Apuntes de

CÁLCULO NUMÉRICO II

Curso 2008/2009

Septiembre de 2008

Índice general

Tema 1

Elementos de Álgebra lineal. Normas.

1.1. Normas

Sea V un espacio vectorial sobre un cuerpo $\mathbb{K}$ de escalares ($\mathbb{K} = \mathbb{R}$ o $\mathbb{C}$)

Definición 1.1 *Una norma sobre V (norma vectorial) es una aplicación $\|\cdot\| : V \to \mathbb{R}_+$ que verifica*

1. $\|v\| \geq 0, \quad \forall v \in V \ y \ \|v\| = 0 \Leftrightarrow v = \theta$
2. $\|\alpha v\| = |\alpha|\|v\|, \quad \forall \alpha \in \mathbb{K}, \ \forall v \in V$
3. $\|u + v\| \leq \|u\| + \|v\|, \quad \forall u, v \in V$ *(desigualdad triangular)*

Al par $(V, \|\cdot\|)$ se le llama espacio normado.

Propiedades que se deducen de esta definición son:

1. $\|u - v\| \leq \|u\| + \|v\|, \quad \forall u, v \in V$
2. $| \ \|u\| - \|v\| \ | \leq \|u \pm v\|, \quad \forall u, v \in V$

Si V es normado, se puede convertir en espacio métrico para la distancia

$$d(u, v) = \|u - v\|, \quad \forall u, v \in V$$

La base de entornos de la topología es

$$\{B(a, \delta), \ a \in V, \ \delta \in \mathbb{R}_+\} \quad \text{donde} \quad B(a, \delta) = \{x \in V : \ \|x - a\| < \delta\}$$

Es también inmediato probar que la aplicación $\|\cdot\| : (V, \|\cdot\|) \to (\mathbb{R}_+, |\cdot|)$ es continua.

Definición 1.2 *Dos normas son equivalentes sobre V si inducen el mismo espacio topológico.*

Son resultados importantes y conocidos

Teorema 1.1 *$\|\cdot\|_1$ y $\|\cdot\|_2$ son equivalentes sobre V si y solo si existen dos constantes $C_1,\ C_2 > 0$ tales que*

$$C_1\|v\|_1 \leq \|v\|_2 \leq C_2\|v\|_1, \quad \forall v \in V$$

Teorema 1.2 *Si V es de dimensión finita, todas las normas que se pueden definir sobre V son equivalentes.*

Si V tiene dimensión n, dada una base de V se puede identificar V con $\mathbb{K}^N$ mediante sus componentes en dicha base: $v = (v_1, ..., v_n)^t$

Ejemplos 1.1 *Ejemplos de normas vectoriales son*

1. $\|v\|_1 = \sum_{i=1}^{n} |v_i|$
2. $\|v\|_2 = \left(\sum_{i=1}^{n} |v_i|^2\right)^{1/2}$ *(norma euclídea)*
3. $\|v\|_p = \left(\sum_{i=1}^{n} |v_i|^p\right)^{1/p}$ *para $p \geq 1$ (p-norma)*
4. $\|v\|_\infty = \max_{1\leq i\leq n} |v_i|$ *(norma del máximo, norma uniforme) ($\lim_{p\to\infty} \|v\|_p = \|v\|_\infty$)*

En V pueden definirse productos escalares a través de $\mathbb{K}$. Los usuales son:

1. Si $\mathbb{K} = \mathbb{R}$, el producto escalar euclídeo viene dado por

$$(\cdot,\cdot) : V \times V \to \mathbb{R}, \quad (u, v) = u \cdot v = v^t u = u^t v = \sum_{i=1}^{n} u_i v_i$$

2. Si $\mathbb{K} = \mathbb{C}$, el producto escalar hermítico viene dado por

$$(\cdot,\cdot) : V \times V \to \mathbb{C}, \quad (u, v) = u \cdot v = v^* u = \overline{u^* v} = \sum_{i=1}^{n} u_i \overline{v_i}$$

donde $\overline{u_i}$ es el conjugado de u_i, u^t es el vector traspuesto de u y u^* es el vector adjunto de u, es decir el conjugado traspuesto.

Según esto, la norma inducida por el producto escalar es la norma euclídea. Estos productos escalares son los que se utilizarán para hablar de bases ortogonales u ortonormales a lo largo del curso. En la base ortonormal, por ejemplo, se verificará

$$(u_i, u_j) = 0, \quad i \neq j; \quad \|u_i\|_2 = 1, \quad i, j = 1, ..., n$$

Definición 1.3 *Sea $(V, \|\cdot\|)$ un espacio normado. Se dice que $\{v_k\} \subset V$ converge a $v \in V$, y se denota $v_k \to v$ o $\lim_{k\to+\infty} v_k = v$ si $\lim_{k\to+\infty} \|v_k - v\| = 0$. v se llama el límite de $\{v_k\} \subset V$.*

Si la dimensión de V es finita, la equivalencia de las normas implica que la convergencia de una sucesión es independiente de la norma elegida. Si se considera cualquiera de las normas del ejemplo anterior, se ve que la convergencia de una sucesión equivale a la convergencia por componentes

$$u_k \to u \quad \text{en } \mathbb{K}^n \iff u_k^i \to u^i \quad \text{en } \mathbb{K},\ 1 \leq i \leq n$$

1.2. Normas matriciales

Sea $\mathcal{M}$ el anillo de las matrices de orden n sobre $\mathbb{K}$.

Definición 1.4 *Una norma matricial es una aplicación $\|\cdot\| : \mathcal{M} \to \mathbb{R}_+$ que verifica:*

1. $\|A\| \geq 0, \quad \forall A \in \mathcal{M}$ *y* $\|A\| = 0 \Leftrightarrow A = \theta$
2. $\|\alpha A\| = |\alpha|\|A\|, \quad \forall \alpha \in \mathbb{K},\ \forall A \in \mathcal{M}$
3. $\|A + B\| \leq \|A\| + \|B\|, \quad \forall A, B \in \mathcal{M}$
4. $\|AB\| \leq \|A\|\ \|B\|, \quad \forall A, B \in \mathcal{M}$

Nótese que una matriz de $\mathcal{M}$ puede considerarse como un vector de n^2 componentes en $\mathbb{K}$, pero la propiedad d) diferencia las normas matriciales de las vectoriales.

Ejemplos 1.2 *Sea $A = (a_{ij})_{1\leq i,j\leq n} \in \mathcal{M}$.*

1. *Son ejemplos de normas matriciales las siguientes:*
   - $\|A\|_1 = \sum_{i,j=1}^{n} |a_{ij}|$ *(norma 1)*

- $\|A\|_2 = \left(\sum_{i,j=1}^{n} |a_{ij}|^2 \right)^{1/2} = \|A\|_{ES}$ *(norma de Erhard Schmidt)*
- $\|A\|_p = \left(\sum_{i,j=1}^{n} |a_{ij}|^p \right)^{1/p}, \quad p \in [1,2]$ *(p-norma matricial)*

2. No es norma matricial la siguiente $\|A\| = \max\limits_{1 \le i,j \le n} |a_{ij}|$.

Antes de ver las primeras propiedades de las normas matriciales, recordamos los siguientes conceptos previos:

Definición 1.5 *Se dice que* $\lambda \in \mathbb{R}$ *o* $\mathbb{C}$ *es una autovalor o valor propio de* A *si*

$$\exists v \in V,\ v \neq \theta :\ Av = \lambda v$$

En tal caso, v *es un autovector o vector propio asociado a* λ.

Puesto que

$$Av = \lambda v \Leftrightarrow (\lambda I - A)v = \theta \Leftrightarrow |\lambda I - A| = 0$$

resulta que los autovalores de A son las raíces del polinomio característico $p_A(\lambda) = |\lambda I - A|$. Son por tanto n números reales o complejos distintos como máximo; si la matriz es real, los autovalores complejos aparecen por parejas conjugadas.

Definición 1.6 *Se llama espectro de* A *y se denota* $\mathrm{sp}(A)$ *al conjunto de los autovalores de* A.

Definición 1.7 *Se llama radio espectral de* A *a* $\rho(A) = \max\{|\lambda_i(A)|,\ i = 1, ..., n\}$.

Proposición 1.1 *Son propiedades de las normas matriciales las siguientes*

1. $\|A^k\| \le \|A\|^k, \quad \forall A \in \mathcal{M},\ \forall k\ \mathbb{N}$
2. $\|I\| \ge 1$
3. *Si* $\|A\| < 1$, *entonces* $A^k \to \theta$
4. $\rho(A) \le \|A\|$, *para cualquier norma matricial.*

Demostración:

1. Es consecuencia inmediata de la propiedad d) de las normas matriciales.
2. Sigue de la anterior haciendo $A = I$ y $k = 2$.

3. Hay que probar que $\lim_{k\to+\infty} \|A^k - \theta\| = 0$. Pero

$$\|A^k\| \leq \|A\|^k \to 0, \quad \text{por ser } \|A\| < 1$$

4. Sea $\lambda \in \text{sp}(A)$ y v un autovector asociado. Entonces

$$A(v|\theta|...|\theta) = \lambda(v|\theta|...|\theta) \Rightarrow \|A(v|\theta|...|\theta)\| = \|\lambda(v|\theta|...|\theta)\| \Rightarrow$$

$$|\lambda| \ \|(v|\theta|...|\theta)\| \leq \|A\| \ \ \|(v|\theta|...|\theta)\|$$

y $\|(v|\theta|...|\theta)\| > 0$ porque esta matriz es no nula. De modo que $|\lambda| \leq \|A\|$. Como esto vale para cualquier $\lambda \in \text{sp}(A)$, sigue la propiedad. ◇

Hay que hacer notar que puede darse la desigualdad estricta. Así, por ejemplo, si $A = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$, entonces, $\rho(A) = 0 < \|A\|$ para cualquier norma matricial.

1.3. Normas consistentes

Definición 1.8 *Se dice que una norma matricial es consistente con una norma vectorial si*

$$\|Av\| \leq \|A\| \ \|v\|, \quad \forall A \in \mathcal{M}, \ \forall v \in V$$

Proposición 1.2 *Dada una norma matricial cualquiera, siempre existe una norma vectorial con la que es consistente.*

Demostración: Sea $\|\cdot\|$ una norma matricial y $v \in V$ un vector cualquiera. Definimos

$$\|v\| = \|(v|\theta|...|\theta)\|$$

Evidentemente se trata de una norma vectorial, y además

$$\|Av\| = \|(Av|\theta|...|\theta)\| = \|A(v|\theta|...|\theta)\| \leq \|A\| \ \|(v|\theta|...|\theta)\| = \|A\| \ \|v\|$$

c.q.d. ◇

Ejemplo 1.1 *Puede comprobarse que si se aplica la proposición anterior a la p-norma matricial, $p \in [1, 2]$, resulta la p-norma vectorial.*

Teorema 1.3 *(Inversión de matrices de la forma $I \pm B$).*
Sea $\|\cdot\|$ una norma matricial y $B \in \mathcal{M}$ tal que $\|B\| < 1$. Entonces, $I \pm B$ es invertible y

$$\frac{\|I\|}{\|I\| + \|B\|} \leq \|(I \pm B)^{-1}\| \leq \frac{\|I\|}{1 - \|B\|} \tag{3.1}$$

Demostración: Haremos la demostración para $I + B$; es análoga para $I - B$. Supongamos que $\exists u \neq \theta$ tal que $(I + B)u = \theta$. Entonces, se tiene

$$(I + B)u = \theta \Rightarrow -u = Bu \Rightarrow -1 \in \text{sp } (B) \Longrightarrow 1 \leq \rho(B) \leq \|B\|$$

que es contradictorio con la hipótesis.
Además, de la igualdad $(I + B)^{-1}(I + B) = I$ se deducen

a)

$$(I + B)^{-1} = I - (I + B)^{-1}B \Rightarrow \|(I + B)^{-1}\| \leq \|I\| + \|(I + B)^{-1}\| \, \|B\| \Rightarrow$$

$$\|(I \pm B)^{-1}\| \leq \frac{\|I\|}{1 - \|B\|}$$

b)

$$\|I\| \leq \|(I + B)^{-1}\| \, \|(I + B)\| \leq \|(I + B)^{-1}\|(\|I\| + \|B\|) \Rightarrow$$

$$\frac{\|I\|}{\|I\| + \|B\|} \leq \|(I \pm B)^{-1}\|$$

c.q.d. ◇

Corolario 1.1 *Sean $A \in \mathcal{M}$ invertible y $B \in \mathcal{M}$ tales que $\|B\| \, \|A^{-1}\| < 1$. Entonces, $A + B$ es invertible y*

$$\|(A + B)^{-1}\| \leq \frac{\|I\| \, \|A^{-1}\|}{1 - \|A^{-1}\| \, \|B\|}$$

Demostración: Tenemos que $\|A^{-1}B\| \leq \|A^{-1}\| \, \|B\| < 1$. Por su parte, $A + B = A(I + A^{-1}B)$ es invertible porque A lo es por hipótesis y $I + A^{-1}B$ lo es por el Teorema 1.3. Por el mismo,

$$\|(A + B)^{-1}\| = \|(I + A^{-1}B)^{-1}A^{-1}\| \leq \|(I + A^{-1}B)^{-1}\| \, \|A^{-1}\| \leq$$

$$\frac{\|I\| \, \|A^{-1}\|}{1 - \|A^{-1}B\|} \leq \frac{\|I\| \, \|A^{-1}\|}{1 - \|A^{-1}\| \, \|B\|}$$

c.q.d. ◇

1.4. Teorema de Schur

Recordamos algunas definiciones

Definición 1.9 *Sea $A = (a_{ij})_{1\leq i,j\leq n} \in \mathcal{M}$. Se llaman*

- *matriz traspuesta de A a $A^t = (a_{ji})$*
- *matriz adjunta de A a $A^* = \overline{A^t} = (\overline{a_{ji}})$*

Definición 1.10 *Sea $A \in \mathcal{M}$. Se dice que*

- *A es simétrica si A es real y $A = A^t$.*
- *A es hermítica si $A = A^*$.*
- *A es ortogonal si A es real y $AA^t = A^tA = I$.*
- *A es unitaria si $A^*A = AA^* = I$.*
- *A es normal si $A^*A = AA^*$.*

Nota:

Si A es unitaria, sus columnas constituyen una base ortonormal de $\mathbb{K}^n$ y recíproca- mente.

■

Definición 1.11 *Una matriz $A \in \mathcal{M}$ se dice triangularizable si es semejante a una matriz triangular, es decir, si existen $B \in \mathcal{M}$ regular y $T \in \mathcal{M}$ triangular tales que $T = B^{-1}AB$.*

Teorema 1.4 *(Schur). Dada $A \in \mathcal{M}_n$, existen $U \in \mathcal{M}_n$ unitaria y $T \in \mathcal{M}_n$ triangular tales que $U^*AU = T$. Es decir, toda matriz es semejante a una matriz triangular con matriz de paso unitaria.*

Notas:

1. Los elementos de la diagonal de T son los autovalores de A. En efecto, A y T son semejantes y tienen el mismo polinomio característico y las matrices triangulares tienen por autovalores los elementos de su diagonal.
2. A consecuencia de la nota anterior, cabe que una matriz real tenga todos sus autovalores complejos y por tanto que la descomposición $A = UTU^*$ sea de matrices complejas.

3. Se hace la demostración obteniendo una T triangular superior. Si se quisiera obtener una triangular inferior basta aplicar el Teorema a A^*. En efecto, si existe U unitaria tal que $U^*A^*U = T$, se deduce tomando adjuntos que $U^*AU = T^*$ y T^* es triangular inferior.

4. Las matrices U y T no son únicas. Considérese, por ejemplo, el caso $A = I$

■

Demostración: Se hace por inducción sobre n, la dimensión de la matriz. Si $n = 1$, el resultado es trivial. Supongámoslo cierto para $n - 1$.
Sea $\lambda \in \text{sp}(A)$ y v un autovector asociado normalizado. Mediante un proceso de ortonormalización de Gram-Schmidt de una base que lo contenga, obtenemos una base ortonormal $\{v, v^2, ..., v^n\}$. Denotemos $V = (v|v^2|...|v^n)$ que es una matriz unitaria. Entonces

$$AV = A(v|v^2|...|v^n) = (\lambda v|Av^2|...|Av^n)$$

Si expresamos cada vector Av^j en la base que tenemos, se obtienen

$$Av^j = \alpha_j v + b_{2j}v^2 + ... + b_{nj}v^n, \quad j = 2, ..., n$$

de modo que se puede escribir

$$AV = (\lambda v|Av^2|...|Av^n) = (v|v^2|...|v^n) \begin{pmatrix} \lambda & \alpha_2 & \dots & \alpha_n \\ 0 & & & \\ \vdots & & B & \\ 0 & & & \end{pmatrix}$$

Por inducción puede probarse que los autovalores de B son los de A menos el que se ha considerado ya, λ.
Por la hipótesis de inducción para B, sabemos que existen $W_{n-1} \in \mathcal{M}_{n-1}$ unitaria y $T_{n-1} \in \mathcal{M}_{n-1}$ triangular superior tales que $BW_{n-1} = W_{n-1}T_{n-1}$. Definimos

$$W \in \mathcal{M}_n, \quad W = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & & & \\ \vdots & & W_{n-1} & \\ 0 & & & \end{pmatrix}$$

Esta matriz es unitaria, pues, en efecto

$$W^*W = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & & & \\ \vdots & & W^*_{n-1} & \\ 0 & & & \end{pmatrix} \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & & & \\ \vdots & & W_{n-1} & \\ 0 & & & \end{pmatrix} = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & & & \\ \vdots & & I_{n-1} & \\ 0 & & & \end{pmatrix} = I_n$$

Entonces, vamos a probar que $U = VW$ es la matriz unitaria que necesitamos

$$AVW = V \begin{pmatrix} \lambda & \alpha_2 & \dots & \alpha_n \\ 0 & & & \\ \vdots & & B & \\ 0 & & & \end{pmatrix} \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & & & \\ \vdots & & W_{n-1} & \\ 0 & & & \end{pmatrix} =$$

$$= V \begin{pmatrix} \lambda & \beta_2 & \dots & \beta_n \\ 0 & & & \\ \vdots & & BW_{n-1} & \\ 0 & & & \end{pmatrix} = V \begin{pmatrix} \lambda & \beta_2 & \dots & \beta_n \\ 0 & & & \\ \vdots & & W_{n-1}T_{n-1} & \\ 0 & & & \end{pmatrix} =$$

$$= V \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & & & \\ \vdots & & W_{n-1} & \\ 0 & & & \end{pmatrix} \begin{pmatrix} \lambda & \beta_2 & \dots & \beta_n \\ 0 & & & \\ \vdots & & T_{n-1} & \\ 0 & & & \end{pmatrix} = VWT$$

siendo T triangular superior. Basta llamar ahora $U = VW, \quad U \in \mathcal{M}_n$ que es unitaria por ser el producto de dos matrices unitarias y comprobar que se verifica $AU = UT$. $\diamond$

Corolario 1.2 *Sea $A \in \mathcal{M}$. Entonces, A es normal si y solo si existe $U \in \mathcal{M}$ unitaria tal que $U^*AU = D$, siendo D diagonal. Es decir, las matrices normales son las matrices diagonalizables con matriz de paso unitaria.*

Demostración: Supongamos que A es normal y sean U unitaria y T triangular superior tales que $U^*AU = T$. Entonces,

- T es normal porque

$$TT^* = U^*AUU^*A^*U = U^*A^*AU = U^*A^*UU^*AU = T^*T$$

- T es diagonal, porque

$$\begin{cases} (T^*T)_{11} = |t_{11}|^2 \\ (TT^*)_{11} = \sum_{k=1}^{n} |t_{1k}|^2 \end{cases} \Rightarrow t_{1k} = 0, \quad k = 2, ..., n$$

y en general

$$\begin{cases} (T^*T)_{ii} = |t_{ii}|^2 \\ (TT^*)_{ii} = \displaystyle\sum_{k=i}^{n} |t_{ik}|^2 \end{cases} \Rightarrow t_{ik} = 0, \quad k = i+1, ..., n, \quad i = 1, ..., n-1$$

lo que prueba que T es diagonal.

Recíprocamente, sean U unitaria y D diagonal tales que $U^*AU = D$. Entonces, $U^*A^*U = D^*$ y

$$\begin{cases} DD^* = U^*AUU^*A^*U = U^*AA^*U = \text{diag } (|\lambda_i|^2) \\ D^*D = U^*A^*UU^*AU = U^*A^*AU = \text{diag } (|\lambda_i|^2) \end{cases} \Rightarrow$$

$$U^*AA^*U = U^*A^*AU \Rightarrow AA^* = A^*A$$

c.q.d. ◇

Nota:

La matriz de paso está constituida por los autovectores de A. De modo que dada una matriz normal existe siempre una base ortonormal de autovectores asociados. ■

Corolario 1.3 *Se verifica*

1. *Los autovalores de las matrices hermíticas y simétricas son reales*
2. *det* $(A) = \displaystyle\prod_{i=1}^{n} \lambda_i(A), \quad \forall A \in \mathcal{M}$
3. $\lambda_i(A^k) = (\lambda_i(A))^k, \quad i = 1, ..., n, \ k \in \mathbb{N}$. *En particular,* $\rho(A^k) = \rho(A)^k$.

Demostración:

1. Si A es hermítica (o simétrica si es real), entonces es normal y por el Corolario 1.2, existe U unitaria tal que $U^*AU = D = \text{diag } (\lambda_i(A))$. Pero

   $$D^* = U^*A^*U = U^*AU = D$$

   de modo que D es hermítica (o simétrica) y

   $$\lambda_i(A) = \overline{\lambda_i(A)} \Rightarrow \lambda_i(A) \in \mathbb{R}, \quad i = 1, ..., n$$

2. Basta tomar determinantes en la igualdad del Teorema de Schur.

3. Por el Teorema de Schur $A^k = UT^kU^*$. Basta ahora tener en cuenta que T^k es también una matriz triangular cuya diagonal tiene por elementos los de la diagonal de T elevados a k.

◇

De forma similar, para matrices simétricas el razonamiento se puede hacer en $\mathbb{R}$. Como en el Corolario 1.2 se deduce que

Corolario 1.4 *Si $A \in \mathcal{M}_n(\mathbb{R})$, entonces A es simétrica si y solo si existe una matriz real ortogonal $\mathcal{O}$ y una matriz diagonal D tales que $\mathcal{O}^tA\mathcal{O} = D$. Es decir, las matrices simétricas son las matrices reales diagonalizables con matriz de paso ortogonal.*

1.5. El Teorema de Courant-Fisher

Definición 1.12 *Sea $A \in \mathcal{M}_n(\mathbb{C})$. Se llama cociente de Rayleigh de A a la aplicación*

$$R_A : \mathbb{C}^n \setminus \{\theta\} \to \mathbb{C}, \quad R_A(v) = \frac{v^*Av}{v^*v}, \ v \neq \theta$$

Proposición 1.3 *1. El cociente de Rayleigh de una matriz hermítica toma sólo valores reales.*

2. Se verifica que

$$R_A(\alpha v) = R_A(v), \quad \forall \alpha \in \mathbb{C} \setminus \{\theta\}, \ \forall v \in \mathbb{C}^n \setminus \{\theta\}$$

Demostración:

1. Se tiene siempre que

$$\overline{v^*Av} = v^t\overline{A}\overline{v} = (A^*v)^t\overline{v} = ((A^*v)^t\overline{v})^t = v^*A^*v$$

y en consecuencia, es siempre cierto que $\overline{R_A(v)} = R_{A^*}(v)$. Si A es hermítica

$$\overline{R_A(v)} = R_A(v) \Rightarrow R_A(v) \in \mathbb{R}, \quad \forall v \neq \theta$$

2. Dados $\alpha \in \mathbb{C} \setminus \{\theta\}$ y $v \in V \setminus \{\theta\}$

$$R_A(\alpha v) = \frac{(\alpha v)^*A(\alpha v)}{(\alpha v)^*(\alpha v)} = \frac{|\alpha|^2 v^*Av}{|\alpha|^2 v^*v} = R_A(v)$$

c.q.d. ◇

Se recuerda que toda matriz hermítica es diagonalizable con autovalores reales y para la que siempre es posible encontrar una base ortonormal de autovectores asociados.

Teorema 1.5 *(Courant-Fisher). Sea $A \in \mathcal{M}_n(\mathbb{C})$ hermítica de autovalores $\lambda_1 \leq ... \leq \lambda_n$ y $\{p^1, ..., p^n\}$ una base ortonormal de autovectores asociados. Para $k = 1, ...n$, denotamos $\mathcal{V}_k = \{$subespacios de $\mathbb{C}^n$ de dimension $k\}$ y $V_k = \langle p^1, ...p^k \rangle$ y $V_0 = \mathcal{V}_0 = \{\theta\}$. Entonces, se verifica*

1. $\lambda_k = R_A(p^k), \quad k = 1, ..., n$
2. $\lambda_k = \max\limits_{v \in V_k \setminus \{\theta\}} R_A(v)$. *En particular,* $\lambda_n = \max\limits_{v \in V \setminus \{\theta\}} R_A(v)$.
3. $\lambda_k = \min\limits_{\substack{v \perp V_{k-1} \\ v \neq \theta}} R_A(v)$. *En particular,* $\lambda_1 = \min\limits_{v \in V \setminus \{\theta\}} R_A(v)$
4. $\lambda_k = \min\limits_{W \in \mathcal{V}_k} \max\limits_{v \in W \setminus \{\theta\}} R_A(v)$
5. $\lambda_k = \max\limits_{W \in \mathcal{V}_{k-1}} \min\limits_{\substack{v \perp W \\ v \neq \theta}} R_A(v)$
6. $\{R_A(v) : \ v \in V \setminus \{\theta\}\} = [\lambda_1, \lambda_n]$

Demostración: No justificamos los apartados 4) y 5). (Ver libro P. G. Ciarlet [2]).

1. $$R_A(p^k) = \frac{(p^k)^* A p^k}{(p^k)^* p^k} = \frac{(p^k)^* \lambda_k p^k}{(p^k)^* p^k} = \lambda_k$$

   Además, si v^k es un autovector cualquiera asociado a λ_k, también se tiene que $R_A(v^k) = \lambda_k$.

2. Sea $v \in V_k \setminus \{\theta\}$. Entonces, $v = \alpha_1 p^1 + ... + \alpha_k p^k$ y

   $$R_A(v) = \frac{(\alpha_1 p^1 + ... + \alpha_k p^k)^* A (\alpha_1 p^1 + ... + \alpha_k p^k)}{(\alpha_1 p^1 + ... + \alpha_k p^k)^* (\alpha_1 p^1 + ... + \alpha_k p^k)} = \frac{\sum\limits_{i=1}^{k} \lambda_i |\alpha_i|^2}{\sum\limits_{i=1}^{k} |\alpha_i|^2} \leq \lambda_k$$

   Por tanto $R_A(v) \leq \lambda_k, \quad \forall v \in V_k \setminus \{\theta\}$ lo que junto con la propiedad a) implica que $\lambda_k = \max\limits_{v \in V_k \setminus \{\theta\}} R_A(v)$.

3. Sea $v \in V_{k-1}^{\perp} \setminus \{\theta\}$. Entonces, $v = \alpha_k p^k + ... + \alpha_n p^n$ y

   $$R_A(v) = \frac{(\alpha_k p^k + ... + \alpha_n p^n)^* A (\alpha_k p^k + ... + \alpha_n p^n)}{(\alpha_k p^k + ... + \alpha_n p^n)^* (\alpha_k p^k + ... + \alpha_n p^n)} = \frac{\sum\limits_{i=k}^{n} \lambda_i |\alpha_i|^2}{\sum\limits_{i=k}^{n} |\alpha_i|^2} \geq \lambda_k$$

Por tanto $R_A(v) \geq \lambda_k, \quad \forall v \in V_{k-1}^{\perp} \setminus \{\theta\}$ lo que junto con la propiedad a) implica que $\lambda_k = \min_{v \in V_{k-1}^{\perp} \setminus \{\theta\}} R_A(v)$.

4. Es evidente que $R_A(V \setminus \{\theta\}) \subset [\lambda_1, \lambda_n]$. Veamos la inclusión contraria. Sea $\partial B_1 = \{z \in V : \ |z| = 1\}$. Tomando $\alpha = \dfrac{1}{\|v\|}$ en la Proposición 1.3 b), se obtiene que $R_A(V \setminus \{\theta\}) = R_A(\partial B_1)$. Se considera entonces la aplicación $v \in \partial B_1 \mapsto R_A(v) \in \mathbb{R}$ que es continua. Como ∂B_1 es conexo, también lo es $R_A(\partial B_1)$. Por tanto, $R_A(V \setminus \{\theta\})$ es un intervalo (que son los conexos de $\mathbb{R}$) que contiene a $\lambda_1 = R_A(p^1)$ y a $\lambda_n = R_A(p^n)$. De modo que $[\lambda_1, \lambda_n] \subset R_A(V \setminus \{\theta\})$.

◇

Corolario 1.5 *Si $A \in \mathcal{M}_n(\mathbb{C})$ es hermítica, entonces*

$$\lambda_1 v^* v \leq v^* A v \leq \lambda_n v^* v \qquad \forall v \in \mathbb{C}^n.$$

Demostración: Inmediata. ◇

1.6. Matrices definidas positivas

Sea $A \in \mathcal{M}$ una matriz hermítica.

Definición 1.13 *Se dice que A es semidefinida positiva (resp. definida positiva) si*

$$v^* A v \geq 0 \quad (resp\ v^* A v > 0), \quad \forall v \in \mathbb{C}^n \setminus \{\theta\}$$

Análogamente se definen las matrices semidefinidas negativas y definidas negativas

Lema 1.1 *Se verifica*

1. *Si A es definida positiva, entonces A es regular.*
2. *Si $A \in M$ cualquiera, entonces A^*A y AA^* son hermíticas y semidefinidas positivas.*
3. *AA^* y A^*A son definidas positivas si y solo si A es regular.*

Demostración:

1. Si A es singular, $\exists v \neq \theta : \ Av = \theta$. Entonces, para ese vector $v^*Av = 0$, en contradicción con la hipótesis.

2. Es trivial que AA^* y A^*A son hermíticas. Además

$$\forall v \in \mathbb{C}^n \setminus \{\theta\}, \quad \begin{cases} v^*(AA^*)v = (A^*v)^*(A^*v) = \|A^*v\|_2^2 \geq 0 \\ v^*(A^*A)v = (Av)^*(Av) = \|Av\|_2^2 \geq 0 \end{cases}$$

3. De la expresión anterior

   $\forall v \in \mathbb{C}^n \setminus \{\theta\},\ v^*(AA^*)v = \|A^*v\|_2^2 > 0 \Longleftrightarrow Av \neq \theta,\ \forall v \in \mathbb{C}^n \setminus \{\theta\} \Longleftrightarrow A$ es regular

   Análogamente el otro.

◇

Teorema 1.6 *(Caracterización de las matrices definidas positivas). Sea $A \in \mathcal{M}$ hermítica. Entonces,*

1. *A es definida positiva si y solo si $\lambda_i(A) > 0,\ i = 1, ..., n$*
2. *A es semidefinida positiva si y solo si $\lambda_i(A) \geq 0,\ i = 1, ..., n$*

(Hay un resultado análogo para matrices definidas y semidefinidas negativas).

Demostración: Sigue de que

$$\forall v \in \mathbb{C}^n \setminus \{\theta\}, \quad v^*Av = R_A(v)(v^*v)$$

siendo el segundo de los factores siempre positivo y el rango del primero de ellos igual a $[\lambda_1, \lambda_n]$. ◇

Nota:

Si A es definida o semidefinida positivas, entonces $\rho(A) = \lambda_n(A)$. ■

Corolario 1.6 *Se verifica que*

1. *$\lambda_i(A^*A) \geq 0,\ i = 1, ..., n$*
2. *$\lambda_i(A^*A) > 0,\ i = 1, ...n$ si y solo si A es regular.*

Demostración: Inmediato. ◇

1.7. Normas subordinadas

Definición 1.14 *Dada una norma vectorial $\|\cdot\|$ sobre $\mathbb{C}^n$, se llama norma matricial subordinada a la norma vectorial a la aplicación*

$$\|\cdot\| : \mathcal{M}(\mathbb{C}) \to \mathbb{R}_+, \quad \|A\| = \sup_{v\in\mathbb{C}^n\setminus\{\theta\}} \frac{\|Av\|}{\|v\|}$$

Proposición 1.4 *La anterior aplicación es efectivamente una norma matricial. Además, el supremo se alcanza y pueden darse las siguientes definiciones equivalentes*

$$\|A\| = \max_{\substack{v\in\mathbb{C}^n\\ \|v\|\leq 1}} \|Av\| = \max_{\substack{v\in\mathbb{C}^n\\ \|v\|=1}} \|Av\| = \text{ínf}\{M > 0 : \ \|Av\| \leq M\|v\|, \ \forall v \in \mathbb{C}^n\}$$

Demostración: En problemas. ◇**Notas:**

1. Para toda norma matricial subordinada, $\|I\| = 1$.
2. Existen, por tanto, normas matriciales que no son subordinadas a ninguna norma vectorial. Por ejemplo, $\|\cdot\|_{ES}$ no es subordinada porque $\|I\|_{ES} = \sqrt{n} \neq 1$, si $n \geq 2$.
3. Es claro que $\|Av\| \leq \|A\| \cdot \|v\|, \quad \forall v \in \mathbb{C}^n$. Por tanto, la norma subordinada es consistente con la norma matricial dada.

■

Teorema 1.7 *Sea $A \in \mathcal{M}_n(\mathbb{C})$.*

1. *La norma matricial subordinada a la norma vectorial $\|\cdot\|_1$ se llama norma columna y viene dada por*
$$\|A\|_C = \sup_{v\in\mathbb{C}^n\setminus\{\theta\}} \frac{\|Av\|_1}{\|v\|_1} = \max_j \sum_{i=1}^n |a_{ij}|$$
2. *La norma matricial subordinada a la norma vectorial $\|\cdot\|_\infty$ se llama norma fila y viene dada por*
$$\|A\|_F = \sup_{v\in\mathbb{C}^n\setminus\{\theta\}} \frac{\|Av\|_\infty}{\|v\|_\infty} = \max_i \sum_{j=1}^n |a_{ij}|$$
3. *La norma matricial subordinada a la norma vectorial $\|\cdot\|_2$ se llama norma espectral y viene dada por*
$$\|A\|_S = \sup_{v\in\mathbb{C}^n\setminus\{\theta\}} \frac{\|Av\|_2}{\|v\|_2} = \sqrt{\rho(A^*A)}$$

Demostración: Los apartados 1) y 2) se demuestran en problemas.

Ya que A^*A es hermítica y semidefinida positiva, tiene sus autovalores ≥ 0, de modo que pueden ordenarse en la forma $0 \leq \lambda_1(A^*A) \leq \cdots \leq \lambda_n(A^*A)$. Entonces

$$\|A\|_S^2 = \sup_{v\in\mathbb{C}^n\setminus\{\theta\}} \frac{\|Av\|_2^2}{\|v\|_2^2} = \sup_{v\in\mathbb{C}^n\setminus\{\theta\}} \frac{v^*A^*Av}{v^*v} = \sup_{v\in\mathbb{C}^n\setminus\{\theta\}} R_{A^*A}(v) = \lambda_n(A^*A) = \rho(A^*A)$$

por el Teorema de Courant-Fisher. ◇

Proposición 1.5 *Si A es normal, entonces $\|A\|_S = \rho(A)$.*

Demostración: Por el Corolario 1.2, existe U unitaria tal que

$$\begin{cases} U^*AU = \text{diag } (\lambda_i(A)) \\ U^*A^*U = \text{diag } (\overline{\lambda_i(A)}) \end{cases} \Rightarrow U^*A^*AU = \text{diag } (|\lambda_i(A)|^2) \Rightarrow$$

$$\lambda_i(A^*A) = |\lambda_i(A)|^2, \quad i = 1, ..., n$$

Por tanto, $\rho(A^*A) = \rho(A)^2$. ◇

Proposición 1.6 *La norma espectral es invariante por transformaciones unitarias, es decir, dada $A \in \mathcal{M}(\mathbb{C})$,*

$$\|A\|_S = \|AU\|_S = \|UA\|_S = \|U^*AU\|_S, \quad \forall U, \text{ unitaria}$$

Demostración: Basta probar que

$$\rho(A^*A) = \rho(U^*A^*AU) = \rho(A^*U^*UA)$$

Es evidente la igualdad entre el primer y tercer término. La primera igualdad sigue de que el espectro de una matriz es invariante por semejanzas. ◇

Veamos ahora que se puede aproximar superiormente el radio espectral de una matriz dada mediante normas matriciales de la matriz convenientemente elegidas. Para ello es necesario previamente el siguiente

Lema 1.2 *Denotemos $\|\cdot\|$ una norma vectorial en $\mathbb{C}^n$ y la norma matricial subordinada en $\mathcal{M}_n(\mathbb{C})$ y sea $H \in \mathcal{M}_n(\mathbb{C})$ una matriz regular. Consideremos la aplicación*

$$\|\cdot\|_H : \mathbb{C}^n \to \mathbb{R}_+, \quad \|v\|_H = \|H^{-1}v\|$$

Entonces

1. *$\|\cdot\|_H$ es una norma vectorial en $\mathbb{C}^n$.*

2. La norma matricial subordinada a ella viene dada por

$$\|\cdot\|_H : \mathcal{M}_n(\mathbb{C}) \to \mathbb{R}_+, \quad \|B\|_H = \sup_{v\neq\theta} \frac{\|Bv\|_H}{\|v\|_H} = \|H^{-1}BH\|$$

Demostración: En problemas ◇

El anterior resultado se lee como sigue en un caso particular: sea H una matriz regular; la aplicación

$$\|\cdot\|_H : \ \mathbb{C}^n \longrightarrow \mathbb{R}_+, \quad \|v\|_H = \|H^{-1}v\|_\infty$$

es una norma vectorial y su norma matricial subordinada es la aplicación

$$\|\cdot\|_H : \ \mathcal{M}_n(\mathbb{C}) \longrightarrow \mathbb{R}_+, \quad \|B\|_H = \|H^{-1}BH\|_F$$

Teorema 1.8 *Dada $A \in \mathcal{M}(\mathbb{C})$, y $\varepsilon > 0$, existe una norma matricial subordinada, $\|\cdot\|$, tal que $\|A\| \leq \rho(A) + \varepsilon$.*

Demostración: Supongamos dadas $A \in \mathcal{M}(\mathbb{C})$, y $\varepsilon > 0$. Por el Teorema de Schur, existe U unitaria tal que $U^{-1}AU = T$, siendo

$$T = \begin{pmatrix} \lambda_1 & t_{12} & \dots & t_{1n} \\ 0 & \lambda_2 & \dots & t_{2n} \\ \cdot & \cdot & \dots & \cdot \\ 0 & 0 & \dots & \lambda_n \end{pmatrix} \quad \text{siendo } \lambda_i = \lambda_i(A)$$

Se introduce la matriz $D_\delta = \text{diag}\ (1, \delta, \delta^2, ..., \delta^{n-1})$ donde $\delta \neq 0$ es un parámetro que se fijará posteriormente. Se tiene que D_δ es regular y se verifica que

$$(UD_\delta)^{-1}A(UD_\delta) = D_\delta^{-1}TD_\delta = \begin{pmatrix} \lambda_1 & \delta t_{12} & \delta^2 t_{13} & \dots & \delta^{n-1}t_{1n} \\ 0 & \lambda_2 & \delta t_{23} & \dots & \delta^{n-2}t_{2n} \\ \cdot & \cdot & \cdot & \dots & \cdot \\ 0 & 0 & 0 & \dots & \delta t_{n-1,n} \\ 0 & 0 & 0 & \dots & \lambda_n \end{pmatrix}$$

Si se aplica el Lema 1.2 a la norma vectorial $\|\cdot\|_\infty$ y a su correspondiente norma matricial subordinada (la norma fila), resulta que la aplicación

$$\|\cdot\| : \mathcal{M}(\mathbb{C}) \to \mathbb{R}_+, \quad \|B\| = \|(UD_\delta)^{-1}B(UD_\delta)\|_F$$

es una norma matricial subordinada a una norma vectorial. En esta norma

$$\|A\| = \max_{1\leq i\leq n}\{|\lambda_i| + |\delta t_{i,i+1}| + \cdots + |\delta^{n-i}t_{in}|\} \leq$$

$$\max_{1\le i\le n} |\lambda_i| + \max_{1\le i\le n-1} \{|\delta t_{i,i+1}| + \cdots + |\delta^{n-i} t_{in}|\} \le \rho(A) + \varepsilon$$

escogiendo $\delta > 0$ para que el segundo sumando sea $\le \varepsilon$. ◇

El siguiente resultado da condiciones necesarias y suficientes para que la sucesión formada por las potencias sucesivas de una matriz converja a la matriz nula. Es un resultado fundamental para la convergencia de los métodos iterativos de resolución de sistemas lineales.

Teorema 1.9 *Sea $B \in \mathcal{M}$. Son equivalentes las siguientes afirmaciones*

1. $\lim_{k\to+\infty} B^k = \theta$
2. $\lim_{k\to+\infty} B^k v = \theta, \quad \forall v \in \mathbb{K}^n$
3. $\rho(B) < 1$
4. *existe una norma matricial subordinada tal que* $\|B\| < 1$

Demostración:

1) ⇒ 2) Sea $\|\cdot\|$ una norma matricial consistente con la norma vectorial dada en $\mathbb{K}^n$. Entonces

$$\forall v \in \mathbb{K}^n, \quad \|B^k v\| \le \|B^k\| \, \|v\| \to 0$$

por la hipótesis 1).

2) ⇒ 3) Supongamos que $\rho(B) \ge 1$. Entonces, existen $\lambda \in \mathrm{sp}(B)$, $|\lambda| \ge 1$ y $v \ne \theta$ tales que $Bv = \lambda v$. Pero

$$B^2 v = \lambda B v = \lambda^2 v \Rightarrow \ldots \Rightarrow B^k v = \lambda^k v$$

que no convergería a θ contra la hipótesis por ser $|\lambda| \ge 1$.

3) ⇒ 4) Según el Teorema 1.8, para cada $\varepsilon > 0$ existe una norma matricial subordinada tal que $\|B\| \le \rho(B) + \varepsilon$. Entonces, basta elegir ε tal que $\rho(B) + \varepsilon < 1$.

4) ⇒ 1) Se vió en la Proposición 1.1.

◇

Teorema 1.10 *(Lema de Neumann) Sea $B \in \mathcal{M}(\mathbb{K})$ y supongamos que $\rho(B) < 1$. Entonces $I - B$ es regular y $(I - B)^{-1} = \sum_{n=0}^{\infty} B^n$ convergiendo la serie. Recíprocamente, si esta serie converge, entonces $\rho(B) < 1$.*

Demostración: En efecto, si $\rho(B) < 1$, por el Teorema 1.3 es conocido que $I - B$ es invertible, y por el Teorema 1.9 se sabe también que $\lim_{n\to\infty} B^n = \theta$. Por otra parte, es fácil de comprobar que $I - B^{k+1} = (I - B)\sum_{n=0}^{k} B^n$; tomando límites con $k \to \infty$, resulta

$$I = (I - B)\sum_{n=0}^{\infty} B^n \Rightarrow (I - B)^{-1} = \sum_{n=0}^{\infty} B^n$$

Recíprocamente, si la serie converge, necesariamente $\lim_{n\to\infty} B^n = \theta$ y por el Teorema 1.9, $\rho(B) < 1$. ◇

Por último indicamos un resultado útil para el estudio de la convergencia de los métodos iterativos de resolución de sistemas lineales.

Teorema 1.11 *Sea $B \in \mathcal{M}$ y $\|\cdot\|$ una norma matricial cualquiera. Entonces*

$$\lim_{k\to+\infty} \|B^k\|^{1/k} = \rho(B)$$

Demostración: Por el Corolario 1.3 se tiene que $\rho(B) = \rho(B^k)^{1/k}$. Y por la Proposición 1.1, $\rho(B^k) \leq \|B^k\|$. De modo que $\rho(B) \leq \|B^k\|^{1/k}$. En consecuencia, para probar la tesis bastará justificar que

$$\text{para cualquier } \varepsilon > 0 \text{ fijo, } \exists k_0 : \ \forall k > k_0 \text{ se tiene } \|B^k\|^{1/k} < \rho(B) + \varepsilon$$

o equivalentemente, que para $k > k_0$ se tiene que

$$\frac{\|B^k\|}{(\rho(B) + \varepsilon)^k} < 1$$

En efecto, dado $\varepsilon > 0$, consideremos la matriz $B_\varepsilon = \dfrac{1}{\rho(B) + \varepsilon} B$ que está bien definida (aunque pudiera ser $\rho(B) = 0$). Ya que $\lambda_i(B_\varepsilon) = \dfrac{1}{\rho(B) + \varepsilon}\lambda_i(B)$, será $\rho(B_\varepsilon) < 1$ y por el Teorema 1.9,

$$\lim_{k\to+\infty} B_\varepsilon^k = \lim_{k\to+\infty} \frac{1}{(\rho(B) + \varepsilon)^k} B^k = \theta$$

De modo que

$$\exists k_0 : \ \forall k > k_0, \ \frac{\|B^k\|}{(\rho(B) + \varepsilon)^k} < 1$$

cqd. ◇

Tema 2

Métodos Iterativos de resolución de Sistemas Lineales

2.1. Introducción

Los métodos directos de resolución de los sistemas lineales se ejecutan a través de un número finito de pasos y generarían una solución exacta si no fuera por los errores de redondeo. Por el contrario, un método indirecto da lugar a una sucesión de vectores que idealmente converge a la solución. El cálculo se detiene cuando se encuentra una solución aproximada con cierto grado de precisión fijado de antemano.

Los métodos indirectos suelen ser iterativos, es decir, para obtener la sucesión de aproximaciones de la solución se utiliza repetidamente un proceso sencillo. Los métodos iterativos son apropiados para sistemas lineales grandes y con frecuencia muy eficientes en el caso de matrices huecas. Esta suele ser la situación en la resolución numérica de las ecuaciones en derivadas parciales.

Comenzamos dando un ejemplo para entender los procedimientos que veremos a continuación.

Ejemplo: Sea el sistema

$$\begin{pmatrix} 7 & -6 \\ -8 & 9 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 3 \\ -4 \end{pmatrix}$$

cuya solución es $x_1 = \frac{1}{5} = 0{,}2$ y $x_2 = -\frac{4}{15} = -0{,}266$. Inicialmente se eligen x_1^0 y x_2^0 como valores iniciales. La k-ésima iteración podría venir dada por

$$\begin{cases} x_1^k = \frac{1}{7}(6x_2^{k-1} + 3) \\ x_2^k = \frac{1}{9}(8x_1^{k-1} - 4) \end{cases}$$

Este procedimiento se conoce como el método de Jacobi. Algunos valores que se obtienen son

k	x_1^k	x_2^k
0	0,000000	0,000000
10	0,148651	−0,198201
20	0,186516	−0,249088
30	0,196615	−0,262154
40	0,199131	−0,265508
50	0,199777	−0,266369

Podemos modificar el método de modo que se considere en cada iteración el valor más reciente de x_1^k para la segunda ecuación. Este método, que se llama de Gauss-Seidel, se escribiría así

$$\begin{cases} x_1^k = \dfrac{1}{7}(6x_2^{k-1} + 3) \\ x_2^k = \dfrac{1}{9}(8x_1^k - 4) \end{cases}$$

Algunos valores obtenidos por este procedimiento son

k	x_1^k	x_2^k
0	0,000000	0,000000
10	0,219773	−0,249088
20	0,201304	−0,265308
30	0,200086	−0,266590
40	0,200006	−0,266662
50	0,200000	−0,266666

Observamos que ambos métodos convergen al mismo límite, pero que el segundo lo hace más rápidamente. En contraste con los métodos directos, la precisión que se obtiene en la solución depende del momento en que se detenga el proceso.

En general, dado un sistema lineal $Au = b$, utilizar un método iterativo consiste en ir obteniendo términos de una sucesión $\{u_k\}$ que sean solución de

$$\begin{cases} u_0 \in \mathbb{K}^n \quad \text{arbitrario} \\ u_{k+1} = Bu_k + c, \quad k \geq 0 \end{cases}$$

para cierta matriz B y cierto vector c. Hemos de estudiar si el método converge, es decir, si $\lim_{k\to\infty} u_k = u$, solución del sistema, y su velocidad de convergencia.

2.2. Generalidades sobre la convergencia de los Métodos Iterativos

Consideremos el sistema lineal

(SL) $$Au = b, \quad A \text{ invertible}$$

Supongamos que somos capaces de encontrar una matriz B y un vector c tales que $I - B$ sea invertible y tal que la única solución del sistema lineal

$$u = Bu + c$$

sea la de (SL). Entonces se puede definir el método iterativo

$$\begin{cases} u_0 \in \mathbb{K}^n \quad \text{arbitrario} \\ u_{k+1} = Bu_k + c, \quad k \geq 0 \end{cases}$$

a) Convergencia del método:

Si denotamos $e_k = u_k - u$, el método converge (globalmente) si y solo si $\lim_{k\to\infty} e_k = 0$ (para cada $u_0 \in K^n$).

Teorema 2.1 *Las siguientes proposiciones son equivalentes:*

a) El método iterativo es convergente.

b) $\rho(B) < 1$

c) $\|B\| < 1$*, para alguna norma matricial subordinada.*

Demostración: Se tiene

$$e_k = u_k - u = Bu_{k-1} + c - (Bu + c) = B(u_{k-1} - u) = Be_{k-1}$$

y, por tanto

$$e_k = Be_{k-1} = B^2 e_{k-2} = ... = B^k e_0$$

El método iterativo será directo si $\exists K \in \mathbb{N}$ tal que $B^K = \theta$. Será convergente si

$$\lim_{k\to\infty} B^k v = \theta, \quad \forall v \in \mathbb{K}^n$$

El Teorema sigue ahora del Teorema 1.9 . ◇

b) Velocidad de convergencia:

Entre todos los métodos iterativos aplicables a (SL), nos preguntamos cuál es el que tiene mayor velocidad de convergencia. Veremos dos casos

i) Supongamos B normal y $\|\cdot\|_2$.
En estas condiciones

$$\|e_k\|_2 = \|B^k e_0\|_2 \leq \|B^k\|_s \cdot \|e_0\|_2 = \rho(B^k)\|e_0\|_2 = \rho(B)^k \|e_0\|_2$$

La primera desigualdad es óptima por la definición de la norma espectral. La siguiente igualdad sigue de ser B normal (Proposición 1.5). La última igualdad, del Corolario 1.3.

Por tanto, en el caso de matrices normales, el método es más rápido en el sentido de la norma espectral cuanto más pequeño sea $\rho(B)$.

ii) Caso general: B cualquiera y cualquier norma vectorial.
Veremos que la conclusión es la misma en el sentido que, asintóticamente, $\|e_k\|$ se comporta en el peor de los casos como $\rho(B)^k$.

Teorema 2.2 *Sea $\|\cdot\|$ una norma vectorial cualquiera. Sea u la solución del sistema $u = Bu + c$. Se considera el método iterativo*

$$\begin{cases} u_0 \in \mathbb{K}^n \quad \text{arbitrario} \\ u_{k+1} = Bu_k + c, \quad k \geq 0 \end{cases}$$

Entonces

i) $\lim_{k\to\infty} \{ \sup_{\|u_0 - u\|=1} \|u_k - u\|^{1/k} \} = \rho(B)$

ii) Si el método es convergente, se verifica la siguiente estimación

$$\|u_k - u\| \leq \frac{\|B\|^k}{1 - \|B\|} \|u_1 - u_0\|$$

para la norma matricial subordinada tal que $\|B\| < 1$ y la norma vectorial de la que aquélla es subordinada.

Demostración: i) Sea $\|\cdot\|$ la norma matricial subordinada. Entonces

$$\sup_{\|u_0 - u\|=1} \|u_k - u\| = \sup_{\|e_0\|=1} \|e_k\| = \sup_{\|e_0\|=1} \|B^k e_0\| = \max_{\|e_0\|=1} \|B^k e_0\| = \|B^k\|$$

Por tanto

$$\sup_{\|e_0\|=1} \|e_k\|^{1/k} = \|B^k\|^{1/k} \to \rho(B)$$

según Teorema 1.11.
ii) Sabemos que existe una norma matricial subordinada a una norma vectorial para la que $\|B\| < 1$. Tomando ambas

$$u_k - u_{k-1} = B(u_{k-1} - u_{k-2}) = ... = B^{k-1}(u_1 - u_0)$$

Por tanto

$$\|u_k - u_{k-1}\| \leq \|B^{k-1}\| \cdot \|u_1 - u_0\| \leq \|B\|^{k-1} \cdot \|u_1 - u_0\|$$

Entonces, si $m > k$, se tendrá

$$\|u_m - u_k\| \leq \|u_m - u_{m-1}\| + ... + \|u_{k+1} - u_k\| \leq (\|B\|^{m-1} + ... + \|B\|^k)\|u_1 - u_0\|$$

De modo que para todo $m > k$ se tiene

$$\|u_m - u_k\| \leq \frac{\|B\|^k}{1 - \|B\|}\|u_1 - u_0\|$$

y tomando límites con $m \to \infty$ sigue el resultado.

•

A consecuencia del apartado i) del Teorema anterior sigue

dado $\varepsilon > 0$, $\exists l : \ \forall k \geq l$, se verifica $\|e_k\| \leq (\rho(B) + \varepsilon)^k$, $\forall e_0$ tal que $\|e_0\| = 1$

El método aumenta su velocidad de convergencia cuanto menor sea $\rho(B)$.

2.3. Métodos de Jacobi, Gauss-Seidel y relajación por puntos.

Estos métodos son casos particulares del método iterativo siguiente. Sea

$$(SL) \qquad Au = b, \quad A \text{ regular}$$

Supongamos que podemos escribir $A = M - N$ siendo M "fácil de invertir", en el sentido de que el sistema lineal de matriz M sea fácil de resolver. En la práctica, M va a ser casi diagonal o triangular. Entonces

$$Au = b \Longleftrightarrow Mu = Nu + b \Longleftrightarrow u = (M^{-1}N)u + M^{-1}b$$

siendo, por tanto,

$$B = M^{-1}N, \quad c = M^{-1}b \quad \text{y} \quad I - B = I - M^{-1}N = M^{-1}A, \text{ regular}$$

Se asocia el método iterativo siguiente

$$\begin{cases} u_0, \quad \text{dado} \\ u_{k+1} = (M^{-1}N)u_k + M^{-1}b, \quad k \geq 0 \end{cases}$$

que será convergente si y solo si $\rho(M^{-1}N) < 1$. En la práctica, resolveremos los sistemas lineales sucesivos en la forma

$$Mu_{k+1} = Nu_k + b, \quad k \geq 0$$

Para resolver el método iterativo, hay que invertir la parte M de la matriz A. Intuitivamente parece que cuanto más se parezca M a A, mejor será el método, pero más difícil será de calcular. En el caso límite, $M = A,\ N = \theta$ y la primera iteración da la solución exacta $u_1 = A^{-1}b$.

En el caso de los métodos de Jacobi y Gauss-Seidel la descomposición $A = M - N$ es disjunta en el sentido de que $m_{ij} = a_{ij}$ o $m_{ij} = 0$. En el método de relajación, la descomposición es no disjunta.

Supondremos en lo que sigue la hipótesis

$$(H) \qquad a_{ii} \neq 0, \quad i = 1, ..., n.$$

Haremos la siguiente descomposición por puntos de A

$$A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ . & . & \dots & . \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{pmatrix} = D - E - F$$

siendo

$$D = \begin{pmatrix} a_{11} & 0 & \dots & 0 \\ 0 & a_{22} & \dots & 0 \\ . & . & \dots & . \\ 0 & 0 & \dots & a_{nn} \end{pmatrix}$$

$$E = \begin{pmatrix} 0 & 0 & \dots & 0 \\ -a_{21} & 0 & \dots & 0 \\ . & . & \dots & . \\ -a_{n1} & -a_{n2} & \dots & 0 \end{pmatrix}, \quad F = \begin{pmatrix} 0 & -a_{12} & \dots & -a_{1n} \\ 0 & 0 & \dots & -a_{2n} \\ . & . & \dots & . \\ 0 & 0 & \dots & 0 \end{pmatrix}$$

A) Metodo de Jacobi por puntos

Se define tomando $M = D, \quad N = E + F$.

El método es

$$\begin{cases} u_0, \quad \text{arbitrario} \\ u_{k+1} = D^{-1}(E+F)u_k + D^{-1}b, \quad \forall k \geq 0 \end{cases}$$

Se llamará matriz de Jacobi a

$$J = D^{-1}(E+F) = I - D^{-1}A$$

El método convergerá si y solo si $\rho(J) < 1$.

El cálculo efectivo se lleva a cabo del modo siguiente

$$\begin{pmatrix} u_1^{k+1} \\ u_2^{k+1} \\ . \\ u_n^{k+1} \end{pmatrix} = - \begin{pmatrix} 1/a_{11} & 0 & \dots & 0 \\ 0 & 1/a_{22} & \dots & 0 \\ . & . & \dots & . \\ 0 & . & \dots & 1/a_{nn} \end{pmatrix} \cdot$$

$$\cdot \left[\begin{pmatrix} 0 & a_{12} & \dots & a_{1n} \\ a_{21} & 0 & \dots & a_{2n} \\ . & . & \dots & . \\ a_{n1} & a_{n2} & \dots & 0 \end{pmatrix} \begin{pmatrix} u_1^k \\ u_2^k \\ . \\ u_n^k \end{pmatrix} - \begin{pmatrix} b_1 \\ b_2 \\ . \\ b_n \end{pmatrix} \right]$$

Así pues,

$$u_i^{k+1} = \frac{1}{a_{ii}}[b_i - (a_{i1}u_1^k + \ldots + a_{i,i-1}u_{i-1}^k + a_{i,i+1}u_{i+1}^k + \ldots + a_{in}u_n^k)], \ 1 \leq i \leq n$$

Observaciones:

1) Para calcular u_i^{k+1} se utilizan $n-1$ componentes del vector $u_k = (u_i^k)$. Por tanto, u_k ha de guardarse en la memoria durante el cálculo de u_{k+1}. Se usan $2n$ registros de memoria en cada iteración, n para u_k y n para u_{k+1}.
2) Los pasos a seguir para el cálculo de u_i^{k+1} son los siguientes:
   1. $s = \sum_{j=1}^{n} a_{ij}u_j^k$
   2. $s = s - a_{ii}u_i^k - b_i, \quad i = 1, ..., n$
   3. $u_i^{k+1} = -\frac{s}{a_{ii}}$

■

Parece razonable pensar que el método se mejorará si se va "actualizando"el cálculo de u^{k+1} con las componentes de este vector ya obtenidas. Es decir, para obtener u_i^{k+1} se pueden utilizar las $u_j^{k+1},\ j<i$ ya calculadas. Así se usarán además solo n lugares de memoria puesto que los u_i^{k+1} van reemplazando a los valores de u_i^k. Este método se conoce con el nombre de

B) Metodo de Gauss − Seidel por puntos

Se define tomando $M = D - E, \quad N = F$. La matriz $D - E$ es invertible por la hipótesis (H).

El método es

$$\begin{cases} u_0, \quad \text{arbitrario} \\ u_{k+1} = (D-E)^{-1}Fu_k + (D-E)^{-1}b, \quad \forall k \geq 0 \end{cases}$$

Se llamará matriz de Gauss-Seidel a

$$\mathcal{L}_1 = (D-E)^{-1}F$$

El método convergerá si y solo si $\rho(\mathcal{L}_1) < 1$.

El cálculo efectivo se lleva a cabo del modo siguiente.

$$(D-E)u_{k+1} = Fu_k + b \Longrightarrow Du_{k+1} = Eu_{k+1} + Fu_k + b \Longrightarrow$$

$$u_{k+1} = D^{-1}(Eu_{k+1} + Fu_k + b)$$

$$\begin{pmatrix} u_1^{k+1} \\ u_2^{k+1} \\ . \\ u_n^{k+1} \end{pmatrix} = - \begin{pmatrix} 1/a_{11} & 0 & \dots & 0 \\ 0 & 1/a_{22} & \dots & 0 \\ . & . & \dots & . \\ 0 & . & \dots & 1/a_{nn} \end{pmatrix} \cdot$$

$$\cdot \left[\begin{pmatrix} 0 & 0 & \dots & 0 \\ a_{21} & 0 & \dots & 0 \\ . & . & \dots & . \\ a_{n1} & a_{n2} & \dots & 0 \end{pmatrix} \begin{pmatrix} u_1^{k+1} \\ u_2^{k+1} \\ . \\ u_n^{k+1} \end{pmatrix} + \begin{pmatrix} 0 & a_{12} & \dots & a_{1n} \\ 0 & 0 & \dots & a_{2n} \\ . & . & \dots & . \\ 0 & 0 & \dots & 0 \end{pmatrix} \begin{pmatrix} u_1^k \\ u_2^k \\ . \\ u_n^k \end{pmatrix} - \begin{pmatrix} b_1 \\ b_2 \\ . \\ b_n \end{pmatrix} \right] .$$

Así pues,

$$u_i^{k+1} = \frac{1}{a_{ii}}[b_i - \sum_{j<i} a_{ij}u_j^{k+1} - \sum_{j>i} a_{ij}u_j^k],\ 1 \leq i \leq n.$$

Observación:

En este método, además de necesitar menos memoria, se "invierte"más parte de la matriz A que en el de Jacobi, por lo que es razonable pensar que será más rápido. Pero hay ejemplos en que el método de Jacobi converge y el de Gauss-Seidel no. ■

C) Metodo de relajacion por puntos

Consideremos la siguiente descomposición de D

$$D = \frac{1}{\omega}D + \left(1 - \frac{1}{\omega}\right) D, \quad \omega \in \mathbb{R} \setminus \{0\}$$

De esta forma

$$A = \frac{1}{\omega}D - E - \left(\frac{1-\omega}{\omega}\right) D - F$$

y se pueden tomar

$$M = \frac{1}{\omega}D - E, \quad N = \left(\frac{1-\omega}{\omega}\right) D + F$$

de modo que se ha pasado parte de la diagonal D a la matriz N. La matriz M es invertible por la hipótesis (H).

El método iterativo obtenido es

$$\begin{cases} u_0, \quad \text{arbitrario} \\ u_{k+1} = \left(\frac{1}{\omega}D - E\right)^{-1} \left[\left(\frac{1-\omega}{\omega}D + F\right) u_k + b\right], \quad \forall k \geq 0 \end{cases}$$

Se llamará matriz de relajación a

$$\mathcal{L}_\omega = \left(\frac{1}{\omega}D - E\right)^{-1} \left(\frac{1-\omega}{\omega}D + F\right) = (D - \omega E)^{-1}[(1-\omega)D + \omega F]$$

El método convergerá si y solo si $\rho(\mathcal{L}_\omega) < 1$.

El cálculo efectivo que se lleva a cabo es el siguiente:

$$\left(\frac{1}{\omega}D - E\right) u_{k+1} = \left(\frac{1-\omega}{\omega}D + F\right) u_k + b$$

$$(D - \omega E)u_{k+1} = ((1-\omega)D + \omega F)u_k + \omega b$$

$$Du_{k+1} = Du_k + \omega[Eu_{k+1} - (D-F)u_k + b]$$

$$u_{k+1} = u_k + \omega D^{-1}(Eu_{k+1} - (D-F)u_k + b)$$

$$\begin{pmatrix} u_1^{k+1} \\ u_2^{k+1} \\ . \\ u_n^{k+1} \end{pmatrix} = \begin{pmatrix} u_1^{k} \\ u_2^{k} \\ . \\ u_n^{k} \end{pmatrix} - w \begin{pmatrix} 1/a_{11} & 0 & \dots & 0 \\ 0 & 1/a_{22} & \dots & 0 \\ . & . & \dots & . \\ 0 & . & \dots & 1/a_{nn} \end{pmatrix} \cdot$$

$$\cdot \left[\begin{pmatrix} 0 & 0 & \dots & 0 \\ a_{21} & 0 & \dots & 0 \\ . & . & \dots & . \\ a_{n1} & a_{n2} & \dots & 0 \end{pmatrix} \begin{pmatrix} u_1^{k+1} \\ u_2^{k+1} \\ . \\ u_n^{k+1} \end{pmatrix} + \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ 0 & a_{22} & \dots & a_{2n} \\ . & . & \dots & . \\ 0 & 0 & \dots & a_{nn} \end{pmatrix} \begin{pmatrix} u_1^{k} \\ u_2^{k} \\ . \\ u_n^{k} \end{pmatrix} - \begin{pmatrix} b_1 \\ b_2 \\ . \\ b_n \end{pmatrix} \right].$$

Así pues,

$$u_1^{k+1} = u_1^k - \frac{\omega}{a_{11}}[a_{11}u_1^k + a_{12}u_2^k + \ldots + a_{1n}u_n^k - b_1]$$

y una vez hallados u_j^{k+1} para $j < i$, se determina

$$u_i^{k+1} = u_i^k - \frac{\omega}{a_{ii}}[a_{i1}u_i^{k+1} + \ldots + a_{i,i-1}u_{i-1}^{k+1} + a_{ii}u_i^k + a_{i,i+1}u_{i+1}^k + \ldots + a_{in}u_n^k - b_i]$$

Observaciones:

1) El método de relajación para $\omega = 1$ coincide con el de Gauss-Seidel. De ahí la notación usada para la matriz de Gauss-Seidel.
2) Aunque en principio el parámetro ω podría ser un número real no nulo, se probará (Teorema 3.5) que para que el método converja es necesario que $\omega \in (0, 2)$. El método se llamará de sobrerrelajación si $\omega \in (1, 2)$ y de subrrelajación si $\omega \in (0, 1)$.
3) Se verá que $\rho(\mathcal{L}_\omega)$ es una función continua de ω. Entonces, el estudio del método consiste en
   a) Determinar un intervalo $I \subset \mathbb{R} \setminus \{0\}$, tal que $\forall \omega \in I,\ \rho(\mathcal{L}_\omega) < 1$
   b) Determinar $\omega_0 \in I$ tal que
   $$\rho(\mathcal{L}_{\omega_0}) \approx \inf_{\omega \in I} \rho(\mathcal{L}_\omega)$$
4) Para ciertos valores del parámetro de relajación se obtiene una convergencia más rápida que para $\omega = 1$ y por tanto un tiempo de cálculo menor que para el método de Gauss-Seidel. El número de operaciones es similar en ambos métodos. No obstante, hay que tener en cuenta el tiempo utilizado en la estimación preliminar del parámetro ω_0 para comparar la eficacia de ambos métodos.

■

2.4. Métodos Iterativos por bloques

Supongamos la matriz A descompuesta por bloques de forma que los bloques diagonales sean cuadrados y escribamos

$$\begin{cases} A = D_B - E_B - F_B \\ D_B \quad \text{formada por los bloques diagonales} \\ -E_B \quad \text{formada por los bloques subdiagonales} \\ -F_B \quad \text{formada por los bloques superdiagonales} \end{cases}$$

Se recuerda el siguiente resultado

Proposición 2.1 *Si A es una matriz triangular por bloques, entonces se verifica que* $det\ (A) = \prod_{i=1}^{N} det\ (A_{ii})$. *En particular, si A es diagonal por bloques se tiene que A es invertible si y solo si A_{ii} es invertible para cada i.*

Demostración: En problemas.

Se establece la hipótesis

(H_B) Las matrices A_{ii} son invertibles para cada i

La Proposición anterior asegura que D_B, $D_B - E_B$ y $D_B - \omega E_B$ son invertibles. Entonces se pueden definir los métodos de Jacobi, Gauss-Seidel y relajación por bloques de modo análogo al descrito por puntos:

a) Método de Jacobi por bloques

$$u_{k+1} = D_B^{-1}(E_B + F_B)u_k + D_B^{-1}b, \quad J_B = D_B^{-1}(E_B + F_B)$$

b) Método de Gauss-Seidel por bloques

$$u_{k+1} = (D_B - E_B)^{-1}F_B u_k + (D_B - E_B)^{-1}b, \quad \mathcal{L}_{B,1} = (D_B - E_B)^{-1}F_B$$

c) Método de relajación por bloques

$$u_{k+1} = \left(\frac{1}{\omega}D_B - E_B\right)^{-1}\left(\frac{1-\omega}{\omega}D_B + F_B\right)u_k + \left(\frac{1}{\omega}D_B - E_B\right)^{-1}b$$

$$\mathcal{L}_{B,\omega} = \left(\frac{1}{\omega}D_B - E_B\right)^{-1}\left(\frac{1-\omega}{\omega}D_B + F_B\right)$$

Observación:

Parece que los métodos por bloques deben converger más rápidamente que los métodos por puntos porque invierten más parte de la matriz A. No obstante, en cada iteración es necesario resolver N sistemas lineales cuyas matrices son A_{ii}. Por tanto, se utilizarán métodos por bloques si la aceleración de la convergencia compensa el tiempo de resolución de los sistemas lineales en cada iteración. ■

2.5. Resultados de convergencia para Métodos Iterativos

Para fijar ideas, consideremos $\mathbb{K} = \mathbb{C}$ (es análogo si $\mathbb{K} = \mathbb{R}$).

Supongamos que los métodos iterativos están bien planteados. En el caso de los tres métodos descritos en las preguntas anteriores significa que se verifican las hipótesis (H) o (H_B) según sean por puntos o por bloques.

Supondremos que A es una matriz hermítica y definida positiva. En tal caso, es conocido el siguiente resultado:

Lema 2.1 *Sea la matriz A definida positiva. Entonces*

a) $a_{ii} > 0$ para $i = 1, ...n$.

b) En cualquier descomposición por bloques de A que tenga los bloques diagonales cuadrados, éstos son también matrices definidas positivas.

Por tanto para una matriz definida positiva, se verifica la hipótesis (H). Por otra parte, como una matriz definida positiva es no singular, se verifica también la hipótesis (H_B).

La primera condición suficiente de convergencia de carácter general es

Teorema 2.3 *(Householder). Sea A una matriz hermítica y definida positiva y sea $A = M - N$, con M regular. Si la matriz $M^* + N$ es definida positiva, entonces $\rho(M^{-1}N) < 1$.*

Demostración: Comenzamos verificando que $M^* + N$ es siempre hermítica; en efecto

$$(M^* + N)^* = M + N^* = A + N + N^* = (A^* + N^*) + N = M^* + N$$

Basta demostrar que $\|M^{-1}N\| < 1$ para alguna norma matricial (Proposición 1.1). Consideraremos la norma matricial subordinada a cierta norma vectorial. En concreto, la aplicación

$$\|\cdot\|_A : \ \mathbb{C}^n \longrightarrow \mathbb{R}_+, \quad \|v\|_A = (v^* A v)^{1/2}$$

es una norma vectorial por ser A definida positiva (Ver problemas). La norma matricial subordinada, verifica

$$\|M^{-1}N\| = \max_{\|v\|_A = 1} \|M^{-1}Nv\|_A = \|M^{-1}Nv_0\|_A$$

para algún $v_0 \in \mathbb{C}^n$ que verifica $\|v_0\|_A = 1$. Pero

$$M^{-1}N = M^{-1}(M - A) = I - M^{-1}A$$

de modo que

$$M^{-1}Nv_0 = v_0 - w_0, \quad \text{siendo} \quad w_0 = M^{-1}Av_0 \neq \theta$$

por ser $M^{-1}A$ regular y $v_0 \neq \theta$. Entonces

$$\|M^{-1}Nv_0\|_A^2 = \|v_0 - w_0\|_A^2 = (v_0^* - w_0^*)A(v_0 - w_0) =$$

$$= v_0^*Av_0 - v_0^*Aw_0 - w_0^*Av_0 + w_0^*Aw_0 = 1 - (v_0^*Aw_0 + w_0^*Av_0 - \|w_0\|_A^2)$$

Escribiendo esta expresión solo en función de w_0, se obtiene

$$v_0 = A^{-1}Mw_0 \Longrightarrow v_0^* = w_0^*M^*(A^{-1})^* = w_0^*M^*A^{-1},$$

la última igualdad, por ser A hermítica. De modo que

$$\left\{ \begin{array}{l} v_0^*Aw_0 = w_0^*M^*A^{-1}Aw_0 = w_0^*M^*w_0 \\ \\ w_0^*Av_0 = w_0^*AA^{-1}Mw_0 = w_0^*Mw_0 \end{array} \right. \Longrightarrow$$

$$\|M^{-1}Nv_0\|_A^2 = 1 - w_0^*(M^* + M - A)w_0 = 1 - w_0^*(M^* + N)w_0 < 1$$

por ser $M^* + N$ definida positiva y $w_0 \neq \theta$. $\diamond$

Aplicamos este Teorema para dar una condición suficiente de convergencia para el método de relajación.

Teorema 2.4 *(Criterio de Ostrowski-Reich). Si A es hermítica y definida positiva, entonces el método de relajación por puntos o por bloques converge si $0 < \omega < 2$. En particular, el método de Gauss-Seidel es convergente.*

Demostración: Como se indicó en el Lema anterior, los métodos de relajación por puntos o por bloques están bien definidos, pues por ser A definida positiva se verifican las hipótesis (H) y (H_B). Se tiene entonces que

$$A = M - N = \left(\frac{1}{\omega}D_B - E_B\right) - \left(\frac{1-\omega}{\omega}D_B + F_B\right) \longrightarrow$$

$$M^* + N = \left(\frac{1}{\omega}D_B^* - E_B^*\right) + \left(\frac{1-\omega}{\omega}D_B + F_B\right) = \frac{1}{\omega}D_B + \frac{1-\omega}{\omega}D_B = \frac{2-\omega}{\omega}D_B$$

porque al ser A hermítica se verifica

$$D_B = D_B^*, \quad E_B = F_B^*, \quad F_B = E_B^*$$

siendo eventualmente los bloques de dimensión 1.

Al aplicar el Teorema de Householder, queda

$$M^* + N \text{ es definida positiva } \iff \frac{2-\omega}{\omega} D_B \text{ es definida positiva} \iff 0 < \omega < 2$$

puesto que D_B es definida positiva. ◇

Observación:

La aplicación del Teorema de Householder al método de Jacobi no da ninguna condición fácilmente explotable. ■

Veremos ahora que independientemente de cualquier hipótesis sobre A, la condición $0 < \omega < 2$ es necesaria para la convergencia del método de relajación.

Teorema 2.5 *(de Kahan). Supuestas las hipótesis (H) o (H_B), se verifica siempre que*

$$\rho(\mathcal{L}_\omega) \geq |\omega - 1|, \quad \omega \neq 0$$

$$\rho(\mathcal{L}_{B,\omega}) \geq |\omega - 1|, \quad \omega \neq 0$$

Por tanto, si el método de relajación converge, entonces $\omega \in (0, 2)$.

Demostración: Haremos el razonamiento para el método por bloques, siendo análogo por puntos. Sabemos que

$$\mathcal{L}_{B,\omega} = \left(\frac{1}{\omega} D_B - E_B\right)^{-1} \left(\frac{1-\omega}{\omega} D_B + F_B\right)$$

Por tanto

$$\det\left(\mathcal{L}_{B,\omega}\right) = \prod_{i=1}^{n} \lambda_i(\mathcal{L}_{B,\omega}) = \frac{\det\left(\dfrac{1-\omega}{\omega} D_B + F_B\right)}{\det\left(\dfrac{1}{\omega} D_B - E_B\right)} = \frac{\left(\dfrac{1-\omega}{\omega}\right)^n \det\left(D_B\right)}{\left(\dfrac{1}{\omega}\right)^n \det\left(D_B\right)} \Longrightarrow$$

$$\det\left(\mathcal{L}_{B,\omega}\right) = (1-\omega)^n$$

En consecuencia,

$$\rho(\mathcal{L}_{B,\omega}) \geq \Big|\prod_{i=1}^{n} \lambda_i(\mathcal{L}_{B,\omega})\Big|^{1/n} = |1-\omega|$$

◇

La existencia de una estructura tridiagonal por bloques o por puntos de A permite comparar de forma más precisa los radios espectrales de la matriz de Jacobi y de la matriz del método de relajación. Comenzamos probando el siguiente:

Lema 2.2 *Sea $\mu \in \mathbb{C} \setminus \{0\}$ y $A(\mu)$ una matriz tridiagonal por bloques de la forma*

$$A(\mu) = \begin{pmatrix} B_1 & \mu^{-1}C_1 & \theta & \theta & \dots & \theta \\ \mu A_2 & B_2 & \mu^{-1}C_2 & \theta & \dots & \theta \\ \theta & \ddots & \ddots & \ddots & \ddots & \vdots \\ \dots & \dots & \dots & \dots & \dots & \dots \\ \theta & \theta & \dots & \mu A_{N-1} & B_{N-1} & \mu^{-1}C_{N-1} \\ \theta & \theta & \theta & \dots & \mu A_N & B_N \end{pmatrix}$$

Entonces, det $(A(\mu))$ = det $(A(1))$

Demostración: Se introduce la matriz diagonal

$$Q(\mu) = \begin{pmatrix} \mu I_1 & \theta & \dots & \theta & \theta \\ \theta & \mu^2 I_2 & \dots & \theta & \theta \\ \dots & \dots & \ddots & \dots & \dots \\ \dots & \dots & \dots & \mu^{N-1} I_{N-1} & \theta \\ \dots & \dots & \dots & \dots & \mu^N I_N \end{pmatrix}$$

donde I_j es la matriz identidad del mismo orden que B_j. Entonces, puede comprobarse que

$$A(\mu) = Q(\mu)A(1)Q(\mu)^{-1}$$

de donde se obtiene el resultado. ◇

Teorema 2.6 *(Comparación de los métodos de Jacobi y Gauss-Seidel). Sea A tridiagonal por bloques. Entonces, los radios espectrales de las matrices de Jacobi y Gauss-Seidel por bloques correspondientes se relacionan de la forma*

$$\rho(\mathcal{L}_{B,1}) = \rho(J_B)^2$$

de manera que los dos métodos convergen o divergen simultáneamente. Cuando convergen, el método de Gauss-Seidel converge más rápidamente que el de Jacobi.

Nota:

En particular, si A es tridiagonal por puntos, se verifica $\rho(\mathcal{L}_1) = \rho(J)^2$. ■

Demostración: Los autovalores de la matriz de Jacobi $J_B = D_B^{-1}(E_B + F_B)$ son los ceros del polinomio característico

$$p_{J_B}(\lambda) = \det\,(\lambda I - D_B^{-1}(E_B + F_B))$$

que son los ceros del polinomio

$$q_{J_B}(\lambda) = \det\,(\lambda D_B - (E_B + F_B)) = \det\,(D_B) p_{J_B}(\lambda)$$

Análogamente, los autovalores de la matriz de Gauss-Seidel $\mathcal{L}_{B,1} = (D_B - E_B)^{-1} F_B$ son los ceros del polinomio característico

$$p_{\mathcal{L}_{B,1}} = \det\,(\lambda I - (D_B - E_B)^{-1} F_B)$$

que son los ceros de

$$q_{\mathcal{L}_{B,1}}(\lambda) = \det\,(\lambda D_B - \lambda E_B - F_B) = \det\,(D_B - E_B) p_{\mathcal{L}_{B,1}}(\lambda)$$

Gracias al Lema 3.2 y por ser A tridiagonal por bloques, se tiene que $\forall \lambda \in \mathbb{C} \setminus \{0\}$,

$$q_{\mathcal{L}_{B,1}}(\lambda^2) = \det\,(\lambda^2 D_B - \lambda^2 E_B - F_B) = \det\,(\lambda^2 D_B - \lambda E_B - \lambda F_B) =$$

$$\lambda^n \det\,(\lambda D_B - E_B - F_B) = \lambda^n q_{J_B}(\lambda)$$

Esta igualdad es válida también para $\lambda = 0$, porque $q_{\mathcal{L}_{B,1}}(0) = 0$. Por tanto,

$$q_{\mathcal{L}_{B,1}}(\lambda^2) = \lambda^n q_{J_B}(\lambda), \quad \forall \lambda \in \mathbb{C}$$

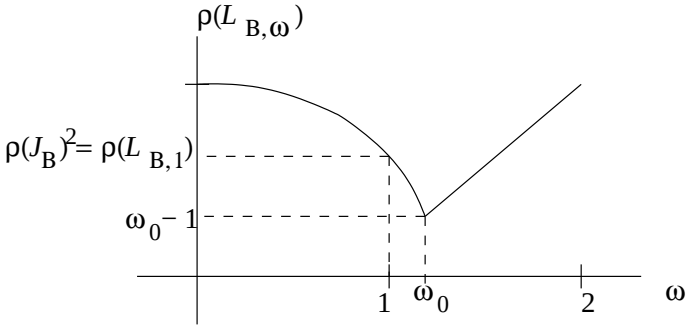
Luego, si

$$\alpha \in \text{sp}\,(\mathcal{L}_{B,1}),\ \alpha \neq 0 \Longrightarrow \{\sqrt{\alpha}, -\sqrt{\alpha}\} \in \text{sp}\,(J_B)$$

$$\beta \in \text{sp}\,(J_B),\ \beta \neq 0 \Longrightarrow \beta^2 \in \text{sp}\,(\mathcal{L}_{B,1}) \quad \text{y} \quad -\beta \in \text{sp}\,(J_B)$$

Existe una biyección entre los autovalores no nulos de $\mathcal{L}_{B,1}$ y pares de autovalores opuestos de J_B, de donde sigue el Teorema. ◇

Teorema 2.7 *(Comparación de los métodos de Jacobi y relajación). Sea A tridiagonal por bloques y supongamos que sp $(J_B) \subset \mathbb{R}$. Entonces, el método de Jacobi por bloques y el método de relajación por bloques para $0 < \omega < 2$ convergen o divergen simultáneamente. Cuando convergen, la función $\omega \in (0,2) \longmapsto \rho(\mathcal{L}_{B,\omega})$ es de la forma*



con $\omega_0 = \dfrac{2}{1+\sqrt{1-\rho(J_B)^2}}$. *De modo que el método es óptimo para* ω_0 *siendo* $\rho(\mathcal{L}_{B,\omega_0}) = \omega_0 - 1$.

Demostración: Es similar a la del Teorema 3.6 pero con detalles más técnicos debido a la mayor complejidad de la matriz de relajación. (cf. Ciarlet [2] pp 107 y ss). ◇

El siguiente teorema da una condición suficiente cómoda para que se verifiquen las hipótesis del Teorema anterior y ocurra una de las dos alternativas posibles.

Teorema 2.8 *Sea* A *hermítica definida positiva y tridiagonal por bloques. Entonces, el método de Jacobi por bloques y el método de relajación por bloques para* $0 < \omega < 2$ *convergen simultáneamente. La función* $\omega \in (0,2) \longmapsto \rho(\mathcal{L}_{B,\omega})$ *es de la forma dada en el Teorema anterior con* $\omega_0 = \dfrac{2}{1+\sqrt{1-\rho(J_B)^2}}$. *Así, si* $\rho(J_B) > 0$, *entonces*

$$\rho(\mathcal{L}_{B,\omega_0}) = \min_{0<\omega<2} \rho(\mathcal{L}_{B,\omega}) = \omega_0 - 1 < \rho(\mathcal{L}_{B,1}) = \rho(J_B)^2;$$

si $\rho(J_B) = 0$, *entonces*

$$\omega_0 = 1 \quad y \quad \rho(\mathcal{L}_{B,1}) = \rho(J_B) = 0$$

Demostración: Comenzamos verificando que sp $(J_B) \in \mathbb{R}$ para aplicar el Teorema 3.7. En efecto, sea $\alpha \in$ sp (J_B); entonces, existe un vector $v \neq \theta$ tal que

$$D_B^{-1}(E_B + F_B)v = \alpha v \Longrightarrow (E_B + F_B)v = \alpha D_B v \Longrightarrow Av = (1-\alpha)D_B v \Longrightarrow$$

$$v^* A v = (1-\alpha) v^* D_B v$$

Ya que A es hermítica definida positiva, sigue que también D_B es hermítica definida positiva de modo que $v^*Av > 0$ y $v^*D_Bv > 0$ al ser $v \neq \theta$. De aquí se deduce que $1-\alpha > 0$ y en particular que $\alpha \in \mathbb{R}$.

El Teorema 3.7 asegura que los métodos de relajación y Jacobi convergen o divergen simultáneamente. Pero el método de relajación converge por el criterio de Ostrowski-Reich, de modo que ambos convergen y se aplican los Teoremas 3.6 y 3.7. ◇

Observación:

En realidad cualquier matriz puede ser considerada tridiagonal por bloques. Basta considerarla

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}.$$

■

Tema 3

Condicionamiento

3.1. Condicionamiento de sistemas lineales

Intuitivamente parece razonable pensar que al resolver un problema lineal, pequeñas variaciones de los datos deben traducirse en pequeñas variaciones de las soluciones obtenidas. Sin embargo veremos en ejemplos que esto no es siempre así. Diremos en estos casos que nos encontramos ante un problema mal condicionado.

Ejemplo: Es debido a Wilson. Consideremos el sistema lineal $Au = b$, siendo

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

cuya solución es $u^t = (1, 1, 1, 1)$. Si denotamos

$$A' = \begin{pmatrix} 10 & 7 & 8{,}1 & 7{,}2 \\ 7{,}08 & 5{,}04 & 6 & 5 \\ 8 & 5{,}98 & 9{,}89 & 9 \\ 6{,}99 & 4{,}99 & 9 & 9{,}98 \end{pmatrix}, \quad b' = \begin{pmatrix} 32{,}1 \\ 22{,}9 \\ 33{,}1 \\ 30{,}9 \end{pmatrix}$$

y consideramos el sistema $Av = b'$, la solución es $v^t = (9{,}2, -12{,}6, 4{,}5, -1{,}1)$ y si consideramos el sistema $A'w = b$, la solución es $w^t = (-81, 137, -34, 22)$.

Nos planteamos el estudio del sistema lineal cuadrado bien definido siguiente.

$$\begin{cases} \text{Dados } b \in \mathbb{R}^n \text{ y } A \in \mathcal{M}_n(\mathbb{R}) \text{ invertible,} \\ \text{Hallar } u \in \mathbb{R}^n \text{ tal que } Au = b. \end{cases}$$

Si los datos del problema están afectados de error, es decir, si tenemos $A + \delta A$ en vez de A y/o $b + \delta b$ en vez de b, la solución será $u + \delta u$ en vez de u. Nos interesa estudiar el error de condicionamiento δu en relación con los errores de los datos.

Distinguiremos las dos posibilidades siguientes:

1. El condicionamiento con respecto al segundo miembro: $b \to b + \delta b$.
2. El condicionamiento respecto de la matriz: $A \to A + \delta A$.

El problema general se resuelve combinando ambos resultados.

La herramienta que se utiliza para resolver el problema se introduce en la siguiente definición.

Definición 3.1 *Denotemos $\|\cdot\|$ una norma vectorial cualquiera y su norma matricial subordinada. Sea $A \in \mathcal{M}_n(\mathbb{R})$ invertible. Se llama número de condición de A respecto de la norma matricial a*

$$cond\ (A) = \|A\| \cdot \|A^{-1}\|$$

Si A no es invertible, se define $\text{cond}(A) = +\infty$.

3.1.1. Condicionamiento respecto del segundo miembro

Vamos a comparar las soluciones de

$$Au = b, \qquad \text{y} \qquad Av = b + \delta b$$

Teorema 3.1 *Sea $A \in \mathcal{M}_n$ invertible, u la solución de $Au = b$ y $u + \delta u$ la solución de $Av = b + \delta b$, con $b \neq \theta$. Entonces, se verifica*

$$\frac{\|\delta u\|}{\|u\|} \leq cond\ (A) \frac{\|\delta b\|}{\|b\|}$$

donde la norma vectorial que aparece es aquélla de la que es subordinada la norma matricial que define el número de condición de la matriz. Además, la desigualdad es óptima, es decir, existen b y δb no nulos tales que se verifica la igualdad.

Demostración: En efecto:

$$Au = b \Longrightarrow \|b\| = \|Au\| \leq \|A\| \cdot \|u\| \Longrightarrow \|u\| \geq \frac{\|b\|}{\|A\|}$$

$$A(u + \delta u) = b + \delta b \Longrightarrow A(\delta u) = \delta b \Longrightarrow \delta u = A^{-1}(\delta b) \Longrightarrow \|\delta u\| \leq \|A^{-1}\| \cdot \|\delta b\|$$

De donde sigue que los errores relativos verifican

$$\varepsilon_r(u) = \frac{\|\delta u\|}{\|u\|} \leq \frac{\|A^{-1}\| \cdot \|\delta b\|}{\|b\|/\|A\|} = \text{cond } (A)\frac{\|\delta b\|}{\|b\|} = \text{cond } (A)\, \varepsilon_r(b)$$

Para lograr la igualdad, basta considerar que por ser $\|\cdot\|$ una norma matricial subordinada, existen

$$u_0 \in \mathbb{C}^n \setminus \{\theta\}: \quad \|Au_0\| = \|A\| \cdot \|u_0\|$$

$$\delta b_0 \in \mathbb{C}^n \setminus \{\theta\}: \quad \|A^{-1}(\delta b_0)\| = \|A^{-1}\| \cdot \|\delta b_0\|$$

Tomando $b_0 = Au_0 \neq \theta$ y $\delta u_0 = A^{-1}(\delta b_0)$, resulta que todas las desigualdades anteriores son igualdades. ◇

3.1.2. Condicionamiento respecto de la matriz

Teorema 3.2 *a) Sea $A \in \mathcal{M}_n$ invertible, u la solución de $Au = b$ con $b \neq \theta$. Sea $\delta A \in \mathcal{M}_n$ tal que $(A + \delta A)v = b$ tenga una solución, a la que denotamos $u + \delta u$. Entonces, se verifica*

$$\frac{\|\delta u\|}{\|u + \delta u\|} \leq \textit{cond } (A)\frac{\|\delta A\|}{\|A\|}$$

donde la norma vectorial que aparece es aquélla de la que es subordinada la norma matricial que define el número de condición de la matriz y que aparece en el segundo miembro. Además, la desigualdad es óptima, es decir, existen b y δA no nulos tales que se verifica la igualdad.

b) Si $\|\delta A\| \cdot \|A^{-1}\| < 1$, entonces se verifica

$$\frac{\|\delta u\|}{\|u\|} \leq \textit{cond } (A)\frac{\|\delta A\|}{\|A\|}[1 + O(\|\delta A\|)]$$

donde $O(\cdot)$ es el símbolo de Landau.

Demostración: a) Se tiene

$$(A + \delta A)(u + \delta u) = b \Longrightarrow Au + A(\delta u) + (\delta A)(u + \delta u) = b \Longrightarrow$$

$$\delta u = -A^{-1}(\delta A)(u + \delta u) \Longrightarrow \|\delta u\| \leq \|A^{-1}\| \cdot \|\delta A\| \cdot \|u + \delta u\| \Longrightarrow$$

$$\varepsilon_r'(u) = \frac{\|\delta u\|}{\|u + \delta u\|} \leq \frac{\|A^{-1}\| \cdot \|\delta A\| \cdot \|u + \delta u\|}{\|u + \delta u\|} = \text{cond } (A)\frac{\|\delta A\|}{\|A\|} = \text{cond } (A)\varepsilon_r(A)$$

Nótese que si el sistema $(A + \delta A)v = b$ tiene más de una solución, todas ellas verifican la estimación anterior.

Para obtener la igualdad puede procederse como sigue. Busquemos δA en la forma $\delta A = \beta I$ para cierto $\beta \neq 0$. Sabemos que existe

$$w_0 \in \mathbb{C}^n \setminus \{\theta\}: \quad \|A^{-1}w_0\| = \|A^{-1}\| \cdot \|w_0\|$$

Tomamos

$$\begin{cases} \delta u_0 = A^{-1}w_0 \\ u_0 + \delta u_0 = w_0 \Rightarrow u_0 = w_0 - A^{-1}w_0 \\ b_0 = Au_0 = Aw_0 - w_0 \end{cases}$$

Con todo esto, la condición $(A + \delta A)(u_0 + \delta u_0) = b_0$ obliga a que

$$(A + \beta I)w_0 = Aw_0 - w_0 \Longrightarrow \beta = -1$$

Con estos valores, las anteriores desigualdades son igualdades.

b) Sigue ahora del Corolario 1.1 (Tema 1) que $A + \delta A$ es invertible. Por tanto

$$(A + \delta A)(u + \delta u) = b \Longrightarrow (\delta A)u + (A + \delta A)(\delta u) = \theta \Longrightarrow \delta u = -(A + \delta A)^{-1}(\delta A)u$$

$$\|\delta u\| \leq \|(A + \delta A)^{-1}\| \cdot \|\delta A\| \cdot \|u\| \leq \frac{\|I\| \cdot \|A^{-1}\|}{1 - \|A^{-1}\| \cdot \|\delta A\|}\|\delta A\| \cdot \|u\|$$

de modo que recordando que la norma es subordinada, resulta

$$\|\delta u\| \leq \text{cond } (A) \cdot \frac{\|\delta A\|}{\|A\|} \cdot \frac{1}{1 - \|A^{-1}\| \cdot \|\delta A\|} \cdot \|u\|$$

Si se escribe el Teorema del Valor Medio de la función $f(r) = \dfrac{1}{1-r}$ para $|r| < 1$, se tiene

$$f(r) = f(0) + f'(\xi)r, \quad 0 < \xi < r \Longrightarrow f(r) = 1 + \frac{1}{(1-\xi)^2}r, \quad 0 < \xi < r$$

de modo que

$$\frac{1}{1 - \|A^{-1}\| \cdot \|\delta A\|} = 1 + \frac{1}{(1-\xi)^2}\|A^{-1}\| \cdot \|\delta A\|, \quad 0 < \xi < \|A^{-1}\| \cdot \|\delta A\|$$

El último término del segundo miembro es una expresión que tiende a 0 cuando $\|\delta A\|$ tiende a 0, es decir, es un término que en la notación de Landau se puede escribir como $0(\|\delta A\|)$. Así pues

$$\frac{\|\delta u\|}{\|u\|} \leq \text{cond } (A) \cdot \frac{\|\delta A\|}{\|A\|}[1 + 0(\|\delta A\|)]$$

$\diamond$

3.2. Número de condición de una matriz

En los dos teoremas precedentes hemos visto que el error relativo sobre el resultado está mayorado por el error relativo sobre los datos multiplicado por el número de condición, y que esta cota es óptima. En el segundo caso, cuando $\|\delta A\|$ es suficientemente pequeño puede considerarse $\dfrac{\|\delta u\|}{\|u\|}$ en lugar de $\dfrac{\|\delta u\|}{\|u+\delta u\|}$ que es un error relativo más natural. Como consecuencia de ello podemos considerar el número de condición como un indicador de la sensibilidad de la solución de un sistema lineal respecto a las variaciones de los datos, propiedad que se llama condicionamiento del sistema lineal considerado. Así, un sistema está bien o mal condicionado según que su número de condición sea pequeño o grande.

En la práctica, el número de condición utilizado corresponde a alguna de las normas matriciales subordinadas introducidas anteriormente, especialmente, las de las normas subordinadas a $\|\cdot\|_1$, $\|\cdot\|_2$, $\|\cdot\|_\infty$, es decir, $\|\cdot\|_C$, $\|\cdot\|_S$, $\|\cdot\|_F$. Se denotarán $\text{cond}_C(A)$, $\text{cond}_2(A)$, $\text{cond}_F(A)$ respectivamente.

El siguiente resultado recoge una serie de propiedades del número de condición.

Teorema 3.3 *1) $\forall A \in \mathcal{M}_n$, invertible, se verifica*

$$cond\ (A) \geq 1, \quad y \quad cond\ (A) \geq \frac{\text{máx}_{1\leq i\leq n}\,|\lambda_i(A)|}{\text{mín}_{1\leq i\leq n}\,|\lambda_i(A)|}$$

$$\begin{cases} cond\ (A) &= cond\ (A^{-1}) \\ cond\ (\alpha A) &= cond\ (A), \quad \forall \alpha \in \mathbb{K}\setminus\{0\} \end{cases}$$

2) Si $A \in \mathcal{M}_n$ es invertible y normal, entonces

$$cond_2(A) = \frac{\text{máx}_{1\leq i\leq n}\,|\lambda_i(A)|}{\text{mín}_{1\leq i\leq n}\,|\lambda_i(A)|}$$

3) Si $A \in \mathcal{M}_n$ es unitaria (u ortogonal) entonces, $cond_2(A) = 1$.

4) $cond_2(A)$ es invariante ante transformaciones unitarias, es decir,

$$UU^* = I \Longrightarrow cond_2(A) = cond_2(AU) = cond_2(UA) = cond_2(U^*AU), \quad \forall A \in \mathcal{M}_n$$

Demostración: 1) Se tiene

$$I = AA^{-1} \Longrightarrow 1 = \|I\| = \|AA^{-1}\| \leq \|A\|\cdot\|A^{-1}\| = \text{cond}\ (A)$$

Por otra parte,

$$\text{cond}\ (A) = \|A\|\cdot\|A^{-1}\| \geq \rho(A)\rho(A^{-1}) = \max_{1\leq i\leq n}|\lambda_i(A)|\frac{1}{\text{mín}_{1\leq i\leq n}\,|\lambda_i(A)|}$$

Las demás propiedades son evidentes.

2) Como A es regular, $\mu_i(A) > 0, \quad i = 1, ..., n$. Sabemos que $\|A\|_S = \mu_n(A)$. Por otra parte,

$$\|A^{-1}\|_S^2 = \rho((A^{-1})^* A^{-1}) = \rho((AA^*)^{-1}) = \rho((A^*A)^{-1}) =$$

$$\max_{1\leq i\leq n} \lambda_i((A^*A)^{-1}) = \frac{1}{\min_{1\leq i\leq n} \lambda_i(A^*A)} = \frac{1}{\mu_1(A)^2}$$

Por tanto

$$\text{cond}_2(A) = \mu_n(A)\frac{1}{\mu_1(A)}$$

3) Por ser A normal,

$$\|A\|_S = \rho(A) = \max_{1\leq i\leq n} |\lambda_i(A)|$$

Como A^{-1} es también normal,

$$\|A^{-1}\|_S = \rho(A^{-1}) = \frac{1}{\min_{1\leq i\leq n} |\lambda_i(A)|}$$

de donde sigue el resultado.

4) Si A es unitaria (u ortogonal), se tiene que $|\lambda_i(A)| = 1, \quad i = 1, ..., n$. Como además A es normal y se tiene 3), resulta que $\text{cond}_2(A) = 1$.

5) Basta recordar que la norma espectral de una matriz es invariante ante transformaciones unitarias. ◇

Observaciones:

1) La desigualdad, cond $(A) \geq 1$ indica que un sistema lineal estará tanto mejor condicionado cuanto más próximo esté el número de condición a 1. Y a su vez esto depende de que los módulos de los autovalores de la matriz estén próximos o no. Así en el ejemplo de Wilson de la primera pregunta se puede comprobar que $\lambda_1 \approx 0{,}01$ y $\lambda_4 \approx 30{,}28$.

2) De las propiedades 1) y 3) se deduce que para una matriz normal, $\text{cond}_2(A) \leq$ cond (A). Es decir, el mejor número de condición para una matriz normal es el $\text{cond}_2(A)$.

3) Según 3), para una matriz normal el número de condición será grande si son muy distantes los módulos extremos de sus autovalores. Pero para una matriz que no sea normal, el número de condición puede ser grande aunque sus autovalores tengan módulos iguales, porque la propiedad 1) es una desigualdad.

4) De 4) se deduce que las matrices unitarias están muy bien condicionadas. La posibilidad de emplear transformaciones unitarias sin que varíe el número de condición, hace que se utilicen matrices unitarias y ortogonales como matrices auxiliares en algunos métodos de resolución de sistemas lineales (matrices de Householder).

5) Debido a que la norma espectral de una matriz puede ser complicada de obtener, puede ser útil en ocasiones la siguiente estimación

$$\text{cond}_2(A) = \|A\|_S \cdot \|A^{-1}\|_S \leq \|A\|_{ES} \cdot \|A^{-1}\|_{ES}$$

6) En general suele ser necesario manejar acotaciones como la de la nota anterior en vez de manejar el número de condición. El conocimiento de este número necesita el de A^{-1} que suele ser difícil de obtener.

■

3.3. Número de condición y error de redondeo o truncamiento en un sistema lineal

Sea u^* una aproximación por redondeo o truncamiento de la solución u de $Au = b$. Llamaremos $r = Au^* - b$ al vector residual que es la magnitud que usualmente puede medirse. Puede uno pensar que si $\|r\|$ es pequeño, también lo será $\|u - u^*\|$. Pero esto siempre no es así como muestra el siguiente ejemplo

Ejemplo. Consideremos el sistema

$$\begin{pmatrix} 1 & 2 \\ 1{,}0001 & 2 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = \begin{pmatrix} 3 \\ 3{,}0001 \end{pmatrix}$$

que tiene como solución única $u^t = (1, 1)$. La aproximación $(u^*)^t = (3, 0)$ tiene un vector residual

$$r = \begin{pmatrix} 3 \\ 3{,}0001 \end{pmatrix} - \begin{pmatrix} 1 & 2 \\ 1{,}0001 & 2 \end{pmatrix} \begin{pmatrix} 3 \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ -0{,}0002 \end{pmatrix}$$

De modo que $\|r\|_\infty = 0{,}0002$ mientras que $\|u - u^*\|_\infty = 2$

En realidad este problema es un enfoque distinto de algo ya visto. Si consideramos que u^* es la solución de un problema de la forma $Au^* = b^*$, siendo b^* una aproximación de b, $b^* = b + \delta b$, resulta que $r = \delta b$ y estamos ante el estudio del condicionamiento de un sistema lineal respecto del segundo miembro. Se puede, pues, afirmar, por el Teorema 2.1 que

$$\frac{\|u - u^*\|}{\|u\|} \leq \text{cond } (A) \frac{\|r\|}{\|b\|}$$

Nota:

El anterior sistema está mal condicionado. Puede comprobarse que

$$A = \begin{pmatrix} 1 & 2 \\ 1{,}0001 & 2 \end{pmatrix}, \quad A^{-1} = \begin{pmatrix} -10000 & 10000 \\ 5000{,}5 & -5000 \end{pmatrix}$$

$$\text{cond}_F(A) = \|A\|_F \cdot \|A^{-1}\|_F = 3{,}0001 \cdot 20000 = 60002$$

■

3.4. Precondicionamiento

Según hemos visto, un sistema lineal está tanto mejor condicionado cuanto más próximo está a 1 el número de condición de su matriz. La filosofía del precondicionamiento es reemplazar la resolución del sistema $Au = b$ por la del sistema equivalente

$$C^{-1}Au = C^{-1}b$$

donde se elige C^{-1} de modo que cond $(C^{-1}A) <$ cond (A). Es claro que la mejor elección posible es $C = A$ porque cond $(C^{-1}A) =$ cond $(I) = 1$, pero si se conoce A^{-1} entonces la solución del sistema lineal es inmediata.

La idea es, pues, buscar una C "fácil de obtener"para la cual el nuevo número de condición disminuya. Hay pocos métodos de precondicionamiento generales, sino que suelen estar más bien adaptados al método de resolución de sistemas que se utilice. Veremos a continuación uno muy sencillo que se puede utilizar de forma general.

Definición 3.2 *Una matriz $A \in \mathcal{M}_n$ se dice equilibrada por filas si $\sum_{j=1}^{n} |a_{ij}|$ no depende de i (es decir, esta suma es constante).*

La equilibración por filas es un proceso que se consigue multiplicando la matriz por la izquierda por una matriz diagonal regular.

Proposición 3.1 *Toda matriz regular se convierte en una equilibrada al multiplicarla por la izquierda por cierta matriz diagonal regular.*

Demostración: En efecto, sea $B = (b_{ij})$ una matriz regular dada. Buscamos la matriz $D = \text{diag}\,(d_{ii}), \quad d_{ii} > 0$ para que DB sea equilibrada. Se tiene

$$DB = \begin{pmatrix} d_{11} & & \\ & \ddots & \\ & & d_{nn} \end{pmatrix} \begin{pmatrix} b_{11} & \dots & b_{1n} \\ . & \dots & . \\ b_{n1} & \dots & b_{nn} \end{pmatrix} = \begin{pmatrix} d_{11}b_{11} & \dots & d_{11}b_{1n} \\ . & \dots & . \\ d_{nn}b_{n1} & \dots & d_{nn}b_{nn} \end{pmatrix}$$

Si se desea, por ejemplo, que

$$d_{ii}\sum_{j=1}^{n}|b_{ij}| = \alpha$$

basta tomar

$$d_{ii} = \frac{\alpha}{\sum_j |b_{ij}|}, \quad i = 1, ..., n$$

◇

En este caso, la matriz D es la matriz C^{-1} del caso general.

Proposición 3.2 *Sea $B \in \mathcal{M}_n$ una matriz regular y $D \in \mathcal{M}_n$ una matriz diagonal regular tal que DB sea equilibrada por filas. Entonces, $cond_F(DB) \leq cond_F(B)$.*

Demostración: Supongamos que $\|DB\|_F = \alpha$. Se tiene que

$$\|B\|_F = \max_i \sum_j |b_{ij}| = \frac{\alpha}{\min_i |d_{ii}|} = \alpha\|D^{-1}\|_F$$

Entonces

$$\mathrm{cond}_F(DB) = \|DB\|_F\|(DB)^{-1}\|_F \leq$$
$$\alpha\|B^{-1}\|_F\|D^{-1}\|_F = \|B^{-1}\|_F\|B\|_F = \mathrm{cond}_F(B)$$

◇

Vamos a comprobar ahora que esta matriz es la mejor matriz diagonal que podemos tomar para disminuir el número de condición fila de la matriz.

Proposición 3.3 *Sea $A \in \mathcal{M}_n$ una matriz regular y $D_1,\ D_2 \in \mathcal{M}_n$ matrices diagonales regulares tales que D_1A sea equilibrada por filas y D_2A no. Entonces, $cond_F(D_1A) \leq cond_F(D_2A)$.*

Demostración: Ya que $D_1A = D_1D_2^{-1}D_2A$, basta aplicar la Proposición 2.2 a $D = D_1D_2^{-1}$ y a $B = D_2A$. ◇

Ejemplo: Es fácil comprobar que si

$$A = \begin{pmatrix} 1 & 10^8 \\ 2 & 0 \end{pmatrix}, \quad \text{entonces} \quad A^{-1} = \begin{pmatrix} 0 & 1/2 \\ 10^{-8} & -10^{-8}/2 \end{pmatrix}$$

y que

$$\|A\|_F = 1 + 10^8, \quad \|A^{-1}\|_F = 1/2, \quad \mathrm{cond}_F(A) = \frac{1 + 10^8}{2}$$

Si se precondiciona por filas para, por ejemplo, $\alpha = 1$, resulta

$$DA = \begin{pmatrix} (1+10^8)^{-1} & 10^8(1+10^8)^{-1} \\ 1 & 0 \end{pmatrix}, \quad (DA)^{-1} = \begin{pmatrix} 0 & 1 \\ 1 + 10^{-8} & -10^{-8} \end{pmatrix}$$

y, por tanto,

$$\text{cond}_F(DA) = 1 + 2 \cdot 10^{-8}$$

Nota:

Algunos autores, sobre todo cuando el método de resolución que se utiliza es el de Gauss, llaman equilibración por filas al proceso que consiste en dividir cada fila de A por el máximo de los valores absolutos de los elementos de dicha fila. De esta forma se consigue que el máximo valor absoluto en cada fila de la nueva matriz sea 1. ■

3.5. Condicionamiento de un problema de autovalores

Comenzaremos planteando el problema que pretendemos estudiar con el siguiente ejemplo. Se considera la matriz de orden n

$$A(\varepsilon) = \begin{pmatrix} 0 & 0 & \dots & 0 & \varepsilon \\ 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & \dots & 0 & 0 \\ . & . & \ddots & . & . \\ 0 & 0 & \dots & 1 & 0 \end{pmatrix}$$

Su polinomio característico es $\det(\lambda I - A) = \lambda^n - \varepsilon$. Para $\varepsilon = 0$, se tiene que $\text{sp}(A) = \{0\}$, mientras que para $\varepsilon \neq 0$, el espectro de A está constituido por las raíces n-simas de ε. Por ejemplo, si $n = 40$ y $\varepsilon = 10^{-40}$, los autovalores de $A(\varepsilon)$ tienen de módulo 10^{-1}; es decir, la variación de los autovalores medida en el plano complejo es igual a la variación de ε multiplicada por 10^{39}. Observamos que pequeñas variaciones de los datos provocan grandes variaciones en los resultados. Pretendemos estudiar cómo afectan al cálculo de autovalores pequeñas variaciones de la matriz. Nos restringiremos al caso de las matrices diagonalizables.

Teorema 3.4 *(Bauer-Fike). Sea $A \in \mathcal{M}_n$ diagonalizable, P una matriz regular tal que $P^{-1}AP = diag\ (\lambda_i(A))$ y $\|\cdot\|$ una norma matricial subordinada que verifique que para cualquier matriz diagonal,*

$$\|\text{diag}\ (d_i)\| = \max_i |d_i|$$

Entonces, para cualquier matriz δA, se verifica

$$\text{sp}\ (A + \delta A) \subset \cup_{i=1}^n D_i, \quad \textit{siendo} \quad D_i = \{z \in \mathbb{C} :\ |z - \lambda_i(A)| \leq cond\ (P)\|\delta A\|\}$$

Demostración: Sea $\lambda \in \text{sp } (A + \delta A)$. Entonces, $A + \delta A - \lambda I$ es una matriz singular.

Si $\lambda = \lambda_j(A)$ para algún j, el resultado es trivial. Supongamos, pues, que $\lambda \neq \lambda_i(A),\ i = 1, ..., n$. Entonces, $D - \lambda I$ es invertible y podemos escribir

$$P^{-1}(A + \delta A - \lambda I)P = D - \lambda I + P^{-1}(\delta A)P = (D - \lambda I)[I + (D - \lambda I)^{-1}P^{-1}(\delta A)P]$$

Como el primer miembro es singular y el primer factor del segundo es regular, se deduce que $I + (D - \lambda I)^{-1}P^{-1}(\delta A)P$ es singular y del Teorema de Inversión de Matrices, sigue que

$$1 \leq \|(D - \lambda I)^{-1}P^{-1}(\delta A)P\|$$

y, por tanto,

$$1 \leq \|(D - \lambda I)^{-1}\| \cdot \|P^{-1}\| \cdot \|\delta A\| \cdot \|P\| = \text{cond } (P) \cdot \|(D - \lambda I)^{-1}\| \cdot \|\delta A\|$$

Por la hipótesis sobre la norma matricial, $\|(D - \lambda I)^{-1}\| = \dfrac{1}{\text{mín}_i\, |\lambda_i(A) - \lambda|}$. De modo que

$$1 \leq \frac{1}{\text{mín}_i\, |\lambda_i(A) - \lambda|}\text{cond } (P)\|\delta A\| \Longrightarrow \exists j :\ |\lambda - \lambda_j(A)| \leq \text{cond } (P)\|\delta A\|$$

◇

El número de condición que interviene ahora es el de la matriz de paso a la matriz diagonal. Hay muchas matrices de paso posibles; ello lleva a definir

Definición 3.3 *Se llama número de condición de A respecto del problema de autovalores a*

$$\Gamma(A) = \text{ínf}\{\text{cond } (P) :\ P^{-1}AP = \textit{diag } (\lambda_i(A))\}$$

Corolario 3.1 *En las condiciones del Teorema 2.4, se verifica*

$$\textit{sp } (A + \delta A) \subset \cup_{i=1}^{n}\{z \in \mathbb{C} :\ |z - \lambda_i(A)| \leq \Gamma(A)\|\delta A\|\}$$

Notas:

1) La propiedad que se le pide a la norma matricial en el Teorema de Bauer-Fike es verificada por las normas subordinadas más usuales, por ejemplo, por $\|\cdot\|_F$, $\|\cdot\|_C$, $\|\cdot\|_S$.

2) En principio, cond (A) y $\Gamma(A)$ no están relacionados.

3) Cuando A es normal (en particular simétrica), sabemos que es diagonalizable con matriz de paso unitaria. Y es sabido que si P es unitaria $\text{cond}_2(P) = 1$. Por tanto $\Gamma_2(A) = 1$ también. Es decir, las matrices normales están muy bien condicionadas para el problema de valores propios.

■

En general, tan solo puede afirmarse que los autovalores de $A+\delta A$ están en la unión de las bolas centradas en cada autovalor de A. En el caso particular de las matrices normales sabemos que $\Gamma_2(A) = 1$ y que

$$\text{sp } (A + \delta A) \subset \cup_{i=1}^{n}\{z \in \mathbb{C} : \ |z - \lambda_i(A)| \leq \|\delta A\|_S\}$$

Pero puede afirmarse más si las matrices A y δA son hermíticas: se sabe en qué bola está cada autovalor de la matriz perturbada.

Teorema 3.5 *Sean A y δA dos matrices hermíticas (simétricas). Sean $\alpha_1 \leq \alpha_2 \leq ... \leq \alpha_n$ los autovalores de A y $\beta_1 \leq \beta_2 \leq ... \leq \beta_n$ los autovalores de $A + \delta A$. Entonces*

$$|\alpha_j - \beta_j| \leq \|\delta A\|_S, \quad j = 1, ..., n$$

Demostración: Aplicaremos el Teorema de Courant-Fisher. Denotamos $\{p^1, ..., p^n\}$ una base ortonormal de autovectores de A asociados a $\{\alpha_1, ...\alpha_n\}$. Sea $V_k = \langle p^1, ..., p^k\rangle$ y $\mathcal{V}_k$ el conjunto de subespacios de dimensión k de $\mathbb{C}^n$. Se tiene por dicho Teorema

$$\beta_k = \min_{W\in\mathcal{V}_k} \max_{v\in W\setminus\{\theta\}} R_{A+\delta A}(v) \leq \max_{v\in V_k\setminus\{\theta\}} R_{A+\delta A}(v) = \max_{v\in V_k\setminus\{\theta\}} (R_A(v) + R_{\delta A}(v)) \leq$$

$$\max_{v\in V_k\setminus\{\theta\}} R_A(v) + \max_{v\in V_k\setminus\{\theta\}} R_{\delta A}(v) = \alpha_k + \max_{v\in V_k\setminus\{\theta\}} R_{\delta A}(v) \leq \alpha_k + \max_{v\in\mathbb{C}^n\setminus\{\theta\}} R_{\delta A}(v)$$

Pero

$$\max_{v\in\mathbb{C}^n\setminus\{\theta\}} R_{\delta A}(v) = \lambda_n(\delta A) \leq \rho(\delta A) = \|\delta A\|_S;$$

por tanto

$$\beta_k \leq \alpha_k + \|\delta A\|_S$$

Intercambiando A y $A + \delta A$ se obtiene

$$\alpha_k \leq \beta_k + \|\delta A\|_S$$

y, por tanto

$$|\alpha_k - \beta_k| \leq \|\delta A\|_S$$

◇

Tema 4

Métodos de Descenso para la resolución de Sistemas Lineales

4.1. Métodos de Descenso

Consideramos de nuevo el problema de hallar la solución de

$$Au = b, \quad A \in \mathcal{M}(\mathbb{R}) \text{ simetrica definida positiva}, \quad b \in \mathbb{R}^n$$

Denotemos $\bar{u}$ la solución del mismo. Definimos

Definición 4.1 *Dado un vector $u \in \mathbb{R}^n$, se llama residuo de u a*

$$r(u) = b - Au = A(\bar{u} - u)$$

En lo que sigue denotaremos $(\cdot, \cdot)$ el producto escalar euclídeo en $\mathbb{R}^n$. Consideremos la forma cuadrática

$$q: \ \mathbb{R}^n \longrightarrow \mathbb{R}, \quad q(u) = (u, Au) - 2(u, b)$$

y también la aplicación

$$E(u) = (A(\bar{u} - u), \bar{u} - u)$$

Se tiene entonces la siguiente

Proposición 4.1 *La forma cuadrática q y la aplicación E alcanzan su mínimo (si lo hacen) en los mismos puntos.*

Demostración: Esto sigue de que ambas expresiones se diferencian en una constante. En efecto, por ser A simétrica,

$$E(u) = (A(\bar{u} - u), \bar{u} - u) = (A\bar{u}, \bar{u}) - (A\bar{u}, u) - (Au, \bar{u}) + (Au, u) =$$

$$(A\bar{u}, \bar{u}) - 2(u, A\bar{u}) + (u, Au) = q(u) + (A\bar{u}, \bar{u})$$

$\diamond$

Observaciones:

1) Nótese que se puede expresar

$$E(u) = (r(u), A^{-1}r(u))$$

2) También se tiene la siguiente desigualdad

$$E(u) = \|\bar{u} - u\|_2^2 R_A(\bar{u} - u) \geq \lambda_1 \|\bar{u} - u\|_2^2 > 0$$

siendo λ_1 el menor autovalor de A.

El resultado fundamental que sugiere los métodos de descenso es

Proposición 4.2 *Sea A simétrica definida positiva. Entonces son equivalentes los siguientes problemas*

1. *Hallar la solución $\bar{u}$ del sistema $Au = b$.*
2. *Obtener el vector u^* que proporciona el mínimo de $q(u)$.*

Demostración: Veamos el comportamiento de q a lo largo de la recta

$$\alpha \in \mathbb{R} \longmapsto v + \alpha p$$

donde $v,\ p$ son dos vectores fijos no nulos cualesquiera y α es un escalar. Se tiene por ser A simétrica

$$q(v + \alpha p) = (v + \alpha p, A(v + \alpha p)) - 2(v + \alpha p, b) =$$
$$(v, Av) + \alpha(v, Ap) + \alpha(p, Av) + \alpha^2(p, Ap) - 2(v, b) - 2\alpha(p, b) =$$
$$q(v) + 2\alpha(p, Av) - 2\alpha(p, b) + \alpha^2(p, Ap) = q(v) + 2\alpha(p, Av - b) + \alpha^2(p, Ap)$$

Se obtiene una parábola en α de coeficiente principal positivo por ser A definida positiva. De modo que tendrá un mínimo en $\hat{\alpha}$ que puede calcularse

$$\frac{d}{d\alpha} q(v + \alpha p) = 2(p, Av - b) + 2\alpha(p, Ap) = 0 \Longrightarrow$$

$$\hat{\alpha} = -\frac{(p, Av - b)}{(p, Ap)} = \frac{(p, r(v))}{(p, Ap)}$$

El valor del mínimo a lo largo de la recta es

$$q(v + \hat{\alpha}p) = q(v) + \hat{\alpha}[2(p, Av - b) + \hat{\alpha}(p, Ap)] =$$

$$q(v) + \hat{\alpha}[2(p, Av - b) + (p, r(v))] = q(v) - \hat{\alpha}(p, r(v)) = q(v) - \frac{(p, r(v))^2}{(p, Ap)}$$

La expresión $(p, r(v))^2 > 0$ salvo en los casos en que $r(v) = 0$ o que p sea perpendicular a $r(v)$; en todos los demás hay una disminución del valor de q al pasar de v a $v + \hat{\alpha}p$.

Probamos ahora el enunciado:

a) Sea $\bar{u}$ la solución del sistema $Au = b$. Si tomamos $v = \bar{u}$, entonces $r(\bar{u}) = \theta$ y

$$q(\bar{u} + \hat{\alpha}p) = q(\bar{u}) \leq q(\bar{u} + \alpha p)$$

cualesquiera que sean p y α. En $\bar{u}$ se obtiene el mínimo de q.

b) Sea u^* el vector donde q alcanza su mínimo y supongamos que $r(u^*) \neq \theta$. Entonces puede escogerse un p tal que $(p, r(u^*)) \neq 0$, y puede determinarse el correspondiente $\hat{\alpha}$ para el que

$$q(u^* + \hat{\alpha}p) < q(u^*)$$

en contradicción con que el mínimo se alcanza en u^*.

◇

Esta Proposición sugiere un método indirecto para resolver $Au = b$

Definición 4.2 *Un método de descenso es un método indirecto que se construye eligiendo en la k-sima iteración una dirección $p_k \neq \theta$ y un escalar α_k de manera que*

$$u_{k+1} = u_k + \alpha_k p_k, \quad y \quad q(u_{k+1}) < q(u_k)$$

Según hemos visto, fijada la dirección p_k y escogiendo la elección óptima para α_k queda

$$u_{k+1} = u_k + \frac{(r_k, p_k)}{(Ap_k, p_k)} p_k$$

donde hemos denotado $r_k = r(u_k)$.

Proposición 4.3 *Para cualquier elección de $p_k \neq \theta$ y para el α_k óptimo, se verifica:*

a) $r_{k+1} = r_k - \hat{\alpha}_k A p_k, \quad \forall k \geq 0$

b) $(p_k, r_{k+1}) = 0, \quad \forall k \geq 0$

Demostración: De la definición de u_{k+1} se obtiene

$$Au_{k+1} = Au_k + \frac{(r_k, p_k)}{(Ap_k, p_k)} Ap_k \Longrightarrow Au_{k+1} - b = Au_k - b + \hat{\alpha}_k Ap_k$$

que es la relación a).

Multiplicando esta igualdad por p_k queda

$$(p_k, r_{k+1}) = (p_k, r_k) - \hat{\alpha}_k (p_k, Ap_k) = 0$$

por la definición de $\hat{\alpha}_k$. ◇

4.1.1. Interpretacion geométrica de los métodos de descenso

Puede hacerse una interpretación geométrica en $\mathbb{R}^2$ o $\mathbb{R}^3$ que permite intuir lo que hace el método en $\mathbb{R}^n$.

En $\mathbb{R}^2$, $E(u) = Cte$ es una elipse. Al variar la constante se obtiene una familia de elipses homotéticas cuyo centro es $\bar{u}$.

En la etapa k del método se determina u_k y por tanto se determina una elipse de la familia: la que tiene de ecuación $E(u) = E(u_k)$. Escogida ahora una dirección p_k cualquiera, existe una única elipse de la familia que es tangente a la recta que pasa por u_k y tiene de dirección p_k. El punto de tangencia, que es interior a la elipse $E(u) = E(u_k)$ y por tanto más próximo a $\bar{u}$, es u_{k+1} y se obtiene como $u_{k+1} = u_k + \hat{\alpha}_k p_k$. Nótese que si la dirección p_k que se escoge es la tangente a la elipse $E(u) = E(u_k)$, entonces $u_{k+1} = u_k$.

Algoritmo de los Métodos de Descenso con paso óptimo:

$$\begin{cases} u_0 \in \mathbb{R}^n,\ p_0 \in \mathbb{R}^n,\ r_0 = b - Au_0 \quad \text{(inicialización)} \\ \text{En la etapa } k, \text{ conocidos} \quad u_k, p_k, r_k \in \mathbb{R}^n \\ \hat{\alpha}_k = \dfrac{(r_k, p_k)}{(Ap_k, p_k)}, \quad k \geq 0 \\ r_{k+1} = r_k - \hat{\alpha}_k A p_k, \quad k \geq 0 \\ u_{k+1} = u_k + \hat{\alpha}_k p_k, \quad k \geq 0 \end{cases}$$

4.1.2. Condicion suficiente de convergencia

De los resultados anteriores se tiene

$$E(u_{k+1}) = E(u_k) - 2\hat{\alpha}_k(r_k, p_k) + \hat{\alpha}_k^2(Ap_k, p_k) \Longrightarrow E(u_{k+1}) = E(u_k) - \frac{(p_k, r_k)^2}{(p_k, Ap_k)}$$

$$E(u_{k+1}) = E(u_k)(1 - \gamma_k), \quad \gamma_k = \frac{(r_k, p_k)^2}{(r_k, A^{-1}r_k)(Ap_k, p_k)}$$

Lema 4.1 *Para cualquier elección de $p_k \neq \theta$ y para el α_k óptimo, se tiene*

$$\gamma_k \geq \frac{1}{cond_2(A)} \left(\frac{r_k}{\|r_k\|_2}, \frac{p_k}{\|p_k\|_2} \right)^2, \quad \forall k \geq 0$$

Demostración: Por ser A simétrica, $\|A\|_S = \sup_{v \neq \theta} R_A(v)$. De modo que

$$\text{cond}_2(A) = \|A\|_S \|A^{-1}\|_S \geq \frac{(Ap_k, p_k)}{\|p_k\|_2^2} \cdot \frac{(r_k, A^{-1}r_k)}{\|r_k\|_2^2}$$

de donde se deduce el Lema. ◇

Teorema 4.1 *Consideremos el método de descenso*

$$\begin{cases} u_0 \in \mathbb{R}^n \\ u_{k+1} = u_k + \dfrac{(r_k, p_k)}{(Ap_k, p_k)} p_k, \quad k \geq 0 \end{cases}$$

donde las direcciones p_k son tales que existe un número $\mu > 0$ independiente de k que verifica $\left(\dfrac{r_k}{\|r_k\|_2}, \dfrac{p_k}{\|p_k\|_2} \right)^2 \geq \mu > 0$. *Entonces, se verifica que* $\lim\limits_{k \to +\infty} u_k = \bar{u}$, *es decir, la sucesión* $\{u_k\}$ *converge hacia la solución que minimiza* $E(u)$.

Demostración: Por hipótesis

$$1 - \gamma_k \leq 1 - \frac{\mu}{\text{cond}_2(A)}$$

Además, por la desigualdad de Cauchy-Schwarz, $0 < \mu \leq 1$, y es sabido que $\text{cond}_2(A) \geq 1$. De modo que $0 \leq 1 - \dfrac{\mu}{\text{cond}_2(A)} < 1$.

Entonces,

$$E(u_k) \leq E(u_{k-1}) \left(1 - \frac{\mu}{\text{cond}_2(A)} \right) \leq ... \leq E(u_0) \left(1 - \frac{\mu}{\text{cond}_2(A)} \right)^k$$

y, por tanto,

$$\lim_{k \to \infty} \|\bar{u} - u_k\|_2^2 \leq \frac{1}{\lambda_1} \lim_{k \to \infty} E(u_k) = 0$$

◇

Nota: Este Teorema da una condición suficiente de convergencia, a saber, que p_k no se haga asintóticamente ortogonal a r_k. Una posible elección evidente es escoger $p_k = r_k$. Es lo que se hace en el método siguiente.

4.1.3. Metodo del gradiente:

El método de gradiente con paso óptimo consiste en tomar $p_k = r_k$

$$\begin{cases} u_0 \in \mathbb{R}^n, \ r_0 = b - Au_0 \quad \text{(inicialización)} \\ u_{k+1} = u_k + \dfrac{\|r_k\|_2^2}{(Ar_k, r_k)} r_k, \quad k \geq 0 \\ r_{k+1} = r_k - \dfrac{\|r_k\|_2^2}{(Ar_k, r_k)} Ar_k, \quad k \geq 0 \end{cases}$$

En este caso se verifica que el método es convergente pues

$$E(u_{k+1}) = E(u_k)(1 - \gamma_k), \quad \gamma_k = \frac{\|r_k\|_2^2}{(r_k, A^{-1}r_k)(Ar_k, r_k)} \quad \text{y} \quad \left(\frac{r_k}{\|r_k\|_2}, \frac{r_k}{\|r_k\|_2} \right) = 1 (= \mu)$$

Además, puede probarse que el número de iteraciones necesarias para conseguir que $\frac{E(u_k)}{E(u_0)} \leq \varepsilon$ es del orden de $k \approx \frac{1}{4}\text{cond}_2(A) \ln\left(\frac{1}{\varepsilon}\right)$; es decir, el número de iteraciones es proporcional a $\text{cond}_2(A)$.

Si $\text{cond}_2(A)$ es grande, lo que sucede cuando los valores propios tienen módulos extremos muy diferentes, los elipsoides $E(u) = cte$ son muy achatados y la convergencia es lenta. Es lo que sucede en el caso de las matrices que proceden de un operador diferencial cuyo número de condición suele aumentar al afinar la discretización. Por eso este método tiene poco interés práctico y se buscan métodos más eficaces. La idea es intentar elegir p_k que apunte hacia el centro de los elipsoides.

4.1.4. Metodo de gradiente conjugado

Ahora no elegimos $r_k = p_k$, razonamos de la siguiente forma: fijado p_{k-1}, sabemos que con la elección óptima de α_k, se tiene:

$$(p_{k-1}, r_k) = 0 \Longrightarrow (p_{k-1}, A(\bar{u} - u_k)) = 0, \quad \forall k \geq 1$$

Si queremos que $u_{k+1} \approx \bar{u}$, la dirección $p_k = \frac{1}{\hat{\alpha}_k}(u_{k+1} - u_k)$ debe verificar algo similar a la igualdad anterior. Por ello exigiremos que

$$(p_{k-1}, Ap_k) = 0 \Longrightarrow (p_k, Ap_{k-1}) = 0, \quad \forall k \geq 1.$$

Consideraremos $p_0 = r_0$ y, en la etapa k, escogeremos p_{k+1} en el plano formado por p_k y r_{k+1}, es decir

$$p_{k+1} = r_{k+1} + \beta_{k+1} p_k, \quad \forall k \geq 0$$

con β_{k+1} a elegir tal que $(p_{k+1}, Ap_k) = 0$.

Se verifican entonces

Lema 4.2 *En las condiciones anteriores, se tiene:*

a) $(r_k, p_k) = \|r_k\|_2^2, \quad \forall k \geq 0$

b) $(r_k, r_{k+1}) = 0, \quad \forall k \geq 0$

c)

$$\begin{cases} \beta_0 = 0 \\ \beta_{k+1} = \dfrac{\|r_{k+1}\|_2^2}{\|r_k\|_2^2}, \quad \forall k \geq 0. \end{cases}$$

Demostración: a) Se tiene

$$(r_k, p_k) = (r_k, r_k + \beta_k p_{k-1}) = (r_k, r_k) + \beta_k (r_k, p_{k-1}) = \|r_k\|_2^2, \quad \forall k \geq 1$$

por la Proposición 4.3 b). Para $k = 0$ sigue tomando $p_0 = r_0$.

b) Por la Proposición 4.3 a), $r_{k+1} = r_k - \hat{\alpha}_k A p_k$. Por tanto,

$$(r_k, r_{k+1}) = (r_k, r_k) - (r_k, \hat{\alpha}_k A p_k) = (r_k, r_k) - \frac{(r_k, p_k)}{(Ap_k, p_k)}(r_k, Ap_k) =$$

$$(r_k, r_k)\left(1 - \frac{(r_k, Ap_k)}{(Ap_k, p_k)}\right) = (r_k, r_k)\frac{(p_k - r_k, Ap_k)}{(Ap_k, p_k)} = \|r_k\|_2^2 \frac{\beta_k (p_{k-1}, Ap_k)}{(Ap_k, p_k)} = 0$$

Esta igualdad sale para $k \geq 1$; para $k = 0$ resulta tomando también $p_0 = r_0$.

c) Vamos a pedir que $\forall k \geq 0, \quad (p_{k+1}, Ap_k) = 0$. Así se puede determinar β_k. En efecto,

$$(Ap_k, p_{k+1}) = 0 \Longrightarrow (Ap_k, r_{k+1} + \beta_{k+1} p_k) = 0 \Longrightarrow \beta_{k+1} = -\frac{(Ap_k, r_{k+1})}{(Ap_k, p_k)}$$

Aplicando la Proposición 4.3 a) resulta

$$\beta_{k+1} = -\frac{\hat{\alpha}_k (r_k - r_{k+1}, r_{k+1})}{\hat{\alpha}_k (r_k - r_{k+1}, p_k)} = \frac{-(r_k, r_{k+1}) + (r_{k+1}, r_{k+1})}{(r_k, p_k) - (r_{k+1}, p_k)}$$

Utilizando el Lema 4.2 b), la Proposición 4.3 b) y el Lema 4.2 a) resulta

$$\beta_{k+1} = \frac{\|r_{k+1}\|_2^2}{\|r_k\|_2^2}, \quad \forall k \geq 0$$

◇

Ello origina el siguiente método, denominado método de gradiente conjugado:

$$\begin{cases} \text{Se inicializa } u_0 \in \mathbb{R}^n, \quad p_0 = r_0 = b - A u_0, \\ \text{dados } u_k, p_k, r_k \in \mathbb{R}^n, \quad \text{se obtienen :} \\ \alpha_k = \dfrac{\|r_k\|_2^2}{(Ap_k, p_k)} \\ u_{k+1} = u_k + \alpha_k p_k \\ r_{k+1} = r_k - \alpha_k A p_k \\ \beta_{k+1} = \dfrac{\|r_{k+1}\|_2^2}{\|r_k\|_2^2} \end{cases} \qquad \text{para } k = 0, 1, 2, ...$$

El test de parada de las iteraciones se hace sobre $\|r_k\|_2$.

La prueba del teorema de convergencia se apoya en el siguiente Lema.

Lema 4.3 *En las condiciones anteriores, se verifica,*

a) $(r_{k+1}, p_i) = 0, \qquad i = 0, ..., k$

b) $(p_{k+1}, Ap_i) = 0, \qquad i = 0, ..., k$

c) $(r_{k+1}, r_i) = 0, \qquad i = 0, ..., k$

Demostración: Nótese que de la Proposición 3.4 a), sigue que

$$r_k \in \langle r_{k-1}, Ap_{k-1} \rangle$$

Por otra parte,

$$p_{k+1} = r_{k+1} + \beta_{k+1} p_k = r_k - \hat{\alpha}_k A p_k + \beta_{k+1} p_k \Rightarrow A p_k \in \langle r_k, p_k, p_{k+1} \rangle$$

Ello implica que

$$r_k \in \langle p_0, ..., p_k \rangle, \qquad A p_k \in \langle p_0, ..., p_{k+1} \rangle, \qquad \forall k \geq 0$$

Procedemos por inducción. El resultado es conocido para $k = 1$. Supongámoslo cierto para k.

a) Hay que probar que $(r_{k+1}, pi) = 0, \quad i = 0, ..., k$. Se tiene

$$(r_{k+1}, p_i) = (r_k, p_i) - \hat{\alpha}_k (A p_k, p_i i)$$

Para $i = 0, ..., k-1$, sigue de la hipótesis de inducción que cada sumando es 0. Para $i = k$, es la Proposición 3.4 b).

b) Hay que probar que $(p_{k+1}, Ap_i) = 0, \quad i = 0, ..., k$. Se tiene

$$(p_{k+1}, Ap_i) = (r_{k+1}, Ap_i) + \beta_{k+1} (p_k, Ap_i)$$

Para $i = k$, es la condición que se le exige al método de gradiente conjugado. Para $i = 0, ..., k-1$, se tiene que el segundo sumando es 0 por la hipótesis de inducción, mientras que el primero es también 0, porque

$$Ap_i \in \langle p_0, ..., p_{i+1} \rangle \subset \langle p_0, ..., p_k \rangle$$

y, por el apartado a), $(r_{k+1}, p_j) = 0, \quad j = 0, ..., k$.

c) Hay que probar que $(r_{k+1}, r_i) = 0, \quad i = 0, ..., k$. Se tiene

$$(r_{k+1}, r_i) = (r_k, r_i) - \hat{\alpha}_k (A p_k, r_i)$$

Para $i = k$ es el Lema 3.4 b). Para $i = 0, ..., k-1$, se tiene que el primer sumando es 0 por la hipótesis de inducción, mientras que el segundo es también 0, porque

$$r_i \in \langle p_0, ..., p_i \rangle \Longrightarrow A r_i \in \langle A p_0, ..., A p_i \rangle$$

y el la hipótesis de inducción asegura que $(p_k, Ap_j) = 0, \quad j = 0, ..., k-1$. ◇

Tenemos el resultado siguiente

Teorema 4.2 *El método de gradiente conjugado es exacto en un máximo de n iteraciones.*

Demostración: Del Lema 4.3 c) sigue que los vectores distintos $\{p_0, p_1, ..., p_j\}$ son linealmente independientes por ser ortogonales respecto del producto escalar $(\cdot, A\cdot)$.

Al realizar el método de gradiente conjugado, puede suceder que

-) $r_k = \theta$, para algún $k = 0, ..., N-1$. Entonces, el método es exacto en la iteración k.

-) $r_k \neq \theta$ para $i = 1, ..., N-1$. Entonces $\{p_0, ..., p_{N-1}\}$ son diferentes y por ser linealmente independientes forman base de $\mathbb{R}^N$. De modo que por el Lema 4.3 a), sigue que r_N es ortogonal a una base. Por tanto $r_N = \theta$ y el método es exacto en la iteración N

$\diamond$

En teoría, pues, el método de gradiente conjugado es un método directo, pero debido a los errores de redondeo se convierte en un método indirecto. Se prueba que el número de iteraciones necesarias para hacer que $\dfrac{E(u_k)}{E(u_0)} < \varepsilon$ es del orden de $\dfrac{1}{2}\log\dfrac{2}{\varepsilon}\sqrt{\text{cond}_2(A)} + 1$, es decir, el número de iteraciones es proporcional a $\sqrt{\text{cond}_2(A)}$, lo que disminuye el número de iteraciones con respecto al método de gradiente. Sin embargo, el número puede seguir siendo excesivo si la matriz está mal condicionada, en cuyo caso se acude a técnicas de precondicionamiento.

Tema 5

Localización y aproximación de autovalores y autovectores

5.1. Introducción

Nos preocupamos en este Tema de dar métodos que permitan aproximar el conjunto de valores propios de una matriz. Es sabido que los autovalores de una matriz $A = (a_{ij})$ son las raices de la ecuación característica

$$p(\lambda) = |A - \lambda I| = 0$$

que es un polinomio de grado n. Recíprocamente, puede comprobarse que la ecuación polinómica general

$$x^n + a_{n-1}x^{n-1} + ... + a_1 x + a_0 = 0 \tag{4.1}$$

es la ecuación característica de la matriz (llamada matriz de Frobenius)

$$A = \begin{pmatrix} -a_{n-1} & -a_{n-2} & \dots & -a_1 & -a_0 \\ 1 & 0 & \dots & 0 & 0 \\ \cdot & \cdot & \dots & \cdot & \cdot \\ 0 & 0 & \dots & 1 & 0 \end{pmatrix} \tag{4.2}$$

de modo que las raíces de (4.1) son los autovalores de (4.2). Esta consideración lleva a deducir que los métodos de cálculo de valores propios solo pueden ser iterativos pues la existencia de un método directo equivaldría a afirmar que se pueden calcular las raíces de un polinomio arbitrario en un número finito de operaciones elementales en contradicción con el Teorema de Abel relativo a la imposibilidad de resolver por radicales una ecuación de grado ≥ 5.

El cálculo del polinomio característico es muy costoso. Por ello no se encuentran métodos que calculen valores propios a partir del polinomio característico; más bien al contrario,

para calcular raíces de polinomios de grado elevado suele ser frecuente utilizar los métodos que se van a indicar ahora a su matriz asociada (matriz de Frobenius).

Un método que se revela efectivo es el método de la potencia, aplicable a matrices diagonalizables y que tengan un autovalor de módulo máximo. Una vez aproximado éste, mediante una técnica de deflación se puede ir aproximando el de mayor módulo de los que quedan y así se va procediendo sucesivamente. Hay variantes del método que cubren el caso en que hay autovalores de módulos iguales. El método además está bien adaptado para la aproximación de los autovectores.

Para matrices simétricas generales se tiene el método de Jacobi y para matrices simétricas tridiagonales el de Givens que permite el cálculo aproximado con una precisión arbitraria de un valor propio en un rango dado.

Para matrices no simétricas, hay también un método disponible que se apoya en una descomposición factorial de la matriz llamada descomposición QR y que permite aplicar un método parecido al de Jacobi.

Observación:

Existe un método debido a Householder que reduce cualquier matriz simétrica, A, a una matriz simétrica tridiagonal, P^tAP, mediante una matriz ortogonal, P. De este manera, el método de Householder-Givens es aplicable a toda matriz simétrica. ■

El marco natural de este problema es el cuerpo $\mathbb{C}$. Por ello, supondremos que $A \in \mathcal{M}_n(\mathbb{C})$ y particularizaremos los resultados obtenidos si $A \in \mathcal{M}_n(\mathbb{R})$

5.2. Localización de autovalores

Son conocidos diversos resultados de localización de autovalores. Resultados parciales son los Teoremas 2.4 (Bauer-Fike) y 2.5.

De la Proposición 1.1 sigue en particular que

Proposición 5.1 *Se verifica que si $A \in \mathcal{M}_n$, entonces*

$$\forall\, \lambda \in \text{sp } (A), \quad |\lambda| \leq \text{mín}\{\|A\|_F, \|A\|_C\}$$

El resultado general más conocido es

Teorema 5.1 (círculos de Gerschgorin) *Sea $A = (a_{ij}) \in \mathcal{M}_n(\mathbb{C})$ y denotemos*

$$P_i = \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}|, \qquad Q_j = \sum_{\substack{i=1 \\ i\neq j}}^{n} |a_{ij}|$$

Entonces

a) $sp\ (A) \subset \bigcup_{i=1}^{n} C_i, \quad donde \quad C_i = \{z \in \mathbb{C} :\ |z - a_{ii}| \leq P_i\}$

b) $sp\ (A) \subset \bigcup_{j=1}^{n} D_j, \quad donde \quad D_j = \{z \in \mathbb{C} :\ |z - a_{jj}| \leq Q_j\}$

c) Si $\hat{S}$ es una unión de m discos (C_i o D_j) que es disjunta con los restantes discos, entonces $\hat{S}$ contiene precisamente m autovalores de A contando su multiplicidad (Teorema de Brauer).

Demostración: a) Por reducción al absurdo, supongamos que

$$\lambda \notin \bigcup_{i=1}^{n} C_i \Longrightarrow \lambda \notin C_i, \quad \forall\, i = 1, ..., n \Longrightarrow |\lambda - a_{ii}| > P_i, \quad \forall\, i = 1, ..., n$$

Ello implica que la matriz $\lambda I - A$ es estrictamente diagonalmente dominante por filas y, por tanto, regular. De modo que $\lambda \notin \text{sp}\ (A)$.

b) Es análogo

c) Consideremos la familia de matrices

$$A_t = D + tB, \quad \text{siendo} \quad D = \text{diag}\ (a_{ii}),\ B = A - D,\ t \in [0, 1]$$

Obsérvese que $A_0 = D$ y $A_1 = A$. Consideremos los conjuntos

$$C_i(t) = \{z \in \mathbb{C} : |z - a_{ii}| \leq tP_i\}, \quad i = 1, ..., n$$

Supondremos sin pérdida de generalidad que $\hat{S} = \cup_{i=1}^{m} C_i$ y denotaremos

$\hat{S}(t) = \bigcup_{i=1}^{m} C_i(t)$

Nótese que para cada $t \in [0, 1]$, $C_i(t) \subset C_i$.

Por tanto $\hat{S}(t) \subset \hat{S}$ y $\bigcup_{i=m+1}^{n} C_i(t) \subset \bigcup_{i=m+1}^{n} C_i$. De modo que $\hat{S}(t)$ y $\bigcup_{i=m+1}^{n} C_i(t)$ son también disjuntos para cada t. Por el apartado a), los autovalores de A_t se encuentran en la unión de dichos conjuntos.

Para $t = 0$ hay m autovalores en $\hat{S}$ que son $a_{11}, ..., a_{mm}$. Los autovalores son soluciones de la ecuación característica que es una ecuación polinómica cuyos coeficientes son funciones continuas de t. De modo que hay curvas continuas que emanan de esos puntos y que contienen a los autovalores de A_t para los distintos valores de t. Esas curvas no pueden salir de $\hat{S}(t)$, porque tendrían que saltar a $\cup_{i=m+1}^{n} C_i(t)$. De modo que se mantienen en $\hat{S}(t)$ para cada $t \in [0, 1]$ y por tanto en $\hat{S}$. ■

5.3. Método de la Potencia

El método de la potencia permite calcular aproximaciones sucesivas del autovalor de módulo máximo (si existe solo uno) así como de autovectores asociados a él.

Sea $A \in \mathcal{M}_n$ diagonalizable y supongamos que existe un autovalor de módulo máximo (que no tiene por qué ser simple). Denotemos

$$|\lambda| > |\lambda_2| \geq |\lambda_3| \geq ... \geq |\lambda_m| \qquad \text{con } m \leq n$$

y sean $\{v_2, v_3, ..., v_m\}$ el conjunto de autovectores asociados a $\lambda_2, \lambda_3, ..., \lambda_m$. De modo que si denotamos por $V_\lambda(A)$ al subespacio propio correspondiente al autovalor λ, se tiene $\mathbb{C}^n = V_\lambda(A) \oplus \langle v_2, ..., v_m \rangle$; esto es

$$\forall\, u \in \mathbb{C}^n,\ \exists!\ v \in V_\lambda(A),\ \exists!\ \alpha_2, ..., \alpha_m :\ u = v + \sum_{i=2}^{m} \alpha_i v_i$$

Se define el siguiente método iterativo

$$\begin{cases} u_0 \in \mathbb{C}^n \setminus \theta \quad \text{arbitrario} \\ u_{k+1} = Au_k, \quad k \geq 0 \end{cases}$$

Nótese que $u_k = A^k u_0,\ k \geq 0$. Supondremos que $u_k \neq \theta$ para todo $k \geq 0$ (en caso de que exista k_0 tal que $u_{k_0} \neq \theta$ y $u_{k_0+1} = \theta$, se tiene que $\lambda_m = 0$ y u_{k_0} es un autovector asociado).

Teorema 5.2 *Sea $A \in \mathcal{M}_n$ diagonalizable y tal que sus autovalores verifican $|\lambda| > |\lambda_2| \geq |\lambda_3| \geq ... \geq |\lambda_m|$. Sea $u_0 \notin \langle v_2, ..., v_m \rangle$ y se construye la sucesión de términos no nulos $u_{k+1} = Au_k,\ k \geq 0$. Entonces*

a) Para cualquier aplicación lineal $\phi : \mathbb{C}^n \longrightarrow \mathbb{C}$ tal que $\phi(x) \neq 0$ para cada $x \in V_\lambda(A) \setminus \{\theta\}$, se verifica

$$\lim_{k \to +\infty} \frac{\phi(u_{k+1})}{\phi(u_k)} = \lambda$$

b) Existe el límite

$$\lim_{k \to +\infty} \frac{u_k}{\lambda^k} = v$$

siendo v un autovector asociado a λ.

Nota:

Las aplicaciones lineales $\phi : \mathbb{C}^n \longrightarrow \mathbb{C}$ se pueden identificar con un vector $a \in \mathbb{C}^n$ como $\phi(u) = (u, a)$ para cada $u \in \mathbb{C}^n$. Por ejemplo, $\phi_i(u) = u_i,\ i = 1, \ldots, n$. ■

Demostración: Sea $u_0 \in \mathbb{C}^n$. Sabemos que se puede escribir $u_0 = v + \sum_{i=2}^{m} \alpha_i v_i$. Entonces,

$$u_k = A^k u_0 = A^k v + \sum_{i=2}^{m} \alpha_i A^k v_i = \lambda^k v + \sum_{i=2}^{m} \alpha_i \lambda_i^k v_i$$

Luego

$$u_k = \lambda^k \left[v + \sum_{i=2}^{m} \alpha_i \left(\frac{\lambda_i}{\lambda} \right)^k v_i \right]$$

Ya que $|\frac{\lambda_i}{\lambda}| < 1$ para $i = 2, ..., m$, se tiene que

$$\varepsilon_k = \sum_{i=2}^{m} \alpha_i \left(\frac{\lambda_i}{\lambda} \right)^k v_i \longrightarrow \theta \quad \text{si} \quad k \to +\infty$$

y, por tanto, que

$$\frac{u_k}{\lambda^k} = v + \varepsilon_k \longrightarrow v \quad \text{cuando} \quad k \to +\infty$$

Esto prueba b).

Para probar a), tomamos ϕ en la expresión de u_k,

$$\phi(u_k) = \lambda^k [\phi(v) + \phi(\varepsilon_k)]$$

por ser ϕ lineal. Para k suficientemente grande, $\phi(u_k)$ es no nula porque el primer sumando es no nulo y el segundo tiende a cero. De modo que

$$\lim_{k \to +\infty} \frac{\phi(u_{k+1})}{\phi(u_k)} = \lambda \cdot \lim_{k \to +\infty} \frac{\phi(v) + \phi(\varepsilon_{k+1})}{\phi(v) + \phi(\varepsilon_k)} = \lambda$$

■

Notas:

1) Obsérvese que para $k \geq 1$

$$\frac{\phi(u_{k+1})}{\phi(u_k)} = \lambda \cdot \frac{\phi(v) + \sum_{i=2}^{m} \alpha_i \left(\frac{\lambda_i}{\lambda} \right)^{k+1} \phi(v_i)}{\phi(v) + \sum_{i=2}^{m} \alpha_i \left(\frac{\lambda_i}{\lambda} \right)^k \phi(v_i)} = \lambda \left[1 + O\left(|\frac{\lambda_2}{\lambda}|^k \right) \right]$$

Por tanto, la convergencia es más rápida cuanto menor sea $|\frac{\lambda_2}{\lambda}|$.

2) Si la aplicación ϕ se anula sobre $V_\lambda(A)$, entonces el cociente $\dfrac{\phi(u_{k+1})}{\phi(u_k)}$ tiene también un límite que se puede calcular. En efecto, por la linealidad, se tiene

$$\phi(\varepsilon_k) = \frac{1}{\lambda^k}\sum_{i=2}^{m} \alpha_i {\lambda_i}^k \phi(v_i)$$

y, por tanto,

$$\frac{\phi(u_{k+1})}{\phi(u_k)} = \frac{\sum_{i=2}^{m} \alpha_i {\lambda_i}^{k+1}\phi(v_i)}{\sum_{i=2}^{m} \alpha_i {\lambda_i}^{k}\phi(v_i)}$$

Si, por ejemplo, el primer índice para el que $\alpha_i\phi(v_i)$ no se anula, corresponde a un autovalor de módulo estrictamente superior al de los posteriores, entonces el límite del cociente es ese autovalor.

3) No obstante lo anterior, en la práctica, los errores de redondeo hacen que sea no nula ϕ sobre $V_\lambda(A)$ y que la sucesión converja hacia λ.

4) Una elección frecuente de ϕ en la literatura es $\phi(u) = u_i$, $i = 1, ..., n$. Una vez calculados los primeros vectores u_k, se puede tomar $\phi(u) = u_i$ para i una componente donde los valores de u_k en módulo sean grandes, con lo que es poco probable que esta ϕ se anule sobre algún u_k posterior.

5) Nótese que en general λ^k tiende a 0 o no está acotado; y, por tanto, teniendo en cuenta la expresión de u_k en función de λ^k, lo mismo le pasa a las componentes de los vectores u_k que se van obteniendo. Por ello conviene normalizar los vectores u_k, de modo que el proceso queda

$$\text{Se da } u_0; \qquad v_0 = \frac{u_0}{\|u_0\|}$$

$$u_1 = Av_0 = \frac{Au_0}{\|u_0\|}; \qquad v_1 = \frac{u_1}{\|u_1\|} = \frac{Au_0}{\|Au_0\|}$$

y en general

$$u_k = \frac{A^k u_0}{\|A^{k-1}u_0\|}; \qquad v_k = \frac{u_k}{\|u_k\|} = \frac{A^k u_0}{\|A^k u_0\|}$$

De este modo es fácil comprobar

$$\lim_{k\to+\infty} \frac{\phi(u_{k+1})}{\phi(v_k)} = \lim_{k\to+\infty} \frac{\phi(A^{k+1}u_0/\|A^k u_0\|)}{\phi(A^k u_0/\|A^k u_0\|)} = \lim_{k\to+\infty} \frac{\phi(A^{k+1}u_0)}{\phi(A^k u_0)} = \lambda$$

según el Teorema 5.2.

Por su parte,

$$\lim_{k\to+\infty} v_k = \lim_{k\to+\infty} \frac{A^k u_0}{\lambda^k} \frac{\lambda^k}{\|A^k u_0\|}$$

Si $\lambda > 0$, $\lambda^k = |\lambda|^k$ y

$$\lim_{k\to+\infty} v_k = \frac{v}{\|v\|}$$

Si $\lambda < 0$, $\lambda^k = (-1)^k |\lambda|^k$ y $\{v_k\}$ tiene dos puntos de acumulación; la subsucesión de los términos pares tiende a $\frac{v}{\|v\|}$ y la de los impares tiende a su opuesto.

Si λ no es real, $\lambda^k = |\lambda|^k \exp(i\varphi)^k$ ($|\exp(i\varphi)| = 1$); en este caso, la sucesión no tiene límite.

6) Para matrices simétricas la convergencia se acelera utilizando la sucesión de los cocientes de Rayleigh de u_k. Se procede así: dado u_0 inicial, se toman

$$v_k = \frac{u_k}{\|u_k\|}; \qquad u_{k+1} = A v_k$$

$$R_A(u_k) = \frac{u_k^* A u_k}{u_k^* u_k} = \left(\frac{u_k^*}{||u_k||^2}\right) A \left(\frac{u_k}{||u_k||^2}\right) = v_k^* u_{k+1}$$

De las convergencias de las sucesiones anteriores se sigue que

$$R_A(u_k) = \lambda \left[1 + O\left(|\frac{\lambda_2}{\lambda}|^{2k}\right)\right].$$

Luego $\lim_{k\to+\infty} R_A(u_k) = \lambda$ y la velocidad de convergencia aumenta al ser $\left|\frac{\lambda_2}{\lambda}\right|^2 < \left|\frac{\lambda_2}{\lambda}\right|$.

■

5.4. Método de Givens

Es un método para aproximar valores propios de matrices tridiagonales simétricas. Consideremos la matriz tridiagonal simétrica general

$$B = \begin{pmatrix} b_1 & c_1 & 0 & \dots & 0 & 0 & 0 \\ c_1 & b_2 & c_2 & \dots & 0 & 0 & 0 \\ . & . & . & . & . & . & . \\ 0 & 0 & 0 & \dots & c_{n-2} & b_{n-1} & c_{n-1} \\ 0 & 0 & 0 & \dots & 0 & c_{n-1} & b_n \end{pmatrix}$$

Denotemos las submatrices principales de B

$$B_i = \begin{pmatrix} b_1 & c_1 & 0 & \dots & 0 & 0 & 0 \\ c_1 & b_2 & c_2 & \dots & 0 & 0 & 0 \\ . & . & . & . & . & . & . \\ 0 & 0 & 0 & \dots & c_{i-2} & b_{i-1} & c_{i-1} \\ 0 & 0 & 0 & \dots & 0 & c_{i-1} & b_i \end{pmatrix}, \quad 1 \leq i \leq n, \ (B_n = B)$$

Desarrollando $|\lambda I - B_i|$ por los elementos de la última fila, es fácil comprobar que los polinomios característicos $p_i(\lambda)$ de B_i verifican la siguiente relación de recurrencia

$$\begin{cases} p_0(\lambda) = 1 \\ p_1(\lambda) = \lambda - b_1 \\ p_i(\lambda) = (\lambda - b_i)p_{i-1}(\lambda) - c_{i-1}^2 p_{i-2}(\lambda), \quad 2 \leq i \leq n \end{cases}$$

Si para cierto j es $c_j = 0$, entonces

$$B = \begin{pmatrix} B_j & \theta \\ \theta & \hat{B} \end{pmatrix}$$

y $\det(\lambda I - B) = \det(\lambda I - B_j) \cdot \det(\lambda I - \hat{B})$. Los autovalores de B son, pues, los de B_j y de $\hat{B}$. Por tanto, podemos suponer, sin pérdida de generalidad que

$$c_i \neq 0, \quad i = 1, ..., n-1$$

Teorema 5.3 *Supongamos $c_i \neq 0 \ \forall\, i$. Los polinomios $p_i(\lambda)$ tienen las siguientes propiedades:*

1) $\lim_{\lambda \to +\infty} p_i(\lambda) = +\infty$

$$\lim_{\lambda \to -\infty} p_i(\lambda) = \begin{cases} +\infty, & \text{si } i \text{ es par} \\ -\infty, & \text{si } i \text{ es impar} \end{cases}$$

2) Si $p_i(\lambda_0) = 0$, entonces, $p_{i-1}(\lambda_0)p_{i+1}(\lambda_0) < 0$, para $1 \leq i \leq n$.

3) El polinomio $p_i(\lambda)$ tiene i raíces reales distintas que separan las $i+1$ raíces reales y distintas del polinomio $p_{i+1}(\lambda)$. Esto, para $1 \leq i \leq n-1$.

Demostración:

1) Sigue de que el término principal de $p_i(\lambda)$ es λ^i.

2) La fórmula recurrente permite escribir que

$$p_{i+1}(\lambda) = (\lambda - b_{i+1})p_i(\lambda) - c_i^2 p_{i-1}(\lambda), \quad 1\lambda ei \leq n-1$$

Si $p_i(\lambda_0) = 0$, se verificará que $p_{i+1}(\lambda_0) = -c_i^2 p_{i-1}(\lambda_0)$. Si $p_{i-1}(\lambda_0) \neq 0$, entonces ya está demostrada la afirmación (porque $c_i^2 > 0$). Si $p_{i-1}(\lambda_0) = 0$, entonces aplicando escalonadamente hacia atrás la fórmula recurrente, se obtendría que $p_i(\lambda_0) = p_{i-1}(\lambda_0) = ... = p_0(\lambda_0) = 0$ lo que es absurdo, pues $p_0(\lambda_0) = 1$.

3) Hacemos la demostración por inducción sobre i.

$p_1(\lambda) = \lambda - b_1$ tiene una única raíz real $\lambda_1^{(1)} = b_1$. Teniendo en cuenta 2), $p_0(\lambda_1^{(1)}) \cdot p_2(\lambda_1^{(1)}) < 0$. Por tanto, $p_2(\lambda_1^{(1)}) < 0$, lo que junto con el apartado 1) y el teorema de Bolzano justifica que existe dos raíces de $p_2(\lambda)$: $\lambda_1^{(2)}$ y $\lambda_2^{(2)}$ que verifican

$$\lambda_1^{(2)} > \lambda_1^{(1)} > \lambda_2^{(2)}.$$

Supongamos la propiedad cierta hasta $i = k$, es decir, el polinomio $p_i(\lambda)$ tiene i raíces reales distintas que separan las $i+1$ raíces del polinomio $p_{i+1}(\lambda)$ para $1 \leq i \leq k$. Sean $\lambda_1^{(k+1)}, \lambda_2^{(k+1)}, ..., \lambda_{k+1}^{(k+1)}$ las raíces de $p_{k+1}(\lambda)$. Se tiene

$$\lambda_1^{(k+1)} > \lambda_1^{(k)} > \lambda_2^{(k+1)} > ... > \lambda_k^{(k)} > \lambda_{k+1}^{(k+1)}.$$

Como $\lim_{\lambda \to +\infty} p_k(\lambda) = +\infty$ y $p_k(\lambda_1^{(k)}) = 0$, debe ser $p_k(\lambda_1^{(k+1)}) > 0$. Como $p_k(\lambda_2^{(k)}) = 0$, debe ser $p_k(\lambda_2^{(k+1)}) < 0$ (ya que p_k tiene raíces simples), etc.

Por otro lado, según 2), $p_k(\lambda_i^{(k+1)}) \cdot p_{k+2}(\lambda_i^{(k+1)}) < 0 \; \forall\, i$, así pues,

$$p_{k+2}(\lambda_1^{(k+1)}) < 0, \quad p_{k+2}(\lambda_2^{(k+1)}) > 0 \quad ...$$

Como $\lim_{\lambda \to +\infty} p_{k+2}(\lambda) = +\infty$, existirá una raíz $\lambda_1^{(k+2)}$ de $p_{k+2}(\lambda)$ mayor que $\lambda_1^{(k+1)}$. Análogamente, en cada intervalo $(\lambda_{i+1}^{(k+1)}, \lambda_i^{(k+1)})$, $i = 1, 2, ..., k+1$ encontraremos una raíz de $p_{k+2}(\lambda)$. Por último, si k es par, $\lim_{\lambda \to -\infty} p_{k+2}(\lambda) = +\infty$ y $p_{k+2}(\lambda_{k+1}^{(k+1)}) < 0$, si k es impar, $\lim_{\lambda \to -\infty} p_{k+2}(\lambda) = -\infty$ y $p_{k+2}(\lambda_{k+1}^{(k+1)}) > 0$. En cualquier caso, existirá una raíz $\lambda_{k+2}^{(k+2)}$ de $p_{k+2}(\lambda)$ que cumple $\lambda_{k+2}^{(k+2)} < \lambda_{k+1}^{(k+1)}$.

■

Una sucesión de polinomios que verifica las propiedades 1), 2) y 3) del Teorema 5.3 se llama sucesión de Sturm. Estas sucesiones verifican la siguiente propiedad

Teorema 5.4 *Dada una sucesión de Sturm* $\{p_i(\lambda)\}$, $i = 1, ..., n$ *y dado un número* $\mu \in \mathbb{R}$, *se denota*

$$\operatorname{Sgn} p_i(\mu) = \begin{cases} \operatorname{sgn} p_i(\mu) & si \;\; p_i(\mu) \neq 0 \\ \operatorname{sgn} p_{i-1}(\mu) & si \;\; p_i(\mu) = 0 \end{cases} \qquad i = 0, ..., n$$

y denotemos por $V(\mu)$ el número de cambios de signo en la sucesión

$$\{\operatorname{Sgn} p_0(\mu), \operatorname{Sgn} p_1(\mu), ..., \operatorname{Sgn} p_n(\mu)\}$$

Entonces, el número de raíces del polinomio $p_n(\lambda)$ en el intervalo $[a, b]$ es $V(a) - V(b)$, supuesto que $p_n(a)p_n(b) \neq 0$.

Demostración: Probaremos que para cualquier $a \in \mathbb{R}$, el número de raíces de $p_n(\lambda)$ mayores que a es $V(a)$, de donde seguirá la conclusión del Teorema. La demostración se hace por inducción sobre i. Denotaremos en este Teorema $\lambda_j^{(i)}$ para $j = 1, ..., i$, las i raíces del polinomio $p_i(\lambda)$ en orden decreciente. Verificamos la inducción.

1. Para $i = 1$,

   a) Si $a < \lambda_1^{(1)}$, la sucesión de signos es $\{+, -\}$ y por tanto $V(a) = 1$.

   b) Si $a \geq \lambda_1^{(1)}$, la sucesión de signos es $\{+, +\}$ y por tanto $V(a) = 0$.

2. Cierto para $i = k$ siendo m el número de raíces del polinomio $p_k(\lambda)$ mayores que a, que viene dado por el número de cambios de signo de la sucesión $\{\operatorname{Sgn} p_0(a), ..., \operatorname{Sgn} p_k(a)\}$.

   Sea $i = k + 1$. Resulta que

   $$\lambda_k^{(k)} < ... < \lambda_{m+1}^{(k)} \leq a < \lambda_m^{(k)} < ... < \lambda_1^{(k)}$$

   Además sabemos que

   $$\lambda_{k+1}^{(k+1)} < \lambda_k^{(k)} < ... < \lambda_{m+1}^{(k)} < \lambda_{m+1}^{(k+1)} < \lambda_m^{(k)} < \lambda_m^{(k+1)} < ... < \lambda_1^{(k)} < \lambda_1^{(k+1)}$$

   Hay que probar que el número de raíces de $p_{k+1}(\lambda)$ mayores que a es el número de cambios de signo de la sucesión $\{\operatorname{Sgn} p_0(a), ..., \operatorname{Sgn} p_k(a), \operatorname{Sgn} p_{k+1}(a)\}$. Consideraremos para ello tres casos posible

   a) $a \neq \lambda_{m+1}^{(k)}, \lambda_{m+1}^{(k+1)}$.
   Si $\lambda_{m+1}^{(k)} < a < \lambda_{m+1}^{(k+1)}$, el número de raíces de $p_{k+1}(\lambda)$ mayores que a es $m + 1$, mientras que el número de raíces de $p_k(\lambda)$ mayores que a es m. Pero $\operatorname{Sgn} p_k(a) = \operatorname{sgn} (-1)^m$ y $\operatorname{Sgn} p_{k+1}(a) = \operatorname{sgn} (-1)^{m+1}$, de donde sigue el resultado.
   Si $\lambda_{m+1}^{(k+1)} < a < \lambda_m^{(k)}$, entonces el número de raíces de $p_k(\lambda)$ y de $p_{k+1}(\lambda)$ mayores que a es m. Y efectivamente, $\operatorname{Sgn} p_k(a) = \operatorname{Sgn} p_{k+1}(a) = \operatorname{sgn} (-1)^m$.

   b) $a = \lambda_{m+1}^{(k+1)}$.
   En este caso el número de raíces de $p_{k+1}(\lambda)$ y de $p_k(\lambda)$ mayores que a es m. Y por definición, $\operatorname{Sgn} p_{k+1}(a) = \operatorname{Sgn} p_k(a)$.

c) $a = \lambda_{m+1}^{(k)}$.
En este caso el número de raíces de $p_{k+1}(\lambda)$ mayores que a es $m+1$ y el de $p_k(\lambda)$ mayores que a es m. Pero por construcción, $\operatorname{Sgn} p_k(a) = \operatorname{Sgn} p_{k-1}(a)$ y es sabido que $\operatorname{Sgn} p_{k+1}(a) \neq \operatorname{Sgn} p_{k-1}(a)$ (usando el apartado 2 del Teorema anterior).

■

Nota:

La anterior propiedad se verifica, evidentemente, para cada $p_i(\lambda)$, pero para calcular los autovalores de B hay que aplicarlo a $p_n(\lambda)$. ■

Así pues, con los resultados de la sección 5.2 pueden localizarse los autovalores (que son reales y distintos). Con ayuda de este resultado, pueden separarse en intervalos. Entonces, para aproximar los autovalores, puede aplicarse un método de dicotomía para seguir afinando los intervalos todo lo que se quiera o un método como el de Newton . Nótese que la recurrencia sirve para evaluar el polinomio característico en puntos particulares sin necesidad de obtenerlo de forma genérica y también para evaluar las derivadas necesarias para el método de Newton; ya que se verifica

$$\begin{cases} p_0'(\lambda) = 0 \\ p_1'(\lambda) = 1 \\ p_i'(\lambda) = p_{i-1}(\lambda) + (\lambda - b_i)p_{i-1}'(\lambda) - c_{i-1}^2 p_{i-2}'(\lambda), \quad 2 \leq i \leq n \end{cases}$$

Tema 6

Resolución de Sistemas de Ecuaciones no Lineales

6.1. Introducción

Sea $D \subset \mathbb{R}^n$ y sean $f, g : D \subset \mathbb{R}^n \to \mathbb{R}^n$ funciones continuas. Consideraremos en este tema sistemas (algebraicos) no lineales, que escribimos en forma homogénea o en forma de punto fijo, como:

$$(SH) \qquad f(x) = \theta \Longleftrightarrow \begin{cases} f_1(x_1, ..., x_n) = 0 \\ . \quad . \quad . \quad . \quad . \\ f_n(x_1, ..., x_n) = 0 \end{cases}$$

$$(SPF) \qquad x = g(x) \Longleftrightarrow \begin{cases} x_1 = g_1(x_1, ..., x_n) \\ . \quad . \quad . \quad . \quad . \\ x_n = g_n(x_1, ..., x_n) \end{cases}$$

Las definiciones de solución α, de método localmente convergente hacia α y de método globalmente convergente en D hacia α son análogos a los del caso escalar.

Diremos que un método iterativo tiene orden de convergencia al menos p hacia la solución α si es localmente convergente hacia α y

$$\exists\, k_0 \in \mathbb{N},\ \exists\, C > 0 : \quad \|x_{k+1} - \alpha\| \leq C\|x_k - \alpha\|^p, \quad \forall\, k \geq k_0.$$

Si $p = 1$ se exige $C < 1$. Nótese que la desigualdad anterior es independiente de la norma vectorial elegida para $p > 1$.

6.2. Método de Aproximaciones Sucesivas

Consideremos el sistema no lineal en forma de punto fijo

$$(SPF) \qquad x = g(x) \Longleftrightarrow \begin{cases} x_1 = g_1(x_1, ..., x_n) \\ . \quad . \quad . \quad . \quad . \\ x_n = g_n(x_1, ..., x_n) \end{cases}$$

para el que se define el método de aproximaciones sucesivas

$$(MAS) \qquad \begin{cases} x_0 \in \mathbb{R}^n \quad \text{dado} \\ x_{k+1} = g(x_k) \quad \forall\, k \geq 0 \end{cases}$$

Se tiene el siguiente

Teorema 6.1 (Convergencia global y estimación del error) *Sea $D \subset \mathbb{R}^n$ cerrado y $g:\ D \subset \mathbb{R}^n \to \mathbb{R}^n$ tal que*

1. *$g(D) \subset D$ (D es g-invariante)*
2. *$\exists\, L \in (0,1):\ \|g(x) - g(y)\| \leq L\|x - y\|\ \forall\, x, y \in D$ (g es L-contractiva en D)*

Entonces:

1. *Existe un único $\alpha \in D$ solución del (SPF).*
2. *El (MAS) es globalmente convergente hacia α.*
3. *Se tienen las siguientes estimaciones de error*

$$\|x_k - \alpha\| \leq \frac{L^k}{1-L}\|x_1 - x_0\| \quad \forall\, k \geq 1 \quad \textit{(a posteriori)}$$

$$\|x_{k+1} - \alpha\| \leq L\|x_k - \alpha\| \quad \forall\, k \geq 1 \quad \textit{(a priori)}$$

En particular, la convergencia es al menos lineal.

Demostración: Es evidente que el (MAS) define una sucesión $\{x_k\}_{k\geq 1} \subset D$.

Dado $k > 1$, se verifica

$$\|x_{k+1} - x_k\| = \|g(x_k) - g(x_{k-1})\| \leq L\|x_k - x_{k-1}\| \leq ... \leq L^k\|x_1 - x_0\|$$

Por tanto, dados $k, n \in \mathbb{N}$, por ejemplo $n > k$, se tiene

$$\|x_n - x_k\| \leq \|x_n - x_{n-1}\| + ... + \|x_{k+1} - x_k\| \leq$$

$$\leq (L^{n-1} + ... + L^k)\|x_1 - x_0\| \leq (L^{n-k-1} + ... + 1)L^k\|x_1 - x_0\| \leq \frac{L^k}{1-L}\|x_1 - x_0\|$$

De aquí se deduce que la sucesión $\{x_k\}$ es de Cauchy y, por tanto, convergente. De modo que

$$\exists\, \alpha \in D \text{ (por ser } D \text{ cerrado) tal que } \lim_{k\to+\infty} x_k = \alpha$$

Tomando límites en el (MAS) sigue que $\alpha = g(\alpha)$, de modo que α es solución de la ecuación. Además es la única solución posible porque si hubiera dos soluciones, α_1 y α_2,

$$\|\alpha_1 - \alpha_2\| = \|g(\alpha_1) - g(\alpha_2)\| \leq L\|\alpha_1 - \alpha_2\| < \|\alpha_1 - \alpha_2\|$$

lo que es absurdo.

Las estimaciones siguen de las anteriores desigualdades. ∎

Nótese que la condición de contractividad depende de la norma que se elija. Para asegurar que g es contractiva, suele ser útil la siguiente condición suficiente

Lema 6.1 *Sea $D \subset \mathbb{R}^n$ convexo y compacto y sea $g \in C^1(D)$ (es decir, que existe un abierto G tal que $D \subset G$ y $g \in C^1(G)$). Si*

$$\max_{x\in D} \|g'(x)\| \leq L \quad (g'(x) = (\tfrac{\partial g_i(x)}{\partial x_j})_{ij})$$

para alguna norma matricial $\|A\|$ (que es consistente con alguna norma vectorial $\|u\|$), entonces

$$\|g(x) - g(y)\| \leq L\|x - y\|, \quad \forall\, x, y \in D.$$

Nota:

En particular,

$$\|g(x) - g(y)\|_2 \leq \left(\max_{x\in D} \|g'(x)\|_S\right) \|x - y\|_2,$$

$$\|g(x) - g(y)\|_1 \leq \left(\max_{x\in D} \|g'(x)\|_C\right) \|x - y\|_1,$$

$$\|g(x) - g(y)\|_\infty \leq \left(\max_{x\in D} \|g'(x)\|_F\right) \|x - y\|_\infty.$$

∎

Demostración: Para cada $i = 1, \ldots, n$, si fijamos $x, y \in D$, gracias a la convexidad de D, podemos definir las funciones

$$h_i : [0, 1] \to \mathbb{R} \quad h_i(s) = g_i(x + s(y - x)).$$

Entonces, usando la regla de la cadena y el desarrollo hasta orden 1 de una función real con resto integral,

$$g_i(y) - g_i(x) = h_i(1) - h_i(0) = \int_0^1 h_i'(s)\, ds = \int_0^1 \nabla g_i(x + s(y-x)) \cdot (y-x)$$

donde $\nabla g_i = (\partial_{x_j} g_i)_{j=1,...,n}$ (vector gradiente). Acotando la norma de la integral por la integral de la norma y usando la consistencia de la norma matricial

$$\|g(y) - g(x)\| \leq \int_0^1 \|\nabla g(x + s(y-x))\|\, \|y - x\|\, ds$$

de donde se deduce la estimación del lema tomando maximo en $s \in [0,1]$. ■

Teorema 6.2 (Convergencia local) *Sea $D \subset \mathbb{R}^n$ y $g : D \subset \mathbb{R}^n \to \mathbb{R}^n$ tal que*

1. $\exists\, \alpha \in$ *int*$(D) : \alpha = g(\alpha)$
2. $g \in C^1(D)$ *y* $\|g'(\alpha)\| < 1$ *para alguna norma matricial consistente.*

Entonces, $\exists\, \rho > 0$ tal que el (MAS) converge hacia α, $\forall\, x_0 \in \overline{\mathrm{B}}(\alpha, \rho)$, con convergencia al menos lineal.

Demostración: Como $\|g'(\alpha)\| < 1$ y las derivadas parciales son continuas, $\exists\, \mathrm{L} \in (0,1)$, $\exists\, \rho > 0$: $\|g'(x)\| \leq L\ \forall\, x \in \overline{\mathrm{B}}(\alpha, \rho)$. De acuerdo con el lema anterior, g es contractiva en $\overline{\mathrm{B}}(\alpha, \rho)$.

Para aplicar el Teorema 6.1 en $\overline{\mathrm{B}}(\alpha, \rho)$ basta probar que $g(\overline{\mathrm{B}}) \subset \overline{\mathrm{B}}$. En efecto, si $\overline{x} \in \overline{\mathrm{B}} \Rightarrow \|\overline{x} - \alpha\| \leq \rho$; entonces

$$\|g(\overline{x}) - \alpha\| = \|g(\overline{x}) - g(\alpha)\| \leq L\|\overline{x} - \alpha\| \leq L\rho < \rho \Rightarrow g(\overline{x}) \in \overline{\mathrm{B}}(\alpha, \rho).$$

■

Corolario 6.1 (Orden cuadrático) *En las condiciones del teorema anterior, si suponemos además, que $g \in C^2(D)$ y $g'(\alpha) = 0$, entonces se tiene convergencia al menos cuadrática.*

Demostración: En efecto, si $g \in C^2(D)$ según el Teorema de Taylor podemos escribir en un entorno de α

$$g_i(x) - g_i(\alpha) = \frac{1}{2} \sum_{j,k=1}^{n} \frac{\partial^2 g_i(\xi)}{\partial x_j \partial x_k} (x_j - \alpha_j)(x_k - \alpha_k), \quad i = 1, ..., n.$$

Considerando la bola $\overline{\mathrm{B}} = \overline{\mathrm{B}}(\alpha;\rho)$ de la demostración del Teorema anterior, si definimos

$$M_{ijk} = \max_{\xi\in\overline{\mathrm{B}}} |\frac{\partial^2 g_i(\xi)}{\partial x_j \partial x_k}| < +\infty \quad \text{y} \quad M = \max_{ijk} M_{ijk},$$

se tendrá

$$|g_i(x) - g_i(\alpha)| \leq \frac{Mn^2}{2}\|x-\alpha\|_\infty^2$$

Y por tanto

$$\|x_{k+1}-\alpha\|_\infty = \|g(x_k)-g(\alpha)\|_\infty \leq \frac{Mn^2}{2}\|x_k-\alpha\|_\infty^2$$

consiguiéndose así la convergencia al menos cuadrática. ■

6.3. Método de Newton

Sea $f : D \subset \mathbb{R}^n \to \mathbb{R}^n$ derivable. Se considera el sistema no lineal homogéneo

$$(SH) \qquad f(x) = \theta$$

Por razones análogas a las del caso escalar (se trataba de buscar un esquema de segundo orden), (SH) se escribe como el sistema de punto fijo

$$(SPF) \qquad x = x - (f'(x))^{-1} f(x)$$

que será equivalente al (SH) si la matriz jacobiana $f'(x)$ es regular. El método de Newton es el (MAS) asociado al anterior (SPF):

$$(MN) \qquad \begin{cases} x_0 \in \mathbb{R}^n, \quad \text{dado} \\ x_{k+1} = x_k - (f'(x_k))^{-1} f(x_k), \quad \forall\, k \geq 0 \end{cases}$$

Desde el punto de vista algorítmico, el método consiste en, dado x_k,

1. hallar la solución, δ_k, del sistema lineal $f'(x_k)\delta_k = f(x_k)$ (de matriz $f'(x_k)$),
2. hacer $x_{k+1} = x_k - \delta_k$.

Teorema 6.3 (Convergencia local del (MN)) *Sea $D \subset \mathbb{R}^n$ y $f : D \subset \mathbb{R}^n \to \mathbb{R}^n$ tal que*

1. *$\exists\, \alpha \in$ int (D) : $f(\alpha) = 0$*
2. *$f \in C^2(D)$ y $f'(\alpha)$ es regular (es decir, det $f'(s) \neq 0$).*

Entonces, $\exists\, r > 0$ *tal que* $\forall\, x_0 \in B(\alpha, r)$*, el (MN) converge hacia* α*. Además, si* $f \in C^3(B(\alpha, r))$*, entonces la convergencia es al menos cuadrática.*

Demostración: Consideremos el método de Newton como un MAS para

$$(SPF) \qquad x = g(x) \quad \text{con} \quad g(x) = x - (f'(x))^{-1} f(x)$$

Como $f \in C^2(D)$ y $f'(\alpha)$ es regular,

$$\exists\, \rho > 0 : \quad f'(x) \quad \text{es regular} \quad \forall\, x \in B(\alpha, \rho)$$

De modo que $g \in C^1(B(\alpha, \rho))$.

Si comprobamos que $g'(\alpha) = \theta$, aplicando el Teorema 6.2 obtendremos el resultado. Denotemos

$$\delta(x) = (f'(x))^{-1} f(x) \Rightarrow g(x) = x - \delta(x)$$

Se verifica entonces

$$f'(x)\delta(x) = f(x) \Longleftrightarrow f_i(x) = \sum_{j=1}^{n} \frac{\partial f_i}{\partial x_j}(x) \cdot \delta_j(x), \quad i = 1, ..., n$$

Derivando esta expresión respecto de x_k, resulta

$$\frac{\partial f_i}{\partial x_k}(x) = \sum_{j=1}^{n} \left(\frac{\partial^2 f_i}{\partial x_k \partial x_j}(x)\delta_j(x) + \frac{\partial f_i}{\partial x_j}(x)\frac{\partial \delta_j}{\partial x_k}(x) \right).$$

Evaluando en $x = \alpha$,

$$\frac{\partial f_i}{\partial x_k}(\alpha) = \sum_{j=1}^{n} \frac{\partial f_i}{\partial x_j}(\alpha)\frac{\partial \delta_j}{\partial x_k}(\alpha), \quad i, k = 1, ..., n.$$

Escritas estas igualdades en forma matricial, resultan

$$f'(\alpha) = f'(\alpha) \cdot \delta'(\alpha) \Longrightarrow \delta'(\alpha) = Id$$

De modo que

$$g'(\alpha) = Id - \delta'(\alpha) = \theta.$$

Si $f \in \mathrm{C}^3(\mathrm{B}(\alpha, r))$, entonces $g \in \mathrm{C}^2(\mathrm{B}(\alpha, r))$ y el Corolario 6.1 nos proporciona el resultado de convergencia al menos cuadrática. ■

Desde el punto de vista numérico, el método de Newton es muy costoso porque en cada etapa hay que resolver un sistema lineal con matrices diferentes. Por ello se introduce una variante.

Variante de Whittaker

$$(MW) \qquad \begin{cases} x_0 \quad \text{dado} \\ x_{k+1} = x_k - M^{-1} f(x_k), \quad k \geq 0 \end{cases}$$

siendo M una matriz fija. Así en cada etapa hay que resolver el sistema lineal

$$M\delta_k = f(x_k)$$

lo que permite aplicar un mismo método directo (descomposición LU o Cholesky si M es definida positiva) en todas las etapas. Se puede tomar por ejemplo, $M = f'(x_0)$.

La variante de Whittaker no tiene ya convergencia cuadrática.

Una posibilidad, en el caso de converegencia muy lenta, es actualizar la matriz M con $f'(x_k)$ en algunas iteraciones.