuma norma matricial (induzida) que coincida com o raio espectral da matriz, como se verifica no exemplo seguinte.

Exemplo 2.1.32. Consideremos a matriz

$$A = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}.$$

Verifica-se imediatamente que $r_\sigma(A) = 0$. No entanto, nenhuma norma de A pode ter este valor, já que A não é uma matriz nula. Podemos, porém, para qualquer valor de ϵ, definir a norma $N(\epsilon)$ de que fala o teorema 2.1.31. Para isso, basta considerar $\|B\|_{N(\epsilon)} = \epsilon\|B\|_1, \forall B \in L^n$. Nesse caso, teremos $\|A\|_{N(\epsilon)} = \epsilon$, o que está de acordo com o teorema.

Exemplo 2.1.33. Consideremos a matriz

$$A = \begin{pmatrix} 1 & 3 \\ 0 & 2 \end{pmatrix}.$$

A sua equação característica é $(1-\lambda)(2-\lambda) = 0$, de onde os seus valores próprios são $\lambda_1 = 1, \lambda_2 = 2$. Logo, $r_\sigma(A) = 2$. Comparemos este valor com o de algumas normas de A:

$$\begin{aligned} \|A\|_\infty &= max(4,2) = 4; \\ \|A\|_1 &= max(5,1) = 5; \\ \|A\|_2 &= \sqrt{r_\sigma(A^*A)} = \sqrt{3\sqrt{5}+7} \approx 3.7\,. \end{aligned}$$

2.1.4 Problemas

1. Seja N uma norma num espaço linear E^n. Mostre que

$$\|x - y\|_N \geq |\|x\|_N - \|y\|_N|, \forall x, y \in E^n.$$

2. Prove que a função

$$N(x) = \left(\sum_{i=1}^{n} |x_i|^2 \right)^{1/2},$$

definida no espaço linear E^n, é uma norma (a norma euclidiana).

Sugestão : Utilize a desigualdade de Cauchy-Schwarz para demonstrar a desigualdade triangular.

3. Seja M uma norma matricial, associada a uma certa norma vectorial V.

   (a) Prove que $\|I\|_M = 1$, onde I é a matriz identidade.

   (b) Mostre que, se A é invertível, então

$$\|A^{-1}\|_M \geq \frac{1}{\|A\|_M}$$

.

4. Demonstre que a norma de Frobenius não está associada a nenhuma norma vectorial.

5. Demonstre as seguintes relações entre normas:

$$\|x\|_\infty \leq \|x\|_1 \leq n \|x\|_\infty$$

6. Seja A uma matriz quadrada de ordem n . Supondo que são conhecidos os valores próprios de A, determine

   (a) os valores próprios de A^{-1} (admitindo que A é invertível).

   (b) os valores próprios de A^m, $m = 1, 2, \ldots$.

   (c) os valores próprios de $A + c\,I$, onde c é uma constante. .

7. Para qualquer matriz quadrada, prove que

   (a) $\|A\|_2 = r_\sigma(A)$, se A for hermitiana.

   (b) $\|A\,U\|_2 = \|U\,A\|_2 = \|A\|$, se U for unitária.

8. Seja A uma matriz quadrada de ordem n e representemos por F a norma de Frobenius.

   (a) Mostre que $F(AU) = F(UA) = F(A)$, qualquer que seja a matriz unitária U.

   (b) Se A for hermitiana, prove que

   $$F(A) = \sqrt{\sum_{i=1}^{n} \lambda_i^2},$$

   onde λ_i são os valores próprios de A.

   (c) Ainda no caso de A ser hermitiana, demonstre a seguinte relação entre normas:

   $$\frac{1}{\sqrt{n}}\,F(A) \leq \|A\|_2 \leq F(A)$$

9. Seja A uma matriz quadrada de ordem n e seja $B = A^*\,A$.

   (a) Prove que B é hermitiana.

   (b) Mostre que todos os valores próprios de B são não negativos.

10. Seja Q uma matriz quadrada não singular de ordem n.

    (a) Mostre que a função $V(x) = \|Q^{-1}x\|_\infty$ define uma norma no espaço vectorial E^n.

    (b) Verifique que a norma matricial M, associada à norma V da alínea anterior, tem a seguinte expressão :

    $$\|A\|_M = \|Q^{-1}\,A\,Q\|_\infty$$

2.2 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 ($b \in \mathbf{R^n}$) e a matriz ($A \in L^n$), 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:

$$\tilde{A} \;=\; A \,+\, \delta A, \tag{2.2.1}$$

onde δA representa uma matriz de norma pequena (em comparação com a de A). Analogamente, representemos por $\tilde{b}$ uma perturbação do segundo membro do sistema:

$$\tilde{b} \;=\; b \,+\, \delta b. \tag{2.2.2}$$

Se substituirmos A por $\tilde{A}$ e b por $\tilde{b}$ no sistema inicial, obteremos um novo sistema, cuja solução representaremos por $\tilde{x}$. Usando uma notação semelhante à de 2.2.1 e 2.2.2, podemos escrever

$$\tilde{x} \;=\; x \,+\, \delta x, \tag{2.2.3}$$

onde δx representa o erro da solução aproximada $\tilde{x}$. O nosso objectivo é escolhendo uma certa norma vectorial e a norma matricial associada, estimar o erro relativo de $\tilde{x}$ em função dos erros relativos de $\tilde{A}$ e $\tilde{b}$. Por outras palavras, queremos exprimir o quociente $\frac{||\delta x||}{||x||}$ em função de $\frac{||\delta A||}{||A||}$ e de $\frac{||\delta b||}{||b||}$.

Definição 2.2.1. 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 $\delta A \;=\; 0$. Nesse caso, temos

$$A\,(x \,+\, \delta x) \;=\; b \,+\, \delta b. \tag{2.2.4}$$

De 2.2.4 obtém-se

$$\delta x \;=\; A^{-1}\,\delta b. \tag{2.2.5}$$

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

$$\|\delta x\| \;\leq\; \|A^{-1}\|\,\|\delta b\|. \tag{2.2.6}$$

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

$$\|b\| \;\leq \|A\|\,\|x\|, \tag{2.2.7}$$

logo

$$\frac{1}{\|x\|} \;\leq\; \frac{\|A\|}{\|b\|}. \tag{2.2.8}$$

Multiplicando cada um dos membros de 2.2.6 pelo membro correspondente de 2.2.8, obtém-se

$$\frac{\|\delta x\|}{\|x\|} \;\leq\; \|A\|\,\|A^{-1}\|\frac{\|\delta b\|}{\|b\|}. \tag{2.2.9}$$

Obtivemos assim a estimativa que procurávamos para o erro relativo da solução, em função do erro relativo do segundo membro. A desigualdade 2.2.9 leva-nos à definição de número de condição de uma matriz.

Definição 2.2.2. Seja A uma matriz invertível. Chama-se *número de condição de* A o seguinte valor

$$cond(A) \;=\; \|A\|\,\|A^{-1}\|. \tag{2.2.10}$$

O número de condição de uma matriz depende da norma considerada. Para se especificar a norma, usa-se geralmente o índice p (igual ao da norma considerada) e escreve-se

$$cond_p(A) \;=\; \|A\|_p\,\|A^{-1}\|_p. \tag{2.2.11}$$

Como resulta da desigualdade 2.2.9, 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) de §2.1.4). 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

$$cond_*(A) \;=\; r_\sigma(A)\, r_\sigma(A^{-1}). \tag{2.2.12}$$

De acordo com o teorema 2.1.30, podemos escrever

$$cond_*(A) \;\leq\; cond_p(A), \tag{2.2.13}$$

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 próprios de A^{-1} são os inversos dos valores próprios de A (ver problema 6-b) de §2.1.4), o número de condição $cond_*(A)$ pode escrever-se sob a forma

$$cond_*(A) \;=\; \frac{\max_{\lambda_i \in \sigma(A)} |\lambda_i|}{\min_{\lambda_i \in \sigma(A)} |\lambda_i|}. \tag{2.2.14}$$

No caso de a matriz A ser hermitiana, então , como sabemos do problema 7-a) de §2.1.4, a sua norma euclidiana coincide com o raio espectral, pelo que podemos escrever:

$$cond_2(A) \;=\; cond_*(A). \tag{2.2.15}$$

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 2.2.3. Consideremos o sistema linear $A\,x \;=\; b$, onde A é uma matriz invertível. Sejam δA e δb definidos pelas igualdades 2.2.1 e 2.2.2,

respectivamente, e suponhamos que

$$\|\delta A\| \leq \frac{1}{\|A^{-1}\|}. \tag{2.2.16}$$

Então verifica-se a seguinte desigualdade:

$$\frac{\|\delta x\|}{\|x\|} \leq \frac{cond(A)}{1 - cond(A)\frac{\|\delta A\|}{\|A\|}} \left\{ \frac{\|\delta A\|}{\|A\|} + \frac{\|\delta b\|}{\|b\|} \right\}. \tag{2.2.17}$$

Observação . Note-se que a desigualdade 2.2.9, obtida anteriormente, é um caso particular de 2.2.17, que se obtém fazendo $\delta A = 0$. A demonstração do teorema 2.2.3 encontra-se, por exemplo, em [1].□

A desigualdade 2.2.17 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 2.2.4 Consideremos o sistema linear $A\,x = b$, onde

$$A = \begin{bmatrix} 10 & 7 & 8 & 7 \\ 7 & 5 & 6 & 5 \\ 8 & 6 & 10 & 9 \\ 7 & 5 & 9 & 10 \end{bmatrix}, \quad b = \begin{bmatrix} 32 \\ 23 \\ 33 \\ 31 \end{bmatrix}. \tag{2.2.18}$$

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, hermitiana 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

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

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

$$\tilde{A} = \begin{bmatrix} 10 & 7 & 8.1 & 7.2 \\ 7.08 & 5.04 & 6 & 5 \\ 8 & 5.98 & 9.89 & 9 \\ 6.99 & 5 & 9 & 9.98 \end{bmatrix}, \tag{2.2.19}$$

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 iniciail é ainda mais acentuada. Entretanto, a norma da perturbação é relativamente pequena:

$$\|\delta A\|_\infty = max(0.3, 0.12, 0.13, 0.04) = 0.3,$$

de onde resulta

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

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

$$A^{-1} = \begin{bmatrix} 25 & -41 & 10 & -6 \\ -41 & 68 & -17 & 10 \\ 10 & -17 & 5 & -3 \\ -6 & 10 & -3 & 2 \end{bmatrix}. \tag{2.2.20}$$

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 2.2.9, aplicada a este caso, diz-nos que

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

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 2.2.17, uma vez que, para a perturbação considerada, não se verifica a condição

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

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.

Problemas

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 $\mathbf{R^n}$ tais que

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

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

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

   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 $a \in \mathbf{R}$. Suponhamos que, ao resolver o sistema $Ax = b$, 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.

2.3 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.

2.3.1 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, $Ax = b$, a um sistema equivalente, $Ux = b'$, 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

$$A = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix}. \tag{2.3.1}$$

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:

$$A^{(1)} = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ 0 & a_{22}^{(1)} & \dots & a_{2n}^{(1)} \\ \dots & \dots & \dots & \dots \\ 0 & a_{n2}^{(1)} & \dots & a_{nn}^{(1)}. \end{bmatrix}. \tag{2.3.2}$$

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

$$a_{ij}^{(1)} = a_{ij} - m_{i1}\, a_{1j}, \tag{2.3.3}$$

onde

$$m_{i1} = \frac{a_{i1}}{a_{11}}. \tag{2.3.4}$$

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

$$A^{(k)} = \begin{bmatrix} a_{11} & a_{12} & \dots & \dots & \dots & a_{1n} \\ 0 & a_{22}^{(1)} & \dots & \dots & \dots & a_{2n}^{(1)} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & 0 & a_{kk}^{(k-1)} & \dots & a_{kn}^{(k-1)} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & 0 & a_{nk}^{(k)} & \dots & a_{nn}^{(k)} \end{bmatrix}. \tag{2.3.5}$$

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

$$a_{ij}^{(k)} = a_{ij}^{(k-1)} - m_{ik}\, a_{kj}^{(k-1)}, \quad i = k+1, \dots, n; \quad j = k+1, \dots, n; \tag{2.3.6}$$

onde

$$m_{ik} = \frac{a_{ik}^{(k-1)}}{a_{kk}^{(k-1)}} \tag{2.3.7}$$

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

$$A^{(n-1)} = \begin{bmatrix} a_{11} & a_{12} & \dots & \dots & \dots & a_{1n} \\ 0 & a_{22}^{(1)} & \dots & \dots & \dots & a_{2n}^{(1)} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & 0 & a_{kk}^{(k-1)} & \dots & a_{kn}^{(k-1)} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & \dots & \dots & 0 & a_{nn}^{(n-1)} \end{bmatrix}. \tag{2.3.8}$$

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. 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 2.3.8, 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:

$$b_i^{(1)} = b_i - m_{i1}\, b_1, \quad i = 2, 3, \dots, n. \tag{2.3.9}$$

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

$$b_i^{(k)} = b_i - m_{ik}\, b_k^{(k-1)}, \quad i = k+1, \dots, n. \tag{2.3.10}$$

Os coeficientes m_{ik} são dados pelas fórmulas 2.3.4 e 2.3.7.

3. Resolução do sistema triangular superior. Depois de reduzido o sistema inicial à forma triangular superior, com a matriz 2.3.8, a sua solução

obtém-se facilmente, mediante um processo de substituições sucessivas:

$$\begin{array}{rcl} x_n & = & \dfrac{b_n^{(n-1)}}{a_{nn}^{(n-1)}}; \\ x_{n-1} & = & \dfrac{b_{n-1}^{(n-2)} - a_{n-1,n}^{(n-2)}\, x_n}{a_{n-1,n-1}^{(n-2)}}; \\ \ldots \quad \ldots & & \ldots \\ x_1 & = & \dfrac{b_1 - \sum_{i=2}^{n} a_{1,i}\, x_i}{a_{1,1}}. \end{array} \tag{2.3.11}$$

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 2.3.3 e 2.3.6.

No 1º passo, a fórmula 2.3.3 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 2.3.4, 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

$$M(n) \;=\; AS(n) \;=\; \sum_{k=1}^{n-1}(n-k)^2 \;=\; \frac{n(n-1)(2n-1)}{6}, \tag{2.3.12}$$

Quanto às divisões , o número total é

$$D(n) \;=\; \sum_{k=1}^{n-1}(n-k) \;=\; \frac{n(n-1)}{2}. \tag{2.3.13}$$

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

$$TO(n) = M(n) + AS(n) + D(n) \approx \frac{2n^3}{3} + O(n^2). \tag{2.3.14}$$

2. Transformação do segundo membro. Quando transformamos o vector b, aplicamos a fórmula 2.3.10 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

$$M(n) = AS(n) = \sum_{k=1}^{n-1}(n-k) = \frac{n(n-1)}{2}, \tag{2.3.15}$$

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

$$TO(n) = M(n) + AS(n) \approx n^2. \tag{2.3.16}$$

3. Resolução do sistema triangular. Para resolver o sistema triangular, efectuamos a série de substituições 2.3.11. 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,

$$TO(n) = M(n) + AS(n) + D(n) \approx n^2. \tag{2.3.17}$$

Até agora, ao falarmos do método de Gauss, não entrámos em consideração com com os erros cometidos durante os cálculos. Em §2.2, 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 2.3.7, 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

$$c_k = \max_{k \leq i \leq n} \left| a_{ik}^{(k-1)} \right| . \tag{2.3.18}$$

Se o máximo no segundo membro de 2.3.18 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 linha i e só de-

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

pois 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

$$c_k = \max_{k \leq i,j \leq n} \left| a_{ij}^{(k-1)} \right| . \tag{2.3.19}$$

Sejam i' e j', respectivamente, os valores dos índices i e j para os quais é atingido o máximo no segundo membro de 2.3.19. 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 2.3.1. Consideremos o sistema linear $A\,x = b$, onde

$$A = \begin{bmatrix} 10^{-6} & 0 & 1 \\ 1 & 10^{-6} & 2 \\ 1 & 2 & -1 \end{bmatrix}, \quad b = \begin{bmatrix} 1 \\ 3 \\ 2 \end{bmatrix}. \tag{2.3.20}$$

Se procurarmos resolvê-lo, utilizando o método da eliminação de Gauss, chegamos ao sistema equivalente $U\,x\,=\,b'$,onde

$$U\,=\,\begin{bmatrix} 10^{-6} & 0 & 1 \\ 0 & 10^{-6} & 2-10^{6} \\ 0 & 0 & 2\times 10^{12}-5\times 10^{6}-1 \end{bmatrix},\quad b'\,=\,\begin{bmatrix} 1 \\ 3-10^{6} \\ 2\times 10^{12}-7\times 10^{6}+2 \end{bmatrix}. \tag{2.3.21}$$

Na realidade, a matriz U e o vector b' que vamos obter não são exactamente os da fórmula 2.3.21, 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:

$$\tilde{U}\,=\,\begin{bmatrix} 1.00000\times 10^{-6} & 0 & 1.00000 \\ 0 & 1.00000\times 10^{-6} & -0.999998\times 10^{6} \\ 0 & 0 & 1.99999\times 10^{12} \end{bmatrix},\quad \tilde{b}\,=\,\begin{bmatrix} 1 \\ -0.999997\times 10^{6} \\ 1.99999\times 10^{12} \end{bmatrix}. \tag{2.3.22}$$

Assim, ao resolvermos o sistema 2.3.22 por substituição ascendente, obtemos sucessivamente:

$$\begin{aligned} \tilde{x_3} &= \frac{1.99999\times 10^{12}}{1.99999\times 10^{12}} = 1.00000; \\ \tilde{x_2} &= \frac{-0.999997\times 10^{6}+0.999998\times 10^{6}\,\tilde{x_3}}{1.00000\times 10^{-6}} = 1.00000\times 10^{6}; \\ \tilde{x_1} &= \frac{1.00000-1.00000\,\tilde{x_3}}{1.00000\times 10^{-6}} = 0. \end{aligned} \tag{2.3.23}$$

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

$$A^{(1)} = \begin{bmatrix} 1 & 10^{-6} & 2 \\ 0 & -10^{-12} & 1 - 2 \times 10^{-6} \\ 0 & 2 - 10^{-6} & -3 \end{bmatrix}. \tag{2.3.24}$$

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:

$$A^{(2)} = \begin{bmatrix} 1.00000 & 1.00000 \times 10^{-6} & 2.00000 \\ 0 & 2.00000 & -3.00000 \\ 0 & 0 & 9.99998 \times 10^{-1} \end{bmatrix}, \quad b^{(2)} = \begin{bmatrix} 3.00000 \\ -1.00000 \\ 9.99997 \times 10^{-1} \end{bmatrix}. \tag{2.3.25}$$

Resolvendo o sistema 2.3.25, obtém-se a seguinte solução :

$$\begin{aligned} x_3 &= \frac{9.99997 \times 10^{-1}}{1.00000} &&= 9.99999 \times 10^{-1}; \\ x_2 &= \frac{-1.00000 + 3.00000\, x_3}{2.00000} &&= 1.00000; \\ x_1 &= 3.00000 - 2.00000\, x_3 - 1.00000 \times 10^{-6}\, x_2 &&= 9.99999 \times 10^{-1}. \end{aligned} \tag{2.3.26}$$

Esta solução é bastante diferente da que obtivemos, quando não foi utilizada a pesquisa de pivot. Se substituirmos estes valores no sistema 2.3.20, 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.

2.3.2 Métodos de Factorização

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

Definição 2.3.2.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 $Ax = b$ reduz-se a dois sistemas lineares com matrizes triangulares:

$$\begin{aligned} Lg &= b, \\ Ux &= g, \end{aligned}$$

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 :

$$a_{ij} = \sum_{k=1}^{n} l_{ik}\, u_{kj}; \quad i = 1, \ldots, n; \; j = 1, \ldots, n; \tag{2.3.27}$$

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.

Factorização de Doolittle

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

$$l_{ii} = 1; \quad i = 1, \ldots, n. \tag{2.3.28}$$

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 2.3.28, podemos escrever:

$$a_{ij} = \sum_{k=1}^{i} l_{ik}\, u_{kj} = \sum_{k=1}^{i-1} l_{ik}\, u_{kj} + u_{ij}; \quad i = 1, \ldots, n; \; j = i, \ldots, n. \tag{2.3.29}$$

Da fórmula 2.3.29 resulta imediatamente que

$$u_{ij} = a_{ij} - \sum_{k=1}^{i-1} l_{ik}\, u_{kj}. \tag{2.3.30}$$

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 2.3.29, temos

$$a_{ij} = \sum_{k=1}^{j} l_{ik}\, u_{kj} = \sum_{k=1}^{j-1} l_{ik}\, u_{kj} + l_{ij}\, u_{jj}, \quad i = 1, \ldots, n;\ j = i, \ldots, n. \quad (2.3.31)$$

Daqui resulta

$$l_{ij} = \frac{a_{ij} - \sum_{k=1}^{j-1} l_{ik}\, u_{kj}}{u_{jj}}. \quad (2.3.32)$$

Utilizando as fórmulas 2.3.30 e 2.3.32, 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,

$$det\, A = (-1)^{Nt}\, det\, L\ det\, U, \quad (2.3.33)$$

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 2.3.8. 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 2.3.7. 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:

$$L = \begin{bmatrix} 1 & 0 & \dots & 0 \\ m_{21} & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ \dots & m_{n-1,n-2} & 1 & 0 \\ \dots & m_{n,n-2} & m_{n,n-1} & 1 \end{bmatrix}. \tag{2.3.34}$$

Teorema 2.3.3. As matrizes L e U, dadas, respectivamente, pelas fórmulas 2.3.34 e 2.3.8,formam uma factorização LU da matriz A, idêntica à factorização de Doolittle.

Demonstração . Vamos demonstrar as igualdades

$$a_{ij}^{(i-1)} = u_{ij}, \quad i = 1, \dots, n; \; j = i, \dots, n; \tag{2.3.35}$$

$$m_{ij} = l_{ij}, \; j = 1, \dots, n; \; i = j, \dots, n. \tag{2.3.36}$$

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

$$a_{1j} = u_{1j}, \quad j = 1, \dots, n; \quad m_{i1} = \frac{a_{i1}}{a_{11}}, \; i = 2, \dots, n. \tag{2.3.37}$$

Vamos agora supor, por indução , que a igualdade 2.3.35 se verifica para as linhas da matriz U, com índice $k = 1, \dots, i-1$, e que a igualdade 2.3.36 se verifica para todas colunas de L, com índice $k = 1, \dots, j-1$. 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 2.3.6 do método de Gauss, temos

$$a_{ij}^{(k)} = a_{ij}^{(k-1)} - m_{ik}\, a_{kj}^{(k-1)}, \quad (k = 1, \dots, n-1), \tag{2.3.38}$$

onde se subentende que $a_{ij}^{(0)} = a_{ij}, \; i = 1, \dots, n; \; j = 1, \dots, n.$ Aplicando a

fórmula 2.3.38, sucessivamente, com $k = 1, \ldots, i-1$ obtém-se:

$$\begin{array}{lllll} a_{ij}^{(1)} & = & a_{ij} - m_{i1}\, a_{1j}; & & \\ a_{ij}^{(2)} & = & a_{ij}^{(1)} - m_{i2}\, a_{2j}^{(1)} & = & a_{ij} - m_{i1}\, a_{1j} - m_{i2}\, a_{2j}^{(1)}; \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ a_{ij}^{(i-1)} & = & a_{ij}^{(i-2)} - m_{i,i-1}\, a_{i-1,j}^{(i-2)} & = & a_{ij} - \sum_{k=1}^{i-1} m_{ik}\, a_{kj}^{(k-1)}. \end{array} \quad (2.3.39)$$

Se, de acordo com a hipótese da indução , substituirmos os coeficientes $m_{i,k}$ e $a_{k,j}^{(k-1)}$, no segundo membro de 2.3.39, por l_{ik} e u_{kj}, obtemos a fórmula 2.3.30, 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 2.3.7, têm a forma

$$m_{ij} = \frac{a_{ij}^{(j-1)}}{a_{jj}^{(j-1)}};\ i = j, \ldots, n. \quad (2.3.40)$$

Do mesmo modo como deduzimos a fórmula 2.3.39, podemos mostrar que

$$a_{ij}^{(j-1)} = a_{ij} - \sum_{k=1}^{j-1} m_{ik}\, a_{kj}^{(k-1)}. \quad (2.3.41)$$

Se, no segundo membro da fórmula 2.3.40, substituirmos o numerador de acordo com 2.3.41, obtemos

$$m_{ij} = \frac{a_{ij} - \sum_{k=1}^{j-1} m_{ik}\, a_{kj}^{(k-1)}}{a_{jj}^{(j-1)}};\ i = j, \ldots, n. \quad (2.3.42)$$

Mas, de acordo com a hipótese da indução , podemos substituir, no segundo membro de 2.3.42, $a_{kj}^{(k-1)}$ por u_{kj},$k = 1, \ldots, j$ e m_{ik} por l_{ik},$k = 1 \ldots, i$. Então o segundo membro de 2.3.42 fica igual ao segundo membro de 2.3.32, 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.□

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.

Factorização de Crout

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

$$u_{ii} = 1, \quad i = 1, \ldots, n. \tag{2.3.43}$$

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

$$a_{ij} = \sum_{k=1}^{j} l_{ik}\, u_{kj} = \sum_{k=1}^{j-1} l_{ik}\, u_{kj} + l_{ij}, \quad j = 1, \ldots, n;\ i = j, \ldots, n. \tag{2.3.44}$$

Daqui obtém-se imediatamente

$$l_{ij} = a_{ij} - \sum_{k=1}^{j-1} l_{ik}\, u_{kj}. \tag{2.3.45}$$

No que diz respeito à matriz L, partimos da igualdade

$$a_{ij} = \sum_{k=1}^{i} l_{ik}\, u_{kj} = \sum_{k=1}^{i-1} l_{ik}\, u_{kj} + l_{ii}\, u_{ij}; \quad i = 1, \ldots, n;\ j = 1, \ldots, i. \tag{2.3.46}$$

Da igualdade 2.3.46 resulta

$$u_{ij} = \frac{a_{ij} - \sum_{k=1}^{i-1} l_{ik}\, u_{kj}}{l_{ii}}. \tag{2.3.47}$$

As fórmulas 2.3.45 e 2.3.47, 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 2.3.33.

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

$$A = \begin{bmatrix} 2 & 1 & 3 \\ -2 & -1 & 1 \\ 2 & 4 & 2 \end{bmatrix}, \quad b = \begin{bmatrix} 5 \\ -1 \\ 4 \end{bmatrix}. \tag{2.3.48}$$

Comecemos por factorizar A segundo o método de Doolittle. Como resulta da fórmula 2.3.30, neste caso a primeira linha de U é igual à primeira linha de A, ou seja,

$$u_{11} = 2, \quad u_{12} = 1, \quad u_{13} = 3. \tag{2.3.49}$$

Calculando os elementos da primeira coluna de L, de acordo com a fórmula 2.3.32, obtemos

$$l_{11} = 1, \quad l_{21} = \frac{a_{21}}{u_{11}} = -1, \quad l_{31} = \frac{a_{31}}{u_{11}} = 1. \tag{2.3.50}$$

Passemos ao cálculo da segunda linha de U. Temos

$$\begin{array}{rcl} u_{22} & = & a_{22} - l_{21}\, u_{12} = 0; \\ u_{23} & = & a_{23} - l_{21}\, u_{13} = 4. \end{array} \tag{2.3.51}$$

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

$$\begin{array}{rcl} u'_{22} & = & u_{23}, \\ u'_{23} & = & u_{22}. \end{array} \tag{2.3.52}$$

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

$$\begin{array}{rcccl} l_{32} & = & \dfrac{a'_{32} - l_{31}\, u'_{12}}{u'_{22}} & = & \dfrac{a_{33} - l_{31}\, u_{13}}{u_{23}} = -\dfrac{1}{4}; \\ u'_{33} & = & a'_{33} - l_{31}\, u'_{13} - l_{32}\, u'_{23} & = & a_{32} - l_{31}\, u_{12} - l_{32}\, u_{22} = 3. \end{array} \tag{2.3.53}$$

Recapitulando, obtivemos a seguinte factorização de A:

$$L = \begin{bmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 1 & -\frac{1}{4} & 1 \end{bmatrix}, \quad U' = \begin{bmatrix} 2 & 3 & 1 \\ 0 & 4 & 0 \\ 0 & 0 & 3 \end{bmatrix}. \tag{2.3.54}$$

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:

$$\begin{array}{rcccccl} g_1 & = & b_1 & = & 5; & & \\ -g_1 + g_2 & = & b_2 & \Leftrightarrow & g_2 & = & 4; \\ g_1 - g_2/4 + g_3 & = & b_3 & \Leftrightarrow & g_3 & = & 0. \end{array} \tag{2.3.55}$$

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

$$\begin{array}{rcccccl} 3x_2 & = & g_3 & \Leftrightarrow & x_2 & = & 0; \\ 4x_3 & = & g_2 & \Leftrightarrow & x_3 & = & 1. \\ 2x_1 + 3x_3 + x_2 & = & g_1 & \Leftrightarrow & x_1 & = & 1. \end{array} \tag{2.3.56}$$

Se, em vez do método de Doolittle, quisermos aplicar a factorização de Crout, teremos de basear os cálculos nas fórmulas 2.3.45 e 2.3.47. Nesse caso,

a primeira coluna de L fica igual à primeira coluna de A. Para a primeira linha de U, obtém-se

$$u_{11} = 1; \;\; u_{12} = \frac{a_{12}}{l_{11}} = \frac{1}{2}; \;\; u_{13} = \frac{a_{13}}{l_{11}} = \frac{3}{2}. \tag{2.3.57}$$

Na segunda coluna de L, obtemos:

$$\begin{array}{rcl} l_{22} & = & a_{22} - l_{21}\,u_{12} \;\; = \;\; 0; \\ l_{32} & = & a_{32} - l_{31}\,u_{12} \;\; = \;\; 3. \end{array} \tag{2.3.58}$$

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

$$\begin{array}{rcl} l'_{22} & = & l_{32} \;\; = \;\; 3; \\ l'_{32} & = & l_{22} \;\; = \;\; 0. \end{array} \tag{2.3.59}$$

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

$$\begin{array}{rcl} u_{23} & = & \dfrac{a'_{23} - l'_{21}u_{13}}{l'_{22}} \;\; = \;\; -\frac{1}{3}; \\ l'_{33} & = & a'_{33} - l'_{31}u_{13} - l'_{32}u_{23} \;\; = \;\; 4. \end{array} \tag{2.3.60}$$

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

$$L' = \begin{bmatrix} 2 & 0 & 0 \\ 2 & 3 & 0 \\ -2 & 0 & 4 \end{bmatrix}, \;\; U = \begin{bmatrix} 1 & \frac{1}{2} & \frac{3}{2} \\ 0 & 1 & -\frac{1}{3} \\ 0 & 0 & 1 \end{bmatrix}. \tag{2.3.61}$$

A partir de qualquer uma das factorizações de A obtidas, utilizando a fórmula 2.3.33, 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).

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:

$$A = L\,L^T. \tag{2.3.62}$$

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 2.3.5. Seja A uma matriz simétrica e definida positiva. Então existe uma matriz real, triangular inferior, L, tal que se verifica a factorização 2.3.62.

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

$$\sum_{i=1,j=1}^{n} a_{ij}\,x_i\,x_j > 0, \quad \forall x \in \mathbf{R^n}, x \neq 0. \tag{2.3.63}$$

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 2.3.63 resulta

$$\sum_{i=1,j=1}^{n-1} a_{ij}\, x_i\, x_j > 0; \tag{2.3.64}$$

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

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

$$L = \begin{bmatrix} \hat{L} & 0 \\ \gamma^T & z \end{bmatrix}, \tag{2.3.65}$$

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

$$L\, L^T = \begin{bmatrix} \hat{L} & 0 \\ \gamma^T & z \end{bmatrix} \begin{bmatrix} \hat{L}^T & \gamma \\ 0 & z \end{bmatrix} = A = \begin{bmatrix} \hat{A} & c \\ c^T & d \end{bmatrix}, \tag{2.3.66}$$

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

$$\hat{A} = \hat{L}\, \hat{L}^T. \tag{2.3.67}$$

Resta provar que existe um vector $\gamma \in \mathbf{R}^{n-1}$ e um número real z, para os quais se verifica a igualdade 2.3.66. 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 2.3.66 é verdadeira. Para que a igualdade 2.3.66 seja verdadeira, deve verificar-se

$$det\, A = (\, det\, \hat{L})^2\, z^2 > 0. \tag{2.3.68}$$

Uma vez que $det\, A > 0$,a equação 2.3.68 tem duas raizes e o valor de z pode ser determinado, escolhendo a raiz positiva. Assim, o teorema fica demonstrado, por indução . □

Observação . Note-se que o problema da factorização de Cholesky admite mais do que uma solução . Isso verifica-se pela equação 2.3.68, a qual admite duas raizes reais:

$$z_{1,2} = \pm \frac{\sqrt{det\, A}}{det\, \hat{L}} \tag{2.3.69}$$

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 2.3.62 resulta

$$a_{ij} = \sum_{k=1}^{j} l_{ik}\, l_{jk} = \sum_{k=1}^{j-1} l_{ik}\, l_{jk} + l_{ij}\, l_{jj}, j = 1, \ldots, n;\ i = j, \ldots, n. \tag{2.3.70}$$

No caso de $i = j$, da igualdade 2.3.70 obtém-se a fórmula para os elementos da diagonal principal de L:

$$l_{ii} = \sqrt{a_{ii} - \sum_{k=1}^{i-1} l_{ik}^2};\ \ i = 1, \ldots, n. \tag{2.3.71}$$

De acordo com o teorema 2.3.5, todos os elementos da diagonal principal de L são reais, pelo que o segundo membro de 2.3.71 é sempre real. Uma vez calculado l_{jj}, podemos obter as restantes componentes da j-ésima coluna de L. Da fórmula 2.3.70 resulta:

$$l_{ij} = \frac{a_{ij} - \sum_{k=1}^{j-1} l_{ik}\, l_{jk}}{l_{jj}};\ \ i = j+1, \ldots, n. \tag{2.3.72}$$

Assim, usando as fórmulas 2.3.71 e 2.3.72, alternadamente, pode ser obtida a factorização de Cholesky da matriz A.

Exemplo 2.3.6. Consideremos a matriz de ordem n com a forma geral

$$A = \begin{bmatrix} 4 & 2 & 0 & \dots & 0 \\ 2 & 5 & 2 & \dots & 0 \\ 0 & 2 & 5 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & 2 & 5 & 2 \\ 0 & \dots & 0 & 2 & 5 \end{bmatrix}. \tag{2.3.73}$$

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 2.3.71 e 2.3.72 e verificar se elas são aplicáveis. Comecemos, como sempre, pela componente l_{11}. De acordo com 2.3.71, o seu valor é o seguinte:

$$l_{11} = \sqrt{a_{11}} = 2. \tag{2.3.74}$$

As restantes componentes desta coluna são dadas pela fórmula 2.3.72:

$$\begin{aligned} l_{21} &= \frac{a_{21}}{l_{11}} = 1; \\ l_{k1} &= \frac{a_{k1}}{l_{11}} = 0; \quad k = 3, \dots, n. \end{aligned} \tag{2.3.75}$$

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

$$\begin{aligned} l_{jj} &= 2; \\ l_{j+1,j} &= 1; \\ l_{i,j} &= 0; \quad i = j+2, \dots, n. \end{aligned} \tag{2.3.76}$$

Para a primeira coluna, as fórmulas 2.3.76 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 2.3.71, podemos escrever

$$l_{jj} = \sqrt{a_{jj} - \sum_{k=1}^{j-1} l_{jk}^2} = \sqrt{a_{jj} - l_{j,j-1}^2} = 2. \tag{2.3.77}$$

Depois, aplicando a fórmula 2.3.72, obtemos

$$\begin{aligned} l_{j+1,j} &= \frac{a_{j+1,j}}{l_{jj}} = 1; \\ l_{i,j} &= 0; \qquad\qquad i = j+2, \ldots, n. \end{aligned} \tag{2.3.78}$$

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

$$L = \begin{bmatrix} 2 & 0 & 0 & \ldots & 0 \\ 1 & 2 & 0 & \ldots & 0 \\ 0 & 1 & 2 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & \ldots & 1 & 2 & 0 \\ 0 & \ldots & 0 & 1 & 2 \end{bmatrix}. \tag{2.3.79}$$

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

$$det\, A = (det\, L)^2 = (l_{11}\, l_{22}\, \ldots\, l_{nn})^2 = (2^n)^2 = 4^n. \tag{2.3.80}$$

Uma vez que a fórmula 2.3.80 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.

2.3.3 Problemas

1. Mostre que se a matriz A for tridiagonal, isto é se se verificar $a_{ij} = 0$, se $|i-j| > 1$, então as matrizes L e U da sua factorização pelo método de Crout satisfazem as condições

$$\begin{aligned} l_{ij} &= 0, \quad se\ i < j\ ou\ i > j+1; \\ u_{ij} &= 0, \quad se\ i > j\ ou\ i < j+1. \end{aligned}$$

2. Considere a matriz quadrada de ordem n com a forma geral

$$A = \begin{bmatrix} 9 & 3 & 0 & \dots & 0 \\ 3 & 10 & 3 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & 3 & 10 & 3 \\ 0 & \dots & 0 & 3 & 10 \end{bmatrix}.$$

   (a) Obtenha a forma geral da factorização de Crout desta matriz.

   (b) Com base na factorização obtida, calcule $DetA$.

   (c) Resolva o sistema $A\,x = b$, onde $b = (0, 0, ..., 1)^T$.

   (d) Prove que esta matriz é definida positiva e determine a sua factorização de Cholesky.

3. Considere o sistema linear

$$\begin{cases} 10^{-6}\,x + y = 0.5 \\ x + y = 1. \end{cases}$$

   (a) Resolva o sistema pelo método da eliminação de Gauss.

   (b) Suponha que o sistema é resolvido numa calculadora onde os números são representados num sistema de vírgula flutuante, apenas com 6 dígitos na mantissa. Que solução obteria nesse caso? Compare com a solução exacta.

   (c) Suponha que o sistema é resolvido na mesma máquina, mas usando *pesquisa parcial de pivot*. Qual é o resultado nestas condições? Compare com o resultado da alínea anterior e comente.

4. Considere o sistema linear

$$\begin{cases} x & + & 10^6\, y & = & 0.5 \times 10^6 \\ x & + & y & = & 1. \end{cases}$$

   (a) Mostre que este sistema é equivalente ao da alínea anterior.

   (b) Será que, neste caso, a pesquisa parcial de pivot permite superar os efeitos dos erros de arredondamento, como acontecia na alínea anterior? Justifique.

   (c) Resolva o sistema, utilizando o método da pesquisa total de pivot. Comente.

2.4 Métodos 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.

2.4.1 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 2.4.1. 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 2.4.2. Dizemos que um método iterativo Ψ a p passos, definido sobre x, é convergente para um certo $x \in E$, num certo domínio $A \subset X^p$, 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 em E).

Note-se que, tal como foi provado no parágrafo 2.1, 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.

Exemplo 2.4.3 . Consideremos um método iterativo a 1 passo, definido sobre $X = \mathbf{R}$, com a função iteradora $\phi(x) = x^2$. Se for dado um certo elemento $\xi_0 \in \mathbf{R}$, fica inteiramente definida uma sucessão $\{x^{(k)}\}_{k=0}^{\infty}$, com a seguinte forma:

$$\begin{aligned} x^{(0)} &= \xi_0, \\ x^{(1)} &= \phi(x^0) = (x^{(0)})^2 = \xi_0^2, \\ x^{(k)} &= \xi_0^{2^k},\ k = 2, 3, \ldots \end{aligned} \tag{2.4.1}$$

Vejamos em que domínio converge este método iterativo. É evidente que, se $|\xi_0| < 1$, então $x^k \to 0$. Ou seja, o método é convergente para 0 no domínio $A = \{\xi_0 \in \mathbf{R} : |\xi_0| < 1\}$. Se tivermos $\xi_0 = 1$, então $x^{(k)} = 1, \forall k \in N$, ou seja, o método converge para 1. Se $\xi_0 = -1$, então $x^{(k)} = (-1)^k, \forall k \in N$, pelo que o método diverge. Por último, se tivermos $|\xi_0| > 1$, então a sucessão $\{x^{(k)}\}_{k=0}^{\infty}$ é ilimitada, pelo que o método também diverge nesse caso.

Exemplo 2.4.4. Consideremos o método a 2 passos, definido em $\mathbf{R}$, com a função iteradora $\phi(x,y) = 3y - 2x$. Para qualquer par ordenado $(\xi_0, \xi_1) \in \mathbf{R^2}$,este método gera uma sucessão $\{x^{(k)}\}_{k=0}^{\infty}$, com os seguintes elementos:

$$\begin{aligned} x^{(0)} &= \xi_0, \\ x^{(1)} &= \xi_1, \\ x^{(k+2)} &= \phi(x^{(k)}, x^{(k+1)}) = 3x^{(k+1)} - 2x^{(k)},\ k = 0, 1, 2, \ldots \end{aligned} \tag{2.4.2}$$

A fim de estudar a convergência deste método, consideremos a equação com diferenças

$$x^{(k+2)} - 3x^{(k+1)} + 2x^{(k)} = 0. \tag{2.4.3}$$

A equação característica correspondente é

$$\lambda^2 - 3\lambda + 2 = 0 \tag{2.4.4}$$

e as suas raizes são $\lambda_1 = 1, \lambda_2 = 2$. Por conseguinte, a solução geral da equação 2.4.3 tem a forma

$$x^{(k)} = C_1 + C_2\, 2^k, \tag{2.4.5}$$

onde as constantes C_1 e C_2 são determinadas, para cada solução particular, a partir dos elementos iniciais (ξ_0, ξ_1), mediante a resolução do sistema de equações :

$$\begin{aligned} C_1 + C_2 &= \xi_0 \\ C_1 + 2\,C_2 &= \xi_1. \end{aligned}$$

A solução deste sistema é

$$\begin{aligned} C_2 &= \xi_1 - \xi_0 \\ C_1 &= 2\,\xi_0 - \xi_1. \end{aligned} \tag{2.4.6}$$

Da forma da solução geral 2.4.5 resulta que a sucessão $\{x^{(k)}\}_{k=0}^{\infty}$ é convergente se e só se $C_2 = 0$. Substituindo na fórmula 2.4.6, conclui-se que o método iterativo considerado tem o seguinte domínio de convergência:

$$A = \{(\xi_0, \xi_1) \in \mathbf{R^2} : \xi_1 - \xi_0 = 0\}.$$

Para valores iniciais dentro desse domínio, o método iterativo gera uma sucessão constante, com a forma

$$x^{(k)} = C_1 = 2\,\xi_0 - \xi_1 = \xi_0,\ k = 0, 1, \ldots$$

Se os valores iniciais não pertencerem a A, então o método iterativo gera uma sucessão ilimitada, logo divergente.

Para efeitos práticos, interessa-nos não só que um método iterativo seja convergente, mas também que convirja rapidamente, isto é, que nos permita alcançar a precisão necessária com um número não muito grande de iterações. Quanto a este aspecto, os métodos iterativos distinguem-se entre si pela *ordem de convergência.*

Definição 2.4.5. Seja $\{x^{(k)}\}_{k=0}^{\infty}$ uma sucessão convergente para x num espaço normado E. Diz-se que esta sucessão tem *ordem de convergência não inferior a* r (onde $r \in \mathbf{R}^+$) se existir uma constante positiva C_r tal que

$$\|x^{(k+1)} - x\| \leq C_r \|x^{(k)} - x\|^r,\ \forall k \in \mathbf{N}. \tag{2.4.7}$$

Se r for o maior de todos os números reais, para os quais a sucessão $\{x^{(k)}\}_{k=0}^{\infty}$ tem ordem de convergência não inferior a r, então diz-se que esta sucessão tem *ordem de convergência* r.

Definição 2.4.6 Diz-se que uma sucessão *converge linearmente* ou tem *convergência linear* se tiver ordem de convergência $r = 1$.

Definição 2.4.7 Diz-se que uma sucessão tem *convergência supralinear* se tiver ordem de convergência $r > 1$. Em particular, se tiver ordem de convergência $r = 2$, diz-se que tem *convergência quadrática.*

Definição 2.4.8. Seja $\{x^{(k)}\}_{k=0}^{\infty}$ uma sucessão convergente para x num espaço normado E, com ordem de convergência r. Chama-se *factor as-*

simptótico de convergência ao ínfimo de todas as constantes C_r, para as quais é satisfeita a condição 2.4.7.

Exemplo 2.4.9. Consideremos de novo o método iterativo do exemplo 2.4.3 . Seja ξ_0 tal que $|\xi_0| < 1$. Nesse caso, de acordo com o que verificámos, a sucessão $\{x^{(k)}\}_{k=0}^{\infty}$, gerada pelo método iterativo, converge para $x = 0$. Determinemos a ordem de convergência. De acordo com a fórmula 2.4.1 , temos

$$\|x^{(k+1)} - x\| = |x^{(k+1)}| = |x^{(k)}|^2. \tag{2.4.8}$$

Logo, verifica-se

$$\|x^{(k+1)} - x\| = \|x^{(k)} - x\|^2, \forall k \in \mathbf{N}. \tag{2.4.9}$$

De acordo com a definição 2.4.5, esta sucessão tem ordem de convergência não inferior a 2. Como facilmente se verifica, para $r > 2$ não existe nenhuma constante C_r para a qual seja satisfeita a condição 2.4.7. Por conseguinte, a sucessão considerada tem ordem de convergência 2 ou, de acordo com a definição 2.4.7, tem *convergência quadrática.*

Da igualdade 2.4.9 resulta que a sucessão satisfaz a condição 2.4.7 com a constante $C_2 = 1$, mas não a satisfaz com nenhuma constante inferior a 1. Logo, de acordo com a definição 2.4.8, o *factor assimptótico de convergência desta sucessão é* 1.

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

Definição 2.4.10. 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

$$\max_{n \in N} \|x^{(n)} - y^{(n)}\| \leq C \max_{i=1,\ldots,n} \|\xi_i - \eta_i\| \ \forall \xi, \eta \in A^p, \tag{2.4.10}$$

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

Por outras palavras, um método iterativo diz-se estável se, a dois conjuntos próximos de valores iniciais, ξ e η, ele faz corresponder sucessoes igualmente próximas. Note-se que se um método iterativo for convergente, em todo o espaço, para um certo elemento x, então ele também é estável em todo o espaço (nesse caso, diz-se *incondicionalmente estável*). Mas se ele só for convergente num certo subespaço X, então pode não ser estável nesse sub-espaço.

Nos próximos parágrafos, vamos analisar alguns métodos iterativos para o cálculo aproximado da solução do sistema linear

$$A\,x \;=\; b, \tag{2.4.11}$$

onde $A \in L^n$ e $b \in L^n$. Suponhamos, como habitualmente, que a matriz A é não singular, pelo que o sistema 2.4.11 tem uma única solução .

O primeiro passo para a dedução de um método iterativo consiste, geralmente, em reduzir o sistema a uma forma adequada. Com este fim, vamos supor que a matriz A pode ser decomposta na soma de duas matrizes M e N, que apresentam certas propriedades especiais. Então o sistema 2.4.11 pode ser escrito, de novo, sob a forma

$$M\,x \;=\; b \;-\; N\,x. \tag{2.4.12}$$

Os sistemas 2.4.11 e 2.4.12 são equivalentes. Entretanto, a forma do sistema 2.4.12 sugere-nos que a sua solução pode ser aproximada por um método iterativo a 1 passo, em que a iterada $x^{(k+1)}$ pode ser obtida a partir da anterior resolvendo o sistema

$$M\,x^{(k+1)} \;=\; b \;-\; N\,x^{(k)}, k = 0, 1, 2, \ldots \tag{2.4.13}$$

A fórmula 2.4.13 está na base de uma série de métodos iterativos para sistemas lineares. A ideia básica destes métodos consiste em definir as matrizes M e N de tal modo que a solução do sistema 2.4.13 seja imediata. Nos próximos parágrafos, analisaremos dois casos particulares de métodos baseados na fórmula 2.4.13, que são conhecidos como método de Jacobi e método de Gauss-Seidel.

2.4.2 Método de Jacobi

O método de Jacobi é um exemplo de um método que se baseia na fórmula 2.4.13. Neste caso, o sistema 2.4.11 é reduzido à forma 2.4.12, utilizando a decomposição da matriz A na soma $A = M_J + N_J$, onde as matrizes M_J e N_J têm, repectivamente, as formas

$$M_J = \begin{bmatrix} a_{11} & 0 & \dots & 0 \\ 0 & a_{22} & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & \dots & 0 & a_{nn} \end{bmatrix}, \; N_J = \begin{bmatrix} 0 & a_{12} & \dots & a_{1n} \\ a_{21} & 0 & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & \dots & a_{n-1,n} & 0 \end{bmatrix}. \quad (2.4.14)$$

Ou seja, M_J é uma matriz diagonal, cujos elementos são as entradas da diagonal principal de A, enquanto N_J difere de A apenas por ter a diagonal principal preenchida por zeros. Neste caso, dada a forma simples de M_J, o sistema 2.4.13 pode ser escrito, por componentes, sob a forma

$$a_{ii}\, x_i^{(k+1)} = b_i - \sum_{j=1, j\neq i}^{n} a_{ij}\, x_j, \; i = 1, 2, \dots, n \quad (2.4.15)$$

Do sistema escrito sob a forma 2.4.15 deduz-se imediatamente a função iteradora do método de Jacobi:

$$x_i^{(k+1)} = \frac{1}{a_{ii}} \left[b_i - \sum_{j=1, j\neq i}^{n} a_{ij}\, x_j^{(k)} \right], \; i = 1, 2, \dots, n; \; k = 0, 1, \dots, \quad (2.4.16)$$

Para aplicar o método de Jacobi, como qualquer método iterativo a 1 passo, escolhe-se uma aproximação inicial

$$\xi_0 = x^{(0)} = (x_1^{(0)}, \dots, x_n^{(0)}) \in E^n$$

(em geral, arbitrária) e, a partir dela, constroi-se uma sucessão de vectores $\{x^{(k)}\}_{k=0}^{\infty}$, onde cada termo é obtido a partir do anterior, de acordo com a fórmula 2.4.16.

Exemplo 2.4.11. Consideremos o sistema linear em $\mathbf{R^3}$:

$$\begin{cases} 2x_1 + x_2 = 2 \\ -x_1 + 2x_2 + x_3 = 2 \\ - x_2 + 2x_3 = 1 \end{cases}. \quad (2.4.17)$$

Este sistema pode ser escrito na forma matricial $A\,x = b$, onde

$$A = \begin{bmatrix} 2 & 1 & 0 \\ -1 & 2 & 1 \\ 0 & -1 & 2 \end{bmatrix},\ b = \begin{bmatrix} 2 \\ 2 \\ 1 \end{bmatrix}. \tag{2.4.18}$$

Logo, a fórmula 2.4.16 tem, neste caso, o seguinte aspecto

$$\begin{aligned} x_1^{(k+1)} &= \frac{1}{2}\left[2 - x_2^{(k)}\right], \\ x_2^{(k+1)} &= \frac{1}{2}\left[2 + x_1^{(k)} - x_3^{(k)}\right], \\ x_3^{(k+1)} &= \frac{1}{2}\left[1 + x_2^{(k)}\right];\ k = 0, 1, \ldots \end{aligned}$$

Assim, se a aproximação inicial for, por exemplo, $x^{(0)} = (1, 1, 1)^T$, obtém-se a primeira iterada com a forma

$$\begin{aligned} x_1^{(1)} &= \frac{1}{2}[2 - 1] = \frac{1}{2}, \\ x_2^{(1)} &= \frac{1}{2}[2 + 1 - 1] = 1, \\ x_3^{(1)} &= \frac{1}{2}[1 + 1] = 1. \end{aligned}$$

As iteradas seguintes podem ser calculadas exactamente do mesmo modo.

Note-se que as fórmulas do método de Jacobi não podem ser aplicadas se algum dos elementos da diagonal principal de A for nulo. No entanto, esta dificuldade pode ser facilmente ultrapassada, se trocarmos a ordem das linhas da matriz de modo a que o(s) elemento(s) nulos deixem de estar na diagonal principal. 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. Para qualquer método, baseado na fórmula 2.4.13, os erros das iteradas, que representaremos por $e^{(k)}$, $k = 0, 1, \ldots$, satisfazem a igualdade

$$e^{(k+1)} = x^{(k+1)} - x = -M^{-1}\,N(x^{(k)} - x);\ k = 0, 1, 2, \ldots \tag{2.4.19}$$

A igualdade 2.4.19 também pode ser escrita na forma

$$e^{(k+1)} = C\, e^{(k)};\ k = 0, 1, 2, \ldots \tag{2.4.20}$$

onde

$$C = -M^{-1}\, N. \tag{2.4.21}$$

Como veremos adiante, é das propriedades da matriz C que depende a convergência dos métodos iterativos que estamos a analisar. No caso do método de Jacobi, atendendo à forma das matrizes M_J e N_J, dadas pelas igualdades 2.4.14, deduz-se que a matriz C tem a forma

$$C = C_J = -M_J^{-1}\, N_J = \begin{bmatrix} 0 & -\frac{a_{12}}{a_{11}} & \ldots & -\frac{a_{12}}{a_{1n}} \\ -\frac{a_{21}}{a_{22}} & 0 & \ldots & -\frac{a_{2n}}{a_{22}} \\ \ldots & \ldots & \ldots & \ldots \\ -\frac{a_{n1}}{a_{nn}} & \ldots & -\frac{a_{n-1,n}}{a_{nn}} & 0 \end{bmatrix}. \tag{2.4.22}$$

Vejamos então que propriedades deve apresentar a matriz C para garantir a convergência de um método iterativo. Em primeiro lugar, notemos que da igualdade 2.4.20 resulta imediatamente uma formula que exprime o erro de qualquer iterada através do erro da aproximação inicial:

$$e^{(k)} = C^k\, e^{(0)};\ k = 0, 1, 2, \ldots \tag{2.4.23}$$

Definição 2.4.12 Uma matriz $C \in L^n$, que representa uma aplicação linear no espaço vectorial E^n, diz-se *convergente* se e só se

$$\lim_{k\to\infty} C^k\, x = 0,\ \forall x \in E^n. \tag{2.4.24}$$

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 2.4.13.

Teorema 2.4.13. Seja $\{x^{(k)}\}_{k=0}^{\infty}$ uma sucessão em E^n, gerada pela fórmula 2.4.13, com certas matrizes M e N, tais que $M + N = A$. Então esta sucessão converge para a solução do sistema $Ax = b$, qualquer que seja a aproximação inicial ξ_0, se e só se a matriz $C = -M^{-1}\, N$ 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 2.4.23 e 2.4.24, temos

$$\lim_{k\to\infty} e^{(k)} = \lim_{k\to\infty} C^k\, e^{(0)} \;=\; 0, \tag{2.4.25}$$

qualquer que seja o vector $e^{(0)} \in E^n$, independentemente da norma considerada. Isto significa que o método iterativo converge, qualquer que seja a aproximação inicial $x^{(0)} \in E^n$.

Condição necessária. Suponhamos que C não é convergente. Então existe um vector $v \in E^n$ 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 2.4.23, 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 $x^{(0)} \;=\; x + v$. □

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 2.4.14. Seja $C \in L^n$. Se existir uma norma matricial M, associada a uma certa norma vectorial V, tal que

$$\|C\|_M \;<\; 1 \tag{2.4.26}$$

então a matriz C é convergente.

Demonstração . Seja x um vector arbitrário de E^n. Então , de acordo com as propriedades das normas matriciais, temos

$$\|C^k\, x\|_V \;\leq\; \|C^k\|_M\, \|x\|_V \;\leq\; (\|C\|_M)^k\, \|x\|_V. \tag{2.4.27}$$

Da desigualdade 2.4.27 resulta imediatamente que, se $\|C\|_V \;<\; 1$, então

$$\lim_{k\to\infty} \|C^k\, x\|_V \;=\; 0, \tag{2.4.28}$$

o que significa, por definição , que a matriz C é convergente.□

Teorema 2.4.15. Para que a matriz $C \in L^n$ seja convergente é necessário e suficiente que

$$r_\sigma(C) < 1. \tag{2.4.29}$$

Demonstração .*Condição suficiente.* Se tivermos $r_\sigma(C) = \rho < 1$, de acordo com o teorema 2.1.31, para qualquer $\epsilon > 0$, existe uma norma matricial $N(\epsilon)$ tal que

$$\|C\|_{N(\epsilon)} \leq \rho + \epsilon. \tag{2.4.30}$$

Se considerarmos $\epsilon = \frac{1-\rho}{2}$, obtemos

$$\|C\|_{N(\epsilon)} \leq \frac{\rho + 1}{2} < 1. \tag{2.4.31}$$

Da desigualdade 2.4.31 resulta, pelo teorema 2.4.14, que a matriz C é convergente.

Condição necessária. Suponhamos que a condição 2.4.29 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

$$\|C^k\, v\| = \|\lambda^k\, v\| = |\lambda|^k\, \|v\|. \tag{2.4.32}$$

Da igualdade 2.4.32, 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.□

Se aplicarmos os teoremas 2.4.14 e 2.4.15 ao caso particular do método de Jacobi, obtemos critérios de convergência para este método. A fim de formularmos estes critérios numa forma mais concisa, vamos introduzir mais algumas definições .

Definição 2.4.16. Diz-se que a matriz $A \in L^n$ tem a *diagonal estritamente dominante por linhas*, se forem satisfeitas as condições :

$$\sum_{j=1, j\neq i}^{n} |a_{ij}| < |a_{ii}|,\ i = 1, \ldots, n. \tag{2.4.33}$$

Definição 2.4.17. Diz-se que a matriz $A \in L^n$ tem a *diagonal estritamente dominante por colunas*, se forem satisfeitas as condições :

$$\sum_{i=1, i\neq j}^{n} |a_{ij}| < |a_{jj}|, \ j = 1, \ldots, n. \tag{2.4.34}$$

Dos teoremas 2.4.13 e 2.4.14 resulta o seguinte critério de convergência para o método de Jacobi.

Teorema 2.4.18. Se a matriz A tiver a diagonal estritamente dominante por linhas, então o método de Jacobi converge para a solução do sistema $Ax = b$, qualquer que seja a aproximação inicial $x^{(0)} \in E^n$.

Demonstração . Se a matriz A tiver a diagonal estritamente dominante por linhas, das desigualdades 2.4.33 resulta

$$\sum_{j=1, j\neq i}^{n} \frac{|a_{ij}|}{|a_{ii}|} < 1 \ i = 1, \ldots, n. \tag{2.4.35}$$

De acordo com a forma da matriz C_J, dada por 2.4.22, as desigualdades 2.4.35 implicam

$$\|C_J\|_\infty = \max_{i=1,\ldots,n} \sum_{j=1, j\neq i}^{n} \frac{|a_{ij}|}{|a_{ii}|} < 1, \tag{2.4.36}$$

De acordo com o teorema 2.4.14, a condição 2.4.36 garante que a matriz C_J é convergente. Isto, por sua vez, implica, de acordo com o teorema 2.4.13, 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 2.4.19. Se a matriz A tiver a diagonal estritamente domi nante por colunas, então o método de Jacobi converge para a solução do sistema $Ax = b$, qualquer que seja a aproximação inicial $x^{(0)} \in E^n$.

Demonstração . Suponhamos que a matriz A satisfaz as condições 2.4.34. Seja D uma matriz diagonal com a forma $D = diag(a_{11}, \ldots, a_{nn})$.

Então , à semelhança do que se faz no problema 10 do parágrafo 2.1.4, podemos definir uma norma matricial M pela igualdade

$$\|C\|_M \;=\; \|D\,C\,D^{-1}\|_1,\ \forall C \in L^n. \tag{2.4.37}$$

Das condições 2.4.34 obtem-se facilmente que

$$\|C_J\|_M \;=\; \|D\,C_J\,D^{-1}\|_1 \;<\; 1. \tag{2.4.38}$$

De acordo com o teoremas 2.4.14 e 2.4.13, da desigualdade 2.4.38 resulta que o método de Jacobi converge para a solução do sistema $A\,x \;=\; b$, qualquer que seja a aproximação inicial $x^{(0)} \in E^n$. $\square$.

Exemplo 2.4.20 Voltemos ao sistema do exemplo 2.4.11 . Recordemos que a matriz A daquele sistema tem a forma

$$A \;=\; \begin{bmatrix} 2 & 1 & 0 \\ -1 & 2 & 1 \\ 0 & -1 & 2 \end{bmatrix}. \tag{2.4.39}$$

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 2.4.18 e 2.4.19 não são , neste caso, aplicáveis. Vejamos se é possível aplicar directamente o teorema 2.4.15.

A matriz C_J, neste caso, tem a forma:

$$C_J \;=\; \begin{bmatrix} 0 & -1/2 & 0 \\ 1/2 & 0 & -1/2 \\ 0 & 1/2 & 0 \end{bmatrix}. \tag{2.4.40}$$

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, pelos teoremas 2.4.13 e 2.4.15, podemos concluir que o método de Jacobi converge para a solução do sistema considerado, qualquer que seja a aproximação inicial.

2.4.3 Método de Gauss-Seidel

Um outro método iterativo que se enquadra igualmente no esquema definido pela fórmula 2.4.13 é o método de Gauss-Seidel. Este método caracteriza-se pela decomposição da matriz A do sistema na forma
$A = M_S + N_S$, onde M_S e N_S são ,respectivamente, as matrizes:

$$M_S = \begin{bmatrix} a_{11} & 0 & \dots & 0 \\ a_{21} & a_{22} & \dots & 0 \\ \dots & \dots & \dots & \dots \\ a_{n,1} & a_{n,2} & \dots & a_{n,n} \end{bmatrix}, \; N_S = \begin{bmatrix} 0 & a_{12} & \dots & \dots & a_{1n} \\ 0 & 0 & a_{23} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & \dots & \dots & a_{n-1,n} \\ 0 & \dots & \dots & \dots & 0 \end{bmatrix}. \tag{2.4.41}$$

Neste caso, não sendo a matriz M_S diagonal, a obtenção de fórmulas explícitas, por componente, para cada iterada, não é imediata. Para obtermos estas fórmulas, vamos decompor a matriz M_S, por sua vez, na soma $M_S = T + D$, onde

$$T = \begin{bmatrix} 0 & 0 & \dots & 0 \\ a_{21} & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ a_{n,1} & \dots & a_{n-1,n} & 0 \end{bmatrix}, \; D = \begin{bmatrix} a_{11} & 0 & \dots & 0 \\ 0 & a_{22} & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & a_{n,n} \end{bmatrix}. \tag{2.4.42}$$

Então a fórmula iteradora 2.4.13 pode ser escrita sob a forma

$$(T + D)\, x^{(k+1)} = b - N_S\, x^{(k)}. \tag{2.4.43}$$

Desta equação , por sua vez, obtém-se

$$D\, x^{(k+1)} = b - N_S\, x^{(k)} - T\, x^{(k+1)}. \tag{2.4.44}$$

Da fórmula 2.4.44 é fácil obter uma fórmula iteradora explícita, análoga à fórmula 2.4.15, que define o método de Jacobi:

$$x_i^{(k+1)} = \frac{1}{a_{ii}} \left(b_i - \sum_{j=1}^{i-1} a_{ij}\, x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij}\, x_j^{(k)} \right), \; i = 1, 2, \dots, n; \; k = 0, 1, \dots. \tag{2.4.45}$$

Note-se que no segundo membro da fórmula 2.4.45 aparecem componentes do vector $x^{(k+1)}$, que se pretende definir com esta fórmula. Essas componentes

são as que têm o índice $j = 1, \ldots, i-1$. Logo, para se poder calcular a componente com um certo índice i, têm de estar calculadas todas as componentes de índice inferior. Ou seja, a fórmula 2.4.45 só pode ser aplicada por ordem crescente do índice i. Esta particularidade distingue o método de Gauss-Seidel do de Jacobi. Isto explica que o método de Gauss-Seidel também seja conhecido como *método das substituições sucessivas*, enquanto o método de Jacobi, segundo esta terminologia, é o *método das substituições simultâneas*.

Exemplo 2.4.21 . Vamos mostrar como se aplica o método de Gauss-Seidel no caso do sistema $Ax = b$, considerado no exemplo 2.4.11. Recordemos que a matriz A e o vector b deste sistema são , respectivamente,

$$A = \begin{bmatrix} 2 & 1 & 0 \\ -1 & 2 & 1 \\ 0 & -1 & 2 \end{bmatrix}, \; b = \begin{bmatrix} 2 \\ 2 \\ 1 \end{bmatrix}. \tag{2.4.46}$$

Por conseguinte, a fórmula 2.4.45, aplicada a este sistema, adquire a forma

$$\begin{aligned} x_1^{(k+1)} &= \frac{1}{2}\left[2 - x_2^{(k)}\right], \\ x_2^{(k+1)} &= \frac{1}{2}\left[2 + x_1^{(k+1)} - x_3^{(k)}\right], \\ x_3^{(k+1)} &= \frac{1}{2}\left[1 + x_2^{(k+1)}\right]; \; k = 0, 1, \ldots \end{aligned}$$

Se utilizarmos de novo a aproximação inicial $x^{(0)} = (1, 1, 1)$, a primeira iterada, neste caso, será:

$$\begin{aligned} x_1^{(1)} &= \frac{1}{2}\left[2 - 1\right] = \frac{1}{2}, \\ x_2^{(1)} &= \frac{1}{2}\left[2 + \frac{1}{2} - 1\right] = \frac{3}{4}, \\ x_3^{(1)} &= \frac{1}{2}\left[1 + \frac{3}{4}\right] = \frac{7}{8}. \end{aligned}$$

Vamos agora averiguar as condições que garantem a convergência do

método de Gauss-Seidel. Representemos por C_S a matriz

$$C_S = -M_S^{-1} N_S. \tag{2.4.47}$$

De acordo com o teorema 2.4.13, 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 2.4.15, é necessário e suficiente que o raio espectral daquela matriz seja menor que 1. Notemos, no entanto, que, diferentemente do que acontecia no caso do método de Jacobi, a matriz M_S não é diagonal, pelo que a sua inversa, em geral, não pode ser expressa por uma fórmula simples. Torna-se conveniente, portanto, conhecer alguns critérios que nos permitem averiguar a convergência do método de Gauss-Seidel, sem ter que calcular a matriz C_S.

Proposição 2.4.22. Sejam M_S e N_S as matrizes definidas pelas fórmulas 2.4.41, sendo M_S uma matriz não singular. O método de Gauss-Seidel converge para a solução do sistema $A\,x = b$, qualquer que seja a aproximação inicial, se e só se todas as raizes da equação

$$Det(\lambda\, M_S + N_S) = 0 \tag{2.4.48}$$

forem, em módulo, menores que 1.

Demonstração . A equação dos valores próprios da matriz C_S tem a forma

$$Det(-M_S^{-1} N_S - \lambda\, I) = 0. \tag{2.4.49}$$

Multiplicando ambos os membros de 2.4.49 por $-det\, M_S$ obtém-se a equação 2.4.48. Logo, as duas equações são equivalentes, ou seja, os valores próprios de C_S são as raizes da equação 2.4.48. Por conseguinte, o raio espectral de C_S é menor que 1 se e só se todas as raizes da equação 2.4.49 forem, em módulo, menores que 1. Logo, de acordo com o teorema 2.4.15, esta condição é necessária e suficente para a convergência do método de Gauss-Seidel. □

Exemplo 2.4.23. Voltando ao sistema do exemplo 2.4.21,verifiquemos se este sistema satisfaz a condição do teorema 2.4.22. Neste caso, temos

$$\lambda\, M_S + N_S = \begin{bmatrix} 2\lambda & 1 & 0 \\ -\lambda & 2\lambda & 1 \\ 0 & -\lambda & 2\lambda \end{bmatrix}. \tag{2.4.50}$$

Logo, a equação 2.4.48, para este sistema, tem a forma

$$Det(\lambda\, M_S \,+\, N_S) \,=\, 8\lambda^3 \,+\, 4\lambda^2 = 4\lambda^2\,(2\lambda \,+\, 1) \,=\, 0. \tag{2.4.51}$$

As raizes desta equação são

$$\lambda_1 \,=\, \lambda_2 \,= 0, \;\; \lambda_3 = -1/2. \tag{2.4.52}$$

Daqui, pelo teorema 2.4.22, podemos concluir que o método de Gauss-Seidel converge para a solução do sistema $A\,x \,=\, b$, qualquer que seja a aproximação inicial. Também podemos constatar que $r_\sigma(C_S) \,=\, |\lambda_3| = 1/2$.

Pode-se mostrar ainda 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 :

$$\alpha_i \,= \sum_{j=1}^{i-1} \left|\frac{a_{ij}}{a_{ii}}\right|, \;\; \beta_i = \sum_{j=i+1}^{n} \left|\frac{a_{ij}}{a_{ii}}\right|. \tag{2.4.53}$$

Sendo conhecidos α_i e β_i,$i \,=\, 1, \ldots, n$, define-se a grandeza η através da fórmula

$$\eta \,= \max_{i=1,\ldots,n} \left(\frac{\beta_i}{1-\alpha_i}\right). \tag{2.4.54}$$

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 2.4.24. 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

$$\|e^{(k)}\|_\infty \,\leq\, \eta^k\, \|e^{(0)}\|_\infty. \tag{2.4.55}$$

Demonstração . Da fórmula 2.4.45 deduz-se facilmente que o erro da k-ésima iterada do método de Gauss-Seidel satisfaz a igualdade

$$e_i^{(k+1)} \,=\, \frac{1}{a_{ii}} \left(-\sum_{j=1}^{i-1} a_{ij}\, e_j^{(k+1)} \,-\, \sum_{j=i+1}^{n} a_{ij}\, e_j^{(k)}\right), \;\; i = 1, 2, \ldots, n; \;\; k = 0, 1, \ldots. \tag{2.4.56}$$

Tomando o módulo de ambos os membros da igualdade 2.4.56 e entrando em conta com as definições das grandezas α_i e β_i, obtém-se

$$|e_i^{(k+1)}| \leq \alpha_i \|e^{(k+1)}\|_\infty + \beta_i \|e^{(k)}\|_\infty, \; i = 1, 2, \ldots, n; \; k = 0, 1, \ldots. \quad (2.4.57)$$

Seja m o índice para o qual se verifica $|e_m^{(k+1)}| = \|e^{(k+1)}\|_\infty$. Então , escrevendo a desigualdade 2.4.57, com $i = m$, obtém-se

$$\|e^{(k+1)}\|_\infty \leq \alpha_m \|e^{(k+1)}\|_\infty + \beta_m \|e^{(k)}\|_\infty, \; k = 0, 1, \ldots. \quad (2.4.58)$$

Daqui resulta

$$\|e^{(k+1)}\|_\infty (1 - \alpha_m) \leq \beta_m \|e^{(k)}\|_\infty, \; k = 0, 1, \ldots. \quad (2.4.59)$$

Visto que $\alpha_m < 1$, podemos dividir ambos os membros de 2.4.59 por $1 - \alpha_m$ e obter

$$\|e^{(k+1)}\|_\infty \leq \frac{\beta_m}{1 - \alpha_m} \|e^{(k)}\|_\infty \leq \eta \|e^{(k)}\|_\infty, \; k = 0, 1, \ldots. \quad (2.4.60)$$

Da desigualdade 2.4.60 resulta a estimativa do erro 2.4.55. Por outro lado, uma vez que a matriz tem a diagonal estritamente dominante, por linhas, $\eta < 1$. Logo, a desigualdade 2.4.55 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 . □

2.4.4 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 2.4.20, se verifica

$$\|e^{(k+1)}\|_V \leq \|C\|_M \|e^{(k)}\|_V. \tag{2.4.61}$$

Daqui resulta, de acordo com a definição 2.4.5, que se um método deste tipo convergir, ele tem, pelo menos, *ordem de convergência 1.* A rapidez de convergência depende, naturalmente, das propriedades da matriz C e da aproximação inicial escolhida. Nalguns casos especiais, pode acontecer que a ordem de convergência seja superior a 1 ou até 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

$$c_1 = \lim_{k\to\infty} \frac{\|e^{(k+1)}\|_V}{\|e^{(k)}\|_V}. \tag{2.4.62}$$

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 2.4.62, 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

$$x^{(k+1)} - x^{(k)} = -e^{(k+1)} + e^{(k)} = -C\,e^{(k)} + C\,e^{(k-1)} = C(x^{(k)} - x^{(k-1)}). \tag{2.4.63}$$

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 2.4.20. Logo, se o limite 2.4.62 existir, também existe o limite

$$c_1' = \lim_{k\to\infty} \frac{\|x^{(k+1)} - x^{(k)}\|_V}{\|x^{(k)} - x^{(k-1)}\|_V}. \tag{2.4.64}$$

e os dois limites (c_1 e c_1') têm o mesmo valor, para a maioria das aproximações iniciais. A fórmula 2.4.64 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

$$\|e^{(k+1)}\|_V = \|x^{(k+1)} - x\|_V \leq c_2\,\|x^{(k)} - x\|_V. \tag{2.4.65}$$

Por outro lado, da desigualdade triangular, temos

$$\|x^{(k)} - x\|_V \leq \|x^{(k)} - x^{(k+1)}\|_V + \|x^{(k+1)} - x\|_V \tag{2.4.66}$$

Das desigualdades 2.4.66 e 2.4.65 resulta

$$\|x^{(k)} - x\|_V \leq \|x^{(k)} - x^{(k+1)}\|_V + c_2\|x^{(k)} - x\|_V, \tag{2.4.67}$$

de onde

$$\|x^{(k)} - x\|_V\,(1 - c_2) \leq \|x^{(k)} - x^{(k+1)}\|_V. \tag{2.4.68}$$

Uma vez que $c_2 < 1$, por construção , da desigualdade 2.4.68 obtém-se

$$\|e^{(k)}\|_V = \|x^{(k)} - x\|_V \leq \frac{\|x^{(k)} - x^{(k+1)}\|_V}{1 - c_2}. \qquad (2.4.69)$$

De 2.4.69, utilizando 2.4.65, obtém-se também

$$\|e^{(k+1)}\|_V = \|x^{(k+1)} - x\|_V \leq \frac{c_2}{1 - c_2}\|x^{(k)} - x^{(k+1)}\|_V. \qquad (2.4.70)$$

A desigualdade 2.4.70 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.

Exemplo 2.4.25 . Regressando, mais uma vez, ao sistema linear que foi analisado no exemplo 2.4.21, 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 2.1, enquanto os resultados obtidos para o método de Gauss-Seidel se encontram na tabela 2.2. 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_S, respectivamente (ver exemplos 2.4.20 e 2.4.23). Com base nestes valores, podemos obter estimativas do erro para cada um dos métodos.

Para o método de Jacobi, de acordo om a fórmula 2.4.69, considerando $c_2 = 0.70711$, temos

$$\|e^{(9)}\|_2 \leq \frac{c_2}{1 - c_2}\|x^{(9)} - x^{(8)}\|_2 \leq 0.0242.$$

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

Tabela 2.1: Resultados do método de Jacobi para o exemplo 2.4.25

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

Tabela 2.2: Resultados do método de Gauss-Seidel para o exemplo 2.4.25

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_S 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 2.4.18 e 2.4.24, 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

$$\|e^{(k)}\|_\infty \leq \mu^k \|e^{(0)}\|_\infty, \ k = 1, 2, \ldots \tag{2.4.71}$$

onde $\mu = \|C_J\|_\infty$. Recordando a forma da matriz C_J, dada por 2.4.22, e as definições das grandezas α_i e β_i, dadas por 2.4.53, podemos concluir que

$$\mu = \max_{i=1,\ldots,n} (\alpha_i + \beta_i). \tag{2.4.72}$$

Por outro lado, para o método de Gauss-Seidel, segundo o teorema 2.4.24, é válida a estimativa do erro

$$\|e^{(k)}\|_\infty \leq \eta^k \|e^{(0)}\|_\infty, \ k = 1, 2, \ldots \tag{2.4.73}$$

Para estabelecer uma relação entre a rapidez de convergência dos dois métodos, basta-nos portanto comparar o parâmetro μ da fórmula 2.4.71 com o parâmetro η da fórmula 2.4.73. Atendendo à definição de μ segundo a fórmula 2.4.72, 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

$$\mu \geq \eta, \tag{2.4.74}$$

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 2.4.26. 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 2.4.53, temos

$$\alpha_i = \beta_i = 2/5, \quad i = 1, 2, \dots, n.$$

Daqui, pelas fórmulas 2.4.72 e 2.4.54, obtém-se imediatamente

$$\mu = 4/5, \quad \eta = 2/3.$$

Assim, neste exemplo verifica-se a desigualdade 2.4.74. 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 2.4.10). 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 .

2.4.5 Optimização de Métodos Iterativos

Como vimos nos parágrafos anteriores, os métodos de Jacobi e de Gauss-Seidel não são aplicáveis a qualquer sistema linear e, mesmo quando teoricamente aplicáveis, podem ser pouco eficientes, se a sua convergência for lenta. Neste parágrafo, vamos introduzir novos métodos iterativos, que têm em comum o facto de dependerem de um certo parâmetro, por vezes chamado o parâmetro de *relaxação* . O nosso objectivo é, variando esse parâmetro, obter um método iterativo adequado para cada sistema linear.

Método da Iteração Simples.Este método também é conhecido como método de relaxação linear. Começaremos por reduzir o sistema $A\,x = b$ à forma

$$x = x - \omega(A\,x - b), \tag{2.4.75}$$

onde ω é um número real. Para qualquer $\omega \neq 0$, o sistema 2.4.75 é equivalente ao inicial. Do sistema, escrito nesta nova forma, deduzimos imediatamente a fórmula iteradora

$$x^{(k+1)} = x^{(k)} - \omega(A\,x^{(k)} - b), \quad k = 0, 1, \ldots \tag{2.4.76}$$

Vamos analisar a convergência do método da iteração simples em função do parâmetro ω. Da fórmula 2.4.76 e do sistema 2.4.75 obtém-se a seguinte fórmula para o erro do método:

$$e^{(k+1)} = e^{(k)} - \omega\,A\,e^{(k)} = C(\omega)e^{(k)}, \quad k = 0, 1, \ldots \tag{2.4.77}$$

onde

$$C(\omega) = I - \omega\,A. \tag{2.4.78}$$

Seja $r(\omega)$ o raio espectral de $C(\omega)$. Com base no teorema 2.4.15, podemos podemos afirmar que o método converge, para qualquer aproximação inicial, se e só se for satisfeita a condição $r(\omega) < 1$. Assim, dado um certo sistema linear, o nosso objectivo é averiguar para que valores de de ω é satisfeita esta condição e determinar o valor ω_{opt} que minimiza $r(\omega)$. A este valor corresponde, em princípio, a maior rapidez de convergência. No caso geral, trata-se de um problema complexo, que não pode ser resolvido analiticamente. Na maior parte dos casos, limitamo-nos, portanto, a determinar empiricamente o valor ω_{opt}.Isto significa que se testam diferentes valores de ω, sempre com a mesma aproximação inicial, até se encontrar aquele que nos permite atingir uma dada precisão com um número mínimo de iterações .

Entretanto, nalguns casos particulares, é possível determinar *a priori* os valores admissíveis de ω e o valor ω_{opt}, que minimiza $r(\omega)$. Consideremos, por exemplo, o caso em que *todos os valores próprios de A são reais positivos* . Numeremos esses valores próprios de tal modo que

$$0 < \lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n.$$

Consideremos agora a matriz $C(\omega)$, definida por 2.4.78 . Por construção , os valores próprios desta matriz são

$$\mu_i = 1 - \omega\lambda_i, \quad i = 1, \ldots, n.$$

De acordo com a definição de raio espectral, temos

$$r(\omega) = \max_{i=1,\ldots,n} |\mu_i| = \max_{i=1,\ldots,n} |1 - \omega\,\lambda_i|.$$

Uma vez que todos os valores próprios de A são positivos e estão compreendidos entre λ_1 e λ_n, é evidente que, para valores de ω, positivos e próximos de 0, se verifica

$$0 \leq \mu_n \leq \mu_{n-1} < \ldots \leq \mu_1.$$

Daqui resulta que, para tais valores de ω, temos

$$r(\omega) = \mu_1 = 1 - \omega\,\lambda_1.$$

Por outro lado, para valores suficientemente altos de ω , verifica-se que todos os valores μ_i são negativos e

$$0 \leq -\mu_1 \leq -\mu_2 < \ldots \leq -\mu_n.$$

Por conseguinte, para tais valores de ω, o raio espectral de $C(\omega)$ é dado por

$$r(\omega) = -\mu_n = \omega\,\lambda_n - 1.$$

Daqui, e entrando em conta com o significado geométrico deste problema, que está representado na figura 2.1, podemos concluir que a função $r(\omega)$ tem a seguinte expressão analítica:

$$r(\omega) = \begin{cases} \mu_1 = 1 - \omega\lambda_1, \; se\,\omega \leq \omega_{opt}, \\ -\mu_n = \omega\lambda_n - 1, \; se\,\omega > \omega_{opt}. \end{cases} \tag{2.4.79}$$

Na fórmula 2.4.79, o valor ω_{opt} representa o ponto onde $r(\omega)$ deixa de ser dado pela função $1 - \omega\lambda_1$ (decrescente) e passa a ser dado pela função $\omega\lambda_n - 1$ (crescente). Por conseguinte, este valor é precisamente aquele que minimiza a função r_ω. Este valor determina-se resolvendo a equação

$$1 - \omega_{opt}\lambda_1 = \omega_{opt}\lambda_n - 1,$$

de onde

$$\omega_{opt} = \frac{2}{\lambda_n + \lambda_1}. \tag{2.4.80}$$

Da igualdade **??** resulta imediatamente que o mínimo da função $r(\omega)$ tem o valor

$$r(\omega_{opt}) = \frac{\lambda_n - \lambda_1}{\lambda_n + \lambda_1}. \tag{2.4.81}$$

Para determinar o intervalo de valores admissíveis de ω, isto é, aqueles para os quais $r(\omega) < 1$, temos de resolver a equação

$$\omega_{sup}\lambda_n - 1 = 1,$$

de onde

$$\omega_{sup} = \frac{2}{\lambda_n}. \tag{2.4.82}$$

Conhecido o valor de ω_{sup}, podemos afirmar que *o método da iteração simples converge para a solução do sistema, qualquer que seja a aproximação inicial, desde que o valor de ω pertença ao intervalo* $]0, \omega_{sup}[$. A análise que acabámos de fazer tem uma interpretação geométrica simples que é ilustrada pelo gráfico da fig.**??**

Exemplo 2.4.28.Consideremos o sistema $Ax = b$, onde

$$A = \begin{bmatrix} 2 & 1 & 1 \\ 1 & 3 & 1 \\ 1 & 2 & 2 \end{bmatrix}. \tag{2.4.83}$$

Os valores próprios de A são :

$$\lambda_1 = 1, \quad \lambda_2 = 3 - \sqrt{3} = 1.268, \quad \lambda_3 = 3 + \sqrt{3} = 4.732.$$

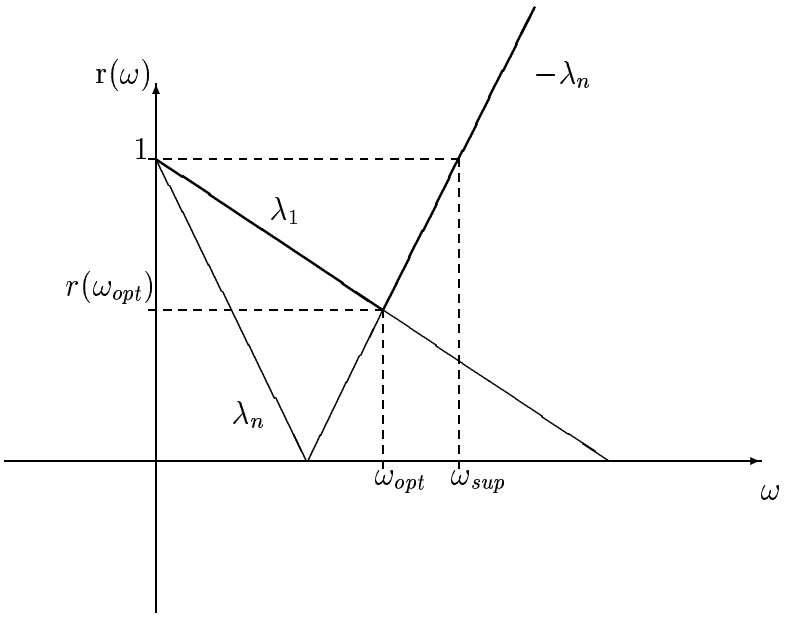

Figura 2.1: Interpretação geométrica da optimização do parâmetro ω para o método da iteração simples.

Uma vez que todos os valores próprios de A são positivos, podemos aplicar a este sistema os resultados que acabámos de obter. De acordo com estes resultados, o método da iteração simples converge para solução do sistema considerado, qualquer que seja a aproximação inicial, desde que o valor de ω pertença ao intervalo $]0, \omega_{sup}[$. Neste caso, o valor de ω_{sup}, dado pela fórmula **??**, é

$$\omega_{sup} = \frac{2}{\lambda_3} = 0.4226. \tag{2.4.84}$$

Além disso, de acordo com a fórmula **??**, o valor de ω que proporciona a convergência mais rápida, neste caso, é

$$\omega_{opt} = \frac{2}{\lambda_3 + \lambda_1} = 0.3489. \tag{2.4.85}$$

O raio espectral de $C(\omega)$, correspondente a ω_{opt}, tem o valor

$$r(\omega_{opt}) = \frac{\lambda_3 - \lambda_1}{\lambda_3 + \lambda_1} = 0.6511. \tag{2.4.86}$$

Método de Jacobi Modificado. Com base no método de Jacobi, é possível definir um novo método iterativo que também pode ser optimizado mediante a escolha adequada de um certo parâmetro. Este método é conhecido como método de Jacobi modificado. Para deduzirmos a respectiva fórmula iteradora, começamos por escrever o sistema $A\,x = b$ sob a forma

$$x = x - \omega\, D^{-1}(A\,x - b), \tag{2.4.87}$$

onde ω é, como antes, um parâmetro real e D representa, como habitualmente, a matriz diagonal $D = diag(a_{11}, a_{22}, \ldots, a_{nn})$. Sempre que $\omega \neq 0$, o sistema **??** é equivalente ao sistema inicial. A partir do sistema, escrito na forma **??**, a fórmula iteradora do método de Jacobi modificado obtém-se pelo processo habitual:

$$x^{(k+1)} = x^{(k)} - \omega\, D^{-1}(A\,x^{(k)} - b),\ k = 0, 1, \ldots \tag{2.4.88}$$

Note-se que, no caso de $\omega = 1$, a fórmula **??** se reduz à fórmula do método de Jacobi habitual. Da fórmula **??** obtêm-se, por componentes, as seguintes

fórmulas:

$$x_i^{(k+1)} \;=\; \frac{1}{a_{ii}}\left(\omega\, b_i \;+(1-\omega)x_i^{(k)} \;-\; \omega \sum_{j=1,j\neq i}^{n} a_{ij}\, x_j^{(k)}\right),\; i=1,\ldots,n;\; k=0,1,\ldots \tag{2.4.89}$$

A fórmula 2.4.20, que relaciona os erros das iteradas do método de Jacobi, pode ser generalizada para o método modificado:

$$e^{(k+1)} \;=\; C_J(\omega)\, e^{(k)}, \tag{2.4.90}$$

onde $C_J(\omega)$ é a seguinte matriz:

$$C_J(\omega) \;=\; \begin{bmatrix} 1-\omega & -\omega\frac{a_{12}}{a_{11}} & \ldots & -\omega\frac{a_{1n}}{a_{11}} \\ -\omega\frac{a_{21}}{a_{22}} & 1-\omega & \ldots & -\omega\frac{a_{2n}}{a_{22}} \\ \ldots & \ldots & \ldots & \ldots \\ -\omega\frac{a_{n1}}{a_{nn}} & -\omega\frac{a_{n2}}{a_{nn}} & \ldots & 1-\omega \end{bmatrix}. \tag{2.4.91}$$

O processo de optimização , neste caso, consiste em determinar o valor de ω que minimiza o raio espectral de $C_J(\omega)$. Nem sempre este valor pode ser determinado analiticamente. Vamos ver num exemplo concreto como pode ser resolvido este problema.

Exemplo 2.4.29. Consideremos de novo o sistema do exemplo 2.4.28. Neste caso, a matriz $C_J(\omega)$ tem a forma

$$C_J(\omega) \;=\; \begin{bmatrix} 1-\omega & -\omega\frac{1}{2} & -\omega\frac{1}{2} \\ -\omega\frac{1}{3} & 1-\omega & -\omega\frac{1}{3} \\ -\omega\frac{1}{2} & -\omega & 1-\omega \end{bmatrix}. \tag{2.4.92}$$

Os valores próprios de $C_J(\omega)$ são

$$\lambda_1 \;=\; \lambda_2 \;=\; 1 \;-\; \frac{\omega}{2},\;\; \lambda_3 \;=\; 1 \;-\; 2\,\omega.$$

Os gráficos destes valores próprios, como funções de ω, estão representados na fig. **??**. Com o auxílio destes gráficos, podemos conluir que a função $r(\omega)$, neste caso, tem a forma:

$$r(\omega) \;=\; \begin{cases} \lambda_1 \;=\; 1 \;-\; \omega/2,\; se\; \omega \leq \omega_{opt}, \\ -\lambda_3 \;= 2\omega - \;1,\; se\; \omega > \omega_{opt}. \end{cases} \tag{2.4.93}$$

O valor de ω_{opt} obtém-se resovendo a equação

$$1 - \omega_{opt} = 2\,\omega_{opt} - 1, \tag{2.4.94}$$

de onde resulta

$$\omega_{opt} = \frac{4}{5}.$$

Como já sabemos, é para este valor de ω que se obtém, em princípio, a convergência mais rápida. O raio espectral de $C_J(\omega)$, neste caso, tem o valor

$$r(\omega_{opt}) = 1 - \frac{\omega_{opt}}{2} = \frac{3}{5}.$$

A convergência do método de Jacobi modificado, neste caso, está garantida para qualquer valor de ω, pertencente ao intervalo $]0, \omega_{sup}[$, sendo o valor de ω_{sup} obtido a partir da equação

$$2\omega_{sup} - 1 = 1.$$

Por conseguinte, $\omega_{sup} = 1$, valor que corresponde ao método de Jacobi habitual. Isto significa que, no caso do sistema considerado, a convergência do método de Jacobi habitual não está garantida para qualquer aproximação inicial.

Método das Relaxações Sucessivas. Este termo é utilizado para designar uma generalização do método de Gauss-Seidel, de aplicação frequente. Representemos por $x^{(k)}$ a k-ésima iterada do método das relaxações sucessivas. A fim de escrever numa forma concisa as fórmulas deste método, vamos introduzir o vector auxiliar $z^{(k+1)}$, que se define do seguinte modo:

$$z^{(k+1)} = D^{-1}(b - T\,x^{(k+1)} - N_S\,x^{(k)}),\ k = 0, 1, \ldots \tag{2.4.95}$$

onde D, T e N_S são as matrizes definidas pelas fórmulas 2.4.41 e 2.4.42. Conhecendo o vector $z^{(k+1)}$, a iterada $x^{(k+1)}$ define-se pela fórmula

$$x^{(k+1)} = \omega\,z^{(k+1)} + (1 - \omega)\,x^{(k)},\ k = 0, 1, \ldots \tag{2.4.96}$$

onde ω representa, como habitualmente, um parâmetro real, que pode ser utilizado para optimizar o método. É evidente que, para $\omega = 1$, este método

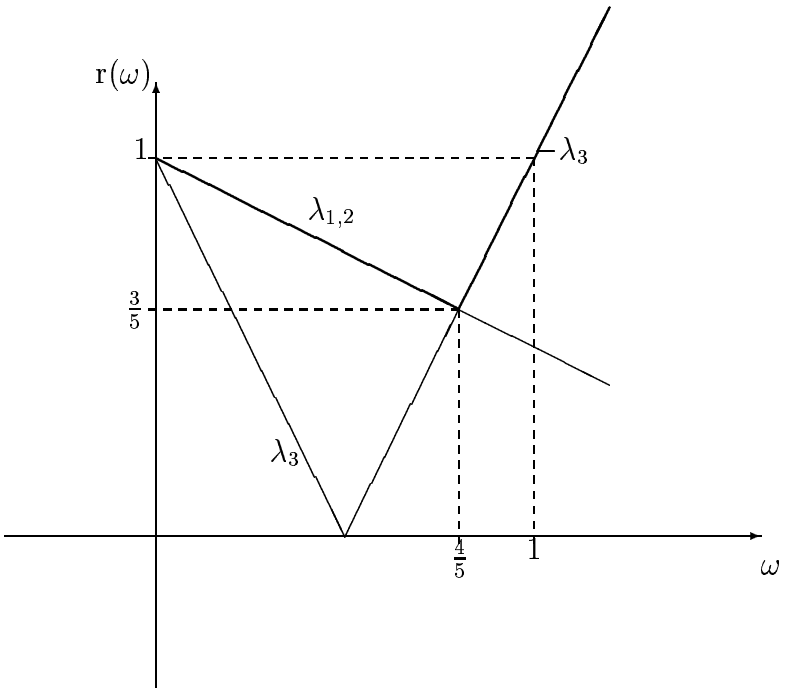

Figura 2.2: Determinação geométrica de $r(\omega)$ para a matriz A do exemplo 2.4.29.

coincide com o de Gauss-Seidel. Vamos deduzir a forma da matriz $C_S(\omega)$, que é uma generalização da matriz C_S para qualquer valor de ω. Substituindo a fórmula **??** em **??**, temos

$$x^{(k+1)} = \omega\, D^{-1}(b - T\, x^{(k+1)} - N_S\, x^{(k)}) + (1-\omega)\, x^{(k)},\ k = 0, 1, \ldots \quad (2.4.97)$$

Da igualdade **??** obtém-se

$$(I + \omega D^{-1}\, T)\, x^{(k+1)} = \omega\, D^{-1} b - \omega\, D^{-1}\, N_S\, x^{(k)} + (1-\omega)\, x^{(k)},\ k = 0, 1, \ldots \quad (2.4.98)$$

Admitindo que a matriz $I + \omega D^{-1}\, T$ é invertível, multiplicamos ambos os membros de **??** à esquerda pela sua inversa e obtemos

$$x^{(k+1)} = (I + \omega D^{-1}\, T)^{-1}(\omega\, D^{-1} b - \omega\, D^{-1}\, N_S\, x^{(k)} + (1-\omega)\, x^{(k)}),\ k = 0, 1, \ldots \quad (2.4.99)$$

Finalmente, podemos escrever a igualdade **??** sob a forma

$$x^{(k+1)} = (I + \omega D^{-1}\, T)^{-1} \omega\, D^{-1} b + C_S(\omega)\, x^{(k)}, \quad (2.4.100)$$

onde

$$C_S(\omega) = (I + \omega D^{-1}\, T)^{-1}(-\omega\, D^{-1}\, N_S + (1-\omega)\, I). \quad (2.4.101)$$

Para optimizar o método das relaxações sucessivas, temos de escolher ω de modo a minimizar o raio espectral de $C_S(\omega)$. De um modo geral, não é possível determinar analiticamente ω_{opt}, uma vez que, mesmo para sistemas pequenos, a determinação de $r(\omega)$ envolve cálculos demasiado complicados. No entanto, são conhecidos alguns factos que facilitam a determinação de ω_{opt} por meios empíricos. Pode-se provar que, para qualquer matriz quadrada A, *uma condição necessária para a convergência do método das relaxações sucessivas é que o valor de ω pertença ao intervalo]0,2[. No caso de a matriz A ser simétrica e ter todos os valores próprios positivos, a mesma condição é necessária e suficiente para a convergência do método.*

Exemplo 2.4.30. A título de exemplo, aplicámos os três métodos iterativos estudados neste parágrafo (iteração simples, Jacobi modificado e relaxações sucessivas) à resolução do sistema $A\, x = b$, onde A é a matriz considerada nos exemplos 2.4.28 e 2.4.29, $b = (4, 5, 5)$. Foi utilizada sempre

a aproximação inicial $x^{(0)} = (0, 0, 0)$, com diferentes valores de ω dentro do intervalo $]0, 2[$. O critério de paragem utilizado foi

$$\|x^{(k+1)} - x^{(k)}\|_2 \leq 10^{-5}.$$

O número de iterações efectuadas em cada caso está indicado na tabela 2.3 (só foram considerados os casos em que o critério de paragem foi alcançado com menos de 200 iterações).

Os resultados obtidos, pelos métodos da iteração simples e de Jacobi modificado, estão de acordo com o estudo teórico efectuado nos exemplos 2.4.28 e 2.4.29. Com efeito, no caso do método de Jacobi modificado, o número mínimo de iterações (26) é obtido com $\omega = 0.8$, o que coincide com o valor de ω_{opt} anteriormente obtido. No caso do método da iteração simples, o número mínimo de iterações registado corresponde a $\omega = 0.3$, o que também está de acordo com o valor de ω_{opt} calculado ($\omega_{opt} = 0.349$). Os intervalos de valores admissíveis de ω também foram confirmados pela experiência. Quanto ao método das relaxações sucessivas, verificou-se que tem um intervalo de valores admissíveis de ω maior que qualquer dos outros dois métodos (aproximadamente $]0, 1.7[$). Em particular, confirmámos que o método de Gauss-Seidel é convergente para este sistema. Além disso, constatou-se que o número mínimo de iterações , entre os valores de ω testados, corresponde a $\omega = 1.1$, tendo-se registado, para este valor de ω, 11 iterações .

ω	(1)	(2)	(3)
0.1	169	175	61
0.2	86	94	31
0.3	56	64	**21**
0.4	38	49	114
0.5	26	39	
0.6	25	33	
0.7	22	28	
0.8	19	**26**	
0.9	16	58	
1.0	13		
1.1	**11**		
1.2	12		
1.3	14		
1.4	17		
1.5	19		
1.6	31		
1.7	64		

Tabela 2.3: Comparação da rapidez de métodos iterativos na resolução do sistema do exemplo 2.4.30. (1) Método das relaxações sucessivas. (2) Método de Jacobi modificado. (3) Método da iteração simples.

2.4.6 Problemas

1. Considere o sistema linear

$$\begin{cases} x + z & = & 2 \\ -x + y & = & 0 \\ x + 2y - 3z & = & 0 \end{cases}.$$

   (a) Prove que o método de Jacobi converge para a solução exacta deste sistema, qualquer que seja a aproximação inicial.

   (b) Mostre que, no caso de se usar o método de Gauss-Seidel, não está garantida a convergência para qualquer aproximação inicial. Indique uma aproximação inicial $x^{(0)}$ (diferente da solução exacta), tal que a sucessão $\{x^{(k)}\}$ seja convergente; e uma aproximação inicial $\tilde{x}^{(0)}$, partindo da qual o método divirja.

2. Consideremos um sistema $A\,x = b$, onde A é uma matriz quadrada de ordem 2 e seja

$$r = \frac{a_{12}\,a_{21}}{a_{11}\,a_{22}}.$$

   (a) Prove que a convergência do método de Jacobi para a solução do sistema está garantida desde que se verifique $|r| < 1$.

   (b) Para que valores de r está garantida a convergência do método de Gauss-Seidel?

   (c) Suponhamos que, para ambos os métodos, a convergência está garantida e que existe o limite

$$c_1 = \lim_{k\to\infty} \frac{\|e^{(k+1)}\|_1}{\|e^{(k)}\|_1}.$$

   Determine c_1, para cada um dos métodos.

   (d) Nas condições da alínea anterior, partindo de uma aproximação inicial arbitrária $x^{(0)}$, quantas iterações é necessário efectuar (utilizando cada um dos métodos) para obter uma aproximação $x^{(k)}$, tal que $\|e^{(k)}\| \leq \epsilon$?

3. Seja A uma matriz triangular inferior, não singular, de ordem n. Pretende-se resolver um certo sistema $A\,x = b$, partindo de uma aproximação inicial arbitrária.

(a) Se aplicarmos o método de Gauss-Seidel, podemos garantir que a solução exacta é obtida com um número finito de iterações .Justifique e diga quantas.

(b) A mesma pergunta, em relação ao método de Jacobi.

4. Seja A uma matriz quadrada real, de ordem 2, tal que os seus valores próprios são complexos:

$$\lambda_{1,2} = a \pm ib.$$

Considerando a solução de um sistema linear com a matriz A, pelo método da iteração simples, determine

(a) o intervalo de valores de ω, para os quais está garantida a convergência do método.

(b) o valor ω_{opt}, para o qual se obtém, em princípio, a maior rapidez de convergência, e o valor correspondente do raio espectral de $C(\omega)$.

5. O método de Jacobi modificado e o das relaxações sucessivas podem ser entendidos como generalizações do método da iteração simples, cujas fórmulas iteradoras têm o aspecto:

$$x^{(k+1)} = x^{(k)} - \omega K(A\,x^{(k)} - b),$$

onde K é uma certa matriz, dependente de A.

(a) Mostre que $K = D^{-1}$, no caso do método de Jacobi modificado.

(b) Deduza a expressão de K, para o método das relaxações sucessivas.

2.5 Métodos Numéricos para o Cálculo de Valores e Vectores Próprios

Os problemas algébricos de valores próprios aparecem frequentemente na matemática aplicada. Em particular, quando se resolvem problemas de valores próprios para equações diferenciais, diferentes tipos de aproximações conduzem, em geral, a problemas algébricos de valores próprios. As dimensões destes são tanto mais elevadas quanto maior for a precisão que se pretende obter na solução do problema inicial. Em §2.1.1, expusemos os principais conceitos relativos ao problema de valores e vectores próprios de matrizes quadradas. No texto que se segue, debruçar-nos-emos sobre alguns aspectos numéricos deste problema.

2.5.1 Localização de Valores Próprios

Consideremos, como habitualmente, uma matriz quadrada A de ordem n. Antes de passarmos ao cálculo numérico dos seus valores próprios, convém fazer uma análise prévia que nos permita ter uma ideia aproximada sobre a natureza e a localização dos mesmos. Recorde-se que, de acordo com a definição de valores próprios (ver §2.1.1), qualquer matriz, real ou complexa, tem n valores próprios em $\mathbb{C}$ (contando com as respectivas multiplicidades). Em particular, se A for uma matriz *hermitiana*, de acordo com o teorema 2.1.8, *estes valores são todos reais*. Além disso, neste caso, sabemos que a matriz A tem n vectores próprios linearmente independentes, que formam uma base ortogonal no espaço $\mathbf{R^n}$. Estes factos, por si só, explicam que as matrizes hermitianas (e, em particular, as reais simétricas) formam uma classe "priveligiada" no que se refere a problemas de valores e vectores próprios. Nos parágrafos que se seguem, veremos que muitos métodos numéricos para cálculo de valores próprios se simplificam bastante quando se trata de matrizes hermitianas, podendo mesmo não ser aplicáveis no caso geral.

O nosso primeiro objectivo é determinar um majorante para o raio espectral da matriz A. Isto significa determinar um círculo no plano complexo que contenha todos os valores próprios de A (ou, no caso de estes serem todos reais, um intervalo que os contenha). De acordo com o teorema 2.1.30, temos

$$r_\sigma(A) \leq \|A\|_M, \tag{2.5.1}$$

onde M representa qualquer norma matricial, associada a uma norma em E^n. Assim, podemos sem grande dificuldade calcular o majorante pretendido. Note-se porém que a majoração 2.5.1 pode ser bastante grosseira, pelo que não podemos tirar dela conclusões válidas sobre a ordem de grandeza dos valores próprios ($\|A\|_M$ pode ser muito maior do que o raio espectral).

O nosso objectivo seguinte, bastante mais ambicioso, é localizar cada um dos valores próprios. Ou seja, para cada valor próprio λ_i determinar um círculo no plano complexo (ou um intervalo real, no caso de matrizes hermitianas) que o contenha. Vamos ver o que nos diz um conhecido teorema sobre esta questão .

Teorema 2.5.1 (Gerschgorin). Seja A uma matriz de ordem n e seja λ um valor próprio de A. Então são válidas as seguintes afirmações :

1. λ pertence, pelo menos, a um dos círculos Z_i, definidos pelas fórmulas:

$$Z_i = \{z \in \mathbb{C} : |z - a_{ii}| \leq r_i\}, \tag{2.5.2}$$

onde

$$r_i = \sum_{j=1, j\neq i}^{n} |a_{ij}|\,, \; i = 1, 2, \ldots, n. \tag{2.5.3}$$

2. Se a reunião de m daqueles círculos formar um conjunto conexo S, disjunto dos restantes $m - n$ círculos, então S contém exactamente m valores próprios de A (contando com as respectivas multiplicidades algébricas).

Demonstração . Seja λ um valor próprio de A e seja x um vector próprio associado a λ. Seja k o índice da componente de x para a qual

$$|x_k| = \max_{1\leq i\leq n} |x_i| = \|x\|_\infty.$$

De acordo com a definição de vector próprio, temos

$$(A\,x)_k = \sum_{j=1}^{n} a_{kj}\,x_j = \lambda\,x_k,$$

de onde resulta

$$(\lambda - a_{kk})x_k = \sum_{j=1, j\neq k}^{n} a_{kj}\, x_j,$$

$$|\lambda - a_{kk}||x_k| \leq \sum_{j=1, j\neq k}^{n} |a_{kj}|\, |x_j| \leq r_k \, \|x\|_\infty. \tag{2.5.4}$$

Dividindo ambos os membros da igualdade 2.5.4 por $\|x\|_\infty$, obtém-se a primeira afirmação do teorema.

Para demonstrar a segunda segunda afirmação consideremos a família de matrizes

$$A(\epsilon) = D + \epsilon E,$$

onde $D = diag(a_{11}, a_{22}, \ldots, a_{nn})$, $E = A - D$, $0 \leq \epsilon \leq 1$. Representemos os valores próprios de $A(\epsilon)$ por $\lambda_i(\epsilon), i = 1, \ldots, n$. Uma vez que $\lambda_i(\epsilon)$ são as raizes do polinómio característico de $A(\epsilon)$, cujos coeficientes são funções contínuas de ϵ, estes valores próprios também são funções contínuas de ϵ. Para cada uma dessas funções verifica-se $\lambda_i(0) = a_{ii}$, $\lambda_i(1) = \lambda_i$. Representemos por $Z_i(\epsilon)$ os círculos

$$Z_i(\epsilon) = \{z \in \mathbb{C} : |z - a_{ii}| \leq \epsilon\, r_i\},$$

onde r_i é definido por 2.5.3. De acordo com a primeira parte do teorema, para cada $\epsilon > 0$ verifica-se $\lambda_i(\epsilon) \in Z_i(\epsilon), i = 1, \ldots, n$. Por outro lado, se para um certo i, $Z_i(\epsilon)$ não intersectar nenhum círculo $Z_j(\epsilon), j \neq i$, então os gráficos das funções correspondentes aos restantes valores próprios não intersectam $Z_i(\epsilon)$, pelo que $Z_i(\epsilon)$ não pode conter outros valores próprios, além de $\lambda_i(\epsilon)$. No caso de m círculos se intersectarem, formando um conjunto conexo S, então esse conjunto contém exactamente m valores próprios. Esta afirmação mantém-se válida para qualquer valor de ϵ, no intervalo $]0, 1]$. No caso de $\epsilon = 1$, obtém-se a segunda afirmação do teorema. □

Corolário. Se, na fórmula 2.4.29, somarmos ao longo de cada coluna de A (e não ao longo de cada linha), obtemos as grandezas:

$$r'_j = \sum_{i=1, i\neq j}^{n} |a_{ij}|\,, \; j = 1, 2, \ldots, n. \tag{2.5.5}$$

Os coeficientes r_i' estão para a matriz A^T como os coeficientes r_i estão para A. Logo, as afirmações do teorema de Gerschgorin aplicam-se a A^T, com os valores de r_i substituídos pelos de r_i'. Mas como os valores próprios de A^T são os mesmos que os de A, podemos concluir que as afirmações do teorema de Gerschgorin são válidas, quer com os valores de r_i, quer com os de r_i'. Podemos assim aplicar o teorema nas duas formas e combinar os resultados de modo a obter uma localização mais precisa dos valores próprios.

Exemplo 2.5.2. Sendo dada a matriz

$$A = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 2 & 1/2 & 0 \\ 0 & 1 & 6 & 1 \\ 0 & 0 & 2 & 7 \end{bmatrix},$$

vamos localizar os seus valores próprios, segundo o teorema de Gerschgorin. Neste caso, os círculos Z_i são dados pelas fórmulas (ver figura 2.3):

$$\begin{aligned} Z_1 &= \{z \in \mathbb{C} : |z - 1| \leq 1\}, \\ Z_2 &= \{z \in \mathbb{C} : |z - 2| \leq 3/2\}, \\ Z_3 &= \{z \in \mathbb{C} : |z - 6| \leq 2\}, \\ Z_4 &= \{z \in \mathbb{C} : |z - 7| \leq 2\}. \end{aligned}$$

Os círculos Z_1 e Z_2 intersectam-se, formando o conjunto $S_1 = Z_1 \bigcup Z_2$. Por seu lado, os círculos Z_3 e Z_4 intersectam-se entre si mas não intersectam Z_1 nem Z_2. Se representarmos a sua reunião por S_2, estamos na presença de dois conjuntos S_1 e S_2, cada um dos quais é conexo. Logo, de acordo com o teorema de Gerschgorin, temos dois valores próprios de A em S_1 e outros dois em S_2.

Para podermos aplicar também o corolário do teorema de Gerschgorin, calculemos os valores de r_i', dados por 2.5.5:

$$r_1' = r_1 = 1, \quad r_2' = 2, \quad r_3' = 5/2, \quad r_4' = 1.$$

Neste caso, a reunião dos conjuntos Z_i', $i = 1, \ldots, 4$ forma um conjunto conexo S, ao qual pertencem todos os valores próprios de A. Combinando as afirmações do teorema de Gerschgorin e do corolário, podemos afirmar que *a matriz A tem 2 valores próprios em S_1 (que está contido em S) e outros dois em $S_2 \bigcap S$* .

Figura 2.3: Interpretação geométrica do teorema de Gerschgorin para o exemplo 2.5.2.

2.5.2 Condicionamento do Problema de Valores Próprios

Já vimos em §2.2 que pequenos erros relativos na matriz de um sistema linear podem dar origem grandes erros relativos na sua solução . Nesse caso, dizíamos que o sistema linear era mal condicionado. O condicionamento do problema de valores próprios define-se de modo parecido, mas considerando os erros em termos absolutos. Seja A uma matriz quadrada de ordem n, cujos valores próprios são $\lambda_1, \ldots, \lambda_n$. Suponhamos que A é substituída por uma certa matriz A', tal que $A' = A + E$, a que chamaremos a *matriz perturbada*. Pressupõe-se que a matriz E, que representa a *perturbação de* A, tem a norma pequena, em comparação com a de A (pode tratar-se, por exemplo, dos erros de arredondamento que afectam as componentes de A). Seja λ um valor próprio de A'. Naturalmente,diremos que o problema de valores próprios é bem condicionado, se existir, pelo menos, um valor próprio λ_i de A que seja próximo de λ.Neste sentido, é natural relacionar o condicionamento dos valores próprios com o mínimo das distâncias de λ aos valores próprios de A. Mais precisamente, o problema de valores próprios para A diz-se *bem condicionado* se, a pequenos valores de $\|E\|$ corresponderem pequenos valores de $\min_{1\leq i\leq n} |\lambda_i - \lambda|$. O exemplo que a seguir apresentamos mostra que nem sempre assim acontece, ou seja, que, embora a perturbação seja pequena, os valores próprios de A' podem ser bastante distantes dos de A.

Exemplo 2.5.3. Consideremos as matrizes

$$A = \begin{bmatrix} 101 & -90 \\ 110 & -98 \end{bmatrix}, \; A' = \begin{bmatrix} 101 - \epsilon & -90 - \epsilon \\ 110 & -98 \end{bmatrix}.$$

Partindo do princípio que ϵ tem um valor próximo de 0, a norma de E é pequena em comparação com a de A, mais precisamente, $\|E\|_1 = \epsilon$. Verifica-se, por cálculo directo, que os valores próprios de A são

$$\lambda_1 = 1, \;\; \lambda_2 = 2,$$

enquanto que os de A' são

$$\lambda'_{1,2} = \frac{1}{2}\left(3 - \epsilon \pm \sqrt{1 - 828\epsilon + \epsilon^2}\right). \tag{2.5.6}$$

Desenvolvendo o segundo membro de 2.5.6 em série de Taylor, quando $\epsilon \to 0$, obtêm-se as seguintes fórmulas aproximadas:

$$\begin{array}{l} \lambda_1' \approx 1 - \epsilon/2 + 207\epsilon + O(\epsilon^2), \\ \lambda_2' \approx 2 - \epsilon/2 - 207\epsilon + O(\epsilon^2). \end{array} \tag{2.5.7}$$

Vemos assim que

$$\begin{array}{l} \min_{i=1,2} |\lambda_i - \lambda_1'| = |\lambda_1 - \lambda_1'| \approx 206.5\epsilon, \\ \min_{i=1,2} |\lambda_i - \lambda_2'| = |\lambda_2 - \lambda_2'| \approx 207.5\epsilon. \end{array}$$

Ou seja, um pequeno erro nas componentes de A pode gerar erros muito maiores nos seus valores próprios. A razão do aumento, neste caso, atinge mais de 200 vezes, em termos absolutos, pelo que se pode dizer que *o problema de valores próprios para esta matriz é mal condicionado.*

Quando se trata de aplicar métodos numéricos para a determinação de valores e vectores próprios de uma matriz, é importante averiguar previamente o condicionamento do problema, pois, se o problema for mal condicionado, até os erros de arredondamento podem provocar alterações significativas dos resultados numéricos. A seguir, formulamos um dos mais conhecidos teoremas sobre o condicionamento do problema de valores próprios.

Teorema 2.5.4 (Bauer-Fike). Suponhamos que A é uma matriz que tem a forma canónica de Jordan diagonal, isto é, existe uma matriz não singular P tal que

$$P^{-1} A P = diag(\lambda_1, \lambda_2, \ldots, \lambda_n) = D.$$

Seja E uma perturbação de A e λ um valor próprio da matriz perturbada $A' = A + E$. Então , para qualquer norma matricial M (associada a uma certa norma vectorial) verifica-se

$$\min_{1 \le i \le n} |\lambda - \lambda_i| \le \|P\|_M \, \|P^{-1}\|_M \, \|E\|_M. \tag{2.5.8}$$

A demonstração deste teorema encontra-se em [1]. □

Uma consequência imediata do teorema anterior consiste em que, se A for uma matriz *hermitiana*, para qualquer valor próprio λ de A' verifica-se

$$\min_{1\leq i\leq n} |\lambda - \lambda_i| \leq \|E\|_2. \tag{2.5.9}$$

Esta majoração resulta da desigualdade 2.5.8, atendendo a que, quando A é hermitiana, P é unitária (ver teorema 2.1.8) e, por conseguinte, $\|P\|_2 = 1$. Isto permite-nos afirmar que *o problema de valores próprios é bem condicionado para qualquer matriz hermitiana*. No caso geral, o condicionamento do problema de valores próprios, sempre que se verifiquem as condições do teorema 2.5.3, depende da norma de P.

Definição 2.5.5. Seja A uma matriz que satisfaz as condições do teorema 2.5.4 e M uma norma matricial. Chama-se *número de condição do problema de valores próprios para* A (na norma M) à grandeza

$$K(A) = \|P\|_M \|P^{-1}\|_M.$$

Exemplo 2.5.6. Consideremos de novo a matriz A do exemplo 2.5.3. O teorema 2.5.4 pode ser aplicado a esta matriz, uma vez que todos os seus valores próprios são reais e distintos. Para o aplicarmos, precisamos de conhecer a matriz P, a qual é constituída pelos vectores próprios de A. Sabendo que $\lambda_1 = 1$, obtém-se facilmente um vector próprio associado a λ_1, com a forma $v_1 = (1, 10/9)$. Para o valor próprio $\lambda_2 = 2$, um vector próprio associado será $v_2 = (1, 11/10)$. Assim, a matriz P tem, neste caso, a forma

$$P = \begin{bmatrix} 1 & 1 \\ 10/9 & 11/10 \end{bmatrix}.$$

Para averiguar o condicionamento dos valores próprios de A falta-nos determinar P^{-1}:

$$P^{-1} = \begin{bmatrix} -99 & 90 \\ 100 & -90 \end{bmatrix}.$$

Assim, se considerarmos, por exemplo, a norma por colunas, temos

$$K(A) = \|P\|_1 \|P^{-1}\|_1 = 417.9.$$

Se considerarmos a perturbação de A introduzida no exemplo 2.5.3, temos

$$\|E\|_1 = \epsilon.$$

Asssim, segundo o teorema 2.5.3,

$$\min_{i=1,2} |\lambda - \lambda_i| \leq \|P\|_1 \|P^{-1}\|_1 \|E\|_1 = 417.9\,\epsilon. \tag{2.5.10}$$

Esta majoração está de acordo com os resultados obtidos no exemplo 2.5.3.

2.5.3 Método das Potências

Um dos métodos mais simples para a determinação de valores e vectores próprios de matrizes é o *método das potências*. Este método, no entanto, só nos permite determinar directamente um dos valores próprios da matriz (o de maior módulo), pelo que nem sempre é indicado. Seja A uma matriz real quadrada de ordem n, cuja forma canónica de Jordan é diagonal. Representemos por λ_i os valores próprios de A, ordenando-os de tal forma que

$$|\lambda_1| \geq |\lambda_2| \geq \ldots \geq |\lambda_n|.$$

Representaremos por x_i um vector próprio associado ao valor próprio λ_i, i=1,. . . ,n.

Definição 2.5.7.Dizemos que λ_1 é o valor próprio dominante de A se λ_1 for o único valor para o qual se verifica

$$|\lambda_1| > |\lambda_i|, \; i = 2, \ldots, n.$$

(Admite-se o caso de o valor próprio dominante ser múltiplo).

Note-se que uma matriz pode não ter nenhum valor próprio dominante, no caso de se verificar $|\lambda_1| = |\lambda_2|$, mas $\lambda_1 \neq \lambda_2$. Neste parágrafo, consideraremos apenas matrizes com um valor próprio dominante λ_1. Este valor próprio será sempre real, pois doutro modo o seu complexo conjugado também seria valor próprio de A e λ_1 não seria dominante. Se a matriz não tiver um valor próprio dominante, então o método das potências não é, de um modo geral, aplicável.

O método das potências é um método iterativo, segundo o qual se constroi uma sucessão de vectores $\{z^{(k)}\}_{k=0}^{\infty}$ que converge para o vector próprio x_1. Paralelamente, determina-se uma sucessão numérica $\{\lambda_1^{(k)}\}_{k=0}^{\infty}$, que converge para λ_1. Assim, começa-se por escolher um vector $z^{(0)}$ (em geral, aleatório) que constitui a aproximação inicial do vector próprio. Este vector deve apenas satisfazer a condição

$$\|z^{(0)}\|_\infty = \max_{i=1,\ldots,n} |z_i^{(0)}| = 1.$$

Para construir cada aproximação $z^{(k)}$, determina-se previamente o vector auxiliar $w^{(k)}$, através da fórmula

$$w^{(k)} = A\, z^{(k-1)}. \tag{2.5.11}$$

Conhecido o vector $w^{(k)}$, a iterada $z^{(k)}$ define-se do seguinte modo:

$$z^{(k)} = \frac{w^{(k)}}{\alpha_k}, \tag{2.5.12}$$

onde α_k é a componente de $w^{(k)}$ que tem maior módulo (se houver mais do que uma componente nessas condições escolhe-se a que tiver menor índice). As relações de recorrência 2.5.11 e 2.5.12 definem o esquema iterativo do método das potências. A fórmula 2.5.12 garante que, para qualquer n, se tem $\|z^{(k)}\| = 1$. Verifiquemos que a sucessão $\{z^{(k)}\}_{k=0}^{\infty}$ converge para um vector colinear com x_1.

Começamos por demonstrar, por indução , que, para um termo genérico da sucessão , se verifica

$$z^{(k)} = \sigma_k \frac{A^k z^{(0)}}{\|A^k z^{(0)}\|_\infty}, \; k = 1, 2, \ldots, \tag{2.5.13}$$

onde σ_k é um factor de sinal, isto é, só pode ter o valor 1 ou -1.

Vejamos o caso $k = 1$. De acordo com as fórmulas 2.5.11 e 2.5.12,

$$z^{(1)} = \frac{w^{(1)}}{\alpha_1} = \sigma_1 \frac{A z^{(0)}}{\|A z^{(0)}\|_\infty}. \tag{2.5.14}$$

Admitamos agora que a fórmula 2.5.13 é válida para $k = m \geq 1$. Então podemos escrever

$$z^{(m+1)} = \frac{w^{(m+1)}}{\alpha_{m+1}} = \mu \frac{A z^{(m)}}{\|A z^{(m)}\|_\infty} = \mu\sigma_m \frac{A^{m+1} z^{(0)}}{\|A^{m+1} z^{(0)}\|_\infty}, \tag{2.5.15}$$

onde $\mu = \pm 1$. Considerando $\sigma_{m+1} = \mu\sigma_m$, a igualdade 2.5.15 é a aplicação da fórmula 2.5.13 no caso $k = m + 1$. Assim, a fórmula 2.5.13 fica demonstrada por indução . □

Vamos agora utilizar a fórmula 2.5.13 para demonstrar a convergência do método das potências. Uma vez que, por hipótese, os vectores próprios de A formam uma base em $\mathbf{R^n}$, podemos escrever

$$z^{(0)} = \sum_{j=1}^{n} \alpha_j x_j. \tag{2.5.16}$$

Vamos admitir que o coeficiente α_1 deste desenvolvimento é diferente de zero, o que, em geral, se verifica, se $z^{(0)}$ for aleatório. Utilizando a igualdade 2.5.16, obtém-se

$$\begin{aligned} A^k\, z^{(0)} &= \sum_{j=1}^{n} \alpha_j\, A^k\, x_j = \\ &= \sum_{j=1}^{n} \alpha_j\, \lambda_j^k\, x_j = \\ &= \lambda_1^k \left[\alpha_1\, x_1 + \sum_{j=2}^{n} \alpha_j\, (\tfrac{\lambda_j}{\lambda_1})^k\, x_j\right]. \end{aligned} \tag{2.5.17}$$

Atendendo a que λ_1 é o valor próprio dominante, temos $\frac{|\lambda_j|}{|\lambda_1|} < 1$, j=2,...,n. Por conseguinte, $\left(\frac{|\lambda_j|}{|\lambda_1|}\right)^k \to 0$, quando $k \to \infty,\ j = 2, \ldots, n$. De 2.5.17 e 2.5.13 conclui-se então que

$$z^{(k)} \approx \sigma_k \frac{\lambda_1^k\, \alpha_1\, x_1}{|\lambda_1|^k\, \alpha_1\, \|x_1\|_\infty} = \hat{\sigma_k}\, \frac{x_1}{\|x_1\|_\infty}, \tag{2.5.18}$$

onde $\hat{\sigma_k} = \pm 1$. Se o vector x_1 tiver uma única componente de módulo máximo, então, a partir de um certo k, teremos $\hat{\sigma_k} = const.$ No caso de haver mais do que uma componente de x_1 com o módulo máximo, então o sinal $\hat{\sigma_k}$ depende da iterada. Em qualquer dos casos, podemos sempre assegurar que o método das potências converge, verificando-se

$$\lim_{k\to\infty} \hat{\sigma_k} z^{(k)} = x_1', \tag{2.5.19}$$

onde x_1' representa o vector próprio normalizado: $x_1' = \frac{x_1}{\|x_1\|_\infty}$. Para obtermos uma estimativa da norma do erro de $z^{(k)}$, podemos utilizar a igualdade 2.5.17. Considerando no segundo membro de 2.5.17 o primeiro termo do somatório e ignorando os restantes (que convergem mais rapidamente para 0), obtém-se

$$\|z^{(k)} - \hat{\sigma_k}\, x_1'\|_\infty \leq C \left|\frac{\lambda_2}{\lambda_1}\right|^k, \tag{2.5.20}$$

onde C é uma constante real (não depende de k). Concluímos, portanto, que *a sucessão* $\{\hat{\sigma_k}\, z^{(k)}\}_{k=0}^{\infty}$ *converge linearmente para* x_1'.

Entretanto, o método das potências permite-nos também construir uma sucessão numérica λ_1^k que converge para λ_1. Com esse fim defina-se

$$\lambda_1^{(k)} = \frac{w_m^{(k)}}{z_m^{(k-1)}},\ k = 1, 2, \ldots \tag{2.5.21}$$

onde m é o índice de uma das componentes (em geral, escolhe-se a componente de maior módulo de $z^{(k-1)}$, para garantir que o denominador não se anula). A fim de obtermos uma estimativa do erro de $\lambda_1^{(k)}$ procede-se do mesmo modo que no caso de $z^{(k)}$.Transformando a igualdade 2.5.21 de acordo com 2.5.11 e 2.5.13, temos

$$\lambda_1^{(k)} = \frac{(A\, z^{(k-1)})_m}{z_m^{(k-1)}} = \frac{(A^k\, z^{(0)})_m}{(A^{k-1}\, z^{(0)})_m}. \tag{2.5.22}$$

Aplicando a fórmula 2.5.17 no numerador e no denominador de 2.5.22, obtém-se

$$\lambda_1^{(k)} = \frac{\lambda_1^k \left[\alpha_1\, x_1 + \sum_{j=2}^n \alpha_j (\lambda_j/\lambda_1)^k\right]_m}{\lambda_1^{k-1} \left[\alpha_1\, x_1 + \sum_{j=2}^n \alpha_j (\lambda_j/\lambda_1)^k\right]_m}, \tag{2.5.23}$$

de onde se conclui que

$$\lambda_1^{(k)} = \lambda_1 \left[1 + O\left(\left|\frac{\lambda_2}{\lambda_1}\right|^k\right)\right]. \tag{2.5.24}$$

Por conseguinte, a sucessão $\{\lambda_1^{(k)}\}_{k=0}^{\infty}$ também converge linearmente para λ_1.

Nas fórmulas 2.5.20 e 2.5.24 verifica-se que o factor assimptótico de convergência, em ambos os casos, é $k_\infty = \frac{|\lambda_2|}{|\lambda_1|}$, o que significa que a convergência do método das potências é tanto mais rápida quanto mais pequeno for este quociente. Uma vez que, em regra, o valor de λ_2 não é conhecido, não é possível determinar, a priori, o factor assimptótico de convergência. No entanto, se forem conhecidas três aproximações sucessivas de λ_1, nomeadamente $\lambda_1^{(k-1)}$, $\lambda_1^{(k)}$ e $\lambda_1^{(k+1)}$, é possível obter uma estimativa de k_∞, utilizando uma fórmula análoga à que se aplica no caso de métodos iterativos para sistemas lineares (ver §2.4.4). Assim, se for conhecido que $\lim_{k\to\infty} \lambda_1^{(k)} = \lambda_1$, define-se $R^{(k)}$ pela fórmula

$$R^{(k)} = \frac{\lambda_1^{(k+1)} - \lambda_1^{(k)}}{\lambda_1^{(k)} - \lambda_1^{(k-1)}}, \quad k = 2, 3, \ldots. \tag{2.5.25}$$

Uma vez que $\lambda_1^{(k)}$ converge linearmente para λ_1, podemos afirmar que o módulo de $R^{(k)}$ converge para o factor assimptótico k_∞. Assim, a partir de uma certa ordem k, os valores de $|R^{(k)}|$ dão -nos boas aproximações de k_∞.

Estas aproximações podem ser utilizadas, em particular, para obter estimativas do erro, à semelhança do que fizemos no caso dos métodos iterativos para sistemas lineares. Mais precisamente, repetindo as operações que efectuámos nas fórmulas 2.4.66 a 2.4.70, obtêm-se as seguintes estimativas do erro:

$$|\lambda_1^{(k)} - \lambda_1| \leq \frac{1}{1 - k_\infty} |\lambda_1^{(k+1)} - \lambda_1^{(k)}|, \tag{2.5.26}$$

$$|\lambda_1^{(k+1)} - \lambda_1| \leq \frac{k_\infty}{1 - k_\infty} |\lambda_1^{(k+1)} - \lambda_1^{(k)}|. \tag{2.5.27}$$

Assim, conhecendo apenas a diferença entre as duas últimas iteradas e um valor aproximado de k_∞, obtido com base em $R^{(k)}$, podemos majorar o erro de qualquer aproximação de λ_1, obtida com base no método das potências.

Exemplo 2.5.8 . Consideremos a matriz real

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

Os seus valores próprios podem ser determinados directamente: $\lambda_1 = 6$, $\lambda_2 = 1$. Utilizámos este exemplo simples para ver como funciona o método das potências. Partindo da aproximação inicial $z^{(0)} = (1, 0)$, obtivemos os valores de $w^{(k)}$,$z^{(k)}$ e $\lambda_1^{(k)}$, de acordo com as fórmulas 2.5.11, 2.5.12 e 2.5.21, respectivamente. Os valores obtidos estão representados na tabela 2.4. Nesta tabela, além das sucessivas aproximações do vector próprio x_1, são apresentadas as aproximações $\lambda_1^{(k)}$ do valor próprio dominante e, a partir de $k = 2$, os valores do quociente $R^{(k)}$, definido pela fórmula 2.5.25. Com base na fórmula 2.5.27, obtém-se a seguinte estimativa do erro de $\lambda_1^{(6)}$:

$$|\lambda_1^{(6)} - \lambda_1| \leq \frac{0.166}{1 - 0.166} |\lambda_1^{(6)} - \lambda_1^{(5)}| = 0.000641. \tag{2.5.28}$$

Comparando o valor aproximado $\lambda_1^{(6)}$ com o valor exacto de λ_1, verifica-se que a estimativa obtida está muito próxima do valor real do erro. Isto está relacionado com o facto de utilizarmos uma boa aproximação para k_∞. Com efeito, o valor $R^{(6)}$, utilizado na estimativa do erro, é muito próximo do valor exacto de k_∞ que, para este exemplo, é $k_\infty = \frac{\lambda_2}{\lambda_1} = \frac{1}{6}$.

k	$w_1^{(k)}$	$w_2^{(k)}$	$z_1^{(k)}$	$z_2^{(k)}$	$\lambda_1^{(k)}$	$R^{(k)}$
1	2	-2	1	-1	2	
2	4	-7	-0.571429	1	7	
3	-3.14286	6.14286	-0.511628	1	6.14286	0.136
4	-3.02326	6.02326	-0.501931	1	6.02326	0.1622
5	-3.00386	6.00386	-0.500321	1	6.00386	0.1626
6	-3.00064	6.00064	-0.500053	1	6.00064	0.1656

Tabela 2.4: Resultados da aproximação numérica, pelo método das potências, do valor próprio dominante e de um vector próprio associado (exemplo 2.5.8)
.

2.5.4 Aceleração de Convergência

Quando se utilizam métodos iterativos para aproximar as raizes de uma certa equação (em paticular, os valores próprios de uma matriz) assumem grande importância as chamadas técnicas de aceleração de convergência, que nos permitem melhorar significativamente a precisão dos resultados, sem que isso implique um grande aumento do volume dos cálculos.

A ideia fundamental da *aceleração de convergência*, que hoje em dia constitui um domínio em expansão da análise numérica, consiste em transformar uma sucessão inicial dada (em geral, aproximações sucessivas da grandeza que se pretende calcular) noutra sucessão, que converge para o mesmo limite, mas mais rapidamente.

Por exemplo, suponhamos que é dada uma certa sucessão de números reais $\{x^{(n)}\}_{n\in N}$, da qual se sabe que converge linearmente para um certo $x \in \mathbf{R}$. A partir dessa sucessão , pode obter-se uma nova sucessão $\{y^{(n)}\}_{n\in N}$, que converge também para x, mas com uma ordem de convergência igual a 2 (quadrática). Uma transformação de sucessões deste tipo chama-se um *método de extrapolação* . Um dos métodos de extrapolação mais conhecidos, que se aplica a sucessões de congergência linear, é a *extrapolação de Aitken.*

A extrapolação de Aitken é aplicável sempre que os termos da sucessão inicial satisfazem a igualdade aproximada

$$x^{(n+1)} - x \approx k_\infty (x^{(n)} - x), \tag{2.5.29}$$

onde x é o limite e k_∞ é o factor assimptótico de convergência da sucessão. No caso de utilizarmos o método das potências, as aproximações sucessivas do valor próprio dominante, como vimos no parágrafo anterior, convergem linearmente e satisfazem uma igualdade aproximada, idêntica a 2.5.29. Vimos também que, no caso da convergência linear, o factor assimptótico de convergência pode ser estimado, uma vez conhecidos três termos consecutivos da sucessão, pela seguinte igualdade aproximada:

$$k_\infty \approx \frac{x^{(n+1)} - x^{(n)}}{x^{(n)} - x^{(n-1)}}. \tag{2.5.30}$$

Substituindo 2.5.30 em 2.5.29 obtém-se

$$x^{(n+1)} - x \approx \frac{x^{(n+1)} - x^{(n)}}{x^{(n)} - x^{(n-1)}} (x^{(n)} - x). \tag{2.5.31}$$

Note-se que, se a igualdade 2.5.31 se verificar exactamente para uma certa sucessão (como acontece para as progressões geométricas), dela pode obter-se imediatamente o limite dessa sucessão, resolvendo-a como uma equação em ordem a x. Nesse caso obteremos

$$x = x^{(n+1)} - \frac{(x^{(n+1)} - x^{(n)})^2}{(x^{(n+1)} - x^{(n)}) - (x^{(n)} - x^{(n-1)})}, \quad n = 1, 2, \ldots \qquad (2.5.32)$$

Na maioria dos casos que nos interessam, a igualdade 2.5.31 é apenas aproximada, pelo que dela apenas podemos obter um valor aproximado de x. Representando esse valor por $y^{(n+1)}$, temos

$$y^{(n+1)} = x^{(n+1)} - \frac{(x^{(n+1)} - x^{(n)})^2}{(x^{(n+1)} - x^{(n)}) - (x^{(n)} - x^{(n-1)})}, \quad n = 1, 2, \ldots \quad (2.5.33)$$

A extrapolação de Aitken consiste em substituir a sucessão inicial $\{x^{(n)}\}_{n\in N}$ pela nova sucessão $\{y^{(n)}\}_{n\in N}$, cujos termos são obtidos pela fórmula 2.5.33. Este método de extrapolação é estudado com profundidade em [4]. Em particular, está provado que, se a sucessão inicial for convergente, a nova sucessão converge para o mesmo limite e que, se a convergência da sucessão inicicial for linear, então a convergência de $\{y^{(n)}\}_{n\in N}$ é pelo menos, quadrática.

Daqui resulta o grande interesse prático da extrapolação de Aitken, quando se utilizam métodos iterativos de convergência linear, como é o caso do método das potências. No exemplo que se segue, vamos ver como a extrapolação de Aitken permite acelerar a convergência do método das potências, aplicado a um caso concreto.

Exemplo 2.5.9. Voltando à matriz do exemplo 2.5.8, consideremos a sucessão $\{\lambda_1^{(k)}\}_{k\in N}$ que aproxima o valor próprio dominante daquela matriz. Na tabela 2.5 estão repressentados vários termos da sucessão $\lambda_1^{(k)}$, bem como os respectivos erros. Além disso, nesta tabela estão representados os termos correspondentes da sucessão $\mu^{(k)}$, obtidos mediante a extrapolação de Aitken, seguidos dos seus erros. Pela análise desta tabela verifica-se que a extrapolação de Aitken permite obter resultados muito mais precisos, com um custo relativamente baixo em termos de cálculo.

k	$\lambda_1^{(k)}$	$\lvert\lambda_1 - \lambda_1^{(k)}\rvert$	$\mu^{(k)}$	$\lvert\lambda_1 - \mu^{(k)}\rvert$
2	7	1		
3	6.14286	0.14286		
4	6.02326	0.02326	6.00387	0.00387
5	6.00386	0.00386	6.00010	0.00010
6	6.00064	0.00064	6.00000	< 0.000005

Tabela 2.5: Aceleração da convergência do método das potências através da extrapolação de Aitken.

Capítulo 3

Equações e Sistemas Não-Lineares

Neste capítulo, trataremos de métodos iterativos para a determinação de raizes de equações e sistemas não-lineares. Naturalmente, o cálculo destas raizes levanta problemas mais complexos que a resolução dos sistemas lineares, estudados no capítulo anterior. A própria existência de raizes reais de uma equação não-linear é, por vezes, difícil de averiguar. A determinação destas raizes através de fórmulas explícitas é, na grande maioria dos casos, impossível.

Tomemos como exemplo uma equação algébrica de grau n, cuja forma geral é:

$$a_n\, x^n \,+\, a_{n-1}\, x^{n-1} + \,\ldots\, + a_1\, x \,+\, a_0 \,=\, 0.$$

Embora seja conhecido que uma equação deste tipo tem n raizes complexas, nem sempre é fácil determinar quantas dessas raizes são reais e, no caso de $n > 4$, não existem fórmulas resolventes para as calcular.

Se considerarmos o caso mais geral de uma equação com a forma

$$f(x) \,=\, 0,$$

onde f é uma certa função real, de variável real, não é simples, na maioria dos casos, saber se a equação tem raizes reais ou quantas tem. Assim, quer para averiguar a existéncia de soluções de uma equação ou sistema não-linear, quer para aproximar essas soluções , somos levados a fazer um estudo prévio da função f, aplicando conhecimentos de análise matemática. Um dos

instrumentos de análise mais importantes a que vamos recorrer é o *teorema do ponto fixo de Banach*, de que falaremos no próximo parágrafo.

3.1 Método do Ponto Fixo

Ao estudarmos os métodos iterativos para sistemas lineares, vimos que uma das maneiras de resolvermos um sistema consistia em reduzi-lo à forma

$$x = g + C\,x, \tag{3.1.1}$$

onde $g \in \mathbf{R^n}$,$C \in L^n$. Este procedimento pode ser generalizado para equações ou sistemas não-lineares, ficando neste caso a equaçao (ou o sistema) reduzido à forma

$$x = G(x), \tag{3.1.2}$$

onde G é uma função, definida num certo conjunto X.

Definição 3.1.1. Seja X um conjunto num certo espaço normado E e G uma aplicação de X em E. Então $x \in X$ diz-se um ponto fixo de G se e só se $x = G(x)$.

Assim, calcular as raizes da equação (3.1.2) equivale a determinar os pontos fixos de G.

Um método iterativo muito simples, utilizado para aproximar o(s) ponto(s) fixos de uma dada função G é o chamado *método do ponto fixo*, que consiste no seguinte. Escolhendo como aproximação inicial um certo $x^{(0)} \in X$, constroi-se uma sucessão $\{x^{(k)}\}_{k\in\mathbf{N}}$, tal que

$$x^{(k+1)} = G(x^{(k)}), \quad k = 0, 1, \ldots \tag{3.1.3}$$

Se a função G, a que chamaremos *função iteradora*, for contínua em X, verifica-se facilmente que, caso a sucessão considerada convirja, ela só pode convergir para um ponto fixo de G. Com efeito, admitamos que a sucessão converge e representemos o seu limite por z. Então podemos escrever

$$G(z) = G\left(\lim_{k\to\infty} x^{(k)}\right) = \lim_{k\to\infty} G(x^{(k)}) = \lim_{k\to\infty} x^{(k+1)} = z, \tag{3.1.4}$$

donde resulta que z é um ponto fixo de G.

No entanto, o método do ponto fixo nem sempre converge. A sua convergência depende da função iteradora e da aproximação inicial escolhida. Vejamos um exemplo simples.

$x^{(0)}$	-2	-1	2
$x^{(1)}$	-1.8646	-1.6321	5.3891
$x^{(2)}$	-1.8451	-1.8045	216.997
$x^{(3)}$	-1.8420	-1.8355	1.7×10^{94}
$x^{(4)}$	-1.8415	-1.8405	
$x^{(5)}$	-1.8414	-1.8413	
$x^{(6)}$	-1.8414	-1.8414	
$x^{(7)}$	-1.8414	-1.8414	

Tabela 3.1: Sucessões obtidas pelo método do ponto fixo para o exemplo 3.1.2

Exemplo 3.1.2. Seja $E = \mathbf{R}$. Consideremos a equação

$$f(x) = e^x - x - 2 = 0. \tag{3.1.5}$$

Esta equação também pode ser escrita na forma

$$x = e^x - 2. \tag{3.1.6}$$

Neste caso, determinar as raizes da equação (3.1.5) equivale a calcular os pontos fixos da função $G(x) = e^x - 2$. Uma vez que $G(-1) < -1$ e $G(-2) > -2$, pelo teorema do valor intermédio conclui-se que esta função tem um ponto fixo no intervalo $]-2,-1[$. Designemo-lo z_1. De modo análogo se verifica que ela tem um ponto fixo no intervalo $]1,2[$, que designaremos z_2. Por outro lado, pelo teorema de Rolle a equação (3.1.5) não pode ter mais que duas raizes, visto que $f'(x) = e^x - 1$ só se anula em $x = 0$. Logo a função G, neste caso, tem exactamente dois pontos fixos, que são as raizes da equação (3.1.5). Tentemos obter aproximações de z_1 e z_2 através do método do ponto fixo. Se tomarmos como aproximação inicial $x^{(0)} = -2$, obtemos a sucessão que está representada na primeira coluna da tabela (3.1). Se tomarmos como aproximação inicial $x^{(0)} = -1$, obtemos a sucessão que se encontra na segunda coluna. Em ambos os casos, os números obtidos indicam que as sucessões convergem para o mesmo limite, aproximadamente igual a -1.8414. De acordo com o que acima dissemos, este é o valor do ponto fixo z_1. Se, no entanto, tomarmos como aproximação inicial $x^{(0)} = 2$, obtemos uma

sucessão divergente. Se tentarmos outras aproximações iniciais, verificaremos que as respectivas sucessões ou convergem para z_1, ou divergem. Em nenhum caso se obtém uma sucessão que convirja para z_2.

A fim de investigar as condições que garantem a convergência do método do ponto fixo, vamos recorrer a um teorema importante da análise, conhecido como o *teorema do ponto fixo*. Antes de o podermos formular, necessitamos de introduzir algumas definições.

Definição 3.1.3. Seja E um espaço normado, $X \subset E$ e G uma função de X em E. A função G diz-se lipschitziana em X, se e só se existir uma constante q, tal que

$$\|G(x_1) - G(x_2)\| \leq q\,\|x_1 - x_2\|, \quad \forall x_1, x_2 \in X. \tag{3.1.7}$$

Ao ínfimo de todas as constantes q, para as quais a desigualdade (3.1.7) se verifica, chama-se *constante de Lipschitz* de G em X e representa-se por $L_{G,X}$.

Definição 3.1.4. Diz-se que G é uma contracção (ou uma função contractiva) em X se e só se G for lipschitziana em X e $L_{G,X} < 1$.

Exemplo 3.1.5. Seja $E = \mathbf{R}$ e $G(x) = x^2$. Verifiquemos para que valores de r a função G é contractiva em $X = [-r, r]$. Temos

$$|G(x_1) - G(x_2)| = |x_1^2 - x_2^2| = |x_1 - x_2|.|x_1 + x_2|. \tag{3.1.8}$$

Se x_1 e x_2 pertencerem a X, podemos escrever

$$|x_1 + x_2| \leq r + r = 2r. \tag{3.1.9}$$

Substituindo (3.1.9) em (3.1.8), obtém-se

$$|G(x_1) - G(x_2)| \leq 2r\,|x_1 - x_2|, \tag{3.1.10}$$

donde se conclui que G é lipschitziana em X, com a constante $L_{G,X} = 2r$. Por conseguinte, se $r < 1/2$, podemos afirmar que G é contractiva em X.

No caso de funções de uma variável real, a condição de contractividade pode ser expressa noutros termos, tornando-se mais fácil a sua verificação.

Teorema 3.1.6. Seja G uma função com domínio em $X = [a, b]$ e valores reais e suponhamos que $G \in C^1([a, b])$. Então G é contractiva em X se e só se

$$\max_{x \in [a,b]} |G'(x)| < 1. \tag{3.1.11}$$

Demonstração. Pelo teorema de Lagrange, quaisquer que sejam x_1 e x_2, pertencentes a $[a, b]$, existe $\xi \in [x_1, x_2]$, tal que

$$|G(x_1) - G(x_2)| = |G'(\xi)||x_1 - x_2|. \tag{3.1.12}$$

Então podemos afirmar que a constante de Lipschitz de G é

$$L_G = \max_{x \in [a,b]} |G'(x)| < 1, \tag{3.1.13}$$

donde se conclui que G é contractiva em $[a, b]$.

Demonstremos agora a inversa. Suponhamos que existe em $[a, b]$ um certo y, tal que $|G'(y)| \geq 1$. Entretanto, pelo teorema de Lagrange, para qualquer $h > 0$, existe $\theta \in [0, 1]$ tal que

$$|G(y + h) - G(y)| = |G'(y + \theta h)|\, h. \tag{3.1.14}$$

Visto que G' é contínua em $[a, b]$,para qualquer $\rho < 1$, existe h_0 tal que $|G'(y + \theta h_0)| > \rho$. Escrevendo a desigualdade (3.1.14) com $h = h_0$, obtém-se

$$|G(y + h_0) - G(y)| = |G'(y + \theta h_0)|\, h_0 > \rho\, h_0. \tag{3.1.15}$$

A desigualdade (3.1.15) implica que G não é contractiva em $[a, b]$, ficando assim demonstrado o teorema. □

No caso mais geral em que G é uma função de $\mathbf{R^n}$ em $\mathbf{R^n}$, com derivadas parciais contínuas, a contractividade pode ser verificada através da sua matriz jacobiana.

Definição 3.1.7. Seja G uma função vectorial, tal que

$$G(x) = (G_1(x), G_2(x), \ldots, G_n(x)),$$

onde G_i é uma função escalar com domínio em $X \subset \mathbf{R^n}$. Se em X existirem as derivadas parciais $\frac{\partial G_i}{\partial x_j}, i, j \in \{1, ..., n\}$ chama-se *jacobiana de* G e representa-se por J_G a matriz

$$J_G(x) = \begin{bmatrix} \frac{\partial G_1}{\partial x_1} & \cdots & \frac{\partial G_1}{\partial x_n} \\ \frac{\partial G_2}{\partial x_1} & \cdots & \frac{\partial G_2}{\partial x_n} \\ \cdots & \cdots & \cdots \\ \frac{\partial G_n}{\partial x_1} & \cdots & \frac{\partial G_n}{\partial x_n} \end{bmatrix}. \tag{3.1.16}$$

Teorema 3.1.8. Seja X um conjunto convexo [1] em $\mathbf{R^n}$ e G uma função vector com domínio em X e valores em $\mathbf{R^n}$. Então, se as derivadas parciais $\frac{\partial G_i}{\partial x_j}, i, j \in \{1, ..., n\}$ forem contínuas em X e se

$$sup_{x \in X} \|J_G(x)\|_\infty < 1, \tag{3.1.17}$$

G é contractiva em X (segundo a norma do máximo).

Demonstração . Sejam x_1 e x_2 dois elementos de X. Segundo o teorema de Lagrange para funções de n variáveis, para cada função G_i existe um ponto ξ_i, pertencente ao segmento $[x_1, x_2]$, tal que

$$G_i(x_1) - G_i(x_2) = (\nabla G_i(\xi_i), x_1 - x_2), \tag{3.1.18}$$

onde

$$\nabla G_i(x) = \left(\frac{\partial G_i}{\partial x_1}, \ldots, \frac{\partial G_i}{\partial x_n}\right), \quad i \in \{1, \ldots, n\}, \tag{3.1.19}$$

e os parêntesis curvos no segundo membro de (3.1.18) denotam o produto interno de vectores. Note-se que todos os pontos ξ_i pertencem a X, uma vez

[1] Um conjunto X diz-se convexo se, para quaisquer x_1, x_2 pertencentes a X, todos os pontos do segmento $[x_1, x_2]$ também pertencerem a X.

que este conjunto é convexo. De (3.1.18) e (3.1.19) obtém-se

$$|G_i(x_1) - G_i(x_2)| \leq \max_{j=1,\ldots,n} |x_{1,j} - x_{2,j}| \sum_{j=1}^{n} \left|\frac{\partial G_i}{\partial x_j}(\xi_i)\right| = \|\nabla G_i(\xi_i)\|_1 \, \|x_1 - x_2\|_\infty, \quad i \in \{1, \ldots, n\}. \tag{3.1.20}$$

Seja i' o índice para o qual se verifica

$$|G_{i'}(x_1) - G_{i'}(x_2)| = \|G(x_1) - G(x_2)\|_\infty.$$

No caso de $i = i'$, a desigualdade (3.1.20) toma o aspecto

$$\|G(x_1) - G(x_2)\|_\infty \leq \|\nabla G_{i'}(\xi_{i'})\|_1 \, \|x_1 - x_2\|_\infty. \tag{3.1.21}$$

Atendendo a que

$$\|\nabla G_{i'}(\xi_{i'})\|_1 \leq \max_{1 \leq i \leq n} \sum_{j=1}^{n} \left|\frac{\partial G_i}{\partial x_j}(\xi_{i'})\right| = \|J_G(\xi_{i'})\|_\infty < 1, \tag{3.1.22}$$

de (3.1.21) resulta que G é contractiva em X, segundo a norma do máximo. □

Nalguns casos, pode ser mais cómodo considerar em $\mathbf{R^n}$ outras normas que não a do máximo, por exemplo, a norma $\|.\|_1$. Por isso enunciamos aqui o seguinte teorema, análogo a 3.1.8.

Teorema 3.1.9. Seja X um conjunto convexo em $\mathbf{R^n}$ e G uma função vectorial com domínio em X e valores em $\mathbf{R^n}$. Então , se as derivadas parciais $\frac{\partial G_i}{\partial x_j}, i, j \in \{1, ..., n\}$ forem contínuas em X e se

$$sup_{x \in X} \|J_G(x)\|_1 < 1, \tag{3.1.23}$$

G é contractiva em X (segundo a norma $\|\|_1$).

Para não nos tornarmos repetitivos, não apresentamos a demonstração deste teorema, que pode ser obtida por um raciocínio semelhante ao do teorema anterior.

Estamos agora em condições de formular o teorema do ponto fixo, para espaços normados de dimensão finita, por exemplo, os espaços $\mathbf{R^n}$.

Teorema 3.1.10 (*Teorema do ponto fixo*). Seja E um espaço normado de dimensão finita e X um sub-conjunto fechado de E. Seja G uma função contractiva em X, tal que $G(X) \subset X$. Então são verdadeiras as afirmações

1. G tem um único ponto fixo z em X.
2. Se $\{x^{(k)}\}_{k\in\mathbf{N}}$ for uma sucessão de termos em E tal que $x^{(0)} \in X$ e $x^{(k+1)} = G(x^{(k)}), \forall k \in \mathbf{N}$, então $x^{(k)} \to z$.
3. Se G satisfaz em X a desigualdade (3.1.7) com a constante $q < 1$, então são válidas as desigualdades

$$\|x^{(n+1)} - z\| \leq q\|x^{(n)} - z\|, \quad \forall n \in \mathbf{N}. \tag{3.1.24}$$

e

$$\|x^{(m)} - z\| \leq \frac{q^m}{1-q}\|x^{(1)} - x^{(0)}\|, \forall m \in \mathbf{N}. \tag{3.1.25}$$

Demonstração . Em primeiro lugar, note-se que se $x^{(0)} \in X$, então $x^{(k)} \in X$, $\forall k$, visto que $G(X) \subset X$. Comecemos por provar que a sucessão de que se fala no ponto 2 é convergente. Para tanto, basta provar que ela é uma sucessão de Cauchy. Uma vez que G é contractiva em X, existe uma constante $q < 1$, tal que

$$\|G(x_1) - G(x_2)\| \leq q\,\|x_1 - x_2\|, \quad \forall x_1, x_2 \in X. \tag{3.1.26}$$

Em particular, se tivermos dois termos consecutivos da sucessão considerada verifica-se

$$\|x^{(k+1)} - x^{(k)}\| \leq q\,\|x^{(k)} - x^{(k-1)}\|, \quad \forall k \in \mathbf{N}. \tag{3.1.27}$$

Sejam agora $x^{(m)}$ e $x^{(n)}$ dois termos quaisquer da sucessão , com $n > m$. Então podemos escrever

$$\begin{aligned}\|x^{(n)} - x^{(m)}\| = \|x^{(n)} - x^{(n-1)} + x^{(n-1)} - x^{(n-2)} + \ldots + x^{(m+1)} - x^{(m)}\| \leq \\ \|x^{(n)} - x^{(n-1)}\| + \|x^{(n-1)} - x^{(n-2)}\| + \ldots + \|x^{(m+1)} - x^{(m)}\|.\end{aligned} \tag{3.1.28}$$

Das desigualdades (3.1.27) e (3.1.28) obtém-se

$$\|x^{(n)} - x^{(m)}\| \leq (q^{n-m-1} + \ldots + q + 1)\,\|x^{(m+1)} - x^{(m)}\| \leq q^m\,(q^{n-m-1} + \ldots + q + 1))\,\|x^{(1)} - x^{(0)}\|. \tag{3.1.29}$$

A soma que figura no segundo membro de (3.1.29) é a soma de uma progressão geométrica de razão q. Como $q < 1$, podemos escrever

$$q^m \sum_{k=0}^{n-m-1} q^k < \frac{q^m}{1-q}, \quad \forall n \in \mathbf{N}. \tag{3.1.30}$$

Substituindo (3.1.30) em (3.1.29), obtém-se

$$\|x^{(m)} - x^{(n)}\| < \frac{q^m}{1-q}\|x^{(1)} - x^{(0)}\|, \quad \forall n > m. \tag{3.1.31}$$

Da desigualdade (3.1.31) resulta que $\forall \epsilon > 0$, existe $n_0 \in \mathbf{N}$ tal que

$$\|x^{(m)} - x^{(n)}\| < \epsilon, \quad \forall m, n > n_0. \tag{3.1.32}$$

Daqui resulta, por definição que a sucessão considerada é uma sucessão de Cauchy e, logo, converge. Representemos por z o seu limite. Uma vez que X é fechado, $z \in X$. Provemos agora que z é um ponto fixo de G. Utilizando o facto de G ser contractiva, podemos escrever

$$\|x^{(m+1)} - G(z)\| = \|G(x^{(m)}) - G(z)\| \leq q\|x^{(m)} - z\|, \quad \forall m. \tag{3.1.33}$$

Logo $\|x^{(m+1)} - G(z)\| \to 0$, de onde resulta que $x^{(m)} \to G(z)$, quando $m \to \infty$. Por conseguinte, $G(z) = z$, tal como se pretendia demonstrar. Fica assim demonstrado o ponto 2 do teorema. A desigualdade (3.1.24), por sua vez, resulta de (3.1.33). Quanto à desigualdade (3.1.25), ela obtém-se de (3.1.31), se fizermos n tender para infinito. Resta-nos provar que z é o único ponto fixo de G em X. Suponhamos que existem dois pontos fixos de G em X e representemo-los por z e z'. Uma vez que G é contractiva, temos

$$\|G(z') - G(z)\| = \|z' - z\| \leq q\|z' - z\|, \tag{3.1.34}$$

donde

$$\|z' - z\|\,(1 - q) \leq 0. \tag{3.1.35}$$

Uma vez que $1 - q > 0$, de (3.1.35) resulta que $z' = z$. $\square$

Exemplo 3.1.11. Consideremos de novo a equação do exemplo 3.1.2:

$$e^x - x - 2 = 0. \tag{3.1.36}$$

Vimos que esta equação tem uma e só uma raiz no intervalo $X = [1, 2]$, a qual representámos por z_1. A equação (3.1.36) pode ser escrita na forma $G(x) = x$, onde $G(x) = e^x - 2$. Verifiquemos se a função G satisfaz em X as condições do teorema do ponto fixo. Uma vez que G é diferenciável e se verifica

$$\max_{x \in X} |G'(x)| \ = \ e^{-1} \ < \ 1, \tag{3.1.37}$$

Podemos afirmar, de acordo com o teorema 3.1.8, que G é contractiva em X e a sua constante de Lipschitz é$L_{G,X} = e^{-1}$. Resta-nos verificar que $G(X) \subset X$. Neste caso, uma vez que G é cresente em X, basta-nos verificar os valorees da função nos extremos do intervalo. Assim temos

$$G(-1) = e^{-1} - 2 = -1.632; \quad G(-2) = e^{-2} - 2 = -1.865. \tag{3.1.38}$$

Uma vez que $G(-1) \in X$ e $G(-2) \in X$, podemos concluir que $G(X) \subset X$. Logo, podemos aplicar o teorema do ponto fixo, o qual nos diz que a função G tem um único ponto fixo em X.Por construção , esse ponto fixo é a raiz z_1 da equação (3.1.36).Segundo o mesmo teorema, o método do ponto fixo vai convergir para z_1, qualquer que seja $x^{(0)} \in X$. Em particular, as sucessões que estão representados nas primeiras duas colunas da tabela (3.1) convergem para z_1. Com base na desigualdade (3.1.25), podemos obter a seguinte estimativa do erro:

$$|x^{(m)} - z_1| \ \leq \ \frac{e^{-m}}{1 - e^{-1}} |x^{(1)} - x^{(0)}|, \forall m \in \mathbf{N}. \tag{3.1.39}$$

Em particular, no caso de $x^{(0)} = -2$ e $m = 4$, obtém-se

$$|x^{(4)} - z| \ \leq \ 0.003923; \tag{3.1.40}$$

no caso de $x^{(0)} = -1$ e $m = 4$, obtém-se

$$|x^{(4)} - z| \ \leq \ 0.0183. \tag{3.1.41}$$

No que diz respeito à raiz z_2 , situada no intervalo $[1,2]$, é fácil verificar que a função G considerada não é contractiva em nenhuma vizinhança desta raiz. Com efeito, verifica-se

$$\min_{x\in[1,2]} |G'(x)| \;=\; e^1 \;>\; 1. \tag{3.1.42}$$

Neste caso, podemos afirmar que o método do ponto fixo, com a função iteradora considerada, não converge para z_2, qualquer que seja a aproximação inicial.

Exemplo 3.1.12. Consideremos o sistema de duas equações

$$\left\{ \begin{array}{rcrcl} 3x_1 & + & x_2^2 & = & 0 \\ x_1^2 & + & 3x_2 & = & 1 \end{array} \right. \tag{3.1.43}$$

Vamos utilizar o teorema do ponto fixo para provar que este sistema tem uma única raiz no conjunto

$$X \;=\; \{(x_1,x_2)\in \mathbf{R}^2 : -1/3 \le x_1 \le 0;\; 0 \le x_2 \le 1/3\} \tag{3.1.44}$$

Com esse fim, o sistema (3.1.43) pode ser reescrito na forma $x = G(x)$, onde

$$\begin{array}{rcl} G_1(x_1,x_2) & = & -\frac{x_2^2}{3} \\ G_2(x_1,x_2) & = & \frac{1-x_1^2}{3}. \end{array} \tag{3.1.45}$$

Verifiquemos se a função G, definida por (3.1.45), satisfaz as condições do teorema do ponto fixo em X. Em primeiro lugar, constata-se que o conjunto X é um quadrado, contendo a sua fronteira, pelo que é convexo e fechado. Além disso, vemos que as derivadas parciais de G_1 e G_2 são contínuas em X, de tal modo que a jacobiana de G tem a forma

$$J_G(x_1,x_2) \;=\; \left[\begin{array}{cc} 0 & -\frac{2x_2}{3} \\ -\frac{2x_1}{3} & 0 \end{array} \right] \tag{3.1.46}$$

Assim, podemos escrever

$$\|J_G(x_1,x_2)\|_\infty \;=\; max\left(\frac{2|x_2|}{3}, \frac{2|x_1|}{3}\right) \tag{3.1.47}$$

de onde resulta que

$$\|J_G(x_1, x_2)\|_\infty \leq \frac{2}{9} < 1 \ \forall (x_1, x_2) \in X. \tag{3.1.48}$$

Com base no teorema 1.8 podemos, portanto, afirmar que G é contractiva em X (segundo a norma do máximo) com a constante $q = \frac{2}{9}$. Para podermos aplicar o teorema do ponto fixo, precisamoas também de verificar que $G(X) \subset X$.

Seja $x = (x_1, x_2) \in X$. Então

$$\begin{array}{rcl} G_1(x_1, x_2) & = & -\frac{x_2^2}{3} \in [-1/3, 0] \\ G_2(x_1, x_2) & = & \frac{1 - x_1^2}{3} \in [0, 1/3]. \end{array} \tag{3.1.49}$$

Por conseguinte, temos $(G_1(x_1, x_2), G_2(x_1, x_2)) \in X$, de onde se conclui que $G(X) \subset X$.

Logo, a função G satisfaz as condições do teorema do ponto fixo em X. Assim, podemos garantir que aquela função tem um único ponto fixo em X, que, por construção , será a única raiz do sistema (3.1.43) no referido conjunto.

Podemos também aplicar o método do ponto fixo para aproximar a raiz considerada. Tomemos como aproximação inicial qualquer ponto do conjunto X, por exemplo, a origem das coordenadas: $x^{(0)} = (0, 0)$. Então obtêm-se as seguintes aproximações sucessivas:

$$\begin{array}{rcl} x_1^{(1)} & = & G_1(0, 0) = 0 \\ x_2^{(1)} & = & G_2(0, 0) = \frac{1}{3} \end{array}, \tag{3.1.50}$$

$$\begin{array}{rcl} x_1^{(2)} & = & G_1(0, 1/3) = -\frac{1}{27} \\ x_2^{(2)} & = & G_2(0, 1/3) = \frac{1}{3} \end{array}. \tag{3.1.51}$$

Tentemos agora obter uma estimativa do erro da iterada $x^{(2)}$. De acordo com a desigualdade (3.1.25), podemos escrever

$$\|x^{(2)} - z\|_\infty \leq \frac{q^2}{1 - q} \|x^{(1)} - x^{(0)}\|_\infty, \tag{3.1.52}$$

onde $q = \frac{2}{9}$. Neste caso, temos $\|x^{(1)} - x^{(0)}\|_\infty = 1/3$; logo

$$\|x^{(2)} - z\|_\infty \leq \frac{4}{63}\frac{1}{3} = \frac{4}{189}. \tag{3.1.53}$$

Esta última estimativa poderia ser refinada se, em vez da desigualdade (3.1.25), aplicássemos a desigualdade

$$\|x^{(m+1)} - z\|_\infty \leq \frac{q}{1-q}\|x^{(m+1)} - x^{(m)}\|_\infty, \tag{3.1.54}$$

que também se pode deduzir do teorema do ponto fixo. Nesse caso obteríamos

$$\|x^{(2)} - z\|_\infty \leq \frac{q}{1-q}\|x^{(2)} - x^{(1)}\|_\infty = \frac{2}{189}. \tag{3.1.55}$$

Bibliografia

[1] **K.E.Atkinson**. *An Introduction to Numerical Analysis. John Whiley & Sons, 1989.*

[2] **O.A. Axelsson and V.A.Barker**.*Finite Element Solution of Boundary-Value Problems.* Academic Press, 1985.

[3] **N.S. Bakhvalov** .*Numerical Methods.* Moscow, Mir, 1977.

[4] **C.Brezinski and M.Zaglia**.*Extrapolation Methods, Theory and Practice*, Elsevier Science, 1991.

[5] **M.J. Carpentier**.*Análise Numérica.* IST,1993.

[6] **P.Henrici**.*Applied and Computational Complex Analysis*,Vol. I, New York, Wiley,1974.

[7] **L.C.Loura** . *Análise Numérica I — apontamentos da cadeira.* IST, 1992.

[8] **L.T. Magalhães**. *Álgebra Linear.* I.S.T., Lisboa, 1987.

[9] **J. Ortega**.*Numerical Analysis - A Second Course.* Academic Press, New York and London, 1972.