Matemática Computacional
Cap. 3 - Resolução Numérica de Sistemas de Equações

Sumário

1 Normas matriciais

No que se segue, $\mathbb{K}$ designa um corpo de escalares, que será sempre $\mathbb{R}$ ou $\mathbb{C}$, e E designa um espaço vetorial sobre $\mathbb{K}$. Recordamos:

Definição 1.1. *Uma função $\|\cdot\| : E \to \mathbb{R}$ diz-se uma* norma *sobre E se*

(i) $\|x\| \geq 0$, $\forall x \in E$,

(ii) $\|x\| = 0 \Leftrightarrow x = 0$,

(iii) $\|\alpha x\| = |\alpha|\|x\|$, $\forall \alpha \in \mathbb{K}$, $\forall x \in E$,

(iv) $\|x + y\| \leq \|x\| + \|y\|$, $\forall x, y \in E$.

Um espaço vetorial onde está definida uma norma diz-se um espaço normado*, sendo habitual escrever $(E, \|\cdot\|)$.*

Estaremos particularmente interessados nos casos $E = \mathbb{R}^N$ e $E = \mathbb{R}^{N\times N}$, os mais relevantes para o estudo de sistemas de equações.

Exemplo 1.1. *As seguintes funções são normas sobre $\mathbb{K}^N$*

(i) Normas−p, para $p \geq 1$

$$\|x\|_p := \left(\sum_{i=1}^{N} |x_i|^p\right)^{\frac{1}{p}}.$$

A norma−2, ou norma euclidiana, $\|x\|_2 := \left(\sum_{i=1}^{N} |x_i|^2\right)^{\frac{1}{2}}$ está associada ao produto interno $(x, y) := \sum_{i=1}^{N} x_i\overline{y_i} = x \cdot \overline{y}$, ou seja, $\|x\|^2 = (x, x)$. É claro que quando $\mathbb{K} = \mathbb{R}$, fica simplesmente $(x, y) := \sum_{i=1}^{N} x_i y_i = x \cdot y$. É válida a desigualdade de Cauchy-Schwarz

$$|(x, y)| \leq \|x\|_2\|y\|_2 \quad (x, y \in \mathbb{K}^N).$$

(i) Norma do máximo

$$\|x\|_\infty := \max_{1\leq i\leq N} |x_i|.$$

Exemplo 1.2. *Seja $\|\cdot\|$ uma norma arbitrária em $\mathbb{K}^N$ e $M \in \mathbb{C}^{N\times N}$ uma matriz não-singular arbitrária. Então*

$$x \mapsto \|Mx\|$$

define uma norma em $\mathbb{K}^N$.

Num espaço normado $(E, \|\cdot\|)$, as noções topológicas elementares de ponto interior, exterior ou fronteiro, conjunto limitado, aberto, fechado, vizinhança de um ponto, etc, surgem como generalização das mesmas noções em $\mathbb{R}$, após se introduzir a bola aberta

$$B(a, R) := \{x \in E;\ \|x - a\| < R\}$$

e a bola fechada

$$\overline{B(a, R)} := \{x \in E;\ \|x - a\| \leq R\}.$$

Definição 1.2. *Diz-se que duas normas sobre E, $\|\cdot\|$ e $\|\cdot\|_*$, são equivalentes se*

$$\exists \underline{C}, \overline{C} > 0 \ :\ \underline{C}\|x\|_* \leq \|x\| \leq \overline{C}\|x\|_*,\ \forall x \in E.$$

Teorema 1.1. *Num espaço de dimensão finita, todas as normas são equivalentes.*

Demonstração. Ver [11]. □

Exemplo 1.3. *São válidas:*

1. $\|x\|_\infty \leq \|x\|_p \leq \sqrt[p]{N}\|x\|_\infty$, $\forall x \in \mathbb{K}^N$, *para qualquer* $p \geq 1$;
2. $\|x\|_2 \leq \|x\|_1 \leq \sqrt{N}\|x\|_2$, $\forall x \in \mathbb{K}^N$.

No contexto da Análise Numérica, é fundamental a noção de convergência. A convergência de uma sucessão de elementos do espaço pode ser estudada através da convergência de uma sucessão de números reais:

Definição 1.3. *Seja $(E, \|\cdot\|)$ um espaço normado. Sejam $\{x^{(n)}\}_{n\in\mathbb{N}}$ uma sucessão de elementos de E e $x \in E$. Diz-se que a sucessão converge para x, e escrevemos $x^{(n)} \to x$, se*

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

ou seja, se

$$\forall \varepsilon > 0\ \exists p \in \mathbb{N} :\ n > p \Rightarrow \|x^{(n)} - x\| < \varepsilon.$$

Atendendo ao Teorema 1.1, quando referirmos a convergência de uma sucessão num espaço de dimensão finita, não é necessário indicar nenhuma norma em particular.

Ainda no contexto da Análise Numérica, é fundamental estimar os erros das aproximações calculadas. Assim:

Definição 1.4. *Seja $(E, \|\cdot\|)$ um espaço normado e sejam $x, \tilde{x} \in E$. Se $x \approx \tilde{x}$, define-se*
Erro de $\tilde{x}$ em relação a x: $e_{\tilde{x}} = x - \tilde{x}$;
Erro absoluto de $\tilde{x}$: $\|e_{\tilde{x}}\| = \|x - \tilde{x}\|$;
Erro relativo de $\tilde{x}$: $\|\delta_{\tilde{x}}\| = \|x - \tilde{x}\|/\|x\|$ $(x \neq 0)$;
Erro relativo percentual de $\tilde{x}$: $100\%\|\delta_{\tilde{x}}\|$ $(x \neq 0)$.

Vamos agora introduzir as normas matriciais. Começamos por uma norma que surge naturalmente se identificarmos $\mathbb{K}^{N\times N}$ com $\mathbb{K}^{N^2}$, a *norma de Frobenius*, ou *norma de Schur* por estarmos a considerar matrizes quadradas:

$$\|A\|_{Fb} = \left(\sum_{i,j=1}^{N} |a_{ij}|^2\right)^{\frac{1}{2}} \quad (A \in \mathbb{K}^{N\times N}).$$

Definição 1.5. *Uma norma $\|\cdot\|_M$ em $\mathbb{K}^{N\times N}$ diz-se* compatível *com a norma vetorial $\|\cdot\|_V$ em $\mathbb{K}^N$ se*

$$\|Ax\|_V \leq \|A\|_M \|x\|_V, \, \forall A \in \mathbb{K}^{N\times N}, \, \forall x \in \mathbb{K}^N.$$

Exemplo 1.4. *A norma de Frobenius é compatível com a norma-2:*

$$\|Ax\|_2^2 = \sum_{j=1}^{N} |(Ax)_j|^2 = \sum_{j=1}^{N} |(Ax, e_j)|^2 = \sum_{j=1}^{N} |(e_j^{\mathsf{T}} A, x)|^2 \leq \sum_{j=1}^{N} \|e_j^{\mathsf{T}} A\|_2^2 \|x\|_2^2$$

pela desigualdade de Cauchy-Schwarz, e

$$\sum_{j=1}^{N} \|e_j^{\mathsf{T}} A\|_2^2 \|x\|_2^2 \leq \|x\|_2^2 \sum_{j=1}^{N} \left(\sum_{i=1}^{N} |(e_j^{\mathsf{T}} A)_i|^2\right) = \|x\|_2^2 \sum_{i,j=1}^{N} |a_{ij}|^2 = \|x\|_2^2 \|A\|_{Fb}^2,$$

para qualquer $A \in \mathbb{K}^{N\times N}$ e qualquer $x \in \mathbb{K}^N$.

Definição 1.6. *Uma norma $\|\cdot\|_M$ em $\mathbb{K}^{N\times N}$ diz-se* regular *se*

$$\|AB\|_M \leq \|A\|_M \|B\|_M, \, \forall A, \, B \in \mathbb{K}^{N\times N}.$$

Exemplo 1.5. *A norma de Frobenius é regular:*

$$\|AB\|_{Fb}^2 = \sum_{i,j=1}^{N} |(AB)_{ij}|^2 = \sum_{i,j=1}^{N} |e_i^{\top} ABe_j|^2 = \sum_{i,j=1}^{N} |(e_i^{\top} A, Be_j)|^2 \leq \sum_{i,j=1}^{N} \|e_i^{\top} A\|_2^2 \|Be_j\|_2^2$$

$$= \left(\sum_{i=1}^{N} \|e_i^{\top} A\|_2\right) \left(\sum_{j=1}^{N} \|Be_j\|_2^2\right) = \left(\sum_{i,k=1}^{N} |a_{ik}|^2\right) \left(\sum_{k,j=1}^{N} |b_{kj}|^2\right) = \|A\|_{Fb}^2 \|B\|_{Fb}^2$$

Sendo $\|\cdot\|_V$ uma norma vetorial, tem-se

$$\sup_{x\in\mathbb{K}^N\setminus\{0\}} \frac{\|Ax\|_V}{\|x\|_V} < +\infty,\ \forall A\in\mathbb{K}^{N\times N}$$

e a função $\|\cdot\|_M : \mathbb{K}^{N\times N}\to\mathbb{R}$ definida por

$$\|A\|_M = \sup_{x\in\mathbb{K}^N\setminus\{0\}} \frac{\|Ax\|_V}{\|x\|_V} \tag{1}$$

satisfaz todas as condições da Definição 1.1. Assim, podemos introduzir:

Definição 1.7. *A norma* $\|\cdot\|_M : \mathbb{K}^{N\times N}\to\mathbb{R}$ *definida por* (1) *diz-se a* norma matricial induzida pela norma vetorial $\|\cdot\|_V$.

É fácil ver que

$$\|A\|_M = \sup_{x\in\mathbb{K}^N\setminus\{0\}} \|A\frac{x}{\|x\|_V}\|_V = \max_{x\in\mathbb{K}^N,\,\|x\|_V=1} \|Ax\|_V. \tag{2}$$

Qualquer norma matricial induzida é regular e compatível com a norma vetorial que lhe dá origem. Além disso, $\|I\| = 1$ em qualquer norma matricial induzida.

Exemplo 1.6. *Seja* $I\in\mathbb{K}^{N\times N}$ *a matriz identidade. Para a norma de Frobenius tem-se* $\|I\|_{Fb} = \sqrt{N}$*, pelo que esta norma não é induzida por nenhuma norma vetorial. Além disso, o valor da norma aumenta à medida que a dimensão do espaço de matrizes aumenta.*

Vamos agora averiguar a possibilidade de calcular as normas matriciais induzidas $\|\cdot\|_p$, $p\in[1,\infty]$, através de cálculos diretos com as componentes a_{ij} da matriz A. Será possível deduzir fórmulas mais simples do que (1) e (2) para calcular as normas matriciais?

No caso das normas $\|A\|_1$ e $\|A\|_\infty$, $A\in\mathbb{K}^{N\times N}$, tem-se

1. $\|A\|_1 = \max_{1\le j\le N}\sum_{i=1}^N |a_{ij}|$, ou seja, é uma *norma por colunas*:

$$\begin{aligned}
(i)\,\|A\|_1 &= \max_{x\in\mathbb{K}^N,\,\|x\|_1=1}\|Ax\|_1 = \max_{x\in\mathbb{K}^N,\,\|x\|_1=1}\sum_{i=1}^N |(Ax)_i| = \max_{x\in\mathbb{K}^N,\,\|x\|_1=1}\sum_{i=1}^N |\sum_{j=1}^N a_{ij}x_j| \\
&\le \max_{x\in\mathbb{K}^N,\,\|x\|_1=1}\sum_{i=1}^N\sum_{j=1}^N |a_{ij}||x_j| \le \max_{x\in\mathbb{K}^N,\,\|x\|_1=1}\sum_{j=1}^N\left(\sum_{i=1}^N |a_{ij}|\right)|x_j| \\
&\le \max_{1\le j\le N}\sum_{i=1}^N |a_{ij}| \max_{x\in\mathbb{K}^N,\,\|x\|_1=1}\sum_{j=1}^N |x_j| = \max_{1\le j\le N}\sum_{i=1}^N |a_{ij}|;
\end{aligned}$$

   (ii) $\max_{1\le j\le N}\sum_{i=1}^N |a_{ij}| = \sum_{i=1}^N |a_{ij^*}| = \|Ae_{j^*}\|_1 \le \|A\|_1$.

2. $\|A\|_\infty = \max_{1\leq i\leq N} \sum_{j=1}^{N} |a_{ij}|$, tratando-se de uma *norma matricial por linhas* (ver, por exemplo, [1]).

No que se segue, usaremos a mesma notação $\|\cdot\|$ para normas vetoriais e normas matriciais induzidas.

2 Relações entre normas matriciais e raio espetral

Começamos por recordar a noção de raio espetral.

Definição 2.1. *Se $A \in \mathbb{K}^{N\times N}$, $\sigma(A) \subset \mathbb{C}$ designa o* espetro *de A, ou seja, o conjunto de todos os valores próprios da matriz A. A $\varrho(A) = \max_{\lambda\in\sigma(A)} |\lambda|$ chama-se* raio espetral *de A.*

A primeira relação entre normas matriciais e raio espetral envolve a norma $\|\cdot\|_2$:

$$\|A\|_2 = (\varrho(A^*A))^{\frac{1}{2}}, \forall A \in \mathbb{K}^{N\times N}.$$

De facto, para $x \in \mathbb{K}^N$ com $\|x\|_2 = 1$, tem-se

$$\|Ax\|_2^2 = (Ax, \overline{Ax}) = x^*(A^*A)x.$$

A matriz A^*A é hermiteana, com valores próprios $\lambda_1, ..., \lambda_N \geq 0$. Usando a decomposição $A^*A = U^*DU$ com U unitária e $d_{ij} = \delta_{ij}\lambda_i$, obtém-se

$$\|Ax\|_2^2 = (Ax, \overline{Ax}) = x^*(A^*A)x = (Ux)^*D(Ux) = y^*Dy = (\sqrt{D}y, \overline{\sqrt{D}y}) = \|\sqrt{D}y\|_2^2$$

onde $\|y\|_2 = \|x\|_2 = 1$. Então

$$\max_{x\in\mathbb{K}^N,\, \|x\|_2=1} \|Ax\|_2 = \max_{y\in\mathbb{K}^N,\, \|y\|_2=1} \|\sqrt{D}y\|_2 = \max_{1\leq i\leq N} \sqrt{|\lambda_i|}$$

O raio espetral pode ser entendido como o ínfimo de todas as normas matriciais induzidas, de acordo com os seguintes resultados:

Teorema 2.1. *(i) Qualquer que seja a norma matricial induzida $\|\cdot\|$, tem-se*

$$\varrho(A) \leq \|A\|, \forall A \in \mathbb{K}^{N\times N}.$$

(ii) Para cada $A \in \mathbb{K}^{N\times N}$ e cada $\varepsilon > 0$, existe uma norma matricial induzida $\|\cdot\|$ tal que

$$\|A\| \leq \varrho(A) + \varepsilon.$$

Demonstração. (i) O caso $\mathbb{K} = \mathbb{C}$ é simples. Seja $\|\cdot\|$ uma norma matricial induzida e seja $A \in \mathbb{C}^{N\times N}$. Para qualquer $\lambda \in \sigma(A)$ e $v \in \mathbb{C}^N \setminus \{0\}$ tal que $Av = \lambda v$, tem-se

$$\|A\| = \sup_{x\in\mathbb{C}^N\setminus\{0\}} \frac{\|Ax\|}{\|x\|} \geq \frac{\|Av\|}{\|v\|} = \frac{\|\lambda v\|}{\|v\|} = |\lambda|$$

pelo que $\varrho(A) \leq \|A\|$.

No caso $\mathbb{K} = \mathbb{R}$, se $\varrho(A) = |\lambda|$ e $\lambda \in \mathbb{R}$, a demonstração anterior continua válida, pois podemos tomar um vetor próprio de A, $v \in \mathbb{R}^N \setminus \{0\}$, para deduzir

$$\|A\| = \sup_{x\in\mathbb{R}^N\setminus\{0\}} \frac{\|Ax\|}{\|x\|} \geq \frac{\|Av\|}{\|v\|} = \frac{\|\lambda v\|}{\|v\|} = |\lambda| = \varrho(A). \tag{3}$$

No entanto, se $\varrho(A) = |\lambda|$ com $\lambda \in \mathbb{C} \setminus \mathbb{R}$, não podemos usar um vetor próprio complexo v em (3) uma vez que o supremo é tomado em $\mathbb{R}^N \setminus \{0\}$ apenas. Seja então $v = x + iy$, com $x, y \in \mathbb{R} \setminus \{0\}$ um vetor próprio associado ao valor próprio complexo $\lambda = |\lambda|e^{i\theta}$. Tem-se

$$Re(e^{i\varphi}v) = \cos(\varphi)x - \sin(\varphi)y \neq 0, \text{ para todo } \varphi \in [0, 2\pi]$$

pelo que

$$\min_{0\leq\varphi\leq 2\pi} \|Re(e^{i\varphi}v)\| =: \|Re(e^{i\varphi_0}v)\| = \|\cos(\varphi_0)x - \sin(\varphi_0)y\| > 0.$$

Pondo $u := Re(e^{i\varphi_0}v)$, obtém-se

$$\|A\| \geq \frac{\|Au\|}{\|u\|} = \frac{\|Re(e^{i\varphi_0}Av)\|}{\|u\|} = |\lambda|\frac{\|Re(e^{i(\varphi_0+\theta)}v)\|}{\|u\|} \geq |\lambda|.$$

(ii) Dado $A \in \mathbb{K}^{N\times N}$, sejam $\lambda_1, ..., \lambda_k$ os valores próprios distintos de A, cada um deles com multiplicidade algébrica $m_a(\lambda_j)$ e multiplicidade geométrica $m_g(\lambda_j)$. Pelo Teorema da decomposição de Jordan, existe uma matriz invertível $S \in \mathbb{C}^{N\times N}$ tal que $A = SJS^{-1}$, em que $J = diag(J^{(1)}, ..., J^{(k)}) \in \mathbb{C}^{N\times N}$ é uma forma canónica de Jordan, onde cada bloco $J^{(j)} \in \mathbb{C}^{m_a(\lambda_j)\times m_a(\lambda_j)}$, por sua vez, é uma matriz diagonal por blocos, com o número de blocos igual a $m_g(\lambda_j)$, sendo cada um desses blocos da forma

$$J_q(\lambda_j) = \begin{bmatrix} \lambda_j & 1 & 0 & \dots & 0 \\ 0 & \lambda_j & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & 0 & \lambda_j & 1 \\ 0 & \cdots & \cdots & 0 & \lambda_j \end{bmatrix} \in \mathbb{C}^{q\times q}.$$

Pondo $D := diag(1, \varepsilon, ..., \varepsilon^{N-1})$, tem-se

$$\|D^{-1}JD\|_\infty \leq \varrho(A) + \varepsilon. \tag{4}$$

Para terminar, define-se a norma vetorial $\|x\| := \|D^{-1}S^{-1}x\|_\infty$ (ver Exemplo 1.2) e verifica-se que a correspondente norma matricial é dada por

$$\begin{aligned} \|A\| &= \sup_{x\in\mathbb{K}^N\setminus\{0\}} \frac{\|D^{-1}S^{-1}Ax\|_\infty}{\|D^{-1}S^{-1}x\|_\infty} = \sup_{x\in\mathbb{K}^N\setminus\{0\}} \frac{\|D^{-1}JS^{-1}x\|_\infty}{\|D^{-1}S^{-1}x\|_\infty} \\ &= \sup_{x\in\mathbb{K}^N\setminus\{0\}} \frac{\|D^{-1}JD(D^{-1}S^{-1}x)\|_\infty}{\|D^{-1}S^{-1}x\|_\infty} = \sup_{x\in\mathbb{K}^N\setminus\{0\}} \frac{\|D^{-1}JDy\|_\infty}{\|y\|_\infty} = \|D^{-1}JD\|_\infty. \end{aligned}$$

Combinando este resultado com (4), obtém-se a relação pretendida: $\|A\| \leq \varrho(A) + \varepsilon$. □

Teorema 2.2. *(Fórmula de Gelfand). Seja $A \in \mathbb{K}^{N\times N}$. Para qualquer norma matricial induzida $\|\cdot\|$, tem-se*

$$\varrho(A) = \lim_{n\to\infty} \|A^n\|^{1/n}.$$

Demonstração. Ver [11]. □

Aparentemente, a fórmula de Gelfand poderia ser útil do ponto de vista computacional, como processo de aproximação do raio espetral de uma matriz, uma vez que podemos tomar, com qualquer norma matricial induzida,

$$\varrho(A) \approx \|A^n\|^{1/n}, \text{ para } n \text{ suficientemente grande.} \tag{5}$$

Vejamos um exemplo:

Exemplo 2.1. *Seja*

$$C = \begin{bmatrix} 0 & -0.7 & -0.8 & -0.7 \\ -1.4 & 0 & -1.2 & -1 \\ -0.8 & -0.6 & 0 & -0.9 \\ -0.7 & -0.5 & -0.9 & 0 \end{bmatrix}$$

Vamos recorrer ao MATLAB *para calcular vários termos $\|C^n\|^{1/n}$, na norma $\|\cdot\|_1$, por exemplo. Obtém-se*

```
>> norm(C^10,1)^(1/10)
ans = 2.505819350463062

>> norm(C^20,1)^(1/20)
```

```
ans = 2.490761143816249

>> norm(C^100,1)^(1/100)
ans = 2.478778174813123

>> norm(C^200,1)^(1/200)
ans = 2.477284362866631

>> norm(C^500,1)^(1/500)
ans = 2.476388507844573

>> norm(C^600,1)^(1/600)
ans = 2.476288988397730

>> norm(C^700,1)^(1/700)
ans = 2.476217905527190

>> norm(C^800,1)^(1/800)
ans = Inf
```

o que mostra que a fórmula de Gelfand pode não ser um processo eficaz para aproximar (numericamente, com muita precisão) o raio espetral de uma matriz. Contudo, parece ser $\varrho(C) \approx 2.47....$

Mas a relação assintótica

$$\varrho(A)^n \approx \|A^n\|, \text{ para } n \text{ suficientemente grande}$$

irá, mais à frente, ajudar a caracterizar a rapidez de convergência dos métodos iterativos para sistemas lineares.

3 Teorema do ponto fixo em $\mathbb{R}^N$

Este teorema, bem como as suas generalizações a espaços de Banach, são de fundamental importância em Análise Numérica e Matemática Aplicada. O Teorema do ponto fixo de Banach fornece (i) existência e unicidade de ponto fixo para funções contrativas, (ii) um algoritmo para o cálculo aproximado desse ponto fixo, o chamado método do ponto fixo (iii)

majorações *a priori* e *a posteriori* para os erros das aproximações fornecidas pelo método do ponto fixo.

No que se segue, $\|\cdot\|$ designa uma norma específica sobre $\mathbb{R}^N$. Uma sucessão de termos em $\mathbb{R}^N$ será escrita na forma $\left\{x^{(n)}\right\}_{n\in\mathbb{N}_0}$, exceto, eventualmente, quando $N = 1$. Sabemos que toda a sucessão de Cauchy em $\mathbb{R}^N$, i.e., toda a sucessão que satisfaz

$$\forall \varepsilon > 0\ \exists p \in \mathbb{N} :\ m, n > p \Rightarrow \|x^{(m)} - x^{(n)}\| < \varepsilon,$$

é sucessão convergente em $\mathbb{R}^N$.

Definição 3.1. *Um elemento $z \in X$ diz-se um* ponto fixo *da função $G : X \subseteq \mathbb{R}^N \to \mathbb{R}^N$ se $G(z) = z$.*

Definição 3.2. *Uma função $G : X \subseteq \mathbb{R}^N \to \mathbb{R}^N$ diz-se* Lipschitziana *ou* função de Lipschitz *se*

$$\exists L \geq 0 :\ \|G(x) - G(y)\| \leq L\|x - y\|,\ \forall\, x, y \in X.$$

Ao ínfimo de todas as constantes L que satisfazem esta relação chama-se constante de Lipschitz *de G.*

A função $G : X \subseteq \mathbb{R}^N \to \mathbb{R}^N$ diz-se uma contração *se*

$$\exists\, 0 \leq L < 1 :\ \|G(x) - G(y)\| \leq L\|x - y\|,\ \forall\, x, y \in X.$$

Neste caso, a constante de Lipschitz diz-se a constante de contratividade *de G.*

Note-se que, apesar de todas as normas em $\mathbb{R}^N$ serem equivalentes, uma função pode ser contrativa em relação a uma norma e não o ser noutra norma.

Estamos agora em condições de enunciar e demonstrar o Teorema do ponto fixo em $\mathbb{R}^N$.

Teorema 3.1. *Seja $X \subseteq \mathbb{R}^N$ não vazio e fechado. Se $G : X \to \mathbb{R}^N$ é uma função tal que*

(i) $G(X) \subseteq X$;

(ii) G *é uma contração em* X

então

1. *G tem um e um só ponto fixo em X, $z = G(z)$;*
2. *O método do ponto fixo $x^{(n+1)} = G(x^{(n)})$, $n \in \mathbb{N}_0$, converge para z, qualquer que seja $x^{(0)} \in X$;*

3. São válidas as majorações para os erros $z - x^{(n)}$

$$\|z - x^{(n)}\| \leq L^n \|z - x^{(0)}\|,$$

$$\|z - x^{(n)}\| \leq \frac{L^n}{1-L} \|x^{(1)} - x^{(0)}\|,$$

$$\|z - x^{(n+1)}\| \leq \frac{L}{1-L} \|x^{(n+1)} - x^{(n)}\|,\ n \in \mathbb{N}_0,$$

onde $0 \leq L < 1$ *é a constante de contratividade de* G *em* X*.*

Demonstração. Consideremos uma sucessão $\left\{x^{(n)}\right\}_{n\in\mathbb{N}_0}$ definida por

$$\begin{cases} x^{(0)} \in X, \\ x^{(n+1)} = G(x^{(n)}),\ n \in \mathbb{N}_0 \end{cases} \tag{6}$$

onde $x^{(0)}$ é arbitrário, e provemos que se trata de uma sucessão de Cauchy. Convém observar desde já que a hipótese $G(X) \subseteq X$ permite concluir que $x^{(n)} \in X$, para todo $n \in \mathbb{N}_0$, quando $x^{(0)} \in X$. Como G é uma contração em X, tem-se:

$$\|x^{(n+1)} - x^{(n)}\| = \|G(x^{(n)}) - G(x^{(n-1)})\| \leq L\|x^{(n)} - x^{(n-1)}\|, \quad \forall n \in \mathbb{N}.$$

Isto permite mostrar, por indução, que

$$\|x^{(n+1)} - x^{(n)}\| \leq L^n \|x^{(1)} - x^{(0)}\|, \quad \forall n \in \mathbb{N}_0. \tag{7}$$

Sejam $n, m \in \mathbb{N}$ com $m > n$. Começando por aplicar a propriedade telescópica dos somatórios e usando depois (7), obtém-se

$$\begin{aligned} \|x^{(m)} - x^{(n)}\| &= \left\| \sum_{k=n}^{m-1} (x^{(k+1)} - x^{(k)}) \right\| \leq \sum_{k=n}^{m-1} \left\| x^{(k+1)} - x^{(k)} \right\| \\ &\leq \sum_{k=n}^{m-1} L^k \|x^{(1)} - x^{(0)}\| = \|x^{(1)} - x^{(0)}\| L^n \frac{1 - L^{m-n}}{1-L} \\ &= \frac{\|x^{(1)} - x^{(0)}\|}{1-L} |L^n - L^m|. \end{aligned}$$

Portanto $\left\{x^{(n)}\right\}_{n\in\mathbb{N}_0}$ é uma sucessão de Cauchy, uma vez que $\left\{L^n\right\}_{n\in\mathbb{N}_0}$ o é, pois $0 \leq L < 1$.

Como X é fechado, a sucessão $\left\{x^{(n)}\right\}_{n\in\mathbb{N}_0}$ é convergente para um elemento de X. Pondo $z := \lim_{n\to\infty} x^{(n)}$ e passando ao limite em $x^{(n+1)} = G(x^{(n)})$, usando o facto de G ser contínua em X, obtém-se $G(z) = z$. Fica assim provado a *existência* de ponto fixo de G e a *convergência* do método do ponto fixo (6).

Para provar a *unicidade* do ponto fixo, suponhamos que existiam dois pontos fixos de G, $z, \zeta \in X$. Então

$$\|z - \zeta\| = \|G(z) - G(\zeta)\| \leq L\|z - \zeta\|$$

obtendo-se

$$(1 - L)\|z - \zeta\| \leq 0.$$

Como $1 - L > 0$, conclui-se que $\|z - \zeta\| \leq 0$, e portanto $z = \zeta$.

Quanto às fórmulas de majoração dos erros das aproximações, passando ao limite $m \to \infty$, com n fixo, em

$$\|x^{(m)} - x^{(n)}\| \leq \|\frac{L^n - L^m}{1 - L}\|x^{(1)} - x^{(0)}\|$$

obtém-se a estimativa de erro *a priori*:

$$\|z - x^{(n)}\| \leq \frac{L^n}{1 - L}\|x^{(1)} - x^{(0)}\|.$$

A validade da fórmula de erro *a posteriori* resulta de

$$\begin{aligned}\|x^{(n+m)} - x^{(n+1)}\| &= \left\|\sum_{k=1}^{m-1}(x^{(n+1+k)} - x^{(n+k)})\right\| \\ &\leq \sum_{k=1}^{m-1}\|x^{(n+1+k)} - x^{(n+k)})\| \leq \sum_{k=1}^{m-1} L^k\|x^{(n)} - x^{(n+1)}\| = \frac{L - L^m}{1 - L}\|x^{(n)} - x^{(n+1)}\|\end{aligned}$$

onde $m > 1$, e de fazer $m \to \infty$ (com n fixo) em

$$\|x^{(n+m)} - x^{(n+1)}\| \leq \frac{L - L^m}{1 - L}\|x^{(n+1)} - x^{(n)}\|.$$

Finalmente, usando $z = G(z)$ e $x^{(k)} = G(x^{(k-1)})$, obtém-se

$$\|z - x^{(k)}\| = \|G(z) - G(x^{(k-1)})\| \leq L\|z - x^{(k-1)}\|$$

donde resulta

$$\|z - x^{(n)}\| \leq L^n\|z - x^{(0)}\|.$$

□

Note-se que, para $N = 1$, obtemos uma extensão do teorema do ponto fixo antes estabelecido apenas para intervalos limitados. No teorema que se segue, usamos a notação do capítulo anterior para as sucessões de números reais.

Teorema 3.2. *Seja X um subconjunto fechado, não vazio de $\mathbb{R}$. Se $g : X \to \mathbb{R}$ é uma função tal que*

(i) $g(X) \subseteq X$;

(ii) g é uma contração em X, com constante de contratividade L

então

1. *g tem um e um só ponto fixo em X, $z = g(z)$;*
2. *o método do ponto fixo $x_{n+1} = g(x_n)$, $n \in \mathbb{N}_0$, converge para z, qualquer que seja $x_0 \in X$;*
3. *são válidas as seguintes majorações para os erros $z - x_n$:*

$$|z - x_n| \leq L^n |z - x_0|,$$
$$|z - x_n| \leq \frac{L^n}{1 - L} |x_1 - x_0|,$$
$$|z - x_{n+1}| \leq \frac{L}{1 - L} |x_{n+1} - x_n|.$$

Exemplo 3.1. *A sucessão de números reais definida por $x_{n+1} = \sqrt{2 + x_n}$, $n = 0, 1, 2...$, e $x_0 \in [-2, +\infty[$, converge para $z = 2$, independentemente de x_0 no intervalo referido.*

Seja $g(x) := \sqrt{2 + x}$ $(x \geq -2)$. Uma observação, antes de aplicar o Teorema do ponto fixo, é a seguinte: qualquer que seja $x \in [-2, +\infty[$, tem-se $g(x) \in [0, +\infty[$, pelo que basta estudar a equação e o método do ponto fixo neste último intervalo. Agora, $g \in C^1([0, +\infty[)$, satisfaz

$$g([0, +\infty[) \subseteq [0, +\infty[,$$
$$|g'(x)| = g'(x) = \frac{1}{2\sqrt{2 + x}} \leq \frac{1}{2\sqrt{2}} < 1, \ \forall x \in [0, +\infty[.$$

Pelo Teorema do ponto fixo, g tem um e um só ponto fixo, $z \in [0, +\infty[$ e a sucessão definida por $x_{n+1} = g(x_n)$ converge para z, qualquer que seja a iterada inicial escolhida em $[0, +\infty[$, ou como já vimos, em $[-2, +\infty[$. Tem-se

$$z = \lim_{n \to \infty} x_{n+1} = \lim_{n \to \infty} \sqrt{2 + x_n} = \sqrt{2 + \lim_{n \to \infty} x_n} = \sqrt{2 + z}.$$

Então

$$z = \sqrt{2 + z} \Leftrightarrow z^2 = 2 + z \wedge z > 0 \Leftrightarrow z = \frac{1 \pm 3}{2} \wedge z > 0 \Leftrightarrow z = 2.$$

4 Resolução numérica de sistemas lineares

Os sistemas lineares surgem com bastante frequência na Matemática Aplicada. Ocorrem como formulações diretas de problemas do "mundo real" e surgem também como parte da análise numérica de outros problemas. Como exemplos desta última situação, temos a resolução de sistemas de equações não lineares, de equações diferenciais ordinárias e parciais, equações integrais e problemas de otimização.

4.1 Condicionamento de sistemas lineares

Antes de estudarmos os métodos iterativos para resolver sistemas lineares, vamos perceber qual é a influência dos erros de arredondamento e perturbações nos dados na solução calculada para esses problemas. Começamos com dois exemplos ilustrativos da propagação de erros em sistemas lineares.

Exemplo 4.1. *O sistema linear (muito simples)*

$$\begin{cases} x_1 + x_2 = 2 \\ x_1 + 1.00001\, x_2 = 2.00001 \end{cases}$$

tem $x = [1 \;\; 1]^\mathsf{T}$ *como única solução.*

Consideremos uma pequena perturbação nos dados

$$\begin{cases} x_1 + x_2 = 2 \\ x_1 + 1.00001\, x_2 = 2 \end{cases} \tag{8}$$

ou seja, o vetor $b = [2 \;\; 2.00001]^\mathsf{T}$ *foi arredondado para* $\tilde{b} = [2 \;\; 2]^\mathsf{T}$*. O erro relativo percentual deste arredondamento é muito pequeno e pode ser dado por*

$$100\%\|\delta_{\tilde{b}}\|_\infty = 100\%\frac{\|[0 \;\; 0.00001]^\mathsf{T}\|_\infty}{\|[2 \;\; 2.00001]^\mathsf{T}\|_\infty} \approx 0.0005\%.$$

No entanto, o sistema perturbado (8) *tem solução* $\tilde{x} = [2 \;\; 0]^\mathsf{T}$*, muito diferente de* $x = [1 \;\; 1]^\mathsf{T}$*, o que é confirmado pelo erro relativo percentual*

$$100\%\|\delta_{\tilde{x}}\|_\infty = 100\%\frac{\|[-1 \;\; 1]^\mathsf{T}\|_\infty}{\|[1 \;\; 1]^\mathsf{T}\|_\infty} = 100\%.$$

Se considerarmos agora uma pequena perturbação na matriz do sistema, temos que o novo sistema

$$\begin{cases} x_1 + x_2 = 2 \\ x_1 + 1\, x_2 = 2.00001 \end{cases}$$

não tem solução (a matriz do sistema ficou singular). Note-se que, sendo A a matriz do sistema original e $\tilde{A}$ a matriz deste sistema, se tem

$$100\%\|\delta_{\tilde{A}}\|_\infty = 100\%\frac{\|\begin{bmatrix} 0 & 0 \\ 0 & 0.00001 \end{bmatrix}\|_\infty}{\|\begin{bmatrix} 1 & 1 \\ 1 & 1.00001 \end{bmatrix}\|_\infty} \approx 0.0005\%$$

mas a naturaza da matriz foi completamente alterada pelo arredondamento efetuado.

Se considerarmos agora as seguintes pequenas perturbações

$$\begin{cases} x_1 + x_2 = 2 \\ x_1 + 1\,x_2 = 2 \end{cases}$$

obtemos um sistema que possui um número infinito de soluções. Também neste caso, pequenas perturbações nos dados do sistema (8) *produziram grandes alterações no resultado.*

Exemplo 4.2. *O sistema linear*

$$\begin{cases} 10\,x_1 + 7\,x_2 + 8\,x_3 + 7\,x_4 = 32 \\ 7\,x_1 + 5\,x_2 + 6\,x_3 + 5\,x_4 = 23 \\ 8\,x_1 + 6\,x_2 + 10\,x_3 + 9\,x_4 = 33 \\ 7\,x_1 + 5\,x_2 + 9\,x_3 + 10\,x_4 = 31 \end{cases} \tag{9}$$

tem solução $x = [1\ \ 1\ \ 1\ \ 1]^\mathsf{T}$.

Consideremos a perturbação

$$\begin{cases} 10\,x_1 + 7\,x_2 + 8\,x_3 + 7\,x_4 = 32.1 \\ 7\,x_1 + 5\,x_2 + 6\,x_3 + 5\,x_4 = 22.9 \\ 8\,x_1 + 6\,x_2 + 10\,x_3 + 9\,x_4 = 33.1 \\ 7\,x_1 + 5\,x_2 + 9\,x_3 + 10\,x_4 = 30.9 \end{cases}$$

ou seja, o vetor $b = [32\ \ 23\ \ 33\ \ 31]^\mathsf{T}$ foi substituído por $\tilde{b} = [32.1\ \ 22.9\ \ 33.1\ \ 30.9]^\mathsf{T}$, sendo o erro relativo percentual de $\tilde{b}$

$$100\%\|\delta_{\tilde{b}}\|_\infty = 100\%\frac{\|[-0.1\ \ 0.1\ \ -0.1\ \ 0.1]^\mathsf{T}\|_\infty}{\|[32\ \ 23\ \ 33\ \ 31]^\mathsf{T}\|_\infty} \approx 0.30303\%.$$

No entanto, o sistema perturbado tem solução $\tilde{x} = [9.2 \;\; -12.6 \;\; 4.5 \;\; -1.1]^\mathsf{T}$ *muito diferente de* $x = [1 \;\; 1 \;\; 1 \;\; 1]^\mathsf{T}$. *Com efeito, para o erro relativo percentual desta perturbação tem-se um erro enorme*

$$100\%\|\delta_{\tilde{x}}\|_\infty = 100\%\frac{\|[-8.2 \;\; 13.6 \;\; -3.5 \;\; 2.1]^\mathsf{T}\|_\infty}{\|[1 \;\; 1 \;\; 1 \;\; 1]^\mathsf{T}\|_\infty} = 1360\%.$$

Consideremos agora uma pequena perturbação na matriz do sistema

$$\begin{cases} 10\,x_1 + 7\,x_2 + 8.1\,x_3 + 7.2\,x_4 = 32 \\ 7.08\,x_1 + 5.04\,x_2 + 6\,x_3 + 5\,x_4 = 23 \\ 8\,x_1 + 5.98\,x_2 + 9.89\,x_3 + 9\,x_4 = 33 \\ 6.99\,x_1 + 5\,x_2 + 9\,x_3 + 9.98\,x_4 = 31 \end{cases}$$

O erro relativo percentual de $\tilde{A}$ *pode ser dado por*

$$100\%\|\delta_{\tilde{A}}\|_\infty = 100\%\frac{\max\{0.3, 0.12, 0.13, 0.03\}}{\max\{32, 23, 33, 31\}} \approx 0.9\%.$$

O sistema perturbado tem solução $\tilde{x} = [-81 \;\; 137 \;\; -34 \;\; 22]^\mathsf{T}$, *muito diferente de* $x = [1 \;\; 1 \;\; 1 \;\; 1]^\mathsf{T}$ *e tem-se*

$$100\%\|\delta_{\tilde{x}}\|_\infty = 100\%\frac{\|[82 \;\; -136 \;\; 35 \;\; -21]^\mathsf{T}\|_\infty}{\|[1 \;\; 1 \;\; 1 \;\; 1]^\mathsf{T}\|_\infty} = 13600\%.$$

A questão natural que os exemplos acima colocam é: Como explicar estes fenómenos de propagação de erros de forma tão drástica?

Começamos por analisar o caso mais simples, em que apenas o vetor b do sistema linear $Ax = b$ está afetado de erro. Suponhamos que $A \in \mathbb{R}^{N\times N}$ é não singular e considere-se o sistema perturbado $A\tilde{x} = \tilde{b}$. Seja $\|\cdot\|$ uma norma matricial induzida por uma norma vetorial, a qual será também designada por $\|\cdot\|$. Tem-se

$$x - \tilde{x} = A^{-1}(b - \tilde{b})$$

donde

$$\|x - \tilde{x}\| \leq \|A^{-1}\|\|b - \tilde{b}\|.$$

Então

$$\frac{\|x - \tilde{x}\|}{\|x\|} \leq \frac{\|A^{-1}\|\|b - \tilde{b}\|}{\|x\|}$$

e como $\|b\| \leq \|A\|\|x\|$ obtém-se

$$\|\delta_{\tilde{x}}\| \leq \|A\|\|A^{-1}\|\|\delta_{\tilde{b}}\|.$$

Definição 4.1. *Seja $A \in \mathbb{R}^{N\times N}$ não singular e $\|\cdot\|$ uma norma matricial induzida por uma norma vetorial. À quantidade $\|A\|\|A^{-1}\|$ chama-se* número de condição *da matriz A e escreve-se $cond(A) = \|A\|\|A^{-1}\|$.*

Assim, para o sistema $A\tilde{x} = \tilde{b}$ tem-se

$$\|\delta_{\tilde{x}}\| \leq cond(A)\|\delta_{\tilde{b}}\|.$$

Para matrizes A cujo número de condição seja elevado, um pequeno erro relativo no vector b *pode provocar* um grande erro relativo na solução do sistema. Se cond(A) for pequeno, podemos concluir que háo bom condicionamento do sistema. Note-se que o número de condição satisfaz $cond(A) \geq 1$, pois

$$cond(A) = \|A\|\|A^{-1}\| \geq \|AA^{-1}\| = \|I\| = 1,$$

e portanto, dizer que o número de condição é pequeno significa que este é da ordem da unidade. No caso geral, tem-se:

Teorema 4.1. *Seja $A \in \mathbb{R}^{N\times N}$ uma matriz não singular e $b \in \mathbb{R}^N$. Considere-se o sistema linear $Ax = b$ e um sistema perturbado $\tilde{A}\tilde{x} = \tilde{b}$ com $\|\delta_{\tilde{A}}\| < 1/cond(A)$, onde $\|\cdot\|$ é uma norma matricial induzida por uma norma vetorial. Então $\tilde{A}$ é não singular (e consequentemente, o sistema $\tilde{A}\tilde{x} = \tilde{b}$ admite uma e uma só solução) e tem-se*

$$\|\delta_{\tilde{x}}\| \leq \frac{cond(A)}{1 - cond(A)\|\delta_{\tilde{A}}\|} \left(\|\delta_{\tilde{A}}\| + \|\delta_{\tilde{b}}\|\right).$$

Demonstração. (i) Começamos por mostrar que a condição $\|A^{-1}\|\|\delta_{\tilde{A}}\| < 1/cond(A)$ garante que $\tilde{A}$ é não-singular. Note-se que

$$\|\delta_{\tilde{A}}\| < 1/cond(A) \Leftrightarrow \|A - \tilde{A}\|\|A^{-1}\| < 1$$

pelo que

$$\|\delta_{\tilde{A}}\| < 1/cond(A) \Rightarrow \|I - A^{-1}\tilde{A}\| = \|A^{-1}(A - \tilde{A})\| \leq \|A^{-1}\|\|A - \tilde{A}\| < 1.$$

Vamos mostrar que, para um vetor arbitrário $a \in \mathbb{R}^N$, o sistema $\tilde{A}z = a$ tem uma e uma só solução. Atendendo a que

$$\tilde{A}z = a \Leftrightarrow 0 = -\tilde{A}z + a \Leftrightarrow 0 = A^{-1}(a - \tilde{A}z) \Leftrightarrow z = (I - A^{-1}\tilde{A})z + A^{-1}a,$$

basta mostrar que a função $G(y) := (I - A^{-1}\tilde{A})y + A^{-1}a$ tem um e um só ponto fixo em $\mathbb{R}^N$. Como

$$\|G(z) - G(y)\| = \|(I - A^{-1}\tilde{A})(z - y)\| \leq \|(I - A^{-1}\tilde{A})\|\|z - y\|, \, \forall z, y \in \mathbb{R}^N$$

e $\|(I - A^{-1}\tilde{A})\| < 1$, a função G é uma contração em $\mathbb{R}^N$. O resultado pretendido é assegurado pelo Teorema do ponto fixo aplicado a G em $X = \mathbb{R}^N$. Neste caso, é trivial que $G(x) \subseteq X$.

(ii) Atendendo a que

$$\tilde{A}\tilde{x} = \tilde{b} \Leftrightarrow \tilde{x} = (I - A^{-1}\tilde{A})\tilde{x} + A^{-1}\tilde{b},$$

e $x = A^{-1}b$, tem-se

$$x - \tilde{x} = (A^{-1}\tilde{A} - I)\tilde{x} + A^{-1}(b - \tilde{b}) = (I - A^{-1}\tilde{A})(x - \tilde{x}) + (A^{-1}\tilde{A} - I)x + A^{-1}(b - \tilde{b}).$$

Então

$$\|x - \tilde{x}\| \leq \|I - A^{-1}\tilde{A}\|\|x - \tilde{x}\| + \|A^{-1}\tilde{A} - I\|\|x\| + \|A^{-1}\|\|b - \tilde{b}\|$$

e de

$$\|I - A^{-1}\tilde{A}\| \leq \|A^{-1}\|\|A - \tilde{A}\| = cond(A)\|\delta_{\tilde{A}}\|$$

resulta

$$\frac{\|x - \tilde{x}\|}{\|x\|} \leq cond(A)\|\delta_{\tilde{A}}\|\frac{\|x - \tilde{x}\|}{\|x\|} + cond(A)\|\delta_{\tilde{A}}\| + \frac{\|A^{-1}\|\|b - \tilde{b}\|}{\|x\|}.$$

Como $\|b\| \leq \|A\|\|x\|$ tem-se

$$\frac{\|A^{-1}\|\|b - \tilde{b}\|}{\|x\|} \leq cond(A)\|\delta_{\tilde{b}}\|$$

e ainda

$$(1 - cond(A)\|\delta_{\tilde{A}}\|)\|\delta_{\tilde{x}}\| \leq cond(A)\|\delta_{\tilde{A}}\| + cond(A)\|\delta_{\tilde{b}}\|.$$

Se $\|A^{-1}\|\|\delta_{\tilde{A}}\| < 1/cond(A)$ então

$$\|\delta_{\tilde{x}}\| \leq \frac{cond(A)\|\delta_{\tilde{A}}\| + cond(A)\|\delta_{\tilde{b}}\|}{1 - cond(A)\|\delta_{\tilde{A}}\|}.$$

□

Estamos agora em condições de explicar o que acontece nos exemplos apresentados. Para o Exemplo 1, tem-se

$$\|A\|_\infty = \max\{2, 2.00001\} = 2.00001,$$

$$\|A^{-1}\|_\infty = \| \begin{bmatrix} 100001 & -100000 \\ -100000 & 100000 \end{bmatrix} \|_\infty = \max\{200001, 200000\} = 200001,$$

$$cond(A) = 2.00001 \times 200001 = 400004 \approx 4 \times 10^5.$$

Assim, o número de condição de A é muito grande. Por exemplo, quando $e_{\tilde{A}} = 0$, fica

$$\|\delta_{\tilde{x}}\|_\infty \leq 400004 \|\delta_{\tilde{b}}\|_\infty,$$

o que significa que os erros relativos em b *podem ser* ampliados 400004 vezes.

Note-se que $\|\delta_{\tilde{A}}\|_\infty cond_\infty(A) = 2.00001 > 1$ e $\tilde{A}$ é singular.

Para o Exemplo 2:

$$\|A\|_\infty = 33$$

$$\|A^{-1}\|_\infty = \| \begin{bmatrix} 25 & -41 & 10 & -6 \\ -41 & 68 & -17 & 10 \\ 10 & -17 & 5 & -3 \\ -6 & 10 & -3 & 2 \end{bmatrix} \|_\infty \max\{82, 136, 35, 21\} = 136$$

$$cond(A) = 33 \cdot 136 = 4488$$

donde

$$\|\delta_{\tilde{x}}\|_\infty \leq 4488 \|\delta_{\tilde{b}}\|_\infty$$

quando $e_{\tilde{A}} = 0$. Neste caso, os erros relativos em b *podem ser* ampliados 4488 vezes.

Para corrigir ou minimizar o efeito do mau condicionamento de sistemas lineares, pode-se tentar [6]:

- Mudar a escala da matriz;
- Usar estratégias de pivot na resolução dos sistemas por métodos diretos;
- Usar métodos iterativos e técnicas de pré-condicionamento;
- Aplicar técnicas de regularização (por exemplo, a regularização de Tikhonov).

4.2 Métodos iterativos para sistemas lineares

Consideremos um sistema de equações lineares $Ax = b$, com $A \in \mathbb{R}^{N\times N}$ e $b \in \mathbb{R}^N$. A partir de uma decomposição aditiva da matriz A, $A = M_A + N_A$, onde M_A se supõe não singular e facilmente invertível, por exemplo, diagonal, tridiagonal, triangular,..., podemos escrever

$$Ax = b \Leftrightarrow M_A x = -N_A x + b \Leftrightarrow x = -M_A^{-1} N_A x + M_A^{-1} b \Leftrightarrow x = Cx + d$$

onde

$$C := -M_A^{-1} N_A = -M_A^{-1}(A - M_A) = I - M_A^{-1} A$$
$$d := M_A^{-1} b.$$

o que conduz ao método do ponto fixo $x^{(n+1)} = Cx^{(n)} + d$, $n \in \mathbb{N}_0$, para aproximar a solução do sistema $Ax = b$.

Definição 4.2. *Sejam $C \in \mathbb{R}^{N\times N}$ e $d \in \mathbb{R}^N$ tais que $Ax = b \Leftrightarrow x = Cx + d$. Neste caso, diz-se que o método iterativo $x^{(n+1)} = Cx^{(n)} + d$, $n \in \mathbb{N}_0$, é* consistente com o sistema linear *$Ax = b$. A matriz C chama-se* matriz de iteração *do método.*

Como corolário do Teorema do ponto fixo, temos a seguinte condição suficiente de convergência:

Teorema 4.2. *Nas condições da Definição 4.2, se $\|C\| < 1$ para alguma norma matricial induzida, então*

1. *Existe um e um só $z \in \mathbb{R}^N$ tal que $z = Cz + d$, ou seja, o sistema $Ax = b$ tem uma única solução;*

2. *O método do ponto fixo $x^{(n+1)} = Cx^{(n)} + d$, $n \in \mathbb{N}_0$, converge para z, qualquer que seja $x^{(0)} \in \mathbb{R}^N$;*

3. *São válidas as seguintes majorações para os erros $z - x^{(n)}$:*

$$\|z - x^{(n)}\| \leq \|C\|^n \|z - x^{(0)}\|,$$

$$\|z - x^{(n)}\| \leq \frac{\|C\|^n}{1 - \|C\|} \|x^{(1)} - x^{(0)}\|,$$

$$\|z - x^{(n+1)}\| \leq \frac{\|C\|}{1 - \|C\|} \|x^{(n+1)} - x^{(n)}\|,\ n \in \mathbb{N}_0.$$

Demonstração. Aplica-se o Teorema do ponto fixo com $X = \mathbb{R}^N$ e $G(x) := Cx + d$. É óbvio que $G(\mathbb{R}^N) \subseteq \mathbb{R}^N$ e a condição de contratividade é fácil de estabelecer:

$$\|G(x) - G(y)\| = \|C(x-y)\| \leq \|C\|\|x-y\|, \, \forall \, x, y \in \mathbb{R}^N$$

Assim, se $L := \|C\| < 1$ são satisfeitas as hipóteses do Teorema 3.1. □

Vejamos agora uma condição necessária e suficiente de convergência para o método iterativo anterior.

Teorema 4.3. *Sejam $C \in \mathbb{R}^{N\times N}$, $d \in \mathbb{R}^N$ e suponhamos que $z = Cz + d$. O método do ponto fixo $x^{(n+1)} = Cx^{(n)} + d$, $n \in \mathbb{N}_0$, converge para z, qualquer que seja $x^{(0)} \in \mathbb{R}^N$, se e só se $\varrho(C) < 1$. Se $\varrho(C) = 0$ então z é obtido ao fim de um número finito de iterações (no máximo n).*

Demonstração. (i) Se $\varrho(C) < 1$ então existe $\varepsilon > 0$ tal que $\varrho(C) + \varepsilon < 1$. Pelo Teorema 2.1, existe uma norma matricial induzida $\|\cdot\|$ tal que $\|C\| \leq \varrho(C) + \varepsilon < 1$ e, pelo Teorema 4.2, esta condição é suficiente para a existência e unicidade de z e para a convergência do método iterativo.

(ii) Se $\varrho(C) \geq 1$, sejam $\lambda \in \mathbb{C}$ com $|\lambda| \geq 1$ e $v \in \mathbb{C}^N \setminus \{0\}$ tais que $Cv = \lambda v$. Os erros $z - x^{(n)}$ satisfazem

$$z - x^{(n)} = C^n(z - x^{(0)}) \quad (n \in \mathbb{N}_0).$$

Se $v \in \mathbb{R}^N$ e escolhermos $x^{(0)} = z - v$, de

$$z - x^{(n)} = C^n(z - x^{(0)}) = C^n v = \lambda^n v$$

resulta

$$\|z - x^{(n)}\| = |\lambda|^n \|v\| \geq \|v\|, \, \forall n \in \mathbb{N}_0,$$

para qualquer norma em $\mathbb{R}^N$. Nesta situação, não há convergência. Se não for possível considerar vetores próprios $v \in \mathbb{R}^N$, podemos, à semelhança da demonstração do Teorema 2.1, tomar $u := Re(e^{i\varphi_0}v)$, com

$$\|Re(e^{i\varphi_0}v)\| = \|\cos(\varphi_0)x - \sin(\varphi_0)y\| = \min_{0\leq\varphi\leq 2\pi} \|Re(e^{i\varphi}v)\|$$

obtendo

$$C^n u = Re(e^{i\varphi_0}C^n v) = |\lambda|^n Re(e^{i(\varphi_0+n\theta)}v)$$

e portanto

$$\|C^n u\| = |\lambda|^n \|Re(e^{i(\varphi_0+n\theta)}v)\| \geq |\lambda|^n \|u\|$$

para qualquer norma em $\mathbb{R}^N$. Assim, se escolhermos $x^{(0)} = z - u$, a sucessão $x^{(n+1)} = Cx^{(n)} + d$, $n \in \mathbb{N}_0$, não converge para z pois

$$\|z - x^{(n)}\| \geq |\lambda|^n \|u\| \geq \|u\|, \forall n \in \mathbb{N}_0.$$

(iii) Se $\varrho(C) = 0$ então $\lambda = 0$ é o único valor próprio de C e o polinómio característico de p é dado por $p(t) = t^N$. Pelo Teorema de Caley-Hamilton, tem-se $p(C) = C^N = 0$ e portanto

$$z - x^{(N)} = C^N(z - x^{(0)}) = 0.$$

□

É interessante notar que, ao definir a matriz de iteração $C = -M_A^{-1}N_A$ a partir da decomposição $A = M_A + N_A$, a matriz M_A deve satisfazer dois requisitos algo contraditórios: por um lado, do ponto de vista da eficiência e custo computacional, M_A deve ser "muito mais" fácil de inverter do que A, mas por outro lado, atendendo à condição de convergência $\varrho(I - M_A^{-1}A) < 1$, M_A deve ser próxima de A, sendo a rapidez de convergência tanto maior quanto mais próxima M_A estiver de A. A propósito da rapidez de convergência, do Teorema 4.2 e de (5), resulta

$$\|z - x^{(n)}\| \approx \varrho(C)^n \|z - x^{(0)}\|, \text{ para } n \text{ suficientemente grande,}$$

pelo que $\varrho(C)$ pode ser entendido como uma medida da rapidez de convergência dos métodos iterativos da forma $x^{(n+1)} = Cx^{(n)} + d$, $n \in \mathbb{N}_0$.

4.3 Métodos de Jacobi, Gauss-Seidel e SOR

Consideramos um sistema $Ax = b$, com $A \in \mathbb{R}^{N\times N}$ e $b \in \mathbb{R}^N$ em que os elementos da diagonal principal de A são todos diferentes de zero: $a_{ii} \neq 0$, $i = 1, ..., N$. Se A é não-singular, tal é possível através de eventual troca de linhas na matriz, ou mais rigorosamente, troca de equações no sistema linear. A partir das igualdades

$$x_i = \frac{1}{a_{ii}} \left(b_i - \sum_{j=1, j\neq i}^{N} a_{ij} x_j \right), \quad i = 1, ..., N,$$

podemos definir uma sucessão $\left\{x^{(n)}\right\}_{n\in\mathbb{N}_0} \subset \mathbb{R}^N$ de aproximações da solução do sistema através de

$$x_i^{(n+1)} = \frac{1}{a_{ii}} \left(b_i - \sum_{j=1, j\neq i}^{N} a_{ij} x_j^{(n)} \right), \quad i = 1, ..., N, \ n = 0, 1, ... \tag{10}$$

Esta é a forma computacional do *método de Jacobi*, um dos métodos iterativos mais simples, e que apesar de ser lento, voltou a ser popular devido às possibilidades computacionais que oferece em termos de processamento paralelo.

Exemplo 4.3. *Consideremos o sistema linear com solução única*

$$\begin{aligned} &4x_1 - 2x_2 + x_3 = 1 \\ &-2x_1 + 10x_2 - 0.5x_3 = 2 \\ &x_1 - 0.5x_2 + 2x_3 = 3 \end{aligned}$$

O método de Jacobi para este sistema fica

$$\begin{aligned} x_1^{(k+1)} &= 0.25 + 0.5x_2^{(k)} - 0.25x_3^{(k)} \\ x_2^{(k+1)} &= 0.2 + 0.2x_1^{(k)} + 0.05x_3^{(k)} \\ x_3^{(k+1)} &= 1.5 - 0.5x_1^{(k)} + 0.25x_2^{(k)}, \quad k = 0, 1, 2, ... \end{aligned}$$

Tomando a iterada inicial $x^{(0)} = [1\ 0\ 0]^\mathsf{T}$ *obtém-se*

$$\begin{aligned} x^{(1)} &= [0.25,\ 0.2 + 0.2,\ 1.5 - 0.5]^\mathsf{T} = [0.25,\ 0.4,\ 1]^\mathsf{T} \\ x^{(2)} &= [0.25 + 0.5 \times 0.4 - 0.25,\ 0.2 + 0.2 \times 0.25 + 0.05,\ 1.5 - 0.5 \times 0.25 + 0.25 \times 0.4]^\mathsf{T} \\ &= [0.2,\ 0.3,\ 1.475]^\mathsf{T} \\ x^{(3)} &= [0.03125,\ 0.31375,\ 1.475]^\mathsf{T} \\ x^{(4)} &= [0.038125,\ 0.28,\ 1.5628125]^\mathsf{T} \end{aligned}$$

Será que esta sucessão converge para a solução exata do sistema?

A matriz de iteração do método de Jacobi é dada por:

$$C = \begin{bmatrix} 0 & 0.5 & -0.25 \\ 0.2 & 0 & 0.05 \\ -0.5 & 0.25 & 0 \end{bmatrix}$$

Tem-se, por exemplo,

$$\|C\|_\infty = \max\{0.75, 0.25\} = 0.75 < 1,$$

pelo que o método de Jacobi converge qualquer que seja a iterada inicial.

Exemplo 4.4. *Consideremos o sistema* (9) *e a aproximação inicial* $x^{(0)} = [0\ 0\ 0\ 0]^\mathsf{T}$. *O método de Jacobi escreve-se*

$$\begin{aligned} x_1^{(k+1)} &= 3.3 - 0.7x_2^{(k)} - 0.8x_3^{(k)} - 0.7x_4^{(k)} \\ x_2^{(k+1)} &= 4.6 - 1.4x_1^{(k)} - 1.2x_3^{(k)} - x_4^{(k)} \\ x_3^{(k+1)} &= 3.3 - 0.8x_1^{(k)} - 0.6x_2^{(k)} - 0.9x_4^{(k)} \\ x_4^{(k)} &= 3.1 - 0.7x_1^{(k)} - 0.5x_2^{(k)} - 0.9x_3^{(k)}, \end{aligned} \tag{11}$$

Obtém-se sucessivamente

$$\begin{aligned} x^{(1)} &= [3.2,\ \ 4.6,\ \ 3.3,\ \ 3.1]^{\mathsf{T}}, \\ x^{(2)} &= [-4.83,\ \ -6.94,\ \ -4.81,\ \ -4.41]^{\mathsf{T}}, \\ x^{(3)} &= [14.993,\ \ 21.544,\ \ 15.297,\ \ 14.28]^{\mathsf{T}}, \\ x^{(4)} &= [-34.1144,\ \ -49.0266,\ \ -34.4728,\ \ -31.9344]^{\mathsf{T}}, \end{aligned}$$

e parece que não há convergência para a solução exata, pois as iteradas estão a afastar-se de $[1\ 1\ 1\ 1]^{\mathsf{T}}$ *. De facto, como* (11) *se escreve em forma matricial como*

$$x^{(k+1)} = \begin{bmatrix} 3.3 \\ 4.6 \\ 3.3 \\ 3.1 \end{bmatrix} + \begin{bmatrix} 0 & -0.7 & -0.8 & -0.7 \\ -1.4 & 0 & -1.2 & -1 \\ -0.8 & -0.6 & 0 & -0.9 \\ -0.7 & -0.5 & -0.9 & 0 \end{bmatrix} x^{(k)}$$

vê-se imediatamente que a matriz de iteração neste caso é dada por

$$C_J = \begin{bmatrix} 0 & -0.7 & -0.8 & -0.7 \\ -1.4 & 0 & -1.2 & -1 \\ -0.8 & -0.6 & 0 & -0.9 \\ -0.7 & -0.5 & -0.9 & 0 \end{bmatrix}$$

Recordamos o Exemplo 2.1, onde a aplicação da fórmula de Gelfand sugeria que $\varrho(C_J) > 1$*. De facto, pondo* $p(\lambda) := \det(C - \lambda I)$*, tem-se*

$$p(-2.5) > 0, \quad p(-2.4) < 0$$

o que confirma a existência de um valor próprio de C *no intervalo* $]-2.5, -2.4[$*. De acordo com o Teorema 4.3, existem iteradas iniciais com as quais não há convergência do método de Jacobi.*

O método de Gauss-Seidel, utiliza as componentes mais atualizadas à medida que estas vão sendo calculadas, sendo definido pelas fórmulas

$$x_i^{(n+1)} = \frac{1}{a_{ii}} \left(b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(n+1)} - \sum_{j=i+1}^{N} a_{ij} x_j^{(n)} \right), \quad i = 1, ..., N,\ n = 0, 1, ... \tag{12}$$

Exemplo 4.5. *Consideremos novamente o sistema do Exemplo 4.3*

$$\begin{aligned} 4x_1 - 2x_2 + x_3 &= 1 \\ -2x_1 + 10x_2 - 0.5x_3 &= 2 \\ x_1 - 0.5x_2 + 2x_3 &= 3 \end{aligned}$$

O método de Gauss-Seidel para este sistema escreve-se

$$\begin{aligned} x_1^{(k+1)} &= 0.25 + 0.5x_2^{(k)} - 0.25x_3^{(k)} \\ x_2^{(k+1)} &= 0.2 + 0.2x_1^{(k+1)} + 0.05x_3^{(k+1)} \\ x_3^{(k+1)} &= 1.5 - 0.5x_1^{(k+1)} + 0.25x_2^{(k+1)}, \quad k = 0, 1, 2, ... \end{aligned}$$

Tomando a iterada inicial $x^{(0)} = [1\ 0\ 0]^\mathsf{T}$*, tal como na aplicação do método de Jacobi, obtém-se sucessivamente*

$$\begin{aligned} x^{(1)} &= [0.25,\ 0.25,\ 1.4375]^\mathsf{T} \\ x^{(2)} &= [0.015625,\ 0.275,\ 1.5609375]^\mathsf{T} \\ x^{(3)} &= [-0.002734375,\ 0.2775,\ 1.5707421875]^\mathsf{T} \\ x^{(4)} &= [-0.003935546875,\ 0.27775,\ 1.5714052734375]^\mathsf{T} \end{aligned}$$

Será que há convergência para a solução exata do sistema?

Exemplo 4.6. *Consideremos novamente o sistema* (9) *e a aproximação inicial* $x^{(0)} = [0\ 0\ 0\ 0]^\mathsf{T}$*. O método de Gauss-Seidel*

$$\begin{aligned} x_1^{(k+1)} &= 3.3 - 0.7x_2^{(k)} - 0.8x_3^{(k)} - 0.7x_4^{(k)} \\ x_2^{(k+1)} &= 4.6 - 1.4x_1^{(k+1)} - 1.2x_3^{(k)} - x_4^{(k)} \\ x_3^{(k+1)} &= 3.3 - 0.8x_1^{(k+1)} - 0.6x_2^{(k+1)} - 0.9x_4^{(k)} \\ x_4^{(k)} &= 3.1 - 0.7x_1^{(k+1)} - 0.5x_2^{(k+1)} - 0.9x_3^{(k+1)}, \quad k = 0, 1, 2, ... \end{aligned}$$

produz

$$\begin{aligned} x^{(1)} &= [3.2,\ 0.12,\ 0.668,\ 0.1988]^\mathsf{T}, \\ x^{(2)} &= [2.442440,\ 0.180184,\ 1.0590176,\ 0.34708416]^\mathsf{T}, \\ x^{(3)} &= [1.983698208,\ 0.2049172288,\ 1.27771535232,\ 0.459008822912]^\mathsf{T}, \\ x^{(4)} &= [1.71307948195,\ 0.20942147958,\ 1.39077558607,\ 0.54443559538]^\mathsf{T}, \end{aligned}$$

e ainda

$$x^{(100)} = [1.426881107655231,\quad 0.296920022708473,\quad 1.174654939447426,\quad 0.895533767784419],$$

$$x^{(1000)} = [1.026140353243326,\quad 0.956946422370679,\quad 1.010695113143166,\quad 0.993602939715483],$$

$$x^{(5000)} = [1.000000106223449,\quad 0.999999825048291,\quad 1.000000043460459,\quad 0.999999974005027].$$

Parece que há convergência para a solução exata, mas de forma muito lenta...

Para estudar com mais detalhe a convergência dos métodos apresentados e estimar os erros das iteradas com base nos Teoremas 4.2 e 4.3, é conveniente recorrer à forma matricial dos métodos. Para tal, considera-se as matrizes L_A, D_A, $U_A \in \mathbb{R}^{N\times N}$ que são a parte triangular estritamente inferior de A, a parte diagonal de A e a parte triangular estritamente superior de A, respetivamente, ou seja,

$$(L_A)_{ij} = \begin{cases} a_{ij},\ i > j, \\ 0,\ i \leq j \end{cases} \qquad (D_A)_{ij} = \begin{cases} a_{ij},\ i = j, \\ 0,\ i \neq j \end{cases} \qquad (U_A)_{ij} = \begin{cases} a_{ij},\ i < j, \\ 0,\ i \geq j \end{cases}$$

É fácil verificar que a matriz de iteração do método de Jacobi é dada por

$$C_J = -D_A^{-1}(L_A + U_A) = -D_A^{-1}(A - D_A) = I - D_A^{-1}A.$$

Quanto ao método de Gauss-Seidel, tem-se

$$C_{GS} = -(L_A + D_A)^{-1}U_A = -(L_A + D_A)^{-1}(A - (L_A + D_A)) = I - (L_A + D_A)^{-1}A.$$

O programa MATLAB para a aplicação do método de Jacobi a um sistema linear $Ax = b$ que a seguir se apresenta utiliza o critério de paragem $\|x^{(n)} - x^{(n-1)}\|_\infty < tol$ e a indicação do número máximo $nmax$ de iterações a realizar. Também é habitual considerar o critério de paragem $\|x^{(n)} - x^{(n-1)}\|_\infty / \|x^{(n)}\|_\infty < tol$. Supõe-se que tol será pequeno, por exemplo 10^{-6}, mas positivo. Também se supõe que a convergência do método foi verificada antes de se aplicar o programa para resolver o sistema. Recordamos que a atualização das componentes é feita simultaneamente ou em paralelo.

```
function [x,niter] = Jacobi(A,b,x0,tol,nmax)
n = length(b);
xant=x0(:);
x=zeros(n,1);
dif=2*tol;
niter=0;
while (niter<nmax) && (dif>=tol)
   for i=1:n
      x(i)=(b(i) - A(i,1:n)*xant)/A(i,i) + xant(i);
   end
   dif=norm(x-xant,inf);
   xant=x;
   niter=niter+1;
end
end
```

Quanto à implementação do método de Gauss-Seidel, onde a atualização das componentes é feita de modo sequencial, podemos considerar o seguinte programa:

```
function [x,niter] = Gauss_Seidel(A,b,x0,tol,nmax)
n = length(b);
x=x0(:);
dif=2*tol;
niter=0;
while (niter<nmax) && (dif>=tol)
   xant=x;
   for i = 1:n
      x(i) = (b(i)-A(i,1:n)*x)/A(i,i) + x(i);
   end
   dif=norm(x-xant,inf);
   niter=niter+1;
end
end
```

Exemplo 4.7. *Os resultados do Exemplo 4.6 foram obtidos executando, por exemplo,*

```
>> [x,niter] = Gauss_Seidel(A,b,[0;0;0;0],10^(-19),1000)

>> [x,niter] = Gauss_Seidel(A,b,[0;0;0;0],10^(-19),5000)
```

Recordamos que a partir de uma decomposição aditiva da matriz A, $A = M_A + N_A$, com M_A não singular e invertível, podemos escrever

$$Ax = b \Leftrightarrow x = Cx + d$$

com $C := -M_A^{-1}N_A$ e $d := M_A^{-1}b$, e usar esta via para construir outros métodos iterativos além dos de Jacobi e Gauss-Seidel. Por exemplo, para um parâmetro real $\omega \neq 0$, podemos tomar

$$M_\omega = L_A + \frac{1}{\omega}D_A,\ N_\omega = \left(1 - \frac{1}{\omega}\right) D_A + U_A$$

e considerar o método iterativo $x^{(n+1)} = C_\omega x^{(n)} + d_\omega$, com

$$C_\omega = -\left(L_A + \frac{1}{\omega}D_A\right)^{-1}\left[U_A + \left(1 - \frac{1}{\omega}\right) D_A\right] = -\left(\omega L_A + D_A\right)^{-1}\left[\omega U_A + (\omega - 1)D_A\right] \tag{13}$$

e

$$d_\omega = \left(L_A + \frac{1}{\omega}D_A\right)^{-1} b.$$

Note-se que no caso $\omega = 1$, recuperamos o método de Gauss-Seidel. Ao método definido por $x^{(n+1)} = C_\omega x^{(n)} + d_\omega$ chama-se *método de relaxação* ou *método das relaxações sucessivas*, sendo habitual a designação de método SOR, de *Successive Over Relaxation*.

Para efeitos de implementação computacional, tal como no caso do método de Gauss-Seidel, é possível evitar a inversão da matriz $L_A + \frac{1}{\omega} D_A$. De facto, basta notar que

$$\begin{gathered}
x^{(n+1)} = C_\omega x^{(n)} + d_\omega \\
\Leftrightarrow \\
\left(L_A + \tfrac{1}{\omega} D_A\right) x^{(n+1)} = -\left[U_A + \left(1 - \tfrac{1}{\omega}\right) D_A\right] x^{(n)} + b \\
\Leftrightarrow \\
\left(\omega L_A + D_A\right) x^{(n+1)} = -\left[\omega U_A + (\omega - 1) D_A\right] x^{(n)} + \omega b \\
\Leftrightarrow \\
D_A x^{(n+1)} = -\omega L_A x^{(n+1)} - \left[\omega U_A + (\omega - 1) D_A\right] x^{(n)} + \omega b \\
\Leftrightarrow \\
x^{(n+1)} = (1 - \omega) x^{(n)} - \omega D_A^{-1} L_A x^{(n+1)} - \omega D_A^{-1} U_A x^{(n)} + \omega D_A^{-1} b.
\end{gathered}$$

Comparando com o algoritmo do método de Gauss-Seidel (12), obtemos o seguinte algoritmo para o método SOR: dado $x^{(0)} \in \mathbb{R}^N$, para $n = 0, 1, ...$, calcular

$$\begin{aligned}
\widehat{x}_i^{(n+1)} &= \frac{1}{a_{ii}} \left(b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(n+1)} - \sum_{j=i+1}^{n} a_{ij} x_j^{(n)} \right), \\
x_i^{(n+1)} &= (1 - \omega) x_i^{(n)} + \omega \widehat{x}_i^{(n+1)}, \qquad i = 1, ..., n,
\end{aligned} \tag{14}$$

O método de relaxação visa, mediante escolha adequada do parâmetro ω, obter uma convergência mais rápida. Como iremos ver já a seguir, a condição $0 < \omega < 2$ é necessária para a convergência. Para além disso, para cada sistema $Ax = b$, existe um valor ótimo ω^*, no sentido que o raio espetral da matrix C_ω é minimizado para $\omega = \omega^*$. Contudo, na prática, tal parâmetro é, em geral, difícil de calcular, sendo habitualmente obtido por tentativa e erro, experimentando vários valores de ω e observando o efeito na rapidez de convergência. Para mais detalhes sobre o método SOR e outros métodos de relaxação, ver, p. ex., [3, 9].

Na análise de convergência, é conveniente utilizar a matriz de iteração do método SOR, dada por (13), sendo válido o seguinte resultado:

Teorema 4.4. *(de Kahan) Se o método SOR é convergente com qualquer iterada inicial* $x^{(0)} \in \mathbb{R}^N$ *então* $\omega \in]0, 2[$.

Demonstração. Para qualquer $\omega \neq 0$, tem-se $\varrho(C_\omega) \geq |\omega - 1|$, pois

$$\begin{aligned}
\varrho(C_\omega) = \max_{\lambda \in \sigma(C_\omega)} |\lambda| &\geq \left(\prod_{\lambda \in \sigma(C_\omega)} |\lambda|^{m_a(\lambda)} \right)^{\frac{1}{n}} \\
&= |\det(C_\omega)|^{\frac{1}{n}} = \left(\frac{|\det((\omega-1)D_A)|}{|\det(D_A)|} \right)^{\frac{1}{n}} \\
&= \left(\frac{\prod_{i=1}^n |(\omega-1)|a_{ii}|}{\prod_{i=1}^n |a_{ii}|} \right)^{\frac{1}{n}} = |\omega - 1|
\end{aligned}$$

onde $m_a(\lambda)$ designa a multiplicidade algébrica do valor próprio λ. Aqui usámos o facto de o determinante de uma matriz quadrada (real ou complexa) ser o produto dos seus valores próprios (complexos) repetidos de acordo com a multiplicidade algébrica. Na hipótese de convergência, tem-se $\varrho(C_\omega) < 1$, e como $|\omega - 1| \leq \varrho(C_\omega)$, conclui-se que $|\omega - 1| < 1$, ou seja $0 < \omega < 2$. □

4.4 Sistemas com matriz de diagonal estritamente dominante

Os Teoremas 4.2 e 4.3 baseiam-se na matriz de iteração do método iterativo para assegurar a convergência desse método para a solução do sistema $Ax = b$. No entanto, os algoritmos dos métodos de Jacobi, Gauss-Seidel e SOR não requerem explicitamente as respetivas matrizes de iteração, uma vez que as entradas da matriz A e as componentes do vetor b são usadas em cálculos diretos para obter as sucessivas iteradas dos métodos.

Vamos agora deduzir condições que permitam assegurar a convergência dos métodos acima referidos analisando diretamente a matriz A. Começamos por tentar perceber o que significa a condição $\|C_J\|_\infty < 1$ em termos de propriedades da matriz A. É fácil ver que

$$(C_J)_{ij} = \begin{cases} -\dfrac{a_{ij}}{a_{ii}},\ i \neq j, \\ 0,\ i = j \end{cases}$$

pelo que

$$\|C_J\|_\infty < 1 \Leftrightarrow \max_{1 \leq i \leq N} \sum_{j=1}^N |c_{ij}| < 1 \Leftrightarrow \max_{1 \leq i \leq N} \sum_{j=1, j \neq i}^N \frac{|a_{ij}|}{|a_{ii}|} < 1 \Leftrightarrow \sum_{j=1, j \neq i}^N \frac{|a_{ij}|}{|a_{ii}|} < 1,\ \forall i \in \{1, ..., N\}$$

ou seja, se a matriz A satisfaz

$$|a_{ii}| > \sum_{j=1, j \neq i}^N |a_{ij}|,\ \forall i \in \{1, ..., N\}$$

então, pelo Teorema 4.2, o sistema $Ax = b$ tem solução única e o método de Jacobi é convergente qualquer que seja a iterada inicial.

Definição 4.3. *Diz-se que $A \in \mathbb{R}^{N\times N}$ tem diagonal estritamente dominante por linhas (resp. por colunas) se*

$$|a_{ii}| > \sum_{j=1, j\neq i}^{N} |a_{ij}|,\ \forall i \in \{1, ..., N\} \text{ (resp. } |a_{jj}| > \sum_{i=1, i\neq j}^{N} |a_{ij}|,\ \forall j \in \{1, ..., N\}).$$

Uma consequência importante do resultado anterior, deduzido no contexto da aplicação do método de Jacobi a sistemas lineares, é que as matrizes de diagonal estritamente dominante por linhas são não-singulares. Consequentemente, também as matrizes de diagonal estritamente dominante por colunas são não-singulares.

Teorema 4.5. *Seja $A \in \mathbb{R}^{N\times N}$ com diagonal principal estritamente dominante por linhas ou por colunas. Então A é não singular e os métodos de Jacobi e Gauss-Seidel são convergentes para a solução do sistema $Ax = b$, qualquer que seja a iterada inicial $x^{(0)} \in \mathbb{R}^N$.*

Demonstração. Acabámos de mostrar que, se A tem diagonal estritamente dominante por linhas, então o método de Jacobi é convergente. O mesmo acontece se A tiver diagonal estritamente dominante por colunas. De facto,

$$\lambda \in \sigma(C_J) \Leftrightarrow \det(\lambda I - C_J) = 0 \Leftrightarrow \det(D_A^{-1}) \det(\lambda D_A + (L_A + U_A)) = 0$$

ou seja,

$$\lambda \in \sigma(C_J) \Leftrightarrow \det(\lambda D_A + (L_A + U_A)) = 0 \Leftrightarrow \lambda D_A + (L_A + U_A) \text{ é singular,}$$

e se $|\lambda| \geq 1$ e A tem diagonal principal estritamente dominante por colunas então

$$|\lambda a_{jj}| = |\lambda||a_{jj}| > |\lambda| \sum_{i=1, i\neq j}^{N} |a_{ij}| \geq \sum_{i=1, i\neq j}^{N} |a_{ij}|.$$

Mas isto significa que $\lambda D_A + (L_A + U_A)$ tem também a diagonal estritamente dominante por colunas, logo é não singular. Assim, terá de ser $|\lambda| < 1$ e o método é convergente, pelo Teorema 4.3.

No caso do método de Gauss-Seidel aplicado a A de diagonal estritamente dominante por linhas, o raciocínio é semelhante. Por um lado,

$$\lambda \in \sigma(C_{GS}) \Leftrightarrow \det(\lambda I - C_{GS}) = 0 \Leftrightarrow \det((L_A + D_A)^{-1}) \det(\lambda(L_A + D_A) + U_A) = 0$$

ou seja,

$$\lambda \in \sigma(C_{GS}) \Leftrightarrow \det(\lambda(L_A + D_A) + U_A) = 0 \Leftrightarrow \lambda(L_A + D_A) + U_A \text{ é singular,}$$

mas, por outro lado, se $|\lambda| \geq 1$ e A tem diagonal principal estritamente dominante por linhas então

$$|\lambda a_{ii}| = |\lambda||a_{ii}| > |\lambda| \sum_{j=1, j\neq i}^{N} |a_{ij}| = |\lambda| \sum_{j=1}^{i-1} |a_{ij}| + |\lambda| \sum_{j=i+1}^{N} |a_{ij}| \geq \sum_{j=1}^{i-1} |\lambda a_{ij}| + \sum_{j=i+1}^{N} |a_{ij}|$$

e $\lambda(L_A + D_A) + U_A$ tem diagonal estritamente dominante por linhas, logo é não-singular. Analogamente se mostra que se A tem diagonal principal estritamente dominante por colunas então o método de Gauss-Seidel converge, bastando, mais uma vez, usar as desigualdades

$$|\lambda a_{jj}| = |\lambda||a_{jj}| > |\lambda| \sum_{i=1, i\neq j}^{N} |a_{ij}| \geq \sum_{i=1, i\neq j}^{N} |(\lambda(L_A + D_A) + U_A)_{ij}|,$$

quando se supõe que $|\lambda| \geq 1$. □

Convém notar que um sistema linear, na sua formulação original, pode não ter diagonal estritamente dominante, mas poderá ser possível, através de trocas de linhas ou de colunas, obter um sistema equivalente onde esta característica se verifique. Contudo, se for efetuada troca de colunas, as variáveis correspondentes devem também ser trocadas.

Exemplo 4.8. *Aplicando o método de Jacobi ao sistema linear*

$$\begin{aligned} &x_1 - 0.5x_2 + 2x_3 = 3 \\ &-2x_1 + 10x_2 - 0.5x_3 = 2 \\ &4x_1 - 2x_2 + x_3 = 1 \end{aligned}$$

Fica

$$\begin{aligned} x_1^{(k+1)} &= 3 + 0.5x_2^{(k)} - 2x_3^{(k)} \\ x_2^{(k+1)} &= 0.2 + 0.2x_1^{(k)} + 0.05x_3^{(k)} \\ x_3^{(k+1)} &= 1 - 4x_1^{(k)} + 2x_3^{(k)}, \quad k = 0, 1, 2, ... \end{aligned}$$

Iterando a partir de $x^{(0)} = [1\ 0\ 0]^\mathsf{T}$*, obtém-se*

$$\begin{aligned} x^{(1)} &= [3,\ \ 0.4,\ \ -3]^\mathsf{T}, \\ x^{(2)} &= [9.199999999999999,\ \ 0.65,\ \ -10.199999999999999]^\mathsf{T}, \\ x^{(3)} &= [23.724999999999998,\ \ 1.53,\ \ -34.5]^\mathsf{T}. \end{aligned}$$

Para o sistema equivalente

$$\begin{aligned} 4x_1 - 2x_2 + x_3 &= 1 \\ -2x_1 + 10x_2 - 0.5x_3 &= 2 \\ x_1 - 0.5x_2 + 2x_3 &= 3 \end{aligned}$$

correspondente a uma troca de linhas que produz uma matriz de diagonal estritamente dominante, há convergência do método de Jacobi, como ilustra o Exemplo 4.3. Naquele exemplo, recorremos à matriz de iteração para justificar a convergência do método, mas tal poderia ter sido feito, de forma mais simples, recorrendo ao Teorema 4.5, pois a matriz

$$\begin{bmatrix} 4 & -2 & 1 \\ -2 & 10 & -0.5 \\ 1 & -0.5 & 2 \end{bmatrix}$$

tem diagonal estritamente dominante: $4 > |-2| + 1$, $10 > |-2| + |-0.5|$, $2 > 1 + |-0.5|$.

4.5 Sistemas com matriz simétrica e definida positiva

A classe das matrizes simétricas e definidas positivas aparece com frequência nas aplicações, nomeadamente, associada ao método dos mínimos quadrados, à discretização de certos operadores diferenciais e ao método dos elementos finitos para certas equações diferenciais.

Definição 4.4. *Uma matriz $A \in \mathbb{R}^{N\times N}$ diz-se definida positiva se*

$$x^{\mathsf{T}} A x > 0,\ \forall x \in \mathbb{R}^N \setminus \{0\}.$$

Recordamos algumas propriedades das matrizes definidas positivas:

- $A \in \mathbb{R}^{N\times N}$ é definida positiva $\Leftrightarrow$ $A + A^{\mathsf{T}}$ é definida positiva.
- $A \in \mathbb{R}^{N\times N}$ é definida positiva $\Leftrightarrow$ $\det(A_k) > 0$, para todo $k \in \{1, ..., N\}$, onde as submatrizes $A_k \in \mathbb{R}^{k\times k}$ consistem nos elementos das primeiras k linhas e k colunas da matriz A.

Para matrizes simétricas tem-se ainda:

- $A \in \mathbb{R}^{N\times N}$ é definida positiva $\Leftrightarrow$ todos os valores próprios de A são positivos.
- $A \in \mathbb{R}^{N\times N}$ é definida positiva $\Rightarrow$ $z^* A z > 0$, $\forall z \in \mathbb{C}^N \setminus \{0\}$.

Teorema 4.6. *(de Householder-John) Seja $A \in \mathbb{R}^{N\times N}$ e sejam M_A, $N_A \in \mathbb{R}^{N\times N}$ tais que $A = M_A + N_A$ e M_A é não singular. Seja $C := -M_A^{-1} N_A$. Se A é uma matriz simétrica e as matrizes A e $M_A - N_A^{\mathsf{T}}$ são definidas positivas, então $\varrho(C) < 1$.*

Demonstração. Sejam $\lambda \in \sigma(C)$ e $v \in \mathbb{C} \setminus \{0\}$ um vetor próprio associado a este valor próprio. Tem-se

$$Cv = \lambda v \Leftrightarrow -M_A^{-1}N_A v = \lambda v \Leftrightarrow -N_A v = \lambda M_A v \Leftrightarrow (\lambda - 1)N_A v = \lambda A v \Leftrightarrow N_A v = \frac{\lambda}{\lambda - 1} A v.$$

Note-se que $\lambda \neq 1$, pois, caso contrário, seria $Av = 0$, o que é impossível dado A ser não-singular. Tem-se também

$$v^* N_A v = \frac{\lambda}{\lambda - 1} v^* A v \quad \text{e} \quad v^* N_A^\mathsf{T} v = \frac{\overline{\lambda}}{\overline{\lambda} - 1} v^* A v,$$

onde usámos o facto de A ser simétrica.

Usando as hipóteses sobre A e $M_A - N_A^\mathsf{T}$, ambas matrizes simétricas e definidas positivas,

$$\begin{aligned} v^*(M_A - N_A^\mathsf{T})v &= v^* A v - v^* N_A v - v^* N_A^\mathsf{T} v = v^* A v - \frac{\lambda}{\lambda - 1} v^* A v - \frac{\overline{\lambda}}{\overline{\lambda} - 1} v^* A v \\ &= \left(1 - \frac{\lambda}{\lambda - 1} - \frac{\overline{\lambda}}{\overline{\lambda} - 1}\right) v^* A v = \frac{1 - |\lambda|^2}{|\lambda|^2} v^* A v > 0, \end{aligned}$$

e como $v^* A v > 0$, conclui-se que $(1 - |\lambda|^2)/|\lambda|^2 > 0$, ou seja $|\lambda| < 1$. □

Como corolário, temos:

Teorema 4.7. *(de Ostrowshi-Reich) Seja $A \in \mathbb{R}^{N \times N}$ simétrica e definida positiva e suponhamos que $0 < \omega < 2$. Então o método SOR é convergente para a solução dos sistema $Ax = b$, qualquer que seja a iterada inicial $x^{(0)} \in \mathbb{R}^N$.*

Demonstração. Basta aplicar o Teorema de Householder-John a $M_\omega = L_A + \frac{1}{\omega} D_A$ e $N_\omega = \left(1 - \frac{1}{\omega}\right) D_A + U_A$ e verificar que $M_\omega - N_\omega^\mathsf{T}$ é definida positiva:

$$M_\omega - N_\omega^\mathsf{T} = L_A + \frac{1}{\omega} D_A - \left(1 - \frac{1}{\omega}\right) D_A - U_A^\mathsf{T} = L_A + \frac{1}{\omega} D_A - \left(1 - \frac{1}{\omega}\right) D_A - L_A = \left(\frac{2}{\omega} - 1\right) D_A,$$

onde usámos $U_A^\mathsf{T} = L_A$ por A ser simétrica. Assim, $M_\omega - N_\omega^\mathsf{T}$ é definida positiva se e só se

$$\frac{2 - \omega}{\omega} > 0 \Leftrightarrow 0 < \omega < 2,$$

pois $a_{ii} = e_i^\mathsf{T} A e_i > 0$, $i = 1, ..., N$, por A ser definida positiva. □

Note-se que o Teorema 4.7 inclui o caso do método de Gauss-Seidel, e portanto, este método é convergente quando A é simétrica e definida positiva. No que diz respeito à aplicação do método de Jacobi a sistemas com matriz simétrica e definida positiva, tal é só deverá ser feito com uma condição adicional sobre A.

Teorema 4.8. *Seja $A \in \mathbb{R}^{N\times N}$ simétrica e definida positiva tal que $2D_A - A$ é definida positiva. Então o método de Jacobi é convergente para a solução do sistema $Ax = b$, qualquer que seja a iterada inicial $x^{(0)} \in \mathbb{R}^N$.*

Demonstração. O resultado é consequência de $M_A - N_A^{\mathsf{T}} = D_A - (A - D_A)^{\mathsf{T}} = D_A - (A - D_A) = 2D_A - A$. □

4.6 Algumas considerações sobre a aproximação de valores próprios

4.6.1 Método das potências

Este método gera uma sucessão de números $\lambda^{(k)}$, convergente para λ_1, valor próprio dominante da matriz $A \in \mathbb{R}^{n\times n}$, e uma sucessão de vetores que converge para um vetor próprio de A associado a λ_1.

O algoritmo consiste em:

(i) Escolher um vector $x^{(0)} \in \mathbb{R}^n$, tal que $\|x^{(0)}\|_\infty = 1$.

(ii) Para $k = 1, 2, 3...$, calcular:

$$z^{(k)} = Ax^{(k-1)};\ \lambda^{(k)} = \frac{z_i^{(k)}}{x_i^{(k-1)}}\ (i \text{ qualquer índice tal que } x_i^{(k-1)} \neq 0);\ x^{(k)} = \frac{z^{(k)}}{\lambda^{(k)}}.$$

(iii) Continuar o processo até que $|\lambda^{(k)} - \lambda^{(k-1)}| < \varepsilon$ onde ε é uma tolerância dada. Em alternativa, pode-se usar o critério de paragem: $|\lambda^{(k)} - \lambda^{(k-1)}|/|\lambda^{(k)}| < \varepsilon$.

Relativamente à **convergência** do método das potências, considere-se uma base ortonormada, $\{v^{(1)}, ..., v^{(n)}\}$, formada por vetores próprios de A. Escolhe-se $x^{(0)} \in \mathbb{R}^n$, tal que $\|x^{(0)}\|_\infty = 1$ e *com componente não nula relativamente ao vetor próprio associado ao valor próprio dominante*, ou seja,

$$x^{(0)} = \alpha_1 v^{(1)} + ... + \alpha_n v^{(n)} \qquad (\alpha_1 \neq 0).$$

Tem-se

$$z^{(1)} = Ax^{(0)}$$
$$x^{(1)} = \frac{z^{(1)}}{\lambda^{(1)}} = \frac{1}{\lambda^{(1)}} Ax^{(0)}$$
$$z^{(2)} = Ax^{(1)} = \frac{1}{\lambda^{(1)}} A^2 x^{(0)}$$
$$x^{(2)} = \frac{z^{(2)}}{\lambda^{(2)}} = \frac{1}{\lambda^{(1)}\lambda^{(2)}} A^2 x^{(0)}$$

e em geral:

$$z^{(k)} = \frac{1}{\lambda^{(1)} \dots \lambda^{(k-1)}} A^k x^{(0)},$$
$$x^{(k)} = \frac{1}{\lambda^{(1)} \dots \lambda^{(k)}} A^k x^{(0)},\ k = 2, 3, ...$$

Então

$$\begin{aligned}
\lim_{k\to\infty} \lambda^{(k)} &= \lim_{k\to\infty} \frac{z_i^{(k)}}{x_i^{(k-1)}} = \lim_{k\to\infty} \frac{(A^k x^{(0)})_i^{(k)}}{(A^{k-1} x^{(0)})_i^{(k-1)}} \\
&= \lim_{k\to\infty} \frac{\lambda_1^k \left[\alpha_1 v^{(1)} + \alpha_2 \left(\frac{\lambda_2}{\lambda_1}\right)^k v^{(2)} + \dots + \alpha_n \left(\frac{\lambda_n}{\lambda_1}\right)^k v^{(n)} \right]_i}{\lambda_1^{k-1} \left[\alpha_1 v^{(1)} + \alpha_2 \left(\frac{\lambda_2}{\lambda_1}\right)^{k-1} v^{(2)} + \dots + \alpha_n \left(\frac{\lambda_n}{\lambda_1}\right)^{k-1} v^{(n)} \right]_i} = \lambda_1
\end{aligned}$$

e como

$$x^{(k)} = \frac{\lambda_1^k}{\lambda^{(1)} \dots \lambda^{(k)}} \left[\alpha_1 v^{(1)} + \alpha_2 \left(\frac{\lambda_2}{\lambda_1}\right)^k v^{(2)} + \dots + \alpha_n \left(\frac{\lambda_n}{\lambda_1}\right)^k v^{(n)} \right],$$

$x^{(k)}$ tende a alinhar-se na mesma direção que $v^{(1)}$, quando $k \to \infty$.

Convém ainda fazer as seguintes observações sobre o método das potências:

- Em geral, não se sabe se uma matriz possui apenas um valor próprio dominante.
- Na prática não é possível assegurar a condição $\alpha_1 \neq 0$ em $x^{(0)}$, uma vez que não se conhece o vetor próprio $v^{(1)}$.
- Os erros de arredondamento que surgem na execução computacional do método das potências podem ajudar, pois podem conduzir ao aparecimento de uma componente não nula na direção de $v^{(1)}$. (Podemos dizer que este é um dos raros casos em que os erros de arredondamento nos ajudam!)

5 Resolução numérica de sistemas não-lineares

5.1 Método de Newton generalizado

Começamos com um exemplo e a sua resolução através do Mathematica.

Exemplo 5.1. *Consideramos o sistema de equações não lineares*

$$\begin{cases} x^2 + xy = 10 \\ y + 3xy^2 = 57 \end{cases}$$

No Mathematica, a função `FindRoot` *permite resolver (numericamente) equações não line-*

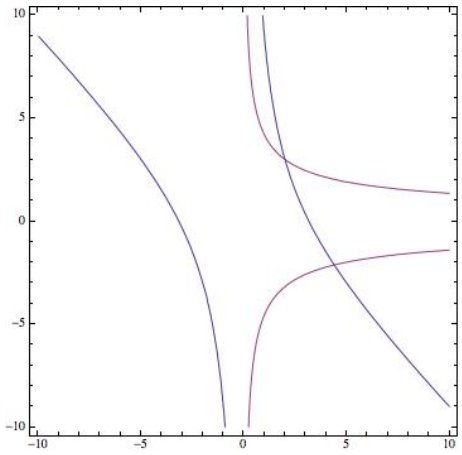

Figura 1: Localização das raízes através de um gráfico

ares. Neste caso,

$$In : \texttt{FindRoot}[\{x^2 + x * y == 10, y + 3x * y^2 == 57\}, \{\{x, 1.5\}, \{y, 3.5\}\}]$$
$$Out : \{x{-}{>}\ 2., y{-}{>}\ 3.\}$$

$$In : \texttt{FindRoot}[\{x^2 + x * y == 10, y + 3x * y^2 == 57\}, \{\{x, 5\}, \{y, -1\}\}]$$
$$Out : \{x{-}{>}\ 4.39374, y{-}{>}\ -2.11778\}$$

O comando FindRoot baseia-se no método de Newton, cujo algoritmo é dado por

$$\begin{cases} x^{(0)} \text{ dado} \\ J_f(x^{(k)})\Delta x^{(k)} = -f(x^{(k)}) \\ x^{(k+1)} = x^{(k)} + \Delta x^{(k)}, \quad k = 0, 1, 2, ... \end{cases}$$

onde J_f designa a matriz jacobiana de f.

Tem-se o seguinte resultado sobre a convergência local do método de Newton.

Teorema 5.1. *Seja $f \in C^2(V_z)$, onde V_z é uma vizinhança de z, zero de f tal que*

$$\det(J_f(z)) \neq 0.$$

Então o método de Newton converge para z desde que $x^{(0)}$ esteja suficientemente perto de z.

Demonstração. Ver [8]. □

Exemplo 5.2. *Consideremos novamente o sistema*

$$\begin{cases} x_1^2 + x_1 x_2 = 10 \\ x_2 + 3x_1 x_2^2 = 57 \end{cases} \Leftrightarrow f(x) = 0$$

com

$$f : \mathbb{R}^2 \to \mathbb{R}^2$$
$$f(x_1, x_2) = \begin{bmatrix} x_1^2 + x_1 x_2 - 10 \\ x_2 + 3x_1 x_2^2 - 57 \end{bmatrix}.$$

A matriz jacobiana de f é dada por

$$J_f(x_1, x_2) = \begin{bmatrix} 2x_1 + x_2 & x^{(1)} \\ 3x_2^2 & 1 + 6x_1 x_2 \end{bmatrix}.$$

Vamos aplicar o método de Newton com $x^{(0)} = [1.5 \;\; 3.5]^\mathsf{T}$.

A primeira iterada do método de Newton $x^{(1)} = x^{(0)} + \Delta x^{(0)}$ obtém-se resolvendo o sistema linear

$$J_f(x^{(0)})\Delta x^{(0)} = -f(x^{(0)}).$$

Tem-se

$$f(x^{(0)}) = \begin{bmatrix} -2.5 \\ 1.625 \end{bmatrix} \qquad J_f(x^{(0)}) = \begin{bmatrix} 6.5 & 1.5 \\ 36.75 & 32.5 \end{bmatrix}$$

e

$$\Delta x^{(0)} = [0.536029 \quad -0.656125]^{\mathsf{T}}$$

pelo que

$$x^{(1)} = x^{(0)} + \Delta x^{(0)} = [2.03603 \;\; 2.84388]^{\mathsf{T}}.$$

Analogamente, $x^{(2)} = x^{(1)} + \Delta x^{(1)}$*, onde* $\Delta x^{(1)}$ *é a solução do sistema linear*

$$J_f(x^{(1)})\Delta x^{(1)} = -f(x^{(1)}).$$

Com $x^{(1)} = [2.03603 \;\; 2.84388]^{\mathsf{T}}$*, tem-se*

$$f(x^{(1)}) = \begin{bmatrix} -0.0643568 \\ -4.756 \end{bmatrix} \qquad J_f(x^{(1)}) = \begin{bmatrix} 6.91594 & 2.03603 \\ 24.263 & 35.7413 \end{bmatrix}$$

pelo que

$$\Delta x^{(1)} = [-0.0373293 \;\; 0.158408]^{\mathsf{T}}$$

e

$$x^{(2)} = x^{(1)} + \Delta x^{(1)} = [1.9987 \;\; 3.00229]^{\mathsf{T}}.$$

A terceira iterada é ainda mais precisa:

$$x^{(3)} = [1.99999... \;\; 2.99999...]^{\mathsf{T}}.$$

Note-se que $x = [2 \;\; 3]^{\mathsf{T}}$ *é solução do sistema.*

5.2 Método do ponto fixo para sistemas de equações não-lineares

Recordamos:

Definição 5.1. *Um conjunto* $X \subseteq \mathbb{R}^N$ *diz-se convexo se*

$$t \in [0,1],\ x, y \in X \Rightarrow tx + (1-t)y \in X,$$

i.e., se x *e* y *pertencem a* X *então todos os pontos do segmento de reta que une* x *e* y *também pertencem a* X*.*

Teorema 5.2. *Seja* $X \subseteq \mathbb{R}^N$ *não vazio, fechado e convexo. Se* $G \in C^1(X; \mathbb{R}^N)$ *e satisfaz*

(i) $G(X) \subseteq X$*;*

(ii) $L := \sup_{x \in X} \|J_G(x)\| < 1$*, onde* $\|\cdot\|$ *é uma norma matricial induzida por uma norma vetorial, que será representada também por* $\|\cdot\|$*,*

então

1. *G tem um e um só ponto fixo em X, $z = G(z)$;*
2. *O método do ponto fixo $x^{(n+1)} = G(x^{(n)})$, $n \in \mathbb{N}_0$, converge para z, qualquer que seja $x^{(0)} \in X$;*
3. *São válidas as majorações para os erros $z - x^{(n)}$*

$$\|z - x^{(n)}\| \leq L^n \|z - x^{(0)}\|,$$

$$\|z - x^{(n)}\| \leq \frac{L^n}{1-L} \|x^{(1)} - x^{(0)}\|,$$

$$\|z - x^{(n+1)}\| \leq \frac{L}{1-L} \|x^{(n+1)} - x^{(n)}\|,\ n \in \mathbb{N}_0.$$

Demonstração. Para cada par $(x, y) \in X$, seja $\psi : [0, 1] \to \mathbb{R}^N$ dada por

$$\psi(t) = G(y + t(x - y)).$$

Tem-se $\psi \in C^1([0, 1]; \mathbb{R}^N)$ e

$$\psi_i'(t) = \sum_{j=1}^{N} (x - y)_j \frac{\partial G_i}{\partial x_j}(y + t(x - y))),\ i = 1, ..., N,$$

donde

$$G(x) - G(y) = \psi(1) - \psi(0) = \int_0^1 \psi'(t)dt = \int_0^1 J_G(y + t(x - y))(x - y)dt.$$

Como

$$\begin{aligned} \|G(x) - G(y)\| &= \| \int_0^1 J_G(y + t(x - y))(x - y)dt\| \\ &\leq \int_0^1 \|J_G(y + t(x - y))\| \|x - y\| dt \\ &\leq \sup_{t\in[0,1]} \|J_G((y + t(x - y))(x - y))\| \|x - y\|,\ \forall x, y \in X, \end{aligned}$$

se $L := \sup_{x\in X} \|J_G(x)\| < 1$ então G é contrativa em X. □

O Teorema do ponto fixo em $\mathbb{R}^N$ pode ser imediatamente generalizado ao caso de um espaço de Banach (ver, p. ex., [3]). Diz-se que um espaço normado $(E, \|\cdot\|)$ é espaço de Banach se toda a sucessão de Cauchy for convergente.

Referências

[1] C. Alves, Fundamentos de Análise Numérica, 2001.

[2] R. Bagnara, A unified proof for the convergence of Jacobi and Gauss-Seidel methods, SIAM Review, Vol. 37, No. 1, 93–97, 1995.

[3] R. Kress, Numerical Analysis, Springer-Verlag, 1998.

[4] E. Kreyszig, Introductory Functional Analysis with Applications, John Wiley, 1978.

[5] L. Loura, Tópicos de Análise Numérica, Instituto Superior Técnico, 1990.

[6] A. Neumaier, Solving ill-conditioned and singular linear systems: a tutorial on regularization, SIAM Review, Vol. 40, No. 3, 636–666, 1998.

[7] P. Olver, Numerical Analysis Lecture Notes, 2008
(online: http://www.users.math.umn.edu/ olver/num-/lni.pdf).

[8] J. M. Ortega, Numerical Analysis: a second course, Classics in Applied Mathematics; Vol. 3, SIAM, 1990.

[9] A. Quarteroni, R. Sacco e F. Saleri, Cálculo Científico com Matlab e Octave, Springer-Verlag, 2007 (traduzido por Adélia Sequeira).

[10] F. Romeiras, Matemática Computacional, Apontamentos das aulas, 2008.

[11] J. A. Tropp, An elementary proof of the spectral radius formula for matrices, 2001
(online: http://users.cms.caltech.edu/ jtropp/notes/Tro01-Spectral-Radius.pdf)