# CÁLCULO NUMÉRICO

Departamento de Matemática Aplicada
Universidad de Salamanca

Luis Ferragut Canals

3-febrero-2008
Rev: 30-abril-2013

# Índice general

# Capítulo 1

# Resolución numérica de ecuaciones de una variable

En este tema se discute uno de los problemas básicos del cálculo numérico: Dada una función $f$ de una variable real a valores reales, hallar los valores de la variable $x$ que satisfagan la ecuación $f(x) = 0$. Éste es uno de los problemas más antiguos en aproximación numérica y sigue siendo hoy en día objeto de investigación. Los procedimientos que estudiaremos son algunos de los más clásicos, como el método de la bisección, el método de punto fijo o el método de Newton-Raphson, básicamente desarrollado por Newton hace aproximadamente 300 años y los derivados de éste que son el origen de los más recientes métodos de Quasi-Newton para resolver sistemas de ecuaciones no lineales, que se estudiarán en la segunda parte de esta asignatura.

El problema de hallar las raíces de una ecuación, $f(x) = 0$, aparece frecuentemente en el trabajo científico. Por ejemplo, en teoría de la difracción de la luz necesitamos las raíces de la equación

$$x - \tan x = 0 \tag{1.1}$$

En el cálculo de órbitas planetarias necesitamos las raíces de la ecuación de Kepler

$$x - a \sin x = b \tag{1.2}$$

para varios valores de $a$ y $b$. En teoría de la combustión

$$x = \delta e^{\gamma x} \tag{1.3}$$

para varios valores de $\gamma$ y $\delta$.

El problema general, planteado en el caso más sencillo de una función de variable real y cuya imagen sea un número real, es el siguiente:

Dada una función $f : \mathcal{R} \longrightarrow \mathcal{R}$, encontrar los valores de $x$ para los cuáles $f(x) = 0$. A continuación consideraremos varios de los procedimientos estándar para resolver este problema.

## 1.1. Método de la bisección

Se basa en la aplicación reiterada del teorema de Bolzano. Si f es una función continua definida sobre un intervalo cerrado $[a, b]$ tal que $f(a).f(b) < 0$ entonces $f$ debe tener un cero en $]a, b[$.

El método de la bisección explota esta propiedad de la siguiente manera:

(a) Hacemos $c = \frac{a+b}{2}$

(b)
- Si $f(a).f(c) < 0$, entonces $f$ tiene un cero en $]a, c[$ hacemos $b \longleftarrow c$
- Si $f(a).f(c) > 0$, entonces $f(c).f(b) < 0$ y $f$ tiene un cero en $]c, b[$ hacemos $a \longleftarrow c$
- Si $f(a).f(c) = 0$, est claro que $f(c) = 0$ y hemos encontrado un cero.

En las dos primeras situaciones del punto 2, hemos reducido el problema a la búsqueda de ceros en un intervalo de longitud la mitad que la del intervalo original.

La situación $f(c) = 0$ es poco probable que se dé en la práctica, debido a los errores de redondeo. Así el criterio para concluir no debe depender de que $f(c) = 0$, sino que permitiremos una tolerancia razonable, tal como $|f(c)| < 10^{-10}$ si trabajamos en doble precisión (Depende del ordenador).

**Pseudo-código del algoritmo de la bisección**

- input $a, b, M, \delta, \varepsilon$
- $u \longleftarrow f(a)$
- $v \longleftarrow f(b)$
- $e \longleftarrow b - a$
- output $a, b, u, v$
- if $sign(u) = sign(v)$ then STOP
- For $k = 1, ..., M$ do
  - $e \longleftarrow \frac{e}{2}$
  - $c \longleftarrow a + e$
  - $w \longleftarrow f(c)$
  - output $k, c, w, e$
  - if $|e| < \delta$ or $|w| < \varepsilon$ then STOP
  - if $sign(w) \neq sign(u)$ then

- $b \longleftarrow c$
- $v \longleftarrow w$

- else
  - $a \longleftarrow c$
  - $u \longleftarrow w$
- endif

- end

Varias de las partes de este pseudo-código necesitan explicación adicional. En primer lugar el punto medio $c$ se calcula como $c \longleftarrow a + \frac{b-a}{2}$ en lugar de $c \longleftarrow \frac{a+b}{2}$. Al hacerlo así se sigue la estrategia general de que , al efectuar cálculos numéricos, es mejor calcular una cantidad añadiendo un pequeño término de corrección a una aproximación obtenida previamente. En segundo lugar, es mejor determinar si la función cambia de signo en el intervalo recurriendo a que $sign(w) \neq sign(u)$ en lugar de utilizar $w.u < 0$ ya que ésta última requiere una multiplicación innecesaria. Por otra parte $e$ corresponde al cálculo de la cota del error que se establece más adelante.

En el algoritmo hay tres criterios que pueden detener la ejecución:

- $M$, señala el máximo número de pasos que permitirá el usuario. Un algoritmo correctamente diseñado tiene que ser finito.
- Por otra parte la ejecución del programa se puede detener, ya sea cuando el error es suficientemente pequeño o cuando lo es el valor de $f(c)$. Los parámetros $\delta$ y $\varepsilon$ controlan esta situación. Se pueden dar ejemplos en los que se satisface uno de los dos criterios sin que el otro se satisfaga.

**Teorema 1.1** : Análisis del error

Sea $f$ continua en $[a,b] = [a_0,b_0]$ con $f(a).f(b) < 0$. Sean $[a_0,b_0],[a_1,b_1],...,[a_n,b_n]$ los intervalos sucesivos generados por el método de la bisección. Entonces los límites $\lim_{n\to\infty} a_n$, $\lim_{n\to\infty} b_n$ existen, son iguales y representan un cero de $f$. Si $r = \lim_{n\to\infty} c_n$ con $c_n = \frac{a_n+b_n}{2}$, entonces

$$|r - c_n| \leq \frac{b_0 - a_0}{2^{n+1}}$$

**Demostración:**

Por la propia construcción del algoritmo, tenemos,

$$a_0 \leq a_1 \leq ... \leq b_0$$

$$b_0 \geq b_1 \geq ... \geq a_0$$

$$b_{n+1} - a_{n+1} = \frac{b_n - a_n}{2} \quad n \geq 0$$

La sucesión $\{a_n\}$ converge debido a que es creciente y está acotada superiormente.

La sucesión $\{b_n\}$ converge por ser decreciente y estar acotada inferiormente.

También tendremos

$$b_n - a_n = \frac{b_{n-1} - a_{n-1}}{2} = ... = \frac{b_0 - a_0}{2^n}$$

Así

$$\lim b_n - \lim a_n = \lim \frac{b_0 - a_0}{2^n} = 0$$

Si escribimos $r = \lim a_n = \lim b_n$, tomando límites en la desigualdad $f(a_n).f(b_n) < 0$, resulta $f(r)^2 = f(r).f(r) \leq 0$, es decir $f(r) = 0$.

Finalmente, en la etapa en la que se ha construido el intervalo $[a_n, b_n]$, si se detiene en este momento el algoritmo, sabemos que la raiz de la ecuación se encuentra en ese intervalo. La mejor estimación para esa raiz ser el punto medio $c_n = \frac{a_n+b_n}{2}$ y el error cometido verificará

$$|r - c_n| \leq \frac{b_n - a_n}{2} \leq \frac{1}{2}\frac{b_0 - a_0}{2^n} = \frac{b_0 - a_0}{2^{n+1}}$$

■

**Ejercicios**

(a) Encontrar la raiz más cercana a cero de la ecuación

$$e^x = \sin x \tag{1.4}$$

Indicación: Antes de aplicar rutinariamente el procedimiento anterior conviene hacer un análisis sencillo del problema planteado y tratar de localizar la posible solución. Por ejemplo, esbozando las gráficas de $e^x$ y de $\sin x$, resulta evidente que no existen raices positivas de $f(x) = e^x - \sin x$.

(b) Supongamos que el método de la bisección se inicia con el intervalo $[50, 63]$. Cuántos pasos deben darse para calcular una raiz con una precisión relativa de $10^{-12}$.

## 1.2. El método de punto fijo

Utilizaremos este método para resolver ecuaciones de la forma $x = f(x)$. Observemos que si queremos hallar las raíces de una ecuación $g(x) = 0$, podemos ponerla de la forma anterior, por ejemplo, haciendo $f(x) = x - g(x)$ o más generalmente $f(x) = x - \rho(x)g(x)$ donde $\rho(x) \neq 0$, es una función adecuadamente elegida, que puede ser constante o no.

De manera más precisa el problema planteado es el siguiente:

dada $f : [a, b] \longrightarrow [a, b]$ función continua, hallar $x \in [a, b]$ tal que $x = f(x)$

**Teorema 1.2** de existencia

El problema anterior tiene al menos una solución.

**Demostración:**

Si $a = f(a)$ o $b = f(b)$ entonces $a$ o $b$ es una solución. Supongamos pues que $a \neq f(a)$ y que $b \neq f(b)$.

Pongamos $g(x) = x - f(x)$, tendremos, $g(a) = a - f(a) < 0$ y $g(b) = b - f(b) > 0$. Por el teorema de Bolzano existe al menos $\bar{x} \in ]a, b[$ tal que $g(\bar{x}) = 0$, es decir, $\bar{x} = f(\bar{x})$. ■

**Teorema 1.3** de unicidad Supongamos además que se verifica la siguiente hipótesis:

Existe $k < 1$ tal que $|f(x) - f(y)| \leq k|x - y|$ Para todo $x, y \in [a, b]$, entonces el punto fijo $\bar{x}$ es único.

**Demostración:**

Sean $\bar{x_1}$ y $\bar{x_2}$ dos puntos fijos de $f$, $\bar{x_1} \neq \bar{x_2}$, es decir, $\bar{x_1}, \bar{x_2} \in [a, b]$, $\bar{x_1} = f(\bar{x_1})$ y $\bar{x_2} = f(\bar{x_2})$.

$$|\bar{x_1} - \bar{x_2}| = |f(\bar{x_1}) - f(\bar{x_2})| \leq k|\bar{x_1} - \bar{x_2}| < |\bar{x_1} - \bar{x_2}|$$

■

**Observación:** Si $f$ es diferenciable y existe un número $k < 1$ tal que $|f'(x)| < k$ para todo $x \in [a, b]$, entonces para $\xi \in [a, b]$, resulta $|f(x) - f(y)| = |f'(\xi)||x - y| \leq k|x - y|$.

**Algoritmo de punto fijo**

- $x_0 \in [a, b]$, arbitrario.
- Obtenido $x_n$, se calcula $x_{n+1} = f(x_n)$.
- Control de error y de parada.

**Pseudo-código**

- input $x_0,\ M,\ \varepsilon$
- $x \leftarrow x_0$
- For $k = 1, ..., M$ do

- $x_1 \leftarrow x$
- $x \leftarrow f(x)$
- $e \leftarrow |x - x_1|$
- output $k,\ x,\ e$
- if $e < \varepsilon$ STOP

▪ end

**Comentario:** El control del error se puede sustituir por otro, por ejemplo, controlando el error relativo $|\frac{x-x_1}{x}|$ siempre que estemos seguros que la solución buscada no esté cerca de $x = 0$.

**Teorema 1.4** de convergencia

Sea $f : [a, b] \longrightarrow [a, b]$ continua y tal que

$$|f(x) - f(y)| < k|x - y| \quad \forall x, y \in [a, b], \quad k < 1$$

entonces la sucesión $x_n$ generada por el algoritmo de punto fijo verifica

$$\lim_{n\to\infty} x_n = \bar{x}$$

siendo $\bar{x}$ el único punto fijo de $f$.

**Demostración:**

$$|x_n - \bar{x}| = |f(x_n) - f(\bar{x})| \leq |x_{n-1} - \bar{x}| \leq ... \leq k^n|x_0 - \bar{x}|$$

de donde

$$\lim_{n\to\infty} |x_n - \bar{x}| \leq |x_0 - \bar{x}| \lim_{n\to\infty} k^n = 0$$

pues $k < 1$. ■

**Teorema 1.5** : Análisis del error

$f$ con las hipótesis del teorema anterior, entonces

$$|x_n - \bar{x}| \leq \frac{k^n}{1-k}|x_1 - x_0|$$

**Demostración:**

$$|x_{n+1} - x_n| = |f(x_n) - f(x_{n-1})| \leq k|x_n - x_{n-1}| \leq ... \leq k^n|x_1 - x_0|$$

Para $m > n \geq 1$ tendremos,

$$\begin{aligned} |x_m - x_n| &= |x_m - x_{m-1} + x_{m-1} - x_{m-2} + x_{m-2} - ... + x_{n+1} - x_n| \\ &\leq |x_m - x_{m-1} + |x_{m-1} - x_{m-2}| + ... + |x_{n+1} - x_n| \\ &\leq (k^{m-1} + k^{m-2} + ... + k^n)|x_1 - x_0| \\ &\leq k^n(1 + k + ... + k^{m-n-1})|x_1 - x_0| \end{aligned}$$

Pasando al límite cuando $m \to \infty$ se obtiene

$$|x_n - \bar{x}| \leq \frac{k^n}{1-k}|x_1 - x_0|$$

■

**Observación:** El torema anterior permite demostrar la existencia de punto fijo sin hacer uso del torema de Bolzano. En efecto, la estimación

$$|x_m - x_n| \leq \frac{k^n}{1-k}|x_1 - x_0|$$

demuestra que la sucesión $x_1, x_2, ..., x_n, ...$ es de Cauchy, por lo tanto convergente; sea $\lim_{n\to\infty} x_n = \bar{x}$. Pasando al límite en la expresión $x_{n+1} = f(x_n)$, resulta $\bar{x} = f(x)$. ■

**Comentario:** Lo anterior es un ejemplo del torema más general del punto fijo de Banach en espacios métricos completos, siguiente:

Sea $f : \mathcal{M} \longrightarrow \mathcal{M}$ una aplicación de un espacio métrico completo $\mathcal{M}$ en sí mismo tal que

$$d(f(x), f(y)) \leq kd(x, y), \quad k < 1$$

Entonces $f$ tiene un único punto fijo que se puede calcular con el algoritmo anterior.

**Definición:** Orden de convergencia, noción de convergencia lineal, cuadrática y orden $\alpha$.

Supongamos que $(x_n), n = 1, ...$ es una sucesión convergente cuyo límite es $p$. Sea $e_n = x_n - p$. Si existen dos constantes $\lambda > 0$ y $\alpha > 0$ tales que

$$lim_{n\to\infty} \frac{|e_{n+1}|}{|e_n|^\alpha} = \lambda$$

diremos que $(x_n)$ converge hacia $p$, con orden $\alpha$. En particular:

- Si $\alpha = 1$, diremos que la convergencia es lineal.
- Si $\alpha = 2$, diremos que la convergencia es cuadrática.

- Si $1 < \alpha < 2$, diremos que la convergencia es superlineal.

**Orden de convergencia del método de punto fijo**

El método de punto fijo tiene convergencia lineal si $f'$ es continua y $f'(\bar{x}) \neq 0$ siendo $\bar{x}$ el punto fijo de $f$. En efecto,

$$e_{n+1} = x_{n+1} - \bar{x} = f(x_n) - f(\bar{x}) = f'(\xi_n)(x_n - \bar{x}) = f'(\xi_n)e_n$$

donde $\xi_n \in [x_n, \bar{x}]$, finalmente

$$lim_{n\to\infty}\frac{|e_{n+1}|}{|e_n|} = lim_{n\to\infty}|f'(\xi_n)| = |f'(\bar{x})| = \lambda > 0$$

**Ejercicios**

(a) Sea $f : \mathcal{R} \longrightarrow \mathcal{R}$, continua. Podemos afirmar que $f$ tiene un punto fijo ?

(b) Demostrar que las siguientes funciones son contractivas (Lipschitzianas de constante menor que 1) en los intervalos indicados y determinar el valor óptimo de la constante de Lipschitz.

- $\frac{1}{1+x^2}$ en un intervalo arbitrario.
- $\frac{1}{2}x$ sobre $1 \leq x \leq 5$
- $arctan(x)$ sobre un intervalo que excluye el 0.
- $|x|^{3/2}$ sobre $|x| \leq 1/3$

(c) Si en una calculadora se mano se introduce un número y se presiona repetidamente la tecla que corresponde al coseno. Qué número aparecerá eventualmente ? Razónalo.

(d) Sea $p$ un número positivo. Cuál es el valor de la siguiente expresión ?

$$x = \sqrt{p + \sqrt{p + ...}}$$

(e) En astronomía se conoce como ecuación de Kepler a la siguiente expresión $x = y - \varepsilon sen(y)$ con $0 \leq \varepsilon \leq 1$. Demostrar que para cada $x \in [0, \pi]$ existe una $y$ que satisface la ecuación y aplicar el algoritmo de punto fijo para resolverla.

## 1.3. El método de Newton

Consideremos de nuevo el problema de buscar las raíces de una ecuación del tipo $f(x) = 0$.

**Construcción del método de Newton**

Supongamos que $\bar{x}$ es una raiz de la ecuación anterior y supongamos además que $f$ es dos veces derivable con continuidad. Si $x$ es una aproximación de $\bar{x}$, podemos escribir,

$$0 = f(\bar{x}) = f(x) + f'(x)(x - \bar{x}) + \frac{1}{2}f''(\xi)(\bar{x} - x)^2$$

Si $x$ está cerca de $\bar{x}$, $(\bar{x} - x)^2$ es un número pequeño y podremos despreciar el último término frente a los otros, y $\bar{x}$ vendrá dado aproximadamente por

$$\bar{x} \approx x - \frac{f(x)}{f'(x)}$$

Como hemos despreciado el término cuadrático este valor no será exactamente $\bar{x}$, pero es de esperar que será una mejor aproximación que el valor $x$ de partida. De ahí el siguiente

**Algoritmo de Newton**

- $x_0$, valor cercano a $\bar{x}$.
- Calculado $x_n$, obtenemos $x_{n+1}$,

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$$

**Interpretación gráfica:** Se deja como ejercicio.

**Pseudocódigo**

(a) input $x_0, M, \delta, \varepsilon$

(b) $v \leftarrow f(x_0)$

(c) output $0, x_0, v$

(d) if $|v| < \varepsilon$ then STOP

(e) for $k = 1, ..., M$ do

- $x_1 \leftarrow x_0 - \frac{v}{f'(x_0)}$
- $v \leftarrow f(x_1)$
- output $k, x_1, v$
- if $|x_1 - x_0| < \delta$ or $|v| < \varepsilon$ then STOP
- $x_0 \leftarrow x_1$

(f) end

**Observación:** Cualquier programa que se base en el método de Newton requerirá un subprograma o procedimiento para calcular $f(x)$ y $f'(x)$.

**Ejercicios**

(a) Aplicar el método de Newton para encontrar el cero negativo de la función $f(x) = e^x - 1{,}5 - arctan(x)$.

(b) Dar un ejemplo gráfico de no convergencia del método de Newton.

**Análisis numérico del método de Newton**

Demostraremos, para una clase amplia de problemas, que el método de Newton converge cuadráticamente hacia la raiz de una ecuación no lineal con tal que la aproximación inicial sea suficientemente buena. Sin embargo puede no haber convergencia para una aproximación inicial no cercana a la solución. Necesitaremos primero un lema sobre funciones.

**Lema 1.1** Sea $I$ un intervalo abierto de la recta real y $f : I \longrightarrow \mathcal{R}$ una función verificando la propiedad,

$\exists \gamma \geq 0$ tal que $\forall x, y \in I$ verifica

$$|f'(x) - f'(y)| \leq \gamma |x - y|$$

Entonces,

$$|f(y) - f(x) - f'(x)(y - x)| \leq \frac{\gamma}{2}|y - x|^2$$

**Demostración:**

$$f(y) - f(x) = \int_x^y f'(z)dz$$

$$f(y) - f(x) - f'(x)(y - x) = \int_x^y [f'(z) - f'(x)]dz$$

haciendo el cambio de variable $z = x + t(y - x)$, $dz = (y - x)dt$, resulta

$$f(y) - f(x) - f'(x)(y - x) = \int_0^1 [f'(x + t(y - x)) - f'(x)](y - x)dt$$

$$\begin{aligned} |f(y) - f(x) - f'(x)(y - x)| &\leq |y - x| \int_0^1 \gamma |t(y - x)|dt \\ &\leq \gamma |y - x|^2 \int_0^1 tdt = \frac{\gamma}{2}|y - x|^2 \end{aligned}$$

**Observación:** El lema anterior permite estimar el resto de un desarrollo de Taylor de orden 1 sin necesidad de utilizar las derivadas segundas de $f$.

**Teorema 1.6** de convergencia local del método de Newton

Sea $f : I \longrightarrow \mathcal{R}$, $I$ intervalo abierto de $\mathcal{R}$ y sea $f'$ verificando

(a) $\exists \gamma \geq 0$ tal que $\forall x, y \in I$ verifica
$$|f'(x) - f'(y)| \leq \gamma |x - y|$$

(b) Para algún $\rho > 0$, $|f'(x)| \geq \rho$ para todo $x \in I$

(c) $f(x) = 0$ tiene una solución $\bar{x} \in I$

Entonces, existe algún $\eta > 0$ tal que si $|x_0 - \bar{x}| < \eta$, la sucesión $x_0, x_1, ...$ generada por el algoritmo de Newton
$$x_{n+1} = x_n - \frac{f(x)}{f'(x_n)}$$
existe y converge hacia $\bar{x}$. Además para $n = 0, 1, ...$
$$|x_{n+1} - \bar{x}| \leq \frac{\gamma}{2\rho} |x_n - \bar{x}|^2$$
es decir, la convergencia es cuadrática.

**Demostración:**

Sea $\tau \in ]0, 1[$, sea $\bar{\eta}$ el radio del mayor intervalo de centro $\bar{x}$ tal que esté contenido en $I$; definimos $\eta = \text{mín}\{\bar{\eta}, \tau \frac{2\rho}{\gamma}\}$. Utilizando el principio de inducción, demostraremos que para $n = 0, 1, 2, ...$ se verifica
$$|x_{n+1} - \bar{x}| \leq |x_n - \bar{x}| < \eta$$
en efecto, para $n = 0$,
$$\begin{aligned} x_1 - \bar{x} = x_0 - \bar{x} - \frac{f(x_0)}{f'(x_0)} & \\ &= x_0 - \bar{x} - \frac{f(x_0) - f(\bar{x})}{f'(x_0)} \\ &= \frac{1}{f'(x_0)} [f(\bar{x}) - f(x_0) - f'(x_0)(\bar{x} - x_0)] \end{aligned}$$

Haciendo uso del lema anterior,
$$|x_1 - \bar{x}| \leq \frac{\gamma}{2|f'(x_0)|} |x_0 - \bar{x}|^2$$
de donde, utilizando las hipótesis sobre $f'(x)$
$$|x_1 - \bar{x}| \leq \frac{\gamma}{2\rho} |x_0 - \bar{x}|^2$$

Como $|x_0 - \bar{x}| \leq \eta \leq \tau \frac{2\rho}{\gamma}$ tenemos,

$$|x_1 - \bar{x}| \leq \frac{\gamma}{2\rho}|x_0 - \bar{x}||x_0 - \bar{x}| \leq \tau|x_0 - \bar{x}| < \tau\eta < \eta$$

Por lo que $x_1$ verifica las mismas propiedades que $x_0$ y la prueba se completa por inducción. ■

**Ejercicios**

(a) Estudiar, analizando la convergencia y el orden de la misma del método de Newton como algoritmo de punto fijo, aplicado a la función $g(x) = x - \frac{f(x)}{f'(x)}$.

(b) Sea $f$ tal que $f'(\bar{x}) = 0$ y $|f''(x)| \geq \rho > 0$ en un entorno de $\bar{x}$, demostrar que el orden de convergencia del método de Newton es a lo sumo lineal.

(c) Sea $f$ una función que tiene una raiz de multiplicidad $m$ en $\bar{x}$, es decir, $f(\bar{x}) = 0, f'(\bar{x}) = 0, ..., f^{(m-1)}(\bar{x}) = 0, f^{(m)}(\bar{x}) \neq 0$.

- Demostrar que $f(x) = (x - \bar{x})^m q(x)$ donde $\bar{x}$ no es una raiz de $q(x) = 0$, es decir, $q(\bar{x}) \neq 0$.
- Considerar el método de Newton modificado siguiente:

$$x_{n+1} = x_n - \frac{mf(x_n)}{f'(x_n)}$$

y demostrar que la convergencia es cuadrática.

(d)
- Aplicar el método de punto fijo para resolver la ecuación $3ln(x) = x$. ¿ Hacia qué raiz converge eventualmente ?
- Transformar la ecuación adecuadamente para que el algoritmo de punto fijo converja hacia la raiz más próxima a cero.
- Comparar la velocidad de convergencia del método de punto fijo con la obtenida con el método de Newton.

(e) Considerar la ecuación $x = \gamma e^x$

- Determinar para que valores de $\gamma$ la ecuación no tiene solución, tiene una única solución o tiene dos soluciones.
- Analizar el método de punto fijo para resolver dicha ecuación: ¿ Hay convergencia ?, en caso afirmativo, ¿ hacia qué solución. ? ¿Cuál es el orden de convergencia? Distinguir el caso de una solución y de dos soluciones. En el caso de dos soluciones ¿Podemos obtener las dos soluciones por aplicación directa del método de punto fijo?
- Analizar el método de Newton para resolver la anterior ecuación. ¿ Hay convergencia global ? ¿ Cuál es el orden de convergencia ?

(f) Sea $f$ diferenciable, estrictamente creciente, convexa y con una raiz. Demostrar que la raiz es única y que la iteración de Newton es convergente cualquiera que sea el valor inicial elegido.

(g) Utilizar el método de Newton para calcular la raiz cuadrada de 2 y de 3 y verificar que la convergencia es cuadrática.

(h) Calcular la raiz doble de $f(x) = x^2 - 2x + 1$ que es $\bar{x} = 1$. Observar que la convergencia es solo lineal.

## 1.4. Modificaciones del método de Newton

En muchas aplicaciones, $f(x)$ no viene dada por una fórmula explícita, por ejemplo si $f(x)$ es el resultado de algún algoritmo numérico o de un proceso experimental. Como $f'(x)$ no estará en consecuencia disponible, el método de Newton deberá modificarse de modo que únicamente requiera valores de $f(x)$.

Cuando $f'(x)$ no está disponible, podemos remplazarlo por una aproximación suya, por ejemplo, tomando la pendiente de la secante formada a partir de dos puntos sobre la gráfica de la función, es decir, aproximamos la derivada en un punto $x_n$ mediante

$$f'(x_n) \approx a_n = \frac{f(x_n + h_n) - f(x_n)}{h_n}$$

El algoritmo será

- $x_0$, valor cercano a $\bar{x}$
- $x_{n+1} = x_n - \frac{f(x_n)}{a_n}$

donde $a_n = \frac{f(x_n+h_n)-f(x_n)}{h_n}$.

Seguirá funcionando el método ?

Cómo elegir adecuadamente $h_n$ en cada paso ?

**Análisis del error**

Consideremos primero el método anterior donde $a_n$ es una aproximación adecuada de $f'(a_n)$. Bajo

las mismas hipótesis que en el caso del método de Newton tendremos

$$\begin{aligned} x_1 - \bar{x} &= x_0 - \bar{x} - \frac{f(x_0)}{a_0} \\ x_1 - \bar{x} &= a_0^{-1}[f(\bar{x}) - f(x_0) - a_0(\bar{x} - x_0)] \\ &= a_0^{-1}[f(\bar{x}) - f(x_0) - f'(x_0)(\bar{x} - x_0) + (f'(x_0) - a_0)(\bar{x} - x_0)] \\ |x_1 - \bar{x}| &\leq |a_0^{-1}|[\frac{\gamma}{2}|\bar{x} - x_0|^2 + |f'(x_0) - a_0||\bar{x} - x_0|] \end{aligned}$$

Aparece un término adicional que contribuye al error, $|f'(x_0) - a_0||\bar{x} - x_0|$. Depende como elijamos $a_0, a_1, ...$ y de que sean valores suficientemente cercanos a $f'(x_0), f'(x_1), ...$ para que la convergencia no se deteriore. Veamos algunas posibilidades. Consideramos la elección $a_n = \frac{f(x_n+h_n)-f(x_n)}{h_n}$. En primer lugar veamos el siguiente lema.

**Lema 1.2**

$$|a_n - f'(x_n)| \leq \frac{\gamma|h_n|}{2}$$

**Demostración:**

$$|\frac{f(x_n + h_n) - f(x_n)}{h_n} - f'(x_n)| = |\frac{f(x_n + h_n) - f(x_n) - h_n f'(x_n)}{h_n}| \leq \frac{\gamma}{2}|h_n|$$

donde hemos utilizado el lema de la sección anterior.

**Teorema 1.7** de convergencia del método de Newton modificado

Sea $f : I \longrightarrow \mathcal{R}$, $I$ un intervalo abierto de $\mathcal{R}$ y $f'$ verificando la condición de Lipschitz siguiente:

$\exists\gamma \geq 0$ tal que $\forall x, y \in I$ se verifica

$$|f'(x) - f'(y)| \leq \gamma|x - y|$$

Supongamos que para algún $\rho > 0$, $|f'(x)| \geq \rho\ \forall x \in I$. Si $f(x) = 0$ tiene una solución $\bar{x} \in I$, entonces existen dos constantes $\eta$ y $\eta'$ tales que si $(h_n)_n > 0$ es una sucesión de números reales con $0 < |h_n| \leq \eta'$ y si $|x_0 - \bar{x}| < \eta$, entonces la sucesión $x_0, x_1, x_2, ...$ generada por el algoritmo de Newton modificado

$x_{n+1} = x_n - \frac{f(x_n)}{a_n}$, donde $a_n = \frac{f(x_n+h_n)-f(x_n)}{h_n}$ está bien definido y converge linealmente hacia $\bar{x}$.

- Si $\lim h_n = 0$ la convergencia es superlineal.
- Si existe alguna constante $c$ tal que

  $$|h_n| \leq c|x_n - \bar{x}|,$$

  la convergencia es cuadrática.

**Demostración:**

$$|x_1 - \bar{x}| \leq |a_0^{-1}|(\frac{\gamma}{2}|\bar{x} - x_0|^2 + |f'(x_0) - a_0||\bar{x} - x_0|)$$

Según el lema anterior $|f'(x_0) - a_0| \leq \frac{\gamma}{2}|h_0|$, de donde

$$|x_1 - \bar{x}| \leq |a_0^{-1}|(\frac{\gamma}{2}(|\bar{x} - x_0| + |h_0|)|\bar{x} - x_0|$$

Por otra parte como $|f'(x)| \geq \rho > 0 \; \forall x \in I$

$$|f'(x_0)| - |a_0| \leq |f'(x_0) - a_0| \leq \frac{\gamma}{2}|h_0|$$

$$|a_0| \geq |f'(x_0)| - \frac{\gamma}{2}|h_0|$$

y eligiendo $\eta'$ suficientemente pequeño de modo que siendo $|h_0| < \eta'$ resulte $\frac{\gamma}{2}|h_0| < \frac{\rho}{2}$ tendremos

$$|a_0| \geq \rho - \frac{\rho}{2} = \frac{\rho}{2}$$

es decir, $|a_0^{-1}| \leq \frac{2}{\rho}$ y finalmente

$$|x_1 - \bar{x}| \leq \frac{\gamma}{\rho}(|x_0 - \bar{x}| + |h_0|)|x_0 - \bar{x}|$$

Ahora, eligiendo $\eta$ y $\eta'$ de modo que $\frac{\gamma}{\rho}(\eta + \eta') \leq \tau < 1$

$$|x_1 - \bar{x}| \leq \tau|x_0 - \bar{x}|$$

y el razonamiento sigue inductivamente, teniendo

$$|x_{n+1} - \bar{x}| \leq \frac{\gamma}{\rho}(|x_n - \bar{x}| + |h_n|)|x_n - \bar{x}|$$

y

$$|x_{n+1} - \bar{x}| \leq \tau|x_n - \bar{x}| \leq ... \leq \tau^{n+1}|x_0 - \bar{x}|$$

de donde $lim|x_n - \bar{x}| = 0$ y la convergencia es al menos lineal.

Si $\lim_{n\to\infty} |h_n| = 0$ entonces

$$|\frac{x_{n+1} - \bar{x}}{x_n - \bar{x}}| \leq \frac{\gamma}{\rho}(|x_n - \bar{x}| + |h_n|) \to 0$$

y la convergencia es superlineal.

Si $|h_n| \leq c|x_n - \bar{x}|$

$$|x_{n+1} - \bar{x}| \leq \frac{\gamma}{\rho}(1 + c)|x_n - \bar{x}|^2$$

y la convergencia es cuadrática.

**Ejercicio**

(a) Con las mismas hipótesis que el ejercicio anterior demostrar la equivalencia de las condiciones

*a)*

$$\exists c_1 \quad |h_n| \leq c_1 |x_n - \bar{x}|$$

*b)*

$$\exists c_2 \quad |h_n| \leq c_2 |f(x_n)|$$

**Observaciones**

El análisis de la convergencia proporcionado por el teorema anterior nos proporciona información sobre como elegir $h_n$ de manera que el método sea convergente con convergencia lineal, superlineal o cuadrática.

Sin embargo la aritmética con precisión finita puede jugar un importante papel y dar lugar a ciertas limitaciones. Obviamente, en la práctica, $h_n$ no puede ser demasiado pequeño en relación a $x_n$. Por ejemplo, si $fl(x_n) \neq 0$ y $|h_n| \leq |x_n|.\varepsilon$, siendo $\varepsilon$ la precisión del ordenador, entonces $fl(x_n + h_n) = fl(x_n)$ y el numerador en $\frac{f(x_n+h_n)-f(x_n)}{h_n}$ sería nulo. Incluso si $h_n$ es suficientemente grande de modo que $fl(x_n + h_n) \neq fl(x_n)$ hay una pérdidad de precisión al restar números parecidos. Por ejemplo, supongamos que operamos con 5 dígitos significativos y tenemos $f(x_n) = 1{,}0001$ y $f(x_n + h_n) = 1{,}0010$ con $h_n = 10^{-4}$. En este caso $f(x_n + h_n) - f(x_n) = 9{,}0\ 10^{-4}$ y $a_n = 9$. Se han perdido la mayor parte de los dígitos significativos al hacer la diferencia.

## 1.5. El método de la secante

Una manera de aproximar $f'(x_n)$ es utilizar los valores de $f$ en $x_n$ y $x_{n+1}$, es decir,

$$f'(x_n) \approx a_n = \frac{f(x_n) - f(x_{n-1})}{x_n - x_{n-1}}$$

obtenemos así el llamado método de la secante,

- $x_0$, $x_1$
- $x_{n+1} = x_n - f(x_n)\frac{x_n - x_{n-1}}{f(x_n) - f(x_{n-1})}$

y que necesita de una sola evaluación de la función en cada iteración.

**Análisis del error**

Vamos a ver que si hay convergencia esta es superlineal. El método de la secante es un caso particular del método de Newton modificado al aproximar las derivadas de la función mediante el cociente incremental; el método de la secante corresponde a tomar $h_n = x_n - x_{n-1}$.

Según el análisis efectuado anteriormente

$$\begin{aligned} |x_{n+1} - \bar{x}| &\leq \frac{\gamma}{\rho}(|x_n - \bar{x}| + |x_n - x_{n-1}|)|x_n - \bar{x}| \\ &\leq \frac{\gamma}{\rho}(|x_n - \bar{x}|^2 + |x_n - x_{n-1}||x_n - \bar{x}|) \end{aligned}$$

ahora bien, $x_n - x_{n-1} = x_n - \bar{x} + \bar{x} - x_{n-1} = e_n - e_{n-1}$ de donde,

$$\begin{aligned} |e_{n+1}| &\leq \frac{\gamma}{\rho}(|e_n|^2 + |e_n - e_{n-1}||e_n|) \\ &\leq \frac{\gamma}{\rho}(|e_n|^2 + |e_n|^2 + |e_n||e_{n-1}|) \\ \frac{|e_{n+1}|}{|e_n|} &\leq \frac{\gamma}{\rho}(2|e_n| + |e_{n-1}|) \\ \lim \frac{|e_{n+1}|}{|e_n|} &= 0 \end{aligned}$$

es decir, la convergencia es superlineal.

Podemos estimar el orden preciso (valor de $\alpha$ en la definición) de la convergencia. Hemos visto que la convergencia es al menos superlineal, como

$$|e_{n+1}| \leq c_1|e_n|^2 + c_2|e_n||e_{n-1}|$$

podemos escribir,

$$\frac{|e_{n+1}|}{|e_n||e_{n-1}|} \leq c_1\frac{|e_n|}{|e_{n-1}|} + c_2$$

Como la convergencia es superlineal, para $n \to \infty$, $\frac{|e_n|}{|e_{n-1}|} \to 0$ y

$$\frac{|e_{n+1}|}{|e_n||e_{n-1}|} \lesssim c_2$$

poniendo $|e_{n+1}| \approx A|e_n|^\alpha$ con $A \neq 0$ indicará convergencia de orden $\alpha$.

$$\frac{A|e_n|^\alpha}{|e_n||e_{n-1}|} \leq c_2$$

Despejando $|e_{n-1}|$ de la relación $|e_n| \approx A|e_{n-1}|^\alpha$, resulta $|e_{n-1}| = A^{-1/\alpha}|e_n|^{1/\alpha}$, reordenando

$$A|e_n|^{\alpha-1} \approx c_2A^{-1/\alpha}|e_n|^{1/\alpha}$$

y

$$c_2^{-1}A^{1+1/|\alpha} \approx |e_n|^{\frac{1}{\alpha}-\alpha+1}$$

como el primer miembro una constante distinta de cero y en el segundo $|e_n| \to 0$, necesariamente $\frac{1}{\alpha} - \alpha + 1 = 0$, es decir, $\alpha^2 - \alpha - 1 = 0$, de donde finalmente,

$$\alpha = \frac{1+\sqrt{5}}{2} \approx 1{,}62$$

.

**Ejercicios**

(a) Escribir el pseudo-código del método de la secante.

(b) Interpretación gráfica del método de la secante.

(c) Resolver utilizando el método de Newton (con $x_0 = \pi/4.$) y utilizando el método de la secante (con $x_0 = 0{,}5$, $x_1 = \pi/4.$) la ecuación $\cos(x) - x = 0$. Estudiar la eficacia de los dos métodos, comparando el orden de convergencia y el número de evaluaciones funcionales.

# Capítulo 2

# Generalidades sobre el análisis numérico matricial

## 2.1. Los dos principales problemas del cálculo numérico matricial

Los dos principales problemas del cálculo numérico matricial son

(a) Resolución de sistemas lineales:

Hallar $x = [x_1, ..., x_n]^t \in \mathcal{R}^n$ verificando

$$\begin{array}{lll} a_{11}x_1+ & ... & +a_{1n}x_d = b_1 \\ a_{21}x_1+ & ... & +a_{2n}x_d = b_2 \\ & ... & \\ a_{n1}x_1+ & ... & +a_{nn}x_d = b_n \end{array}$$

donde

$$A = \begin{bmatrix} a_{11} & ... & a_{1n} \\ & ... & \\ a_{n1} & ... & a_{nn} \end{bmatrix}.$$

y

$$b = [b_1, ..., b_n]^t \in \mathcal{R}^n$$

o escrito en forma matricial: Dada $A \in \mathcal{R}^{n\times n}$ y $b \in \mathcal{R}^n$, hallar $x \in \mathcal{R}^n$ que verifique

$$Ax = b$$

(b) Cálculo de valores y vectores propios de una matriz:

Dada $A \in \mathcal{R}^{n\times n}$, hallar los pares $(\lambda, v) \in \mathcal{R} \times \mathcal{R}^n$ que verifiquen

$$Av = \lambda v$$

.

## 2.2. Repaso de conceptos y resultados del Álgebra Lineal

### 2.2.1. Notaciones y primeras definiciones

Consideramos $V$ un espacio vectorial de dimensión finita sobre un cuerpo $\mathcal{K}$ (en la práctica $\mathcal{R}$ o $\mathcal{C}$).

Sea $[e_1, ..., e_n]$ una base de $V$.

Todo vector $v \in V$ se puede expresar de la forma $v = \sum v_i e_i$.

Cuando la base $[e_1, ..., e_n]$ está definida sin ambiguedad podemos identificar $V$ con $\mathcal{K}^n$.

$v$ se escribe con notación matricial

$$v = (v_1, ..., v_n)^t = \begin{bmatrix} v_1 \\ . \\ . \\ . \\ v_n \end{bmatrix}.$$

Designaremos mediante $v^t = [v_1, ..., v_n]$ al vector transpuesto de $v$ y mediante $v^* = [\bar{v_1}, ..., \bar{v_n}]$ al vector adjunto, donde $\bar{\alpha}$ designa el número complejo conjugado de $\alpha$.

La aplicación $(.,.) : V \times V \longrightarrow \mathcal{K}$ definida por

$$(u, v) = v^t u = u^t v = \sum_{i=1}^{d} u_i v_i \quad \text{si} \quad \mathcal{K} = \mathcal{R}$$

$$(u, v) = v^* u = \overline{u^* v} = \sum_{i=1}^{d} u_i \bar{v}_i \quad \text{si} \quad \mathcal{K} = \mathcal{C}$$

se llama producto escalar euclídeo si $\mathcal{K} = \mathcal{R}$ y producto escalar hermítico si $\mathcal{K} = \mathcal{C}$.

Dos vectores son ortogonales si $(u, v) = 0$.

Un conjunto de vectores $\{v_1, ..., v_k\}$ son ortonormales si $(v_i, v_j) = \delta_{ij}$.

Sean $V$ y $W$ dos espacios vectoriales sobre el mismo cuerpo $\mathcal{K}$, con bases respectivas $[e_j]_{j=1}^n$ y $[w_i]_{j=1}^m$. Sea $\mathcal{A} : V \longrightarrow W$ una aplicación lineal. Referida a estas bases la aplicación lineal se representa por una matriz de $m$ filas por $n$ columnas $A$:

$$A = \begin{bmatrix} a_{11} & ... & a_{1n} \\ & ... & \\ a_{m1} & ... & a_{mn} \end{bmatrix}.$$

Los elementos $a_{ij}$ de la matriz están definidos de forma única por las relaciones

$$Ae_j = \sum_{i=1}^{m} a_{ij} w_i \quad j = 1, ..., n$$

Dicho de otra manera, la $j-$ésima columna $(a_{1j}, ..., a_{mj})^t$ de la matriz $A$ representa el vector $\mathcal{A}e_j$ en la base $[w_i]_{i=1}^m$.

Una matriz de $m$ filas por $n$ columnas se designa del tipo $(m, n)$ y se denota $\mathcal{A}_{m,n}(\mathcal{K})$ o simplemente $\mathcal{A}_{m,n}$ el espacio vectorial sobre el cuerpo $\mathcal{K}$ formado por las matrices $(m, n)$ de elementos en $\mathcal{K}$.

Dada una matriz $A = (a_{ij})$ $i = 1, ..., m$ $j = 1, ..., n$, $A \in \mathcal{A}_{m,n}(\mathcal{C}$ se define la matriz adjunta por las relaciones

$$(Au, v)_m = (u, A^* v)_n \quad \forall u \in \mathcal{C}^n \ \forall v \in \mathcal{C}^m$$

Estas relaciones implican $(A^*)_{ij} = \bar{a_{ji}}$. análogamente, si $A \in \mathcal{A}_{m,n}(\mathcal{R})$ se define $A^t \in \mathcal{A}_{m,n}(\mathcal{R})$ por las relaciones

$$(Au, v)_m = (u, A^t v)_n \quad \forall u \in \mathcal{R}^n \ \forall v \in \mathcal{R}^m$$

y estas relaciones implican $(A^t)_{ij} = a_{ji}$.

A la composición de aplicaciones $\mathcal{B} \circ \mathcal{A}$ lineales le corresponde el producto de matrices $BA$. Se tienen las relaciones

$$(AB)^t = B^t A^t$$
$$(AB)^* = B^* A^*$$

Cuando el número de filas coincide con el número de columnas las matrices se llaman cuadradas. El conjunto $\mathcal{A}_n = \mathcal{A}_{n,n}$ es el anillo de matrices cuadradas. Para $A \in \mathcal{A}_n$, $A = (a_{ij})$ los elementos $a_{ii}$ se llaman elementos diagonales. La matriz $I = (\delta_{ij})$ se llama matriz unidad. Una matriz es invertible o regular, si existe una matriz (única si existe) denotada $A^{-1}$, llamada inversa de $A$ tal que $AA^{-1} = A^{-1}A = I$. En caso contrario se dice que $A$ es singular. Si $A$ y $B$ son inversibles se tiene:

$$(AB)^{-1} = B^{-1}A^{-1}$$

$$(A^t)^{-1} = (A^{-1})^t$$
$$(A^*)^{-1} = (A^{-1})^*$$

**Definiciones** Sea $A \in \mathcal{A}_n$, $A$ es:

- Simétrica, si $A$ es real y $A = A^t$.
- Hermítica o autoadjunta, si $A = A^*$.
- Ortogonal, si $A$ es real y $AA^t = A^tA = I$, es decir, $A^t = A^{-1}$.
- Unitaria, si $AA^* = A^*A = I$, es decir, $A^* = A^{-1}$
- Normal, si $AA^* = A^*A$

Una matriz $A$ es diagonal si sus términos $a_{ij} = 0 \; \forall i \neq j$. Escribiremos $A = diag(a_{ii}) = diag(a_{11}, ..., a_{nn})$.

Se define la traza de $A$, $tr(A) = \sum_{i=1}^{n} a_{ii}$ y determinante de $A$, $det(A) = \sum_{\sigma} sgn(\sigma) a_{\sigma(1),1} ... a_{\sigma(n),n}$.

### 2.2.2. Valores propios

Los valores propios $\lambda_i = \lambda_i(A), \quad 1 \leq i \leq n$ de una matriz $A$ de orden $n$ son las $n$ raices reales o complejas, distintas o no, del polinomio característico:

$$p_A : \lambda \in \mathcal{C} \longrightarrow p_A(\lambda) = det(A - \lambda I)$$

El espectro $sp(A)$ de la matriz $A$ es el subconjunto $sp(A) = \bigcup_{i=1}^{n} \{\lambda_i(A)\}$ del plano complejo. Recordemos las relaciones

- $tr(A) = \sum_{i=1}^{n} \lambda_i(A)$
- $det(A) = \Pi_{i=1}^{n} \lambda_i(A)$
- $tr(AB) = tr(BA)$
- $tr(A + B) = tr(A) + tr(B)$
- $det(AB) = det(BA) = det(A)det(B)$

Las dos primeras se pueden obtener como consecuencia del lema de Schur que veremos más adelante.

**Definición 2.1** El radio espectral de una matriz $A$ es el número mayor o igual que cero dado por

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

A todo valor propio de $A$ se le asocia al menos un vector $p$ tal que $Ap = \lambda p$ llamado vector propio correspondiente valor propio $\lambda$. Si $\lambda \in sp(A)$ el subespacio $\{v \in V; \; Av = \lambda v\}$ de dimensión al menos 1, se llama subespacio propio correspondiente al valor propio $\lambda$.

### 2.2.3. Reducción de matrices

Sea $V$ en espacio vectorial de dimensión finita $n$. $\mathcal{A} : V \longrightarrow V$ una aplicación lineal representada por una matriz $A$ relativa a una base $[e_i]_{i=1,...,n}$. Relativa a otra base $[w_i]_{i=1,...,n}$ la misma aplicación está representada por la matriz $B = P^{-1}AP$, donde $P$ es la matriz regular cuyo $j-$ésimo vector columna está formado por las componentes de $[w_j]$ en la base $[e_i]$. La matriz $P$ se llama matriz de paso de la base $[e_i]$ a la base $[w_i]$.

La cuestión que se plantea es ahora la de hallar una base $[w_i]$ en la que la matriz $B$ sea lo más sencilla posible, por ejemplo diagonal. S existe tal posibilidad se dice que $A$ es diagonalizable, en cuyo caso los elementos diagonales de $P^{-1}AP$ son los valores propios $\lambda_1, ..., \lambda_n$ de $A$ y la $j-$ésima columna de P son las componentes de un vector propio asociado a $\lambda_j$, en efecto, tenemos

$$P^{-1}AP = \Lambda = diag(\lambda_i) \Leftrightarrow AP = P\Lambda \Leftrightarrow Ap_j = \lambda_j p_j \; 1 \leq j \leq n$$

es decir, el par $(p_j, \lambda_j)$ es la solución de

$$(A - \lambda_j I)p_j = 0$$

como es un sistema homogéneo con solución no trivial $\lambda_j$ es solución de

$$det(A - \lambda I) = 0$$

y $\lambda_j$ es un valor propio según la definición dada anteriormente.

**Definición 2.2** Dos matrices $A$ y $B$ relacionadas de la forma

$$B = P^{-1}AP$$

se dice que son semejantes.

**Propiedades de las matrices semejantes**

- $det(A) = det(B)$
- $tr(A) = tr(B)$
- $sp(A) = sp(B)$

En efecto, la primera es consecuencia de

$$det(B) = det(P^{-1})det(A)det(P) = det(P)^{-1}det(A)det(P) = det(A)$$

La segunda, se obtiene mediante cálculo

$$tr(B) = \sum b_{ii} = \sum_i \sum_{kl} p_{ik}^{-1} a_{kl} p_{li} = \sum_{kl} (a_{kl} \sum_i p_{li} p_{ik}^{-1}) = \sum_{kl} a_{kl}\delta_{kl} = \sum a_{kk} = tr(A)$$

finalmente la tercera,

$$det(B-\lambda I) = det(P^{-1}AP - \lambda P^{-1}P) = det(P^{-1}(A-\lambda I)P) = det(A-\lambda I)$$

**Observación:** Los elementos de la diagonal de una matriz triangular son sus valores propios. En efecto, si

$$B = \begin{bmatrix} b_{11} & b_{12} & ... & b_{1n} \\ 0 & b_{22} & ... & b_{2n} \\ & ... & & \\ 0 & 0 & ... & b_{nn} \end{bmatrix}$$

entonces,

$$det(B-\lambda I) = 0 \Leftrightarrow \Pi_{i=1}^{n}(b_{ii}-\lambda) = 0 \Rightarrow \lambda_i = b_{ii}$$

El siguiente lema es fundamental en el estudio de los valores propios y vectores propios de una matriz.

**Lema de Schur:** Dada una matriz cuadrada $A$, existe una matriz unitaria $U$ tal que $U^{-1}AU$ es triangular. Resulta entonces que

$$U^{-1}AU = \begin{bmatrix} \lambda_1 & b_{12} & ... & b_{1n} \\ 0 & \lambda_2 & ... & b_{2n} \\ & ... & & \\ 0 & 0 & ... & \lambda_n \end{bmatrix}$$

donde $\lambda_1, ..., \lambda_n$ son los valores propios de $A$.

**Demostración:** Procedemos por inducción. El teorema es cierto para $n = 1$ de manera trivial. Suponiendo entonces que es cierto para matrices de orden $n-1$, probaremos que es cierto para matrices de orden $n$.

Sea $u_1, \lambda_1$ un vector propio y el correspondiente valor propio de $A$, donde $u_1$ está normalizado, es decir, $(u_1, u_1) = 1$. Elijamos una base $[u_1, ..., u_n]$ de $\mathcal{C}^n$ ortonormal (que podemos construir, por ejemplo utilizando el método de Gram-Schmidt). Entonces $U = [u_1, ..., u_n]$ es una matriz unitaria y tendremos

$$U^*AU = U^*[Au_1, ..., Au_n] = \begin{bmatrix} u_1^* \\ . \\ . \\ . \\ u_n^* \end{bmatrix} . [\lambda_1 u_1, Au_2, ..., Au_n] = \begin{bmatrix} \lambda_1 & c_{12} & ... & c_{1n} \\ 0 & & & \\ ... & & A_1 & \\ 0 & & & \end{bmatrix}$$

donde $c_{1j} = u_1^* A u_j \quad j = 2, ..., n$ y $A_1$ es de orden $n-1$.

Sea $U_1$ una matriz unitaria de orden $n-1$ tal que $U_1^* A_1 U_1$ es triangular (hipótesis de inducción) y designamos

$$U_2 = \begin{bmatrix} 1 & 0 & ... & 0 \\ 0 & & & \\ ... & & U_1 & \\ 0 & & & \end{bmatrix}$$

Entonces $U_2$ y $UU_2$ son unitarias y obtenemos

$$(UU_2)^* A (UU_2) = U_2^* (U^* A U) U_2 = \begin{bmatrix} 1 & 0 & ... & 0 \\ 0 & & & \\ ... & & U_1^* & \\ 0 & & & \end{bmatrix} \cdot \begin{bmatrix} \lambda_1 & c_{12} & ... & c_{1n} \\ 0 & & & \\ ... & & A_1 & \\ 0 & & & \end{bmatrix} \cdot \begin{bmatrix} 1 & 0 & ... & 0 \\ 0 & & & \\ ... & & U_1 & \\ 0 & & & \end{bmatrix} =$$

$$\begin{bmatrix} \lambda_1 & b_{12} & & ... & b_{1n} \\ 0 & & & & \\ ... & & U_1^* A_1 U_1 & & \\ 0 & & & & \end{bmatrix}$$

donde

$$[\lambda_1, b_{12}, ..., b_{1n}] = [\lambda_1, c_{12}, ..., c_{1n}] U_2$$

resulta entonces

$$U^* A U = \begin{bmatrix} \lambda_1 & b_{12} & ... & b_{1n} \\ 0 & \lambda_2 & ... & b_{2n} \\ & ... & & \\ 0 & 0 & ... & \lambda_n \end{bmatrix}$$

donde $\lambda_1, ..., \lambda_n$ son los valores propios de $A$. ■

El siguiente corolario es fundamental importancia en muchas aplicaciones. Muestra que todos los valores propios de una matriz autoadjunta son reales.

**Corolario 2.1** Si una matriz es autoadjunta, es decir, $A = A^*$, es diagonalizable y sus valores propios son reales. Análogamente si una matriz es simétrica, es diagonalizable y sus valores propios son reales.

**Demostración:** Por el lema de Schur, existe una matriz unitaria $U$ tal que

$$U^* A U = \begin{bmatrix} \lambda_1 & b_{12} & ... & b_{1n} \\ 0 & \lambda_2 & ... & b_{2n} \\ & ... & & \\ 0 & 0 & ... & \lambda_n \end{bmatrix}$$

Como $A$ es autoadjunta

$$\begin{bmatrix} \bar{\lambda_1} & 0 & ... & 0 \\ \bar{b_{12}} & \bar{\lambda_2} & ... & 0 \\ & ... & ... & \\ \bar{b_{1n}} & ... & ... & \bar{\lambda_n} \end{bmatrix} = (U^*AU)^* = U^*A^*U = U^*AU = \begin{bmatrix} \lambda_1 & b_{12} & ... & b_{1n} \\ 0 & \lambda_2 & ... & b_{2n} \\ & & ... & \\ 0 & 0 & ... & \lambda_n \end{bmatrix}$$

por lo tanto $\lambda_i = \bar{\lambda_i}$ y $b_{ij} = 0$ $i \neq j$, es decir, $U^*AU = diag(\lambda_i)$. La demostración para matrices simétricas es análoga. ■

El siguiente corolario caracteriza a las matrices diagonalizables.

**Corolario 2.2** *$A$ es normal si y solo si es diagonalizable y tiene una base de vectores propios ortonormales.*

**Demostración:** Sea $A$ una matriz normal y $U$ una matriz unitaria tales que $T = U^*AU$ es triangular superior. Veamos que $T$ es también normal, en efecto,

$$T^*T = (U^*AU)^*(U^*AU) = U^*A^*UU^*AU = U^*A^*AU$$

$$TT^* = (U^*AU)(U^*AU)^* = U^*AUU^*A^*U = U^*AA^*U$$

como $A^*A = AA^*$ se deduce que $T^*T = TT^*$. Escribamos,

$$T = \begin{bmatrix} t_{11} & t_{12} & ... & t_{1n} \\ 0 & t_{22} & ... & t_{2n} \\ & & ... & \\ 0 & 0 & ... & t_{nn} \end{bmatrix}$$

tenemos,

$$T^*T = \begin{bmatrix} \bar{t_{11}} & 0 & ... & 0 \\ \bar{t_{12}} & \bar{t_{22}} & ... & 0 \\ & ... & ... & \\ \bar{t_{1n}} & ... & ... & \bar{t_{nn}} \end{bmatrix} \cdot \begin{bmatrix} t_{11} & t_{12} & ... & t_{1n} \\ 0 & t_{22} & ... & t_{2n} \\ & & ... & \\ 0 & 0 & ... & t_{nn} \end{bmatrix}$$

$$TT^* = \begin{bmatrix} t_{11} & t_{12} & ... & t_{1n} \\ 0 & t_{22} & ... & t_{2n} \\ & ... & & \\ 0 & 0 & ... & t_{nn} \end{bmatrix} \cdot \begin{bmatrix} \bar{t_{11}} & 0 & ... & 0 \\ \bar{t_{12}} & \bar{t_{22}} & ... & 0 \\ & ... & ... & \\ \bar{t_{1n}} & ... & ... & \bar{t_{nn}} \end{bmatrix}$$

El término $(T^*T)_{11}$ es igual a $|t_{11}|^2$. Por otra parte el término $(TT^*)_{11}$ es igual a $|t_{11}|^2 + |t_{12}|^2 + ... + |t_{1n}|^2$ lo que implica necesariamente que $t_{12} = t_{13} = ... = t_{1n} = 0$. Análogamente demostramos que $t_{23} = t_{24} = ... = t_{2n} = 0$, etc. Por lo tanto $T$ es diagonal y $A$ es diagonalizable.

Recíprocamente, si $A$ es tal que $U^*AU$ es diagonal con $U$ unitaria, entonces $A$ es normal. En efecto, $D = U^*AU$ y $UDU^* = A$ de donde

$$AA^* = (UDU^*)(UDU^*)^* = UDU^*UD^*U^* = UDD^*U^*$$

$$A^*A = (UDU^*)^*(UDU^*) = UD^*U^*UDU^* = UD^*DU^*$$

y por tanto $AA^* = A^*A$ pues $DD^* = D^*D$. ■

## 2.3. Normas vectoriales y normas matriciales

### 2.3.1. Definiciones y ejemplos

**Definición: Norma de un vector**

Sea $v \in \mathcal{K}^d$, $(\mathcal{K} = \mathcal{R}, \mathcal{C})$ se llama norma de un vector a una aplicación

$$\begin{aligned} ||.|| : \quad \mathcal{K}^d &\longrightarrow \mathcal{R} \\ v &\longrightarrow ||v|| \end{aligned} \tag{2.1}$$

verificando las propiedades,

(a) $||v|| \geq 0 \quad \forall\, v \in \mathcal{R}^d$ y $||v|| = 0$ solo si $v = 0$.

(b) $||\alpha v|| = |\alpha| ||v|| \quad \forall \alpha \in \mathcal{R} \quad \forall v \in \mathcal{R}^d$.

(c) $||u + v|| \leq ||u|| + ||v|| \quad \forall u, v \in \mathcal{R}^d$.

**Ejemplos:**

(a) Norma $l_2$: $||v||_2 = (\sum v_i^2)^{1/2}$

(b) Norma $l_1$: $||v||_1 = \sum |v_i|$

(c) Norma $l_\infty$: $||v||_\infty = \text{máx}\, |v_i|$

(d) Norma $l_p$, $p \geq 1$: $||v||_p = (\sum |v_i|^p)^{1/p}$

**Ejercicio:** Verificar que $l_1$, $l_2$ y $l_\infty$ son efectivamente una norma en $\mathcal{K}^d$

Se llaman **Normas euclídeas** a las que derivan de un producto escalar, es decir, $||v|| = (v, v)^{1/2}$, donde $(.,.)$ es una aplicación

$$\begin{aligned} (.,.) : \quad \mathcal{K}^d \times \mathcal{K}^d &\longrightarrow \mathcal{K} \\ u, v &\longrightarrow (u, v) \end{aligned}$$

verificando las propiedades siguientes:

(a) $(u, v) = \overline{(v, u)} \quad \forall u, v \in \mathcal{K}^d$, si $\mathcal{K} = \mathcal{C}$

$(u, v) = (v, u) \quad \forall u, v \in \mathcal{K}^d$, si $\mathcal{K} = \mathcal{R}$

(b) $(\alpha u, v) = \alpha(u, v) \quad \forall \alpha \in \mathcal{K}, \forall u, v \in \mathcal{K}^d$

(c) $(u + v, w) = (u, v) + (u, w) \quad \forall u, v, w \in \mathcal{K}^d$

(d) $(v, v) \geq 0 \quad \forall v \in \mathcal{K}^d$; $(v, v) = 0$ solo si $u = 0$.

**Ejemplo:** $(u, v) = \sum u_i v_i$ en $\mathcal{R}^d$

$(u, v) = \sum u_i \bar{v_i}$ en $\mathcal{C}^d$

La norma $l_2$ es la norma asociada a este producto escalar.

**Convergencia de una sucesión de vectores en $\mathcal{R}^d$ o $\mathcal{C}^d$.**

Una sucesión $(v^k)_k$ de vectores converge hacia un vector $v$ si y solo si $\lim_{k\to\infty} ||v^k - v|| = 0$.

**Teorema 2.1** Las siguientes afirmaciones son equivalentes,

(a) La sucesión $(v^k)_k$ de vectores converge hacia un vector $v$ en una norma determinada.

(b) La sucesión $(v^k)_k$ de vectores converge hacia un vector $v$ en cualquier norma.

(c) Las componentes de los vectores de la sucesión $(v^k)_k$ convergen hacia la correspondiente componente del vector $v$.

**Definición: Norma matricial**

Sea $\mathcal{A}_d$ el anillo de matrices de orden $d \times d$ de términos reales (o complejos). Una norma matricial es una aplicación

$$\begin{aligned} ||.|| : \quad \mathcal{A}_d &\longrightarrow \mathcal{R} \\ A &\longrightarrow ||A|| \end{aligned}$$

verificando las siguientes propiedades:

(a) $||A|| \geq 0 \; \forall A \in \mathcal{A}_d$ y $||A|| = 0$ solo si $A = 0$

(b) $||\alpha A|| = |\alpha| ||A|| \; \forall \alpha \in \mathcal{R}$ (o $\mathcal{C}$) $\forall A \in \mathcal{A}_d$

(c) $||A + B|| \leq ||A|| + ||B|| \; \forall A, B \in \mathcal{A}_d$

(d) $||AB|| \leq ||A||||B|| \; \forall A, B \in \mathcal{A}_d$

**Normas matriciales subordinadas:** Toda norma vectorial induce la correspondiente norma matricial mediante:

$$||A|| = \max_{v \neq 0} \frac{||Av||}{||v||} = \max_{||v||=1} ||Av||$$

**Ejercicios**

(a) Verificar la anterior igualdad.

(b) Una norma subordinada verifica $||Av|| \leq ||A||||v||$

(c) Una norma subordinada es una norma matricial.

(d) Demostrar que si $A \in \mathcal{A}_n$, entonces $||A||_\infty = \max_{||x||_\infty=1} ||A|| = \max_{1 \leq i \leq n} \sum_{j=1,\ldots,n} |a_{ij}|$.

(e) Demostrar que si $A \in \mathcal{A}_n$, entonces $||A||_1 = \max_{||x||_\infty=1} ||Ax|| = \max_{1 \leq j \leq n} \sum_{i=1,\ldots,n} |a_{ij}|$.

### 2.3.2. Caracterización de valores propios y convergencia de sucesiones de matrices

**Definición: Radio espectral de una matriz**

El radio espectral de una matriz $A \in \mathcal{A}_n$ es el número

$$\rho(A) = \max_{\lambda \in Sp(A)} |\lambda|$$

es decir, el máximo valor absoluto del conjunto de valores propios.

**Teorema 2.2** : Caracterización del mínimo y máximo valor propio

Hemos visto como consecuencia del lema de Schur que si $A$ es una matriz simétrica (o autoadjunta) es diagonalizable y sus valores propios son reales, que podremos ordenarlos de la forma $\lambda_1 \leq, ..., \leq \lambda_n$. Entonces, $\lambda_1$ y $\lambda_n$ están caracterizados por

$$\lambda_1 = \min_{v \neq 0} \frac{(Av, v)}{(v, v)}$$

$$\lambda_n = \max_{v \neq 0} \frac{(Av, v)}{(v, v)}$$

**Nota:** El cociente $R_A(v) = \frac{(Av,v)}{(v,v)}$ se llama cociente de Rayleigh asociado al vector $v$.

**Demostración:**

Primero observar que $\max_{v\neq 0} R_A(v) = \max_{||v||_2=1}(Av, v)$ y $\min_{v\neq 0} R_A(v) = \min_{||v||_2=1}(Av, v)$.

Sea $[u_1, u_2, ..., u_n]$ una base de vectores propios ortonormales entre sí.

Para $v$ de norma unidad, tenemos, $v = \sum_{i=1}^{n} \alpha_i u_i$ con $||v||_2^2 = (v, v) = \sum_{i=1}^{n} |\alpha_i|^2 = 1$.

$$R_A(v) = \frac{(Av, v)}{(v, v)} = (Av, v) = \sum_{i=1}^{n} \lambda_i |\alpha_i|^2$$

Mayorando $\lambda_i$ por $\lambda_d$ y minorando por $\lambda_1$ resulta,

$$\lambda_1 = \lambda_1 \sum_{i=1}^{n} |\alpha_i|^2 = R_A(v) \leq \sum_{i=1}^{n} \lambda_n |\alpha_i|^2 = \lambda_n \sum_{i=}^{n} |\alpha_i|^2 = \lambda_n$$

Por otra parte, tomando $v = u_1$

$$R_A(u_1) = \frac{(Au_1, u_1)}{(u_1, u_1)} = \lambda_1$$

y tomando $v = u_n$

$$R_A(u_1) = \frac{(Au_n, u_n)}{(u_n, u_n)} = \lambda_n$$

termina la demostración. ■

**Observación:** Para un par vector propio, valor propio $(u_i, \lambda_i)$ se tiene

$$R_A(u_i) = \lambda_i$$

**Propiedad:** Si una matriz simétrica es definida positiva (resp. semidefinida positiva), es decir,

$$(Av, v) > 0 \quad \forall v \neq 0$$

(resp.$(Av, v) \geq 0 \quad \forall v \neq 0$) sus valores propios son positivos (resp. mayores o igual a cero), en efecto

$$\lambda_i = \frac{(Au_i, u_i)}{(u_i, u_i)} > 0$$

(resp. $\lambda_i = \frac{(Au_i, u_i)}{(u_i, u_i)} \geq 0$).

**Teorema 2.3** Para toda matriz $A$

$$||A||_2 = (\rho(A^*A))^{1/2}$$

**Demostración:** Primero observemos que $A^*A$ es hermítica y semidefinida positiva.

$$||A||_2^2 = \max_{||v||_2=1} ||Av||_2^2 = \max_{||v||_2=1} (Av, Av) = \max_{||v||_2=1} (A^*Av, v) = \max_{v\neq 0} R_{A^*A}(v) = \lambda_n = |\lambda_n| = \rho(A^*A)$$

donde $\lambda_n$ es el máximo valor propio de $A^*A$ que es mayor o igual que cero, pues $A^*A$ es una matriz hermítica semidefinida positiva.

**Teorema 2.4** Para toda matriz $A$ real

$$||A||_2 = (\rho(A^t A))^{1/2}$$

**Demostración:** Primero observemos que $A^t A$ es simétrica y semidefinida positiva.

$$||A||_2^2 = \max_{||v||_2=1} ||Av||_2^2 = \max_{||v||_2=1} (Av, Av) = \max_{||v||_2=1} (A^t Av, v) = \max_{v\neq 0} R_{A^t A}(v) = \lambda_n = |\lambda_n| = \rho(A^t A)$$

donde $\lambda_n$ es el máximo valor propio de $A^t A$ que es mayor o igual que cero, pues $A^t A$ es una matriz simétrica semidefinida positiva.

**Corolario 2.3** Si $A$ es una matriz simétrica $||A||_2 = \rho(A)$.

**Demostración:** $A$ es simétrica si y solo si $A = A^t$. Si $\lambda$ es un valor propio de $A$, $\lambda^2$ es un valor propio de $A^2$, lo que implica $\rho(A)^2 = (\rho(A))^2$, de donde

$$(\rho(A))^2 = \rho(A^2) = \rho(A^t A) = ||A||_2^2$$

y finalmente $||A||_2 = \rho(A)$.■

**ejercicios**

(a) Para toda matriz cuadrada $A$ demostrar

$$\rho(A^* A) = \rho(AA^*)$$

y concluir $||A||_2 = ||A^*||_2$ y que si $Q$ es una matriz unitaria entonces $||Q||_2 = 1$.

(b) Demostrar la invariancia de la norma, $||.||_2$ por transformación unitaria.

(c) Demostrar que si $A$ es normal $||A||_2 = \rho(A)$.

**Teorema 2.5** (a) Sea $A$ una matriz cuadrada cualquiera y $||.||$ una norma matricial, subordinada o no, cualquiera. Entonces,

$$\rho(A) \leq ||A||$$

(b) Dada una matriz cuadrada $A$ y un número $\varepsilon > 0$ existe al menos una norma matricial subordinada tal que

$$||A|| \leq \rho(A) + \varepsilon$$

**Demostración:**

(a) Veamos primero el caso de una norma matricial subordinada. Si $\lambda$ es un valor propio de $A$ y $p$ el correspondiente vector propio, es decir $Ap = \lambda p$,

$$|\lambda|||p|| = ||\lambda p|| = ||Ap|| \leq ||A||||p||$$

de donde, $|\lambda| \leq ||A||$ y $\rho(A) \leq ||A||$. Consideremos ahora el caso de una norma cualquiera, y sea $p$ un vector propio verificando

$$p \neq 0,\ Ap = \lambda p,\ |\lambda| = \rho(A)$$

y sea $q$ un vector tal que la matriz $pq^t$ no sea nula. como

$$\rho(A)||pq^t|| = ||\lambda pq^t|| \leq ||A||||pq^t||$$

resulta $\rho(A) \leq ||A||$.

(b) El lema de Schur nos dice que existe una matriz unitaria tal que $U^*AU$ es triangular, siendo por tanto los elementos de la diagonal los valores propios de $A$.

$$U^*AU = \begin{bmatrix} \lambda_1 & b_{12} & ... & b_{1n} \\ 0 & \lambda_2 & ... & b_{2n} \\ & & ... & \\ 0 & 0 & ... & \lambda_n \end{bmatrix}$$

A todo esclar $\delta \neq 0$ le asociamos la matriz $D_\delta = diag(1, \delta, \delta^2, ..., \delta^{n-1})$ de manera que

$$(UD_\delta)^{-1}A(UD_\delta) = \begin{bmatrix} \lambda_1 & \delta b_{12} & \delta^2 b_{13} & ... & \delta^{n-1}b_{1n} \\ 0 & \lambda_2 & \delta b_{23} & ... & \delta^{n-2}b_{2n} \\ & & & ... & \\ 0 & 0 & ... & \lambda_{n-1} & \delta b_{(n-1)n} \\ 0 & 0 & 0 & ... & \lambda_n \end{bmatrix}$$

Dado $\varepsilon > 0$, fijamos $\delta$ de manera que

$$\sum_{j=i+1}^{n} |\delta^{j-i} b_{ij}| \leq \varepsilon \quad 1 \leq i \leq n-1$$

Entonces la aplicación

$$\begin{aligned} ||.|| : \quad \mathcal{A}_n &\longrightarrow \mathcal{R} \\ A &\longrightarrow ||(UD_\delta)^{-1}A(UD\delta)||_\infty \end{aligned}$$

es la norma buscada. En efecto, tenemos por una parte

$$||A|| \leq \rho(A) + \varepsilon$$

pues $||C||_\infty = \text{máx}_i \sum_j |c_{ij}|$ y es efectivamente una norma matricial subordinada por la norma vectorial

$$v \longrightarrow ||(UD_\delta)^{-1}v||_\infty$$

Aclaremos esto último:

$$||A|| = \max_{||x||=1} ||Ax|| = \max_{||(UD_\delta)^{-1}x||_\infty=1} ||(UD_\delta)^{-1}Ax||_\infty = \max_{||v||_\infty=1} ||(UD_\delta)^{-1}A(UD_\delta)v||_\infty = ||(UD_\delta)^{-1}AUD\delta||_\infty$$

**Sucesiones de matrices:** Dada una matriz $A \in \mathcal{A}_n$ consideramos la sucesión $A^k = AA...A$ $k$ veces. Queremos estudiar en qué condiciones $\lim_{k\to\infty} A^k = 0$, o lo que es lo mismo $\lim_{k\to\infty} ||A^k|| = 0$ o de forma equivalente $\lim_{k\to\infty} a_{ij}^{(k)} = 0$ $1 \leq i, j \leq n$ donde $a_{ij}^{(k)}$ es el término $ij$ de $A^k$.

**Teorema 2.6** Sea $A$ una matriz cuadrada. Las propiedades siguientes son equivalentes,

(a) $\lim_{k\to\infty} A^k = 0$

(b) $\lim_{k\to\infty} A^k v = 0 \quad \forall v \in \mathcal{R}^n$

(c) $\rho(A) < 1$

(d) $||A|| < 1$ para alguna norma subordinada.

**Demostración**

(a) (1) $\Rightarrow$ (2): Sea $||.||$ una norma vectorial y su correspondiente norma matricial subordinada,

$$\forall v \in \mathcal{R}^n \quad ||A^k v|| \leq ||A^k||||v||$$

que implica

$$0 \leq \lim_{k\to\infty} ||A^k v|| \leq ||v|| \lim_{k\to\infty} ||A^k|| = 0$$

(b) (2) $\Rightarrow$ (3): Si $\rho(A)$ fuera mayor o igual que 1, quiere decir que existe un vector propio $v$ y su correspondiente valor propio $\lambda$ tal que $|\lambda| \geq 1$, de donde, la sucesión de vectores $\{A^k v\}$ verifica $A^k v = \lambda^k v$ que no tiene límite nulo en contra de la hipótesis.

(c) (3) $\Rightarrow$ (4): Sea $\rho(A) < 1$, para todo $\varepsilon > 0$ existe una norma matricial subordinada tal que $||A|| \leq \rho(A) + \varepsilon$, tomando $\varepsilon$ suficientemente pequeño tendremos, $||A|| < 1$.

(d) (4) $\Rightarrow$ (1): Si $||A|| < 1$ entonces

$$||A^k|| \leq ||A||^k \to 0 \quad \text{para } k \to \infty$$

**Ejercicios**

(a) Sea $||.||$ una norma matricial subordinada y $B$ una matriz tal que $||B|| < 1$, demostrar que la matriz $I + B$ es no singular y que

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

(b) Sea una matriz de la forma $I + B$ singular, demostrar que entonces $||B|| \geq 1$ para toda norma matricial.

(c) Sea $B$ una matriz cuadrada y $||.||$ una norma matricial cualquiera, entonces

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

(d) Considerar la norma matricial

$$||.||_S : \mathcal{A}_n \longrightarrow \mathcal{R}$$

definida por

$$||A||_S = (\sum_{ij}^{n} |a_{ij}|^2)^{1/2}$$

llamada norma de Schur.

- Demostrar que es una norma matricial, no subordinada a ninguna norma vectorial.
- Demostrar que $||A||_S = (tr(A^*A))^{1/2}$
- Demostrar $||A||_2 \leq ||A||_S \leq \sqrt{n}||A||_2$

**Valores singulares:** Para una matriz rectangular, la noción de valor propio, no tiene sentido. Sin embargo, se puede introducir otro concepto que es el de valor singular.

**Definición 2.3** Dada una matriz rectangular de $m$ filas por $n$ columnas $A$, se llaman valores singulares $\{\mu_i\}$ de $A$ las raices cuadradas positivas de los valores propios de la matriz cuadrada $A^*A$ de orden $n$.

Tenemos el siguiente resultado general que generaliza el obtenido para matrices cuadradas.

**Teorema 2.7** Dada una matriz rectangular de $m$ filas por $n$ columnas $A$. Existen dos matrices cuadradas unitarias $U$ y $V$ de orden $m$ y $n$ respectivamente tales que

$$U^*AV = \Sigma$$

donde $\Sigma$ es una matriz rectangular de $m$ filas por $n$ columnas de la forma

$$\Sigma = \begin{bmatrix} \mu_1 & 0 & ... & 0 \\ 0 & \mu_2 & ... & 0 \\ 0 & ... & ... & 0 \\ 0 & ... & ... & \mu_n \\ 0 & ... & ... & 0 \\ 0 & ... & ... & 0 \end{bmatrix}$$

Donde $\{\mu_i\}$ son los valores singulares de la matriz $A$.

**Demostración:** $A^*A$ es una matriz autoadjunta de orden $n$, entonces existe una matriz unitaria $V$ de orden $n$ tal que

$$V^*(A^*A)V = diag(\mu_i^2)$$

siendo $\{\mu_i\}$ los valores singulares de $A$.

Sea $c_j$ el vector de dimensión $m$ formado por la $j-$ésima columna de $AV$. La igualdad matricial anterior se puede escribir

$$c_i^* c_j = \mu_i^2 \delta_{ij}, \quad 1 \leq j \leq n$$

Sean $\{\mu_1, \mu_2, ..., \mu_r\}$ el conjunto de valores singulares no nulos y $\{\mu_{r+1}, ..., \mu_n\}$ el conjunto, eventualmente vacío de valores singulares nulos.

Tenemos pues, $c_j = 0$ para $r+1 \leq j \leq n$, ya que $||c_j||_2^2 = 0$.

Pongamos $u_j = \frac{c_j}{\mu_j}$, para $1 \leq j \leq r$.

Tenemos pues por construcción $u_i^* u_j = \delta_{ij}$, para $1 \leq i, j \leq r$.

Los $r$ vectores $[u_i]$ pueden completarse para formar una base ortonormal de $\mathcal{C}^m$.

Tendremos entonces $u_i^* u_j = \delta_{ij}$, para $1 \leq i, j \leq m$ y $c_j = \mu_j u_j$, para $1 \leq j \leq n$ pues para $j \leq r$ es la definición de $u_j$ y para $j \geq r+1$ la igualdad anterior es $0 = 0$.

Sea $U$ la matriz cuadrada de orden $m$, cuya $i-$ésima columna está formada por el vector $u_i$. Es una matriz unitaria. Resulta finalmente

$$(U^* AV)_{ij} = u_i^* c_j \quad 1 \leq i \leq m \; 1 \leq j \leq n$$

que se escribe también

$$U^* AV = \Sigma = \begin{bmatrix} \mu_1 & 0 & ... & 0 \\ 0 & \mu_2 & ... & 0 \\ 0 & ... & ... & 0 \\ 0 & ... & \mu_r & 0 \\ 0 & ... & ... & 0 \\ 0 & ... & ... & 0 \end{bmatrix}$$

■

### 2.3.3. Condicionamiento

Supongamos que queremos resolver un sistema de ecuaciones

$$Au = b$$

En la práctica debido a errores de redondeo, falta de precisión en los datos, etc. resolvemos un sistema distinto

$$(A + \Delta A)v = b + \Delta b$$

donde $\Delta A$ y $\Delta b$ son perturbaciones de $A$ y de $b$ respectivamente.

La pregunta que nos hacemos ahora, es estimar o evaluar la diferencia $u - v$ en función de $\Delta A$ y de $\Delta b$. Éste sería el problema de condicionamiento asociado al sistema de ecuaciones anterior. Veamos un ejemplo,

$$A = \begin{bmatrix} 10 & 7 & 8 & 7 \\ 7 & 5 & 6 & 5 \\ 8 & 6 & 10 & 9 \\ 7 & 5 & 9 & 10 \end{bmatrix}.$$

y

$$b = [32\ 23\ 33\ 31]^t$$

$$A + \Delta A = \begin{bmatrix} 10 & 7 & 8{,}1 & 7{,}2 \\ 7{,}08 & 5{,}04 & 6 & 5 \\ 8 & 5{,}98 & 9{,}89 & 9 \\ 6{,}99 & 4{,}99 & 9 & 9{,}98 \end{bmatrix}.$$

y

$$b + \Delta b = [32{,}01\ 22{,}99\ 33{,}01\ 30{,}99]^t$$

es decir

$$\Delta A = \begin{bmatrix} 0 & 0 & 0{,}1 & 0{,}2 \\ 0{,}08 & 0{,}04 & 0 & 0 \\ 0 & -0{,}02 & -0{,}11 & 0 \\ -0{,}01 & -0{,}01 & 0 & -0{,}02 \end{bmatrix}.$$

$$\Delta b = [0{,}01\ -0{,}01\ 0{,}01\ -0{,}01]^t$$

La solución de $Au = b$ es $u = [1\ 1\ 1\ 1]^t$ mientras que la solución de $Az = b + \Delta b$ es $[1{,}82\ -0{,}36\ 1{,}35\ 0{,}79]^t$, es decir $\Delta u = u - z = [0{,}82\ -1{,}36\ 0{,}35\ 0{,}21]^t$

La solución de $(A + \Delta A)z = b$ es $z = [-81\ 137\ -34\ 22]^t$, es decir $\Delta u = u - z = [-0{,}82\ 136\ -35\ 21]^t$ así una perturbación relativa del segundo miembro del orden de $3 \times 10^{-4}$ implica una perturbación relative del segundo miembro del orden de 0,8: El error se amplifica en un factor de 2500 aproximadamente. Una perturbación relativa de la matriz de orden $10^{-2}$ conlleva una perturbación relativa de la solución de orden 80. A continuación vamos a explicar este fenómeno.

**Definición: Condicionamiento de una matriz**

Sea $||.||$ una norma matricial, el condicionamiento de una matriz regular $A$ asociado a la norma dada, es el número

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

En particular

- $cond_1(A) = ||A||_1.||A^{-1}||_1$
- $cond_\infty(A) = ||A||_\infty.||A^{-1}||_\infty$
- $cond_2(A) = ||A||_2.||A^{-1}||_2$

**Propiedades del número de condicionamiento**

(a) $cond(\alpha A) = cond(A), \quad \forall A, \forall \alpha \neq 0$

(b) $cond(A) \geq 1$ si el condicionamiento se calcula para una norma inducida.

(c) $cond_2(A) = \frac{\mu_{max}}{\mu_{min}}$ donde $\mu_{max}$ y $\mu_{min}$ son respectivamente el valor singular máximo y el valor singular mínimo de la matriz $A$.

(d) $cond_2(A) = 1$ si y solo si $A = \alpha Q$ donde $\alpha$ es un escalar y $Q$ es una matriz unitaria.

**Demostración**

(a) $cond(\alpha A) = ||\alpha A||.||(\alpha A)^{-1}|| = |\alpha|||A||.|\alpha^{-1}|||A^{-1}|| = ||A||.||A^{-1}|| = cond(A)$

(b) $1 = ||I|| = ||AA^{-1}|| \leq ||A||.||A^{-1}|| = cond(A)$

(c) $||A||_2 = (\rho(A^*A))^{1/2} = \mu_{max}$

$||A^{-1}||_2 = (\rho((A^{-1})^*A^{-1}))^{1/2} = (\rho(A^*A))^{-1/2} = \frac{1}{\mu_{min}}$

de donde $cond_2(A) = ||A||_2||A^{-1}||_2 = \frac{\mu_{max}}{\mu_{min}}$

(d) Recordemos que para toda matriz $A$ regular existen dos matrices unitarias $U$ y $V$ y una matriz diagonal $\Sigma$ cuyos coeficientes diagonales son los valores singulares $\mu_i$ de $A$ tales que

$$A = U\Sigma V^*$$

Según la propiedad anterior $cond_2(A) = 1$ si y solo si los valores singulares son todos iguales. Sea $\alpha$ este valor, entonces $\Sigma = \alpha I$ y $A = \alpha UV^*$ donde $Q = UV^*$ es una matriz unitaria.

Diremos que una matriz está bien condicionada si su número de condicionamiento no es mucho más grande que 1. La última propiedad muestra que las matrices unitarias son las que poseen el mejor número de condicionamiento posible, de ahí su uso frecuente en algoritmos de análisis numérico.

Vamos ahora a estudiar el efecto de las perturbaciones en función del número de condicionamiento de una matriz.

**Teorema 2.8** Sea $A$ una matriz regular. Sea $u$ y $u+\Delta u$ las soluciones respectivas de los sistemas lineales

$$Au = b \quad A(u + \Delta u) = b + \Delta b$$

tenemos

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

**Demostración:** Restando las dos ecuaciones obtenemos

$$A(\Delta u) = \Delta b \Rightarrow \Delta u = A^{-1}(\Delta b)$$

$$||\Delta u|| = ||A^{-1}(\Delta b)|| \leq ||A^{-1}||||\Delta b||$$

Como $||b|| = ||Au|| \leq ||A||.||u||$, es decir $\frac{1}{||u||} \leq \frac{||A||}{||b||}$ resulta

$$\frac{||\Delta u||}{||u||} \leq ||A||||A^{-1}||\frac{||\Delta b||}{||b||} = cond(A)\frac{||\Delta b||}{||b||}$$

■

La estimación anterior es la mejor posible, es decir, no existe ningún número $k$, $k < cond(a)$ tal que verifique

$$\frac{||\Delta u||}{||u||} \leq k\frac{\Delta b}{||b||}$$

se verifique siempre. En efecto, existen vectores $\Delta b$ y $u$ verificando

$$||A^{-1}(\Delta b)|| = ||A^{-1}||.||\Delta b||$$

$$||Au|| = ||A||.||u||$$

y para estos valores tendremos

$$\frac{||\Delta u||}{||u||} = ||A||.||A^{-1}||\frac{||\Delta b||}{||b||}$$

pues $\Delta u = A^{-1}(\Delta b)$ y $Au = b$. Por ejemplo veamos dos de estos vectores, supongamos por ejemplo que $A$ es simétrica y definida positiva, de valores propios $0 < \lambda_1, ..., \lambda_n$ y sean $u_1, ..., u_n$ una base ortonormal de vectores propios

$$Au_n = \lambda_n u_n$$

$$Au_1 = \lambda_1 u_1$$

tomemos $b = \lambda_n u_n$ y $\Delta b = \lambda_1 u_1$, tendremos $u = u_n$ y $\Delta u = u_1$

$$\frac{||\Delta u||_2}{||u||_2} = \frac{||u_1||_2}{||u_n||_2} = 1$$

$$\frac{||\Delta b||_2}{||b||_2} = \frac{|\lambda_1|}{\lambda_n|} = \frac{1}{cond_2(A)}$$

finalmente,

$$\frac{||\Delta u||_2}{||u||_2} = 1 = cond_2(A)\frac{||\Delta b||}{||b||}$$

Veamos ahora como afecta el condicionamiento de una matriz cuando se modifican sus términos

**Teorema 2.9** Sea $A$ una matriz regular, y sean $u$ y $u + \Delta u$ las soluciones de los dos sistemas lineales

$$Au = b \quad (A + \Delta A)(u + \Delta u) = b$$

tenemos

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

**Demostración:**

$$Au = b = (A + \Delta A)(u + \Delta u) = Au + A(\Delta u) + (\Delta A)(u + \Delta u)$$

$$0 = A(\Delta u) + \Delta A(u + \Delta u)$$

$$\Delta u = -A^{-1}\Delta A(u + \Delta u)$$

$$||\Delta u|| \leq ||A^{-1}||.||\Delta A||.||u + \Delta u||$$

$$\frac{||\Delta u||}{||u + \Delta u||} \leq ||A^{-1}||.||\Delta A|| = cond(A)\frac{||\Delta A||}{||A||}$$

■ **Ejercicios**

(a) Sea $A$ la matriz diagonal de orden $n = 100$ definida por $a_{11} = 1$ y $a_{ii} = 0{,}1\ 2 \leq i \leq n$. Calcular $||A||_2$, $||A^{-1}||_2$, $cond_2(A)$, $det(A)$.

(b) Sea $A$ triangular superior, bidiagonal, definida por $a_{ii} = 1,\ a_{i,i+1} = 2 \quad \forall i\ a_{ij} = 0$ en cualquier otro caso. Es decir,

$$A = \begin{bmatrix} 1 & 2 & 0 & ... & 0 \\ 0 & 1 & 2 & ... & 0 \\ ... & ... & ... & ... & ... \\ 0 & ... & ... & 1 & 2 \\ 0 & ... & ... & 0 & 1 \end{bmatrix} \quad A^{-1} = \begin{bmatrix} 1 & -2 & 4 & ... & (-2)^{i-1} & ... & (-2)^{n-1} \\ 0 & 1 & -2 & ... & ... & ... & (-2)^{n-2} \\ ... & ... & ... & ... & ... & ... & ... \\ 0 & ... & ... & ... & ... & 1 & -2 \\ 0 & ... & ... & ... & ... & ... & 1 \end{bmatrix}$$

- Verificar la expresión de $A^{-1}$.
- Calcular $||A||_\infty$, $||A^{-1}||_\infty$, $||A||_1$, $||A^{-1}||_1$, $cond_\infty(A)$, $cond_1(A)$.
- Calcular $det(A)$.

(c) Con las notaciones anteriores considerar

$$A = \begin{bmatrix} 10 & 7 & 8 & 7 \\ 7 & 5 & 6 & 5 \\ 8 & 6 & 10 & 9 \\ 7 & 5 & 9 & 10 \end{bmatrix} \quad b = \begin{bmatrix} 32 \\ 23 \\ 33 \\ 31 \end{bmatrix} \quad \Delta b = \begin{bmatrix} 0{,}01 \\ -0{,}01 \\ 0{,}01 \\ -0{,}01 \end{bmatrix}$$

Calcular $cond_2(A)$ y comparar la estimación

$$\frac{||Au||_2}{||u||_2} \leq cond_2(A)\frac{||\Delta b||}{||b||}$$

con el verdadero valor de $\frac{||Au||_2}{||u||_2}$.

# Capítulo 3

# Métodos directos para la resolución de ecuaciones lineales

## 3.1. Nociones preliminares. El método de Gauss

### 3.1.1. Descripción del método de Gauss para un sistema de 4 ecuaciones con 4 incógnitas

Consideremos el siguiente problema:

Hallar $x = [x_1, ..., x_4]^t \in \mathcal{R}^n$ verificando

$$\begin{aligned} 2x_1 + 1x_2 + 0x_3 + 4x_4 &= 2 \\ -4x_1 - 2x_2 + 3x_3 - 7x_4 &= -9 \\ 4x_1 + 1x_2 - 2x_3 + 8x_4 &= 2 \\ 0x_1 - 3x_2 - 12x_3 - 1x_4 &= 2 \end{aligned}$$

que escribiremos en forma matricial $Ax = b$ con

$$A = \begin{bmatrix} 2 & 1 & 0 & 4 \\ -4 & -2 & 3 & -7 \\ 4 & 1 & -2 & 8 \\ 0 & -3 & -12 & -1 \end{bmatrix} \quad b = \begin{bmatrix} 2 \\ -9 \\ 2 \\ 2 \end{bmatrix} \quad x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix}$$

Ahora realizaremos una serie de transformaciones en el sistema anterior.

Primera etapa: Poniendo $A_1 = A$, $b_1 = b$ trataremos de eliminar la incógnita $x_1$ de la ecuaciones $2, 3$ y $4$. Para ello multiplicamos la primera ecuación por 2 y la añadimos a la segunda ecuación que se transforma en

$$0x_1 + 0x_2 + 3x_3 + 1x_4 = -5$$

Análogamente, multiplicamos la primera ecuación por $-2$ y la añadimos a la tercera ecuación

$$0x_1 - 1x_2 - 2x_3 + 0x_4 = -2$$

Puesto que $x_1$ no interviene en la cuarta ecuación no modificamos ésta. Al final de esta primera etapa de eliminación de $x_1$ obtenemos un sistema equivalente $A_2x = b_2$ con

$$A_2 = \begin{bmatrix} 2 & 1 & 0 & 4 \\ 0 & 0 & 3 & 1 \\ 0 & -1 & -2 & 0 \\ 0 & -3 & -12 & -1 \end{bmatrix} \quad b = \begin{bmatrix} 2 \\ -5 \\ -2 \\ 2 \end{bmatrix}$$

Segunda etapa: Eliminación de $x_2$. Para eliminar $x_2$ de las ecuaciones tercera y cuarta, no podemos utilizar la segunda ecuación puesto que en ella el coeficiente de $x_2$, que llamamos pivote, es nulo. Por el contrario, se puede utilizar una de las dos últimas ecuaciones. Elijamos, por ejemplo, la tercera, en este caso, es el coeficiente $-1$ que juega el papel de pivote. Puesto que el coeficiente de $x_2$ en la segunda ecuación es ya nulo, $x_2$ está eliminada en esta ecuación. Multipliquemos la tercera ecuación ( la del pivote) por $-3$ y sumemosla a la cuarta, que se convierte en

$$0x_1 + 0x_2 - 6x_3 - x_4 = 8$$

Si permutamos las ecuaciones segunda y tercera, obtenemos el sistema lineal equivalente $A_3x = b_3$ con

$$A_3 = \begin{bmatrix} 2 & 1 & 0 & 4 \\ 0 & -1 & -2 & 0 \\ 0 & 0 & 3 & 1 \\ 0 & 0 & -6 & -1 \end{bmatrix} \quad b = \begin{bmatrix} 2 \\ -2 \\ -5 \\ 8 \end{bmatrix}$$

Tercera etapa: Eliminación de $x_3$ en la cuarta ecuación. Sólo queda eliminar $x_3$ en la última ecuación. Como el coeficiente $x_3$ de la tercera ecuación es no nulo, lo elegiremos como pivote. Se multiplica la tercera ecuación por 2 y la añadimos a la cuarta para obtener

$$0x_1 + 0x_2 + 0x_3 + x_4 = -2$$

¡el sistema de ecuaciones lineales, equivalente, que finalmente hemos obtenido es $A_4x = b_4$ con

$$A_4 = \begin{bmatrix} 2 & 1 & 0 & 4 \\ 0 & -1 & -2 & 0 \\ 0 & 0 & 3 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad b = \begin{bmatrix} 2 \\ -2 \\ -5 \\ -2 \end{bmatrix}$$

Resolución del sistema triangular La matriz $A_4$ es triangular superior. Para obtener la solución del sistema $A_4x = b_4$, se resuelve primero la cuarta ecuación, después la tercera, etc, mediante un proceso llamado de remonte.

$$\begin{aligned} x_4 &= -2 \\ x_3 &= (-5 - x_4)/3 = -1 \\ x_2 &= 2 - 2x_3 = 4 \\ x_1 &= (2 - x_2 - 4x_4)/2 = 3 \end{aligned}$$

Hemos obtenido la solución $x = [3\ 4\ -1\ -2]^t$.

### 3.1.2. Escritura matricial de las operaciones de eliminación

La transformación consistenete en multiplicar la primera componente de $a_1$ de un vector por un número $\alpha$ y añadirlo a la segunad componente $a_2$ se puede representar mediante la multiplicación de una matriz, llamemosla $E_1$ por este vector, en efecto, poniendo

$$E_1 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ \alpha & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

tendremos, designando $a = [a_1\ a_2\ a_3\ a_4]^t$

$$E_1 a = \begin{bmatrix} 1 & 0 & 0 & 0 \\ \alpha & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \end{bmatrix} = \begin{bmatrix} a_1 \\ \alpha a_1 + a_2 \\ a_3 \\ a_4 \end{bmatrix}$$

Análogamente, multiplicar $a_1$ por $\beta$ y añadirlo a $a_3$ se puede expresar

$$E_2 a = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ \beta & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \end{bmatrix} = \begin{bmatrix} a_1 \\ a_2 \\ \beta a_1 + a_3 \\ a_4 \end{bmatrix}$$

Efectuar las dos transformaciones sucesivas (el orden no tiene importancia) equivales a multiplicar el vector por el producto de las dos matrices

$$M_1 = E_2 E_1 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ \beta & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ \alpha & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ \alpha & 1 & 0 & 0 \\ \beta & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

Las anteriores transformaciones son invertibles y

$$E_1^{-1} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -\alpha & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad E_2^{-1} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ -\beta & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad M_1^{-1} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -\alpha & 1 & 0 & 0 \\ -\beta & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

con el significado correspondiente: Si a $a_2 + \alpha a_1$ le restamos $-\alpha a_1$ volvemos a obtener $a_2$. Llamemos $L_1$ a la inversa de $M_1$, $M_1^{-1} = E_1^{-1} E_2^{-1}$. En la primera etapa, a partir del sistema $A_1 x = b_1$, eligiendo $\alpha = 2$ y $\beta = -2$ hemos obtenido el sistema $A_2 x = b_2$ con $M_1 A_1 = A_2$ y $M_1 b_1 = b_2$. De donde deducimos las relaciones

$$A_1 = L_1 A_2 \quad b_1 = L_1 b_2$$

En la segunda etapa, se ha efectuado una permutación de las filas segunda y tercera para obtener un pivote no nulo. Esto se interpreta matricialmente por la multiplicación por la matriz elemental de permutación

$P$ siguiente

$$P = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

Después se ha eliminado $x_2$, que equivale a la multiplicación por la matriz $M_2$, donde

$$M_2 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & -3 & 0 & 1 \end{bmatrix}$$

El sistema obtenido es $A_3x = b_3$ con

$$A_3 = M_2PA_2 \quad b_3 = M_2Pb_2$$

poniendo

$$L_2 = M_2^{-1} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 3 & 0 & 1 \end{bmatrix}$$

se deduce

$$PA_2 = L_2A_3 \quad Pb_2 = L_2b_3$$

Finalmente, la última etapa conduce al sistema $A_4x = b_4$ donde

$$A_4 = M_3A_3 \quad b_4 = M_3b_3$$

o bien

$$A_3 = L_3A_4 \quad b_3 = L_3b_4$$

con

$$L_3 = M_3^{-1} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & -2 & 1 \end{bmatrix}$$

Llamando $U = A_4$ y $c = b_4$ hemos llegado al sistema equivalente

$$Ux = c$$

que es triangular superior.

Interpretación del método de Gauss como método de factorización: Pongamos

$$L_1{}' = PL_1P^t = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -\beta & 1 & 0 & 0 \\ -\alpha & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

donde $\alpha = 2$ y $\beta = -2$ y $L = {L_1}'L_2L_3$. Un cálculo simple da

$$L = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ -2 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 3 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & -2 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ -2 & 0 & 1 & 0 \\ 0 & 3 & -2 & 1 \end{bmatrix}$$

Observemos que para obtener $L$ basta copiar en su sitio respectivo los coeficientes bajo la diagonal de ${L_1}', L_2$ y $L_3$. Calculemos el producto $LU$:

$$LU = {L_1}'L_2L_3A_4 = {L_1}'L_2A_3 = {L_1}'PA_2 = PL_1P^tPA_2 = PL_1A_2 = PA_1 = PA$$

Hemos obtenido así la factorización de la matriz $PA$ en el producto de dos matrices triangulares, una triangular inferior $L$ (del inglés "Lower") con los coeficientes diagonales iguales a 1 y una triangular superior $U$ ( del inglés "Upper"). El cálculo de $P$, $L$ y $U$ puede hacerse independientemente del segundo miembro $b$. A continuación para resolver el sistema $Ax = b$, se resolverá $PAx = LUx = Pb$ resolviendo dos sistemas triangulares

$$\begin{aligned} Ly &= Pb \\ Ux &= y \end{aligned}$$

### 3.1.3. El método de eliminación de Gauss para un sistema regular $N \times N$

Para resolver el sistema lineal $Ax = b$ donde $A$ es una matriz de orden $N$ regular, la vamos a tansformar en $N-1$ etapas en un sistema equivalente $Ux = c$, donde $U$ es una matriz triangular superior.

Describiremos primero la primera etapa y luego la $k$-ésima etapa que sustituye el sistema $A_kx = b_k$ por un sistema equivalente $A_{k+1}x = b_{k+1}$.

Introducimos las matrices ampliadas $\tilde{A}_k$ de formato $N \times (N+1)$ obtenidas añadiendo el segundo miembro $b_k$ a la matriz $A_k$.

$$\tilde{A}_k = [A_k|b_k]$$

(a) Primera etapa: Eliminación de $x_1$ de las ecuaciones 2 a $N$.

Vamos a tener que efectuar divisiones por el coeficiente $A_1(1,1)$ al que llamaremos primer pivote. Si este coeficiente es nulo, se efectúa previamente una permutación de la primera fila con la fila $p$, para conseguir una matriz cuyo coeficiente, de posición $(1,1)$ no sea nulo. Esto es siempre posible, pues $det(A_1) \neq 0$ implica que existe al menos un coeficiente no nulo en la primera columna. Se verá más tarde como, en la práctica, se define la permutación para evitar tener un pivote demasiado pequeño en valor absoluto.

Este intercambio es equivalente a la multiplicación del sistema $A_1 x = b_1$ por una matriz elemental de permutación $P_1$ tal que:

$$P_1 = \begin{bmatrix} 0 & ... & ... & 1 & & & & \\ ... & 1 & ... & ... & & & & \\ ... & ... & 1 & & & & & \\ 1 & ... & ... & 0 & & & & \\ ... & ... & ... & ... & 1 & & & \\ ... & ... & ... & ... & ... & ... & ... & ... \\ & & & & & & & 1 \end{bmatrix}$$

Para escribir las fórmulas que siguen llamaremos también $A_1$ a la matriz de partida incluso si hemos intercambiado la primera fila y la fila $p$. Para $i$ comprendido entre 2 y $N$ restamos de la $i$-ésima fila la primera fila multiplicada por

$$L(i,1) = \frac{A_1(i,1)}{A_1(1,1)}$$

tenemos pues las fórmulas siguientes. Llamando $\tilde{A}_2$ a la matriz transformada

$$\tilde{A}_2(i,j) = \tilde{A}_1(i,j) - L(i,1)\tilde{A}_1(1,j) \quad 2 \leq i \leq N \quad 1 \leq j \leq N+1$$

Para $j = 1$ observamos $\tilde{A}_2(i,1) = 0 \quad 2 \leq i \leq N$ como queríamos. Efectuaremos los cálculos solamente para $2 \leq j \leq N+1$.

$\tilde{A}_2$ posee la estructura siguiente

$$\tilde{A}_2 = \left[\begin{array}{ccccc|c} x & ... & ... & ... & x & x \\ 0 & x & ... & ... & x & . \\ 0 & x & x & ... & x & . \\ ... & ... & ... & ... & ... & . \\ 0 & x & ... & ... & x & x \end{array}\right]$$

Como las primeras filas de $\tilde{A}_1$ y de $\tilde{A}_2$ son iguales podemos escribir

$$\tilde{A}_1(i,j) = L(i,1)\tilde{A}_2(1,j) + \tilde{A}_2(i,j)$$

Definimos ahora el vector $l_1$ de dimensión $N$ y la matriz cuadrada $L_1$ de orden $N$ mediante

$$l_1(1) = 0 \quad l_1(i) = L(i,1) \quad \text{para } 2 \leq i \leq N$$

$$L_1 = \begin{bmatrix} 1 & 0 & ... & ... & 0 \\ l_1(2) & 1 & 0 & ... & 0 \\ l_1(3) & 0 & 1 & ... & 0 \\ ... & ... & ... & ... & ... \\ l_1(N) & 0 & ... & ... & 1 \end{bmatrix} = I + l_1 e_1^t$$

donde $e_1$ es el primer vector de la base canónica de $\mathcal{R}^N$.

$L_1$ es triangular inferior, los coeficientes de su diagonal son iguales a 1, así pues $det(L_1) = 1$. Fuera de la diagonal los únicos términos distintos de cero son los de la primera columna. La multiplicación

por la izquierda de una matriz por $L_1$ equivale a añadir la primera fila multiplicada por $L(i,1)$ a la $i$-ésima fila y esto para $2 \leq i \leq N$. Podemos pues escribir

$$P_1 \tilde{A}_1 = L_1 \tilde{A}_2$$

como $P_1^{-1} = P_1$ tenemos

$$\tilde{A}_1 = P_1 L_1 \tilde{A}_2$$

es decir

$$A_1 = P_1 L_1 A_2 \quad b_1 = P_1 L_1 b_2$$

Como $det(L_1) = 1$ deducimos $det(A_1) = \varepsilon_1 det(A_2)$ donde $\varepsilon_1 = +1$ si no se ha efectuado ninguna permutación y $\varepsilon = -1$ en caso contrario.

(b) $k$-ésima etapa:Eliminación de $x_k$ de las ecuaciones $k+1$ hasta $N$.

El cálculo que se va a describir monstrará, por inducción que al final de la etapa $(k+1)$ la matriz $\tilde{A}_k$ posee la estructura siguiente

$$\tilde{A}_k = [A_k, b_k] = \left[\begin{array}{cccccccccccc|c} x & ... & ... & ... & ... & x & x & & & & x & x \\ 0 & x & ... & ... & ... & x & & & & & & . \\ 0 & 0 & x & ... & ... & x & & & & & & . \\ ... & ... & ... & ... & ... & ... & & & & & & . \\ 0 & 0 & ... & ... & ... & x & & & & & & . \\ 0 & 0 & ... & ... & ... & 0 & x & ... & ... & ... & ... & x & . \\ 0 & 0 & ... & ... & ... & 0 & x & x & ... & ... & ... & x & . \\ 0 & 0 & ... & ... & ... & 0 & x & x & ... & ... & ... & x & . \\ ... & ... & ... & ... & ... & ... & ... & ... & ... & ... & ... & x & . \\ 0 & 0 & ... & ... & ... & 0 & x & x & ... & ... & ... & x & x \end{array}\right] = \left[\begin{array}{cc|c} T_k & X & X \\ O & A_k^{'} & X \end{array}\right]$$

donde los coeficientes de $A_k$ situados debajo de la diagonal, en los $k-1$ primeras columnas son nulos, siendo el bloque $T_k$ de la figura triangular superior. Si consideramos la descomposición de $A_k$ en bloques de dimensión $k-1$ y $N-k+1$ como en la figura, vemos que

$$det(A_k) = det(T_k).det(A_k^{'})$$

Demostraremos más adelante por inducción que $|det(A_k)| = |det(A)| \neq 0$. Ello implica que $det(A_k^{'}) \neq 0$ de donde al menos uno de los coeficientes de la primera columna de $A_k^{'}$, $A_k(i,k) \quad i \geq k$ es distinto de cero. Si el $k$-ésimo pivote, es decir el coeficiente $A_k(k,k)$ es nulo, se puede permutar la $k$-ésima fila con una fila de índice $p \geq k+1$ tal que $A_k(p,k) \neq 0$ y este coeficiente jugará el papel de pivote. Esta permutación es equivalente a la multiplicación del sistema

$$A_k x_x = b_k$$

por una matriz de permutación elemental $P_k$

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

Ahora las $k$ primeras filas no se cambian y para $k+1 \leq i \leq N$ se resta de la $i$-ésima fila la $k$-ésima fila multiplicada por

$$L(i,k) = \frac{A_k(i,k)}{A_k(k,k)}$$

Se define $\tilde{A}_{k+1}$ a partir de $\tilde{A_k}$ por las fórmulas

$$\tilde{A}_{k+1}(i,j) = \tilde{A}_k(i,j) - L(i,k)\tilde{A}_k(k,j) \quad \text{para} \quad k+1 \leq i \leq N \quad 1 \leq j \leq N+1$$

tomando $j = k$

$$A_{k+1}(i,k) = A_k(i,k) - \frac{A_k(i,k)}{A_k(k,k)} = 0$$

que es el resultado buscado. Además para todos los índices $j \leq k-1$ tenemos $A_k(k,j) = 0$ y $A_k(i,j) = 0$ para $i \geq k+1$, de donde $A_{k+1}(i,j) = 0$ para $i \geq k+1$. Es decir los ceros de las $k-1$ primeras columnas de las filas $k+1$ a $N$ se conservan. Observemos que las ceros de las filas 1 a $k$ se conservan pues estas filas no se modifican. La matriz $\tilde{A}_{k+1}$ tiene pues la estructura de la matriz anterior $\tilde{A}_k$ cambiando $k$ por $k+1$.

El número de operaciones efectuado en esta $k$-ésima etapa es:

- $N-k$ divisiones.
- $(N-k)^2$ adiciones y multiplicaciones para los coeficientes $A_{k+1}$.
- $N-k$ adiciones y multiplicaciones para los coeficientes de $b_{k+1}$.

Como la $k$-ésima fila de $\tilde{A}_k$ y de $\tilde{A_{k+1}}$ son iguales podemos escribir

$$\tilde{A}_k(i,j) = L(i,k)\tilde{A}_{k+1}(k,j) + \tilde{A}_{k+1}(i,j) \quad \text{para} \quad k+1 \leq i \leq N \quad 1 \leq j \leq N+1$$

Definimos el vector $l_k$ de dimensión $N$ y la matriz cuadrada $L_k$ de orden $N$ por

$$l_k(i) = 0 \quad \text{para} \quad i \leq k \quad l_k(i) = L(i,k) \quad k+1 \leq i \leq N$$

$$L_k = \begin{bmatrix} 1 & & & & & & & & \\ & 1 & & & & & & & \\ \dots & \dots & \dots & & \dots & \dots & \dots & \mathrm{O} & \dots \\ & & & 1 & & & & & \\ & & & l_k(k+1) & & & & & \\ & \mathrm{O} & & \dots & & & & & \\ & & & l_k(N) & & & & 1 & \end{bmatrix} = I + l_k e_k^t$$

donde $e_k$, es el $k$-ésimo vector de la base canónica.

Tenemos:

- $det(L_k) = 1$
- $P_k \tilde{A}_k = L_k \tilde{A}_{k+1}$
- Como $P_k^2 = I$, $\tilde{A}_k = P_k L_k \tilde{A}_{k+1}$
- $A_k = P_k L_k A_{k+1} \quad b_k = P_K L_K b_{k+1}$
- $$det(A_k) = \varepsilon_k det(A_{k+1}) \quad \varepsilon_k = \begin{cases} 1 & \text{si} \quad P_k = I \\ -1 & \text{si} \quad P_k \neq I \end{cases}$$
- Por inducción tenemos $|det(A_k)| = |det(A)|$.

(c) Resolución del sistema triangular: Al final de la etapa $N-1$, la matriz $A_N$ es triangular superior. Solo falta realizar la etapa de remonte para resolver el sistema lineal $Ux = c$, cuya $i$-ésima ecuación se escribe

$$U(i,i)x_i + \sum_{j=i+1}^{N} U(i,j)x_j = c_i$$

Se calculan las componentes de $x$ por orden de índice decreciente empezando por $x_N$:

$$\begin{cases} x_N = \frac{c_N}{U(N,N)} \\ x_i = \frac{c_i - \sum_{j=i+1}^{N} U(i,j)x_j}{U(i,i)} \end{cases}$$

Como los $U(i,i)$ son los pivotes sucesivos, son todos no nulos y las fórmulas anteriores son válidas. El número de operaciones a efectuar en esta fase es:

- $N$ divisiones.
- $\sum_i^N (N-i) = \frac{N(n-1)}{2}$ adiciones y multiplicaciones.

(d) Cálculo del determinante: $det(A_k) = \varepsilon det(A_{k+1})$ para $1 \leq k \leq N-1$ donde $\varepsilon = -1$ o $\varepsilon = +1$ según se halla realizado una permutación o no en la $k$-ésima etapa.

$$det(A) = det(A_1) = (-1)^p det(A_N) = (-1)^p det(U)$$

siendo $p$ el número de permutaciones de filas realizadas. Como $U$ es triangular superior

$$det(U) = \Pi_{i=1}^{N} U(i,i)$$

(e) Interpretación matricial: $PA = LU$

Hemos visto,

$$A_k = P_k L_k A_{k+1} \quad 1 \leq k \leq N-1$$

sabiendo que $A_1 = A \quad A_N = U$ matriz triangular superior y siendo $P_k$ la matriz de permutación de índices $k$ y $l$ con $l \geq k$. Podemos enunciar el siguiente

**Lema 3.1** $P_p$ matriz de permutación de índices $p$ y $q \geq p$ entonces para $k < p$

$$L_k P_p = P_p L_k^{'} \quad \text{o} \quad P_p L_k P_p = L_k^{'}$$

donde la matriz $L_k^{'}$ obtiene de $L_k$ permutando los coeficientes de las líneas $p$ y $q$ en la $k$-ésima columna.

**Demostración:** para $p > k$ $P_p e_k = e_k$ y $l_k^{'} = P_p l_k$.

$$L_k^{'} = P_p(I + l_k e_k^t)P_p = P_p^2 + l_k^{'} e_k^t = I + l_k^{'} e_k^t$$

donde $l_k^{'}$ se deduce de $l_k$ por permutación de las componentes $p$ y $q$.■

**Lema 3.2** Sea $l_k \quad 1 \leq k \leq N-1$ una sucesión de $N-1$ vectores de dimensión $N$ tales que las $k$ primeras componentes de $l_k$ sean nulas, entonces:

$$(I + l_1 e_1^t)(I + l_2 e_2^t)...(I + l_{N-1} e_{N-1}^t) = I + l_1 e_1^t + l_2 e_2^t + ... + l_{N-1} e_{N-1}^t$$

**Demostración:** Observar que los productos $l_i e_i^t l_j e_j^t$ con $i < j$ son nulos pues $e_i^t l_j = l_j(i) = 0$ para $j > i$.■

**Teorema 3.1** Sea $A$ una matriz regular de orden $N$. Existe una matriz de permutación $P$ y dos matrices $L$ y $U$, $L$ triangular inferior de diagonal unidad, $U$ triangular superior, tales que

$$PA = LU$$

**Demostración:** $A = A_1 = P_1 L_1 P_2 L_2 ... P_{N-1} L_{N-1} U$ Se verifica fácilmente

$$A = P_1 P_2 ... P_{N-1} L_1^{'} L_2^{'} ... L_{N-1}^{'} U$$

donde

$$L_k^{'} = P_{N-1} P_{N-2} .... P_{k-1} L_k P_{k+1} ... P_{N-2} P_{N-1}$$

$$L = L_1^{'} L_2^{'} ... L_{N-1}^{'}$$

$$P = P_{N-1} P_{N-2} ... P_1$$

$$PA = LU$$

### 3.1.4. Existencia y unicidad de $A = LU$

**Teorema 3.2** Sea $A$ una matriz regular de orden $N$. $A$ posee una factorización $A = LU$ donde $L$ es triangular inferior de diagonal unidad y $U$ es triangular superior, si y solo si, todas las submatrices principales $A^{(k)}$ de $A$ son regulares. En ese caso

$$U(k,k) = \frac{det(A^{(k)})}{det(A^{(k-1)})}$$

**Demostración:** Si $A = LU$, descomponemos las tres matrices en bloques de orden $k$ y orden $N - k$

$$\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} = \begin{bmatrix} L_{11} & O \\ L_{21} & L_{22} \end{bmatrix} . \begin{bmatrix} U_{11} & U_{12} \\ O & U_{22} \end{bmatrix}$$

Por definición $A_{11}$ es la submatriz principal de orden $k$

$$A_{11} = L_{11}U_{11} \Rightarrow det(A_{11}) = det(L_{11})det(U_{11}) = \Pi_{i=1}^{k} U(i,i) \neq 0$$

y de ahí

$$U(k,k) = \frac{det(A^{(k)})}{det(A^{(k-1)})}$$

.

Recíprocamente, si los determinantes de las submatrices principales $A^{(k)}$ son todos distintos de cero, veamos que teóricamente podriamos aplicar el algoritmo de eliminación de Gauss sin efectuar permutaciones.

Procedemos por inducción

- Para $k = 1$, como $A^{(1)} = A_{11} \neq 0$, en la primera etapa no necesitamos efectuar ninguna permutación de filas.
- Supongamos que la eliminación sin permutaciones se ha podido realizar hasta la etapa $k-1$. Antes de la etapa $k$-ésima, tendremos

  $$A = A_1 = L_1 L_2 .... L_{k-1} A_k = R A_k$$

  donde $R$ es triangular inferior de diagonal unidad. Descomponiendo en $A$, $R$ y $A_k = B$ en bloques

  $$\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} = \begin{bmatrix} R_{11} & O \\ R_{21} & I \end{bmatrix} . \begin{bmatrix} B_{11} & B_{12} \\ O & B_{22} \end{bmatrix}$$

  resulta $A_{11} = R_{11}B_{11}$, como $det(R_{11}) = 1$

  $$det(B_{11}) = det(A_{11}) = det(A^{(k)}) \neq 0$$

  y $B_{11}$ es triangular superior tenemos todos sus elementos diagonales no nulos, en particular, $B(k,k) = A_k(k,k)$ se puede elegir como pivote en la siguiente etapa. ■

**Teorema 3.3** de unicidad

Si una matriz regular $A$ de orden $N$, posee una factorización $A = LU$, con $L$ triangular inferior de diagonal unidad y $U$ triangular superior, entonces la factorización es única.

**Demostración:** Supongamos $A = L_1U_1 = L_2U_2$, resulta $X = L_2^{-1}L_1 = U_2U_1^{-1}$, de donde necesariamente $X = I$.■

### 3.1.5. Complejidad algoritmica del método de eliminación de Gauss

- Factorización En la etapa $k$-ésima de la factorización, $N-k$ divisiones y $(N-k)^2$ sumas y multiplicaciones. sumando para todo $k = 1, ..., N-1$:

$$\sum_{k=1}^{N-1}(N-k)^2 = \frac{1}{3}N(N-1/2)(N-1)$$

  sumas y multiplicaciones y

$$\sum_{k=1}^{N-1} = \frac{1}{2}N(N-1)$$

  divisiones

- Resolución de un sistema triangular

$$\frac{1}{2}N(N-1)$$

  sumas y multiplicaciones y $N$ divisiones.

En total $O(\frac{N^3}{3})$ para la factorización y $O(N^2)$ para la resolución de los sistemas triangulares. Por ejemplo resolver un sistema de 100 ecuaciones significa realizar del orden de $10^6/3$ operaciones.

**Tema para clase de prácticas:** Requarimientos de memoria del algoritmo de Gauss. Solo se necesita el espacio ocupado por la matriz y el vector segundo miembro inicial y un segundo vector adicional si es necesario realizar permutaciones.

**Ejercicios:**

(a) Influencia de los errores de redondeo. Resolver el sistema siguiente

$$\begin{aligned} \varepsilon x_1 + x_2 &= 1/2 \\ x_1 + x_2 &= 1 \end{aligned}$$

sin realizar permutaciones y realizando una permutación de filas para evitar pivotes pequeños. Comparar con la solución exacta.

## 3.2. El método de Cholesky para matrices simétricas y definidas positiva

Consideramos en esta sección una matrices $A$ simétricas y definidas positiva. Una factorización regular de Cholesky es una factorización de la forma $A = BB^t$ donde $B$ es triangular inferior. Si los coeficientes de la diagonal de $B$ son todos positivos se dice que es una factorización positiva de Cholesky.

**Teorema 3.4** de existencia y unicidad

$A$ posee una factorización regular de Cholesky si y solamente es definida positiva. En este caso posee una factorización positiva única.

**Demostración:**

- Necesidad: Sea $A$ simétrica y supongamos que $A = BB^t$ con $B$ triangular inferior. Entonces $A$ es definida positiva, en efecto

$$(Ax, x) = (BB^t x, x) = (B^t x, B^t x) = ||B^t x||_2 > 0 \quad \forall x \neq 0$$

- Suficiencia: Sea $A$ simétrica y definida positiva, es decir, $(Ax, x) > 0 \quad \forall x \neq 0$. Procedemos por inducción. El resultado es obviamente cierto para matrices de orden 1. Supongamos pues que es cierto para matrices de orden $N-1$ y demostremos que entonces será cierto para matrices de orden $N$. Descomponemos la matriz $A$ en bloques,

$$A = \begin{bmatrix} R & a \\ a^t & \alpha \end{bmatrix}$$

donde $R$ es una matriz de orden $N-1$, $a$ un vector de $\mathcal{R}^N$ y $\alpha$ un número real. Como $A$ es simétrica definida positiva, necesariamente también lo será $R$. Por la hipótesis de inducción que $R = MM^t$ con $M$ triangular inferior de orden $N-1$, por tanto $R$ es simétrica y definida positiva. Busquemos $B$ de la forma

$$B = \begin{bmatrix} M & O \\ r^t & \rho \end{bmatrix}$$

con $r \in \mathcal{R}^{N-1}$ y $\rho \in \mathcal{R}$. Se deber cumplir

$$\begin{bmatrix} R & a \\ a^t & \alpha \end{bmatrix} = \begin{bmatrix} M & O \\ r^t & \rho \end{bmatrix} \cdot \begin{bmatrix} M^t & r \\ O & \rho \end{bmatrix} = \begin{bmatrix} MM^t & Mr \\ r^t M & r^t r + \rho^2 \end{bmatrix}$$

identificando términos

$$a = Mr$$

$$\alpha = r^t r + \rho^2$$

es decir

$$r = M^{-1} a$$

$$\rho^2 = \alpha - r^t r$$

de donde

$$\rho^2 = \alpha - a^t R^{-1} a$$

Basta verificar ahora que $\rho$ es necesariamente un número real, es decir, que $\alpha - a^t R^{-1} a > 0$. En efecto, como $A$ es definida positiva, para todo vector $x \in \mathcal{R}^N$ distinto de cero se tiene

$$x^t \begin{bmatrix} R & a \\ a^t & \alpha \end{bmatrix} x > 0$$

tomando $x = \begin{bmatrix} R^{-1} & -1 \end{bmatrix}^t$

$$\begin{bmatrix} R^{-1}a & -1 \end{bmatrix} \begin{bmatrix} R & a \\ a^t & \alpha \end{bmatrix} \begin{bmatrix} R^{-1}a \\ -1 \end{bmatrix} = \alpha - a^t R^{-1} a$$

de ahí el resultado buscado.

- Queda por verificar la unicidad. Supongamos que existen $B_1$ y $B_2$ matrices triangular inferior tales que

$$A = B_1 B_1^t = B_2 B_2^t$$

resulta

$$X = B_2^{-1} B_1 = B_2^t B_1^{-t}$$

de donde necesariamente $X$ es diagonal y además los términos de la diagonal verifican

$$\frac{b_{ii}^{(1)}}{b_{ii}^{(2)}} = \frac{b_{ii}^{(2)}}{b_{ii}^{(1)}}$$

de donde

$$(b_{ii}^{(1)})^2 = (b_{ii}^{(2)})^2$$

eligiendo la diagonal con términos positivos, resulta

$$b_{ii}^{(1)} = b_{ii}^{(2)}$$

es decir $B_1 = B_2$.■

# Capítulo 4

# Métodos iterativos para la resolución de sistemas de ecuaciones lineales

## 4.1. Generalidades y descripción de algunos métodos

Los métodos directos son eficaces para sistemas de tamaño moderado, por ejemplo $N \approx 1000$ o en el caso de matrices huecas $N \approx 5000\,, 10000$. Para valores significativamente mayores los métodos directos pierden eficacia, no solo porque el número de operaciones necesario crece desmesuradamente sino también porque la acumulación de errores de redondeo puede desvirtuar el resultado.

En el caso de grandes sistemas de ecuaciones, los llamados métodos iterativos, resultan más convenientes. De forma genérica: Para resolver un sistema $Ax = b$, se transforma en otro equivalente (es decir, con la misma solución) que tenga la forma

$$x = Bx + c$$

expresión que sugiere el siguinete método iterativo

$$\begin{cases} x^{(0)}, & \text{arbitrario} \\ x^{(k+1)} = Bx^{(k)} + c \end{cases}$$

Diremos que el método es convergente, si

$$\lim_{k\to\infty} x^{(k)} = x$$

cualquiera que sea el valor inicial $x^{(0)}$ elegido.

**Teorema 4.1** El método iterativo anterior es convergente si $\rho(B) < 1$, o de forma equivalente si $||B|| < 1$ para al menos una norma matricial (que podemos elegir subordinada).

**Demostración:**

$$x = Bx + c$$
$$x^{(k+1)} = Bx^{(k)} + c$$

restando

$$x^{(k+1)} - x = B(x^{(k)} - x)$$

Llamando $e^{(0)} = x^{(0)} - x$ al error inicial y $e^{(k)} = x^{(k)} - x$ al error en la iteración $k$ resulta $e^{(k+1)} = Be^{(k)}$ y también

$$e^{(k)} = Be^{(k-1)} = ... = B^k e^{(0)}$$

de donde

$$||e^{(k)}|| = ||B^k e^{(0)}|| \leq ||B^k||||e^{(0)}|| \leq ||B||^k ||e^{(0)}||$$

Si $||B|| < 1$ entonces

$$\lim_{k\to\infty} ||e^{(k)}|| = 0$$

es decir,

$$\lim_{k\to\infty} x^{(k)} = x$$

■

### 4.1.1. Descripción del método de Jacobi

Supongamos que queremos resolver el sistema

$$\begin{aligned} a_{11}x_1 + \quad & ... \quad + a_{1n}x_d = b_1 \\ a_{21}x_1 + \quad & ... \quad + a_{2n}x_d = b_2 \\ & ... \\ a_{n1}x_1 + \quad & ... \quad + a_{nn}x_d = b_n \end{aligned}$$

que podemos escribir

$$\begin{aligned} a_{11}x_1 &= b_1 - \sum_{j\neq 1} a_{1j}x_j \\ a_{22}x_2 &= b_2 - \sum_{j\neq 2} a_{2j}x_j \\ & ... \\ a_{NN}x_N &= b_N - \sum_{j\neq N} a_{Nj}x_j \end{aligned}$$

El algoritmo de Jacobi se escribe Dado $x^{(0)} \in \mathcal{R}^N$ arbitrario, una vez calculado una aproximación $x^{(k)}$, calculamos $x^{(k)}$ de la manera siguiente:

$$\begin{aligned} x_1^{(k+1)} &= \frac{b_1 - \sum_{j\neq 1} a_{1j} x_j^{(k)}}{a_{11}} \\ x_2^{(k+1)} &= \frac{b_2 - \sum_{j\neq 2} a_{2j} x_j^{(k)}}{a_{22}} \\ &... \\ x_N^{(k+1)} &= \frac{b_N - \sum_{j\neq N} a_{Nj} x_j^{(k)}}{a_{NN}} \end{aligned}$$

Este método está definido solo si $a_{ii} \neq 0$ para $i = 1, ..., N$. La ecuación $i$-ésima

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

se puede escribir también, restando en los dos miembros $x_i^{(k)}$ así:

$$x_i^{(k+1)} - x_i = \frac{b_i - \sum_{j=1}^{N} a_{ij} x_j^{(k)}}{a_{ii}} = \frac{r_i^{(k)}}{a_{ii}}$$

donde hemos designado mediante $r^{(k)}$ al vector residuo

$$r^{(k)} = b - Ax^{(k)}$$

correspondiente al valor $x^{(k)}$.

Vamos a escribir el método anterior en forma matricial. Podremos

$$A = D - E - F$$

donde

- $D$, es la parte diagonal de $A$, $D_{ii} = a_{ii}, \quad i = 1, ..., N$
- $-E$, es la parte estrictamente triangular inferior

$$\begin{cases} (-E)_{ij} = a_{ij} & i > j \\ (-E)_{ij} = 0 & i \leq j \end{cases}$$

- $-F$, es la parte estrictamente triangular superior

$$\begin{cases} (-F)_{ij} = a_{ij} & i < j \\ (-F)_{ij} = 0 & i \geq j \end{cases}$$

Entonces la iteración de Jacobi se escribe

$$x^{(k+1)} = D^{-1}(E+F)x^{(k)} + D^{-1}b$$

o bien,

$$x^{(k+1)} = (I - D^{-1}A)x^{(k)} + D^{-1}b$$

es pues de la forma general con $B = I - D^{-1}A$ y $c = D^{-1}b$.

Pseudocódigo de una iteración del método de Jacobi

- For $i = 1, ..., N$ do
- $s \leftarrow b(i)$
  - For $j = 1, ..., i-1$ do
  - $s \leftarrow s - A(i,j) * x(j)$
  - end
  - For $j = i+1, ..., N$ do
  - $s \leftarrow s - A(i,j) * x(j)$
  - end
- $y(i) \leftarrow \frac{s}{A(i,i)}$
  - For $i = 1, ..., N$ do
  - $x(i) \leftarrow y(i)$
  - end
- end

**Ejercicios**

(a) Considerar el sistema

$$\begin{aligned} 10x_1 + x_2 &= 11 \\ 2x_1 + 10x_2 &= 12 \end{aligned}$$

*a*) Calcular la solución exacta.

*b*) Calcular el radio espectral de la matriz asociada.

*c*) Aplicar el método de Jacobi.

(b) Considerar el sistema

$$\begin{aligned} x_1 + 10x_2 &= 11 \\ 10x_1 + 2x_2 &= 12 \end{aligned}$$

*a*) Calcular la solución exacta.

*b*) Calcular el radio espectral de la matriz asociada.

*c*) Aplicar el método de Jacobi.

### 4.1.2. Descripción del método de Gauss-Seidel

Si observamos con atención la expresión general de una iteración del algoritmo de Jacobi, observaremos que si procedemos en el orden natural, $i = 1, 2, ..., N$, al calcular $x_i^{(k+1)}$, los valores $x_1^{(k+1)}, x_2^{(k+1)}, ..., x_{i-1}^{(k+1)}$ ya los hemos obtenido. Si el método es convergente, tenemos la esperanza que estos $i-1$ valores estén más cerca de la solución que los anteriores $x_1^{(k)}, x_2^{(k)}, ..., x_{i-1}^{(k)}$. Por lo tanto podemos utilizarlos en lugar de estos en la expresión que sirve para calcular $x_i^{(k+1)}$, quedando

$$x_i^{(k+1)} = \frac{b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(k+1)} - \sum_{j=i+1}^{N} a_{ij} x_j^{(k)}}{a_{ii}}$$

Obtenemos así el llamado método de Gauss-Seidel, que podemos escribir de la forma

$$\sum_{j=1}^{i} a_{ij} x_j^{(k+1)} = b_i - \sum_{j=i+1}^{N} a_{ij} x_j^{(k)}$$

o en forma matricial

$$(D - E)x^{(k+1)} = b + Fx^{(k)}$$

y también

$$x^{(k+1)} = (D - E)^{-1} F x^{(k)} + (D - E)^{-1} b$$

En cada iteración del algoritmo de Gauss-Seidel resolvemos un sistema triangular. Veamos el pseudoc'odigo que es más sencillo incluso que el correspondiente al algoritmo de Jacobi.

Pseudocódigo de una iteración del método de Gauss-Seidel

- For $i = 1, ..., N$ do
- $s \leftarrow b(i)$
  - For $j = 1, ..., N$ do
  - $s \leftarrow s - A(i, j) * x(j)$
  - end
- $x(i) \leftarrow x(i) + \frac{s}{A(i,i)}$
- end

donde hemos utilizado la expresión equivalente

$$\delta = x_i^{(k+1)} - x_i^{(k)} = \frac{b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(k+1)} - \sum_{j=i}^{N} a_{ij} x_j^{(k)}}{a(i,i)}$$

$$x_i^{(k+1)} = x_i^{(k)} + \delta$$

### 4.1.3. Métodos de relajación

Se pueden generalizar los dos métodos anteriores de Jacobi y Gauss-Seidel, introduciendo un parámetro $\omega > 0$. Sea $x_i^{(k)}$ ya calculado y $\hat{x}_i^{(k+1)}$ obtenido a partir de $x_i^{(k)}$ por uno de los dos métodos precedentes. Se define entonces la combinación lineal

$$x_i^{(k+1)} = \omega \hat{x}_i^{(k+1)} + (1-\omega) x_i^{(k)}$$

Si el método de partida es el de Jacobi, obtenemos para $i = 1, ..., N$

$$x_i^{(k+1)} = \frac{\omega}{a_{ii}} (b_i - \sum_{j \neq i} a_{ij} x_j^{(k)}) + (1-\omega) x_i^{(k)}$$

o bien multiplicando por $a_{ii}$

$$a_{ii} x_i^{(k+1)} = \omega (b_i - \sum_{j \neq i} a_{ij} x_j^{(k)}) + (1-\omega) a_{ii} x_i^{(k)}$$

y con notación matricial

$$Dx^{(k+1)} = (1-\omega) Dx^{(k)} + \omega b + \omega (E+F) x^{(k)}$$

y también

$$x^{(k+1)} = (I - \omega D^{-1} A) x^{(k)} + \omega D^{-1} b$$

En el caso del método de Gauss-Seidel, el correspondiente método de relajación se llama S.O.R. (Succesive Over Relaxation) y se escribe

$$x_i^{(k+1)} = \frac{\omega}{a_{ii}} (b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(k+1)} - \sum_{j=i+1}^{N} a_{ij} x_j^{(k)}) + (1-\omega) x_i^{(k)}$$

y con notación matricial

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

es decir

$$x^{(k+1)} = (\frac{D}{\omega} - E)^{-1} b + (\frac{D}{\omega} - E)^{-1} (\frac{1-\omega}{\omega} D + F) x^{(k)}$$

Pseudocódigo de una iteración del método de S.O.R.

- For $i = 1, ..., N$ do
- $s \leftarrow b(i)$
  - For $j = 1, ..., N$ do
  - $s \leftarrow s - A(i,j) * x(j)$
  - end
- $x(i) \leftarrow x(i) + \omega \frac{s}{A(i,i)}$
- end

### 4.1.4. Control de parada de las iteraciones

Designemos mediante $r^{(k)}$ al vector residuo correspondiente a la iteración $k$-ésima, es decir,

$$r^{(k)} = b - Ax^{(k)}$$

Un posible control de parada consiste en parar en la $k$-ésima iteración si

$$\frac{||r^{(k)}||}{||b||} \leq \varepsilon$$

para $\varepsilon$ elegido convenientemente pequeño.

Esta relación implica que el error $e^{(k)} = x - x^{(k)}$ verifica

$$\frac{||e^{(k)}||}{||x||} \leq \varepsilon cond(A)$$

siendo $x$ la solución exacta de $Ax = b$. En efecto, como

$$||e^{(k)}|| = ||A^{-1}r^{(k)}|| \leq ||A^{-1}||||r^{(k)}||$$

entonces

$$||e^{(k)}|| \leq \varepsilon ||A^{-1}||.||b|| \leq \varepsilon ||A^{-1}||.||Ax|| \leq \varepsilon cond(A)||x||$$

y de ahí la afirmación realizada.

**Comentarios**

(a) En el método de Jacobi, aparece explícitamente el residuo $r^{(k)} = (r_i^{(k)})$ con $r_i^{(k)} = b_i - \sum_{j=1}^{N} a_{ij}x_j^{(k)}$.

(b) Sin embargo en el método de Gauss Seidel o S.O.R. no aparece el residuo sino el vector $\tilde{r}^{(k)} = (r_i^{(k,i)})$ donde

$$r_i^{(k,i)} = b_i - \sum_{j=1}^{i-1} a_{ij}x_j^{(k+1)} - \sum_{j=i}^{N} a_{ij}x_j^{(k)}$$

Podemos realizar el control de parada con $\tilde{r}^{(k)}$, es decir parar las iteraciones si

$$\frac{||\tilde{r}^{(k)}||}{||b||} \leq \varepsilon$$

lo que evita cálculos suplementarios.

A menudo otro control de parada consiste en interrumpir las iteraciones cuando

$$||x^{(k)} - x^{(k-1)}|| \leq \varepsilon ||x^{(k)}||$$

que es un control cómodo desde el punto de vista del cálculo. Presenta sin embargo el inconveniente que podría darse en ciertos casos en los que se verificase el control sin que $x^{(k)}$ estuviese cerca de la solución $x$. Por ejemplo si para algún $k$ resulta $x^{(k)} = 0$ sin ser ésta la solución buscada.

### 4.1.5. Métodos por bloques

Los métodos precedentes se pueden generalizar a métodos por bloques.

Supongamos que $A$ y el correspondiente sistema de ecuaciones lo descomponemos en bloques

$$\begin{bmatrix} A_{11} & ... & A_{1p} \\ & ... & \\ A_{p1} & ... & A_{pp} \end{bmatrix} \begin{bmatrix} X_1 \\ ... \\ X_p \end{bmatrix} = \begin{bmatrix} B_1 \\ ... \\ B_p \end{bmatrix}$$

donde $A_{11}, ...A_{pp}$ son matrices cuadradas y donde $X_i, B_i$ son vectores cuya dimensión es igual al orden de las matrices $A_{ii}$.

De manera análoga a la efectuada precedentemente se define la descomposición de $A$ por bloques

$$A = D - E - F$$

donde

- $D$, es la parte diagonal de bloques de $A$, $D_{ii} = A_{ii}, \quad i = 1, ..., N$
- $-E$, es la parte estrictamente triangular inferior

$$\begin{cases} (-E)_{ij} = A_{ij} & i > j \\ (-E)_{ij} = 0 & i \leq j \end{cases}$$

- $-F$, es la parte estrictamente triangular superior

$$\begin{cases} (-F)_{ij} = A_{ij} & i < j \\ (-F)_{ij} = 0 & i \geq j \end{cases}$$

Por ejemplo el método de relajación por bloques será

$$A_{ii}(X_i^{(k+1)} - X_i^{(k)}) = \omega(B_i - \sum_{j=1}^{i-1} A_{ij} X_j^{(k+1)} - \sum_{j=i}^{p} A_{ij} X_j^{(k)})$$

Para calcular las componentes del vector $X_i^{(k+1)}$, es necesario resolver el sistema $A_{ii} X_i^{(k+1)} = C_i$ donde $C_i$ es un vector conocido. Es necesario que esto sea relativamente sencillo. Por ejemplo, si utilizamos una factorización $A_{ii} = LU$, bastará hacerla una sola vez para cada bloque $A_{ii}$ y esta factorización se utilizará en cada iteración.

## 4.2. Estudio de la convergencia de los métodos de Jacobi, Gauss-Seidel y S.O.R.

Los métodos anteriores son de la forma general

$$x^{(k+1)} = Bx^{(k)} + c$$

La condición necesaria y suficiente de convergencia es

$$\rho(B) < 1$$

Para el método de Jacobi

$$B = D^{-1}(E + F)$$

Para el método de Gauss-Seidel

$$B = (D - E)^{-1}F$$

Para el método S.O.R.

$$B = (\frac{D}{\omega} - E)^{-1}(\frac{1-\omega}{\omega}D + F)$$

Estos matrices se pueden expresar en función de $L = D^{-1}E$ y de $U = D^{-1}F$ que son respectivamente dos matrices triangulares inferior y superior respectivamente con diagonal nula

$$L = \begin{bmatrix} 0 & 0 & ... & 0 \\ -\frac{a_{21}}{a_{22}} & 0 & ... & 0 \\ & ... & ... & \\ -\frac{a_{N1}}{a_{NN}} & -\frac{a_{N2}}{a_{NN}} & ... & 0 \end{bmatrix} \quad U = \begin{bmatrix} 0 & -\frac{a_{12}}{a_{11}} & ... & -\frac{a_{1N}}{a_{11}} \\ 0 & 0 & ... & -\frac{a_{2N}}{a_{22}} \\ & ... & & \\ 0 & 0 & ... & 0 \end{bmatrix}$$

Tenemos pues para el método de Jacobi

$$I = D^{-1}(E + F) = D^{-1}E + D^{-1}F = L + U$$

para el método de Gauss-Seidel

$$\mathcal{L}_\infty = (D - E)^{-1}F = (I - D^{-1}E)D^{-1}F = (I - L)^{-1}U$$

y para el método S.O.R.

$$\mathcal{L}_\omega = (D - \omega E)^{-1}((1-\omega)D + \omega F) = (I - \omega D^{-1}E)^{-1}((1-\omega)I + \omega D^{-1}F) = (I - \omega L)^{-1}((1-\omega)I + \omega U)$$

**Observación:** $D^{-1}A = D^{-1}(D - E - F) = I - L - U$.

Vamos a ver una condición necesaria para que el radio espectral de la matriz del método S.O.R. sea menor que la unidad.

**Teorema 4.2** Para toda matriz $A$, el radio espectral de la matriz del método de relajación S.O.R. es superior o igual a $|\omega - 1|$ en consecuencia una condición necesaria para que el método sea convergente es $0 < \omega < 2$.

**Demostración:** Los valores propios de la matriz $\mathcal{L}_\omega$ del método de relajación verifican la relación

$$\Pi_{i=1}^N \lambda_i(\mathcal{L}_\omega) = det(\mathcal{L}_\omega) = \frac{det(\frac{1-\omega}{\omega}D + F)}{det(\frac{D}{\omega} - E)} = \frac{(\frac{1-\omega}{\omega})^N \Pi a_{ii}}{(\frac{1}{\omega})^N \Pi a_{ii}} = (1-\omega)^N$$

y como por otra parte

$$\rho(\mathcal{L}_\omega) \geq |\lambda_i|$$

lo que implica

$$\rho^N(\mathcal{L}_\omega) \geq \Pi_{i=1}^N |\lambda_i| = |\omega - 1|^N$$

resulta finalmente

$$\rho(\mathcal{L}_\omega) \geq |\omega - 1|$$

■

**Corolario 4.1** Para toda matriz $A$, una condición necesaria de convergencia del método de S.O.R. es

$$0 < \omega < 2$$

### 4.2.1. Matrices a diagonal dominante

**Definición 4.1** Una matriz $A$ de orden $N$ se llama estrictamente diagonal dominante si

$$|a_{ii}| > \sum_{\substack{j=1 \\ j \neq k}}^{N} |a_{ij}| \quad \text{para} \quad i = 1, ..., N$$

**Teorema 4.3** Si $A$ es una matriz de orden $N$ estrictamente diagonal dominante entonces es no singular.

**Demostración:** Consideremos el sistema de ecuaciones

$$Ax = 0$$

y veamos que tiene como única solución $x = 0$.

Por reducción al absurdo, supongamos que $x = [x_1, ..., x_N]^t$ es una solución distinta de cero. En este caso para algún $k$, $0 < |x_k| = \text{máx}_{1 \leq j \leq N} |x_j|$

Como $\sum_{j=1}^N a_{ij} x_j = 0$ para todo $i = 1, ..., N$, tomando $i = k$ resulta

$$a_{kk} x_k = - \sum_{\substack{j=1 \\ j \neq k}}^{N} a_{kj} x_j$$

de donde

$$|a_{kk}||x_k| \leq \sum_{\substack{j=1 \\ j \neq k}}^{N} |a_{kj}||x_j|$$

es decir

$$|a_{kk}| \leq \sum_{\substack{j=1 \\ j \neq k}}^{N} |a_{kj}| \frac{|x_j|}{|x_k|} \leq \sum_{\substack{j=1 \\ j \neq k}}^{N} |a_{kj}|$$

en contradición con la propiedad de $A$ de ser estrictamente diagonal dominante.■

**Teorema 4.4** Sea $A$, matriz de orden $N$ estrictamente diagonal dominante. Entonces el método de Jacobi para resolver un sistema de ecuaciones lineales asociado a dicha matriz es convergente.

**Demostración:** La matriz de iteración para el método de Jacobi es $B = D^{-1}(E + F) = L + U$. Demostraremos que $||B||_\infty < 1$. En efecto,

$$B = L + U = \begin{bmatrix} 0 & -\frac{a_{12}}{a_{11}} & \dots & \dots & -\frac{a_{1N}}{a_{11}} \\ -\frac{a_{21}}{a_{22}} & 0 & \dots & \dots & -\frac{a_{2N}}{a_{22}} \\ & \dots & \dots & \dots & \\ -\frac{a_{N1}}{a_{NN}} & -\frac{a_{N2}}{a_{NN}} & \dots & \dots & 0 \end{bmatrix}$$

de donde

$$||B||_\infty = \max_{1 \leq i \leq N} \sum_{\substack{j=1 \\ j \neq k}}^{N} |a_{ij}/a_{ii}| = \max_{1 \leq i \leq N} \frac{\sum_{j \neq i} |a_{ij}|}{|a_{ii}|} < 1$$

pues $A$ es estrictamente diagonal dominante. ■

**Teorema 4.5** Sea $A$ una matriz estrictamente diagonal dominante, entonces el método de Gauss-Seidel para resolver un sistema de ecuaciones lineales asociado a dicha matriz es convergente.

**Demostración:** La matriz asociada a la iteración de Gauss-Seidel es

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

Para determinar el radio espectral de $\mathcal{L}_1$, calcularemos primero los valores propios, es decir, las raíces del polinomio característico

$$p(\lambda) = det(\lambda I - \mathcal{L}_1) = 0$$

Observando que $det(I - L) = 1$ resulta

$$\begin{aligned} p(\lambda) &= det(I - L)det(\lambda I - \mathcal{L}_1) \\ &= det(I - L)det(\lambda I - (I - L)^{-1}U) \\ &= det(\lambda(I - L) - U) \\ &= det(\lambda(I - L - \frac{U}{\lambda})) \\ &= \lambda^N det(I - L - \frac{U}{\lambda}) \end{aligned}$$

de donde $p(\lambda) = 0$ si $\lambda = 0$ o bien si $det(I - L - \frac{U}{\lambda}) = 0$.

Queremos demostrar que todas las raíces de $p(\lambda) = 0$ verifican $|\lambda| < 1$. Supongamos por reducción al absurdo que existe al menos una raiz $\lambda$ tal que $|\lambda| \geq 1$. Entonces por una parte $det(I - l - \frac{U}{\lambda}) = 0$ y por otra parte como $A = D - E - F$ es estrictamente diagonal dominante, también lo es $I - L - U$ y lo será también $I - L - \frac{U}{\lambda}$ si $|\lambda| \geq 1$. Por lo tanto $I - L - \frac{U}{\lambda}$ es no singular en contradicción con $det(I - L - \frac{U}{\lambda}) = 0$. ■

### 4.2.2. Matrices hermíticas y definidas postivas

**Teorema 4.6** Sea $A$ hermítica y definida positiva, descompuesta de la forma $A = M - N$ con $M$ no singular. Si la matriz hermítica $M^* + N$ es definida positiva, entonces $B = M^{-1}N = I - M^{-1}A$ verifica $\rho(B) < 1$.

**Demostración:** $M^*+N$ es efectivamente hermítica, en efecto, $M^*+N = A^*+N^*+N = A+N^*+N = M + N^*$.

Vamos a demostrar que $||M^{-1}N|| < 1$ para la norma matricial subordinada a la norma vectorial

$$||.|| : v \longrightarrow (Av, v)^{1/2}$$

que es una norma pues $A$ es hermítica y definida positiva.

Como

$$||M^{-1}N|| = ||I - M^{-1}A|| = \sup_{||v||=1} ||v - M^{-1}Av||$$

y la aplicación

$$v \longrightarrow ||v - M^{-1}Av||$$

definida en el conjunto compacto $\{v \in \mathcal{C}^N; ||v|| = 1\}$ es continua, alcanza su valor máximo para algún $v$ determinado de dicho conjunto. Veamos ahora que para $||v|| = 1$, $||v-M^{-1}Av|| < 1$. En efecto, denotamos $w = M^{-1}Av$ y calculemos $||v - w||$

$$||v - M^{-1}Av||^2 = ||v - w||^2 = (Av, v) - (Av, w) - (Aw, v) + (Aw, w)$$

teniendo en cuenta que $Av = Mw$ y que $A$ es hermítica, resulta $(Av, w) = (Mw, w)$ y $(Aw, v) = (M^*w, w)$, de donde

$$||v - M^{-1}Av||^2 = 1 - ((M^* + N)w, w)$$

y puesto que si $v \neq 0$ y $M$ y $A$ son no singulares, $w = M^{-1}Av \neq 0$, además como $M^* + N$ es definida positiva $((M^* + N)w, w) > 0$ resultando finalmente $||v - M^{-1}Av|| < 1$ y $\rho(M^{-1}N) < 1$. ■

**Corolario 4.2** Si $A$ es hermítica y definida positiva el método S.O.R. por punto o por bloques converge si y solo si $0 < \omega < 2$.

**Demostración:** En el método S.O.R. la descomposición $A = M - N$ es $M = \frac{D}{\omega} - E$, $N = \frac{1-\omega}{\omega}D + F$. De manera que

$$M^* + N = \frac{D^*}{\omega} - E^* + \frac{1-\omega}{\omega}D + F$$

como $A$ es hermítica $D = D^*$ y $F = E^*$ de modo que

$$M^* + N = \frac{2-\omega}{\omega}D$$

Solo resta verificar que $D$ es definida positiva. En efecto, sean $A_{ii}, i = 1, ..., p$ los bloques de $D$. Cada bloque es definido positivo pues, para $v_i \neq 0$

$$(A_{ii}v_i, v_i) = (A\tilde{v_i}, \tilde{v_i}) > 0$$

donde $\tilde{v_i}$ se obtiene prolongando $v_i$ con coordenadas nulas hasta obtener el correspondiente vector de dimensión $N$. Finalmente

$$(Dv, v) = \sum_{i=1}^{p}(A_{ii}v_i, v_i) > 0$$

La matriz $M^* + N = \frac{2-\omega}{\omega}D$ es pues definida positiva si y solo si $0 < \omega < 2$. ■

**Ejercicios**

(a) Considerar la matriz

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

que aperece al aproximar mediante diferencias finitas la ecuación diferencial

$$\begin{aligned} -u'' &= f \quad \text{en} \quad [a, b] \\ u(a) &= u(b) = 0 \end{aligned}$$

Demostrar que el método S.O.R. con $0 < \omega < 2$, y el método de Jacobi convergen al resolver sistemas asociados a ésta matriz.

### 4.2.3. Comparación de los métodos de Jacobi y Gauss-Seidel. Búsqueda del parámetro de relajación óptimo en el método S.O.R.

**Teorema 4.7** Sea $A$ una matriz tridiagonal con bloques diagonales no singulares. Entonces los radios espectrales de las matrices de iteración del método de Jacobi y Gauss-Seidel por bloques están relacionados por

$$\rho(\mathcal{L}_1) = \rho^2(J)$$

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

Empezaremos demostrando un lema.

**Lema 4.1** Sea $A$ una matriz tridiagonal por bloques de orden $N$.

$$A = \begin{bmatrix} A_{11} & A_{12} & 0 & ... & ... & 0 \\ A_{21} & A_{22} & A_{23} & 0 & ... & 0 \\ 0 & A_{32} & A_{33} & A_{34} & ... & 0 \\ ... & ... & ... & ... & ... & ... \\ 0 & ... & ... & ... & A_{p-1,p-1} & A_{p-1,p} \\ 0 & ... & ... & ... & A_{p,p-1} & A_{p,p} \end{bmatrix}$$

donde cada bloque $A_{ii}$ es una matriz de orden $n_i$ y $\sum_{i=1}^{p} n_i = N$. Sea ahora la descomposición por bloques $A = D - E - F$ donde $D$ es la diagonal de bloques

$$\begin{bmatrix} A_{11} & 0 & ... & 0 \\ 0 & A_{22} & ... & 0 \\ ... & ... & ... & ... \\ 0 & ... & 0 & A_{pp} \end{bmatrix}$$

y $-E$ y $-F$ la parte triangular inferior y superior respectivamente excluida la diagonal. Sea $\mu$ un número no nulo y llamemos $A(\mu) = D - \mu E - \frac{1}{\mu} F$ entonces

$$det(A(\mu)) = det(A)$$

**Demostración:** Tenemos

$$A(\mu) = \begin{bmatrix} A_{11} & \frac{1}{\mu}A_{12} & 0 & ... & ... & 0 \\ \mu A_{21} & A_{22} & \frac{1}{\mu}A_{23} & 0 & ... & 0 \\ 0 & \mu A_{32} & A_{33} & \frac{1}{\mu}A_{34} & ... & 0 \\ ... & ... & ... & ... & ... & ... \\ 0 & ... & ... & ... & A_{p-1,p-1} & \frac{1}{\mu}A_{p-1,p} \\ 0 & ... & ... & ... & \mu A_{p,p-1} & A_{p,p} \end{bmatrix}$$

y consideremos la matriz

$$S = \begin{bmatrix} \mu I_1 & 0 & ... & 0 \\ 0 & \mu^2 I_2 & ... & 0 \\ ... & ... & ... & ... \\ 0 & ... & 0 & \mu^p I_p \end{bmatrix}$$

donde $I_i$ es la matriz unidad de orden $n_i$.

$$det(S) = det(\mu I_1) det(\mu^2 I_2) .... det(\mu^p I_p) = \mu^{(n_1 + 2n_2 + ... + pn_p)}$$

por otra parte

$$S^{-1} = \begin{bmatrix} \mu^{-1}I_1 & 0 & ... & 0 \\ 0 & \mu^{-2}I_2 & ... & 0 \\ ... & ... & ... & ... \\ 0 & ... & 0 & \mu^{-p}I_p \end{bmatrix}$$

$$det(S^{-1}) = \mu^{-(n_1+2n_2+...+pn_p)}$$

Veamos que $A(\mu) = SAS^{-1}$, en efecto

$$(SAS^{-1})_{ij} = \sum_{k,l=1}^{p} S_{il}A_{kl}S^{-1}_{lj} = \frac{\mu^i}{\mu^j}A_{ij}$$

de modo que

$$\begin{aligned} &\text{Para} \quad j = i \quad (SAS^{-1})_{ii} = A_{ii} \\ &\text{Para} \quad j = i-1 \quad (SAS^{-1})_{i,i-1} = \mu A_{i,i-1} \\ &\text{Para} \quad j = i+1 \quad (SAS^{-1})_{i,i+1} = \frac{1}{\mu}A_{i,i+1} \end{aligned}$$

de donde finalmente

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

■

**Demostración del teorema**

- La matriz de iteración del método de Jacobi es $J = D^{-1}(E+F)$ y sus valores propios son las raíces de la ecuación

  $$P_J(\lambda) = det(\lambda I - D^{-1}(E+F)) = 0$$

  y también

  $$det(D)P_J(\lambda) = det(D)det(\lambda I - D^{-1}(E+F)) = det(\lambda D - E - F) = 0$$

  por lo tanto los valores propios de $J$ son las raíces de la ecuación

  $$det(\lambda D - E - F) = 0$$

- La matriz de la iteración del método de Gauss-Seidel es $\mathcal{L}_1 = (D-E)^{-1}F$ y sus valores propios son las raíces de la ecuación

  $$P_{\mathcal{L}_1}(\lambda) = det(\lambda I - (D-E)^{-1}F) = 0$$

  y también

  $$det(D-E)P_{\mathcal{L}_1}(\lambda) = det(D-E)det(\lambda I - (D-E)^{-1}F) = det(\lambda D - \lambda E - F) = 0$$

de donde

$$det(D-E)P_{\mathcal{L}_1}(\lambda) = det(\sqrt{\lambda}(\sqrt{\lambda}D - \sqrt{\lambda}E - \frac{1}{\sqrt{\lambda}}F)) = 0$$

es decir, los valores propios de $\mathcal{L}_1$ son las raíces de la ecuación

$$(\sqrt{\lambda})^N det(\sqrt{\lambda}D - E - F) = 0$$

donde hemos aplicado el lema anterior.

Comparando las dos ecuaciones, la que nos da los valores propios de $J$ y de $\mathcal{L}_1$ vemos que si $\lambda$ es un valor propio de $\mathcal{L}_1$ entonces $\pm\sqrt{\lambda}$ lo es de $J$ y también si $\alpha$ es valor propio de $J$, $\alpha^2$ lo es de $\mathcal{L}_1$ y de ahí que

$$\rho(\mathcal{L}_1) = \rho^2(J)$$

■

**Teorema 4.8** : Comparación de los métodos de Jacobi y S.O.R.

Sea $A$ una matriz tridiagonal por bloques con $det(A_{ii}) \neq 0$ y tal que todos los valores propios de la matriz de iteración de Jacobi sean reales. Entonces el método de Jacobi por bloques y el método S.O.R. por bloques para $0 < \omega < 2$ convergen o divergen simultáneamente. Cuando convergen, la función

$$\omega \in ]0,2[ \longrightarrow \rho(\mathcal{L}_\omega)$$

tiene la forma indicada en la figura, donde $\omega_{opt} = \frac{2}{1+\sqrt{1-\rho^2(J)}}$

**Demostración:** Hemos visto anteriormente que para el método de Jacobi, los valores propios $\lambda$ de la matriz de iteración $J = D^{-1}(E+F)$ están caracterizados por

$$P_J(\lambda) = 0 \Leftrightarrow det(\lambda D - E - F) = 0$$

Por otra parte para el método S.O.R. los valores propios $\mu$ de la matriz de iteración

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

son las soluciones de

$$P_{\mathcal{L}_\omega}(\mu) = det(\mu I - (\frac{D}{\omega} - E)^{-1}(\frac{1-\omega}{\omega}D + F)) = 0$$

y de

$$det(\frac{D}{\omega} - E)P_{\mathcal{L}_\omega}(\mu) = det(\mu(\frac{D}{\omega} - E) - (\frac{1-\omega}{\omega}D + F)) = 0$$

o bien de

$$det(\frac{\mu+\omega-1}{\omega}D - \mu E - F) = 0$$

sacando factor común $\sqrt{\mu}$ y aplicando el lema anterior

$$\mu^{N/2} det(\frac{\mu+\omega-1}{\sqrt{\mu}\omega} D - E - F) = 0$$

En resumen: Si $\lambda$ es valor propio de $J$ entonces $\mu$ es valor propio de $\mathcal{L}_\omega$ si se verifica la relación

$$\frac{\mu+\omega-1}{\sqrt{\mu}\omega} = \lambda$$

es decir

$$(\mu+\omega-1)^2 = \mu\omega^2\lambda^2$$

Vamos a estudiar la función $\mu = \mu(\omega, \lambda)$ para cada valor de $\lambda \geq 0$. $\mu$ es solución de la ecuación de segundo grado

$$\mu^2 - (\omega^2\lambda^2 - 2\omega + 2)\mu + (\omega-1)^2 = 0 \tag{4.1}$$

cuyas raíces son

$$\mu_1 = \frac{(\omega^2\lambda^2 - 2\omega + 2) + \omega\lambda\sqrt{\omega^2\lambda^2 - 4\omega + 4}}{2}$$

y

$$\mu_2 = \frac{(\omega^2\lambda^2 - 2\omega + 2) - \omega\lambda\sqrt{\omega^2\lambda^2 - 4\omega + 4}}{2}$$

(a) consideramos el caso $\rho(J) \geq 1$, es decir, existe al menos un valor propio $\lambda \geq 1$. Entonces

$$\begin{aligned} \mu_1 &\geq \frac{1}{2}(\omega^2 - 2\omega + 2) + \frac{\omega}{2}\sqrt{\omega^2 - 4\omega + 4} \\ &= \frac{\omega^2}{2} - \omega + 1 + \omega - \frac{\omega^2}{2} = 1 \end{aligned}$$

por lo tanto $\rho(\mathcal{L}_\omega) \geq 1$.

(b) Consideramos el caso $\rho(J) < 1$. Es decir para todo valor propio $\lambda \geq 0$, $\lambda < 1$. Estudiemos el valor de las raíces $\mu_1$ y $\mu_2$ en función de $\omega$ para $\lambda \geq 0$ fijo. Empecemos estudiando para qué valores de $\omega$ las raíces son reales o complejas. Para ello observemos que las raíces del discriminante de la ecuación de segundo grado (4.1) son

$$\omega_0 = \frac{2 - \sqrt{4 - 4\lambda^2}}{\lambda^2} = \frac{2}{1 + \sqrt{1 - \lambda^2}}$$

$$\omega_1 = \frac{2 + \sqrt{4 - 4\lambda^2}}{\lambda^2} = \frac{2}{1 - \sqrt{1 - \lambda^2}}$$

(ver figura)

Tenemos pues, $1 < \omega_0 < 2 < \omega_1$.

*a*) Estudiemos $\mu(\omega, \lambda)$ para $\omega_0 < \omega < 2$. En este caso $\mu_1$ y $\mu_2$ son complejos conjugados y $|\mu_1| = |\mu_2|$. Como $|\mu_1||\mu_2| = (\omega - 1)^2$, resulta $|\mu_1| = |\mu_2| = \omega - 1$. (Ver figura)

*b*) Para $0 < \omega < \omega_0$, $\mu_1$ y $\mu_2$ son reales y consideremos $\mu_1$ que es el valor propio de mayor valor absoluto. Los valores de $\mu_1$ y $\mu_1'$ (donde la derivada es respecto a $\omega$) en función de distintos valores $\omega$ son para un valor fijo de $\lambda$ son

| $\omega$ | 0 | 1 | $\omega_0$ |
|---|---|---|---|
| $\mu$ | 1 | $\lambda^2$ | $\omega_0 - 1$ |
| $\mu_1'$ | $\lambda - 1$ | | $-\infty$ |

**Ejercicios**

(a) Considerar un sistema de ecuaciones

$$Ax = b$$

asociado a la matriz

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

*a*) Calcular la matriz $\mathcal{J}$ correspondiente a la iteración de Jacobi .

*b*) Calcular para qué valores de $a$ el método de Jacobi será convergente.

*c*) Calcular la matriz $\mathcal{L}_\omega$ correspondiente al método S.O.R. .

*d*) ¿Para qué valores del parámetro de relajación $\omega$ y para qué valores de $a$ el método S.O.R. será convergente ?

*e*) Calcular el paraḿetro óptimo $\omega_{op}$ del método S.O.R. en función de $a$. ¿ Cuál es el radio espectral $\rho_{\mathcal{L}_{\omega_{op}}}$ correspondiente al método S.O.R. con este valor óptimo del parámetro de relajación ?

(b) Considerar el siguiente sistema de ecuaciones

$$\begin{aligned} 4x_1 + 3x_2 &= 24 \\ 3x_1 + 4x_2 - x_3 &= 30 \\ -x_2 + 4x_3 &= -24 \end{aligned} \tag{4.2}$$

*a*) Hallar la matriz de iteración correspondiente al método de Jacobi.

*b*) Demostrar que el método de Jacobi es convergente para resolver este sistema.

*c*) Estimar la velocidad asintótica de convergencia del método de Jacobi.

*d*) Estimar la velocidad asintótica de convergencia del método de Gauss-Seidel.

*e*) Calcular el valor óptimo del parámetro de relajación $\omega$ del correspondiente método S.O.R.

*f*) Estimar la velocidad asintótica de convergencia del método S.O.R. con parámetro óptimo.

*g*) Estimar el número mínimo de iteraciones que hay que realizar con el método de Jacobi, con el método de Gauss-Seidel y con el método S.O.R. con parámetro óptimo para obtener la solución con un error relativo menor que 0,001.

*h*) Resolver el sistema utilizando los tres métodos anteriores y comparar los resultados.

# Capítulo 5

# Optimización sin restricciones

## 5.1. Fundamentos de la optimización

### 5.1.1. Introducción

En el espacio euclídeo $\mathbb{R}^d$ con la norma euclídea que denotamos $||.||$ y que deriva del correspondiente producto escalar que denotaremos mediante $(.,.)$, consideremos

- $K$ un subconjunto cerrado de $\mathbb{R}^d$
- $J : K \longrightarrow \mathbb{R}$ una función sobre $K$
- Consideraremos el problema siguiente: Hallar $u \in K$ tal que

$$J(u) = \inf_{v \in K} J(v)$$

- Si $K = \mathbb{R}^d$ el problema anterior se llama de optimización sin restricciones.
- ¿ En qué condiciones el problema anterior tiene solución?
- Si tiene solución ¿ es ésta única?
- Se dice que $u$ realiza el mínimo global de $J$.

Vamos a tratar de responder a las anteriores preguntas.

**Teorema 5.1** de Weierstrass

Sea $K$ un conjunto compacto de $\mathbb{R}^d$, no vacío y $J : K \longrightarrow \mathbb{R}$ continua en $K$. Entonces el problema:

Hallar $u \in K$ tal que

$$J(u) = \inf_{v \in K} J(v)$$

tiene solución.

**Demostración:**

$K$ compacto y $J$ continua implica que $J(K)$ es compacto en $\mathbb{R}$, por tanto $J(K)$ está acotado inferiormente y existe un extremo inferior, es decir, existe un número $\alpha > -\infty$ tal que $\alpha = \text{ínf}_{v \in K} J(v)$.

Consideremos una sucesión minimizante $(u_n)_n$, es decir, $u_n \in K$ tal que $J(u_n) \to^{n \to \infty} \alpha$. Una tal sucesión la podemos siempre construir tomando por ejemplo $0 < \varepsilon < 1$, $u_n \in K$ tal que

$$\alpha \leq J(u_n) \leq \alpha + \varepsilon^n$$

$(u_n)_n$ es una sucesión en el compacto $K$, por lo tanto existe una subsucesión convergente $(u_\nu)_\nu$. Sea $u$ el límite de $(u_\nu)$. Gracias a la continuidad de $J(.)$ tendremos $\alpha = \text{lím}_{\nu \to \infty} J(u_\nu) = J(u)$. ■

Una variante del anterior teorema de Weierstrass es el siguiente:

**Teorema 5.2** Sea $J : \mathbb{R}^d \to \mathbb{R}$ continua, verificando la propiedad (coercividad)

$$\lim_{||v|| \to \infty} J(v) = \infty$$

Entonces existe $u \in \mathbb{R}^d$ tal que

$$J(u) = \inf_{v \in \mathbb{R}} J(v)$$

**Demostración:**

Llamemos $\alpha = \text{ínf}_{v \in \mathbb{R}^d} J(v)$. En principio $\alpha$ pordría tomar el valor $-\infty$. Consideremos una sucesión $(u_n)_n$ minimizante de elementos de $\mathbb{R}^d$, es decir,

$$J(u_n) \longrightarrow \inf_{v \in \mathbb{R}^d} J(u) = \alpha$$

$$\alpha \leq J(u_n) \leq \alpha + \varepsilon^n$$

si $\alpha$ es finito y si $\alpha = -\infty$ entonces tomaremos $u_n$ de modo que $J(u_n) \to -\infty$.

La sucesión $(u_n)_n$ está acotada, pues si no fuera así, resultaría

$$\lim_{||u_n|| \to \infty} J(u_n) = \infty$$

en contra de la elección de $(u_n)_n$.

Podemos pues extraer una subsucesión convergente. Sea $(u_\nu)_\nu$ tal que $\lim u_\nu = u$; como $J$ es continua resulta $\lim J(u_\nu) = J(u)$ y $\alpha$ es un número finito. ■

**Ejercicio:** Dar 3 contraejemplos en los que el problema anterior no tenga solución debido a que:

(a) $K$ no sea compacto.

(b) $J$ no sea continua.

(c) $J$ no sea coerciva.

### 5.1.2. Extremos relativos y diferenciabilidad

En este apartado consideraremos la noción de extremo relativo y la relacionaremos con las nociones de diferencial.

**Definición 5.1** Sea $A \subset \mathbb{R}^d$ un conjunto abierto y $J : A \longrightarrow \mathbb{R}$. Se dice que $J(.)$ tiene un mínimo relativo en $u \in A$ si existe un entorno $U$ de $u$ tal que

$$\forall v \in U \quad J(u) \leq J(v)$$

Análogamente $J(.)$ tiene un máximo relativo en $u \in A$ si existe un entorno $U$ de $u$ tal que

$$\forall v \in U \quad J(u) \geq J(v)$$

**Teorema 5.3** Condición necesaria de extremo relativo

$A$ abierto de $\mathbb{R}^d$, $J : A \longrightarrow \mathbb{R}$ diferenciable en $A$. Si $J(.)$ tiene un extremo relativo en $u \in A$ (máximo o mínimo) entonces la diferencial de $J(.)$ en $u$ es la aplicación nula, es decir, $DJ(u) = 0$, o bien identificando $DJ(u)$ con un elemento de $\mathbb{R}^d$,

$$(DJ(u), v) = 0 \quad \forall v \in \mathbb{R}^d$$

**Nota 5.1** Referido a la base canónica de $\mathbb{R}^d$ el vector de $\mathbb{R}^d$, $DJ(u)$ se escribirá $J'(u)$, es decir, la matriz Jacobiana de $J(.)$ en el punto $u$. en este caso la matriz jacobiana en un matriz fila. La transpuesta de esta matriz es un vector columna que se suele llamar vector gradiente de $J()$ en $u$ y se escribe $\nabla J(u)$. Cuando identifiquemos $DJ(u)$ con un vector de $\mathbb{R}^d$ escribiremos indistintamente $(DJ(u), v)$ o $(\nabla J(u), v)$.

**Demostración:**

- Caso $d = 1$: Sea $A$ abierto de $\mathbb{R}$, $u \in A$, $f : A \subset \mathbb{R} \longrightarrow \mathbb{R}$, $f(u) \leq f(v) \quad \forall v \in U$, entorno de $u$. Tendremos, por ejemplo

$$f'(u) = \lim_{h \to 0^+} \frac{f(u+h) - f(u)}{h} \geq 0$$

$$f'(u) = \lim_{h \to 0^-} \frac{f(u+h) - f(u)}{h} \leq 0$$

de donde $f'(u) = 0$.

- Caso general $d > 1$

  Fijemos $h$ con norma suficientemente pequeña de modo que $u + th \in U \quad \forall t \in ]-1, +1[$. Introducimos ahora la función

$$f : ]-1, +1[ \longrightarrow \mathbb{R}$$
$$t \longrightarrow J(u + th)$$

  Sea $f = J \circ g$ donde

$$g : \mathbb{R} \longrightarrow \mathbb{R}^d$$
$$t \longrightarrow u + th$$

  Si $J$ tiene un mínimo relativo en $u$, es decir, $J(u) \leq J(u + th)$, entonces $f(0) \leq f(t) \quad \forall t \in ]-1, 1[$. Resulta $f(.)$ tiene un mínimo relativo en 0 y por lo tanto $f'(0) = 0$, es decir, aplicando la regla de la cadena

$$f'(t) = DJ(g(t)) \circ Dg(t) = J'(u + th).g'(t) = J'(u + th).h$$

  de modo que

$$f'(0) = J'(u).h = 0$$

  Como la dirección $h$ es cualquiera resulta $J'(u) = 0$. ■

Vamos ahora a tener en cuenta las derivadas segundas, para obtener una condición necesaria y suficiente de mínimo relativo.

**Teorema 5.4** Condición necesaria de mínimo relativo.

Sea $A$ un abierto de $\mathbb{R}^d$, $J : A \longrightarrow \mathbb{R}$ dos veces diferenciable en $u \in A$; Entonces si $J(.)$ admite un mínnmo relativo en $u$, entonces la diferencial segunda de $J(.)$ en $u$ verifica,

$$D^2 J(u)(h, ) \geq 0 \quad \forall h \in \mathbb{R}^d$$

o de manera equivalente, la matriz Hessiana de $J(.)$ en $u$ es semidefinida positiva.

**Demostración:**

Utilizaremos el desarrollo de Taylor con resto de Taylor-Young.

$$J(u + h) = J(u) + DJ(u)(h) + \frac{1}{2} D^2 J(u)(h, h) + \varepsilon(h)||h||^2$$

donde $\varepsilon(h) \stackrel{h \to 0}{\longrightarrow} 0$. Como $DJ(u) = 0$ resulta,

$$0 \leq J(u + h) - J(u) = \frac{1}{2} D^2 J(u)(h, h) + \varepsilon(h)||h||^2$$

Sustituyendo $h$ por $th$ con $t \in \mathbb{R}$ de modo que $u + th \in A$

$$0 \leq J(u + th) - J(u) = \frac{t^2}{2} D^2 J(u)(h, h) + t^2 \varepsilon(h)||h||^2)$$

Dividiendo por $\frac{t^2}{2}$

$$0 \leq D^2 J(u)(h, h) + \varepsilon(h)||h||^2$$

pasando al límite cuando $t \to 0$, $\varepsilon(th) \to 0$ lo que implica $D^2 J(u)(h, h) \geq 0$.■

Vamos a buscar ahora condiciones suficientes de mínimo relativo. Necesitamos primero una definición y un lema.

**Definición 5.2** Sea $B : \mathbb{R}^d \times \mathbb{R}^d \longrightarrow \mathbb{R}$ una aplicación bilineal. Se dice que $B$ es definida positiva si

$$B(v, v) \geq 0 \quad \forall v \in \mathbb{R}^d$$

y

$$B(v, v) = 0, \quad \text{si y solo si } v = 0$$

**Lema 5.1** Sea $B : \mathbb{R}^d \times \mathbb{R}^d \longrightarrow \mathbb{R}$ una aplicación bilineal y definida positiva, entonces existe una constante $\alpha > 0$ tal que

$$B(v, v) \geq \alpha||v||^2$$

**Demostración:**

Consideremos el conjunto $S = \{v \in \mathbb{R}^d;\ ||v|| = 1\}$ que es compacto (cerrado y acotado) y la aplicación $J : v \to J(v) = B(v, v)$ que es continua sobre el compacto $S$, entonces alcanza el mínimo en el compacto (teorema de Weierstrass). Sea $u$ el punto donde alcanza este mínimo, es decir,

$$\alpha = J(u) = \min_{v \in S} J(v)$$

tendremos $\alpha \neq 0$ pues $u \in S \Rightarrow u \neq 0$. Finalmente $\forall v \in \mathbb{R}^d \quad v \neq 0$, $\frac{v}{||v||} \in S$ y

$$J(\frac{v}{||v||}) = B(\frac{v}{||v||}, \frac{v}{||v||}) \geq \alpha \Rightarrow B(v, v) \geq \alpha||v||^2$$

. ■

**Teorema 5.5** Condición suficiente de mínimo relativo.

Sea $A$ abierto de $\mathbb{R}^d$ y $J : A \longrightarrow \mathbb{R}$ dos veces diferenciable, tal que $DJ(u) = 0$ en $u \in A$.

Si la función $J(.)$ es tal que su diferencial segunda en $u$ es definida positiva entonces $J$ tiene un mínimo relativo en $u$.

**Demostración:**

Consideremos el desarrollo de Taylor-Young en un entorno de $u$.

$$J(u+h) = J(u) + DJ(u)(h) + \frac{1}{2}D^2J(u)(h,h) + \varepsilon(h)||h||^2$$

donde $\varepsilon(h) \stackrel{h\to 0}{\longrightarrow} 0$

$$J(u+h) - J(u) = \frac{1}{2}D^2J(u)(h,h) + \varepsilon(h)||h||^2$$

$$J(u+h) - J(u) \geq \frac{1}{2}\alpha||h||^2 + \varepsilon(h)||h||^2$$

$$J(u+h) - J(u) \geq \frac{1}{2}(\alpha + 2\varepsilon(h))||h||^2$$

Para $||h||$ suficientemente pequeño tendremos $\frac{1}{2}(\alpha + 2\varepsilon(h)) > 0$, es decir $J(u+h) - J(u) \geq 0$.■

**Ejercicio:** verificar que $J : A \longrightarrow \mathbb{R}$ donde

$$A = \{(x,y) \in \mathbb{R}^2 \; x^2 + y^2 < 1\}$$

$$J(x,y) = \log(x^2 + y^2 + 1)$$

tiene un mínimo relativo en $(0,0)$.

### 5.1.3. Extremos y convexidad

Vamos a introducir la convexidad en el estudio de los extremos de funciones. Recordemos primero las nociones de conjunto convexo y de función convexa.

**Definición 5.3** Conjunto convexo.

Un conjunto $K \subset \mathbb{R}^d$ se dice que es convexo si

$$\forall u, v \in K, \quad \text{se verifica} \quad \lambda u + (1-\lambda)v \in K \quad \forall \lambda \in [0,1]$$

**Definición 5.4** Función convexa y estrictamente convexa.

Sea $K \subset \mathbb{R}^d$ un conjunto convexo. Una función $J : K \subset \mathbb{R}^d \longrightarrow \mathbb{R}$ se dice que es convexa en $K$, si para todo $\lambda \in [0,1]$ y para todo $u, v \in K$ se verifica

$$J(\lambda u + (1-\lambda)v) \leq \lambda J(u) + (1-\lambda)J(v)$$

Además se dice que la función es estrictamente convexa si para todo $\lambda \in ]0,1[$ y para todo $u, v \in K$ se verifica

$$J(\lambda u + (1-\lambda)v) < \lambda J(u) + (1-\lambda)J(v)$$

Vamos a estudiar ahora algunas propiedades de la funciones convexas y diferenciables.

**Teorema 5.6** Sea $J : A \subset \mathbb{R}^d \longrightarrow \mathbb{R}$, donde $A$ es un abierto de $\mathbb{R}^d$, $J$ diferenciable en $A$. Sea $K$ una parte convexa de $A$.

(a) $J(.)$ es convexa en $K$ si y solo si

$$J(v) \geq J(u) + (\nabla J(u), v - u) \quad \forall u, v \in K$$

(b) $J(.)$ es estrictamente convexa en $K$ si y solo si

$$J(v) > J(u) + (\nabla J(u), v - u) \quad \forall u, v \in K$$

**Ejercicio:** Interpretar geométricamente las anteriores desigualdades.

**Demostración:**

(a) Si $J(.)$ es convexa en $K$ y $u, v \in K$ entonces para $\lambda \in ]0, 1[$

$$u + \lambda(v - u) \in K$$

$$J(u + \lambda(v - u)) \leq (1 - \lambda)J(u) + \lambda J(v)$$

$$J(u + \lambda(v - u)) - J(u) \leq \lambda(J(v) - J(u))$$

$$\frac{J(u + \lambda(v - u)) - J(u)}{\lambda} \leq J(v) - J(u)$$

haciendo $\lambda \to 0^+$

$$(\nabla J(u), v - u) \leq J(v) - J(u)$$

Recíprocamente, sea $J(.)$ verificando

$$J(v) \geq J(u) + (\nabla J(u), v - u) \quad \forall v, u \in K$$

Sea $\lambda \in ]0, 1[$ y tomemos primero en el lugar de $u$, $v + \lambda(u - v)$, resulta

$$J(v) \geq J(v + \lambda(u - v)) - \lambda(\nabla J(v + \lambda(u - v)), u - v)$$

Ahora en el lugar de $v$ tomemos $u$ y en el lugar de $u$ tomemos $v + \lambda(u - v)$, resulta

$$J(u) \geq J(v + \lambda(u - v)) + (1 - \lambda)(\nabla J(v + \lambda(u - v)), u - v)$$

Multiplicando la primera por $(1 - \lambda)$ y la segunda por $\lambda$ y sumando

$$(1 - \lambda)J(v) + \lambda J(u) \geq J(v + \lambda(u - v))$$

es decir

$$J(\lambda u + (1 - \lambda)v) \leq \lambda J(u) + (1 - \lambda)J(v)$$

(b) Si $J(.)$ es estrictamente convexa, el razonamiento anterior no es aplicable pues la desigualdad estricta se pierde al pasar al límite. Procedemos entonces de la siguiente manera: Sea $\lambda \in ]0,1[$ y $\omega \in ]0,1[$ verificando $\omega > \lambda$, resulta

$$u + \lambda(v-u) = u - \frac{\lambda}{\omega}u + \frac{\lambda}{\omega}u + \lambda(v-u) = \frac{\omega-\lambda}{\omega}u + \frac{\lambda}{\omega}(u + \omega(v-u))$$

de donde

$$J(u+\lambda(v-u)) = J(\frac{\omega-\lambda}{\omega}u + \frac{\lambda}{\omega}(u+\omega(v-u)) \leq \frac{\omega-\lambda}{\omega}J(u) + \frac{\lambda}{\omega}J(u+\omega(v-u))$$

$$J(u+\lambda(v-u)) - J(u) \leq \frac{\lambda}{\omega}(J(u+\omega(v-u) - J(u))$$

$$\frac{J((u+\lambda(v-u)) - J(u)}{\lambda} \leq \frac{J(u+\omega(v-u)) - J(u)}{\omega} < \frac{(1-\omega)J(u) + \omega J(v) - J(u)}{\omega} = J(v) - J(u)$$

pasando al límite cuando $\lambda \to 0$, resulta

$$(\nabla J(u), v-u) \leq \frac{J(u+\omega(v-u)) - J(u)}{\omega} < J(v) - J(u)$$

La recíproca es idéntica al caso anterior.

■

**Corolario 5.1** Sea $J : \mathbb{R}^d \to \mathbb{R}$ una función convexa y diferenciable en $u$ tal que $\nabla J(u) = 0$. Entonces $J(.)$ admite un mínimo en $u$.

**Demostración:**

$J(.)$ convexa y diferenciable implica

$$J(v) \geq J(u) + (\nabla J(u), v-u) \quad \forall v \in \mathbb{R}^d$$

es decir,

$$J(v) \geq J(u) \quad \forall v \in \mathbb{R}^d$$

■

**Ejercicios**

(a) Estudiar los puntos críticos (puntos $u$ donde $\nabla J(u) = 0$) de las funciones

- $$J(x,y) = e^{1+x^2+y^2}$$

- $$J(x,y,z) = x^2 + y^2 + z^2 + xy$$
- $$J(x,y) = x^4 + y^4$$
- $$J(x,y) = x^2 - 2xy + 2y^2$$

(b) Sea $\Omega \subset \mathbb{R}^2$ y $J : \Omega \to \mathbb{R}$ una función diferenciable en $\Omega$ y sea $DJ(a) = 0$ donde $a \in \Omega$. Supongamos además que existe $D^2J(a)$ siendo la matriz Hessiana

$$H(a) = \begin{bmatrix} r & s \\ s & t \end{bmatrix}.$$

Demostrar que si $r > 0$ y $rt - s^2 > 0$ entonces $J(.)$ tiene un mínimo local en $a$

## 5.2. Métodos de gradiente

### 5.2.1. Método de gradiente para la minimización sin restricciones

Sea $J : \mathbb{R}^d \longrightarrow \mathbb{R}$ diferenciable. Consideraremos el problema siguiente: Hallar $u \in \mathbb{R}^d$ tal que

$$J(u) = \inf_{v \in \mathbb{R}^d} J(v)$$

El método de gradiente es

(a) $u^0 \in \mathbb{R}^d$ , arbitrario

(b) Conocido $u^n$ se calcula $u^{n+1}$ de la siguiente manera

$$u^{n+1} = u^n - \rho_n \nabla J(u^n)$$

**Nota 5.2** $d^n = -\nabla u^n$ es la dirección de máximo descenso local. En efecto,

$$J(u^n + \rho_n d^n) = J(u^n) + \rho_n(\nabla J(u^n), d^n) + ...$$

$$J(u^n + \rho_n d^n) - J(u^n) \approx \rho_n(\nabla J(u^n), d^n)$$

El término $(\nabla J(u^n), d^n)$ para $||d^n|| = 1$ toma el valor máximo si $d^n = \frac{\nabla J(u^n)}{||\nabla J(u^n)||}$

Analizaremos la convergencia del anterior método en el caso de funciones elípticas, es decir,

**Definición 5.5** Función elíptica

Una función $J : \mathbb{R}^d \longrightarrow \mathbb{R}$, es elíptica si es diferenciable con continuidad y existe $\alpha > 0$ tal que

$$(\nabla J(v) - \nabla J(u), v - u) \geq \alpha ||v - u||^2, \quad \forall u, v \in \mathbb{R}^d$$

**Teorema 5.7** Sea $J : \mathbb{R}^d \longrightarrow \mathbb{R}$ una función derivable en $\mathbb{R}^d$ tal que

$$\exists \alpha > 0 \quad (\nabla J(v) - \nabla J(u), v - u) \geq \alpha ||v - u||^2, \quad \forall u, v \in \mathbb{R}^d$$

$$\exists \beta \quad ||\nabla J(v) - \nabla J(u)|| \leq \beta ||v - u|| \quad \forall u, v \in \mathbb{R}^d$$

Entonces el problema de optimización: Hallar $u \in \mathbb{R}^d$ tal que

$$J(u) = \inf_{v \in \mathbb{R}^d} J(v)$$

tiene solución única y existen dos números $a$ y $b$ tales que para todo $n \geq 0$ verifican

$$0 < a \leq \rho_n \leq b < \frac{2\alpha}{\beta^2}$$

de modo que el método de gradiente antes descrito es convergente.

Para realizar la demostración necesitamos conocer algunas propiedades de las funciones elípticas.

**Lema 5.2** Una función $J : \mathbb{R}^d \to \mathbb{R}$ elíptica verifica la siguiente desigualdad,

$$J(v) - J(u) \geq (\nabla J(u), v - u) + \frac{\alpha}{2} ||v - u||^2 \quad \forall u, v \in \mathbb{R}^d$$

y por tanto es estrictamente convexa.

**Demostración:**

$$J(v) - J(u) = \int_0^1 (\nabla J(u + t(v - u)), v - u) dt$$

en efecto, la anterior igualdad no es más que

$$f(1) - f(0) = \int_0^1 f'(t) dt$$

para la función $f : \mathbb{R} \to \mathbb{R}$ definida por

$$f(t) = J(u + t(v - u)) \quad u, v \in \mathbb{R}^d$$

resulta pues,

$$\begin{aligned} J(v) - J(u) &= (\nabla J(u), v-u) + \int_0^1 (\nabla J(u + t(v-u)) - \nabla J(u), (v-u))dt \\ &= (\nabla J(u), v-u) + \int_0^1 \frac{1}{t}(\nabla J(u + t(v-u)) - \nabla J(u), t(v-u))dt \\ &\geq (\nabla J(u), v-u) + \int_0^1 \alpha t||v-u||^2 dt = (\nabla J(u), v-u) + \frac{\alpha}{2}||v-u||^2 \end{aligned}$$

■

**Teorema 5.8** Si $J(.)$ es una función elíptica el problema de encontrar $u \in \mathbb{R}^d$ tal que

$$J(u) = \inf_{v \in \mathbb{R}^d} J(v)$$

tiene solución única y se verifica

$$\nabla J(u) = 0$$

**Demostración:**

Si $J(.)$ es elíptica verifica

$$\lim_{||v|| \to \infty} J(v) = \infty$$

en efecto,

$$J(v) - J(u) \geq (\nabla J(u), v-u) + \frac{\alpha}{2}||v-u||^2 \quad \forall u, v \in \mathbb{R}^d$$

en particular, tomando $u = 0$

$$\begin{aligned} J(v) &\geq J(0) + (\nabla J(0), v) + \frac{\alpha}{2}||v||^2 \\ &\geq J(0) - ||\nabla J(0)||||v|| + \frac{\alpha}{2}||v||^2 \end{aligned}$$

La variante del teorema de Weierstrass nos da la existencia de solución. La unicidad se obtiene utilizando la convexidad estricta de $J(.)$. En efecto, supongamos que $u_1$ y $u_2$ son dos soluciones. Por la convexidad estricta

$$J(\frac{u_1 + u_2}{2}) < \frac{1}{2}J(u_1) + \frac{1}{2}J(u_2)$$

Si $\gamma = J(u_1) = J(u_2) = \text{ínf}(J(v)$, entonces $J(\frac{u_1+u_2}{2}) < \gamma$ y $u_1$ y $u_2$ no podrían ser soluciones.

Finalmente, la solución única $u$, es también un mínimo relativo, por tanto verifica $\nabla J(u) = 0$.■

Estamos en condiciones de demostrar el teorema de convergencia del algoritmo de gradiente.

**Demostración del teorema de convergencia:** El algoritmo de gradiente es

(a) $u^0 \in \mathbb{R}^d$ , arbitrario

(b) Conocido $u^n$ se calcula $u^{n+1}$ de la siguiente manera

$$u^{n+1} = u^n - \rho_n \nabla J(u^n)$$

por otra parte, si $u$ es solución, tenemos $\nabla J(u) = 0$ es decir,

$$u = u - \rho_n \nabla J(u)$$

de donde

$$\begin{aligned} ||u - u^{n+1}||^2 &= ||u - u^n - \rho_n(\nabla J(u) - \nabla J(u^n))||^2 \\ &= ||u - u^n||^2 - 2\rho_n(\nabla J(u) - \nabla J(u^n), u - u^n) + \rho_n^2||\nabla J(u) - \nabla J(u^n||^2 \\ &\leq ||u - u^n||^2 - 2\rho_n\alpha||u - u^n||^2 + \beta^2\rho_n^2||u - u^n||^2 \\ &= (1 - 2\rho_n\alpha + \beta^2\rho_n^2)||u - u^n||^2 \end{aligned}$$

siempre podemos elegir $\rho_n$ de modo que

$$\tau(\rho_n) = (1 - 2\rho_n\alpha + \beta^2\rho_n^2) < 1$$

en efecto, la función

$$\tau(\rho) : \rho \longrightarrow 1 - 2\rho\alpha + \beta^2\rho^2$$

alcanza su valor mínimo en $\rho_{min} = \frac{\alpha}{\beta^2}$ y tenemos $0 < \tau(\rho_{min}) < 1$, por tanto siempre existen dos números $a$ y $b$ tales que , si elegimos $a < \rho_n < b$, y denotamos mediante $\gamma^2 = \text{máx}\{\tau(a), \tau(b)\} < 1$ resulta

$$||u - u^{n+1}|| \leq \gamma||u - u^n||$$

de donde

$$||u - u^n|| \leq \gamma^n||u - u^0|| \overset{n\to\infty}{\longrightarrow} 0$$

La convergencia es al menos lineal, pues

$$\lim_{n\to\infty} \frac{||u - u^{n+1}||}{||u - u^n||} \leq \gamma < 1$$

■

### 5.2.2. Método del gradiente con paso óptimo

El método de gradiente con paso óptimo nos da un criterio para elegir el valor de $\rho_n$ en cada iteración. El algoritmo es el siguiente:

(a) $u^0 \in \mathbb{R}^d$ , arbitrario

(b) Conocido $u^n$ se calcula $u^{n+1}$ de la siguiente manera: Se calcula $\rho_n = \rho(u^n)$ resolviendo el problema de minimización de una variable real

$$J(u^n - \rho_n \nabla J(u^n)) = \inf_{\rho \in \mathbb{R}} J(u^n - \rho \nabla J(u^n))$$

obtenido $\rho_n$ se calcula $u^{n+1}$ mediante

$$u^{n+1} = u^n - \rho_n \nabla J(u^n)$$

**Teorema 5.9** Sea $J : \mathbb{R}^d \longrightarrow \mathbb{R}$ elíptica y diferenciable con continuidad, entonces el método de gradiente con paso óptimo es convergente.

**Demostración:**

Supongamos que $\nabla J(u^n) \neq 0$ (en caso contrario el método sería convergente en un número finito de pasos. Consideremos la función

$$f : \rho \in \mathbb{R} \longrightarrow J(u^n - \rho \nabla J(u^n)) \in \mathbb{R}$$

se comprueba sin gran dificultad

- $f(.)$ es estrictamente convexa.
- $f(.)$ verifica $\lim_{\rho \to \infty} f(\rho) = \infty$
- $f'(\rho) = -(\nabla J(u^n - \rho \nabla J(u^n)), \nabla J(u^n))$

por tanto el problema de hallar $\rho_n = \rho(u^n)$ tal que

$$f(\rho_n) = \inf_{\rho \in \mathbb{R}} f(\rho)$$

tiene solución única y está caracterizada por

$$f'(\rho_n) = 0$$

es decir

$$(\nabla J(u^n - \rho_n \nabla J(u^n)), \nabla J(u^n)) = 0$$

o bien

$$(\nabla J(u^{n+1})), \nabla J(u^n)) = 0$$

por tanto las direcciones de descenso consecutivas son ortogonales. A partir de aquí la demostración se realiza en cinco etapas.

(a) $\lim_{n \to \infty} (J(u^n) - J(u^{n+1})) = 0$

(b) $\lim_{n\to\infty} ||u^n - u^{n+1}|| = 0$

(c) $||\nabla J(u^n)|| \leq ||\nabla J(u^n) - \nabla J(u^{n+1})||$

(d) $\lim_{n\to\infty} ||\nabla J(u^n|| = 0$

(e) $\lim_{n\to\infty} ||u^n - u|| = 0$

**Demostración:**

(a) La sucesión $(J(u^n))_{n\geq}$ es decreciente por construcción y acotada inferiormente por $J(u)$. Por lo tanto es convergente, en consecuencia,

$$\lim_{n\to\infty} (J(u^n) - J(u^{n-1})) = 0$$

(b) Tenemos, según se ha visto en el teorema anterior $(\nabla J(u^{n+1}), \nabla J(u^n)) = 0$. Como $u^{n+1} = u^n - \rho(u^n)\nabla J(u^n)$, resulta

$$(\nabla J(u^{n+1}), u^n - u^{n+1}) = 0$$

Por otra parte, la elipticidad de $J(.)$ implica

$$J(u^n) \geq J(u^{n+1}) + (\nabla J(u^{n+1}), u^n - u^{n+1}) + \frac{\alpha}{2}||u^n - u^{n+1}||^2$$

de donde

$$J(u^n) - J(u^{n+1}) \geq \frac{\alpha}{2}||u^n - u^{n+1}||^2$$

de la parte 1 deducimos

$$\lim_{n\to\infty} ||u^n - u^{n+1}|| = 0$$

(c) Por la ortogonalidad de $\nabla J(u^n)$ y de $\nabla J(u^{n+1})$ resulta

$$\begin{aligned} ||\nabla J(u^n)||^2 &= (\nabla J(u^n), \nabla J(u^n) - \nabla J(u^{n+1})) \\ &\leq ||\nabla J(u^n)||.||\nabla J(u^n) - \nabla J(u^{n+1})|| \end{aligned}$$

de donde se obtiene 3) dividiendo ambos miembros de la desigualdad por $||\nabla J(u^n)||$

(d) La sucesión $(J(u^n))_{n\geq 0}$ es acotada. Como $J(.)$ verifica $\lim_{||v||\to\infty} J(v) = \infty$, la sucesión $(u^n)_n$ está acotada y se puede incluir en un conjunto compacto. Por otra parte $J(.)$ es diferenciable con continuidad, es decir, la aplicación

$$DJ(.) : v \in \mathbb{R}^d \longrightarrow DJ(v) \in \mathcal{L}(\mathbb{R}^d; \mathbb{R})$$

es continua. O dicho de otra forma la aplicación

$$\nabla J(.) : v \in \mathbb{R}^d \longrightarrow \nabla J(v) \in \mathbb{R}^d$$

es continua. Por tanto sobre los conjuntos compactos es uniformemente continua. En consecuencia

$$\begin{aligned}\lim_{n\to\infty}||\nabla J(u^n)|| &\leq \lim_{n\to\infty}||\nabla J(u^n)-\nabla J(u^{n+1})||\\ &= \lim_{||u^n-u^{n+1}||\to 0}||\nabla J(u^n)-\nabla J(u^{n+1})||=0\end{aligned}$$

(e)

$$\begin{aligned}\alpha||u^n-u||^2 &\leq (\nabla J(u^n)-\nabla J(u),u^n-u)\\ &= (\nabla J(u^n),u^n-u)\\ &\leq ||\nabla J(u^n||||u^n-u||\end{aligned}$$

de donde finalmente

$$||u^n-u||\leq\frac{1}{\alpha}||\nabla J(u^n)||$$

y finalmente

$$\lim_{n\to\infty}||u^n-u||=0$$

■

**Ejercicios**

(a) Sea $J:\mathbb{R}^d\longrightarrow\mathbb{R}$ dos veces diferenciable en $\mathbb{R}^d$. Demostrar:

*a*) $J(.)$ es convexa si y solo si la diferencial segunda en todo punto, $D^2J(u)$, es semidefinida positiva, es decir,

$$D^2J(u)(v,v)\geq 0\quad\forall v\in\mathbb{R}^d,\forall u\in\mathbb{R}^d$$

*b*) Si la diferencial segunda en todo punto $u\in\mathbb{R}^d$ es definida positiva, es decir,

$$D^2J(u)(v,v)>0\quad\forall v\in\mathbb{R}^d,\forall u\in\mathbb{R}^d$$

entonces, $J(.)$ es estrictamente convexa.

*c*) Encontrar un ejemplo sencillo que permita asegurar que el recíproco de b) no es cierto.

(b) Considerar la función

$$\begin{aligned}J:\quad \mathbb{R}^2 &\longrightarrow\mathbb{R}\\ (x,y) &\longrightarrow (x-2)^4+(x-2y)^2\end{aligned}$$

y considerar el problema

Hallar $u=(x,y)$ tal que

$$J(u)=\inf_{v\in\mathbb{R}^2}J(v)$$

*a*) Demostrar que el problema tiene al menos una solución.

*b*) Hallar una solución y deducir que esta solución es única.

*c*) Aplicar el método del gradiente con paso óptimo para aproximar la solución anterior partiendo de $(0., 3.)$

## 5.3. Método de relajación

Vamos a describir el algoritmo de relajación para resolver el problema:

Hallar $u \in \mathbb{R}^d$ tal que

$$J(u) = \inf_{v \in \mathbb{R}^d} J(v)$$

**Descripción del algoritmo**

(a) Sea $u^0 \in \mathbb{R}^d$ arbitrario.

(b) Obtenido $u^n = (u_1^n, ..., u_d^n) \in \mathbb{R}^d$ calculamos $u^{n+1} = (u_1^{n+1}, ..., u_d^{n+1}) \in \mathbb{R}^d$ en $d$ etapas que son: Para $i = 1, 2, ..., d$ calculamos sucesivamente

$$u^{n+\frac{i}{d}} = (u_1^{n+1}, u_2^{n+1}, ..., u_i^{n+1}, u_{i+1}^n, ..., u_d^n) \in \mathbb{R}^d$$

solución de

$$J(u^{n+\frac{i}{d}}) = \inf_{v=(u_1^{n+1}, u_2^{n+1}, ..., v_i, u_{i+1}^n, ..., u_d^n)} J(v)$$

**Observación:** En cada etapa, se trata de un problema de minimización en una sola variable real. En cada etapa la condición necesaria de mínimo ( y suficiente si $J(.)$ es convexa) es:

$$\frac{\partial J}{\partial x_i}(u_1^{n+1}, ..., u_{i-1}^{n+1}, u_i^{n+1}, u_{i+1}^n, ..., u_d^n) = 0$$

Estudiemos ahora la convergencia del método.

**Teorema 5.10** Si $J : \mathbb{R}^d \longrightarrow \mathbb{R}$ es elíptica y diferenciable con continuidad, el método de relajación es convergente.

**Demostración:** Primero observemos que el algoritmo de relajación está bien definido. En efecto, en cada etapa, $i = 1, ..., d$ se resuelve un problema de optimización, para una función $J_i$ de una sola variable que es elíptica y por tanto tiene solución única.

$$J_i(\xi) = J(u_1^{n+1}, ..., u_{i-1}^{n+1}, \xi, u_{i+1}^n, ..., u_d^n)$$

y la solución de

$$J_i(u_i^{n+1}) = \inf_{\xi\in\mathbb{R}} J_i(\xi)$$

está caracterizada por

$$\frac{\partial J}{\partial x_i}(u^{n+\frac{i}{d}}) = 0$$

La demostración ahora se hace en varias etapas:

(a)

$$\lim_{n\to\infty}(J(u^n) - J(u^{n+1})) = 0$$

En efecto, por la propia construcción del algoritmo tenemos

$$J(u) \le J(u^{n+1}) \le J(u^n) \le J(u^0)$$

es decir $(J(u^n))_n$ es una sucesión decreciente de números reales acotada inferiormente, por tanto convergente. En particular $\lim_{n\to\infty}(J(u^n) - J(u^{n+1})) = 0$

(b)

$$\lim_{n\to\infty} ||u^n - u^{n+1}|| = 0$$

y en particular

$$\lim_{n\to\infty} ||u^{n+\frac{i}{d}} - u^{n+1}|| = 0$$

La elipticidad implica

$$J(u^{n+\frac{i-1}{d}}) - J(u^{n+\frac{i}{d}}) \ge (\nabla J(u^{n+\frac{i}{d}}), u^{n+\frac{i-1}{d}} - u^{n+\frac{i}{d}}) + \frac{\alpha}{2}||u^{n+\frac{i-1}{d}} - u^{n+\frac{i}{d}}||^2$$

como

$$\begin{aligned}(\nabla J(u^{n+\frac{i}{d}}), u^{n+\frac{i-1}{d}} - u^{n+\frac{i}{d}}) &= \Sigma_{j=1}\frac{\partial J}{\partial x_j}(u^{n+\frac{i}{d}})(u_j^{n+\frac{i-1}{d}} - u_j^{n+\frac{i}{d}}) \\ &= \Sigma_{j\neq i}\frac{\partial J}{\partial x_j}(u^{n+\frac{i}{d}})(u_j^{n+\frac{i-1}{d}} - u_j^{n+\frac{i}{d}}) + \frac{\partial J}{\partial x_i}(u^{n+\frac{i}{d}})(u_i^{n+\frac{i-1}{d}} - u_i^{n+\frac{i}{d}}) = 0\end{aligned}$$

pues para $j \neq i$ se tiene $u_j^{n+\frac{i-1}{d}} - u_j^{n+\frac{i}{d}} = 0$ y por otra parte $\frac{\partial J}{\partial x_i}(u^{n+\frac{i}{d}}) = 0$ por la caracterización de la solución del problema de minimización en dimensión uno correspondiente. De modo que,

$$J(u^{n+\frac{i-1}{d}}) - J(u^{n+\frac{i}{d}}) \ge \frac{\alpha}{2}||u^{n+\frac{i-1}{d}} - u^{n+\frac{i}{d}}||^2 = \frac{\alpha}{2}|u_i^n - u_i^{n+1}|^2$$

sumando para $i = 1, ..., d$

$$\Sigma_{i=1}^d J(u^{n+\frac{i-1}{d}}) - J(u^{n+\frac{i}{d}}) \ge \frac{\alpha}{2}\Sigma_{i=1}^d |u_i^n - u_i^{n+1}|^2$$

de donde

$$J(u^n) - J(u^{n+1}) \ge \frac{\alpha}{2}||u^n - u^{n+1}||^2$$

Y finalmente aplicando el resultado de la parte 1 obtenemos el resultado.

(c)

$$\lim_{n\to\infty} |\frac{\partial J}{\partial x_i}(u^{n+\frac{i}{d}}) - \frac{\partial J}{\partial x_i}(u^{n+1})| = 0$$

en efecto, como cada sucesión $(J(u^{n+\frac{i}{d}}))_n$ es decreciente por construcción, $(u^{n+\frac{i}{d}})_n$ es acotada pues $J(.)$ verifica la propiedad $\lim_{||v||\to\infty} J(v) = \infty$ y por otra parte cada derivada parcial $\frac{\partial J}{\partial x_i}(.)$ es continua, y por lo tanto es uniformemente continua en los compactos, resulta

$$\lim_{n\to\infty} ||u^{n+\frac{i}{d}} - u^{n+1}|| = 0 \Rightarrow \lim_{n\to\infty} |\frac{\partial J}{\partial x_i}(u^{n+\frac{i}{d}}) - \frac{\partial J}{\partial x_i}(u^{n+1})| = 0$$

(d)

$$\lim_{n\to\infty} ||u^n - u|| = 0$$

en efecto, tenemos

$$\begin{aligned} \alpha||u^{n+1} - u||^2 &\leq (\nabla J(u^{n+1}) - \nabla J(u), u^{n+1} - u) \\ &= (\nabla J(u^{n+1}), u^{n+1} - u) = \Sigma_{i=1}^n \frac{\partial J}{\partial x_i}(u^{n+1})(u_i^{n+1} - u_i) \\ &\leq (\Sigma_{i=1}^n |\frac{\partial J}{\partial x_i}(u^{n+1})|^2)^{1/2} (\Sigma_{i=1}^n |u_i^{n+1} - u_i|^2)^{1/2} \end{aligned}$$

y finalmente

$$||u^n - u|| \leq \frac{1}{\alpha} \Sigma_{i=1}^n |\frac{\partial J}{\partial x_i}(u^{n+1})|^2)^{1/2}$$

elevando al cuadrado

$$||u^{n+1} - u||^2 \leq \frac{1}{\alpha^2} \Sigma_{i=1}^n |\frac{\partial J}{\partial x_i}(u^{n+1})|^2) = \frac{1}{\alpha^2} \Sigma_{i=1}^n |\frac{\partial J}{\partial x_i}(u^{n+1}) - \frac{\partial J}{\partial x_i}(u^{n+\frac{i}{d}})|^2$$

puesto que $\frac{\partial J}{\partial x_i}(u^{n+\frac{i}{d}}) = 0$, de donde tomando límites cuando $n \to \infty$

$$\lim_{n\to\infty} ||u^{n+1} - u|| \leq \frac{1}{\alpha^2} \Sigma_{i=1}^n \lim_{n\to\infty} |\partial_i J(u^{n+1}) - \partial_i J(u^{n+\frac{i}{d}})|^2 = 0$$

■

## 5.4. Métodos de Newton

Consideraremos en esta sección métodos para resolver ecuaciones de la forma siguiente: Sea $\Omega \in \mathbb{R}^d$ un conjunto abierto. Dada $F : \Omega \longrightarrow \mathbb{R}^d$ queremos hallar $u \in \Omega$ tal que $F(u) = 0$. En particular el método será aplicable a problemas de optimización de una función $J(.)$ diferenciable. Según hemos visto anteriormente si $J(.)$ tiene un extremo realtivo en $u \in \Omega$ entonces $\nabla J(u) = 0$. De hecho los métodos

de gradiente estudiados en las secciones anteriores son métodos de punto fijo para resolver la ecuación $\nabla J(u) = 0$. Vamos ahora a estudiar el método de Newton que no es más que una generalización a varias variables del método de Newton estudiado en el capítulo 1.

Supongamos que $F$ es diferenciable y sea $u^0$ un valor en un entorno de la solución $u$. Pongamos $u = u^0 + h$, tendremos utilizando el desarrollo de Taylor

$$F(u^0 + h) = 0 = F(u^0) + F'(u^0)(h) + \varepsilon(h^2)$$

para valores de $||h||$ pequeños resulta

$$F'(u^0)h \approx -F(u^0)$$

Si ponemos $u^1 = u^0 + h$ esperamos que $u^1$ sea una mejor aproximación de $u$ que $u^0$. De ahí el siguiente algoritmo de Newton:

**Algoritmo de Newton**

(a) $u^0 \in \Omega$ "cercano" a $u$

(b) Obtenido el valor de $u^n$ calculamos $u^{n+1}$ resolviendo

$$F'(u^n)h^n = -F(u^n)$$

$$u^{n+1} = u^n + h^n$$

Vamos a estudiar ahora la convergencia del método de Newton. Será un teorema de convergencia local. Para ello necesitaremos algunos resultados previos.

**Lema 5.3** Sea $||.||$ una norma matricial subordinada y $E$ una matriz cuadrada de orden $n$. Si $||E|| < 1$ entonces existe la matriz inversa de $(I + E)$ y $||(I + E)^{-1}|| \leq \frac{1}{1-||E||}$.

**Demostración:** Considreremos el sistema $x + Ex = 0$, es decir $Ex = -x$, entonces si $||E|| < 1$,

$$||x|| = ||Ex|| \leq ||E||.||x|| < ||x||$$

La única solución de la ecuación anterior es $x = 0$. Por tanto $I + E$ es no singular.

Por otra parte $I = (I + E) - E$, multiplicando por la derecha por $(I + E)^{-1}$, resulta

$$(I + E)^{-1} = I - E(I + E)^{-1}$$

tomando normas

$$||(I + E)^{-1}|| \leq 1 + ||E||||(I + E)^{-1}||$$

y finalmente reordenando y agrupando términos

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

■

**Lema 5.4** Sean $A$ y $B$ matrices cuadradas de orden $n$. Si $A$ es no singular y $||A^{-1}(B-A)|| < 1$ entonces $B$ es no singular y

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

**Demostración:** Pongamos $E = A^{-1}(B-A)$.

$$I + E = I + A^{-1}(B-A) = I + A^{-1}B - I = A^{-1}B$$

Por tanto $A^{-1}B$ es no singular y $(A^{-1}B)^{-1} = B^{-1}A$ y finalmente aplicando el lema anterior

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

de donde teniendo en cuenta $||B^{-1}|| = ||B^{-1}AA^{-1}|| \leq ||B^{-1}A||||A^{-1}||$ resulta

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

■

**Lema 5.5** Sea $\Omega \in \mathbb{R}^d$ un conjunto abierto y convexo y $F : \Omega \subset \mathbb{R}^d \longrightarrow \mathbb{R}$ diferenciable con continuidad en $\Omega$. Para todo $u \in \Omega$ y $h \in \mathbb{R}^d$ tal que $u + h \in \Omega$ se verifica

$$F(u+h) - F(u) = \int_0^1 F'(u+th).h dt$$

**Demostración:** consideremos la función $f : \mathbb{R} \to \mathbb{R}$ definida por $t \in [-1, 1] \to f(t) = F(u+th)$. tenemos por la regla de Barrow

$$f(1) - f(0) = \int_0^1 f'(t) dt$$

que es la expresión buscada teniendo en cuenta que

$$f'(t) = F'(u+th).h$$

■

**Observación:** El lema es válido para $F : \Omega \subset \mathbb{R}^d \to \mathbb{R}^p$ aplicando lo anterior a las p componentes de $F$.

**Lema 5.6** Sea $F : \Omega \subset \mathbb{R}^d \longrightarrow \mathbb{R}^p$ donde $\Omega \in \mathbb{R}^d$ es un conjunto abierto y convexo, una función tal que $DF(.)$ es Lipschitziana, es decir

$$||F'(v) - F'(u)|| \leq \gamma ||v - u|| \quad \forall\, u, v \in \Omega$$

entonces para todo $u+h \in \Omega$

$$||F(u+h) - F(u) - F'(u)h|| \leq \frac{\gamma}{2}||h||^2$$

**Demostración:** Tenemos

$$F(u+h) - F(u) - F'(u)h = \int_0^1 (F'(u+th) - F'(u))hdt$$

tomando normas y mayorando obtenemos el resultado buscado ■

**Teorema 5.11** Sea $\Omega \in \mathbb{R}^d$ un conjunto abierto y convexo y $F : \Omega \subset \mathbb{R}^d \longrightarrow \mathbb{R}^d$ diferenciable con continuidad en $\Omega$. Supongamos que existe un punto $u \in \Omega$ y tres constantes $r > 0$, $\beta > 0$ y $\gamma \geq 0$ tales que

- $F(u) = 0$
- $\mathcal{U}(u,r) = \{v \in \mathbb{R}^d;\ ||v-u|| < r\} \subset \Omega$
- $F'(u)$ es no singular y $||F'(u)^{-1}|| \leq \beta$
- $||F'(v) - F'(u)|| \leq \gamma||v-u|| \quad \forall\, v,u \in \mathcal{U}$

Entonces para $\varepsilon = \text{mín}\{r, \frac{1}{2\beta\gamma}\}$ y para todo $u^0 \in \mathcal{U}(u,\varepsilon) = \{v \in \mathbb{R}^d;\ ||v-u|| < \varepsilon\}$ la sucesión generada por el algoritmo de Newton está bien definida y converge hacia $u$ con convergencia cuadrática, es decir

$$||u^{n+1} - u|| \leq \beta\gamma||u^n - u||^2$$

**Demostración:** tomemos $\varepsilon = min\{r, \frac{1}{2\beta\gamma}\}$ entonces $\forall u^0 \in \mathcal{U}(u,\varepsilon)$ $F'(u^0)$, es no singular, en efecto

$$\begin{aligned} ||F'(u)^{-1}[F'(u^0) - F'(u)]|| &\leq ||F'(u)^{-1}||||F'(u^0) - F'(u)|| \\ &\leq \beta\gamma||u^0 - u|| \leq \beta\gamma\varepsilon \leq \frac{1}{2} \end{aligned}$$

Entonces por la relación del lema 5.4, $F'(u^0)$ es no singular y

$$||F'(u^0)^{-1}|| \leq \frac{||F'(u)^{-1}||}{1 - ||F'(u)^{-1}[F'(u^0) - F'(u)||} \leq 2||F'(u)^{-1}|| \leq 2\beta$$

$u^1$ está bien definido y

$$\begin{aligned} u^1 - u &= u^0 - u - F'(u^0)^{-1}[F(u^0) - F(u)] \\ &= F'(u^0)^{-1}[F(u) - F(u^0) - F'(u^0)(u - u^0)] \end{aligned}$$

$$\begin{aligned} ||u^1 - u|| &\leq ||F(u^0)^{-1}||||F(u) - F(u^0) - F'(u^0)(u - u^0)|| \\ &\leq \beta\gamma||u^0 - u||^2 \end{aligned}$$

Por otra parte como $||u^0 - u|| \leq \frac{1}{2\beta\gamma}$, resulta

$$||u^1 - u|| \leq \frac{1}{2}||u^0 - u||$$

lo que prueba $u^1 \in \mathcal{U}(u, \varepsilon)$ y por inducción probamos que $||u^{n+1} - u|| \leq \frac{1}{2}||u^n - u||$, el método es convergente y la convergencia es cuadrática pues,

$$||u^{n+1} - u|| \leq \beta\gamma||u^n - u||^2$$

■

**Evaluación del coste del método de Newton**

En cada iteración hay que

- Evaluar $F(u^n)$, $d$ evaluaciones funcionales.
- Calcular los términos de $F'(u^n)$, es decir, $d^2$ ($\frac{d(d+1)}{2}$ si $F(u) = \nabla J(u)$) evaluaciones funcionales.
- Resolver un sistema lineal de $d$ ecuaciones con $d$ incógnitas, es decir, del orden de $\frac{d^3}{3}$ ($\frac{d^3}{6}$ si $F(u) = \nabla J(u)$)) operaciones, si lo resolvemos con un método directo.

Muchas veces no se conoce la expresión analítica de las funciones, por lo que no se conoce la expresión de las derivadas parciales. Se recurre entonces al cálculo numérico de estas derivadas mediante diferencias finitas,

$$[F'(u^n)]_{ij} \approx \frac{F_i(u^n + \lambda_n e_j) - F_i(u^n)}{\lambda_n}$$

**Observación:** Si se elige $\lambda_n$ adecuadamente, por ejemplo, $\lambda_n \leq C||F(u^n)||$, la convergencia sigue siendo cuadrática.

**Ejercicios:**

(a) Sea $F : \mathbb{R}^2 \longrightarrow \mathbb{R}^2$ definida por

$$F(x, y) = \begin{bmatrix} x + y - 3 \\ x^2 + y^2 - 9 \end{bmatrix}.$$

Resolver utilizando el método de Newton la ecuación $F(x, y) = 0$ tomando como valor inicial $(x, y)^0 = (1, 5)$.

(b) Resolver utilizando el método de Newton el problema siguiente: Hallar $(\bar{x}, \bar{y}) \in \mathbb{R}^2$ tal que

$$J(\bar{x}, \bar{y}) = \inf_{(x,y)\in\mathbb{R}^2} J(x, y)$$

donde $J(x, y) = (x - 2)^4 + (x - 2y)^2$.

(c)
- Considerar la función $J : \mathbb{R}^2 \to \mathbb{R}^2$ definida por

$$J(x, y) = log(x^2 + y^2 + 1)$$

Demostrar que el problema: Hallar $u = (x, y)^t \in \mathbb{R}^2$ tal que

$$J(u) = \inf_{v\in\mathbb{R}^2} J(v)$$

tiene solución única.
- Comprobar que la Diferencial Segunda de $J(.)$ es definida positiva en el punto solución.
- Calcular la solución del problema anterior utilizando el método de relajación y tomando como valor inicial $u^0 = (1, 1)$.

(d) Considerar el siguiente problema en $\mathbb{R}^d$: hallar $u \in \mathbb{R}^d$ tal que

$$Au + F(u) = f$$

donde $f \in \mathbb{R}^d$, $A$ es una matriz de orden $n$ definida positiva, y por tanto verifica

$$\exists\ \alpha > 0\ (Av, v) \geq \alpha||v||^2\ \forall\ v \in \mathbb{R}^d$$

y $F : \mathbb{R}^d \longrightarrow \mathbb{R}^d$ es Lipschitziana, es decir

$$\exists\ M \geq 0\ ||F(u) - F(v)|| \leq M||u - v||\ \forall\ u, v \in \mathbb{R}^d$$

Demostrar que si $\frac{M}{\alpha} < 1$ el siguiente método de punto fijo converge

- $u^0 \in \mathbb{R}^d$ arbitrario
- Obtenido $u^n$ calculamos $u^{n+1}$ resolviendo

$$Au^{n+1} = f - F(u^n)$$

# Capítulo 6

# Optimización de funciones cuadrátricas

## 6.1. Generalidades sobre las funciones cuadráticas

Una función cuadrática es una función de la forma

$$J : v \in \mathbb{R}^d \longrightarrow J(v) \in \mathbb{R}$$

donde $J(v) = \frac{1}{2}(Av, v) - (b, v)$ siendo $A$ una matriz de $d$ filas por $d$ columnas simétrica y donde $b \in \mathbb{R}^d$

Una función cuadrática $J(.)$, es elíptica si y solo si la matriz asociada $A$ es definida positiva. En efecto, calculemos la diferencial de $J(.)$ en un punto $u$,

$$\begin{aligned} DJ(u)(v) &= (\nabla J(u), v) \\ &= \frac{1}{2}((Au, v) + (Av, u)) - (b, v) \\ &= (Au, v) - (b, v) \end{aligned}$$

pues $A$ es simétrica y $(Av, u) = (v, Au) = (Au, v)$. Por tanto

$$\nabla J(u) = Au - b$$

Si $J(.)$ es elíptica, existe $\alpha > 0$ tal que

$$(\nabla J(u) - \nabla J(v), u - v) \geq \alpha ||u - v||^2 \quad \forall u, v \in \mathbb{R}^d$$

como $\nabla J(u) = Au - b$ y $\nabla J(v) = Av - b$ resulta

$$(Au - Av, u - v) \geq \alpha ||u - v||^2 \quad \forall u, v \in \mathbb{R}^d$$

como $u$ y $v$ son cualesquiera

$$(Av, v) \geq \alpha ||v||^2$$

Recíprocamente, si $A$ es definida positiva, existe $\alpha > 0$ tal que

$$(Av, v) \geq \alpha ||v||^2 \quad \forall v \neq 0$$

por tanto $\forall u, v \in \mathbb{R}^d \quad u \neq v$, entonces

$$(A(u - v), u - v) \geq \alpha ||u - v||^2$$

y finalmente

$$(\nabla J(u) - \nabla J(v), u - v) \geq \alpha ||u - v||^2$$

Si $A$ es definida positiva el mínimo de $J(.)$ está caracterizado por $\nabla J(u) = 0$ es decir, $Au = b$. De modo que el problema de hallar $u \in \mathbb{R}^d$ tal que $J(u) = \text{ínf}_{v \in \mathbb{R}^d} J(v)$ equivale a resolver $Au = b$.

Vamos a considerar ahora la minimización de funciones cuadráticas con matriz asociada $A$ definida positiva, por tanto elípticas. Sabemos que el problema tiene solución única.

## 6.2. Métodos de descenso

### 6.2.1. Método general de descenso

Sea $u^0 \in \mathbb{R}^d$ vector inicial arbitrario. Construiremos una sucesión de vectores $(u^n)_n$ a partir de $u^0$. El paso general para construir $u^{n+1}$ a partir de $u^n$ es

- Fijamos una dirección de descenso $d^n \neq 0$ en el punto $u^n$
- Resolvemos el problema del mínimo siguiente: Hallar $\rho_n = \rho(u^n, d^n)$ tal que

$$J(u^n + \rho_n d^n) = \inf_{\rho \in \mathbb{R}} J(u^n + \rho d^n)$$

  que tiene solución única pues $J(.)$ es elíptica.

- $$u^{n+1} = u^n + \rho_n d^n$$

En el caso de funciones cuadráticas el cálculo de $\rho_n = \rho(u^n, d^n)$ es sencillo, en efecto derivando la función $\rho \to J(u^n + \rho d^n)$ e igualando a 0

$$(\nabla J(u^n + \rho_n d^n), d^n) = 0$$

$$(A(u^n + \rho_n d^n) - b, d^n) = 0$$

despejando $\rho_n$

$$\rho_n = \frac{(b - Au^n, d^n)}{(Ad^n, d^n)} = \frac{(r^n, d^n)}{(Ad^n, d^n)}$$

donde hemos introducido el residuo correspondiente $r^n = b - Au^n$.

**Algoritmo general de descenso**

(a) $u^0 \in \mathbb{R}^d$ arbitrario.

(b) $r^n = b - Au^n$

(c) $\rho_n = \frac{(r^n, d^n)}{(Ad^n, d^n)}$

(d) $u^{n+1} = u^n + \rho_n d^n$

o bien observando que $r^{n+1} = r^n - \rho_n Ad^n$

(a) $u^0 \in \mathbb{R}^d$ arbitrario; $r^0 = b - Au^0$.

(b) $\rho_n = \frac{(r^n, d^n)}{(Ad^n, d^n)}$

(c) $u^{n+1} = u^n + \rho_n d^n$

(d) $r^{n+1} = r^n - \rho_n Ad^n$

Para distintas elecciones $d^n$, obtenemos distintos métodos.

**Ejercicio:** Verificar que si en el método general de descenso elegimos como direcciones de descenso $d^n$ las direcciones de los ejes, es decir, $e_i = (0, ..., 0, 1, 0, ..., 0)$ obtenemos el método de Gauss-Seidel para resolver $Au = b$.

### 6.2.2. Propiedades de convergencia de los métodos de descenso

Vamos a estudiar ahora las Propiedades de los métodos de descenso.

**Propiedad 6.1** Cualquiera que sea la dirección de descenso $d^n$ elegida, para $\rho_n$ óptimo se tiene para todo $n \geq 0$

$$(d^n, r^{n+1}) = 0$$

es decir, la dirección de descenso y el nuevo gradiente de la función son ortogonales.

**Demostración:** Hemos comprobado ya que $r^{n+1} = r^n - \rho_n A d^n$. De donde,

$$\begin{aligned}(d^n, r^{n+1}) &= (d^n, r^n - \rho_n A d^n) = (d^n, r^n) - \rho_n (d^n, A d^n) \\ &= (d^n, r^n) - \frac{(r^n, d^n)}{(A d^n, d^n)} (A d^n, d^n) = 0\end{aligned}$$

**Propiedad 6.2** El problema de minimizar una función $J : v \to \frac{1}{2}(Av, v) - (b, v)$ con $A$ simétrica y definida positiva, es equivalente a minimizar

$$E(v) = (A(v-u), v-u) = ||v-u||_A^2$$

donde $u$ es la solución buscada, es decir, verificando $Au = b$.

**Nota 6.1** Hemos introducido la notación $||v||_A = (Av, v)^{1/2}$ para la norma asociada al producto escalar $u, v \longrightarrow (Au, v)$

**Nota 6.2** $E(.)$ es la función error asociada al valor $v$, más precisamente, es el cuadrado de la norma asociada a la matriz $A$ del error $e = v - u$.

**Demostración:**

$$\begin{aligned}E(v) &= (A(v-u), v-u) = (Av, v) - 2(Au, v) + (Au, u) \\ &= (Av, v) - 2(b, v) + (Au, u) \\ &= 2J(v) + (Au, u)\end{aligned}$$

Como $(Au, u)$ es constante, es decir, independiente de $v$, $E(.)$ y $2J(.)$ y por lo tanto $J(.)$ alcanzan el mínimo en el mismo punto $u$. ■

Demos ahora una interpretación geométrica de los métodos de descenso: Consideremos el caso de funciones cuadráticas en $\mathbb{R}^2$. Las ecuación $E(v) = cte$ es una elipse. Para diferentes valores $E(v) = E(u^n)$ obtenemos una familia de elipses concéntricas centradas en el mínimo de $u$ de la función y que representan las curvas de nivel. El vector $d^n$ es tangente a la elipse $E(v) = E(u^{n+1})$. Como $r^{n+1}$ es ortogonal a $d^n$, $r^{n+1}$ es ortogonal a la tangente de la curva de nivel.

Vamos a estudiar ahora cuáles son las posibles elecciones de $d^n$ que permitan asegurar la convergencia del método de descenso.

**Lema 6.1** Para $d^n \neq 0$ y $r^n \neq 0$

$$E(u^{n+1}) = E(u^n)(1 - \gamma_n)$$

donde

$$\gamma_n = \frac{(r^n, d^n)^2}{(A d^n, d^n)(A^{-1} r^n, r^n)}$$

**Demostración:** Sustituimos el valor de $\rho_n$ en la expresión de $E(u^{n+1})$:

$$\begin{aligned}
E(u^{n+1}) &= E(u^n + \rho_n d^n) = (A(u^n - u + \rho_n d^n), u^n - u + \rho_n d^n) \\
&= (A(u^n - u), u^n - u) + 2\rho_n (A(u^n - u), d^n) + \rho_n^2 (Ad^n, d^n) \\
&= E(u^n) - 2\rho_n (r^n, d^n) + \rho_n^2 (Ad^n, d^n) \\
&= E(u^n) - 2\frac{(r^n, d^n)^2}{(Ad^n, d^n)} + \frac{(r^n, d^n)^2}{(Ad^n, d^n)^2}(Ad^n, d^n) \\
&= E(u^n) - \frac{(r^n, d^n)^2}{(Ad^n, d^n)} \\
&= E(u^n)(1 - \frac{1}{E(u^n)} \frac{(r^n, d^n)^2}{(Ad^n, d^n)})
\end{aligned}$$

Finalmente como

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

obtenemos el resultado. ■

**Observación:** El número $\gamma_n$ es siempre positivo pues $A$ y por lo tanto también $A^{-1}$ son matrices simétricas y definidas positivas.

**Lema 6.2** Mayoración de $\gamma_n$: Cualquiera que sea $d^n \neq 0$, y siendo $\rho_n$ el valor óptimo local tenemos la siguiente relación válida para $n \geq 0$

$$\gamma_n = \frac{(r^n, d^n)^2}{(Ad^n, d^n)(A^{-1} r^n, r^n)} \geq \frac{1}{\kappa(A)} (\frac{r^n}{||r^n||}, \frac{d^n}{||d^n||})^2$$

donde $\kappa(A)$ es el número de condicionamiento de la matriz $A$.

**Demostración:** La matriz $A$ siendo simétrica y definida positiva, tiene todos sus valores propios reales y positivos.

$$0 < \lambda_{min} = \lambda_1 \leq \lambda_2 \leq .... \leq \lambda_d = \lambda_{max}$$

Sabemos $\frac{(Ad^n, d^n)}{||d^n||^2} \leq \lambda_{max}$ es decir $(Ad^n, d^n) \leq \lambda_{max} ||d^n||^2$.

Los valores propios de $A^{-1}$ son

$$0 < \frac{1}{\lambda_{max}} \leq, ..., \leq \frac{1}{\lambda_{min}}$$

de donde $\frac{(A^{-1} r^n, r^n)}{||r^n||^2} \leq \frac{1}{\lambda_{min}}$, es decir $(A^{-1} r^n, r^n) \leq \frac{1}{\lambda_{min}} ||r^n||^2$.

Y finalmente

$$\frac{(Ad^n, d^n)(A^{-1} r^n, r^n)}{||d^n||^2 ||r^n||^2} \leq \frac{\lambda_{max}}{\lambda_{min}} = \kappa(A)$$

de donde

$$\gamma_n \geq \frac{1}{\kappa(A)} \frac{(r^n, d^n)^2}{||r^n||^2 ||d^n||^2}$$

■

El anterior lema nos va a permitir elegir las direcciones de descenso:

**Teorema 6.1** Sea $\rho_n$ el óptimo local; Si para toda dirección $d^n$ se verifica para todo $n \geq 0$

$$(\frac{r^n}{||r^n||}, \frac{d^n}{||d^n||})^2 \geq \mu > 0$$

donde $\mu$ es independiente de $n$, entonces la sucesión $(u^n)_n$ generada por el algoritmo de descenso es convergente hacia la solución que minimiza $E(v)$, y por lo tanto $J(v)$.

**Demostración:** El lema anterior permite escribir

$$E(u^{n+1}) = E(u^n)(1 - \gamma_n) \leq E(u^n)(1 - \frac{\mu}{\kappa(A)})$$

de donde, aplicando recursivamente la relación anterior

$$E(u^n) \leq (1 - \frac{\mu}{\kappa(A)})^n E(u^0)$$

Por otra parte, de la desigualdad de Cauchy-Schwarz

$$0 < \mu \leq (\frac{r^n}{||r^n||}, \frac{d^n}{||d^n||})^2 \leq 1$$

y como $\kappa(A) \geq 1$, resulta $0 \leq 1 - \frac{\mu}{\kappa(A)} < 1$, de donde

$$\lim_{n\to\infty} ||u^n - u||_A^2 = \lim_{n\to\infty} E(u^n) = 0$$

y también

$$0 < \lambda_{min} \leq \frac{(A(u^n - u), u^n - u)}{||u^n - u||^2}$$

$$||u^n - u||^2 \leq \frac{1}{\lambda_{min}} (A(u^n - u), u^n - u) = \frac{1}{\lambda_{min}} E(u^n) \overset{n\to\infty}{\to} 0$$

■

**Observación:** En el marco de los métodos de descenso con $\rho = \rho^n$ óptimo local, el anterior teorema permite dar una condición suficiente en la elección de $d^n$ para asegurar la convergencia: Para todo $n \geq 0$ $d^n$ debe ser no ortogonal a $r^n$.

## 6.3. Método de gradiente con paso óptimo

### 6.3.1. Descipción del método de gradiente con paso óptimo

Entre las direcciones posibles $d^n$ que aseguran la convergencia del método de descenso con elección óptima del parámetro $\rho_n$, una elección obvia es $d^n = r^n$, para la que el parámetro $\mu$ verifica

$$\mu = (\frac{r^n}{||r^n||}, \frac{d^n}{||d^n||})^2 = ||\frac{r^n}{||r^n||}||^2 = 1$$

El método de gradiente consiste en elegir el gradiente como dirección de descenso, concretamente $d^n = -\nabla J(u^n) = -(Au^n - b) = r^n$. El algoritmo de gradiente con paso óptimo se escribe

(a) $u^0 \in \mathbb{R}^d$ arbitrario.

(b) Una vez obtenido $u^n$ se calcula $u^{n+1}$ de la siguiente manera:

- $r^n = b - Au^n$
- $\rho_n = \frac{(r^n, r^n)}{(Ar^n, r^n)}$
- $u^{n+1} = u^n + \rho_n r^n$

para evitar el producto de una matriz por un vector en el cálculo de $r^n$ observemos

$$Au^{n+1} = Au^n + \rho_n Ar^n$$

de donde

$$r^{n+1} = r^n - \rho_n Ar^n$$

y el algoritmo queda de la forma

(a) $u^0 \in \mathbb{R}^d$ arbitrario.

- $r^0 = b - Au^0$
- $\rho_0 = \frac{(r^0, r^0)}{(Ar^0, r^0)}$
- $u^1 = u^0 + \rho_0 r^0$
- $r^1 = r^0 - \rho_0 Ar^0$

(b) Una vez obtenido $u^n$ se calcula $u^{n+1}$ de la siguiente manera:

- $\rho_n = \frac{(r^n, r^n)}{(Ar^n, r^n)}$
- $u^{n+1} = u^n + \rho_n r^n$

- $r^{n+1} = r^n - \rho_n A r^n$

La expresión del error es en este caso

$$E(u^{n+1}) = E(u^n)(1 - \frac{||r^n||^4}{(Ar^n, r^n)(A^{-1}r^n, r^n)})$$

### 6.3.2. Convergencia del método de gradiente con paso óptimo

Vamos a estudiar con más profundidad la convergencia del método de gradiente con paso óptimo. Para ello necesitaremos un lema previo.

**Lema 6.3** Desigualdad de Kantorovich

Sea $A$ una matriz simétrica y definida positiva. Para todo $x \neq 0$, tenemos:

$$1 \leq \frac{(Ax, x)(A^{-1}x, x)}{||x||^4} \leq \frac{(\lambda_{min} + \lambda_{max})^2}{4\lambda_{min}\lambda_{max}}$$

donde $\lambda_{min}$ y $\lambda_{max}$ son los valores propios mínimo y máximo de $A$ respectivamente.

**Demostración:** Si $A$ e simétrica definida positiva (resp. $A^{-1}$) es diagonalizable y tiene todos sus valores propios positivos y se puede elegir una base ortonormal de vectores propios. Sean

$$0 < \lambda_{min} = \lambda_1 \leq, ..., \leq \lambda_d = \lambda_{max}$$

los valores propios de $A$. Respectivamente, los valores propios de $A^{-1}$ son

$$0 < \frac{1}{\lambda_{max}} = \frac{1}{\lambda_d} \leq, ..., \leq \frac{1}{\lambda_1} = \frac{1}{\lambda_{min}}$$

y sea $(v_i)_{i=1}^d$ la base ortonormal de vectores propios. Para $x \in \mathbb{R}^d$, $x = \Sigma_{i=1}^d x_i v_i$; $||x||^2 = \Sigma_{i=1}^d |x_i|^2$ y también $(Ax, x) = \Sigma_{i=1}^d \lambda_i |x_i|^2$ y $(A^{-1}x, x) = \Sigma_{i=1}^d \frac{1}{\lambda_i}|x_i|^2$. De modo que

$$\frac{(Ax, x)(A^{-1}x, x)}{||x||^4} = \frac{\Sigma_{i=1}^d \lambda_i |x_i|^2 \Sigma_{i=1}^d \frac{1}{\lambda_i}|x_i|^2}{\Sigma_{j=1}^d x_j^2 \Sigma_{j=1}^d x_j^2}$$

denotando $\alpha_i = \frac{x_i^2}{\Sigma_{j=1}^d x_j^2}$ para $i = 1, ..., d$ resulta $\alpha_i \geq 0$, $\Sigma_{i=1}^d \alpha_i = 1$ y

$$\frac{(Ax, x)(A^{-1}x, x)}{||x||^4} = \Sigma_{i=1}^d \alpha_i \lambda_i \Sigma_{i=1}^d \alpha_i \frac{1}{\lambda_i}$$

Consideremos en el plano $\mathbb{R}^2$ los puntos $M_i = (\lambda_i, \frac{1}{\lambda_i})$ El punto $M = \Sigma_{i=1}^d \alpha_i M_i = (\Sigma_{i=1}^d \alpha_i \lambda_i, \Sigma_{i=1}^d \alpha_i \frac{1}{\lambda_i})$ combinación lineal convexa de los puntos $M_i$ estará contenido en la zona rayada de la figura. Llamando

$\overline{\lambda} = \Sigma_{i=1}^{d} \alpha_i \lambda_i$, resulta que la ordenada del punto $M$, $\Sigma_{i=1}^{d} \alpha_i \frac{1}{\lambda_i}$ es inferior a la ordenada del punto $\overline{M}$ situado sobre la recta $\overline{M_1 M_d}$. Por otra parte la ordenada del punto $M$ es superior a $\frac{1}{\overline{\lambda}}$. Mediante un sencillo cálculo,

$$\overline{M} = (\overline{\lambda}, \frac{\lambda_1 + \lambda_d - \overline{\lambda}}{\lambda_1 \lambda_d})$$

Es decir,

$$\begin{aligned} 1 = \overline{\lambda}\,(1/\overline{\lambda}) &\leq \Sigma_{i=1}^{d} \alpha_i \lambda_i \Sigma_{i=1}^{d} \alpha_i \frac{1}{\lambda_i} \leq \overline{\lambda} \frac{\lambda_1 + \lambda_d - \overline{\lambda}}{\lambda_1 \lambda_d} \\ &\leq \max_{\lambda_1 \leq \lambda \leq \lambda_d} (\lambda \frac{\lambda_1 + \lambda_d - \lambda}{\lambda_1 \lambda_d}) = \frac{(\lambda_1 + \lambda_d)^2}{4 \lambda_1 \lambda_d} \end{aligned}$$

La función $f(\lambda) = \lambda \frac{\lambda_1 + \lambda_d - \lambda}{\lambda_1 \lambda_d}$ alcanza su máximo para $\lambda = \frac{\lambda_1 + \lambda_d}{2}$ y ese valor máximo vale $\frac{(\lambda_1 + \lambda_d)^2}{4 \lambda_1 \lambda_d}$ ■

**Teorema 6.2** El método del gradiente con paso local óptimo es convergente. El factor de reducción del error en cada paso es menor o igual que $\frac{\kappa(A)-1}{\kappa(A)+1}$.

**Demostración:** Aplicando la desigualdad de Kantorovich para el residuo en la $n$-ésima iteración $r^n$ del método de gradiente tendremos

$$\frac{||r^n||^4}{(Ar^n, r^n)(A^{-1}r^n, r^n)} \geq \frac{4\lambda_{min}\lambda_{max}}{(\lambda_{min} + \lambda_{max})^2} = \frac{4\kappa(A)}{(1 + \kappa(A))^2}$$

entonces

$$E(u^{n+1}) \leq E(u^n)(1 - \frac{4\kappa(A)}{(1 + \kappa(A))^2}) = E(u^n)(\frac{\kappa(A) - 1}{\kappa(A) + 1})^2$$

de donde finalmente

$$E(u^{n+1}) \leq E(u^0)(\frac{\kappa(A) - 1}{\kappa(A) + 1})^{2n}$$

o lo que es lo mismo

$$||u^n - u||_A \leq ||u^0 - u||_A (\frac{\kappa(A) - 1}{\kappa(A) + 1})^n$$

. ■

**Observación:** Si $\kappa(A)$ es próximo a 1, el método converge rápidamente. Cuando $\kappa(A) = 1$ todos los valores propios son iguales. Tomemos $A = \lambda I$ y $E(v) = (A(v-u), v-u) = \lambda(v-u, v-u) = \lambda||v-u||^2$. Es decir la ecuación $E(v) = cte$ es la ecuación de una superficie esférica de centro $u$ (una circunferencia si $d = 2$) y el método converge en una sola iteración.

Por el contrario si $\kappa(A)$ es grande, los valores propios $\lambda_{min}$ y $\lambda_{max}$ son muy diferentes; Los elipsoides $E(v) = cte$ son entonces muy aplastados. La convergencia es lenta. Para que

$$\frac{E(u^n)}{E(u^0)} \leq \varepsilon$$

es suficiente que

$$(\frac{\kappa(A)-1}{\kappa(A)+1})^{2n} \leq \varepsilon$$

es decir

$$2n\ ln(\frac{\kappa(A)+1}{\kappa(A)-1}) \approx ln\frac{1}{\varepsilon}$$

Como para valores de $\kappa(A)$ mucho mayores que 1,

$$ln(\frac{\kappa(A)+1}{\kappa(A)-1}) = ln(1+\frac{2}{\kappa(A)-1}) \approx \frac{2}{\kappa(A)-1} \approx \frac{2}{\kappa(A)}$$

resulta

$$n \approx \frac{\kappa(A)}{4} ln\frac{1}{\varepsilon}$$

El número de iteraciones es proporcional a $\kappa(A)$.

**Ejercicio: Método de gradiente con paso constante**

Puesto que el método de gradiente con paso óptimo , debido al efecto zig-zag y al coste del cálculo de $\rho_n$ puede no resultar óptimo desde un punto de vista global, podemos pensar en un método de gradiente con parámetro constante, donde $u^{n+1}$ se calcula a partir de $u^n$ mediante

(a) $r^n = b - Au^n$

(b) $u^{n+1} = u^n + \alpha r^n$

y donde $\alpha$ es independiente de $n$.

- Demostrar que el error en la iteración $n$-ésima $e^n = u^n - u$ se expresa $e^n = (I - \alpha A)^n e^0$ donde $I$ la matriz identidad y que la condición necesaria y suficiente de convergencia es que el radio espectral de $I - \alpha A$, verifique $\rho(I - \alpha A) < 1$, es decir, que $\alpha$ y los valores propios de $A$, $\lambda_i,\ i = 1, ..., d$ verifiquen

$$|1 - \alpha\lambda_i| < 1,\ i = 1, ..., d$$

- Supongamos que $A$ sea simétrica y definida positiva y sean $0 \leq \lambda_1 < ... \leq \lambda_d$ los valores propios de $A$. Demostrar que la condición de convergencia es

$$0 < \alpha < \frac{2}{\lambda_d}$$

- Demostrar que el valor óptimo de $\alpha$ es

$$\alpha_{opt} = \frac{2}{\lambda_1 + \lambda_d}$$

y el correspondiente radio espectral para este valor óptimo es

$$\rho(I - \alpha_{opt} A) = \frac{\kappa(A)-1}{\kappa(A)+1}$$

## 6.4. Método de Gradiente Conjugado

### 6.4.1. Introducción

En esta sección investigaremos nuevas direcciones de descenso $d^n$ para ser utilizadas con el paso local óptimo, $\rho_n = \frac{(d^n,r^n)}{(Ad^n,d^n)}$. Como hemos demostrado anteriormente $(d^{n-1}, r^n) = 0$.

Vamos a buscar ahora la nueva dirección de descenso $d^n$ en el plano formado por las dos direcciones ortogonales $r^n$ y $d^{n-1}$. Pongamos

$$d^n = r^n + \beta_n d^{n-1}$$

y calculemos el parámetro $\beta_n$ de modo que el factor de reducción del error sea lo más grande posible. Tenemos

$$E(u^{n+1}) = E(u^n)(1 - \frac{(r^n, d^n)^2}{(Ad^n, d^n)(A^{-1}r^n, r^n)})$$

Elegiremos $\beta_n$ de manera que

$$\frac{(r^n, d^n)^2}{(Ad^n, d^n)(A^{-1}r^n, r^n)}$$

sea máximo. Como

$$(r^n, d^n) = (r^n, r^n + \beta_n d^{n-1}) = ||r^n||^2 + \beta_n (r^n, d^{n-1}) = ||r^n||^2$$

elegiremos $d^0 = r^0$ de modo que esta relación se verifique también para $n = 0$. Tendremos pues $(r^n, d^n) = ||r^n||^2$ para todo $n \geq 0$. La determinación del máximo de

$$\frac{(r^n, d^n)^2}{(Ad^n, d^n)(A^{-1}r^n, r^n)} = \frac{||r^n||^4}{(Ad^n, d^n)(A^{-1}r^n, r^n)}$$

se reduce a minimizar $(Ad^n, d^n)$. Desarrollando este término

$$\begin{aligned} (Ad^n, d^n) &= (A(r^n + \beta_n d^{n-1}), r^n + \beta_n d^{n-1}) \\ &= \beta_n^2 (Ad^{n-1}, d^{n-1}) + 2\beta_n (Ad^{n-1}, r^n) + (Ar^n, r^n) \end{aligned}$$

Para que este trinomio sea mínimo, hay que elegir $\beta_n$ de modo que

$$\beta_n (Ad^{n-1}, d^{n-1}) + (Ad^{n-1}, r^n) = 0$$

de donde deducimos

$$\beta_n = -\frac{(Ad^{n-1}, r^n)}{(Ad^{n-1}, d^{n-1})}$$

y también

$$(Ad^{n-1}, r^n + \beta_n d^{n-1}) = 0$$

es decir,

$$(Ad^{n-1}, d^n) = 0$$

**Definición 6.1** Cuando dos vectores $u$ y $v$ verifican la relación $(Au, v) = 0$ se dice que son $A$-conjugados. Cuando $A$ es simétrica y definida positiva la aplicación $u, v \to (Au, v)$ es un producto escalar. La relación para dos vectores de $A$-conjugación no es más que la ortogonalidad con respecto al producto escalar $(A., .)$.

Veamos algunas propiedades que relacionan los residuos sucesivos y el valor de $\beta_n$.

**Lema 6.4** Se tienen las siguientes relaciones, válidas si $r^n \neq 0$ para $i = 0, ..., n$:

$$(r^{n+1}, r^n) = 0 \quad n \geq 0$$

además

$$\beta_0 = 0, \quad \beta_n = \frac{||r^n||^2}{||r^{n-1}||^2} \quad \forall n \geq 1$$

**Demostración:**

$$\begin{aligned}(r^{n+1}, r^n) &= (r^n - \rho_n A d^n, r^n) = ||r^n||^2 - \rho_n (A d^n, r^n) \\ &= ||r^n||^2 - \rho_n (A d^n, d^n - \beta_n d^{n-1}) \\ &= ||r^n||^2 - \rho_n (A d^n, d^n) + \rho_n \beta_n (A d^n, d^{n-1}) = 0\end{aligned}$$

teniendo en cuenta que

$$(Ad^n, d^{n-1}) = 0$$

y que

$$\rho_n = \frac{(r^n, d^n)}{(Ad^n, d^n)} = \frac{||r^n||^2}{(Ad^n, d^n)}$$

Por otra parte $Ad^{n-1} = \frac{1}{\rho_n}(r^{n-1} - r^n)$ para $n \geq 1$ de donde

$$(Ad^{n-1}, r^n) = \frac{1}{\rho_n}((r^{n-1} - r^n), r^n) = -\frac{1}{\rho_n}||r^n||^2$$

y también

$$(Ad^{n-1}, d^{n-1}) = \frac{1}{\rho_n}((r^{n-1} - r^n), d^{n-1}) = \frac{1}{\rho_n}(r^{n-1}, d^{n-1}) = \frac{1}{\rho_n}||r^{n-1}||^2$$

de donde finalmente

$$\beta_n = -\frac{(r^n, Ad^{n-1})}{(d^{n-1}, Ad^{n-1})} = \frac{||r^n||^2}{||r^{n-1}||^2}$$

■

### 6.4.2. Algoritmo de Gradiente Conjugado

El algoritmo de Gradiente Conjugado es el siguiente:

(a) $u^0$ arbitrario y $d^0 = r^0 = b - Au^0$.

Para $n = 0, 1, ...$

(b)

$$\rho_n = \frac{||r^n||^2}{(Ad^n, d^n)}$$

(c)

$$u^{n+1} = u^n + \rho_n d^n$$

$$r^{n+1} = r^n - \rho_n A d^n$$

(d)

$$\beta_{n+1} = \frac{||r^{n+1}||^2}{||r^n||^2}$$

(e)

$$d^{n+1} = r^{n+1} + \beta_{n+1} d^n$$

**Ejercicio:**

Sea $c$ el número medio de coeficientes no nulos por línea de la matriz $A$ de orden $N$. Calcular el número de operaciones de una iteración del algoritmo de Gradiente Conjugado. (Resp: $(c + 5)N + 2$ multiplicaciones y $(c + 4)N - 2$ sumas.)

### 6.4.3. Propiedades del algoritmo de Gradiente Conjugado

**Teorema 6.3** En el método de Gradiente Conjugado, tomando $d^0 = r^0 = b - Au^0$, se verifica, para todo $n \geq 1$ siempre que $r^k \neq 0$ y para todo $k \leq n - 1$, las siguientes propiedades

$$(r^n, d^k) = 0 \quad \forall\, k \leq n - 1$$

$$(d^n, Ad^k) = (Ad^n, d^k) = 0 \quad \forall\, k \leq n - 1$$

$$(r^n, r^k) = 0 \quad \forall\, k \leq n - 1$$

**Demostración:** Utilizaremos el principio de inducción.

(a) Para n=1 se verifica $(r^1, d^0) = (r^1, r^0) = 0$ por la propia definición del método y al propiedades señaladas anteriormente.

Por otra parte

$$(d^1, Ad^0) = (Ad^1, d^0) = 0$$

es la relación de conjugación para dos direcciones de descenso consecutivas.

(b) Supongamos que se verifican las relaciones del teorema para el valor $n$ y veamos que en ese caso se verifican también para $n+1$ y $k \leq n$.

Para la primera relación tenemos en el caso $k = n$, tenemos $(r^{n+1}, d^n) = 0$ que se cumple en todos los métodos de descenso con paso óptimo.

Por otra parte para $k = 1, 2, ..., n-1$ tendremos

$$(r^{n+1}, d^k) = (r^n - \rho_n Ad^n, d^k) = (r^n, d^k) - \rho_n (Ad^n, d^k)$$

siendo estos dos últimos términos nulos por la hipótesis de recurrencia.

Para la segunda relación en el caso $k = n$, tenemos $(Ad^{n+1}, d^n) = 0$ es la relación de conjugación.

Para $k = 1, 2, ..., n-1$ tenemos

$$(d^{n+1}, Ad^k) = (r^{n+1} + \beta_{n+1} d^n, Ad^k) = (r^{n+1}, Ad^k) + \beta_{n+1} (d^n, Ad^k)$$

donde en el último término del segundo miembro $(d^n, Ad^k) = 0$ por la hipótesis de inducción. Además $r^{k+1} = r^k - \rho_k Ad^k$ de donde

$$Ad^k = \frac{1}{\rho_k}(r^k - r^{k+1})$$

resultando

$$(d^{n+1}, Ad^k) = \frac{1}{\rho_k}(r^{n+1}, r^k - r^{k+1}) = \frac{1}{\rho_k}(r^{n+1}, d^k - \beta_k d^{k-1} - d^{k+1} + \beta_{k+1} d^k)$$

siendo todos los términos nulos ya sea por la hipótesis de inducción ya sea por la primera relación.

Para la tercera relación en el caso $k = n$ ya ha sido demostrado en el lema anterior.

Para $k = 1, 2, ..., n-1$ tenemos

$$(r^{n+1}, r^k) = (r^{n+1}, d^k - \beta_k d^{k-1}) = (r^{n+1}, d^k) - \beta_k (r^{n+1}, d^{k-1}) = 0$$

siendo los dos últimos términos nulos por verificarse la primera relación.

■

**Corolario 6.1** El algoritmo de Gradiente Conjugado para la resolución de un sistema con matriz simétrica y definida positiva de orden $N$ converge en al menos $N$ iteraciones.

**Demostración:** O bien $r^k = 0$ y tenemos la convergencia en $k \leq N-1$ iteraciones o por el contrario $r^N$ es ortogonal a $d^0, d^1, ..., d^{N-1}$ que son $N$ vectores linealmente independientes (por ser $A$-ortogonales) del espacio $\mathbb{R}^N$. Necesariamente $r^N = 0$.■

**Observación:** El método de gradiente conjugado, introducido en 1952 por Hestenes y Stiefel es pues teóricamente un método directo pues se obtiene salvo errores de redondeo la solución exacta con un número finito de iteraciones. Sin embargo en la práctica debido a errores de redondeo la relaciones de conjugación no se tienen exactamente y el método se considera como un método iterativo. A continuación estudiaremos como depende el factor de convergencia con el condicionamiento de la matriz $A$. Consideraremos luego técnicas de precondicionamiento que tienen como finalidad mejorar la convergencia. El objetivo es siempre lograr la convergencia con un número de iteraciones considerablemente menor que el número de ecuaciones.

**Definición 6.2** Espacios de Krylov: Llamamos espacio de Krylov de dimensión $k$, $\mathcal{K}_k$, al espacio generado por los vectores $r^0, Ar^0, ..., A^{k-1}r^0$. Es decir

$$\mathcal{K}_k = [r^0, Ar^0, ..., A^{k-1}r^0]$$

**Teorema 6.4** En el método de gradiente conjugado, eligiendo $d^0 = r^0 = b - Au^0$ y siempre que $r^0 \neq 0$ se tiene

$$[r^0, Ar^0, ..., A^k r^0] = [r^0, r^1, ..., r^k]$$

$$[r^0, Ar^0, ..., A^k r^0] = [d^0, d^1, ..., d^k]$$

**Demostración:** Para $k = 1$ como $d^0 = r^0$ y $d^1 = r^1 + \beta_1 d^0$ y por lo tanto $r^1 = d^1 - \beta_1 d^0$ podemos escribir

$$[r^0, r^1] = [d^0, d^1]$$

Por otra parte $r^1 = r^0 - \rho_0 A d^0$ con $\rho_0 = \frac{||r^0||^2}{(Ar^0, r^0)} \neq 0$ así pues $Ar^0 = Ad^0 = \frac{r^0 - r^1}{\rho_0}$ de donde

$$[r^0, Ad^0] = [r^0, Ar^0] = [r^0, r^1]$$

Las relaciones son ciertas para $k = 1$. Supongamos ahora que las relaciones son ciertas para $k$ y veamos que en ese caso también lo son para $k + 1$:

Tendremos por la hipótesis de inducción $r^k \in [r^0, Ar^0, ..., A^k r^0]$ y por otra parte $Ad^k \in A[r^0, Ar^0, ..., A^k r^0] = [Ar^0, A^2 r^0, ..., A^{k+1} r^0]$.

Como $r^{k+1} = r^k - \rho_k A d^k$ tenemos

$$r^{k+1} \in [r^0, Ar^0, ..., A^{k+1} r^0]$$

Recíprocamente demostraremos que $A^{k+1} r^0 \in [r^0, r^1, ..., r^{k+1}]$. En efecto, según se ha visto $r^{k+1}$ es una combinación lineal de términos de la forma $A^i r^0$ para $0 \leq i \leq k+1$, es decir,

$$r^{k+1} = \Sigma_{i=0}^{k+1} \gamma_i A^i r^0 = \Sigma_{i=0}^{k} \gamma_i A^i r^0 + \gamma_{k+1} A^{k+1} r^0$$

Veamos que podemos despejar el término $A^{k+1}r^0$. En efecto, $r^{k+1} \notin [r^0, Ar^0, ..., A^k r^0] = [d^0, d^1, ..., d^k]$ pues $r^{k+1}$ es ortogonal a $d^0, d^1, ..., d^k$ por el teorema anterior.

Podemos pues escribir

$$A^{k+1}r^0 = \frac{1}{\gamma_{k+1}}(r^{k+1} - \Sigma_{i=0}^{k}\gamma_i A^i r^0)$$

Gracias a la hipótesis de inducción

$$\Sigma_{i=0}^{k}\gamma_i A^i r^0 \in [r^0, ..., r^k]$$

de donde

$$A^{k+1}r^0 \in [r^0, ..., r^{k+1}]$$

resumiendo

$$[r^0, ..., r^{k+1}] = [r^0, Ar^0, ..., A^{k+1}r^0]$$

Análogamente se demuestra

$$[d^0, ..., d^{k+1}] = [r^0, Ar^0, ..., A^{k+1}r^0]$$

■

**Teorema 6.5** El valor $u^k$ obtenido en la $k-$ésima iteración del algoritmo de gradiente conjugado verifica

$$E(u^k) \leq E(v) \quad \forall v \in u^0 + \mathcal{K}_k$$

**Demostración:** Como $u^k = u^0 + \Sigma_{i=0}^{k-1}\rho_i d^i \in u^0 + \mathcal{K}_k$ para expresar que $E(u^k)$ es el mínimo de $E(v)$ sobre $u^0 + \mathcal{K}_k$, es necesario y suficiente que

$$E(u^k) \leq E(u^0 + v) \quad \forall v \in \mathcal{K}_k$$

es decir

$$(\nabla E(u^k), v) = 0 \quad \forall v \in \mathcal{K}_k$$

o sea

$$2(r^k, v) = 0 \quad \forall v \in \mathcal{K}_k$$

pero esto es cierto pues $(r^k, r^i) = 0$ para todo $i \leq k-1$ y según el teorema anterior $\mathcal{K}_k = [r^0, ..., r^{k-1}]$ ■

**Teorema 6.6** El valor $u^k$ obtenido en la $k-$ésima iteración del algoritmo de gradiente conjugado verifica

$$E(u^k) = \min_{P_{k-1} \in \mathcal{P}_{k-1}} (A(I - AP_{k-1}(A))e^0, (I - AP_{k-1}(A))e_0)$$

donde $\mathcal{P}_{k-1}$ es el espacio de polinomios de grado inferior o igual a $k-1$ y $e^0 = u^0 - u$.

**Demostración:** Todo $v = u^0 + \mathcal{K}_k$ de escribe $v = u_0 + P_{k-1}(A)r^0$ donde $P_{k-1}$ es un polinomio de grado menor o igual que $k-1$.

Tenemos

$$v - u = e^0 + P_{k-1}(A)r^0 = e^0 - P_{k-1}(A)Ae^0 = (I - AP_{k-1}(A))e^0$$

Por la definición de la función $E(.)$ podemos escribir

$$E(v) = (A(v-u), v-u) = (A(I - AP_{k-1}(A))e^0, (I - AP_{k-1}(A))e^0)$$

Como $E(u^k) = \text{mín}_{v \in u^0 + \mathcal{K}_k} E(v)$ tendremos

$$E(u^k) = \min_{P_{k-1} \in \mathcal{P}_{k-1}} (A(I - AP_{k-1}(A))e^0, (I - AP_{k-1}(A))e^0)$$

■

**Corolario 6.2** Se tiene la relación siguiente

$$E(u^k) \leq (\max_{1 \leq i \leq N} (1 - \lambda_i P_{k-1}(\lambda_i))^2))E(u^0)$$

para todo polinomio $P_{k-1}$ de grado menor o igual que $k-1$ y donde $\lambda_i$ para $i = 1, .., N$ son los valores propios de $A$.

**Demostración:** Siendo $A$ una matriz simétrica definida positiva admite una base ortonormal de vectores propios $(v_1, ..., v_N)$ correspondientes a los valores propios $\lambda_1, ..., \lambda_N$.

En esta base $e^0 = u^0 - u$ se escribe $e^0 = \Sigma_{i=1}^N a_i v_i$ y

$$E(u^0) = (Ae^0, e^0) = \Sigma_{i=0}^N a_i^2 \lambda_i$$

además

$$(I - AP_{k-1}(A))e^0 = \Sigma_{i=0}^N a_i (1 - \lambda_i P_{k-1}(\lambda_i))v_i$$

De donde para todo polinomio de grado menor o igual que $k-1$, tenemos:

$$\begin{aligned} E(u^k) &\leq (A\Sigma_{i=1}^N a_i(1 - \lambda_i P_{k-1}(\lambda_i))v_i, \Sigma_{i=1}^N a_i(1 - \lambda_i P_{k-1}(\lambda_i))v_i) \\ &= \Sigma_{i=1}^N (1 - \lambda_i P_{k-1}(\lambda_i))^2 a_i^2 \lambda_i \\ &\leq [\max_i (1 - \lambda_i P_{k-1}(\lambda_i))^2][\Sigma_{i=1}^N a_i^2 \lambda_i] \\ &= [\max_i (1 - \lambda_i P_{k-1}(\lambda_i))^2]E(u^0) \end{aligned}$$

■

**Corolario 6.3** El valor $u^k$ obtenido en la $k$-ésima iteración del método de Gradiente Conjugado verifica

$$E(u^k) \leq 4(\frac{\sqrt{\kappa(A)}-1}{\sqrt{\kappa(A)}+1})^{2k} E(u^0)$$

**Demostración:** Mayoraremos $\max_i(1-\lambda_i P_{k-1}(\lambda_i))^2$ por $\max_{\lambda_1\leq\lambda\leq\lambda_N}(1-\lambda P_{k-1}(\lambda))^2$

$1-\lambda P_{k-1}(\lambda)$ es un polinomio de grado menor o igual que $k$ que toma el valor 1 en $\lambda = 0$.

Podemos entonces elegir

$$1-\lambda P_{k-1}(\lambda) = \frac{T_k(\frac{\lambda_N+\lambda_1-2\lambda}{\lambda_N-\lambda_1})}{T_k(\frac{\lambda_N+\lambda_1}{\lambda_N-\lambda_1})}$$

donde $T_k$ es el polinomio de Tchebycheff de grado $k$.

Se obtiene entonces la mayoración

$$\begin{aligned} E(u^k) &\leq \frac{1}{T_k^2(\frac{\lambda_N+\lambda_1}{\lambda_N-\lambda_1})} E(u^0) \\ &\leq 4(\frac{\sqrt{\kappa(A)}-1}{\sqrt{\kappa(A)}+1})^{2k} E(u^0) \end{aligned}$$

con $\kappa(A) = \frac{\lambda_N}{\lambda_1}$ ■

**Ejercicios:**

Sea $A$ es simétrica y definida positiva de orden $N$. En los siguientes ejercicios consideramos la aplicación del método de gradiente conjugado para resolver un sistema de ecuaciones

$$Au = b$$

donde $A$ es simétrica y definida positiva.

(a) Dado $\epsilon > 0$ estimar el número de iteraciones que hay que realizar con el método de gradiente conjugado para obtener la solución con un error

$$E(u^n) = (A(u^n-u), u^n-u) \leq \epsilon(A(u^0-u), u^0-u)$$

en función del número de condicionamiento de la matriz $A$.

(b) Supongamos que el residuo inicial $r^0$ es un vector propio de la matriz $A$. Demostrar que el método de gradiente conjugado converge en una sola iteración.

(c) Supongamos que la matriz $A$ tiene todos los valores propios iguales. Demostrar que el método de gradiente conjugado converge en una sola iteración.

(d) Sea $d^0 = r^0 = \Sigma_{i=1}^m \alpha_i v_i$ donde $\{v_i\}_{i=1}^m$ son $m$ vectores propios ($m < N$). Demostrar que el método de gradiente conjugado converge al menos en $m$ iteraciones.

(e) Supongamos que los $N$ valores propios de $A$ están distribuidos de manera que los $N - m$ primeros valores propios están contenidos en un intervalo $[a, b]$, es decir,

$$0 < \lambda_1 \leq, ..., \leq \lambda_{N-m} \leq \lambda_{N-m+1} \leq, ..., \leq \lambda_N$$

$$\lambda_1, ..., \lambda_{N-m} \in [a, b]$$

Demostrar que en la iteración $n$-ésima

$$E(u^n) \leq 4 \left( \frac{\sqrt{\frac{b}{a}} - 1}{\sqrt{\frac{b}{a}} + 1} \right)^{2(n-m)} E(u^0)$$

## 6.5. Precondicionamiento

### 6.5.1. Introducción

Como hemos visto, la rapidez de convergencia en los métodos de gradiente y de gradiente conjugado depende del número de condicionamiento de la matriz $\kappa(A)$ de la matriz $A$ asociada al sistema de ecuaciones.

Cuánto más cercano a 1 sea este número más rápida es la convergencia.

La técnica de precondicionamiento consiste en remplazar la ecuación

$$Au = b$$

por una equivalente

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

$C^{-1}$ se elegir de modo que $\kappa(C^{-1}A)$ sea mucho más pequeño que $\kappa(A)$. En teoría la mejor elección es $C^{-1} = A^{-1}$ pues entonces $\kappa(C^{-1}A) = 1$. En la práctica habrá que encontrar $C^{-1}$ lo más próximo posible a $A^{-1}$ sin que el cálculo de $C^{-1}$ sea demasiado costoso.

### 6.5.2. Algoritmo de gradiente conjugado precondicionado

En general no se aplica directamente el algoritmo al sistema $C^{-1}Au = C^{-1}b$ pues aunque $C^{-1}$ sea una matriz simétrica no necesariamente lo será $C^{-1}A$.

Por ello se escribe el problema de la siguiente manera:

Si $C^{-1}$ es simétrica y definida positiva se puede definir su raiz cuadrada $C^{-1/2}$ simétrica y definida positiva, es decir, $(C^{-1/2})^2 = C^{-1}$.

En lugar de considerar el sistema

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

consideraremos , multiplicando por $C^{1/2}$ ambos lados,

$$C^{-1/2}AC^{-1/2}C^{1/2}u = C^{-1/2}b$$

e introduciendo una nueva variable $\tilde{u} = C^{1/2}u$, el sistema se escribe

$$\tilde{A}\tilde{u} = C^{-1/2}b$$

donde $\tilde{A} = C^{-1/2}AC^{-1/2}$. Aplicaremos el método de gradiente conjugado a este sistema de ecuaciones. Escribamos el algoritmo poniendo

- $\tilde{A} = C^{-1/2}AC^{-1/2}$
- $\tilde{u} = C^{1/2}u$, $\tilde{u}^k = C^{1/2}u^k$
- $\tilde{r}^k = C^{-1/2}b - \tilde{A}\tilde{u}^k = C^{-1/2}b - C^{-1/2}AC^{-1/2}C^{1/2}u^k = C^{-1/2}(b - Au^k) = C^{-1/2}r^k$
- $\tilde{d}^k = C^{1/2}d^k$

El algoritmo se escribe:

(a) $\rho_k = \frac{||\tilde{r}^k||}{(\tilde{A}\tilde{d}^k, \tilde{d}^k)}$

(b) $\tilde{u}^{k+1} = \tilde{u}^k + \rho_k\tilde{d}^k$

(c) $\tilde{r}^{k+1} - \rho_k\tilde{A}\tilde{d}^k$

(d) $\beta_{k+1} = \frac{||\tilde{r}^{k+1}||^2}{||\tilde{r}^k||^2}$

(e) $\tilde{d}^{k+1} = \tilde{r}^{k+1} + \beta_{k+1}\tilde{d}^k$

Expresado en función de las variables originales será:

(a) $\rho_k = \frac{(C^{-1}r^k, r^k)}{(Ad^k, d^k)}$

(b) $u^{k+1} = u^k + \rho_k d^k$

(c) $r^{k+1} = r^k - \rho_k A d^k$

(d) $\beta_{k+1} = \frac{(C^{-1}r^{k+1}, r^{k+1})}{(C^{-1}r^k, r^k)}$

(e) $d^{k+1} = C^{-1}r^{k+1} + \beta_{k+1}d^k$

La inversa de la matriz de precondicionamiento $C$ no se calcula explícitamente. Pra ello se introduce una variable $z^k$ y en lugar de calcular directamente $z^k = C^{-1}r^k$, se resuelve el sistema $Cz^k = r^k$.

Teniendo en cuenta esta última observación el algoritmo de gradiente conjugado precondicionado en su forma práctica será:

$C$ , matriz de precondicionamiento.

$u^0$, $r^0 = b - Au^0$, $Cd^0 = r^0$, $z^0 = d^0$

Para $k = 0, 1, ...$

(a) $\rho_k = \frac{(r^k, z^k)}{(Ad^k, d^k)}$

(b) $u^{k+1} = u^k + \rho_k d^k$

(c) $r^{k+1} - \rho_k A d^k$

(d) $Cz^{k+1} = r^{k+1}$

(e) $\beta_{k+1} = \frac{(r^{k+1}, z^{k+1})}{(r^k, z^k)}$

(f) $d^{k+1} = z^{k+1} + \beta_{k+1}d^k$

Observemos que en cada iteración hay que resolver un sistema de ecuaciones asociado a la matriz $C$. En la práctica se elige $C$ de manera que la resolución de este sistema sea mucho ms fácil que la resolución del sistema original. Este es el caso en que la matriz $C$ es diagonal, o bien se dispone de la factorización de $C = RR^t$ en una matriz triangular superior y su correspondiente transpuesta.

Ejemplos de matrices de precondicionamiento son los siguientes:

- Precondicionador diagonal: Se elige como matriz de precondicionamiento la diagonal de la matriz $A$
- Precondicionadores basados en los métodos iterativos lineales ( Jacobi, SSOR)
- Precondicionadores basados en la factorización incompleta de Cholesky: Se evita el llenado de la matriz manteniendo la estructura de huecos de la matriz original total o parcialmente.

- Precondicionadores multimalla: Generalmente estn ligados al origen del problema, normalmente un problema de Ecuaciones en Derivadas Parciales, aunque también existen precondicionadores multimalla algebraicos.

## 6.6. Anexo: Polinomios de Tchebycheff

**Definición 6.3** Se llama polinomio de Tchebycheff de grado $n$, al polinomio $T_n$ definido por la relación de recurrencia

$$\begin{aligned} T_0(x) &= 1 \\ T_1(x) &= x \\ T_n(x) &= 2xT_{n-1}(x) - T_{n-2}(x) \quad \text{para} \quad n \geq 2 \end{aligned}$$

**Propiedad 6.3** $T_n$ viene dado por

$$\begin{aligned} T_n(x) &= \cos(n \; arc\cos x) \quad \text{para} \quad |x| \leq 1 \\ T_n(x) &= \cosh(n \; arg\cosh x) \quad \text{para} \quad x > 1 \\ T_n(x) &= (-1)^n T_n(x) \quad \text{para} \quad x < -1 \end{aligned}$$

**Demostración:** Para $|x| \leq 1$, pongamos $x = \cos\theta$. Tenemos

$$T_n(x) = \cos n\theta = 2\cos\theta\cos(n-1)\theta - \cos(n-2)\theta$$

que se comprueba fácilmente utilizando el desarrollo del coseno de suma de dos ángulos.

Para $x > 1$ se comprueba análogamente utilizando las funciones hiperbólicas. ■

**Ejemplos**

$$\begin{aligned} T_0(x) &= 1 \\ T_1(x) &= x \\ T_2(x) &= 2x^2 - 1 \\ T_3(x) &= 4x^3 - 3x \\ T_4(x) &= 8x^4 - 8x^2 + 1 \end{aligned}$$

Observamos que el coeficiente del término de grado $n$ es $2^{n-1}$, de modo que el polinomio $T_n(x)/2^{n-1}$ tiene el coeficiente del término de mayor grado igual a 1.

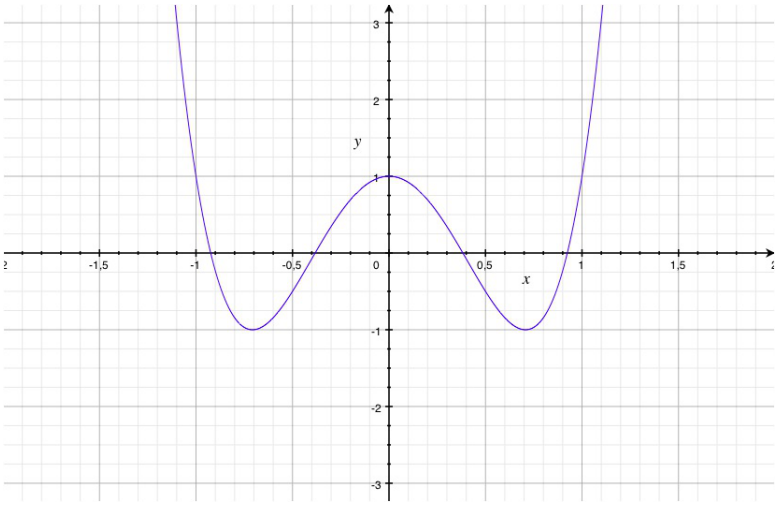


Figura 6.1: Polinomio de Tchebycheff de grado 4

**Propiedad 6.4** Para $n \geq 0$

$$T_n(x) = \frac{1}{2}\left((x + \sqrt{x^2-1})^n + (x - \sqrt{x^2-1})^n\right)$$

**Demostración:** Para $|x| \leq 1$, ponemos $x = \cos\theta$ y utilizamos

$$\cos n\theta = \frac{e^{in\theta} + e^{-in\theta}}{2}$$

junto con

$$e^{in\theta} = (\cos\theta + i\sin\theta)^n = (x - \sqrt{x^2-1})^n$$

$$e^{-in\theta} = (\cos\theta - i\sin\theta)^n = (x + \sqrt{x^2-1})^n$$

Para $x > 1$ ponemos $x = \cosh\theta$ y utilizamos

$$\cosh n\theta = \frac{e^{n\theta} + e^{-n\theta}}{2}$$

■

**Propiedad 6.5** Para $b > a > 0$ tenemos

$$T_n(\frac{b+a}{b-a} \geq \frac{1}{2}\left(\frac{\sqrt{\frac{b}{a}}+1}{\sqrt{\frac{b}{a}}-1}\right)^n$$

**Demostración:** Aplicamos la propiedad anterior y observamos

$$T_n(\frac{b+a}{b-a}) \geq \frac{1}{2}\left(\frac{b+a}{b-a}+\sqrt{\frac{(b+a)^2-(b-a)^2}{(b-a)^2}}\right)^n = \frac{1}{2}\left(\frac{\sqrt{\frac{b}{a}}+1}{\sqrt{\frac{b}{a}}-1}\right)^n$$

■

**Propiedad 6.6** $T_n$ tiene $n$ raíces en $[-1,1]$, $x_k = \cos\frac{2k-1}{n}\frac{\pi}{2} \quad k = 1, ..., n$.

**Demostración:**

Para que

$$T_n(x_k) = \cos n\theta_k = 0$$

es decir, como $x_k = \cos\theta_k$

$$n\theta_k = (2k-1)\frac{\pi}{2} \quad k = 1, ..., n$$

y finalmente

$$x_k = \cos\frac{2k-1}{n}\frac{\pi}{2}$$

■

**Propiedad 6.7** $T_n$ tiene $n+1$ extremos relativos en $[-1,1]$, $x_k^{'} = \cos\frac{k\pi}{n} \quad k = 0, 1, ..., n$ para los cuales $T_n(x_k^{'}) = (-1)^k$

**Demostración:**

Para que $T_n(x_k^{'}) = (-1)^k$

$$x_k^{'} = \cos\frac{k\pi}{n} \quad k = 0, 1, ..., n$$

■

**Teorema 6.7** Propiedad de optimalidad 1

Sea $P_n$ el conjunto de polinomios de grado $n$ cuyo coeficiente de $x^n$ es 1, entonces el polinomio $\frac{T_n}{2^{n-1}}$ verifica

$$\max_{-1\leq x\leq 1}\frac{|T_n(x)|}{2^{n-1}} \leq \max_{-1\leq x\leq 1}|p(x)| \quad \forall p \in P_n$$

**Demostración:**

$\frac{T_n}{2^{n-1}}$ es un elemento de $P_n$ que toma sus valores extremos $\frac{(-1)^k}{2^{n-1}}$, $n+1$ veces en los puntos $x_k^{'} = \cos\frac{k\pi}{n} \quad k = 0, 1, ..., n$.

Por reducción al absurdo supongamos que existe $p \in P_n$ tal que

$$\max_{-1\leq x\leq 1} |p(x)| < \frac{1}{2^{n-1}}$$

Sea $r = \frac{T_n}{2^{n-1}} - p$ que es un polinomio de grado menor o igual que $n-1$.

Entonces $r(x_k^{'}) = \frac{T_n(x_k^{'})}{2^{n-1}} - p(x_k^{'})$ tiene el mismo signo que $(-1)^k$ ya que $|p(x_k^{'})| < \frac{1}{2^{n-1}}$. $r$ cambia de signo $n$ veces en $[-1, 1]$ y tiene por tanto al menos $n$ ceros, y por lo tanto $r = 0$, al ser un polinomio de grado menor o igual que $n-1$. ■

**Teorema 6.8** Propiedad de optimalidad 2

Sea $F_n$ el conjunto de polinomios de grado $n$ tal que $p(\alpha) = 1$ para $|\alpha| > 1$, entonces el polinomio $\frac{T_n}{T_n(\alpha)}$ verifica

$$\max_{-1\leq x\leq 1} \frac{|T_n(x)|}{T_n(\alpha)} \leq \max_{-1\leq x\leq 1} |p(x)| \quad \forall p \in F_n$$

**Demostración:**

$T_n(\alpha) \neq 0$ entonces $\frac{T_n}{T_n(\alpha)} \in F_n$ y

$$\max_{-1\leq x\leq 1} |\frac{T_n(x)}{T_n(\alpha)}| = \frac{1}{T_n(\alpha)}$$

Por reducción al absurdo supongamos que existe $p \in F_n$ tal que

$$\max_{-1\leq x\leq 1} |p(x)| < \frac{1}{|T_n(\alpha)|}$$

Entonces el polinomio $r = \frac{T_n}{T_n(\alpha)} - p$ es un polinomio de grado menor o igual que $n$ que se anula para $x = \alpha$.

Además $T_n(\alpha)r(x_k^{'}) = T_n(x_k^{'}) - T_n(\alpha)p(x_k^{'})$ tiene el mismo signo que $(-1)^k$ para $k = 0, 1, ..., n$. Es decir, tiene al menos $n$ raices en $[-1, 1]$, por tanto $r$ tiene al menos $n+1$ raices y es de grado menor o igual que $n$, en consecuencia $r = 0$. ■

**Corolario 6.4** Sea $G_n$ el conjunto de polinomios de grado $n$ tal que $p(\alpha) = 1$ con $\alpha \notin [a, b]$, $0 < a < b$. Entonces el polinomio

$$q(x) = \frac{T_n(\frac{b+a-2x}{b-a})}{T_n(\frac{b+a-2\alpha}{b-a})}$$

verifica

$$\max_{a\leq x\leq b} |q(x)| \leq \max_{a\leq x\leq b} |p(x)| \quad \forall p \in G_n$$

y

$$\max_{a\leq x\leq b} |q(x)| = \frac{1}{T_n(\frac{b+a-2\alpha}{b-a})}$$

**Demostración:**

Hacemos el cambio de variable

$$\xi = \frac{b+a-2x}{b-a}$$

y aplicamos el teorema anterior. ■

# Capítulo 7

# Cálculo de valores y vectores propios

## 7.1. El método de Jacobi

### 7.1.1. Introducción

El método de Jacobi se emplea para calcular todos los valores propios de una matriz simétrica. Recordemos que si una matriz $A$ es simétrica es diagonalizable y todos sus valores propios son reales. Es decir, si $A$ es simétrica existe una matriz ortogonal $Q$ tal que

$$Q^t AQ = \Lambda = diag(\lambda_1, \lambda_2, ..., \lambda_N)$$

donde $\lambda_1, \lambda_2, ..., \lambda_N$ son los valores propios de $A$ que son reales.

Recordemos que los vectores columna de la matriz $Q$ forman un sistema ortogonal de vectores propios de $A$. El i-ésimo vector columna es un vector propio asociado al valor propio $\lambda_i$.

La idea del método de Jacobi consiste en construir una sucesión de matrices $(Q_k)_k$ ortogonales "elementales" ( es decir sencillas) de manera que la sucesión de matrices, también simétricas,

$$A_{k+1} = Q_k^t A_k Q_k = (Q_1 Q_2 ... Q_k)^t A (Q_1 Q_2 ... Q_k)$$

converga hacia la matriz $diag(\lambda_1, \lambda_2, ..., \lambda_N)$.

En general el método se usa para calcular los valores propios de matrices simétricas de tamaño moderado. Si se quiere calcular los vectores propios asociados a un valor propio se puede usar el método de la potencia inversa con traslación.

### 7.1.2. Descripción del método de Jacobi

El principio de cada transformación elemental $A_k \to A_{k+1}$ es anular dos elementos fuera de la diagonal, en posición simétrica, sean $(A_k)_{pq}$ y $(A_k)_{qp}$ de la matriz $A_k$, siguiendo un procedimiento muy sencillo. Para simplificar la notación designamos mediante $A_k = (a_{ij})$ y $A_{k+1} = (a_{ij})$ a los términos de dos matrices sucesivas. Elegimos $Q_k$ de la forma (matriz de rotación)

$$Q_k = \begin{bmatrix} 1 & & & & & & & \\ & 1 & & & & & & \\ & & \cos\theta & & & \sin\theta & & \\ & & & 1 & & & & \\ & & & & 1 & & & \\ & & -\sin\theta & \ldots & & \cos\theta & & \\ & & & & & & 1 & \\ & & & & & & & 1 \end{bmatrix}$$

donde los términos distintos de cero fuera de la diagonal están en las filas $p$ columna $q$ y fila $q$ y columna $p$ respectivamente. Ajustamos el parámetro $\theta$ de modo que al hacer el producto

$$Q_k^t A_k Q_k$$

los correspondientes términos $b_{pq} = b_{qp} = 0$

La matriz $A_{k+1}$ diferirá de la matriz $A_k$ solo en las filas y columnas $p$ y $q$. Como la matriz es simétrica basta efectuar los cálculos únicamente sobre las líneas $p$ y $q$ de la matriz $A_k$ y deducir por simetría el valor de los términos de las columnas $p$ y $q$.

**Teorema 7.1** (a) Si $A$ una matriz simétrica y $Q$ una matriz de rotación, la matriz

$$B = Q^t A Q = (b_{ij})$$

que es simétrica verifica

$$\Sigma_{i,j=1}^N b_{ij}^2 = \Sigma_{i,j=1}^N a_{ij}^2$$

(b) Si $a_{pq} \neq 0$, existe un único valor de $\theta \in (-\frac{\pi}{4}, 0) \bigcup (0, \frac{\pi}{4})$ tal que $b_{pq} = 0$ y para este valor se tiene entonces

$$\Sigma_{i=1}^N b_{ii}^2 = \Sigma_{i=1}^N a_{ii}^2 + 2a_{pq}^2$$

**Demostración:**

(a) Se verifica fácilmente que la matriz $Q$ es ortogonal. Debido a la invarianza por transformación ortogonal de la norma de Frobenius

$$\Sigma_{i,j} a_{ij}^2 = ||A||_E^2 = ||Q^t A Q||_E = \Sigma_{ij} b_{ij}^2$$

(b) La tranformación de los elementos $(p,p),(p,q),(q,p),(q,q)$ se escribe

$$\begin{bmatrix} b_{pp} & b_{pq} \\ b_{qp} & b_{qq} \end{bmatrix} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} a_{pp} & a_{pq} \\ a_{qp} & a_{qq} \end{bmatrix} \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix}$$

y mediante el razonamiento de la primera parte

$$a_{pp}^2 + a_{qq}^2 + 2a_{pq}^2 = b_{pp}^2 + b_{qq}^2 + 2b_{pq}^2$$

para todo valor de $\theta$. Como

$$b_{pq} = b_{qp} = a_{pq}\cos 2\theta + \frac{a_{pp} - a_{qq}}{2}\sin 2\theta$$

si elegimos $\theta$ de manera que $b_{pq} = b_{qp} = 0$ es decir, $\theta$ solución de la ecuación

$$\operatorname{cotg} 2\theta = \frac{a_{qq} - a_{pp}}{2a_{pq}}$$

tendremos

$$a_{pp}^2 + a_{qq}^2 + 2a_{pq}^2 = b_{pp}^2 + b_{qq}^2$$

Como por otra parte $a_{ii} = b_{ii}$ para $i \neq p$ y $i \neq q$, se obtiene el resultado.

■

### 7.1.3. Cálculo de los términos de la transformación elemental de Jacobi

Efectuando el producto de matrices $B = Q^t AQ$ obtenemos para todo valor del angulo $\theta$

$$\begin{aligned}
b_{ij} &= a_{ij} \quad \text{si} \quad i \neq p,q \quad \text{y} \quad j \neq p,q \\
b_{pi} &= a_{pi}\cos\theta - a_{qi}\sin\theta \quad \text{si} \quad i \neq p,q \\
b_{qi} &= a_{pi}\sin\theta + a_{qi}\cos\theta \quad \text{si} \quad i \neq p,q \\
b_{pp} &= a_{pp}\cos^2\theta + a_{qq}\sin^2\theta - a_{pq}\sin 2\theta \\
b_{qq} &= a_{pp}\sin^2\theta + a_{qq}\cos^2\theta + a_{pq}\sin 2\theta \\
b_{pq} &= b_{qp} = a_{pq}\cos 2\theta + \frac{a_{pp} - a_{qq}}{2}\sin 2\theta
\end{aligned}$$

Llamando $c = \cos\theta$, $s = \sin\theta$ y $t = \tan\theta$ tenemos, gracias a las relaciones entre las funciones trigonométricas, que los elementos de la matriz $B$ están determinados por relaciones algebraicas a partir de los elementos de $A$. De la condición $b_{pq} = b_{qp} = 0$ resulta

$$\nu = \operatorname{cotg} 2\theta = \frac{a_{qq} - a_{pp}}{2a_{pq}}$$

y calculando la raiz de modulo más pequeo del trinomio $t^2 + 2\nu t - 1 = 0$ ( pues $t = \tan\theta$ y $|\theta| \leq \frac{\pi}{4}$) obtenemos sucesivamente las cantidades

$$\begin{aligned} t = -\nu + \mathrm{sg}(\nu)\sqrt{1+\nu^2} \quad &\text{si} \quad \nu \neq 0 \\ t = 1 \quad &\text{si} \quad \nu = 0 \\ c = \frac{1}{\sqrt{1+t^2}} & \\ s = ct & \end{aligned}$$

Los elementos de $B$ vienen dados por las expresiones (ejercicio)

$$\begin{aligned} b_{pi} &= ca_{pi} - sa_{qi} \quad \text{si} \quad i \neq p, q \\ b_{qi} &= ca_{qi} + sa_{pi} \quad \text{si} \quad i \neq p, q \\ b_{pp} &= a_{pp} - ta_{pq} \\ b_{qq} &= a_{qq} + ta_{pq} \end{aligned}$$

### 7.1.4. Algoritmo de Jacobi

(a) Jacobi clásico: En cada paso de eligen $p$ y $q$ de modo que

$$|a^k_{pq}| = \max_{i \neq j} |a^k_{ij}|$$

(b) Jacobi cíclico: Se anulan $a_{pq}$ según el orden

$$(1,2), (1,3), ...., (1,N); (2,3), ..., (2,N); ...(N, N-1)$$

(c) Jacobi con umbral: como en el algortimo de Jacobi cíclico pero saltándose los elementos $a_{pq}$ tales que $|a_{pq}| < \varepsilon$.

**Teorema 7.2** de convergencia: Puesto que la suma de cuadrados de todos los elementos de las matrices $A_k$ permanece constante y en cada iteración la suma de cuadrados de los elementos diagonales aumenta en una cantidad igual a la suma de cuadrados de dos elementos que se transforman en cero, las matrices $A_k$ convergen hacia una matriz diagonal. Esta matriz será la matriz $diag\{\lambda_{\sigma(i)}\}$ para una permutación $\sigma$ de los índices $i = 1, ..., N$.

## 7.2. Matrices tridiagonales simétricas y el método de la bisección

Vamos a ver en esta sección cómo calcular los valores propios de una matriz tridiagonal simétrica. Este método se podrá combinar con los métodos de tridiagonalización de Givens o de Housholder para calcular los valores propios de matrices generales simétricas.

Sea $A$ una matriz tridiagonal simétrica y real

$$A = \begin{bmatrix} a_1 & b_1 & & & \\ b_1 & a_2 & b_2 & & \\ & \ddots & \ddots & \ddots & \ddots \\ & & & a_{N-1} & b_{N-1} \\ & & & b_{N-1} & a_N \end{bmatrix}$$

$A$ es irreducible si $b_i \neq 0$ para $i = 1, 2, ..., N-1$. Cuando la matriz $A$ no sea irreducible, se puede descomponer por bloques bajo la forma

$$A = \begin{bmatrix} A_{11} & 0 \\ 0 & A_{22} \end{bmatrix}$$

y en ese caso los valores propios de $A$ son los de $A_{11}$ y los de $A_{22}$. Vamos a considerar pues solo el caso de matrices irreducibles.

Los valores propios de $A$ son las raíces del polinomio característico $P(\lambda) = \det(A - \lambda I)$.

Designemos mediante $(A - \lambda I)_k$ el menor principal de $A - \lambda I$ de orden $k$, es decir

$$(A - \lambda I)_k = \begin{bmatrix} a_1 - \lambda & b_1 & & & & \\ b_1 & a_2 - \lambda & b_2 & & & \\ & \ddots & \ddots & \ddots & \ddots & \\ & & & a_{k-1} - \lambda & b_{k-1} & \\ & & & b_{k-1} & a_k - \lambda & \end{bmatrix}$$

y sea $P_k$ el polinomio característico de $A_k$. Se verifica la relación de recurrencia siguiente

**Lema 7.1**

$$P_0(\lambda) = 1, \quad P_1(\lambda) = a_1 - \lambda$$

y para $k \geq 2$, los polinomios $P_k$ verifican

$$P_k(\lambda) = (a_k - \lambda)P_{k-1}(\lambda) - b_{k-1}^2 P_{k-2}(\lambda)$$

**Demostración:** Para $k \geq 2$ basta desarrollar el determinante de $A_k - \lambda I_k$ por la última fila. En efecto:

$$\det(A - \lambda I)_k = (a_k - \lambda)\det(A - \lambda I)_{k-1} - b_{k-1}\det B_k$$

donde

$$B_k = \begin{bmatrix} a_1 - \lambda & b_1 & & & & & \\ b_1 & a_2 - \lambda & b_2 & & & & \\ & \ddots & \ddots & \ddots & \ddots & & \\ & & & a_{k-3} - \lambda & b_{k-2} & 0 & \\ & & & b_{k-3} & a_{k-2} - \lambda & 0 & \\ & & & & & b_{k-1} & \end{bmatrix}$$

de donde desarrollando por la última columna

$$\det B_k = b_{k-1} P_{k-2}(\lambda)$$

definiendo $P_0(\lambda) = 1$ y calculando $P_1(\lambda) = a_1 - \lambda$ obtenemos el resultado deseado para $P_k(\lambda)$. ■

**Lema 7.2** Las raices de $P_k$ son simples y reales.

**Demostración:** Las raíces son reales pues son valores propios de una matriz simétrica. La submatriz se $(A - \lambda I)_k$ formada por las filas 2 a $k$ y columnas 1 a $k-1$ es triangular superior. Por ejemplo, en el caso $k = 6$

$$(A - \lambda I)_6 = \begin{bmatrix} a_1 - \lambda & b_1 & 0 & 0 & 0 & 0 \\ b_1 & a_2 - \lambda & b_2 & 0 & 0 & 0 \\ 0 & b_2 & a_3 - \lambda & b_3 & 0 & 0 \\ 0 & 0 & b_3 & a_4 - \lambda & b_4 & 0 \\ 0 & 0 & 0 & b_4 & a_5 - \lambda & b_5 \\ 0 & 0 & 0 & 0 & b_5 & a_6 - \lambda \end{bmatrix}$$

Sus elementos diagonales son para $i = 1$ a $k-1$ los $b_i$ que son no nulos por hipótesis. En consecuencia, para todo valor de $\lambda$, existe una submatriz cuadrada de orden $k-1$ regular extraida de $(A - \lambda I)_k$ (se obtiene quitando la primera fila y la última columna). El rango de $(A - \lambda I)_k$ es pues superior o igual a $k-1$. de modo que la dimensión de $N(A - \lambda I)_k \leq 1$. Esta dimensión vale 1 si $\lambda$ es valor propio. Como $A_k$ es diagonalizable pues es simétrica las raices de $P_k$ son simples. ■

**Lema 7.3** Los polinomios sucesivos $P_k$ no tienen ninguna raiz común.

**Demostración:** Supongamos que $\alpha$ es una raiz de $P_k$ y de $P_{k-1}$. En la relación

$$P_k(\lambda) = (a_k - \lambda) P_{k-1}(\lambda) - b_{k-1}^2 P_{k-2}(\lambda)$$

haciendo $\lambda = \alpha$

$$b_{k-1}^2 P_{k-2}(\alpha) = (a_k - \alpha) P_{k-1}(\alpha) - P_k(\alpha) = 0$$

obtendríamos razonando por inducción que $P_0 = 1$ admite $\alpha$ como raiz, lo que es imposible. ■

**Teorema 7.3** Los polinomios característicos $P_k$ verifican las siguientes propiedades:

(a) $\lim_{\lambda \to \infty} P_k(\lambda) = +\infty$ para $1 \leq k \leq N$

(b) $P_k(\lambda_0) = 0 \quad \Rightarrow P_{k-1}(\lambda_0).P_{k+1}(\lambda_0) < 0 \quad 1 \leq k \leq N-1$

(c) El polinomio $P_k(\lambda)$ posee k raices distintas reales que separan las $k+1$ raices del polinomio $P_{k+1} \quad 1 \leq k \leq N-1$

**Demostración:**

(a) Por inducción, vemos que el coeficiente de mayor grado de $P_k(\lambda)$ es $(-1)^k$, en efecto

- $P_1(\lambda) = a_1 - \lambda$, que tiene coeficiente de $\lambda$ $(-1)^1$
- $P_k(\lambda) = (a_1 - \lambda)P_{k-1}(\lambda) - b_{k-1}P_{k-2}(\lambda)$

Si el coeficiente de $P_{k-1}(\lambda)$ dde mayor grado es $(-1)^{k-1}\lambda^{k-1}$ el de $P_k(\lambda)$ será $(-\lambda)(-1)^{k-1}\lambda^{k-1}0(-1)^k\lambda^k$.

(b) Supongamos que $P_k(\lambda_0) = 0$, entonces

$$P_k(\lambda_0) = (a_k - \lambda)P_{k-1}(\lambda_0) - b_{k-1}^2 P_{k-2}(\lambda_0)$$

poniendo en el lugar de $k$, $k+1$

$$P_{k+1}(\lambda_0) = (a_{k+1} - \lambda)P_k(\lambda_0) - b_k^2 P_{k-1}(\lambda_0)$$

Si $P_k(\lambda_0) = 0$, resulta

$$P_{k+1}(\lambda_0) = -b_k^2 P_{k-1}(\lambda_0)$$

es decir, $P_{k+1}(\lambda_0)$ y $P_{k-1}(\lambda_0)$ son de signos opuestos, pues hemos visto que $\lambda_0$ no puede ser una raiz común de $P_{k+1}$, $P_k$ y $P_{k-1}$.

(c) Es una consecuencia de la parte 1 y de la parte 2.

En efecto, sean $P_0$ y $P_1$. Dibujemos la gráfica de $P_2$. Tenemos $P_2(\lambda) \to \infty$ cuando $\lambda \to \infty$. Sea $a_1$ la raiz de $P_1$. Para $\lambda = a_1$ , resulta $P_2(a_1) < 0$.

Sean $b_1 < b_2$ las dos raíces de $P_2$. Como $P_3(\lambda) \to \infty$ cuando $\lambda \to \infty$ en $b_1$, $P_3(b_1) < 0$ y en $b_2$, $P_3(b_2) > 0$. Razonado inductivamente obtenemos el resultado para todo $k$.

Una sucesión de polinomios verificando las propiedades 1), 2) y 3) del teorema anterior se llama sucesión de Sturm. La siguiente propiedad de las sucesiones de Sturm permite localizar los valores propios.

**Propiedad 7.1** El número de cambios de signo de la sucesión

$$P_0(x), P_1(x), ..., P_N(x)$$

es igual al número de raices del polinomio $P_N$ que son estrictamente menores que $x$. Si $P_l(x) = 0$ se pone $\text{sgn}(P_l(x)) = \text{sgn}(P_{l-1}(x))$.

En consecuencia, el número de valores propios de la matriz $A$ comprendidos en el intervalo $[a, b[$ es igual $w(b) - w(a)$ donde $w(x)$ es igual al número de cambios de signo de la sucesión

$$P_0(x), P_1(x), ..., P_N(x)$$

.

## 7.3. El método de Householder-bisección

### 7.3.1. Introducción

El método de Householder-bisección es un método de cálculo de valores propios particularmente bien adaptado para la búsqueda de valores propios seleccionados de una matriz simétrica, por ejemplo, todos los valores propios situados en un intervalo determinado de antemano. Además este método permite aproximar cada valor propio con una precisión variable según el rango del valor propio considerado, a gusto del usuario. Sin embargo este método no proporciona los vectores propios que habrá que calcularlos por otro método, por ejemplo mediante el método de la potencia inversa con traslación.

El método de Householder-bisección comprende dos etapas:

(a) Dada una matriz simétrica $A$, se determina una matriz ortonormal $P$, de un tipo particular tal que la matriz simétrica $P^tAP$ sea tridiagonal. Esta etapa que solo necesita un número finito de operaciones elementales, se realiza mediante el método de Householder de reducción de una matriz simétrica a la forma tridiagonal.

(b) Se llega así al problema del cálculo de valores propios de una matriz simétrica tridiagonal que se efectúa mediante el método de la bisección.

### 7.3.2. Método de tridiagonalización de Householder

Dada una matriz simétrica $A = A_1$ de orden $d$, se construyen paso a paso $d-2$ matrices ortogonales $H_1, H_2, ..., H_{d-2}$ tales que las matrices

$$A_k = H_{k-1}A_{k-1}H_{k-1} = (H_1H_2...H_{k-1})^tA(H_1H_2...H_{k-1})$$

para $1 \leq k \leq d-2$ sean de la forma

$$A_k = \begin{bmatrix}
x & x & & & & & & & & & \\
x & x & x & & & & & & & & \\
 & x & x & x & & & & & & & \\
 & & x & x & x & & & & & & \\
 & & & x & x & \mathbf{x} & \mathbf{x} & \mathbf{x} & \mathbf{x} & \mathbf{x} & \mathbf{x} \\
 & & & & \mathbf{x} & x & x & x & x & x & x \\
 & & & & \mathbf{x} & x & x & x & x & x & x \\
 & & & & \mathbf{x} & x & x & x & x & x & x \\
 & & & & \mathbf{x} & x & x & x & x & x & x \\
 & & & & \mathbf{x} & x & x & x & x & x & x \\
 & & & & \mathbf{x} & x & x & x & x & x & x
\end{bmatrix}$$

donde se han marcado en negrita parte de la $k-$ésima fila y de la $k-$ésima columna.

Se obtiene así la matriz

$$A_{d-2} = (H_1 H_2 ... H_{d-2})^t A (H_1 H_2 ... H_{d-2})$$

es tridiagonal y semejante a la matriz dada $A$. Por otra parte la matriz ortogonal buscada es

$$P = H_1 H_2 ... H_{d-2}$$

En este método, cada transformación $A_k \to A_{k+1} = H_k^t A_k H_k$ se efectúa con la ayuda de una matriz de la forma

$$H_k = \begin{bmatrix} I_k & 0 \\ 0 & \tilde{H}_k \end{bmatrix}$$

donde $I_k$ es la matriz unidad de orden $k$.

En estas condiciones las propiedades de la multiplicación por bloques de las matrices muestran:

$$\begin{bmatrix} I_k & 0 \\ 0 & \tilde{H}_k^t \end{bmatrix} \begin{bmatrix} x & x & & & & & & & & & & \\ x & x & x & & & & & & & & & \\ & x & x & x & & & & & & & & \\ & & x & x & x & & & & & & & \\ & & & x & x & & & \mathbf{a}_k^t & & & & \\ & & & & & x & x & x & x & x & x & \\ & & & & & x & x & x & x & x & x & \\ & & & & \mathbf{a}_k & x & x & x & x & x & x & \\ & & & & & x & x & x & x & x & x & \\ & & & & & x & x & x & x & x & x & \\ & & & & & x & x & x & x & x & x & \end{bmatrix} \begin{bmatrix} I_k & 0 \\ 0 & \tilde{H}_k \end{bmatrix} =$$

$$\begin{bmatrix} x & x & & & & & & & & & & \\ x & x & x & & & & & & & & & \\ & x & x & x & & & & & & & & \\ & & x & x & & x & & & & & & \\ & & & x & & x & & \mathbf{a}_k^t \tilde{H}_k & & & & \\ & & & & & & x & x & x & x & x & x \\ & & & & & & x & x & x & x & x & x \\ & & & & \tilde{H}_k^t \mathbf{a}_k & & x & x & x & x & x & x \\ & & & & & & x & x & x & x & x & x \\ & & & & & & x & x & x & x & x & x \\ & & & & & & x & x & x & x & x & x \end{bmatrix}$$

Vamos abuscar $\tilde{H}_k$ de manera que $\tilde{H}_k \mathbf{a}_k \in \Re^{d-k}$ tenga todas sus componenetes nulas salvo la primera y de manera que $\tilde{H}_k$ sea lo más sencilla posible y ortogonal. Las llamadas matrices de Hoseholder proporcionan la respuesta.

**Definición 7.1** Matriz de Householder

Se llama matriz de Householder una matriz de la forma

$$H(v) = I - 2\frac{vv^t}{v^t v} \quad \text{para} \quad v \neq 0$$

**Ejercicio**: Comprobar que $H^t(v)H(v) = I$ y que $H(v)$ es una matriz simétrica.

**Teorema 7.4** Sea $a = (a_1, a_2, ...a_d)^t$ un vector de $\mathbb{R}^d$ tal que $\Sigma_{i=2}^d |a_i| > 0$. Existe una matriz de Householder $H$ tal que las $d-1$ últimas componentes del vector $Ha$ son nulas. De manera más concreta

$$H(a + ||a||_2 e_1)a = -||a||_2 e_1$$

$$H(a - ||a||_2 e_1)a = +||a||_2 e_1$$

donde $e_1 = (1, 0, ..., 0)^t$.

**Demostración:** Como $\Sigma_{i=2}^d |a_i| > 0$ los vectores $a \pm ||a||_2 e_1 \neq 0$

$$H(a \pm ||a||_2 e_1)a = a - \frac{2(a \pm ||a||_2 e_1)(a^t \pm ||a||_2 e_1^t)a}{(a^t \pm ||a||_2 e_1^t)(a \pm ||a||_2 e_1)}$$

como

$$(a \pm ||a||_2 e_1)(a^t \pm ||a||_2 e_1^t)a = ||a||(||a|| \pm a_1)(a \pm ||a||e_1)$$

y

$$(a^t \pm ||a||_2 e_1^t)(a \pm ||a||_2 e_1) = 2||a||(||a|| \pm a_1)$$

de donde

$$H(a \pm ||a||_2 e_1)a = a - \frac{2||a||(||a|| \pm a_1)(a \pm ||a||e_1)}{2||a||(||a|| \pm a_1)} = \mp||a||_2 e_1$$

■

**Nota 7.1** Las matrices $H(v)$ no se calculan explícitamente, pues en general lo que se necesita es el producto de una matriz de Houseloder por un vector. Los cálculos se realizan de la siguiente manera:

Dado $a \in \mathbb{R}^d$. Supongamos que queremos encontrar la matriz de Householder $H = H(a \pm ||a||e_1)$ que transforma $a$ en el vector $\pm||a||e_1$ y calcular el producto $Hb$ para un vector dado $b \in \mathbb{R}^d$. Dispondremos los cálculos de la siguiente forma:

(a) Se calcula $||a||$

(b) Se calcula $v = a \pm ||a||e_1$

(c) Se calcula $\frac{v^t v}{2} = ||a||(||a|| \pm a_1)$

(d) Dado un vector $b \in \mathbb{R}^d$ se calcula $v^t b$ y finalmente

$$Hb = b - \frac{v^t b}{v^t v/2} v$$

**Nota 7.2** El signo + o − se elige de forma que la cantidad que aparezca en el denominador, es decir, $v^t v/2$ sea lo más grande posible, para evitar errores de redondeo. Elegimos pues,

$$v = a + ||a||e_1 \quad \text{si} \quad a_1 > 0$$

y

$$v = a - ||a||e_1 \quad \text{si} \quad a_1 < 0$$

### 7.3.3. Aplicación al método de tridiagonalización de Householder

En el paso $k-$ésimo elegimos $H_k = H(v_k) = I - 2\frac{v_k v_k^t}{v_k^t v_k}$ donde

$$v_k = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ a_{k+1,k}^{(k)} \pm (\Sigma_{i=k+1}^{d} |a_{ik}^{(k)}|^2)^{1/2} \\ a_{k+2,k}^{(k)} \\ \vdots \\ a_{d,k}^{(k)} \end{pmatrix}$$

eligiendo el signo + o − el mismo que $a_{k+1,k}^{(k)}$.

Una vez obtenido $v_k$ los elementos $a_{ij}^{(k+1)}$ para $k \leq i, j \leq d$ de la matriz $A_{k+1} = (a_{ij}^{(k+1)})$ se obtienen de la forma siguiente:

(a) $a_{k+1,k}^{(k+1)} = \mp \Sigma_{i=k+1}^{d} |a_{ik}^{(k)}|^2)^{1/2}$

(b) $a_{i,k}^{(k+1)} = a_{k,i}^{(k+1)} = 0 \quad \text{para} \quad k+1 < i \leq d$

(c) Para $k+1 < i, j \leq d$ poniendo $\tilde{A}_{k+1,k+1} = \tilde{H}_k \tilde{A}_{k,k} \tilde{H}_k$ donde

$$\tilde{H}_k = \tilde{I} - \beta \tilde{v} \tilde{v}^t \quad \text{con} \quad \beta = \frac{2}{v^t v}$$

y el símbolo $\backsim$ indica matrices y vectores reducidos de orden $d-k$, tendremos

$$\tilde{A}_{k+1,k+1} = (\tilde{I} - \beta \tilde{v} \tilde{v}^t) \tilde{A}_{k,k} (\tilde{I} - \beta \tilde{v} \tilde{v}^t) =$$
$$\tilde{A}_{k,k} - \beta \tilde{v} \tilde{v}^t \tilde{A}_{k,k} - \beta \tilde{A}_{k,k} \tilde{v} \tilde{v}^t + \beta^2 \tilde{v} (\tilde{v}^t \tilde{A}_{k,k} \tilde{v}) \tilde{v}^t$$

Llamando

$$z = \beta \tilde{A}_{k,k} \tilde{v}$$

$$z^t = \beta \tilde{v}^t \tilde{A}_{k,k}$$

tendremos

$$\tilde{A}_{k+1,k+1} = \tilde{A}_{k,k} - \tilde{v}z^t - z\tilde{v}^t + \beta(z^t\tilde{v})\tilde{v}\tilde{v}^t =$$

$$\tilde{A}_{k,k} - \tilde{v}z^t + \frac{\beta(z^t\tilde{v})\tilde{v}\tilde{v}^t}{2} - z\tilde{v}^t + \frac{\beta(z^t\tilde{v})\tilde{v}\tilde{v}^t}{2} =$$

$$\tilde{A}_{k,k} - \tilde{v}(z^t - \frac{\beta(z^t\tilde{v})}{2}\tilde{v}) - (z - \frac{\beta(z^t\tilde{v})}{2}\tilde{v})\tilde{v}^t =$$

$$\tilde{A}_{k,k} - \tilde{v}q^t - q\tilde{v}^t$$

donde

$$q = z - \frac{\beta(z^t\tilde{v})}{2}\tilde{v}$$

Es decir finalmente,

$$a_{ij}^{(k+1)} = a_{ij}^{(k)} - v_i q_j - v_j q_i \quad \text{para} \quad k+1 \leq i,j \leq d$$