E.T.S. DE INGENIERÍA INFORMÁTICA

Apuntes de

ÁLGEBRA NUMÉRICA

para la titulación de

INGENIERÍA INFORMÁTICA

Curso 2003-2004

por

Fco. Javier Cobos Gavala

DEPARTAMENTO DE

MATEMÁTICA APLICADA I

Contenido

1. Ecuaciones no lineales

Dada una función no nula $f : \mathbf{C} \to \mathbf{C}$, resolver la ecuación $f(x) = 0$ es hallar los valores $\overline{x}$ que anulan a dicha función. A estos valores $\overline{x}$ se les denomina *raíces* o *soluciones* de la ecuación, o también, *ceros* de la función $f(x)$.

Los métodos de resolución de ecuaciones y sistemas de ecuaciones se clasifican en *directos* e *iterados* . Los del primer grupo nos proporcionan la solución mediante un número finito de operaciones elementales, mientras que los iterados producen una sucesión convergente a la solución del problema.

Un ejemplo de método directo es la conocida fórmula de resolución de las ecuaciones de segundo grado $ax^2 + bx + c = 0$, cuyas soluciones vienen dadas por la fórmula

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

Sin embargo, el siglo pasado *Abel* probó que no existe ninguna fórmula equivalente (en término de raíces) para resolver ecuaciones de grado superior a cuatro. Además, si la ecuación no es polinómica no podemos resolverla más que mediante métodos iterados que, incluso en el caso de las polinómicas de grado no superior a cuatro, son más eficientes.

Definición 1.1 Una solución $\overline{x}$ de la ecuación $f(x) = 0$ se dice que tiene *multiplicidad* n si

$$f(\overline{x}) = f'(\overline{x}) = f''(\overline{x}) = \cdots = f^{(n-1}(\overline{x}) = 0 \qquad \text{y} \qquad f^{(n}(\overline{x}) \neq 0$$

Si la multiplicidad de una raíz es 1, diremos que es *simple* .

Todos los métodos numéricos de resolución de ecuaciones presentan dificultades cuando la ecuación tiene raíces múltiples, ya que todos ellos se basan en los cambios de signo de la función y éstos son difícilmente detectables en un entorno de una raíz múltiple.

Ese hecho produce que en estos casos el problema esté mal condicionado.

1.1 Errores y condicionamiento en problemas numéricos

Cualquier problema numérico se resuelve a través de un algoritmo que nos proporciona unos resultados a partir de unos datos iniciales. Es decir, se trata de realizar un proceso del tipo

$$\text{Datos} \Longrightarrow \boxed{\text{Algoritmo}} \Longrightarrow \text{Resultados}$$

Dado que cualquier algoritmo puede cometer errores, no sólo por el algoritmo en sí, sino porque los datos pueden venir afectados de algún tipo de error (redondeo, etc.) es muy importante el estudio de los distintos tipos de error que puedan cometerse con vista a la fiabilidad de los resultados.

Se denomina *error absoluto* de un número x que aproxima a otro $\overline{x}$ a la *distancia* entre ellos. Así, por ejemplo, en el caso real vendrá dado por $|\overline{x}-x|$, es decir, por el valor absoluto de la diferencia entre ambos.

Obsérvese que si sólo disponemos del dato de que el error es, por ejemplo, de 1m. no sabemos nada acerca de la fiabilidad del resultado, ya que no es lo mismo decir que se ha cometido un error de un metro al medir la altura de una persona que al medir la distancia entre dos galaxias.

Debemos reflejar de alguna manera "lo que se está evaluando" en el dato del error. Para ello se utiliza el denominado *error relativo* que es el cociente entre el error absoluto y el objeto evaluado, es decir, en el caso real

$$\left|\frac{\overline{x}-x}{\overline{x}}\right|.$$

En el caso $x=0$ sólo se utiliza el error absoluto. En la mayoría de los procesos numéricos utilizaremos como error el *error absoluto* ya que lo que nos interesa conocer es el número de cifras decimales exactas que posee.

Evidentemente cualquier algoritmo que trabaje con unos datos afectados de algún tipo de error nos proporcionará unos resultados que también vendrán afectados de errores. Estos errores pueden depender sólo de los datos iniciales o también del proceso que se ha realizado.

Supongamos que, queremos evaluar $f(\overline{x})$ y damos un dato aproximado x. Es evidente que, en general, si $x \neq \overline{x}$ será $f(x) \neq f(\overline{x})$.

Dado que $f(\overline{x}) - f(x) \simeq (\overline{x}-x)f'(\overline{x})$, se tiene que

$$|f(\overline{x}) - f(x)| \simeq |\overline{x}-x| \cdot |f'(\overline{x})|$$

por lo que aunque el error del dato sea muy pequeño, si la derivada $f'(\overline{x})$ es muy grande, el resultado obtenido $f(x)$ puede diferir mucho del valor exacto $f(\overline{x})$. Además el problema no está en el algoritmo que se aplique para evaluar $f(x)$ sino en el propio problema a resolver.

Diremos que un problema está *mal condicionado* si pequeños errores en los datos producen grandes errores en los resultados.

Se trata entonces de definir algún número que nos indique el condicionamiento del problema. A éste número lo llamaremos *número de condición* del problema y lo denotaremos por κ. En el ejemplo anterior es evidente que

$$\kappa(x) = |f'(\overline{x})|$$

Para el problema inverso, es decir conocido $f(\overline{x})$ buscar el valor de $\overline{x}$ (resolver una ecuación) se tendrá que

$$|\overline{x} - x| \simeq \frac{1}{|f'(\overline{x})|}|f(\overline{x}) - f(x)|$$

por lo que si $|f'(\overline{x})|$ fuese muy pequeño el problema estaría mal condicionado.

Así pues, un problema (en ambos sentidos) estará mejor condicionado mientras más se acerque a 1 su número de condición.

Respecto al algoritmo que se utiliza para resolver un determinado problema, diremos que es *inestable* cuando los errores que se cometen en los diferentes pasos del algoritmo hacen que el error total que se genera sea muy grande. Si, por el contrario los errores que se producen en los distintos pasos no alteran de forma significativa el resultado del problema, diremos que el algoritmo es *estable*.

Obsérvese que si el algoritmo es inestable no va a generar un resultado fiable, por lo que deberemos utilizar otro algoritmo. Sin embargo, por muy estable que sea el algoritmo, si el problema está mal condicionado lo único que podemos hacer es plantear un problema equivalente al nuestro pero con la seguridad de que se trata de un problema bien condicionado.

Así, por ejemplo, si queremos calcular la tangente de $89°59'$ lo primero que debemos hacer es expresar el ángulo en radianes, por lo que necesariamente debemos redondear el dato (el número π hay que redondearlo), por lo que el dato vendrá afectado de un error de redondeo. Utilicemos el método que utilicemos, dado que $|\operatorname{tg}\ \overline{x} - \operatorname{tg}\ x| \simeq |1 + \operatorname{tg}^2 x| \cdot |\overline{x} - x|$ y $\operatorname{tg}\ \pi = +\infty$ el problema estará mal condicionado. Sin embargo si tenemos en cuenta que $\operatorname{tg}\ (a+b) = \dfrac{\operatorname{tg}\ a + \operatorname{tg}\ b}{1 - \operatorname{tg}\ a \operatorname{tg}\ b}$ podemos reducir nuestro problema al cálculo de

la tangente de $44°59'$ resultando éste un proceso bien condicionado, ya que tg $45 = 1$.

1.2 Método y algoritmo de la bisección: análisis de errores

Este método consiste en la aplicación directa del teorema de Bolzano.

Teorema 1.1 [Teorema de Bolzano] *Si f es una función continua en el intervalo cerrado $[a, b]$ y $f(a) \cdot f(b) < 0$, existe un punto $\alpha \in (a, b)$ en el cual $f(\alpha) = 0$.*

Nuestro problema se reduce a localizarla. Para ello, supongamos que está separada, es decir, que en el intervalo $[a, b]$ es la única raíz que existe. Esto podemos garantizarlo, por ejemplo, viendo que $f'(x) \neq 0$ en todo el intervalo, ya que entonces, el Teorema de Rolle (que se enuncia a continuación) nos garantiza la unicidad de la raíz.

Teorema 1.2 [Teorema de Rolle] *Si $f(x)$ es una función continua en el intervalo $[a, b]$, derivable en (a, b) y $f(a) = f(b)$, existe un punto $\alpha \in (a, b)$ para el que $f'(\alpha) = 0$.*

En efecto, si $f(x)$ tuviese dos raíces α_1 y α_2 en el intervalo $[a, b]$, verificaría las hipótesis del teorema de Rolle en el intervalo $[\alpha_1, \alpha_2] \subset [a, b]$, por lo que debería existir un punto $\alpha \in (\alpha_1, \alpha_2) \Longrightarrow \alpha \in (a, b)$ en el que se anulara la derivada, por lo que si $f'(x) \neq 0$ en todo el intervalo $[a, b]$, no pueden existir dos raíces de la ecuación en dicho intervalo.

Supongamos, sin pérdida de generalidad, que f es creciente en $[a, b]$.

a) Tomamos $\alpha = \dfrac{a+b}{2}$ y $\varepsilon = \dfrac{b-a}{2}$.

b) Si $f(\alpha) = 0$ entonces FIN. α es la raíz exacta.

Si $f(\alpha) > 0$ entonces hacemos $b = \alpha$.

Si $f(\alpha) < 0$ entonces hacemos $a = \alpha$.

Se repite el paso 1, es decir, hacemos $\alpha = \dfrac{a+b}{2}$ y $\varepsilon = \dfrac{b-a}{2}$.

c) Si $\varepsilon < 10^{-k}$ (error prefijado), entonces FIN. El valor de α es la raíz buscada con k cifras decimales exactas.

Si $\varepsilon > 10^{-k}$, entonces repetimos el paso 2.

El error cometido, tomando como raíz de la ecuación el punto medio del intervalo obtenido en la en la iteración n-ésima, viene dado por $\varepsilon_n = \dfrac{b-a}{2^n}$, por lo que si $b-a=1$ y $n=10$ se tiene que $\varepsilon_{10} < \dfrac{1}{2^{10}} < 10^{-3}$, es decir, en 10 iteraciones obtenemos tres cifras decimales exactas.

1.2.1 Algoritmo

Para $i = 1, 2, \ldots, n, \ldots$, $I_i = [a_i, b_i]$ y $m_i = \dfrac{a_i + b_i}{2}$ (punto medio del intervalo I_i) con

$$I_1 = [a, b] \quad \text{y} \quad I_{i+1} = \begin{cases} [a_i, m_i] & \text{si } \operatorname{sig}(f(a_i)) \neq \operatorname{sig}(f(m_i)) \\ [m_i, b_i] & \text{si } \operatorname{sig}(f(b_i)) \neq \operatorname{sig}(f(m_i)) \end{cases}$$

El proceso debe repetirse hasta que

$$f(m_i) = 0 \quad \text{o bien} \quad b_i - a_i < \varepsilon \ \text{ con } \ \varepsilon > 0 \ \text{ prefijado.}$$

Se tiene, por tanto:

Input: $a, b, \varepsilon, f(x)$
Output: m

```
while (b − a)/2 > ε
    m ← a + (b − a)/2
    if f(m) = 0
        a ← m
        b ← m
    end if
    if signf(a) = signf(m)
        a ← m
    end if
    if signf(b) = signf(m)
        b ← m
    end if
end
print m
```

El hecho de calcular el punto medio de $[a, b]$ como $m = a + {}^{(b-a)}/_{2}$ es debido a que para valores muy pequeños de a y b puede darse el caso de que ${}^{(a+b)}/_{2}$ se encuentre fuera del intervalo $[a, b]$.

Ejemplo 1.1 Supongamos que se quiere calcular la raíz cuadrada de 3, para lo que vamos a buscar la raíz positiva de la ecuación $f(x) = 0$ con $f(x) = x^2 - 3$.

Dado que $f(1) = -2 < 0$ y $f(2) = 1 > 0$, el teorema de Bolzano nos garantiza la existencia de una raíz (que además sabemos que es única ya que $f'(x) = 2x$ no se anula en el intervalo $[1, 2]$).

Para obtener la raíz con 14 cifras decimales exactas, es decir, con un error menor que 10^{-14} tendríamos que detener el proceso cuando

$$\frac{2-1}{2^n} < 10^{-14} \Longrightarrow 2^n > 10^{14} \Longrightarrow n \geq 47$$

es decir, tendríamos que detenernos en m_{47} para poder garantizar la precisión exigida. □

Una variante del método de la bisección es el método de la *regula falsi* o de la *falsa posición* consistente en dividir el intervalo $[a, b]$ en dos subintervalos $[a, c] \cup [c, b]$ donde el punto c, a diferencia del punto medio m del método de la bisección, es el punto de corte de la recta secante que pasa por los puntos $(a, f(a))$ y $(b, f(b))$ con el eje de abscisas OX.

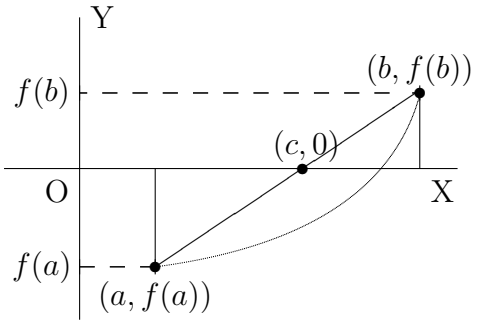


Figura 1.1: Método de la *régula falsi*.

La pendiente de dicha secante viene determinada por

$$m = \frac{f(b) - f(a)}{b - a} = \frac{0 - f(b)}{c - b}$$

Según se utilicen los puntos $(a, f(a))$ y $(b, f(b))$ ó $(c, 0)$ y $(b, f(b))$ respectivamente.

Despejando el valor de c obtenemos que

$$c = b - f(b) \cdot \frac{b-a}{f(b)-f(a)}$$

pudiéndose dar los mismos casos que en el método de la bisección, es decir:

- Si $f(c) = 0$ la raíz buscada es c.
- Si $f(a)$ y $f(c)$ tienen signos contrarios, la raíz se encuentra en el intervalo $[a, c]$.
- Si son $f(c)$ y $f(b)$ los que tienen signos contrarios, la raíz está en el intervalo $[c, b]$.

El algoritmo quedaría de la siguiente forma:

Input: $a, b, \varepsilon, f(x)$
Output: c

$c \leftarrow a$
while abs$f(c) > \varepsilon$
  $c \leftarrow b - f(b) \cdot \dfrac{b-a}{f(b)-f(a)}$
  if $f(c) = 0$
    $a \leftarrow c$
    $b \leftarrow c$
  end if
  if sign$f(a) =$ sign$f(c)$
    $a \leftarrow c$
  end if
  if sign$f(b) =$ sign$f(c)$
    $b \leftarrow c$
  end if
end
print c

Aplicando el algoritmo al Ejemplo 1.1 obtenemos la raíz de 3, con 14 cifras decimales exactas, en sólo 14 iteraciones frente a las 47 necesarias mediante el método de la bisección.

En el método de la *regula falsi* la convergencia suele ser más rápida que en el de la *bisección* aunque ambos métodos son de convergencia lenta siendo necesario, por tanto, buscar otros métodos cuya convergencia sea más rápida.

Estos métodos están basados en el denominado *Teorema del punto fijo* que estudiamos en la siguiente sección.

1.3 Punto fijo e iteración funcional

Ya se comentó que los métodos iterados consisten en crear una sucesión convergente a la solución del problema.

Una función $f : \mathbf{R} \to \mathbf{R}$ se dice *contractiva* si verifica que

$$|f(x_1) - f(x_2)| < |x_1 - x_2| \quad \forall\, x_1, x_2 \in \mathbf{R}$$

Si la función es derivable, basta comprobar que $|f'(x)| \leq q < 1$ cualquiera que sea el valor de $x \in \mathbf{R}$ para poder garantizar que se trata de una función contractiva.

Si se desea resolver la ecuación $f(x) = 0$, se escribe esta de la forma $x = \varphi(x)$, donde $\varphi(x)$ es una función *contractiva*, y partiendo de un determinado valor inicial x_0, se construye la sucesión $x_{n+1} = \varphi(x_n)$. La convergencia de esta sucesión la garantiza el siguiente teorema.

Teorema 1.3 [Teorema del punto fijo] *Dada la ecuación $x = \varphi(x)$ en la que $|\varphi'(x)| \leq q < 1$ cualquiera que sea $x \in [a, b]$ con $\varphi([a, b]) \subseteq [a, b]$ y un punto $x_0 \in [a, b]$, la sucesión $x_0, x_1, \ldots, x_n, \ldots$ en la que $x_{n+1} = \varphi(x_n)$ converge a un valor $\overline{x}$ que es la única solución de la ecuación en dicho intervalo.*

Demostración. Dado que $x_{n+1} = \varphi(x_n)$ y $x_n = \varphi(x_{n-1})$ se tiene que

$$x_{n+1} - x_n = \varphi(x_n) - \varphi(x_{n-1}) = (x_n - x_{n-1})\varphi'(c)$$

con $c \in (x_{n-1}, x_n)$ (x_n, x_{n-1}) y, por tanto, en $[a, b]$, por lo que

$$|x_{n+1} - x_n| \leq q\,|x_n - x_{n-1}|$$

es decir

$$\begin{aligned}
&|x_2 - x_1| \leq q\,|x_1 - x_0| \\
&|x_3 - x_2| \leq q\,|x_2 - x_1| \leq q^2\,|x_1 - x_0| \\
&\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\
&|x_{n+1} - x_n| \leq q^n\,|x_1 - x_0|
\end{aligned}$$

Si construimos la serie

$$x_0 + (x_1 - x_0) + (x_2 - x_1) + \cdots + (x_{n+1} - x_n) + \cdots$$

observamos que la suma parcial $S_{n+1} = x_n$ y que la serie es absolutamente convergente por estar acotados sus términos, en valor absoluto, por los términos de una serie geométrica de razón $q < 1$.

Sea $\overline{x}$ la suma de la serie. Entonces $\overline{x} = \lim S_{n+1} = \lim x_n$, es decir, la sucesión $x_0, x_1, \ldots, x_n, \ldots$ converge a $\overline{x}$.

El valor obtenido $\overline{x}$ es solución de la ecuación, ya que por la continuidad de la función $\varphi(x)$ se verifica que

$$\overline{x} = \lim x_n = \lim \varphi(x_{n-1}) = \varphi(\lim x_{n-1}) = \varphi(\overline{x})$$

Además, es la única solución de la ecuación en el intervalo $[a, b]$, ya que de existir otra $\overline{x}'$ se tendría que

$$\overline{x} - \overline{x}' = \varphi(\overline{x}) - \varphi(\overline{x}') = (\overline{x} - \overline{x}')\varphi'(c) \quad \text{con} \quad c \in (\overline{x}, \overline{x}')$$

por lo que

$$(\overline{x} - \overline{x}')(1 - \varphi'(c)) = 0$$

y dado que el segundo paréntesis es no nulo, se tiene que $\overline{x} = \overline{x}'$. ■

En la Figura 1.2 puede observarse que el método converge si $|\varphi'(x)| \leq q < 1$, mientras que si $|\varphi'(x)| > 1$ el método es divergente.

En los casos (a) y (b), en los que $|\varphi'(x)| \leq q < 1$ el método converge monótonamente en (a) y de forma oscilatoria o en espiral en (b).

En los casos (c) y (d), en los que $|\varphi'(x)| > 1$ el método diverge de forma monótona en (a) y de forma oscilatoria en (b).

1.3.1 Cota del error "a posteriori"

Si $f(x)$ es una función continua en el intervalo $[a, b]$ y derivable en el abierto (a, b), sabemos por el Teorema del Valor Medio que existe un punto $c \in (a, b)$ tal que $\dfrac{f(b) - f(a)}{b - a} = f'(c)$.

Sea $\overline{x}$ una solución de la ecuación $f(x) = 0$ y sea x_n una aproximación de ella obtenida por un método iterado cualquiera. Supongamos $f(x)$ continua en el

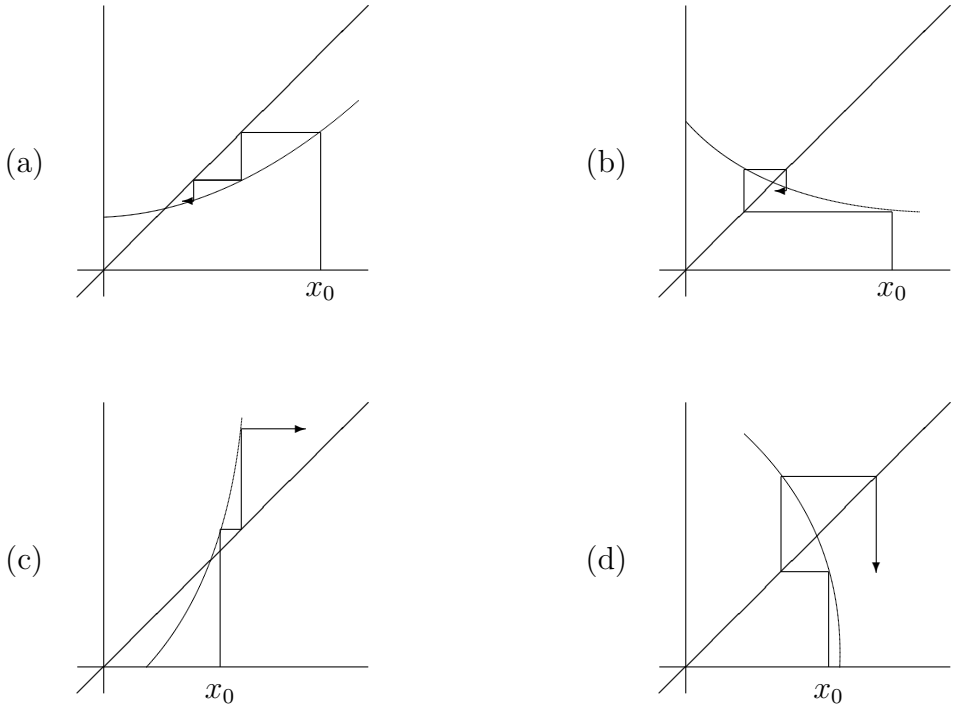


Figura 1.2: Esquema de la convergencia para el teorema del punto fijo.

intervalo cerrado $[x_n, \overline{x}]$ ó $[\overline{x}, x_n]$ (dependiendo de que $\overline{x}$ sea mayor o menor que x_n) y derivable en el abierto.

Existe entonces un punto $c \in (x_n, \overline{x})$ ó $c \in (\overline{x}, x_n)$ tal que

$$\frac{f(\overline{x}) - f(x_n)}{\overline{x} - x_n} = f'(c).$$

Como $f(\overline{x}) = 0$ y $(\overline{x} - x_n) = \varepsilon_n$, nos queda que $\varepsilon_n = -\dfrac{f(x_n)}{f'(c)}$, obteniéndose:

$$|\varepsilon_n| = \frac{|f(x_n)|}{|f'(c)|} \leq \frac{|f(x_n)|}{\min\limits_{x \in \left\{ \begin{smallmatrix} (\overline{x}, x_n) \\ (x_n, \overline{x}) \end{smallmatrix} \right.} |f'(x)|} \leq \frac{|f(x_n)|}{\min\limits_{x \in (a,b)} |f'(x)|} \quad \text{con} \quad \left. \begin{matrix} (\overline{x}, x_n) \\ (x_n, \overline{x}) \end{matrix} \right\} \in (a, b)$$

Lo único que debemos exigir es que la derivada de la función no se anule en ningún punto del intervalo (a, b).

Ejemplo 1.2 El cálculo de la raíz cuadrada de 3 equivale al cálculo de la raíz positiva de la ecuación $x^2 = 3$. Aunque más adelante veremos métodos cuya convergencia es más rápida, vamos a realizar los siguientes cambios:

$$x^2 = 3 \Longrightarrow x + x^2 = x + 3 \Longrightarrow x(1 + x) = 3 + x \Longrightarrow x = \frac{3 + x}{1 + x}$$

Es decir, hemos escrito la ecuación de la forma $x = \varphi(x)$ con

$$\varphi(x) = \frac{3+x}{1+x}$$

Dado que sabemos que la raíz de 3 está comprendida entre 1 y 2 y que

$$|\varphi'(x)| = \frac{2}{(1+x)^2} \leq \frac{2}{2^2} = \frac{1}{2} < 1 \quad \text{para cualquier} \quad x \in [1,2]$$

podemos garantizar que partiendo de $x_0 = 1$ el método convergerá a la raíz cuadrada de 3.

Así pues, partiendo de $x_0 = 1$ y haciendo $x_{n+1} = \dfrac{3+x_n}{1+x_n}$ obtenemos:

x_1	=	2	x_{14}	=	1.73205079844084
x_2	=	1.66666666666667	x_{15}	=	1.73205081001473
x_3	=	1.75000000000000	x_{16}	=	1.73205080691351
x_4	=	1.72727272727273	x_{17}	=	1.73205080774448
x_5	=	1.73333333333333	x_{18}	=	1.73205080752182
x_6	=	1.73170731707317	x_{19}	=	1.73205080758148
x_7	=	1.73214285714286	x_{20}	=	1.73205080756550
x_8	=	1.73202614379085	x_{21}	=	1.73205080756978
x_9	=	1.73205741626794	x_{22}	=	1.73205080756863
x_{10}	=	1.73204903677758	x_{23}	=	1.73205080756894
x_{11}	=	1.73205128205128	x_{24}	=	1.73205080756886
x_{12}	=	1.73205068043172	x_{25}	=	1.73205080756888
x_{13}	=	1.73205084163518	x_{26}	=	1.73205080756888

El error vendrá dado por $\varepsilon_n < \dfrac{|f(x_n)|}{\min\limits_{x\in[1,2]} |f'(x)|}$ donde $f(x) = x^2 - 3$, por lo que

$$\varepsilon_{26} < \frac{|x_{26}^2 - 3|}{2} = 4.884981308350688 \cdot 10^{-15} < 10^{-14}$$

es decir, $\sqrt{3} = 1.73205080756888$ con todas sus cifras decimales exactas. □

Obsérvese que en el Ejemplo 1.1 vimos cómo eran necesarias 47 iteraciones para calcular la raíz cuadrada de 3 (con 14 cifras decimales exactas) mediante el método de la bisección y 14 mediante el de la regula falsi, mientras que ahora hemos necesitado 26. El hecho de que la convergencia de éste último método sea más lenta que cuando se utiliza el método de la regula falsi estriba en la mala elección de la función $\varphi(x)$, por lo que vamos a ver cómo el *método de Newton-Raphson*, generalmente conocido como *método de Newton*, nos permite escribir la ecuación $f(x) = 0$ de la forma $x = \varphi(x)$ de forma que la convergencia sea muy rápida.

1.4 Análisis del método de Newton-Raphson

Si tratamos de resolver la ecuación $f(x) = 0$ y lo que obtenemos no es la solución exacta $\overline{x}$ sino sólo una buena aproximación x_n tal que $\overline{x} = x_n + h$ tendremos que

$$f(\overline{x}) \simeq f(x_n) + h \cdot f'(x_n) \Rightarrow h \simeq -\frac{f(x_n)}{f'(x_n)}$$

por lo que

$$\overline{x} \simeq x_n - \frac{f(x_n)}{f'(x_n)}$$

obteniéndose la denominada fórmula de *Newton-Raphson*

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} \tag{1.1}$$

Si construimos, utilizando la fórmula de Newton-Raphson, la sucesión $\{x_n\}$ y ésta converge, se tendrá que $\lim x_n = \overline{x}$, ya que nos quedaría, aplicando límites en (1.1)

$$\lim x_{n+1} = \lim x_n - \frac{f(\lim x_n)}{f'(\lim x_n)} \Rightarrow f(\lim x_n) = 0$$

siempre que $f'(\lim x_n) \neq 0$, lo cual se verifica si exigimos que la función posea una única raíz en $[a, b]$. Dado que la raíz de la ecuación en el intervalo $[a, b]$ es única, necesariamente $\lim x_n = \overline{x}$.

Este método es también conocido como *método de la tangente*, ya que si trazamos la tangente a la curva $y = f(x)$ en el punto $(x_n, f(x_n))$ obtenemos la recta $y = f(x_n) + f'(x_n)(x - x_n)$, que corta al eje $y = 0$ en el punto de abscisa $x = x_n - \dfrac{f(x_n)}{f'(x_n)}$ que es precisamente el valor de x_{n+1} de la fórmula de Newton-Raphson.

En la Figura1.3 puede observarse cómo actúa geométricamente el método de Newton-Raphson.

Lo más dificultoso del método consiste en el cálculo de la derivada de la función así como la obtención del valor inicial x_0.

Busquemos, a continuación, alguna cota del error.

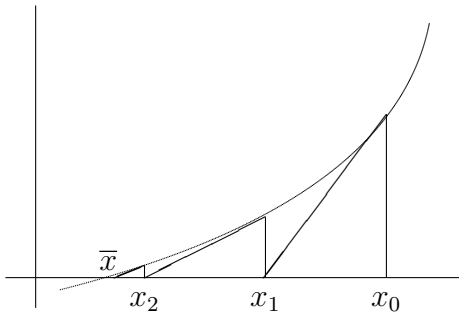


Figura 1.3: Interpretación geométrica del método de Newton.

$$\varepsilon_{n+1} = \overline{x} - x_{n+1} = \overline{x} - \left(x_n - \frac{f(x_n)}{f'(x_n)} \right) = (\overline{x} - x_n) + \frac{f(x_n)}{f'(x_n)} = \varepsilon_n + \frac{f(x_n)}{f'(x_n)}$$

Desarrollando $f(\overline{x})$ en un entorno de x_n se obtiene

$$0 = f(\overline{x}) = f(x_n + \varepsilon_n) = f(x_n) + f'(x_n)\varepsilon_n + \frac{f''(t)}{2!}\varepsilon_n^2 \text{ con } t \in \begin{cases} (\overline{x}, x_n) & \text{si } \overline{x} < x_n \\ (x_n, \overline{x}) & \text{si } \overline{x} > x_n \end{cases}$$

Supuesto que $f'(x_n) \neq 0$ podemos dividir por dicha derivada para obtener

$$0 = \frac{f(x_n)}{f'(x_n)} + \varepsilon_n + \frac{f''(t)}{2f'(x_n)}\varepsilon_n^2 = \varepsilon_{n+1} + \frac{f''(t)}{2f'(x_n)}\varepsilon_n^2$$

por lo que

$$|\varepsilon_{n+1}| = \frac{|f''(t)|}{2\,|f'(x_n)|}\varepsilon_n^2 \leq k \cdot \varepsilon_n^2 \tag{1.2}$$

donde $k \geq \max\limits_{x \in [a,b]} \dfrac{|f''(x)|}{2|f'(x)|}$ siendo $[a, b]$ cualquier intervalo, *en caso de existir*, que contenga a la solución $\overline{x}$ y a todas las aproximaciones x_n.

El método de Newton es, por tanto, un método de segundo orden es decir, el error de una determinada iteración es del orden del cuadrado del de la iteración anterior. En otras palabras, si en una determinada iteración tenemos n cifras decimales exactas, en la siguiente tendremos del orden de $2n$ cifras decimales exactas.

Esta última desigualdad podemos (no queriendo precisar tanto) modificarla para escribir

$$k \geq \frac{\max |f''(x)|}{2 \min |f'(x)|} \quad \text{con} \quad x \in [a, b] \;\text{ y }\; f'(x) \neq 0$$

Supuesta la existencia de dicho intervalo $[a, b]$, el valor de k es independiente de la iteración que se realiza, por lo que

$$k \cdot |\varepsilon_{n+1}| \leq |k \cdot \varepsilon_n|^2 \leq |k \cdot \varepsilon_{n-1}|^4 \leq \cdots \leq |k \cdot \varepsilon_0|^{2^{n+1}}$$

o lo que es lo mismo:

$$|\varepsilon_n| \leq \frac{1}{k} \cdot |k \cdot \varepsilon_0|^{2^n}$$

donde es necesario saber acotar el valor de $\varepsilon_0 = \overline{x} - x_0$.

Es decir, si existe un intervalo $[a, b]$ que contenga a la solución y a todas las aproximaciones x_n se puede determinar *a priori* una cota del error, o lo que es lo mismo, se puede determinar el número de iteraciones necesarias para obtener la solución con un determinado error.

Evidentemente, el proceso convergerá si $|k \cdot \varepsilon_0| < 1$, es decir, si $|\varepsilon_0| < \frac{1}{k}$. En caso de ser convergente, la convergencia es de segundo orden como puede verse en la ecuación (1.2).

1.4.1 Algoritmo

Una vez realizado un estudio previo para ver que se cumplen las condiciones que requiere el método, establecer el valor inicial x_0 y calcular el valor de $m = \min_{x \in [a,b]} |f'(x)|$, el algoritmo es el siguiente

Input: $a, b, x_0, \varepsilon, f(x), m$

Output: x

$x \leftarrow x_0$

$e \leftarrow \text{abs}(\text{f(x)}/\text{m})$

while $e > \varepsilon$

  $x \leftarrow x - \dfrac{f(x)}{f'(x)}$

  $e \leftarrow \text{abs}(\text{f(x)}/\text{m})$

end

print x

Ejemplo 1.3 En el Ejemplo 1.2 calculamos la raíz de 3 con 14 cifras decimales exactas en 26 iteraciones. Vamos a ver cómo se disminuye considerablemente el número de iteraciones cuando se utiliza la fórmula de Newton-Raphson.

Partimos de la ecuación $f(x) = x^2 - 3 = 0$, por lo que la fórmula de Newton-Raphson nos dice que

$$x_{n+1} = \frac{1}{2}\left(x_n + \frac{3}{x_n}\right)$$

Dado que la raíz de 3 es un número comprendido entre 1 y 2 y la función $f'(x) = 2x$ no se anula en dicho intervalo, podemos aplicar el método de Newton tomando como valor inicial $x_0 = 2$. (Más adelante veremos porqué debemos tomar 2 como valor inicial), obteniéndose:

$$\begin{aligned} x_0 &= 2 \\ x_1 &= 1.75000000000000 \\ x_2 &= 1.73214285714286 \\ x_3 &= 1.73205081001473 \\ x_4 &= 1.73205080756888 \end{aligned}$$

El error vendrá dado, al igual que en el Ejercicio 1.2 por $\varepsilon_n < \dfrac{|f(x_n)|}{\min\limits_{x\in[1,2]} |f'(x)|}$, por lo que

$$\varepsilon_4 < \frac{|x_4^2 - 3|}{2} = 4.884981308350688 \cdot 10^{-15} < 10^{-14}$$

es decir, $\sqrt{3} = 1.73205080756888$ con todas sus cifras decimales exactas. □

Nota: La fórmula $x_{n+1} = \dfrac{1}{2}\left(x_n + \dfrac{A}{x_n}\right)$ es conocida como fórmula de *Heron* ya que este matemático la utilizaba para aproximar la raíz cuadrada de un número real positivo A hacia el año 100 a.C. pues sabía que si tomaba un valor inicial x_0 y éste fuese mayor que la raíz buscada, necesariamente A/x_0 sería menor que su raíz y viceversa, por lo que la media entre ambos debía ser una mejor aproximación que x_0. En otras palabras, en cada iteración tomaba como aproximación de la raíz la media aritmética entre el valor x_n anterior y el cociente A/x_n.

Puede observarse cómo la convergencia del método de Newton-Raphson es mucho más rápida que la del método de la bisección y que el de la regula falsi, ya que sólo hemos necesitado 5 iteraciones frente a las 47 que se necesitan en el método de la bisección o las 14 del método de la regula falsi.

1.4.2 Análisis de la convergencia: Regla de Fourier

Hay que tener en cuenta que la naturaleza de la función puede originar dificultades, llegando incluso a hacer que el método no converja.

Ejemplo 1.4 Tratemos de determinar, por el método de Newton-Raphson, la raíz positiva de la función $f(x) = x^{10} - 1$, tomando como valor inicial $x_0 = 0.5$. La fórmula de Newton-Raphson es, en este caso,

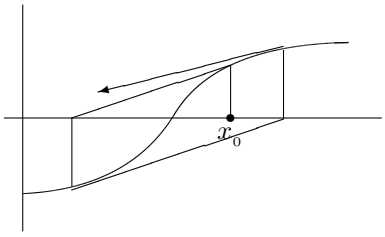
$$x_{n+1} = x_n - \frac{x_n^{10} - 1}{10x_n^9}.$$

Aplicando el algoritmo se obtienen los valores

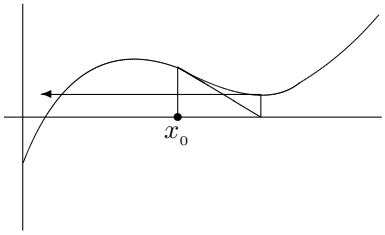
$x_1 = 51.65$ $\qquad$ $x_{20} = 6.97714912329906$
$x_2 = 46.485$ $\qquad$ $x_{30} = 2.43280139954230$
$x_3 = 41.8365$ $\qquad$ $x_{40} = 1.00231602417741$
$x_4 = 37.65285$ $\qquad$ $x_{41} = 1.00002393429084$
$x_5 = 33.887565$ $\qquad$ $x_{42} = 1.00000000257760$
$x_{10} = 20.01026825685012$ $\qquad$ $x_{43} = 1$

Puede observarse que la convergencia es muy lenta y sólo se acelera (a partir de x_{40}) cuando estamos muy cerca de la raíz buscada. □

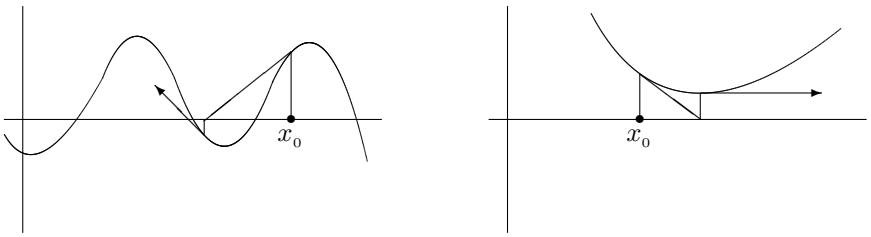
a) Si en las proximidades de la raíz existe un punto de inflexión, las iteraciones divergen progresivamente de la raíz.



b) El método de Newton-Raphson oscila en los alrededores de un máximo o un mínimo local, persistiendo o llegando a encontrarse con pendientes cercanas a cero, en cuyo caso la solución se aleja del área de interés.



c) Un valor inicial cercano a una raíz puede converger a otra raíz muy distante de la anterior como consecuencia de encontrarse pendientes cercanas a cero. Una pendiente nula provoca una división por cero (geométricamente, una tangente horizontal que jamás corta al eje de abscisas).



Estos problemas pueden detectarse previamente a la aplicación del método.

Supongamos que tenemos acotada, en el intervalo $[a, b]$, una única raíz $\overline{x}$ de la ecuación $f(x) = 0$ y que $f'(x)$ y $f''(x)$ no se anulan en ningún punto del intervalo $[a, b]$, es decir, que ambas derivadas tienen signo constante en dicho intervalo.

Obsérvese que si $f(a)f(b) < 0$, dado que $f'(x)$ no se anula en el intervalo (a, b) sabemos, por los teoremas de Bolzano y Rolle, que existe una única raíz en dicho intervalo. Además, por las condiciones exigidas sabemos que no existe, en (a, b) ningún punto crítico (ni extremo relativo ni punto de inflexión), con lo que habremos evitado los problemas expuestos anteriormente.

En cualquiera de los cuatro casos posibles (véase la Figura 1.4), la función cambia de signo en los extremos del intervalo (debido a que la primera derivada no se anula en dicho intervalo), es decir, dado que la segunda derivada tiene signo constante en $[a, b]$, en uno de los dos extremos la función tiene el mismo signo que su segunda derivada.

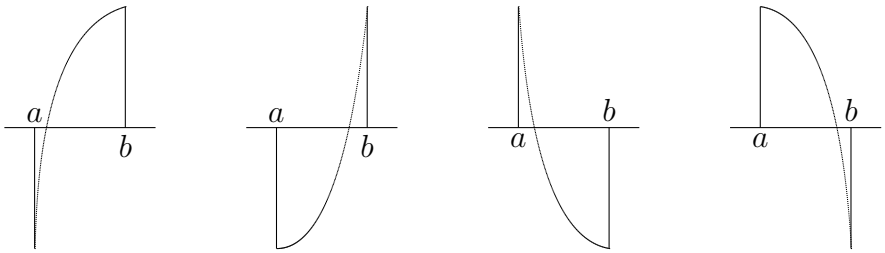
En estos casos, el método de Newton es convergente debiéndose tomar como valor inicial

$$x_0 = \begin{cases} a & \text{si } f(a) \cdot f''(a) > 0 \\ b & \text{si } f(b) \cdot f''(b) > 0 \end{cases}$$

es decir, el extremo en el que la función tiene el mismo signo que su derivada segunda.

Este resultado, que formalizamos a continuación en forma de teorema es conocido como *regla de Fourier*.

Teorema 1.4 [Regla de Fourier] *Sea $f(x)$ una función continua y dos veces derivable $[a, b]$. Si $\operatorname{sig} f(a) \neq \operatorname{sig} f(b)$ y sus dos primeras derivadas $f'(x)$ y $f''(x)$ no se anulan en $[a, b]$ existe una única raíz de la ecuación $f(x) = 0$ en dicho intervalo y se puede garantizar la convergencia del método de Newton-Raphson tomando como valor inicial x_0 el extremo del intervalo en el que la función y su segunda derivada tienen el mismo signo.*



$f'(x) > 0$, $f''(x) < 0$, $x_0 = a$ | $f'(x) > 0$, $f''(x) > 0$, $x_0 = b$ | $f'(x) < 0$, $f''(x) > 0$, $x_0 = a$ | $f'(x) < 0$, $f''(x) < 0$, $x_0 = b$

Figura 1.4: Los cuatro casos posibles

Gracias a que la convergencia es de segundo orden, es posible modificar el método de Newton para resolver ecuaciones que poseen raíces múltiples.

1.4.3 Método de Newton para raíces múltiples

Cuando el método de Newton converge lentamente nos encontramos con una raíz múltiple y, a diferencia de lo que ocurría con otros métodos, podemos modificar el método para acelerar la convergencia.

Sea $\overline{x}$ una raíz de multiplicidad k de la ecuación $f(x) = 0$. En este caso, el método de Newton converge muy lentamente y con grandes irregularidades debido al mal condicionamiento del problema.

Si en vez de hacer $x_{n+1} = x_n - \dfrac{f(x_n)}{f'(x_n)}$ hacemos

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

donde k representa el orden de la primera derivada que no se anula para $x = \overline{x}$ (multiplicidad de la raíz $\overline{x}$), el método sigue siendo de segundo orden.

En la práctica, el problema es que no conocemos k pero a ello nos ayuda la rapidez del método.

Ejemplo 1.5 Para resolver la ecuación $x - \operatorname{sen} x = 0$ comenzamos por expresarla de la forma $x = \operatorname{sen} x$, por lo que las soluciones serán los puntos de intersección de la curva $y = \operatorname{sen} x$ con $y = x$ (Fig.1.5).

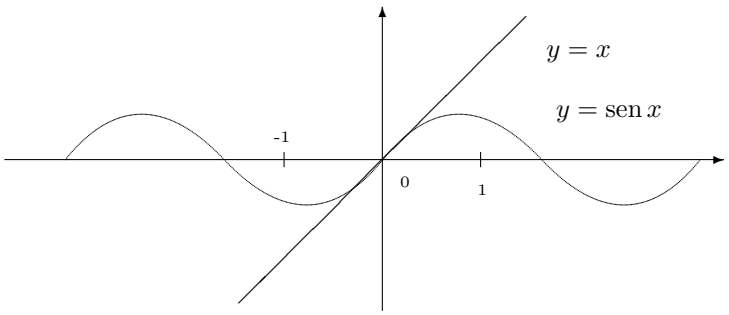


Figura 1.5: Las gráficas de $y = x$ e $y = \operatorname{sen} x$.

Aunque es conocido que la solución de la ecuación es $x = 0$, supondremos que sólo conocemos que está comprendida entre -1 y 1 y vamos aplicar el método de Newton.

$$x_{n+1} = x_n - \frac{x_n - \operatorname{sen} x_n}{1 - \cos x_n} = \frac{\operatorname{sen} x_n - x_n \cos x_n}{1 - \cos x_n}$$

Comenzando con $x_0 = 1$ se obtiene:

$$\begin{array}{l} x_0 = 1 \\ \cdots\cdots\cdots\cdots\cdots\cdots \\ x_{10} = 0'016822799\ldots \quad \left\{ \begin{array}{l} f'(x_{10}) = 0'0001\ldots \\ f''(x_{10}) = 0'016\ldots \\ f'''(x_{10}) = 0'9998\ldots \end{array} \right. \\ \cdots\cdots\cdots\cdots\cdots\cdots \\ x_{20} = 0'0000194\ldots \quad \left\{ \begin{array}{l} f'(x_{20}) = 0'00000001\ldots \\ f''(x_{20}) = 0'0019\ldots \\ f'''(x_{20}) = 0'9999\ldots \end{array} \right. \end{array}$$

Como la convergencia es muy lenta, hace pensar que se trata de una raíz múltiple. Además, como la primera y la segunda derivadas tienden a cero y la tercera lo hace a 1, parece que nos encontramos ante una raíz triple, por lo que aplicamos el método generalizado de Newton.

$$x_{n+1} = x_n - 3\frac{x_n - \operatorname{sen} x_n}{1 - \cos x_n}$$

que comenzando, al igual que antes, por $x_0 = 1$ se obtiene:

$$\begin{aligned} x_0 &= 1 \\ x_1 &= -0'034\ldots \\ x_2 &= 0'000001376\ldots \\ x_3 &= 0'00000000000009\ldots \end{aligned}$$

que se ve que converge rápidamente a $\overline{x} = 0$.

Aplicamos ahora la cota del error a posteriori a este valor y obtenemos:

$\overline{x} = 0 \Longrightarrow f(\overline{x}) = \overline{x} - \operatorname{sen}\overline{x} = 0 \Longrightarrow$ la solución es exacta.

$$\left.\begin{array}{l} f'(x) = 1 - \cos x \Longrightarrow f'(\overline{x}) = 0 \\ f''(x) = \operatorname{sen} x \Longrightarrow f''(\overline{x}) = 0 \\ f'''(x) = \cos x \Longrightarrow f'''(\overline{x}) = 1 \end{array}\right\} \Longrightarrow \text{se trata de una raíz triple.}$$

□

1.5 Un problema mal condicionado: ceros de un polinomio

Si queremos resolver la ecuación $P(x) = 0$ el problema viene condicionado por la función $1/P'(x)$ (su número de condición), por lo que si la derivada es muy pequeña el problema estará mal condicionado, pero si la derivada es muy grande cualquier método se hace muy lento, por lo que lo ideal sería que la derivada estuviese muy próxima a 1, pero claro está, ese derivada ha de estar muy próxima a 1 en todas las raíces del polinomio, cosa que es estadísticamente casi imposible. Por tanto, el cálculo de los ceros de un polinomio es un ejemplo de problema mal condicionado. Sin embargo, vamos a estudiar cómo podemos aproximar un determinado cero del polinomio.

Hemos visto que uno de los mayores problemas que presenta la resolución de una ecuación es encontrarnos que posee raíces múltiples ya que, en un entorno de ellas, o bien la función no cambia de signo, o bien se aproxima *demasiado* a cero y, por tanto, cualquier método puede dar soluciones erróneas.

En el caso de las ecuaciones algebraicas ($P_n(x) = 0$) este problema puede solventarse buscando otra ecuación que posea las mismas raíces que la dada pero todas ellas simples, es decir, eliminando las raíces múltiples.

Por el *teorema fundamental del Álgebra* sabemos que $P_n(x)$ posee n raíces y, por tanto, puede ser factorizado de la forma

$$P_n(x) = a_0(x - x_1)(x - x_2)\cdots(x - x_n)$$

donde $\{x_1, x_2, \ldots, x_n\}$ son los ceros del polinomio.

Si existen raíces múltiples, las podemos agrupar para obtener:

$$P_n(x) = a_0(x - x_1)^{m_1}(x - x_2)^{m_2} \cdots (x - x_k)^{m_k}$$

donde m_i $(i = 1, 2, \ldots k)$ representa la multiplicidad de la raíz x_i y verificándose que $m_1 + m_2 + \cdots m_k = n$

Derivando esta expresión obtenemos:

$$P'(x) = na_0(x - x_1)^{m_1 - 1} \cdots (x - x_k)^{m_k - 1} Q_{k-1}(x)$$

con $Q_{k-1}(x_i) \neq 0 \quad i = 1, 2, \ldots, k$

Por tanto, si $\overline{x}$ es una raíz de la ecuación $P(x) = 0$ con multiplicidad k, es también una raíz de $P'(x) = 0$ pero con multiplicidad $k - 1$.

$$D(x) = \text{mcd}\,[P(x), P'(x)] = (x - x_1)^{m_1 - 1} \cdots (x - x_k)^{m_k - 1}$$

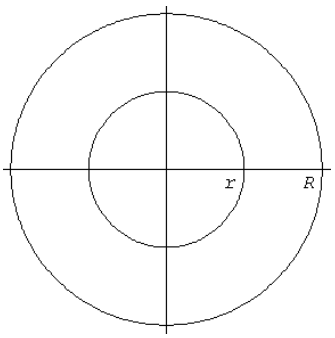
por lo que

$$Q(x) = \frac{P(x)}{D(x)} = a_0(x - x_1)(x - x_2) \cdots (x - x_k)$$

Es decir, hemos encontrado un polinomio cuyas raíces son las mismas que las de $P(x)$ pero todas ellas simples.

Si ya conocemos que una ecuación sólo tiene raíces simples y queremos calcularlas, parece apropiado que un primer paso consista en detectar dónde pueden encontrarse. Así por ejemplo, si son reales, determinar intervalos de una amplitud reducida en los que se encuentren las raíces de la ecuación.

Definición 1.2 Dada una ecuación $f(x) = 0$ (en general compleja) se denomina *acotar las raíces* a buscar dos números reales positivos r y R tales que $r \leq |\overline{x}| \leq R$ para cualquier raíz $\overline{x}$ de la ecuación.



Geométricamente consiste en determinar una corona circular de radios r y R dentro de la cual se encuentran todas las raíces.
En el caso real se reduce a los intervalos $(-R, -r)$ y (r, R).

Veamos, a continuación, una cota para las raíces de una ecuación algebraica.

Proposición 1.5 *Si $\overline{x}$ es una raíz de la ecuación*

$$P(x) = a_0x^n + a_1x^{n-1} + \cdots + a_{n-1}x + a_n = 0$$

se verifica que:

$$|\overline{x}| < 1 + \frac{A}{|a_0|} \qquad \textit{siendo} \qquad A = \max_{i\geq 1}|a_i|$$

Demostración. Sea $P(x) = a_0x^n + a_1x^{n-1} + \cdots + a_{n-1}x + a_n$. Tomando módulos tenemos

$$\begin{aligned}
|P(x)| &= |a_0x^n + a_1x^{n-1} + \cdots + a_{n-1}x + a_n| \geq \\
&\geq |a_0x^n| - |a_1x^{n-1} + \cdots + a_{n-1}x + a_n| \geq \\
&\geq |a_0x^n| - \Big[\,|a_1x^{n-1}| + \cdots + |a_{n-1}x| + |a_n|\,\Big] = \\
&= |a_0x^n| - \Big[\,|a_1|\,|x|^{n-1} + \cdots + |a_{n-1}|\,|x| + |a_n|\,\Big] \geq \\
&\geq |a_0x^n| - A\left[|x|^{n-1} + \cdots + |x| + 1\right]
\end{aligned}$$

(Para considerar el último paréntesis como una progresión geométrica habría que añadir los términos que, probablemente, falten en $P(x)$ y suponer que, además, es $|x| \neq 1$).

$$|P(x)| \geq |a_0|\,|x|^n - A\frac{|x|^n - 1}{|x| - 1}$$

Dado que el teorema es trivial para $|x| < 1$, supondremos que $|x| > 1$ y entonces:

$$|P(x)| > |a_0|\,|x|^n - A\frac{|x|^n}{|x| - 1} = |x|^n\left(|a_0| - \frac{A}{|x| - 1}\right)$$

Como la expresión anterior es cierta para cualquier $|x| > 1$, sea $|\overline{x}| > 1$ con $P(\overline{x}) = 0$. Entonces

$$0 > |\overline{x}|^n\left(|a_0| - \frac{A}{|\overline{x}| - 1}\right) \Longrightarrow |a_0| - \frac{A}{|\overline{x}| - 1} < 0 \Longrightarrow$$

$$|a_0| < \frac{A}{|\overline{x}| - 1} \Longrightarrow |\overline{x}| - 1 < \frac{A}{|a_0|} \Longrightarrow$$

$$|\overline{x}| < 1 + \frac{A}{|a_0|} \qquad \text{con} \qquad |\overline{x}| > 1$$

Es evidente que esta cota también la verifican las raíces $\overline{x}$ con $|\overline{x}| < 1$. ■

Proposición 1.6 [Regla de Laguerre] *Consideremos la ecuación*

$$P(x) = a_0x^n + a_1x^{n-1} + \cdots + a_{n-1}x + a_n = 0$$

Sean $C(x) = b_0x^{n-1} + \cdots + b_{n-2}x + b_{n-1}$ *el cociente y* r *el resto de la división de* $P(x)$ *entre* $x - c$. *Si* $r \geq 0$ *y* $b_i \geq 0$ *para* $0 \leq i \leq n-1$, *el número real* c *es una cota superior para las raíces positivas de la ecuación. (Trivialmente lo es también para las raíces negativas).*

El procedimiento consiste en comenzar con la cota obtenida anteriormente (que no suelen ser muy buena) e ir disminuyéndola hasta afinarla todo lo que podamos.

Ejemplo 1.6 Consideremos la ecuación $2x^4 - 9x^3 - x^2 + 24x + 12 = 0$.

Sabemos que

$$|\overline{x}| < 1 + \frac{A}{|a_0|} \text{ siendo } A = \max_{i\geq 1} |a_i| \Longrightarrow |\overline{x}| < 1 + \frac{24}{2} = 13$$

por lo que cualquier raíz real del polinomio debe estar en el intervalo $(-13, 13)$.

Aplicando la regla de Laguerre obtenemos:

	2	−9	−1	24	12
4		8	−4	−20	16
	2	−1	−5	4	28

	2	−9	−1	24	12
5		10	5	20	240
	2	1	4	48	252

Dado que para $x = 4$ se obtienen valores negativos *no podemos garantizar* que 4 sea una cota superior para las raíces positivas de la ecuación, pero para $x = 5$ todos los valores obtenidos han sido positivos, por lo que podemos garantizar que las raíces reales de la ecuación se encuentran en el intervalo $(-13, 5)$.

No debemos confundir el hecho de que la regla de Laguerre no nos garantice que 4 sea una cota superior de las raíces positivas y 5 sí lo sea, con que la ecuación deba tener alguna raíz en el intervalo $(4, 5)$. En otras palabras, 4 *puede ser* una cota superior para las raíces positivas de la ecuación aunque la regla de Laguerre no lo garantice. De hecho, la mayor de las raíces positivas de nuestra ecuación es $x = 3'56155281280883\ldots < 4$. □

Las cotas obtenidas anteriormente nos delimitan la zona en la que debemos estudiar la existencia de soluciones de la ecuación pero, en realidad, lo que más nos acerca a nuestro problema (resolver la ecuación) es *separar* cada raíz en un intervalo. A este proceso se le conoce como *separación de raíces* y estudiaremos un método que se conoce como *método de Sturm* que nos permite separar las raíces de una ecuación.

1.5.1 Sucesiones de Sturm

Una *sucesión de Sturm* en $[a, b]$ es un conjunto de funciones continuas en dicho intervalo $f_0(x),\ f_1(x), \ldots,\ f_n(x)$ que cumplen las siguientes condiciones:

- $f_n(x) \neq 0$ cualquiera que sea $x \in [a, b]$. Es decir, el signo de $f_n(x)$ permanece constante en el intervalo $[a, b]$.
- Las funciones $f_i(x)$ y $f_{i+1}(x)$ no se anulan simultáneamente. En otras palabras, si $f_i(c) = 0$ entonces $f_{i-1}(c) \neq 0$ y $f_{i+1}(c) \neq 0$.
- Si $f_i(c) = 0$ entonces $f_{i-1}(c)$ y $f_{i+1}(c)$ tienen signos opuestos, es decir, $f_{i-1}(c) \cdot f_{i+1}(c) < 0$. (Engloba al apartado anterior).
- Si $f_0(c) = 0$ con $c \in [a, b]$ entonces $\dfrac{f_0(x)}{f_1(x)}$ pasa de negativa a positiva en c. (Está bien definida en c por ser $f_1(c) \neq 0$ y es creciente en dicho punto).

Teorema 1.7 [Teorema de Sturm]. *Sea $f_0(x),\ f_1(x), \ldots,\ f_n(x)$ una sucesión de Sturm en el intervalo $[a, b]$ y consideremos las sucesiones*

$$\begin{array}{cccc} \mathrm{sig}[f_0(a)] & \mathrm{sig}[f_1(a)] & \cdots & \mathrm{sig}[f_n(a)] \\ \mathrm{sig}[f_0(b)] & \mathrm{sig}[f_1(b)] & \cdots & \mathrm{sig}[f_n(b)] \end{array}$$

teniendo en cuenta que si alguna de las funciones se anula en uno de los extremos del intervalo $[a, b]$ pondremos en su lugar, indistintamente, signo + o - y denotemos por N_1 al número de cambios de signo de la primera sucesión y por N_2 al de la segunda (siempre es $N_1 \geq N_2$).

En estas condiciones, el número de raíces reales existentes en el intervalo $[a, b]$ de la ecuación $f_0(x) = 0$ viene dado por $N_1 - N_2$.

La construcción de una sucesión de Sturm es, en general, complicada. Sin embargo, cuando se trabaja con funciones polinómicas, el problema es mucho

más simple además de que siempre es posible construir una sucesión de Sturm válida para cualquier intervalo.

Dada la ecuación algebraica $P_n(x) = 0$, partimos de

$$f_0(x) = P_n(x) \qquad \text{y} \qquad f_1(x) = P_n'(x)$$

Para determinar las demás funciones de la sucesión vamos dividiendo $f_{i-1}(x)$ entre $f_i(x)$ para obtener

$$f_{i-1}(x) = c_i(x) \cdot f_i(c) + r_i(x)$$

donde $r_i(x)$ tiene grado inferior al de $f_i(x)$ y hacemos

$$f_{i+1}(x) = -r_i(x)$$

Como el grado de $f_i(x)$ va decreciendo, el proceso es finito. Si se llega a un resto nulo (el proceso que estamos realizando es precisamente el del algoritmo de Euclides) la ecuación posee raíces múltiples y se obtiene el máximo común divisor $D(x)$ de $P_n(x)$ y $P_n'(x)$. Dividiendo $P_n(x)$ entre $D(x)$ obtenemos una nueva ecuación que sólo posee raíces simples. La sucesión $f_i(x)/D(x)$ es una sucesión de Sturm para la ecuación $P(x)/D(x) = Q(x) = 0$ que posee las mismas raíces que $P(x) = 0$ pero todas simples.

Si llegamos a un resto constante, no nulo, es éste quien nos determina la finalización del proceso. Hemos obtenido, de esta manera, una sucesión de Sturm válida para cualquier intervalo $[a, b]$.

Nota: Obsérvese que, al igual que en el algoritmo de Euclides, podemos ir multiplicando los resultados parciales de las divisiones por cualquier constante *positiva* no nula, ya que sólo nos interesa el resto (salvo constantes positivas) de la división.

Ejemplo 1.7 Vamos a construir una sucesión de Sturm que nos permita separar las raíces de la ecuación $x^4 + 2x^3 - 3x^2 - 4x - 1 = 0$.

$f_0(x) = x^4 + 2x^3 - 3x^2 - 4x - 1.$ $\qquad f_0'(x) = 4x^3 + 6x^2 - 6x - 4.$

$f_1(x) = 2x^3 + 3x^2 - 3x - 2.$

$$\begin{array}{rl}
2x^4 + 4x^3 - 6x^2 - 8x - 2 & |\underline{2x^3 + 3x^2 - 3x - 2} \\
\underline{-2x^4 - 3x^3 + 3x^2 + 2x} & \quad x + 1 \\
x^3 - 3x^2 - 6x - 2 & \quad \textit{multiplicando por } 2 \\
2x^3 - 6x^2 - 12x - 4 & \\
\underline{-2x^3 - 3x^2 + 3x + 2} & \\
-9x^2 - 9x - 2 &
\end{array}$$

$f_2(x) = 9x^2 + 9x + 2$.

$$\begin{array}{rl} 18x^3 + 27x^2 - 27x - 18 & |\underline{9x^2 + 9x + 2} \\ \underline{-\,18x^3 - 18x^2 - \;\;4x} & \;\;2x + 1 \\ 9x^2 - 31x - 18 & \\ \underline{-\;9x^2 - \;\;9x - \;\;2} & \\ -40x - 20 & \end{array}$$

$f_3(x) = 2x + 1$.

$$\begin{array}{rl} 18x^2 + 18x + 4 & |\underline{2x + 1} \\ \underline{-\,18x^2 - \;\;9x} & \;\;9x + 9 \\ 9x + 4 & \textit{multiplicando por } 2 \\ 18x + 8 & \\ \underline{-\,18x - 9} & \\ -1 & \end{array}$$

$f_4(x) = 1$.

	$-\infty$	-3	-2	-1	0	1	2	$+\infty$
$f_0(x) = x^4 + 2x^3 - 3x^2 - 4x - 1$	$+$	$+$	$-$	$-$	$-$	$-$	$+$	$+$
$f_1(x) = 2x^3 + 3x^2 - 3x - 2$	$-$	$-$	$\pm$	$+$	$-$	$\pm$	$+$	$+$
$f_2(x) = 9x^2 + 9x + 2$	$+$	$+$	$+$	$+$	$+$	$+$	$+$	$+$
$f_3(x) = 2x + 1$	$-$	$-$	$-$	$-$	$+$	$+$	$+$	$+$
$f_4(x) = 1$	$+$	$+$	$+$	$+$	$+$	$+$	$+$	$+$
cambios de signo	4	4	3	3	1	1	0	0

Sabemos, por ello, que existe una raíz en el intervalo $(-3, -2)$, dos raíces en el intervalo $(-1, 0)$ y una cuarta raíz en el intervalo $(1, 2)$.

Como $f_0(-1) = -1 < 0$, $f_0(-0'5) = 0'0625 > 0$ y $f_0(0) = -1 < 0$ podemos separar las raíces existentes en el intervalo $(-1, 0)$ y decir que las cuatro raíces de la ecuación dada se encuentran en los intervalos

$$(-3 - 2) \qquad (-1, -0'5) \qquad (-0'5, 0) \qquad \text{y} \qquad (1, 2) \qquad \square$$

Si, una vez eliminadas las raíces múltiples y separadas las raíces, queremos resolver la ecuación, utilizaremos (excepto en casos muy determinados como el

del Ejemplo 1.4) el método de Newton-Raphson. Al aplicarlo nos encontramos con que tenemos que calcular, en cada paso, los valores de $P(x_n)$ y $P'(x_n)$ por lo que vamos a ver, a continuación, un algoritmo denominado *algoritmo de Horner* que permite realizar dichos cálculos en tiempo lineal.

1.5.2 Algoritmo de Horner

Supongamos un polinomio $P(x)$ y un número real (en general también puede ser complejo) $x_0 \in \mathbf{R}$. Si dividimos $P(x)$ entre $x - x_0$ sabemos que el resto es un polinomio de grado cero, es decir, un número real, por lo que

$$P(x) = (x - x_0)Q(x) + r \quad \text{con} \quad \begin{cases} r \in \mathbf{R} \\ \text{y} \\ \text{gr}[Q(x)] = \text{gr}[P(x)] - 1 \end{cases}$$

Haciendo $x = x_0$ obtenemos que

$$P(x_0) = 0 \cdot Q(x_0) + r \Longrightarrow P(x_0) = r$$

Este resultado es conocido como *teorema del resto* y lo enunciamos a continuación.

Teorema 1.8 [Teorema del resto] *El valor numérico de un polinomio $P(x)$ para $x = x_0$ viene dado por el resto de la división de $P(x)$ entre $x - x_0$.*

Sea

$$P(x) = a_0x^n + a_1x^{n-1} + \cdots + a_{n-1}x + a_n.$$

Si llamamos b_i $(0 \leq i \leq n-1)$ a los coeficientes del polinomio cociente

$$\frac{P(x) - P(x_0)}{x - x_0} = Q(x) = b_0x^{n-1} + b_1x^{n-2} + \cdots + b_{n-2}x + b_{n-1}$$

se tiene que

$$\begin{aligned} b_0 &= a_0 \\ b_1 &= a_1 + x_0b_0 \\ b_2 &= a_2 + x_0b_1 \\ &\vdots \\ b_{n-1} &= a_{n-1} + x_0b_{n-2} \\ r = P(x_0) &= a_n + x_0b_{n-1} \end{aligned}$$

Este procedimiento para calcular el polinomio cociente $Q(x)$ y el valor numérico de $P(x_0)$ es conocido como *algoritmo de Horner* .

Una regla útil para hacer los cálculos a mano es la conocida *regla de Ruffini* que consiste en disponer las operaciones como se indica a continuación.

	a_0	a_1	a_2	$\cdots$	a_{n-1}	a_n
x_0		x_0b_0	x_0b_1	$\cdots$	x_0b_{n-2}	x_0b_{n-1}
	b_0	b_1	b_2	$\cdots$	b_{n-1}	$P(x_0)$

Además, dado que

$$P(x) = Q(x)(x - x_0) + P(x_0) \Longrightarrow P'(x) = Q'(x)(x - x_0) + Q(x)$$

se tiene que

$$P'(x_0) = Q(x_0)$$

y el cálculo de $Q(x_0)$ es análogo al que hemos realizado para $P(x_0)$, es decir, aplicando el algoritmo de Horner a $Q(x)$ obtenemos

$$Q(x) = C(x)(x - x_0) + Q(x_0) \quad \text{donde} \quad Q(x_0) = P'(x_0).$$

Si utilizamos la regla de Ruffini sólo tenemos que volver a dividir por x_0 como se muestra a continuación.

	a_0	a_1	a_2	$\cdots$	a_{n-2}	a_{n-1}	a_n
x_0		x_0b_0	x_0b_1	$\cdots$	x_0b_{n-3}	x_0b_{n-2}	x_0b_{n-1}
	b_0	b_1	b_2	$\cdots$	b_{n-2}	b_{n-1}	$P(x_0)$
x_0		x_0c_0	x_0c_1	$\cdots$	x_0c_{n-3}	x_0c_{n-2}	
	c_0	c_1	c_2	$\cdots$	c_{n-2}	$P'(x_0)$	

Ejemplo 1.8 Consideremos el polinomio $P(x) = 2x^4 + x^3 - 3x^2 + 4x - 5$ y vamos a calcular los valores de P(2) y P'(2). Aplicando reiteradamente al regla de Ruffini obtenemos

	2	1	-3	4	-5
2		4	10	14	36
	2	5	7	18	$\boxed{31}$
2		4	18	50	
	2	9	25	$\boxed{68}$	

por lo que

$$P(2) = 31 \quad \text{y} \quad P'(2) = 68$$

□

Evidentemente, la regla de Ruffini nos es útil para realizar cálculos a mano con una cierta facilidad, pero cuando los coeficientes de $P(x)$ y el valor de x_0 no son enteros sino que estamos trabajando con varias cifras decimales, deja de ser efectivo y debemos recurrir al algoritmo de Horner en una máquina.

1.6 Sistemas de ecuaciones no lineales

Dado un sistema de ecuaciones no lineales

$$\begin{aligned} f_1(x_1, x_2, \ldots, x_m) &= 0 \\ f_2(x_1, x_2, \ldots, x_m) &= 0 \\ &\vdots \\ f_m(x_1, x_2, \ldots, x_m) &= 0 \end{aligned}$$

podemos expresarlo de la forma $f(\mathbf{x}) = 0$ donde $\mathbf{x}$ es un vector de $\mathbf{R}^m$ y f una función vectorial de variable vectorial, es decir:

$$f : D \subset \mathbf{R}^m \to \mathbf{R}^m$$

o lo que es lo mismo, $f = (f_1, f_2, \ldots, f_m)$ con $f_i : \mathbf{R}^m \to \mathbf{R}$ para $1 \leq i \leq m$.

Así, por ejemplo, el sistema

$$\left.\begin{aligned} x^2 - 2x - y + 0'5 &= 0 \\ x^2 + 4y^2 + 4 &= 0 \end{aligned}\right\} \tag{1.3}$$

puede considerarse como la ecuación $f(\mathbf{x}) = 0$ (obsérvese que 0 representa ahora al vector nulo, es decir, que $0 = (0,0)^T$) con $\mathbf{x} = (x, y)^T$ y $f = (f_1, f_2)$ siendo

$$\begin{cases} f_1(\mathbf{x}) = x^2 - 2x - y + 0'5 \\ f_2(\mathbf{x}) = x^2 + 4y^2 + 4 \end{cases}$$

Como hemos transformado nuestro sistema en una ecuación del tipo $f(\mathbf{x}) = 0$, parece lógico que tratemos de resolverla por algún método paralelo a los estudiados para ecuaciones no lineales como puedan ser la utilización del teorema del punto fijo o el método de Newton.

Si buscamos un método iterado basado en el teorema del punto fijo, debemos escribir la ecuación $f(\mathbf{x}) = 0$ de la forma $\mathbf{x} = F(\mathbf{x})$ (proceso que puede

realizarse de muchas formas, la más sencilla es hacer $F(\mathbf{x}) = \mathbf{x} + f(\mathbf{x}))$ para, partiendo de un vector inicial $\mathbf{x}_0$ construir la sucesión de vectores

$$\mathbf{x}_{n+1} = F(\mathbf{x}_n) \qquad \begin{cases} x_{n+1}^1 = F_1(x_n^1, x_n^2, \ldots, x_n^m) \\ x_{n+1}^2 = F_2(x_n^1, x_n^2, \ldots, x_n^m) \\ \quad \vdots \\ x_{n+1}^m = F_m(x_n^1, x_n^2, \ldots, x_n^m) \end{cases}$$

En el ejemplo (1.3) podemos hacer

$$x^2 - 2x - y + 0'5 = 0 \Longrightarrow 2x = x^2 - y + 0'5 \Longrightarrow x = \frac{x^2 - y + 0'5}{2}$$

$$x^2 + 4y^2 - 4 = 0 \Longrightarrow x^2 + 4y^2 + y - 4 = y \Longrightarrow y = x^2 + 4y^2 + y - 4$$

es decir,

$$\mathbf{x} = F(\mathbf{x}) \quad \text{con} \quad \begin{cases} \mathbf{x} = (x, y)^T \\ \quad \text{y} \\ F(\mathbf{x}) = (\dfrac{x^2 - y + 0'5}{2}, x^2 + 4y^2 + y - 4)^T \end{cases}$$

Si $\mathbf{x}$ es una solución de la ecuación y $\mathbf{x}_{n+1}$ es una aproximación obtenida, se tiene que

$$\|\mathbf{x} - \mathbf{x}_{n+1}\| = \|F(\mathbf{x}) - F(\mathbf{x}_n)\| = \|F'(\alpha)(\mathbf{x} - \mathbf{x}_n)\| \leq \|F'(\alpha)\| \cdot \|\mathbf{x} - \mathbf{x}_n\|$$

por lo que si $\|F'(\mathbf{x})\| \leq q < 1$ para cualquier punto de un determinado entorno de la solución, se tiene que

$$\|\mathbf{x} - \mathbf{x}_{n+1}\| \leq \|\mathbf{x} - \mathbf{x}_n\|$$

y la sucesión

$$\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_n, \ldots$$

converge a la única raíz de la ecuación $\mathbf{x} = F(\mathbf{x})$ en la bola considerada (intervalo de $\mathbf{R}^m$).

Es importante observar que:

a) Se ha utilizado el concepto de norma vectorial al hacer uso de $\|\mathbf{x} - \mathbf{x}_n)\|$.

b) Se ha utilizado una versión del teorema del valor medio para varias variables al decir que

$$\|F(\mathbf{x}) - F(\mathbf{x}_n)\| \leq \|F'(\alpha)\|\|(\mathbf{x} - \mathbf{x}_n)\|$$

c) Se ha utilizado el concepto de norma matricial al hacer uso de $\|F'(\alpha)\|$ ya que $F'(\mathbf{x})$ es la matriz jacobiana de la función F, es decir

$$F'(\mathbf{x}) = \begin{pmatrix} \dfrac{\partial F_1}{\partial x_1} & \cdots & \dfrac{\partial F_1}{\partial x_m} \\ \vdots & \ddots & \vdots \\ \dfrac{\partial F_m}{\partial x_1} & \cdots & \dfrac{\partial F_m}{\partial x_m} \end{pmatrix}$$

d) Se supone que $\|F'(\alpha)(\mathbf{x} - \mathbf{x}_n)\| \leq \|F'(\alpha)\| \cdot \|\mathbf{x} - \mathbf{x}_n\|$ o de forma más general, que para cualquier matriz A (cuadrada de orden n) y cualquier vector de $\mathbf{R}^n$ se verifica que

$$\|Ax\| \leq \|A\| \cdot \|x\|$$

e) Que el teorema del punto fijo es generalizable a funciones vectoriales de variable vectorial.

1.6.1 Método de Newton

Consideremos la ecuación $f(\mathbf{x}) = 0$ (donde f es una función vectorial de variable vectorial) equivalente a un sistema de ecuaciones no lineales.

Sea $\mathbf{x}$ la solución exacta del sistema y $\mathbf{x}_n$ una aproximación de ella. Si denotamos por $h = \mathbf{x} - \mathbf{x}_n$ se tiene que

$$\mathbf{x} = \mathbf{x}_n + h$$

y haciendo uso del desarrollo de Taylor obtenemos que

$$0 = f(\mathbf{x}) = f(\mathbf{x}_n + h) \simeq f(\mathbf{x}_n) + f'(\mathbf{x}_n)h$$

de donde

$$h \simeq -f'^{-1}(\mathbf{x}_n)f(\mathbf{x}_n)$$

y, por tanto

$$\mathbf{x} \simeq \mathbf{x}_n - f'^{-1}(\mathbf{x}_n)f(\mathbf{x}_n).$$

Esta aproximación es que utilizaremos como valor de $\mathbf{x}_{n+1}$, es decir

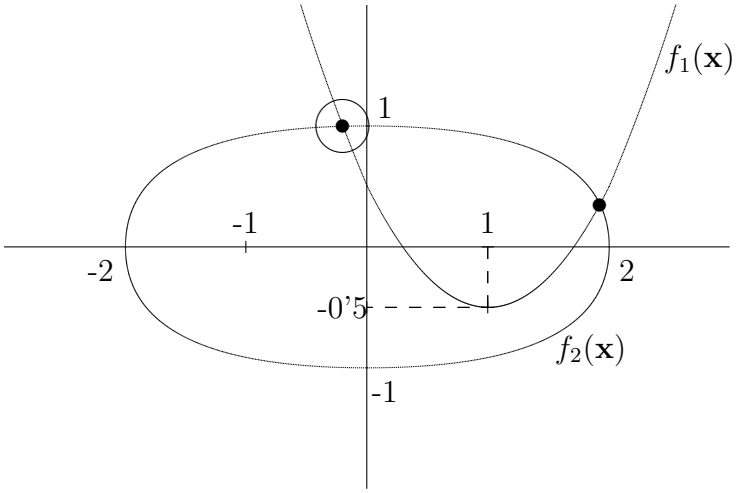
$$\mathbf{x}_{n+1} = \mathbf{x}_n - f'^{-1}(\mathbf{x}_n) f(\mathbf{x}_n)$$

Obsérvese entonces que cada iteración del método de Newton se reduce al cálculo del vector h correspondiente y éste no es más que la solución del sistema de ecuaciones lineales

$$f'(\mathbf{x}_n)h + f(\mathbf{x}_n) = 0$$

En el ejemplo (1.3) se tiene que $f(\mathbf{x}) = 0$ con $\mathbf{x} = (x, y)^T$ y $f = (f_1, f_2)^T$ donde

$$f_1(\mathbf{x}) = x^2 - 2x - y + 0'5 \quad \text{y} \quad f_2(\mathbf{x}) = x^2 + 4y^2 - 4$$



Tomando como valor inicial $\mathbf{x}_0 = (-0'5, 1)^T$ se tiene que

$$f(\mathbf{x}_0) = (0'75, 0'25)^T$$

$$J(\mathbf{x}) = \begin{pmatrix} 2x-2 & -1 \\ 2x & 8y \end{pmatrix} \Longrightarrow f'(\mathbf{x}_0) = J(\mathbf{x}_0) = \begin{pmatrix} -3 & -1 \\ -1 & 8 \end{pmatrix}$$

por lo que debemos resolver el sistema

$$\begin{pmatrix} -3 & -1 \\ -1 & 8 \end{pmatrix} \begin{pmatrix} h_1^1 \\ h_1^2 \end{pmatrix} = \begin{pmatrix} -0'75 \\ -0'25 \end{pmatrix}$$

cuya solución es $h_1 = \begin{pmatrix} h_1^1 \\ h_1^2 \end{pmatrix} = \begin{pmatrix} 0'25 \\ 0 \end{pmatrix}$ y, por tanto

$$\mathbf{x}_1 = \mathbf{x}_0 + h_1 = \begin{pmatrix} -0'25 \\ 1 \end{pmatrix}$$

Aplicando de nuevo el método se obtiene

$$f(\mathbf{x}_1) = \begin{pmatrix} 0'0625 \\ 0'0625 \end{pmatrix} \qquad f'(\mathbf{x}_1) = J(\mathbf{x}_1) = \begin{pmatrix} -0'25 & -1 \\ -0'5 & 8 \end{pmatrix}$$

obteniéndose el sistema

$$\begin{pmatrix} -0'25 & -1 \\ -0'5 & 8 \end{pmatrix} \begin{pmatrix} h_2^1 \\ h_2^2 \end{pmatrix} = \begin{pmatrix} -0'0625 \\ -0'0625 \end{pmatrix}$$

cuya solución es $h_2 = \begin{pmatrix} h_2^1 \\ h_2^2 \end{pmatrix} = \begin{pmatrix} 0'022561\ldots \\ -0'006\ldots \end{pmatrix}$ y, por tanto

$$\mathbf{x}_2 = \mathbf{x}_1 + h_2 = \begin{pmatrix} -0'227439\ldots \\ 0'994\ldots \end{pmatrix}$$

En definitiva, podemos observar que la resolución de un sistema de ecuaciones no lineales mediante el método de Newton se reduce, en cada iteración, a la resolución de un sistema de ecuaciones lineales por lo que el tema siguiente lo dedicaremos al estudio de dichos sistemas de ecuaciones.

1.7 Ejercicios propuestos

Ejercicio 1.1 Dada la ecuación $xe^x - 1 = 0$, se pide:

a) Estudiar gráficamente sus raíces reales y acotarlas.

b) Aplicar el método de la bisección y acotar el error después de siete iteraciones.

c) Aplicar el método de Newton, hasta obtener tres cifras decimales exactas.

Ejercicio 1.2 Probar que la ecuación $x^2 + \ln x = 0$ sólo tiene una raíz real y hallarla, por el método de Newton, con 6 cifras decimales exactas.

Ejercicio 1.3 Eliminar las raíces múltiples en la ecuación $x^6 - 2x^5 + 3x^4 - 4x^3 + 3x^2 - 2x + 1 = 0$. Resolver, exactamente, la ecuación resultante y comprobar la multiplicidad de cada raíz en la ecuación original.

Ejercicio 1.4 Dada la ecuación $8x^3 - 4x^2 - 18x + 9 = 0$, acotar y separar sus raíces reales.

Ejercicio 1.5 Dado el polinomio $P(x) = x^3 + 3x^2 + 2$ se pide:

a) Acotar sus raíces reales.

b) Probar, mediante una sucesión de Sturm, que $P(x)$ sólo posee una raíz real y determinar un intervalo de amplitud 1 que la contenga.

c) ¿Se verifican, en dicho intervalo, las hipótesis del teorema de Fourier? En caso afirmativo, determinar el extremo que debe tomarse como valor inicial x_0 para garantizar la convergencia del método de Newton.

d) Sabiendo que en un determinado momento del proceso de Newton se ha obtenido $x_n = -3.1958$, calcular el valor de x_{n+1} así como una cota del error en dicha iteración.

Ejercicio 1.6 Aplicar el método de Sturm para separar las raíces de la ecuación

$$2x^6 - 6x^5 + x^4 + 8x^3 - x^2 - 4x - 1 = 0$$

y obtener la mayor de ellas con seis cifras decimales exactas por el método de Newton.

Ejercicio 1.7 Se considera el polinomio $P(x) = x^3 - 6x^2 - 3x + 7$.

a) Probar, mediante una sucesión de Sturm, que posee una única raíz en el intervalo $(6, 7)$.

b) Si expresamos la ecuación $P(x) = 0$ de la forma $x = F(x) = \frac{1}{3}(x^3 - 6x^2 + 7)$, ¿podemos asegurar su convergencia?

c) Probar, aplicando el criterio de Fourier, que tomando como valor inicial $x_0 = 7$, el método de Newton es convergente.

d) Aplicando Newton con $x_0 = 7$ se ha obtenido, en la segunda iteración, $x_2 = 6'3039$. ¿Qué error se comete al aproximar la raíz buscada por el valor x_3 que se obtiene en la siguiente iteración?

Ejercicio 1.8 En este ejercicio se pretende calcular $\sqrt[10]{1}$ por el método de Newton. Consideramos, para ello, la función $f(x) = x^{10} - 1$ cuya gráfica se da en la Figura 1.

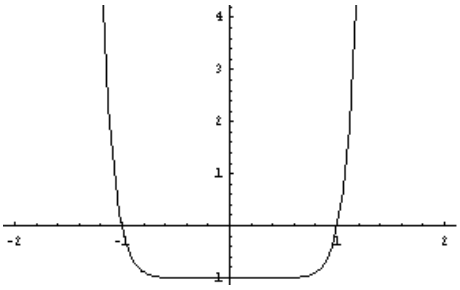


Fig. 1

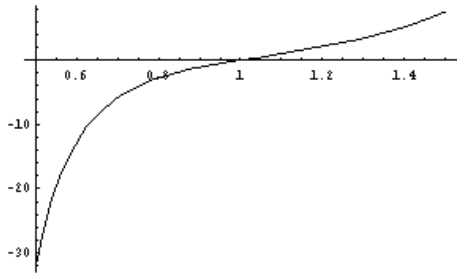


Fig. 2

a) Probar, analíticamente, que en el intervalo $[0'5, 1'5]$ posee una única raíz real.

b) Si tomamos $x_0 = 0'5$ obtenemos la raíz $x = 1$ en la iteración número 43, mientras que si tomamos $x_0 = 1'5$ se consigue el mismo resultado en la iteración número 9. ¿Cómo podríamos haber conocido *a priori* el valor que se debe elegir para x_0?

c) ¿Sabrías justificar el porqué de la extremada lentitud de la convergencia cuando iniciamos el proceso en $x_0 = 0'5$? y ¿por qué sigue siendo lento el proceso si comenzamos en $x_0 = 1'5$? Justifica las respuestas.

d) Dado que en el intervalo $[0'5, 1'5]$ no se anula la función x^5, las raíces de $f(x)$ son las mismas que las de $g(x) = f(x)/x^5 = x^5 - x^{-5}$ cuya gráfica se da en la Figura 2. ¿Se puede aplicar a $g(x)$ la regla de Fourier en dicho intervalo?

e) Si resolvemos, por el método de Newton, la ecuación $g(x) = 0$, ¿ se obtendrá la raíz con mayor rapidez que cuando lo hicimos con $f(x) = 0$? Justifica la respuesta sin calcular las iteraciones.

Ejercicio 1.9 Dada la ecuación $x^7 - 14x + 7 = 0$ se pide:

a) Probar que sólo tiene una raíz real negativa.

b) Encontrar un entero a de tal forma que el intervalo $[a, a+1]$ contenga a la menor de las raíces positivas de la ecuación.

c) ¿Cuál de los extremos del intervalo $[a, a+1]$ debe tomarse como valor inicial para asegurar la convergencia del método de Newton?

d) Aplicar el método de Newton para obtener la menor de las raíces positivas de la ecuación con seis cifras decimales exactas.

Ejercicio 1.10 Sea el polinomio $p(x) = x^4 - x^2 + 1/8$.

a) Utilizar el método de Sturm para determinar el número de raíces reales positivas del polinomio p, así como para separarlas.

b) Hallar los 2 primeros intervalos de la sucesión $([a_1, b_1], [a_2, b_2], \ldots)$ obtenida de aplicar el método de dicotomía para obtener la mayor raíz, r, del polinomio p. Elegir el intervalo $[a_1, b_1]$ de amplitud $1/2$ y tal que uno de sus extremos sea un número entero.

c) Sea la sucesión definida por la recurrencia $x_0 = 1$, $x_{n+1} = F(x_n)$, donde la iteración es la determinada por el método de Newton. Estudiar si la regla de Fourier aplicada al polinomio p en el intervalo $[a_1, b_1]$ del apartado anterior garantiza la convergencia de la sucesión a la raíz r. ¿Y en el intervalo $[a_2, b_2]$?

d) Hallar la aproximación x_1 del apartado anterior, determinando una cota del error cometido.

e) ¿Cuántas iteraciones se deben realizar para garantizar una aproximación de r con veinte cifras decimales exactas?
Indicación: $E_{n+1} = \dfrac{1}{k}(kE_1)^{2^n}$, con $k = \dfrac{\max|f''(x)|}{2\min|f'(x)|}$ en un intervalo adecuado.

Ejercicio 1.11 Dado el polinomio $P(x) = x^4 + 4x^3 - 2x^2 + 4x - 3$ se pide:

a) Acotar las raíces y construir una sucesión de Sturm para probar que sólo posee dos raíces reales, una positiva y otra negativa, dando intervalos de amplitud 1 que las contengan.

b) Partiendo de que la raíz positiva se encuentra en el intervalo $(0, 1)$ y despejando la x del término independiente

$$x = -\frac{1}{4}x^4 - x^3 + \frac{1}{2}x^2 + \frac{3}{4} \iff x = \varphi(x)$$

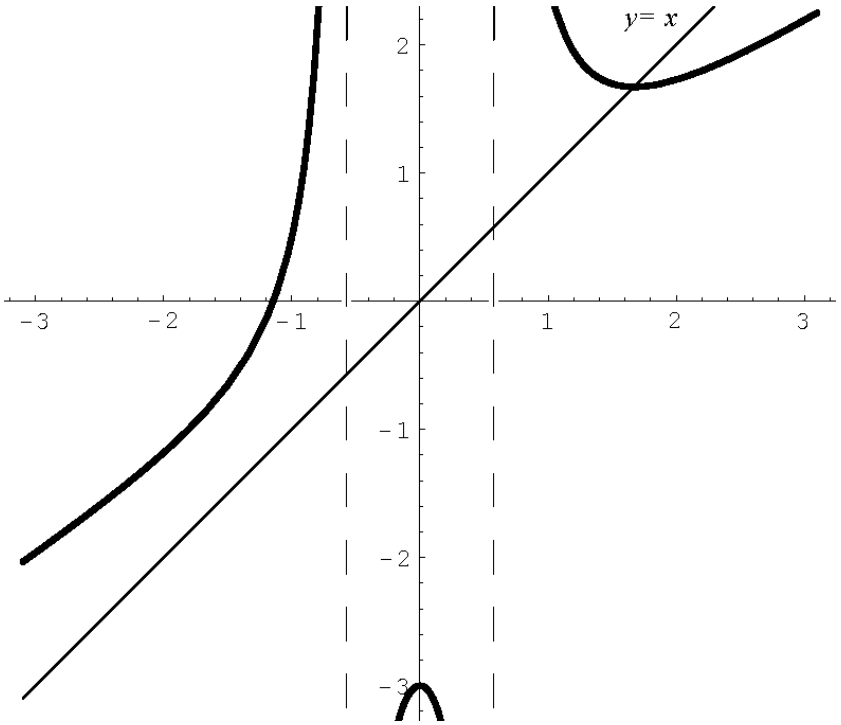
¿se puede asegurar la convergencia de la sucesión $x_1, x_2, \ldots, x_n, \ldots$ definida de la forma $x_1 = 0, \quad x_{n+1} = \varphi(x_n)$?

c) Aplicar Fourier para determinar el valor inicial que debe tomarse para garantizar la convergencia del método de Newton en el cálculo de la raíz negativa. ¿Tenemos las tres cifras exactas si tomamos como raíz -4.646 ?

Ejercicio 1.12 Sea el polinomio $p(x) = -3 - x + x^3$.

a) Utilizar una sucesión de Sturm para probar que el polinomio $p(x)$ sólo tiene una raíz $\alpha \in \mathbb{R}$ y que ésta se encuentra en el intervalo $I = [0, 3]$.

b) Comprobar que la gráfica adjunta se corresponde con la función $y = \varphi(x)$ cuya iteración, $x_{n+1} = \varphi(x_n) = x_n - p(x_n)/p'(x_n)$, es la obtenida con el método de Newton para resolver $p(x) = 0$. Tomando $x_1 = 0$, estudiar geométricamente (sobre el dibujo) si se obtendría una sucesión (x_n) convergente a α. ¿Y empezando en $x_1 = 3$?



c) Tomar un subintervalo de I en el que la regla de Fourier garantice la convergencia del Método de Newton y, con un valor inicial apropiado, obtener una aproximación de α con, al menos, tres cifras decimales exactas.

Ejercicio 1.13 Dado el polinomio $P(x) = x^3 + 6x^2 + 9x + k$ con $k \in \mathbf{R}$ se pide:

a) ¿Puede carecer de raíces reales? ¿y tener dos y sólo dos raíces reales?

b) Utilizar el método de Sturm para probar que tiene una única raíz real si, y sólo si, $k < 0$ o $k > 4$, y que sólo tiene raíces múltiples si $k = 0$ o $k = 4$ no existiendo, en ningún caso, una raíz triple.

c) Para $k = -4$ admite una única raíz real en el intervalo $[0, 1]$. Si tomamos como valor aproximado de la raíz $\mathbf{x} = 0'3553$ ¿de cuántas cifras decimales exactas disponemos?

d) Si, en el caso anterior en que $k = -4$, aplicamos el método de Newton para hallar la raíz del polinomio, ¿cuál de los extremos del intervalo $[0, 1]$ deberíamos tomar como valor inicial x_0 para garantizar la convergencia? y ¿qué valor obtendríamos para x_2?

Ejercicio 1.14 Dados los polinomios $P(x) = 2x^3 - 2x^2 - 2\alpha x + 3\alpha$ y $Q(x) = x^3 - 3x^2 - 3\alpha x + 2\alpha$, se pide:

a) Determinar el valor de α sabiendo que se trata de un entero par y que los valores de dichos polinomios sólo coinciden, para valores positivos de x, en un punto del intervalo $(1, 2)$.

b) Probar (mediante una sucesión de Sturm) que, para $\alpha = -2$, el polinomio $P(x)$ sólo tiene una raíz real, que ésta es positiva, y dar un intervalo de amplitud 1 que la contenga.

c) ¿Verifica el polinomio $P(x)$ las condiciones de Fourier para la convergencia del método de Newton en el intervalo $(1'2,\ 1'3)$?

d) Si tomamos como valor inicial $x_0 = 1'3$, calcular el valor que se obtiene para x_1 dando una cota del error.

2. Sistemas de ecuaciones lineales

2.1 Normas vectoriales y matriciales

Sea E un espacio vectorial definido sobre un cuerpo $\mathbf{K}$. Se define una *norma* como una aplicación, que denotaremos por $\|\ \|$, de E en $\mathbf{R}$ que verifica las siguientes propiedades:

1.- $\|x\| \geq 0 \ \ \forall x \in E$ siendo $\|x\| = 0 \Longleftrightarrow x = 0$. (Definida positiva).

2.- $\|c\,x\| = |c|\,\|x\| \quad \forall c \in \mathbf{K},\ \forall x \in E$. (Homogeneidad).

3.- $\|x + y\| \leq \|x\| + \|y\| \quad \forall x, y \in E$. (Desigualdad triangular).

Un espacio E en el que hemos definido una norma recibe el nombre de *espacio normado*.

Es frecuente que en el espacio E se haya definido también el producto de dos elementos. En este caso, si se verifica que

$$\|x \cdot y\| \leq \|x\|\,\|y\|$$

se dice que la norma es *multiplicativa*. Esta propiedad de las normas es fundamental cuando trabajamos en el conjunto $\mathbf{C}^{n\times n}$ de las matrices cuadradas de orden n. Sin embargo no tiene mucha importancia cuando se trabaja en el espacio $\mathcal{C}[a, b]$ de las funciones continuas en $[a, b]$.

2.1.1 Normas vectoriales

Sea E un espacio normado de dimensión n y sea $\mathcal{B} = \{u_1, u_2, \ldots, u_n\}$ una base suya. Cualquier vector $x \in E$ puede ser expresado de forma única en función

de los vectores de la base $\mathcal{B}$.

$$x = \sum_{i=1}^{n} x_i u_i$$

donde los escalares $(x_1, x_2, \ldots, x_n)$ se conocen como *coordenadas* del vector x respecto de la base $\mathcal{B}$.

Utilizando esta notación, son ejemplos de normas los siguientes:

- **Norma-1** $\|x\|_1 = \sum_{i=1}^{n} |x_i|$
- **Norma euclídea** $\|x\|_2 = \sqrt{\sum_{i=1}^{n} |x_i|^2}$
- **Norma infinito** $\|x\|_\infty = \max_i |x_i|$

Ejemplo 2.1 Para el vector $x = \begin{pmatrix} 2 & 3 & 0 & -12 \end{pmatrix}^T$ se tiene que

$$\|x\|_1 = \sum_{i=1}^{n} |x_i| = 2 + 3 + 0 + 12 = 17$$

$$\|x\|_2 = \sqrt{\sum_{i=1}^{n} |x_i|^2} = \sqrt{2^2 + 3^2 + 0^2 + 12^2} = \sqrt{169} = 13$$

$$\|x\|_\infty = \max_i |x_i| = 12$$

□

2.1.2 Distancia inducida por una norma

Dado un espacio vectorial E, se define una *distancia* como una aplicación $d : E \times E \to \mathbf{R}$ cumpliendo que:

- $d(x, y) \geq 0 \quad \forall x, y \in E$ siendo $d(x, y) = 0 \iff x = y$.
- $d(x, y) = d(y, x) \quad \forall x, y \in E$.
- $d(x, y) \leq d(x, z) + d(z, y) \quad \forall x, y, z \in E$.

Si $\Big(E, \| \ \|\Big)$ es un espacio normado, la norma $\| \ \|$ induce una distancia en E que se conoce como *distancia inducida* por la norma $\| \ \|$ y viene definida por:

$$d(x, y) = \|x - y\|$$

Veamos que, en efecto, se trata de una distancia:

- $d(x, y) \geq 0$ por tratarse de una norma, y además:

  $d(x, y) = 0 \Longleftrightarrow \|x - y\| = 0 \Longleftrightarrow x - y = 0 \Longleftrightarrow x = y.$

- $d(x, y) = \|x - y\| = \|-1(y - x)\| = |-1| \, \|y - x\| = \|y - x\| = d(y, x).$

- $d(x, y) = \|x - y\| = \|x - z + z - y\| \leq \|x - z\| + \|z - y\| =$
  $= d(x, z) + d(z, y).$

2.1.3 Convergencia en espacios normados

Una sucesión de vectores $v_1, v_2, \ldots$ de un espacio vectorial normado $(V, \| \ \|)$ se dice que es *convergente* a un vector v si

$$\lim_{k \to \infty} \|v_k - v\| = 0$$

Esta definición coincide con la idea intuitiva de que la distancia de los vectores de la sucesión al vector límite v tiende a cero a medida que se avanza en la sucesión.

Teorema 2.1 *Para un espacio vectorial normado de dimensión finita, el concepto de convergencia es independiente de la norma utilizada.*

2.1.4 Normas matriciales

Dada una matriz A y un vector x, consideremos el vector transformado Ax. La relación existente entre la norma del vector transformado y la del vector original va a depender de la matriz A. El mayor de los cocientes entre dichas normas, para todos los vectores del espacio, es lo que vamos a definir como norma de la matriz A, de tal forma que de la propia definición se deduce que

$$\|Ax\| \leq \|A\| \|x\|$$

cualquiera que sea el vector x del espacio. (Obsérvese que no es lo mismo que la propiedad multiplicativa de una norma, ya que aquí se están utilizando dos normas diferentes, una de matriz y otra de vector).

Así pues, definiremos

$$\|A\| = \max_{x \in V - \{0\}} \frac{\|Ax\|}{\|x\|} = \max\{\|Ax\| \ : \ \|x\| = 1\}$$

de tal forma que a cada norma vectorial se le asociará, de forma natural, una norma matricial.

- **Norma-1**
  Si utilizamos la norma-1 de vector obtendremos

  $$\|A\|_1 = \max\{\|Ax\|_1 \ : \ \|x\|_1 = 1\}.$$

  Dado que $Ax = y \implies y_i = \sum_{k=1}^{n} a_{ik}x_k$ se tiene que

  $$\|A\|_1 = \max\{\sum_{i=1}^{n}\sum_{k=1}^{n} |a_{ik}x_k| \ : \ \|x\|_1 = 1\}.$$

  Por último, si descargamos todo el peso sobre una coordenada, es decir, si tomamos un vector de la base canónica, obtenemos que

  $$\|A\|_1 = \max_j \sum_{i=1}^{n} |a_{ij}|.$$

- **Norma euclídea**
  Utilizando la norma euclídea de vector se tendrá que

  $$\|A\|_2 = \max\{\sqrt{x^*A^*Ax} \ : \ \sqrt{x^*x} = 1\}$$

  Descargando ahora el peso en los autovectores de la matriz A^*A obtenemos que

  $$\|A\|_2 = \max_i\{\sqrt{x^*\lambda_i x} \ : \ \sqrt{x^*x} = 1\} = \max_i \sqrt{\lambda_i} = \max_i \sigma_i$$

  donde σ_i representa a los valores singulares de la matriz A, es decir, las raíces cuadradas positivas de los autovalores de la matriz A^*A.

- **Norma infinito**
  Utilizando ahora la norma infinito de vector se tiene que

$$\|A\|_\infty = \max\{\sum_{j=1}^{n} |a_{ij}x_j| \ : \ \|x\|_\infty = 1\}.$$

  Como ahora se dará el máximo en un vector que tenga todas sus coordenadas iguales a 1, se tiene que

$$\|A\|_\infty = \max_i \sum_{j=1}^{n} |a_{ij}|.$$

Tenemos pues, que las normas asociadas (algunas veces llamadas *subordinadas*) a las normas de vector estudiadas anteriormente son:

$$\textbf{Norma-1} \qquad \|A\|_1 = \max_{\|x\|_1=1} \|Ax\|_1 = \max_j \sum_{i=1}^{n} |a_{ij}|.$$

$$\textbf{Norma euclídea} \quad \|A\|_2 = \max_{\|x\|_2=1} \|Ax\|_2 = \max_i \sigma_i.$$

$$\textbf{Norma infinito} \quad \|A\|_\infty = \max_{\|x\|_\infty=1} \|Ax\|_\infty = \max_i \sum_{j=1}^{n} |a_{ij}|.$$

Si consideramos la matriz como un vector de $m \times n$ coordenadas, podemos definir (de manera análoga a la norma euclídea de vector) la denominada

$$\textbf{Norma de Frobenius} \quad \|A\|_F = \sqrt{\sum_{i,j} |a_{ij}|^2} = \sqrt{\text{tr } A^*A}$$

Ejemplo 2.2 Para la matriz $A = \begin{pmatrix} 1 & 1 & 2 \\ 2 & -1 & 1 \\ 1 & 0 & 1 \end{pmatrix}$ se tiene:

$$\|A\|_1 = \max_j \sum_{i=1}^{n} |a_{ij}| = \max\{(1+2+1), (1+|-1|+0), (2+1+1)\} = 4$$

El polinomio característico de la matriz $A^TA = \begin{pmatrix} 6 & -1 & 5 \\ -1 & 2 & 1 \\ 5 & 1 & 6 \end{pmatrix}$ es $p(\lambda) = \lambda^3 - 14\lambda^2 + 33\lambda$ cuyas raíces son 0, 3 y 11, por lo que los valores singulares de la matriz A son 0, $\sqrt{3}$ y $\sqrt{11}$ y, por tanto,

$\|A\|_2 = \text{máx}_i\, \sigma_i = \sqrt{11}.$

$\|A\|_\infty = \max\limits_i \sum\limits_{j=1}^{n} |a_{ij}| = \text{máx}\{(1+1+2), (2+|-1|+1), (1+0+1)\} = 4$

$\|A\|_F = \sqrt{\text{tr } A^T A} = \sqrt{6+2+6} = \sqrt{14}.$ □

2.1.5 Transformaciones unitarias

Una matriz $U \in \mathbf{C}^{n\times n}$ se dice *unitaria* si

$$U^*U = UU^* = I$$

es decir, si $U^* = U^{-1}$.

Proposición 2.2 *La norma de Frobenius y la norma euclídea de vector son invariantes mediante transformaciones unitarias.*

Demostración.

- Para la norma de Frobenius de matrices.
  $\|UA\|_F = \sqrt{\text{tr } [(UA)^*(UA)]} = \sqrt{\text{tr } (A^*U^*UA)} = \sqrt{\text{tr } (A^*A)} = \|A\|_F.$
  $\|AU\|_F = \sqrt{\text{tr } [(AU)(AU)^*]} = \sqrt{\text{tr } (AUU^*A^*)} = \sqrt{\text{tr } (AA^*)} =$
  $= \|A^*\|_F = \|A\|_F.$
- Para la norma euclídea de vector.
  $\|Ux\|_2 = \sqrt{(Ux)^*(Ux)} = \sqrt{x^*U^*Ux} = \sqrt{x^*x} = \|x\|_2.$ ■

Lema 2.3 *Las matrices A, A^*, AU y UA, donde U es una matriz unitaria, poseen los mismos valores singulares.*

Demostración. Veamos, en primer lugar que las matrices A^*A y AA^* poseen los mismos autovalores. En efecto:

Sea λ un autovalor de A^*A y x un autovector asociado a λ. Se tiene entonces que $A^*Ax = \lambda x$. Multiplicando, por la izquierda por la matriz A obtenemos que $AA^*Ax = \lambda Ax$ y llamando $y = Ax$ tenemos que $AA^*y = \lambda y$, por lo que λ es un autovalor de AA^* y el vector y es un autovector asociado a λ para la matriz AA^*. Así pues, cualquier autovalor de A^*A lo es también de AA^*.

Razonando de igual forma se obtiene que cualquier autovalor de AA^* lo es también de A^*A, por lo que ambas matrices poseen los mismos autovalores.

a) Como los valores singulares de la matriz A son las raíces cuadradas positivas de los autovalores de A^*A y los de A^* las raíces cuadradas positivas de los autovalores de AA^* (los mismos que los de A^*A), las matrices A y A^* poseen los mismos valores singulares.

b) Los valores singulares de UA son las raíces cuadradas positivas de los autovalores de $(UA)^*(UA) = A^*U^*UA = A^*A$ que eran precisamente los valores singulares de la matriz A.

c) Los valores singulares de AU son (por el primer apartado) los mismos que los de $(AU)^* = U^*A^*$ es decir, las raíces cuadradas positivas de los autovalores de $(U^*A^*)^*(U^*A^*) = AUU^*A^* = AA^*$ cuyos autovalores son los mismos que los de A^*A, por lo que AU tiene también los mismos valores singulares que la matriz A. ■

Proposición 2.4 *La norma euclídea es invariante mediante transformaciones de semejanza unitarias.*

Demostración. Dado que $\|A\|_2 = \max\limits_{x\in V-\{0\}} \dfrac{\|Ax\|_2}{\|x\|_2}$ si U es unitaria, se tiene que:

$$\|U\|_2 = \max_{x\in V-\{0\}} \frac{\|Ux\|_2}{\|x\|_2} = \max_{x\in V-\{0\}} \frac{\|x\|_2}{\|x\|_2} = 1.$$

Es decir, si U es unitaria $\|U\|_2 = \|U^*\|_2 = 1$.

Si $B = U^*AU$ tenemos: $\|B\|_2 \leq \|U^*\|_2 \|A\|_2 \|U\|_2 = \|A\|_2$

Como $A = UBU^*$, es: $\|A\|_2 \leq \|U\|_2 \|B\|_2 \|U^*\|_2 = \|B\|_2$

De ambas desigualdades se deduce que $\|B\|_2 = \|A\|_2$. ■

2.1.6 Radio espectral

Se define el *radio espectral* de una matriz A, y se denota por $\rho(A)$ como el máximo de los módulos de los autovalores de la referida matriz.

$$\rho(A) = \max_i |\lambda_i|$$

Geométricamente representa el radio del círculo mínimo que contiene a todos los autovalores de la matriz.

Teorema 2.5 *El radio espectral de una matriz es una cota inferior de todas las normas multiplicativas de dicha matriz.*

Demostración. Sean $\{\lambda_1, \lambda_2, \ldots, \lambda_r\}$ los autovalores de la matriz A y sean $\{x_1, x_2, \ldots, x_r\}$ autovectores asociados a dichos autovalores.

$$\|Ax_i\| = \|\lambda_i x_i\| = |\lambda_i| \, \|x_i\| \, .$$

Por otra parte sabemos que $\|Ax_i\| \leq \|A\| \, \|x_i\|$. Por tanto:
$|\lambda_i| \, \|x_i\| \leq \|A\| \, \|x_i\|$ siendo $\|x_i\| \neq 0$ por tratarse de autovectores.

Se obtiene entonces que $|\lambda_i| \leq \|A\|$ para cualquiera que sea $i = 1, 2, \ldots, r$, de donde $\max\limits_i |\lambda_i| \leq \|A\|$, es decir: $\rho(A) \leq \|A\|$. ■

Definición 2.1 Sea $A \in \mathbf{K}^{n \times n}$. Se definen los *círculos de Gerschgorin* como los conjuntos

$$\left.\begin{array}{l} C_1 = \{z : \ |z - a_{11}| \leq r_1\} \\ C_2 = \{z : \ |z - a_{22}| \leq r_2\} \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ C_n = \{z : \ |z - a_{nn}| \leq r_n\} \end{array}\right\} \quad \text{en donde} \quad r_i = \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}| \quad \text{para } 1 \leq i \leq n$$

Teorema 2.6 *Sean $C_1, C_2, \ldots, C_n$ los círculos de Gerschgorin de una matriz A. Si λ es un autovalor de A, se verifica que $\lambda \in \bigcup\limits_{i=1}^{n} C_i$.*

Demostración. Si $\lambda \notin \bigcup\limits_{i=1}^{n} C_i$ es porque $\lambda \notin C_i$ para ningún $i = 1, 2, \ldots, n$, por lo que

$$\begin{array}{c} |\lambda - a_{11}| > r_1 \\ |\lambda - a_{22}| > r_2 \\ \ldots\ldots\ldots\ldots \\ \ldots\ldots\ldots\ldots \\ |\lambda - a_{nn}| > r_n \end{array}$$

y, por tanto, la matriz

$$\begin{pmatrix} \lambda - a_{11} & -a_{12} & \cdots & -a_{1n} \\ -a_{21} & \lambda - a_{22} & \cdots & -a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ -a_{n1} & -a_{n2} & \cdots & \lambda - a_{nn} \end{pmatrix} = \lambda I - A$$

es una matriz de Hadamard, por lo que $\det(\lambda I - A) \neq 0$ y, por tanto, λ no puede ser un autovalor de la matriz A en contra de la hipótesis. ■

Obsérvese que como $\det A = \det A^T$ se verifica que

$$\det(\lambda I - A) = \det(\lambda I - A)^T = \det(\lambda I - A^T)$$

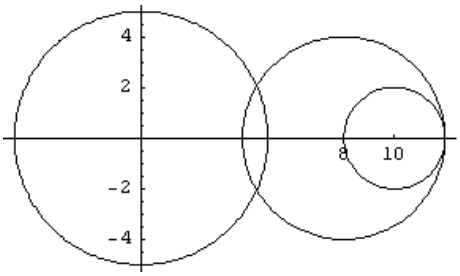
es decir, A y A^T tienen el mismo polinomio característico y, por tanto, los mismos autovalores.

Ello nos lleva a que los círculos de Gerschgorin obtenidos por columnas determinan también el dominio de existencia de los autovalores de la matriz A, es más, la intersección de los círculos obtenidos por filas y los obtenidos por columnas contienen a todos los autovalores de la matriz A.

Teorema 2.7 *Si un conjunto de k discos de Gerschgorin de una matriz A constituyen un dominio conexo aislado de los otros discos, existen exactamente k autovalores de la matriz A en dicho dominio.*

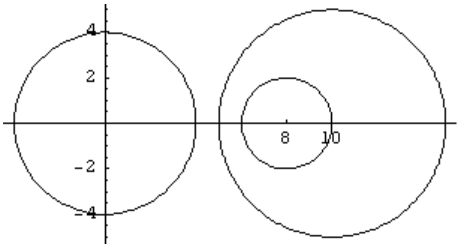
Ejemplo 2.3 Dada la matriz $A = \begin{pmatrix} 0 & 2 & 3 \\ 2 & 8 & 2 \\ 2 & 0 & 10 \end{pmatrix}$, sus círculos de Gerschgorin obtenidos por filas son

$$C_1 = \{z : \ |z - 0| \leq 5\}$$
$$C_2 = \{z : \ |z - 8| \leq 4\}$$
$$C_3 = \{z : \ |z - 10| \leq 2\}$$



Mientras que los obtenidos por columnas son

$$C_1 = \{z : \ |z - 0| \leq 4\}$$
$$C_2 = \{z : \ |z - 8| \leq 2\}$$
$$C_3 = \{z : \ |z - 10| \leq 5\}$$



Estos últimos nos dicen que la matriz tiene, al menos, un autovalor real (hay un disco que sólo contiene a un autovalor y éste debe ser real ya que de ser

complejo el conjugado también sería autovalor de A y al tener el mismo módulo se encontraría en el mismo disco contra la hipótesis de que en dicho disco sólo se encuentra un autovalor), es decir los discos obtenidos por filas no nos dan información sobre el autovalor real, mientras que los obtenidos por columnas nos dicen que existe un autovalor real en el intervalo (-4,4). □

2.2 Sistemas de ecuaciones lineales

En el capítulo anterior se estudiaron métodos iterados para la resolución de ecuaciones no lineales. Dichos métodos se basaban en el teorema del punto fijo y consistían en expresar la ecuación en la forma $x = \varphi(x)$ exigiendo que $\varphi'(x) \leq q < 1$ para cualquier x del intervalo en el cual se trata de buscar la solución.

Para los sistemas de ecuaciones lineales, de la forma $Ax = b$, trataremos de buscar métodos iterados de una forma análoga a como se hizo en el caso de las ecuaciones, es decir, transformando el sistema en otro equivalente de la forma $x = F(x)$ donde $F(x) = Mx + N$. Evidentemente habrá que exigir algunas condiciones a la matriz M para que el método sea convergente (al igual que se exigía que $\varphi'(x) \leq q < 1$ en el caso de las ecuaciones) y estas condiciones se basan en los conceptos de *normas vectoriales* y *matriciales*.

Dada una aplicación $f : \mathbf{R}^m \to \mathbf{R}^n$ y un vector $b \in \mathbf{R}^n$, resolver el sistema de ecuaciones $f(x) = b$ no es más que buscar el conjunto de vectores de $\mathbf{R}^m$ cuya imagen mediante f es el vector b, es decir, buscar la imagen inversa de b mediante f.

Un sistema de ecuaciones se dice *lineal en su componente k-ésima* si verifica que

$$f(x_1, \ldots, x_{k-1}, \alpha x_k^1 + \beta x_k^2, x_{k+1}, \ldots, x_m) = \alpha f(x_1, \ldots, x_{k-1}, x_k^1, x_{k+1}, \ldots, x_m) + \\ + \beta f(x_1, \ldots, x_{k-1}, x_k^2, x_{k+1}, \ldots, x_m)$$

Diremos que un sistema es *lineal* si lo es en todas sus componentes, pudiéndose, en este caso, escribir de la forma $Ax = b$.

Si la aplicación f se define de $\mathbf{C}^m$ en $\mathbf{C}^n$ resulta un sistema complejo que puede ser transformado en otro sistema real. Así, por ejemplo, si el sistema es lineal, es decir, de la forma $Mz = k$ con $M \in \mathbf{C}^{m\times n}$, $x \in \mathbf{C}^{n\times 1}$ y $k \in \mathbf{C}^{m\times 1}$, podemos descomponer la matriz M en suma de otras dos de la forma $M = A + iB$ con $A, B \in \mathbf{R}^{m\times n}$ y análogamente $z = x + iy$ con $x, y \in \mathbf{R}^{n\times 1}$

$k = k_1 + ik_2$ con $k_1, k_2 \in \mathbf{R}^{m\times 1}$, por lo que $(A+iB)(x+iy) = k_1 + ik_2$ es decir

$$\left\{ \begin{array}{l} Ax - By = k_1 \\ Bx + Ay = k_2 \end{array} \right. \Longrightarrow \left(\begin{array}{cc} A & -B \\ B & A \end{array} \right) \left(\begin{array}{c} x \\ y \end{array} \right) = \left(\begin{array}{c} k_1 \\ k_2 \end{array} \right)$$

sistema real de $2m$ ecuaciones con $2n$ incógnitas. Es por ello, que centraremos nuestro estudio en los sistemas reales.

Podemos clasificar los sistemas de ecuaciones lineales atendiendo a

a) **Su tamaño**

a.1) Pequeños: $n \leq 300$ donde n representa el número de ecuaciones.

a.2) Grandes: $n > 300$

(Esta clasificación corresponde al error de redondeo)

b) **Su estructura**

b.1) Si la matriz posee pocos elementos nulos diremos que se trata de un sistema *lleno*.

b.2) Si, por el contrario, la matriz contiene muchos elementos nulos, diremos que la matriz y, por tanto, que el sistema es *disperso* o *sparce*. Matrices de este tipo son las denominadas

- Tridiagonales: $\left(\begin{array}{cccc} a_{11} & a_{12} & 0 & 0 \\ a_{21} & a_{22} & a_{23} & 0 \\ 0 & a_{32} & a_{33} & a_{34} \\ 0 & 0 & a_{43} & a_{44} \end{array} \right)$
- Triangulares superiores: $\left(\begin{array}{cccc} a_{11} & a_{12} & a_{13} & a_{14} \\ 0 & a_{22} & a_{23} & a_{24} \\ 0 & 0 & a_{33} & a_{34} \\ 0 & 0 & 0 & a_{44} \end{array} \right)$
- Triangulares inferiores: $\left(\begin{array}{cccc} a_{11} & 0 & 0 & 0 \\ a_{12} & a_{22} & 0 & 0 \\ a_{31} & a_{32} & a_{33} & 0 \\ a_{41} & a_{42} & a_{43} & a_{44} \end{array} \right)$

En cuanto a los métodos de resolución de sistemas de ecuaciones lineales, podemos clasificarlos en

a) **Métodos directos**

b) **Métodos iterados**

Se denominan *métodos directos* a aquellos métodos que resuelven un sistema de ecuaciones lineales en un número finito de pasos. Se utilizan para resolver sistemas pequeños.

Los denominados *métodos iterados* crean una sucesión de vectores que convergen a la solución del sistema. Estos métodos se utilizan para la resolución de sistemas grandes, ya que al realizar un gran número de operaciones los errores de redondeo pueden hacer inestable al proceso, es decir, pueden alterar considerablemente la solución del sistema.

2.2.1 Número de condición

Un sistema de ecuaciones lineales $Ax = b$ se dice *bien condicionado* cuando los errores cometidos en los elementos de la matriz A y del vector b producen en la solución un error del mismo orden, mientras que diremos que el sistema está *mal condicionado* si el error que producen en la solución del sistema es de orden superior al de los datos. Es decir:

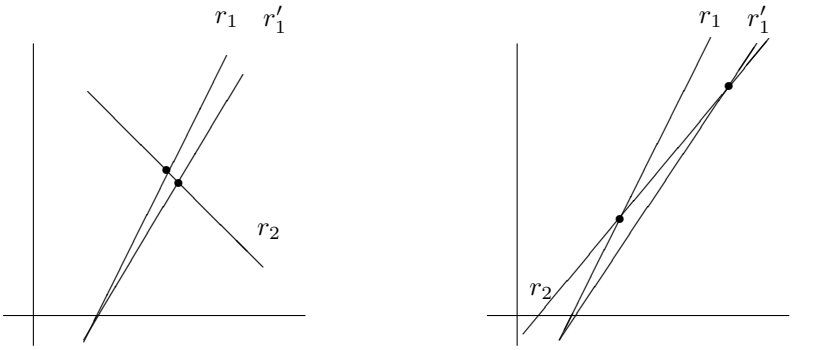
$$\begin{array}{c} \|A - \overline{A}\| < \varepsilon \\ \|b - \overline{b}\| < \varepsilon \end{array} \Longrightarrow \begin{cases} \|x - \overline{x}\| \simeq \varepsilon & \text{sistema bien condicionado} \\ \|x - \overline{x}\| \gg \varepsilon & \text{sistema mal condicionado} \end{cases}$$

Consideremos el sistema cuadrado $Ax = b$ con A regular, es decir, un *sistema compatible determinado*. En la práctica, los elementos de A y de b no suelen ser exactos bien por que procedan de cálculos anteriores, o bien porque sean irracionales, racionales periódicos, etc. Es decir, debemos resolver un sistema aproximado cuya solución puede diferir poco o mucho de la verdadera solución del sistema.

Así, por ejemplo, en un sistema de orden dos, la solución representa el punto de intersección de dos rectas en el plano. Un pequeño error en la pendiente de una de ellas puede hacer que dicho punto de corte se desplace sólo un poco o una distancia considerable (véase la Figura 2.1), lo que nos dice que el sistema está bien o mal condicionado, respectivamente.

Podemos ver que el sistema está mal condicionado cuando las pendientes de las dos rectas son muy similares y que mientras más ortogonales sean las rectas, mejor condicionado estará el sistema.

Se puede observar entonces que si, en un sistema mal condicionado, sustituimos una de las ecuaciones por una combinación lineal de las dos, podemos hacer que el sistema resultante esté bien condicionado.



Sistema bien condicionado Sistema mal condicionado

Figura 2.1: Condicionamiento de un sistema.

Ejemplo 2.4 Si consideramos el sistema

$$\begin{aligned} 3x + 4y &= 7 \\ 3x + 4.00001y &= 7.00001 \end{aligned} \qquad \text{de solución} \qquad \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

y cometemos un pequeño error en los datos, podemos obtener el sistema

$$\begin{aligned} 3x + 4y &= 7 \\ 3x + 3.99999y &= 7.00004 \end{aligned} \qquad \text{de solución} \qquad \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 7.\bar{6} \\ -4 \end{pmatrix}$$

o bien este otro

$$\begin{aligned} 3x + 4y &= 7 \\ 3x + 3.99999y &= 7.000055 \end{aligned} \qquad \text{de solución} \qquad \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 9.\bar{6} \\ -5.5 \end{pmatrix}$$

lo que nos dice que estamos ante un sistema mal condicionado.

Si sustituimos la segunda ecuación por la que resulta de sumarle la primera multiplicada por $-1'0000016$ (la ecuación resultante se multiplica por 10^6 y se divide por $-1'2$) nos queda el sistema

$$\begin{aligned} 3x + 4y &= 7 \\ 4x - 3y &= 1 \end{aligned} \qquad \text{de solución} \qquad \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

siendo éste un sistema bien condicionado. □

El estudio del condicionamiento de un sistema se realiza a través del denominado número de condición que estudiamos a continuación.

Sea A una matriz cuadrada y regular. Se define el *número de condición* de la matriz A y se denota por $\kappa(A)$ como

$$\kappa(A) = \|A\| \cdot \left\|A^{-1}\right\|$$

donde la norma utilizada ha de ser una norma multiplicativa. Este número nos permite conocer el condicionamiento del sistema $Ax = b$.

Dado que en la práctica el cálculo de la matriz inversa A^{-1} presenta grandes dificultades lo que se hace es buscar una cota del número de condición.

$$\kappa(A) = \|A\| \cdot \left\|A^{-1}\right\| < \|A\| \cdot k$$

siendo k una cota de la norma de la matriz inversa.

Si $\|I - A\| < 1$ entonces $\|A^{-1}\| \leq \dfrac{\|I\|}{1 - \|I - A\|}$. En efecto:

$A \cdot A^{-1} = I \Longrightarrow [I - (I - A)]A^{-1} = I \Longrightarrow$

$A^{-1} - (I - A)A^{-1} = I \Longrightarrow A^{-1} = I + (I - A)A^{-1} \Longrightarrow$

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

$\|A^{-1}\| - \|I - A\| \, \|A^{-1}\| \leq \|I\| \Longrightarrow (1 - \|I - A\|) \, \|A^{-1}\| \leq \|I\| \Longrightarrow$

$$\left\|A^{-1}\right\| \leq \frac{\|I\|}{1 - \|I - A\|}$$

Es decir:

$$\kappa(A) \leq \|A\| \cdot k \qquad \text{con} \qquad k = \frac{\|I\|}{1 - \|I - A\|}$$

Debemos tener cuidado con esta acotación ya que si tenemos una matriz casi regular, es decir, con $\det(A) \simeq 0$, quiere decir que tiene un autovalor próximo a cero, por lo que la matriz $I - A$ tiene un autovalor próximo a 1 y será el mayor de todos. En este caso $\|I - A\| \simeq 1$, por lo que $k \to \infty$ y daría lugar a un falso condicionamiento, ya que A no tiene que estar, necesariamente, mal condicionada.

Ejemplo 2.5 Para estudiar el condicionamiento del sistema

$$\begin{aligned} 3x + 4y &= 7 \\ 3x + 4.00001y &= 7.00001 \end{aligned}$$

Se tiene que

$$A = \begin{pmatrix} 3 & 4 \\ 3 & 4.00001 \end{pmatrix} \Longrightarrow \det(A) = 0.00003$$

$$A^{-1} = \frac{1}{0.00003}\begin{pmatrix} 4.00001 & -4 \\ -3 & 3 \end{pmatrix}$$

Utilizando la norma infinito $\|A\|_\infty = \max\limits_i \sum\limits_{j=1}^{n} |a_{ij}|$ se tiene que

$$\left.\begin{array}{l} \|A\|_\infty = 7.00001 \\ \|A^{-1}\|_\infty = \dfrac{8.00001}{0.00003} \end{array}\right\} \Longrightarrow \kappa_\infty(A) \simeq \frac{56}{3}\cdot 10^5 > 1.8\cdot 10^6$$

Se trata pues, de un sistema mal condicionado.

Si utilizamos la norma-1 $\|A\|_1 = \max\limits_j \sum\limits_{i=1}^{n} |a_{ij}|$ obtenemos:

$$\left.\begin{array}{l} \|A\|_1 = 8.00001 \\ \|A^{-1}\|_1 = \dfrac{7.00001}{0.00003} \end{array}\right\} \Longrightarrow \kappa_1(A) \simeq \frac{56}{3}\cdot 10^5 > 1.8\cdot 10^6$$

obteniéndose, también, que se trata de un sistema mal condicionado. □

Propiedades del número de condición $\kappa(A)$.

a) Dado que $\|x\| = \|Ix\| \leq \|I\|\|x\|$, se verifica que $\|I\| \geq 1$, cualquiera que sea la norma utilizada. Como, por otra parte $AA^{-1} = I$ se tiene que

$$1 \leq \|I\| = \|AA^{-1}\| \leq \|A\|\|A^{-1}\| = \kappa(A)$$

es decir $\kappa(A) \geq 1$ cualquiera que sea la matriz cuadrada y regular A.

b) Si $B = zA$, con $z \in \mathbf{C}$ no nulo, se verifica que $\kappa(B) = \kappa(A)$. En efecto:

$$\kappa(B) = \|B\|\left\|B^{-1}\right\| = \|zA\|\left\|\frac{1}{z}A^{-1}\right\| = |z|\,\|A\|\,\frac{\|A^{-1}\|}{|z|} = \|A\|\left\|A^{-1}\right\| = \kappa(A).$$

Dado que $\det(B) = z^n \det(A)$, donde n representa el orden de la matriz A, y $\kappa(B) = \kappa(A)$ se ve que el condicionamiento de una matriz no depende del valor de su determinante.

c) Utilizando la norma euclídea $\| \ \|_2$ se tiene que $\kappa_2(A) = \dfrac{\sigma_n}{\sigma_1}$ donde σ_1 y σ_n representan, respectivamente, al menor y al mayor de los valores singulares de la matriz A.

En efecto: sabemos que los valores singulares σ_i de la matriz A son las raíces cuadradas positivas de los autovalores de la matriz A^*A.

$$\sigma_i = \sqrt{\lambda_i(A^*A)}$$

Si suponemos $\sigma_1 \leq \sigma_2 \leq \cdots \leq \sigma_n$ se tiene que

$$\|A\|_2 = \sqrt{\max_i \lambda_i(A^*A)} = \sigma_n$$

$$\|A^{-1}\|_2 = \sqrt{\max_i \lambda_i\left((A^{-1})^*A^{-1}\right)} = \sqrt{\max_i \lambda_i\left((A^*)^{-1}A^{-1}\right)} =$$

$$= \sqrt{\max_i \lambda_i(AA^*)^{-1}} = \sqrt{\max_i \frac{1}{\lambda_i(AA^*)}} = \sqrt{\frac{1}{\min_i \lambda_i(AA^*)}} =$$

$$= \sqrt{\frac{1}{\min_i \lambda_i(A^*A)}} = \sqrt{\frac{1}{\min_i \sigma_i^2}} \Longrightarrow$$

$$\left\|A^{-1}\right\|_2 = \frac{1}{\sigma_1}$$

Podemos concluir, por tanto, que

$$\left.\begin{array}{r} \|A\|_2 = \sigma_n \\ \|A^{-1}\|_2 = \dfrac{1}{\sigma_1} \end{array}\right\} \Longrightarrow \kappa_2(A) = \frac{\sigma_n}{\sigma_1}$$

En cuanto a su relación con los números de condición obtenidos con otras normas de matriz se tiene que:

$\|A\|_2 \leq \|A\|_F \leq \sqrt{n}\,\|A\|_2 \Longrightarrow \|A^{-1}\|_2 \leq \|A^{-1}\|_F \leq \sqrt{n}\,\|A^{-1}\|_2$

$\kappa_2(A) = \|A\|_2\,\|A^{-1}\|_2 \leq \|A\|_F\,\|A^{-1}\|_F = \kappa(A)_F$

$\kappa_F(A) = \|A\|_F\,\|A^{-1}\|_F \leq \sqrt{n}\sqrt{n}\,\|A\|_2\,\|A^{-1}\|_2 = n\kappa_2(A) \Longrightarrow$

$$\kappa_2(A) \leq \kappa_F(A) \leq n\kappa_2(A)$$

Además: $\begin{cases} \|A\|_2 \leq \sqrt{\|A\|_1\,\|A\|_\infty} \\ \|A^{-1}\|_2 \leq \sqrt{\|A^{-1}\|_1\,\|A^{-1}\|_\infty} \end{cases} \Rightarrow \kappa_2(A) \leq \sqrt{\kappa_1(A)\kappa_\infty(A)}$

d) Una condición necesaria y suficiente para que $\kappa_2(A) = 1$ es que $A = zU$ siendo $z \in \mathbf{C}$ (no nulo) y U una matriz unitaria ($UU^* = U^*U = I$).

$\Leftarrow$) $A = zU \Longrightarrow \kappa_2(A) = 1$. En efecto:
$A = zU \Longrightarrow A^*A = \overline{z}U^*zU = |z|^2 U^*U = |z|^2 I \Longrightarrow$
$\lambda_i(A^*A) = |z|^2$ cualquiera que sea $i = 1, 2, \ldots, n$ y, por tanto,

$$\sigma_1 = \sigma_2 = \cdots = \sigma_n = |z|$$

por lo que

$$\kappa_2(A) = \frac{\sigma_n}{\sigma_1} = 1$$

$\Rightarrow$) $\kappa_2(A) = 1 \Longrightarrow A = zU$.
En efecto: sabemos que si A es diagonalizable existe una matriz regular R tal que $R^{-1}AR = D$ con $D = \text{diag}(\lambda_i)$ (R es la matriz de paso cuyas columnas son los autovectores correspondientes a los autovalores λ_i). Por otra parte sabemos que toda matriz hermítica es diagonalizable mediante una matriz de paso unitaria.
Como la matriz A^*A es hermítica existe una matriz unitaria R tal que

$$R^*A^*AR = \begin{pmatrix} \sigma_1^2 & & & \\ & \sigma_2^2 & & \\ & & \ddots & \\ & & & \sigma_n^2 \end{pmatrix}$$

Como $\kappa_2(A) = \dfrac{\sigma_n}{\sigma_1} = 1 \Longrightarrow \sigma_1 = \sigma_2 = \cdots = \sigma_n = \sigma$, por lo que

$$R^*A^*AR = \sigma^2 I$$

Entonces

$$A^*A = R(\sigma^2 I)R^* = \sigma^2(RIR^*) = \sigma^2 I$$

de donde

$$\left(\frac{1}{\sigma}A^*\right)\left(\frac{1}{\sigma}A\right) = I$$

Llamando $U = \dfrac{1}{\sigma}A$ se tiene que $U^* = \dfrac{1}{\sigma}A^*$, ya que $\sigma \in \mathbf{R} \Rightarrow \overline{\sigma} = \sigma$.
Se tiene entonces que $A = \sigma U$ con $U^*U = \left(\dfrac{1}{\sigma}A^*\right)\left(\dfrac{1}{\sigma}A\right) = I$, es decir, con U unitaria. ■

Los sistemas mejor condicionados son aquellos que tienen sus filas o columnas ortogonales y mientras mayor sea la dependencia lineal existente entres ellas peor es el condicionamiento del sistema.

Al ser $\kappa(AU) = \kappa(UA) = \kappa(A)$ trataremos de buscar métodos de resolución de sistemas de ecuaciones lineales que trabajen con matrices unitarias que no empeoren el condicionamiento del sistema como lo hace, por ejemplo, el método de Gauss basado en la factorización LU. Sin embargo, dado que ha sido estudiado en la asignatura de Álgebra Lineal, comenzaremos estudiando dicho método aunque pueda alterarnos el condicionamiento del problema.

Empezaremos estudiando pues, como *métodos directos*, los basados en la factorización LU y la de *Cholesky*.

2.3 Factorización LU

Al aplicar el método de Gauss al sistema $Ax = b$ realizamos transformaciones elementales para conseguir triangularizar la matriz del sistema. Si este proceso puede realizarse sin intercambios de filas, la matriz triangular superior U obtenida viene determinada por el producto de un número finito de transformaciones fila $F_k F_{k-1} \cdots F_1$ aplicadas a la matriz A. Llamando $L^{-1} = F_k F_{k-1} \cdots F_1$ (ya que el determinante de una transformación fila es ± 1 y, por tanto, su producto es inversible) se tiene que $L^{-1}A = U$, o lo que es lo mismo, $A = LU$. Además, la matriz L es una triangular inferior con *unos* en la diagonal.

Esta factorización es única ya que de existir otra tal que $A = L'U' = LU$ se tendría que $L^{-1}L' = UU'^{-1}$. Como L^{-1} también es triangular inferior con *unos* en la diagonal, el producto $L^{-1}L'$ también es una matriz del mismo tipo. Análogamente, el producto UU'^{-1} resulta ser una triangular superior. El hecho de que $L^{-1}L' = UU'^{-1}$ nos dice que necesariamente $L^{-1}L' = I$, ya que es simultáneamente triangular inferior y superior y su diagonal es de *unos*. Así pues $L^{-1}L' = I$, por lo que $L = L'$ y, por tanto $U = U'$ es decir, la factorización es única.

Debido a la unicidad de la factorización, ésta puede ser calculada por un método directo, es decir, haciendo

$$A = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ l_{21} & 1 & 0 & \cdots & 0 \\ l_{31} & l_{32} & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ l_{n1} & l_{n2} & l_{n3} & \cdots & 1 \end{pmatrix} \begin{pmatrix} u_{11} & u_{12} & u_{13} & \cdots & u_{1n} \\ 0 & u_{22} & u_{23} & \cdots & u_{2n} \\ 0 & 0 & u_{33} & \cdots & u_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & u_{nn} \end{pmatrix}$$

y calculando los valores de los n^2 elementos que aparecen entre las dos matrices.

Así, por ejemplo, para $A = \begin{pmatrix} 3 & 1 & 2 \\ 6 & 3 & 2 \\ -3 & 0 & -8 \end{pmatrix}$ tenemos

$$\begin{pmatrix} 3 & 1 & 2 \\ 6 & 3 & 2 \\ -3 & 0 & -8 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ l_{21} & 1 & 0 \\ l_{31} & l_{32} & 1 \end{pmatrix} \begin{pmatrix} u_{11} & u_{12} & u_{13} \\ 0 & u_{22} & u_{23} \\ 0 & 0 & u_{33} \end{pmatrix} =$$

$$= \begin{pmatrix} u_{11} & u_{12} & u_{13} \\ l_{21}u_{11} & l_{21}u_{12} + u_{22} & l_{21}u_{13} + u_{23} \\ l_{31}u_{11} & l_{31}u_{12} + l_{32}u_{22} & l_{31}u_{13} + l_{32}u_{23} + u_{33} \end{pmatrix}$$

por lo que de la primera fila obtenemos que

$$u_{11} = 3 \qquad u_{12} = 1 \qquad u_{13} = 2$$

de la segunda (teniendo en cuenta los valores ya obtenidos) se tiene que

$$\left.\begin{array}{rcl} 3l_{21} & = & 6 \\ l_{21} + u_{22} & = & 3 \\ 2l_{21} + u_{23} & = & 2 \end{array}\right\} \Longrightarrow \begin{array}{rcl} l_{21} & = & 2 \\ u_{22} & = & 1 \\ u_{23} & = & -2 \end{array}$$

y de la tercera (teniendo también en cuenta los resultados ya obtenidos)

$$\left.\begin{array}{rcl} 3l_{31} & = & -3 \\ l_{31} + l_{32} & = & 0 \\ 2l_{31} - 2l_{32} + u_{33} & = & -8 \end{array}\right\} \Longrightarrow \begin{array}{rcl} l_{31} & = & -1 \\ l_{32} & = & 1 \\ u_{33} & = & -4 \end{array}$$

es decir:

$$\begin{pmatrix} 3 & 1 & 2 \\ 6 & 3 & 2 \\ -3 & 0 & -8 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ -1 & 1 & 1 \end{pmatrix} \begin{pmatrix} 3 & 1 & 2 \\ 0 & 1 & -2 \\ 0 & 0 & -4 \end{pmatrix}$$

Se denominan *matrices fundamentales* de una matriz A, y se denotan por A_k, a las submatrices constituidas por los elementos de A situados en las k primeras filas y las k primeras columnas, es decir:

$$A_1 = (a_{11}) \qquad A_2 = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \qquad A_3 = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}$$

Teorema 2.8 *Una matriz regular A admite factorización LU si, y sólo si, sus matrices fundamentales A_i $(i = 1, 2, \ldots, n)$ son todas regulares.*

Demostración. Supongamos que A admite factorización LU. En ese caso

$$A = \left(\begin{array}{c|c} A_k & \\ \hline & \end{array}\right) = \left(\begin{array}{c|c} L_k & \\ \hline & \end{array}\right)\left(\begin{array}{c|c} U_k & \\ \hline & \end{array}\right) \Longrightarrow$$

$$A_k = L_k U_k \Longrightarrow \det(A_k) = \det(L_k)\det(U_k) = 1 \cdot r_{11} r_{22} \cdots r_{kk} \neq 0$$

ya que, por sea A regular, todos los *pivotes* r_{ii} $\;i = 1, 2, \ldots, n$ son no nulos.

Recíprocamente, si todas las matrices fundamentales son regulares, A admite factorización LU, o lo que es equivalente, se puede aplicar Gauss sin intercambio de filas. En efecto:

Dado que, por hipótesis es $a_{11} \neq 0$, se puede utilizar dicho elemento como pivote para anular al resto de los elementos de su columna quedándonos la matriz

$$A^{(2)} = \begin{pmatrix} a_{11}^{(2)} & a_{12}^{(2)} & \cdots & a_{1n}^{(2)} \\ 0 & a_{22}^{(2)} & \cdots & a_{2n}^{(2)} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & a_{n2}^{(2)} & \cdots & a_{nn}^{(2)} \end{pmatrix}$$

donde $a_{1i}^{(2)} = a_{1i}$ para $i = 1, 2, \ldots, n$.

Si nos fijamos ahora en $a_{22}^2 = a_{22} - a_{12}\dfrac{a_{21}}{a_{11}}$ podemos ver que es no nulo, ya que de ser nulo sería

$$a_{11}a_{22} - a_{12}a_{21} = \det(A_2) = 0$$

en contra de la hipótesis de que *todas* las matrices fundamentales son regulares. Por tanto, podemos utilizar $a_{22}^{(2)} \neq 0$ como nuevo pivote para anular a los elementos de su columna situados bajo él.

Reiterando el procedimiento se puede ver que todos los elementos que vamos obteniendo en la diagonal son no nulos y, por tanto, válidos como pivotes. Es decir, puede aplicarse el método de Gauss sin intercambio de filas. ■

Comprobar si una matriz admite factorización LU estudiando si todas sus matrices fundamentales son regulares es un método demasiado costoso debido al número de determinantes que hay que calcular.

Definición 2.2 Una matriz cuadrada de orden n $\;A = (a_{ij})_{\substack{i\,=\,1,\,2,\,\ldots,\,n \\ j\,=\,1,\,2,\,\ldots,\,n}}$ se dice que es una matriz de *diagonal estrictamente dominante*:

a) Por filas: si $|a_{ii}| > \displaystyle\sum_{\substack{k\,=\,1 \\ k\,\neq\,i}}^{n} |a_{ik}| \quad i = 1, 2, \ldots, n.$

b) Por columnas: si $|a_{ii}| > \sum_{\substack{k=1 \\ k \neq i}}^{n} |a_{ki}| \quad i = 1, 2, \ldots, n.$

Si en vez de $>$ es $\geq$ se dirá que es de diagonal dominante.

Así, por ejemplo, la matriz $A = \begin{pmatrix} 3 & 1 & 1 \\ 0 & 2 & 1 \\ 2 & -1 & 5 \end{pmatrix}$ es de diagonal estrictamente dominante por filas y dominante por columnas.

Una matriz se dice de *Hadamard* si es de diagonal estrictamente dominante.

Teorema 2.9 *Toda matriz de Hadamard es regular.*

Demostración. Supongamos que A es de diagonal estrictamente dominante por filas (de igual forma podría probarse si lo fuese por columnas) y que su determinante fuese nulo. En ese caso, el sistema $Ax = 0$ posee solución no trivial $(\alpha_1, \alpha_2, \ldots, \alpha_n) \neq (0, 0, \ldots, 0)$.

Sea $|\alpha_k| = \max_i |\alpha_i| > 0$ y consideremos la k-ésima ecuación:

$$a_{k1}\alpha_1 + a_{k2}\alpha_2 + \cdots + a_{kk}\alpha_k + \cdots + a_{kn}\alpha_n = 0 \Longrightarrow$$

$$a_{k1}\frac{\alpha_1}{\alpha_k} + a_{k2}\frac{\alpha_2}{\alpha_k} + \cdots + a_{kk} + \cdots + a_{kn}\frac{\alpha_n}{\alpha_k} = 0 \Longrightarrow$$

$$a_{kk} = -a_{k1}\frac{\alpha_1}{\alpha_k} - \cdots - a_{k\,k-1}\frac{\alpha_k - 1}{\alpha_k} - a_{k\,k+1}\frac{\alpha_k + 1}{\alpha_k} - \cdots - a_{kn}\frac{\alpha_n}{\alpha_k} \Longrightarrow$$

$$|a_{kk}| \leq \sum_{\substack{i=1 \\ i \neq k}}^{n} |a_{ki}|\frac{\alpha_i}{\alpha_k} \leq \sum_{\substack{i=1 \\ i \neq k}}^{n} |a_{ki}|$$

en contra de la hipótesis de que A es de diagonal estrictamente dominante por filas. Por tanto, toda matriz de Hadamard, es regular. ■

Teorema 2.10 *Las matrices fundamentales A_k de una matriz A de Hadamard, son también de Hadamard.*

Demostración. La demostración es trivial, ya que si A es de Hadamard se verifica que

$$|a_{ii}| > \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}| \geq \sum_{\substack{j=1 \\ j \neq i}}^{k} |a_{ij}|$$

luego A_k también lo es. ■

Como consecuencia de los Teoremas 2.8, 2.10 y 2.9, podemos deducir el siguiente corolario.

Corolario 2.11 *Toda matriz de Hadamard admite factorización LU.*

Otro tipo de matrices de las que se puede asegurar que admiten factorización LU son las *hermíticas definidas positivas*, ya que las matrices fundamentales de éstas tienen todas determinante positivo, por lo que el Teorema 2.8 garantiza la existencia de las matrices L y U.

2.4 Factorización de Cholesky

Una vez visto el método de Gauss basado en la factorización LU vamos a estudiar otros métodos que se basan en otros tipos de descomposiciones de la matriz del sistema.

Es conocido que toda matriz hermítica y definida positiva tiene sus autovalores reales y positivos y, además, en la factorización LU todos los pivotes son reales y positivos.

Teorema 2.12 [Factorización de Cholesky] *Toda matriz A hermítica y definida positiva puede ser descompuesta de la forma $A = BB^*$ siendo B una matriz triangular inferior.*

Demostración. Por tratarse de una matriz hermítica y definida positiva, sabemos que admite factorización LU. Sea

$$A = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ l_{21} & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ l_{n1} & l_{n2} & \cdots & 1 \end{pmatrix} \begin{pmatrix} u_{11} & u_{12} & \cdots & u_{1n} \\ 0 & u_{22} & \cdots & u_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & u_{nn} \end{pmatrix} =$$

$$= \begin{pmatrix} 1 & 0 & \cdots & 0 \\ l_{21} & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ l_{n1} & l_{n2} & \cdots & 1 \end{pmatrix} \begin{pmatrix} u_{11} & 0 & \cdots & 0 \\ 0 & u_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & u_{nn} \end{pmatrix} \begin{pmatrix} 1 & u_{12}/u_{11} & \cdots & u_{1n}/u_{11} \\ 0 & 1 & \cdots & u_{2n}/u_{22} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix} =$$

$$= L \begin{pmatrix} u_{11} & 0 & \cdots & 0 \\ 0 & u_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & u_{nn} \end{pmatrix} R =$$

$$= L \begin{pmatrix} \sqrt{u_{11}} & 0 & \cdots & 0 \\ 0 & \sqrt{u_{22}} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sqrt{u_{nn}} \end{pmatrix} \begin{pmatrix} \sqrt{u_{11}} & 0 & \cdots & 0 \\ 0 & \sqrt{u_{22}} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sqrt{u_{nn}} \end{pmatrix} R \Rightarrow$$

$$A = BC$$

donde $B = L \begin{pmatrix} \sqrt{u_{11}} & 0 & 0 & \cdots & 0 \\ 0 & \sqrt{u_{22}} & 0 & \cdots & 0 \\ 0 & 0 & \sqrt{u_{33}} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \sqrt{u_{nn}} \end{pmatrix}$ es una matriz triangular inferior y $C = \begin{pmatrix} \sqrt{u_{11}} & 0 & 0 & \cdots & 0 \\ 0 & \sqrt{u_{22}} & 0 & \cdots & 0 \\ 0 & 0 & \sqrt{u_{33}} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \sqrt{u_{nn}} \end{pmatrix} R$ es una triangular superior.

Como A es hermítica, $BC = A = A^* = C^*B^*$, por lo que $(C^*)^{-1}B = B^*C^{-1}$, y dado que $(C^*)^{-1}B$ es triangular inferior y B^*C^{-1} es triangular superior, ambas han de ser diagonales.

Por otra parte, $B = LD$ y $C = DR$, por lo que $C^* = R^*D^* = R^*D$ y, por tanto, $(C^*)^{-1} = D^{-1}(R^*)^{-1}$. Así pues, $(C^*)^{-1}B = D^{-1}(R^*)^{-1}LD$.

Como las matrices diagonales conmutan,

$$(C^*)^{-1}B = D^{-1}D(R^*)^{-1}L = (R^*)^{-1}L.$$

Al ser $(R^*)^{-1}L$ triangular inferior con diagonal de unos y $(C^*)^{-1}B$ diagonal, podemos asegurar que $(R^*)^{-1}L = I$ o, lo que es lo mismo, $R^* = L$. Además, $B^*C^{-1} = I$, por lo que $C = B^*$, luego $A = BB^*$ donde $B = LD$. ■

La unicidad de las matrices L y U implica la unicidad de la matriz B y, por tanto, ésta puede ser calculada por un método directo.

Ejemplo 2.6 Consideremos el sistema

$$\begin{pmatrix} 4 & 2i & 4+2i \\ -2i & 2 & 2-2i \\ 4-2i & 2+2i & 10 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ -4 \end{pmatrix}$$

Realicemos la factorización BB^* directamente, es decir

$$\begin{pmatrix} 4 & 2i & 4+2i \\ -2i & 2 & 2-2i \\ 4-2i & 2+2i & 10 \end{pmatrix} = \begin{pmatrix} b_{11} & 0 & 0 \\ b_{21} & b_{22} & 0 \\ b_{31} & b_{32} & b_{33} \end{pmatrix} \begin{pmatrix} \overline{b_{11}} & \overline{b_{21}} & \overline{b_{31}} \\ 0 & \overline{b_{22}} & \overline{b_{32}} \\ 0 & 0 & \overline{b_{33}} \end{pmatrix}$$

Se obtiene multiplicando, que $|b_{11}|^2 = 4$ por lo que $b_{11} = 2$. Utilizando este resultado tenemos que $2\overline{b_{21}} = 2i$, por lo que $b_{21} = -i$ y que $2\overline{b_{31}} = 4 + 2i$ por lo que $b_{31} = 2 - i$.

Por otro lado, $|b_{21}|^2 + |b_{22}|^2 = 2$, por lo que $|b_{22}|^2 = 1$ y, por tanto, $b_{22} = 1$.

Como $b_{21}\overline{b_{31}} + b_{22}\overline{b_{32}} = 2 - 2i$ tenemos que $1 - 2i + \overline{b_{32}} = 2 - 2i$, es decir, $\overline{b_{32}} = 1$ y, por tanto, $b_{32} = 1$.

Por último, $|b_{31}|^2 + |b_{32}|^2 + |b_{33}|^2 = 10$, por lo que $5 + 1 + |b_{33}|^2 = 10$, es decir $|b_{33}|^2 = 4$ y, por tanto, $b_{33} = 2$. Así pues, el sistema nos queda de la forma

$$\begin{pmatrix} 2 & 0 & 0 \\ -i & 1 & 0 \\ 2-i & 1 & 2 \end{pmatrix} \begin{pmatrix} 2 & i & 2+i \\ 0 & 1 & 1 \\ 0 & 0 & 2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ -4 \end{pmatrix}$$

Haciendo ahora $\begin{pmatrix} 2 & 0 & 0 \\ -i & 1 & 0 \\ 2-i & 1 & 2 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ -4 \end{pmatrix}$, se obtiene

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ -2 \end{pmatrix}$$

y de aquí, que

$$\begin{pmatrix} 2 & i & 2+i \\ 0 & 1 & 1 \\ 0 & 0 & 2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ -2 \end{pmatrix}$$

de donde obtenemos que la solución del sistema es

$$x_1 = 1 \qquad x_2 = 1 \qquad x_3 = -1$$

□

Hemos visto que toda matriz hermítica y definida positiva admite factorización de Cholesky, pero podemos llegar más lejos y enunciar el siguiente teorema (que no probaremos).

Teorema 2.13 *Una matriz hermítica y regular A es definida positiva si, y sólo si, admite factorización de Cholesky.*

Demostración. Si es hermítica y definida positiva admite factorización LU con todos los elementos diagonales de U (pivotes) positivos, por lo que admite factorización de Cholesky.

Recíprocamente, si A admite factorización de Cholesky es $A = BB^*$ por lo que

$$A^* = (BB^*)^* = BB^* = A \implies A \text{ es hermítica}$$

Para cualquier vector x no nulo es

$$x^*Ax = x^*BB^*x = (B^*x)^*(B^*x) = \|B^*x\|^2 \geq 0$$

siendo cero sólo si $B^*x = 0$ pero al ser B^* regular (si no lo fuese tampoco lo sería A en contra de la hipótesis) $B^*x = 0 \Longrightarrow x = 0$ en contra de la hipótesis de que x no es el vector nulo.

Se tiene pues que $x^*Ax > 0$ para cualquier vector x no nulo, es decir, A es definida positiva. ■

2.5 Métodos iterados

Un método iterado de resolución del sistema $Ax = b$ es aquel que genera, a partir de un vector inicial x_0, una sucesión de vectores $x_1, x_2, \ldots$. El método se dirá que es *consistente* con el sistema $Ax = b$, si el límite de dicha sucesión, en caso de existir, es solución del sistema. Se dirá que el método es *convergente* si la sucesión generada por *cualquier* vector inicial x_0 es convergente a la solución del sistema.

Es evidente que si un método es convergente es consistente, sin embargo, el recíproco no es cierto como prueba el siguiente ejemplo.

Ejemplo 2.7 El método $x_{n+1} = 2x_n - A^{-1}b$ es consistente con al sistema $Ax = b$ pero no es convergente. En efecto:

$$x_{n+1} - x = 2x_n - A^{-1}b - x = 2x_n - 2x - A^{-1}b + x = 2(x_n - x) - (A^{-1}b - x)$$

y como $A^{-1}b = x$, se tiene que

$$x_{n+1} - x = 2(x_n - x)$$

Si existe $\lim_{n\to\infty} x_n = x^*$ tendremos que

$$x^* - x = 2(x^* - x) \Longrightarrow x^* - x = 0 \Longrightarrow x^* = x$$

es decir, el límite es solución del sistema $Ax = b$, por lo que el método es consistente.

Sin embargo, de $x_{n+1} - x = 2(x_n - x)$ obtenemos que

$$\|x_{n+1} - x\| = 2\|x_n - x\|$$

es decir, el vector x_{n+1} dista de x el doble de lo que distaba x_n, por lo que el método no puede ser convergente. □

Los métodos iterados que trataremos son de la forma

$$x_{n+1} = Kx_n + c$$

en los que K será la que denominemos *matriz del método* y que dependerá de A y de b y en el que c es un vector que vendrá dado en función de A, K y b.

Teorema 2.14 *Un método iterado, de la forma $x_{n+1} = Kx_n + c$, es consistente con el sistema $Ax = b$ si, y sólo si, $c = (I - K)A^{-1}b$ y la matriz $I - K$ es invertible*

Demostración.

a) Supongamos que el método es consistente con el sistema $Ax = b$.

Como $x = Kx + (I - K)x = Kx + (I - K)A^{-1}b$, se tiene que

$$x_{n+1} - x = K(x_n - x) + c - (I - K)A^{-1}b \tag{2.1}$$

Por ser consistente el método, de existir $x^* = \lim_{n\to\infty} x_n$ ha de ser $x^* = x$. Pasando al límite en la Ecuación (2.1) obtenemos que

$$x^* - x = K(x^* - x) + c - (I - K)A^{-1}b$$

por lo que

$$(I - K)(x^* - x) = c - (I - K)A^{-1}b \tag{2.2}$$

y dado que $x^* = x$ nos queda que $0 = c - (I - K)A^{-1}b$, es decir,

$$c = (I - K)A^{-1}b.$$

Además, dado que $x = Kx + c$, el sistema $(I - K)x = c$ posee solución única x y, por tanto, la matriz $I - K$ es invertible.

b) Si $c = (I - K)A^{-1}b$ y la matriz $I - K$ es invertible, cuando exista $\lim_{n\to\infty} x_n = x^*$ se tendrá de (2.2) que

$$(I - K)(x^* - x) = 0$$

y como $I - K$ es invertible, $x^* = x$, por lo que el método es consistente. ■

Teorema 2.15 *Un método iterado de la forma $x_{n+1} = Kx_n + c$ consistente con el sistema $Ax = b$ es convergente si, y sólo si, $\lim_{n\to\infty} K^n = 0$.*

Demostración.

a) Por tratarse de un método consistente con el sistema $Ax = b$, se verifica que $c = (I - K)A^{-1}b$, por lo que

$$x_{n+1} = Kx_n + (I - K)A^{-1}b$$

restando el vector solución x a ambos miembros, podemos escribir

$$x_{n+1} - x = Kx_n - (K + I - K)x + (I - K)A^{-1}b =$$

$$= K(x_n - x) + (I - K)(A^{-1}b - x)$$

y dado que $A^{-1}b - x = 0$ obtenemos que $x_{n+1} - x = K(x_n - x)$.

Reiterando el proceso se obtiene:

$$x_n - x = K(x_{n-1} - x) = K^2(x_{n-2} - x) = \cdots = K^n(x_0 - x)$$

Pasando al límite

$$\lim_{n\to\infty}(x_n - x) = (\lim_{n\to\infty} K^n)(x_0 - x)$$

Al suponer el método convergente, $\lim_{n\to\infty}(x_n - x) = x - x = 0$, por lo que

$$\lim_{n\to\infty} K^n = 0$$

b) Recíprocamente, si $\lim_{n\to\infty} K^n = 0$, obtenemos que

$$\lim_{n\to\infty}(x_n - x) = 0$$

o lo que es lo mismo,

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

por lo que el método es convergente. ■

Teorema 2.16 *Si para alguna norma matricial es $\|K\| < 1$, el proceso $x_{n+1} = Kx_n + c$, donde $x_0 \in \mathbf{R}^n$ es un vector cualquiera, converge a la solución de la ecuación $x = Kx + c$ que existe y es única.*

Demostración.

a) Veamos, en primer lugar, que la ecuación $x = Kx + c$ posee solución única.

En efecto: $x = Kx + c \Longrightarrow (I - K)x = c$. Este sistema tiene solución única si, y sólo si, el sistema homogéneo asociado $(I - K)z = 0$ admite sólo la solución trivial $z = 0$, es decir, si $I - K$ es invertible.

La solución z no puede ser distinta del vector nulo ya que de serlo, como $\|K\| < 1$ se tiene que al ser $(I - K)z = 0$, o lo que es lo mismo, $z = Kz$

$$\|z\| = \|Kz\| \leq \|K\|\|z\| < \|z\|$$

lo cual es un absurdo, por lo que el sistema homogéneo sólo admite la solución trivial y, por tanto, el sistema completo $x = Kx + c$ posee solución única.

b) Probaremos ahora que la sucesión $\{x_n\}$ converge a x.

Dado que $x_{n+1} - x = (Kx_n + c) - (Kx + c) = K(x_n - x)$, podemos reiterar el proceso para obtener que $x_n - x = K^n(x_0 - x)$ por lo que

$$\|x_n - x\| = \|K^n\|\|x_0 - x\| \leq \|K\|^n\|x_0 - x\|$$

y dado que $\|K\| < 1$, pasando al límite se obtiene

$$\lim_{n\to\infty} \|x_n - x\| = 0 \Longrightarrow \lim_{n\to\infty} x_n = x$$

■

Obsérvese que si $\|K\| < 1$, dado que el radio espectral es una cota inferior de cualquier norma multiplicativa (ver Teorema 2.5) se tiene que $\rho(K) < 1$. El siguiente teorema nos dice que esta es además una condición suficiente para la convergencia del método $x_{n+1} = Kx_n + c$

Teorema 2.17 *Si $\rho(K) < 1$, el proceso $x_{n+1} = Kx_n + c$, donde $x_0 \in \mathbf{R}^n$ es un vector cualquiera, es convergente.*

Demostración. Si los autovalores de K son $\lambda_1, \lambda_2, \ldots, \lambda_m$ los de K^n son $\lambda_1^n, \lambda_2^n, \ldots, \lambda_m^n$. Como $\rho(K) < 1$ se tiene que $|\lambda_i| < 1$ cualquiera que sea $i = 1, 2, \ldots, m$ y, por tanto, la sucesión

$$\lambda_i, \lambda_i^2, \ldots, \lambda_i^n, \ldots$$

converge a cero, lo que nos lleva a que los autovalores de la matriz límite $\lim K^n$ son todos nulos, es decir $\lim K^n = 0$ que era la condición exigida en el Teorema 2.15. ■

Los métodos que vamos a estudiar consisten en descomponer la matriz invertible A del sistema $Ax = b$ de la forma $A = M - N$ de manera que la matriz M sea fácilmente invertible, por lo que reciben el nombre genérico de *métodos de descomposición* . El sistema queda entonces de la forma

$$(M - N)x = b \Longrightarrow Mx = Nx + b \Longrightarrow x = M^{-1}Nx + M^{-1}b$$

es decir, expresamos el sistema de la forma $x = Kx + c$ con $K = M^{-1}N$ y $c = M^{-1}b$. Dado que

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

y la matriz $(I - K) = (I - M^{-1}N) = M^{-1}(M - N) = M^{-1}A$ es invertible, estamos en las condiciones del Teorema 2.14 por lo que el método $x_{n+1} = Kx_n + c$ es consistente con el sistema $Ax = b$. Es decir, si el proceso converge, lo hace a la solución del sistema.

Sabemos también, por el Teorema 2.16, que el proceso será convergente si se verifica que $\|M^{-1}N\| < 1$ para alguna norma.

Para el estudio de los métodos que trataremos a continuación, vamos a descomponer la matriz A de la forma $A = D - E - F$ siendo

$$D = \begin{pmatrix} a_{11} & 0 & 0 & \cdots & 0 \\ 0 & a_{22} & 0 & \cdots & 0 \\ 0 & 0 & a_{33} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & a_{nn} \end{pmatrix} \qquad -E = \begin{pmatrix} 0 & 0 & 0 & \cdots & 0 \\ a_{21} & 0 & 0 & \cdots & 0 \\ a_{31} & a_{32} & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & a_{n3} & \cdots & 0 \end{pmatrix}$$

$$-F = \begin{pmatrix} 0 & a_{12} & a_{13} & \cdots & a_{1n} \\ 0 & 0 & a_{23} & \cdots & a_{2n} \\ 0 & 0 & 0 & \cdots & a_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 0 \end{pmatrix}$$

2.5.1 Método de Jacobi

Consiste en realizar la descomposición $A = M - N = D - (E + F)$. El sistema queda de la forma

$$Ax = b \Longrightarrow Dx = (E + F)x + b \Longrightarrow x = D^{-1}(E + F)x + D^{-1}b$$

La matriz $J = D^{-1}(E + F) = D^{-1}(D - A) = I - D^{-1}A$ se denomina *matriz de Jacobi*.

Teorema 2.18 *Si A es una matriz de diagonal estrictamente dominante por filas, el método de Jacobi es convergente.*

Demostración. La matriz $J = D^{-1}(E + F)$ tiene todos los elementos diagonales nulos y las sumas de los módulos de los elementos de sus filas son todas menores que 1 ya que

$$|a_{ii}| > \sum_{\substack{k=1 \\ k \neq i}}^{n} |a_{ik}| \quad i = 1, 2, \ldots, n \implies \sum_{\substack{k=1 \\ k \neq i}}^{n} \frac{|a_{ik}|}{|a_{ii}|} < 1$$

por lo que los círculos de Gerschgorin están todos centrados en el origen y tienen radios menores que 1 es decir, todos los autovalores de J son menores que 1, por lo que $\rho(J) < 1$ y el método converge. ■

2.5.2 Método de Gauss-Seidel

Este método es el resultado de realizar la descomposición $A = M - N = (D - E) - F$. El sistema nos queda

$$Ax = b \Longrightarrow (D - E)x = Fx + b \Longrightarrow x = (D - E)^{-1}Fx + (D - E)^{-1}b$$

La matriz

$$L_1 = (D-E)^{-1}F = (A+F)^{-1}(A+F-A) = I-(A+F)^{-1}A = I-(D-E)^{-1}A$$

recibe el nombre de *matriz de Gauss-Seidel*.

Teorema 2.19 *Si A es una matriz de diagonal estrictamente dominante por filas, el método de Gauss-Seidel es convergente.*

2.5.3 Métodos de relajación (SOR)

Este método realiza la descomposición

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

El sistema se transforma entonces en

$$\frac{1}{\omega}(D - \omega E)x = \left(\frac{1-\omega}{\omega}D + F\right)x + b \Longrightarrow$$

$$(D - \omega E)x = \Big((1-\omega)D + \omega F\Big)x + \omega b \Longrightarrow$$

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

La matriz del método $L_\omega = (D - \omega E)^{-1}\Big((1-\omega)D + \omega F\Big)$ recibe el nombre de *matriz de relajación.*

- Si $\omega = 1$ la matriz se reduce a $L_1 = (D - E)^{-1}F$, es decir, se trata del método de Gauss Seidel.
- Si $\omega > 1$ se dice que se trata de un método de *sobre-relajación*
- Si $\omega < 1$ se dice que se trata de un método de *sub-relajación*

Teorema 2.20 *Una condición necesaria para que converja el método de relajación es que* $\omega \in (0, 2)$.

Teorema 2.21 *Si A es de diagonal estrictamente dominante por filas, el método de relajación es convergente cualquiera que sea* $\omega \in (0, 1]$.

Teorema 2.22 *Si A es simétrica y definida positiva, el método de relajación converge si, y sólo si,* $\omega \in (0, 2)$

2.6 Métodos del descenso más rápido y del gradiente conjugado

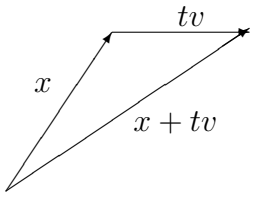
Los métodos que vamos a tratar a continuación son válidos para sistemas $Ax = b$ cuya matriz A es simétrica y definida positiva, es decir, para matrices tales que $A^T = A$ (simétrica) y $x^T Ax > 0$ cualquiera que sea el vector $x \neq 0$ (definida positiva).

Lema 2.23 *Si A es simétrica y definida positiva, el problema de resolver el sistema $Ax = b$ es equivalente al de minimizar la forma cuadrática*

$$q(x) = \langle\, x, Ax \,\rangle - 2\langle\, x, b \,\rangle$$

donde $\langle\, x, y \,\rangle = x^T y$ representa el producto escalar de los vectores x e y.

Demostración. Fijemos una dirección v (*rayo unidimensional*) y vamos a ver cómo se comporta la forma cuadrática q para vectores de la forma $x+tv$ donde t es un escalar.



$$\begin{aligned} q(x+tv) &= \langle\, x+tv, A(x+tv) \,\rangle - 2\langle\, x+tv, b \,\rangle \\ &= \langle\, x, Ax \,\rangle + 2t\langle\, x, Av \,\rangle + t^2\langle\, v, Av \,\rangle - 2\langle\, x, b \,\rangle - 2t\langle\, v, b \,\rangle \\ &= q(x) + 2t\langle\, v, Ax \,\rangle - 2t\langle\, v, b \,\rangle + t^2\langle\, v, Av \,\rangle \\ &= q(x) + 2t\langle\, v, Ax - b \,\rangle + t^2\langle\, v, Av \,\rangle \end{aligned} \tag{2.3}$$

ya que $A^T = A$.

La ecuación (2.3) (ecuación de segundo grado en t con el coeficiente de t^2 positivo, tiene un mínimo que se calcula igualando a cero la derivada

$$\frac{d}{dt} q(x+tv) = 2\langle\, v, Ax - b \,\rangle + 2t\langle\, v, Av \,\rangle$$

es decir, en el punto

$$\hat{t} = \langle\, v, b - Ax \,\rangle / \langle\, v, Av \,\rangle.$$

El valor mínimo que toma la forma cuadrática sobre dicho rayo unidimensional viene dado por

$$\begin{aligned} q(x+\hat{t}v) &= q(x) + \hat{t}\,[2\langle\, v, Ax - b \,\rangle + \hat{t}\langle\, v, Av \,\rangle] \\ &= q(x) + \hat{t}\,[2\langle\, v, Ax - b \,\rangle + \langle\, v, b - Ax \,\rangle] \\ &= q(x) - \hat{t}\langle\, v, b - Ax \,\rangle \\ &= q(x) - \langle\, v, b - Ax \,\rangle^2 / \langle\, v, Av \,\rangle \end{aligned}$$

Esto nos indica que al pasar de x a $x + \hat{t}v$ siempre hay una reducción en el valor de q excepto si $v \perp (b - Ax)$, es decir, si $\langle\, v, b - Ax \,\rangle = 0$. Así pues, si x no es una solución del sistema $Ax = b$ existen muchos vectores v tales que $\langle\, v, b - Ax \,\rangle \neq 0$ y, por tanto, x no minimiza a la forma cuadrática q. Por el

contrario, si $Ax = b$, no existe ningún rayo que emane de x sobre el que q tome un valor menor que $q(x)$, es decir, x minimiza el valor de q. ■

El lema anterior nos sugiere un método para resolver el sistema $Ax = b$ procediendo a minimizar la forma cuadrática q a través de una sucesión de rayos.

En el paso k del algoritmo se dispondrá de los vectores

$$x^{(0)}, x^{(1)}, x^{(2)}, \ldots, x^{(k)}.$$

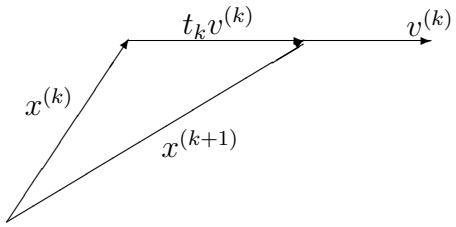
Estos vectores nos permitirán buscar una dirección apropiada $v^{(k)}$ y el siguiente punto de la sucesión vendrá dado por

$$x^{(k+1)} = x^{(k)} + t_k v^{(k)},$$

donde

$$t_k = \frac{\langle\, v^{(k)}, b - Ax^{(k)} \,\rangle}{\langle\, v^{(k)}, Av^{(k)} \,\rangle}$$

Gráficamente, si $\|v^{(k)}\| = 1$, t_k mide la distancia que nos movemos de $x^{(k)}$ para obtener $x^{(k+1)}$



2.6.1 Método del descenso más rápido

Si tomamos $v^{(k)}$ como el gradiente negativo de q en $x^{(k)}$, es decir, como la dirección del residuo $r^{(k)} = b - Ax^{(k)}$ obtenemos el denominado método del descenso más rápido.

Teniendo en cuenta que los diferentes vectores $x^{(i)}$ no es necesario conservarlos, los podemos sobreescribir obteniéndose el siguiente algoritmo:

input x, A, b, n
for $k = 1, 2, 3 \ldots, n$ **do**
  $v \leftarrow b - Ax$
  $t \leftarrow \langle\, v, v \,\rangle / \langle\, v, Av \,\rangle$
  $x \leftarrow x + tv$
  output k, x
end

Obsérvese que a medida que crece el valor de k, el residuo $v = b - Ax$ va disminuyendo, por lo que al encontrarnos en las proximidades de la solución, el cálculo de t se convierte prácticamente en una división de $\frac{0}{0}$ lo que puede alterar considerablemente el valor exacto que debería tomar t y que generalmente nos lleva a que el método diverge.

Este método resulta, en general, muy lento si las curvas de nivel de la forma cuadrática están muy próximas, por lo que no suele utilizarse en la forma descrita.

Sin embargo, utilizando condiciones de ortogonalidad en las denominadas *direcciones conjugadas*, puede ser modificado de forma que se convierta en un método de convergencia rápida que es conocido como método del gradiente conjugado.

2.6.2 Método del gradiente conjugado

Por no profundizar en el concepto de direcciones conjugadas y en cómo se determinan, nos limitaremos a dar el algoritmo correspondiente al método.

Es necesario tener en cuenta las mismas precauciones que en el algoritmo del descenso más rápido, es decir, debido a los errores de cálculo puede resultar un algoritmo divergente.

input $x, A, b, n, \varepsilon, \delta$
$r \leftarrow b - Ax$
$v \leftarrow r$
$c \leftarrow \langle r, r \rangle$
for $k = 1, 2, 3 \ldots, n$ **do**
  if $\langle v, v \rangle^{1/2} < \delta$ **then stop**
  $z \leftarrow Av$
  $t \leftarrow c/\langle v, z \rangle$
  $x \leftarrow x + tv$
  $r \leftarrow r - tz$
  $d \leftarrow \langle r, r \rangle$
  if $d < \varepsilon$ **then stop**
  $v \leftarrow r + (d/c)v$
  $c \leftarrow d$
  output k, x, r
end

2.7 Ejercicios propuestos

Ejercicio 2.1 Estudiar el número de condición de Frobenius de la matriz $A = \begin{pmatrix} a & -b \\ a+\varepsilon & -b \end{pmatrix}$.

Ejercicio 2.2 Dado el sistema:

$$\begin{cases} x + y = 2 \\ 2x + y = 3 \end{cases}$$

a) Calcular su número de condición de Frobenius.

b) Calcular "a" para que el número de condición del sistema resultante de sumarle a la segunda ecuación la primera multiplicada por dicha constante "a", sea mínimo.

Ejercicio 2.3 Dado el sistema:

$$\begin{cases} 3x + 4y = 7 \\ 3x + 5y = 8 \end{cases}$$

a) Calcular su número de condición euclídeo.

b) Sustituir la segunda ecuación por una combinación lineal de ambas, de forma que el número de condición sea mínimo.

Ejercicio 2.4 Comprobar que la matriz:

$$A = \begin{pmatrix} 1 & 2 & 0 & 0 & 0 \\ 1 & 4 & 3 & 0 & 0 \\ 0 & 4 & 9 & 4 & 0 \\ 0 & 0 & 9 & 16 & 5 \\ 0 & 0 & 0 & 16 & 25 \end{pmatrix}$$

admite factorización LU y realizarla.

Ejercicio 2.5 Resolver, por el método de Cholesky, el sistema de ecuaciones:

$$\begin{pmatrix} 6 & -1+3i & 1-2i \\ -1-3i & 3 & -1+i \\ 1+2i & -1-i & 2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} -1-2i \\ 1+i \\ 1-2i \end{pmatrix}$$

Ejercicio 2.6 Dada la matriz $A = \begin{pmatrix} p & -p & 2p \\ -p & p+2 & -1 \\ 2p & -1 & 6p-1 \end{pmatrix}$ se pide:

a) Determinar para qué valores de p es hermítica y definida positiva.

b) Para $p = 1$, efectuar la descomposición de Cholesky y utilizarla para resolver el sistema $Ax = b$ siendo $b = (1 \;\; 0 \;\; 3)^t$

Ejercicio 2.7 Resolver por los métodos de Jacobi, Gauss-Seidel y SOR con $\omega = 1.2$, el sistema:

$$\begin{aligned} 10x_1 - \;\; x_2 + \;\; 2x_3 \;\;\;\;\;\; &= \;\; 6 \\ -x_1 + 11x_2 - \;\; x_3 + 3x_4 &= \;\; 25 \\ 2x_1 - \;\; x_2 + 10x_3 - \;\; x_4 &= -11 \\ 3x_2 - \;\; x_3 + 8x_4 &= \;\; 15 \end{aligned}$$

Ejercicio 2.8 Al resolver por el método de Gauss-Seidel, utilizando MATLAB, el sistema

$$\begin{cases} x - 3y + 5z = 5 \\ 8x - \;\; y - \;\; z = 8 \\ -2x + 4y + \;\; z = 4 \end{cases}$$

observamos que el programa se detiene en la iteración 138 dándonos el vector $(inf \quad inf \quad -inf)^T$.

a) El método de Gauss-Seidel realiza el proceso $x_{n+1} = L_1 x_n + c$. Determina la matriz L_1.

b) Utilizar los círculos de Gerschgorin para estimar el módulo de los autovalores de L_1.

c) Justificar el porqué de la divergencia del método. (Indicación: utilizar el apartado anterior).

d) ¿Existe alguna condición suficiente que deba cumplir la matriz de un sistema para garantizar la convergencia del método de Gauss-Seidel? Hacer uso de ella para modificar el sistema de forma que el proceso sea convergente?

Ejercicio 2.9 Sea $\alpha \in \{0.5, 1.5, 2.5\}$ y consideremos el sistema iterado

$$\begin{pmatrix} x_{n+1} \\ \\ y_{n+1} \end{pmatrix} = \begin{pmatrix} \frac{1}{\alpha} - 1 & 1 \\ \\ -1 & \frac{1}{\alpha} + 1 \end{pmatrix} \begin{pmatrix} x_n \\ \\ y_n \end{pmatrix} + \begin{pmatrix} 1 - \frac{1}{\alpha} \\ \\ 1 - \frac{1}{\alpha} \end{pmatrix}$$

Se pide

a) Resolver el sistema resultante de tomar límites para probar que, en caso de que converja, el límite de la sucesión

$$\left(\begin{pmatrix} x_0 \\ y_0 \end{pmatrix}, \begin{pmatrix} x_1 \\ y_1 \end{pmatrix}, \begin{pmatrix} x_2 \\ y_2 \end{pmatrix} \dots \right)$$

no depende de α.

b) ¿Para qué valores de α converge la sucesión?

c) Para los valores anteriores que hacen que la sucesión sea convergente, ¿con cuál lo hace más rápidamente?

d) Comenzando con el vector $\begin{pmatrix} x_0 \\ y_0 \end{pmatrix} = \begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix}$, aproximar iteradamente el límite de la sucesión utilizando el valor de α que acelere más la convergencia.

Ejercicio 2.10 Sea el sistema $A\mathrm{x} = b$, donde

$$A = \begin{pmatrix} 1000 & 999 \\ 999 & 998 \end{pmatrix}, \quad \mathrm{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \quad \text{y} \quad b = \begin{pmatrix} 1999 \\ 1997 \end{pmatrix}.$$

a) Obtener la factorización LU de la matriz A. ¿Se puede conseguir la factorización de Choleski?

b) Resolver el sistema $A\mathrm{x} = b$ utilizando la factorización $A = LU$ obtenida en el apartado anterior.

c) Calcular $||A||_\infty$, $||A^{-1}||_\infty$ y el número de condición de la matriz $\kappa_\infty(A)$. ¿Se puede decir que está bien condicionada?

d) Comprueba que $||A\mathrm{x}||_\infty = ||A||_\infty$ para la solución $\mathrm{x} = (1, 1)^T$ del sistema $A\mathrm{x} = b$.

¿Cuál es el máximo valor que puede tomar $||A\mathrm{x}||_\infty$, cuando x es un vector unitario para la norma $||\ ||_\infty$?

e) Si se perturba b en $b+\delta b=(1998'99, 1997'01)^T$, calcular $||\delta b||_\infty/||b||_\infty$.

Si $\mathrm{x} + \delta\mathrm{x}$ es la solución obtenida para el nuevo sistema $A\mathrm{x} = b + \delta b$, ¿es el error relativo $||\delta\mathrm{x}||_\infty/||\mathrm{x}||_\infty$ el máximo que se puede cometer?

Indicación: $\dfrac{||\delta\mathrm{x}||_\infty}{||\mathrm{x}||_\infty} \leq \kappa_\infty(A)\dfrac{||\delta b||_\infty}{||b||_\infty}$.

3. Sistemas inconsistentes y sistemas indeterminados

3.1 Factorizaciones ortogonales

Si tratamos de resolver un sistema $Ax = b$ mediante la factorización LU (o la de Cholesky), lo que hacemos es transformar el sistema en $Ax = LUx = b$ para hacer $Ux = L^{-1}b$ que es un sistema triangular que se resuelve por sustitución regresiva. Sin embargo, la matriz del nuevo sistema es $U = L^{-1}A$ y dado que L^{-1} no es una matriz unitaria (ortogonal en el caso real) el número de condición de la matriz del sistema ha cambiado pudiendo estar peor condicionada que la matriz A del sistema original.

Vamos a estudiar, a continuación, otro tipo de factorización $A = QR$ donde R es, al igual que U, una matriz triangular superior, pero donde Q va a ser una matriz unitaria, por lo que el sistema $Ax = b$ lo transformaremos en $Rx = Q^{-1}b = Q^*b$ y, a diferencia del caso anterior, $R = Q^*A$ tiene el mismo número de condición que la matriz A del sistema original, ya que Q^* es unitaria.

3.2 Interpretación matricial del método de Gram-Schmidt: factorización QR

Consideremos la matriz regular $A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix} = (a_1 \ a_2 \ \cdots \ a_n)$ donde a_i representa la columna i-ésima de la matriz A.

Aplicando Gram-Schmidt existe un sistema ortonormal $\{y_1, y_2, \ldots, y_n\}$ tal

que $\mathcal{L}\{y_1, y_2, \ldots, y_k\} = \mathcal{L}\{a_1, a_2, \ldots, a_k\}$, por lo que el vector y_{k+1} pertenece a la variedad $\mathcal{L}^{\perp}\{a_1, a_2, \ldots, a_k\}$.

Sea Q la matriz cuyas columnas son los vectores y_i, $Q = (y_1\ y_2\ \cdots\ y_n)$. Entonces,

$$Q^*A = \begin{pmatrix} y_1^* \\ y_2^* \\ \vdots \\ y_n^* \end{pmatrix} (a_1\ a_2\ \cdots\ a_n)$$

es decir:

$$Q^*A = \begin{pmatrix} \langle\, a_1, y_1 \,\rangle & \langle\, a_2, y_1 \,\rangle & \cdots & \langle\, a_n, y_1 \,\rangle \\ \langle\, a_1, y_2 \,\rangle & \langle\, a_2, y_2 \,\rangle & \cdots & \langle\, a_n, y_2 \,\rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle\, a_1, y_n \,\rangle & \langle\, a_2, y_n \,\rangle & \cdots & \langle\, a_n, y_n \,\rangle \end{pmatrix}$$

Como $y_{k+1} \in \mathcal{L}^{\perp}\{a_1, a_2, \ldots, a_k\}$, se tiene que $\langle\, a_i, y_j \,\rangle = 0$ si, y sólo si, $i < j$, por lo que la matriz Q^*A es una triangular superior.

$$Q^*A = R = \begin{pmatrix} r_{11} & r_{12} & r_{13} & \cdots & r_{1n} \\ 0 & r_{22} & r_{23} & \cdots & r_{2n} \\ 0 & 0 & r_{33} & \cdots & r_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & r_{nn} \end{pmatrix}$$

Como las columnas de Q constituyen un sistema ortonormal de vectores, Q es unitaria, es decir $Q^*Q = I$, por lo que $A = QR$.

El problema que plantea la descomposición QR es que la matriz Q no es otra que la constituida por una base ortonormal obtenida a partir de las columnas de A por el método de Gram-Schmidt. Las transformaciones que se realizan para ortonormalizar los vectores columna de la matriz A mediante el método de Gram-Schmidt son transformaciones no unitarias, por lo que aunque el resultado sea una factorización ortogonal, los pasos que se han dado para llegar a ella han sido transformaciones no ortogonales. Ello nos lleva a tratar de buscar un método por el que podamos realizar una factorización QR en la que todos los pasos que se realicen sean transformaciones ortogonales.

3.3 Rotaciones y reflexiones

Consideremos la matriz

$$A = \begin{pmatrix} a_{11} & \cdots & a_{1i} & \cdots & a_{1j} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots & & \vdots & & \vdots \\ a_{i1} & \cdots & a_{ii} & \cdots & a_{ij} & \cdots & a_{in} \\ \vdots & & \vdots & \ddots & \vdots & & \vdots \\ a_{j1} & \cdots & a_{ji} & \cdots & a_{jj} & \cdots & a_{jn} \\ \vdots & & \vdots & & \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{ni} & \cdots & a_{nj} & \cdots & a_{nn} \end{pmatrix}$$

y tratemos de anular el elemento $a_{ji} \neq 0$. Para ello vamos a aplicar una rotación

$$Q_{ji} = \begin{pmatrix} 1 & \cdots & 0 & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & & \vdots & & \vdots \\ 0 & \cdots & \cos\alpha & \cdots & \operatorname{sen}\alpha & \cdots & 0 \\ \vdots & & \vdots & \ddots & \vdots & & \vdots \\ 0 & \cdots & -\operatorname{sen}\alpha & \cdots & \cos\alpha & \cdots & 0 \\ \vdots & & \vdots & & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \cdots & 0 & \cdots & 1 \end{pmatrix}$$

La matriz $Q_{ji}A$ nos queda

$$Q_{ji}A = \begin{pmatrix} a'_{11} & \cdots & a'_{1i} & \cdots & a'_{1j} & \cdots & a'_{1n} \\ \vdots & \ddots & \vdots & & \vdots & & \vdots \\ a'_{i1} & \cdots & a'_{ii} & \cdots & a'_{ij} & \cdots & a'_{in} \\ \vdots & & \vdots & \ddots & \vdots & & \vdots \\ a'_{j1} & \cdots & a'_{ji} & \cdots & a'_{jj} & \cdots & a'_{jn} \\ \vdots & & \vdots & & \vdots & \ddots & \vdots \\ a'_{n1} & \cdots & a'_{ni} & \cdots & a'_{nj} & \cdots & a'_{nn} \end{pmatrix}$$

con $a'_{ji} = -a_{ii}\operatorname{sen}\alpha + a_{ji}\cos\alpha$, por lo que

- Si $a_{ii} = 0$ se tiene que $a'_{ji} = a_{ji}\cos\alpha$ y si queremos anularlo $a_{ji}\cos\alpha = 0$. Dado que suponemos que $a_{ji} \neq 0$, basta con hacer $\cos\alpha = 0$, es decir: $\alpha = \pi/2$.
- Si $a_{ii} \neq 0$ tendremos que hacer $-a_{ii}\operatorname{sen}\alpha + a_{ji}\cos\alpha = 0$, por lo que $t = \operatorname{tg}\ \alpha = \dfrac{a_{ji}}{a_{ii}}$ y, por tanto,

$$\operatorname{sen}\alpha = \frac{t}{\sqrt{1+t^2}} \qquad \cos\alpha = \frac{1}{\sqrt{1+t^2}}$$

Obsérvese además que los únicos elementos que se alteran en la matriz $Q_{ji}A$ son los correspondientes a la filas y columnas i y j.

Podemos, por tanto, mediante rotaciones, anular todos los elementos subdiagonales y llegar a una matriz R triangular superior

$$Q_k \cdots Q_2 Q_1 A = R \iff Q^* A = R \quad \text{con} \quad Q^* = Q_k \cdots Q_2 Q_1$$

Dado que las matrices Q_i de las rotaciones son ortogonales, su producto también los es, por lo que Q^* es una matriz ortogonal y, por tanto, $A = QR$.

En éste método de factorización QR, a diferencia del aplicado anteriormente mediante el método de Gram-Schmidt todos los pasos que se dan están asociados a transformaciones ortogonales, sin embargo resulta costoso ya que cada rotación hace un único cero en la matriz A, por lo que para una matriz de orden n serían necesarias $\dfrac{n^2 - n}{2}$ rotaciones.

Otra posibilidad es hacer reflexiones en vez de rotaciones, ya que éstas consiguen anular todos los elementos situados por debajo de uno prefijado de una determinada columna. Este tipo de transformaciones vamos a estudiarlas a continuación con las denominadas *transformaciones de Householder*.

3.4 Transformaciones de Householder

Consideremos un espacio vectorial de dimensión n definido sobre un cuerpo $\mathbf{K}$, que denotaremos por $\mathbf{K}^n$ (en general trabajaremos en $\mathbf{R}^n$ ó $\mathbf{C}^n$). Dado un vector $v \in \mathbf{K}^n$ se define la transformación H de Householder asociada al vector v a la que viene definida por la matriz:

$$H = \begin{cases} I \in \mathbf{K}^{n \times n} & \text{si } v = 0 \\ \\ I - \dfrac{2}{v^* v} v v^* & \text{si } v \neq 0 \end{cases}$$

Proposición 3.1 *La transformación H de Householder asociada a un vector $v \in \mathbf{K}^n$ posee las siguientes propiedades:*

a) H es hermítica ($H^* = H$).

b) H es unitaria ($H^* H = H H^* = I$).

c) $H^2 = I$ o lo que es lo mismo, $H^{-1} = H$.

Demostración.

a) $H^* = \left(I - \dfrac{2}{v^*v}vv^*\right)^* = I^* - \overline{\left(\dfrac{2}{v^*v}\right)}(vv^*)^* = I - \dfrac{2}{v^*v}vv^* = H$

Obsérvese que $v^*v = \langle v, v \rangle = \|v\|^2 \in \mathbf{R}$, por lo que $\overline{v^*v} = v^*v$

b) $HH^* = HH = \left(I - \dfrac{2}{v^*v}vv^*\right)\left(I - \dfrac{2}{v^*v}vv^*\right) = I - \dfrac{4}{v^*v}vv^* + \left(\dfrac{2}{v^*v}\right)^2 vv^*vv^* =$

$= I - \dfrac{4}{v^*v}vv^* + \dfrac{4}{(v^*v)^2}v(v^*v)v^* = I - \dfrac{4}{v^*v}vv^* + \dfrac{4}{(v^*v)^2}(v^*v)vv^* = I.$

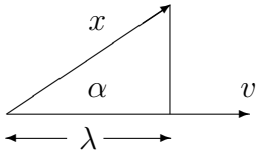
c) Basta observar que, por los apartados anteriores, $H^2 = HH = HH^* = I$.

3.4.1 Interpretación geométrica en $\mathbf{R}^n$

Sean $v \in \mathbf{R}^n$ un vector tal que $\|v\|_2 = 1$ y H la transformación de Householder asociada a él:

$$H = I - 2vv^T$$

Dado un vector $x \in \mathbf{R}^n$ se tiene que
$Hx = \left(I - 2vv^T\right)x = x - 2vv^Tx = x - 2v\langle x, v \rangle =$
$= x - 2v(\|x\| \, \|v\| \cos\alpha) = x - 2v(\|x\| \cos\alpha) =$
$= x - 2\lambda v$
con $\lambda = \|x\| \cos\alpha$, donde α representa el ángulo que forman los vectores x y v.



Sea y el vector simétrico de x respecto del hiperplano perpendicular a v. Podemos observar que $y + \lambda v = x - \lambda v$, o lo que es lo mismo, que $y = x - 2\lambda v = Hx$. Es decir, H transforma a un vector x en su simétrico respecto del hiperplano perpendicular al vector v.

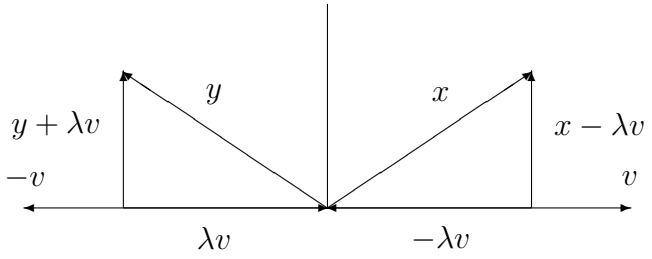
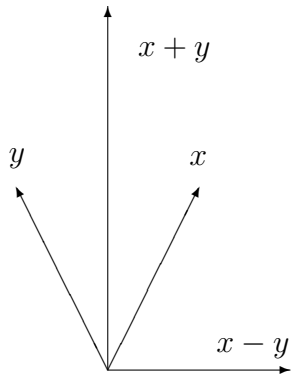


Figura 3.1: x y su transformado.

Es fácil observar que si x es ortogonal a v se verifica que $Hx = x$, así como que $Hv = -v$.

Así pues, si x e y son dos vectores de $\mathbf{R}^n$ tales que $x \neq y$ con $\|x\| = \|y\|$, la transformación de Householder asociada al vector $v = \dfrac{x-y}{\|x-y\|}$ transforma el vector x en y, es decir, $Hx = y$.
En efecto: dado que ambos vectores tienen la misma norma, $\langle\, x+y, x-y \,\rangle = \|x\|^2 - \langle\, x, y \,\rangle + \langle\, y, x \,\rangle - \|y\|^2 = 0$. Además, los vectores x e y son simétricos respecto de la dirección del vector $x+y$, por lo que la transformación de Householder asociada al vector $v = \dfrac{x-y}{\|x-y\|}$ transforma a x en y.



Consideremos los vectores $\begin{cases} x = (x_1, x_2, \ldots, x_n) \\ y = (x_1, x_2, \ldots, x_k, \sqrt{x_{k+1}^2 + \cdots + x_n^2}, 0, \ldots, 0) \end{cases}$ que poseen la misma norma.

Si $v = \dfrac{x-y}{\|x-y\|} = \dfrac{1}{\|x-y\|} \begin{pmatrix} 0 \\ \vdots \\ 0 \\ x_{k+1} - \sqrt{x_{k+1}^2 + \cdots + x_n^2} \\ x_{k+2} \\ \vdots \\ x_n \end{pmatrix}$ la transformación H de Householder asociada a v transforma a x en y.

3.4.2 Householder en $\mathbf{C}^n$

Sean v un vector de $\mathbf{C}^n$ y H la transformación de Householder asociada a él:

$$H = I - \frac{2}{v^*v} vv^*$$

- Si $x = \lambda v$ con $\lambda \in \mathbf{C}$ entonces
$$Hx = \left(I - \frac{2}{v^*v} vv^* \right) \lambda v = \lambda v - \frac{2\lambda}{v^*v} vv^*v = -\lambda v = -x$$
- Si $x \perp v$ se verifica que $\langle\, x, v \,\rangle = v^*x = 0$ y, por tanto
$$Hx = \left(I - \frac{2}{v^*v} vv^* \right) x = x - \frac{2}{v^*v} vv^*x = x$$

Es decir, los vectores ortogonales a v son invariantes mediante H.

Cualquier vector $x \in \mathbf{C}^n$ puede ser descompuesto de forma única en la suma de uno proporcional a v y otro w perteneciente a la variedad W ortogonal a v, es decir $x = \lambda v + w$ con $w \perp v$. Así pues,

$$Hx = H(\lambda v + w) = H(\lambda v) + Hw = -\lambda v + w$$

por lo que Hx es el vector simétrico de x respecto del hiperplano ortogonal a v.

Si $x = (x_1, x_2, \ldots, x_n) \in \mathbf{C}^n$ y pretendemos encontrar un vector v tal que la transformación de Householder H_v asociada a v transforme dicho vector x en otro $y = (y_1, 0, \ldots, 0)$ es evidente que como

$$\|y\|_2 = \|H_v x\|_2 = \|x\|_2$$

por ser H_v unitaria, ambos vectores x e y han de tener igual norma, es decir, ha de verificarse que $|y_1| = \|x\|_2$ o lo que es lo mismo, $y_1 = \|x\|_2 e^{i\alpha}$ con $\alpha \in \mathbf{R}$.

Tomemos un vector v unitario, es decir, tal que $|x_1|^2 + |x_2|^2 + \cdots + |x_n|^2 = 1$. Entonces

$$H_v x = (I - 2vv^*)\, x = x - 2vv^*x = x - (2v^*x)v = y$$

Obligando a que $2v^*x = 1$ se tiene que $x - v = y$, o lo que es lo mismo, que $v = x - y$, es decir:

$$v = \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{pmatrix} = \begin{pmatrix} x_1 - y_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$$

De este vector son conocidas todas sus componentes excepto la primera $v_1 = x_1 - y_1$.

Sabemos que $2v^*x = 1$ y que $v^*v = 1$, por lo que

$$\left.\begin{aligned} 2(\overline{v_1}x_1 + \overline{v_2}x_2 + \cdots + \overline{v_n}x_n) &= 1 \\ \overline{v_1}v_1 + \overline{v_2}v_2 + \cdots + \overline{v_n}v_n &= 1 \end{aligned}\right\} \Rightarrow \quad \left.\begin{aligned} 2(\overline{v_1}x_1 + |x_2|^2 + \cdots + |x_n|^2) &= 1 \\ |v_1|^2 + |x_2|^2 + \cdots + |x_n|^2 &= 1 \end{aligned}\right\} \Rightarrow$$

$$2(\overline{v_1}x_1 + \|x\|^2 - |x_1|^2) = 1 \tag{3.1}$$

$$|v_1|^2 + \|x\|^2 - |x_1|^2 \;\; = 1 \tag{3.2}$$

De la ecuación 3.1 se deduce que $\overline{v_1}x_1$ es un número real, por lo que el argumento del producto $\overline{v_1}x_1$ ha de ser 0 ó π. Como, por otra parte, $\overline{v_1}x_1 = |v_1|\,|x_1|\,e^{i(v_1\hat{}\,x_1)}$, los complejos v_1 y x_1 han de tener igual argumento, por lo que $v_1 = \lambda x_1$.

Llevando este resultado a las ecuaciones 3.1 y 3.2 se obtiene que

$$\left.\begin{array}{r} 2(\lambda\,|x_1|^2 + \|x\|^2 - |x_1|^2) = 1 \\ \lambda^2\,|x_1|^2 + \|x\|^2 - |x_1|^2 \;\; = 1 \end{array}\right\} \Longrightarrow |x_1|^2\,(\lambda^2 - 2\lambda + 1) = \|x\|^2 \Longrightarrow$$

$$\lambda = 1 \pm \frac{\|x\|}{|x_1|} \quad \text{supuesto } x_1 \neq 0$$

(Si $x_1 = 0$ basta con tomar $v = (\|x\|\,, x_2, \ldots, x_n)$).

El vector que buscamos es, por tanto, $v = \begin{pmatrix} \left(1 \pm \dfrac{\|x\|}{|x_1|}\right) x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$ obteniéndose que

$$H_v x = \begin{pmatrix} y_1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \quad \text{con} \quad y_1 = \|x\|\,e^{i\alpha}$$

que resulta ser

$$H_v x = y = x - v = \begin{pmatrix} \mp\dfrac{x_1}{|x_1|}\,\|x\| \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

3.4.3 Factorización QR de Householder

Supongamos que tenemos el sistema $Ax = b$ con $A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix}$.

Sean $x_1 = \begin{pmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{n1} \end{pmatrix}$ e $y_1 = \begin{pmatrix} \sqrt{a_{11}^2 + \cdots + a_{n1}^2} \\ 0 \\ \vdots \\ 0 \end{pmatrix} = \begin{pmatrix} r_{11} \\ 0 \\ \vdots \\ 0 \end{pmatrix}$.

Sea H_1 la transformación de Householder asociada al vector $v_1 = \dfrac{x_1 - y_1}{\|x_1 - y_1\|}$, por lo que $H_1 x_1 = y_1$. Tenemos entonces que

$$H_1 A = \begin{pmatrix} r_{11} & a_{12}^{(1)} & \cdots & a_{1n}^{(1)} \\ 0 & a_{22}^{(1)} & \cdots & a_{2n}^{(1)} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & a_{n2}^{(1)} & \cdots & a_{nn}^{(1)} \end{pmatrix} = \begin{pmatrix} a_1^{(1)} & a_2^{(1)} & \cdots & a_n^{(1)} \end{pmatrix}$$

en la que $a_i^{(1)}$ representa la columna i-ésima de la matriz $H_1 A$.

Busquemos ahora otro vector v_2 tal que la transformación de Householder H_2 asociada a él, deje invariante al vector $a_1^{(1)}$ y transforme al vector $a_2^{(1)}$ en otro de la forma $(r_{12}, r_{22}, 0, \ldots, 0)$.

Como se quiere que deje invariante al vector $a_1^{(1)}$, v_2 ha de ser ortogonal a él, es decir, debe ser de la forma $(0, u_2, \ldots, u_n)$.

Tomando $x_2 = a_2^{(1)} = \begin{pmatrix} a_{21}^{(1)} \\ a_{22}^{(1)} \\ a_{32}^{(1)} \\ \vdots \\ a_{n2}^{(1)} \end{pmatrix}$ e $y_2 = \begin{pmatrix} a_{21}^{(1)} \\ \sqrt{(a_{22}^{(1)})^2 + \cdots + (a_{n2}^{(1)})^2} \\ 0 \\ \vdots \\ 0 \end{pmatrix}$, la transformación H_2 asociada al vector $v_2 = \dfrac{x_2 - y_2}{\|x_2 - y_2\|}$ deja invariante al vector $a_1^{(1)}$ y transforma $x_2 = a_2^{(1)}$ en y_2.

Reiterando al procedimiento se puede triangularizar la matriz A llegar a una matriz triangular superior R. Llegados a ese paso se tiene que

$$H_k H_{k-1} \cdots H_1 A = R \iff Q^* A = R \quad \text{con} \quad Q^* = H_k H_{k-1} \cdots H_1$$

de donde

$$A = QR.$$

Si lo que nos interesa es resolver el sistema aplicamos las transformaciones

al sistema y no sólo a la matriz A.

$$H_k H_{k-1} \cdots H_1 A x = H_k H_{k-1} \cdots H_1 b \iff Rx = b'$$

sistema triangular de fácil resolución.

Ejemplo 3.1 Consideremos la matriz $A = \begin{pmatrix} 1 & -1 & -1 \\ 2 & 0 & 1 \\ -2 & 7 & 1 \end{pmatrix}$.

Como $x_1 = \begin{pmatrix} 1 \\ 2 \\ -2 \end{pmatrix}$ e $y_1 = \begin{pmatrix} \sqrt{1^2+2^2+(-2)^2} \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 3 \\ 0 \\ 0 \end{pmatrix}$, se tiene que

$$v_1 = \frac{x_1 - y_1}{\|x_1 - y_1\|} = \frac{1}{2\sqrt{3}} \begin{pmatrix} -2 \\ 2 \\ -2 \end{pmatrix} = \begin{pmatrix} -\frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{3}} \\ -\frac{1}{\sqrt{3}} \end{pmatrix}$$

$$H_1 = I - \frac{2}{v_1^* v_1} v_1 v_1^* = I - 2 \begin{pmatrix} -\frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{3}} \\ -\frac{1}{\sqrt{3}} \end{pmatrix} \begin{pmatrix} -\frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} & -\frac{1}{\sqrt{3}} \end{pmatrix} =$$

$$= \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} - 2 \begin{pmatrix} \frac{1}{3} & -\frac{1}{3} & \frac{1}{3} \\ -\frac{1}{3} & \frac{1}{3} & -\frac{1}{3} \\ \frac{1}{3} & -\frac{1}{3} & \frac{1}{3} \end{pmatrix} = \begin{pmatrix} \frac{1}{3} & \frac{2}{3} & -\frac{2}{3} \\ \frac{2}{3} & \frac{1}{3} & \frac{2}{3} \\ -\frac{2}{3} & \frac{2}{3} & \frac{1}{3} \end{pmatrix}$$

$$H_1 A = \begin{pmatrix} \frac{1}{3} & \frac{2}{3} & -\frac{2}{3} \\ \frac{2}{3} & \frac{1}{3} & \frac{2}{3} \\ -\frac{2}{3} & \frac{2}{3} & \frac{1}{3} \end{pmatrix} \begin{pmatrix} 1 & -1 & -1 \\ 2 & 0 & 1 \\ -2 & 7 & 1 \end{pmatrix} = \begin{pmatrix} 3 & -5 & -\frac{1}{3} \\ 0 & 4 & \frac{1}{3} \\ 0 & 3 & \frac{5}{3} \end{pmatrix}$$

Ahora son $x_2 = \begin{pmatrix} -5 \\ 4 \\ 3 \end{pmatrix}$ e $y_2 = \begin{pmatrix} -5 \\ \sqrt{4^2+3^2} \\ 0 \end{pmatrix} = \begin{pmatrix} -5 \\ 5 \\ 0 \end{pmatrix}$, por lo que

$$v_2 = \frac{x_2 - y_2}{\|x_2 - y_2\|} = \frac{1}{\sqrt{10}} \begin{pmatrix} 0 \\ -1 \\ 3 \end{pmatrix}$$

$$H_2 = I - \frac{2}{v_2^* v_2} v_2 v_2^* = I - \frac{2}{10} \begin{pmatrix} 0 \\ -1 \\ 3 \end{pmatrix} \begin{pmatrix} 0 & -1 & 3 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \frac{4}{5} & \frac{3}{5} \\ 0 & \frac{3}{5} & -\frac{4}{5} \end{pmatrix}$$

$$H_2 H_1 A = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \frac{4}{5} & \frac{3}{5} \\ 0 & \frac{3}{5} & -\frac{4}{5} \end{pmatrix} \begin{pmatrix} 3 & -5 & -\frac{1}{3} \\ 0 & 4 & \frac{1}{3} \\ 0 & 3 & \frac{5}{3} \end{pmatrix} = \begin{pmatrix} 3 & -5 & -\frac{1}{3} \\ 0 & 5 & \frac{19}{15} \\ 0 & 0 & -\frac{17}{15} \end{pmatrix} \qquad \square$$

Si partimos de una matriz $A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix}$ y tomamos

$$x_1 = a_1 = \begin{pmatrix} a_{11} & a_{21} & \cdots & a_{n1} \end{pmatrix}^T$$

definimos el vector $v_1 = \begin{pmatrix} \left(1 \pm \frac{\|x\|}{|a_{11}|}\right) a_{11} \\ a_{21} \\ \vdots \\ a_{n1} \end{pmatrix}$, obteniéndose que

$$H_1 A = \begin{pmatrix} \alpha_{11} & a_{12}^{(1)} & \cdots & a_{1n}^{(1)} \\ 0 & a_{21}^{(1)} & \cdots & a_{2n}^{(1)} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & a_{n2}^{(1)} & \cdots & a_{nn}^{(1)} \end{pmatrix}$$

Buscamos ahora otro vector v_2 tal que $H_2 \begin{pmatrix} a_{12}^{(1)} \\ a_{22}^{(1)} \\ a_{32}^{(1)} \\ \vdots \\ a_{n2}^{(1)} \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$ y de tal forma que mantenga invariante al vector $\begin{pmatrix} \alpha_{11} & 0 & \cdots & 0 \end{pmatrix}^T$. Bastará coger, para ello, $v_2 = \begin{pmatrix} 0 & v_2 & \cdots & v_n \end{pmatrix}^T$, que es ortogonal a él.

En este caso, la transformación de Householder viene dada por

$$H_2 = I - 2 \begin{pmatrix} 0 \\ v_2 \\ \vdots \\ v_n \end{pmatrix} \begin{pmatrix} 0 & \overline{v_2} & \cdots & \overline{v_n} \end{pmatrix} = I - 2 \begin{pmatrix} 0 & 0 & \cdots & 0 \\ 0 & v_2\overline{v_2} & \cdots & v_2\overline{v_n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & v_n\overline{v_2} & \cdots & v_n\overline{v_n} \end{pmatrix} =$$

$$= \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 - 2v_2\overline{v_2} & \cdots & -2v_2\overline{v_n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & -2v_n\overline{v_2} & \cdots & 1 - 2v_n\overline{v_n} \end{pmatrix} = \left(\begin{array}{c|ccc} 1 & 0 & \cdots & 0 \\ \hline 0 & & & \\ \vdots & & \mathbf{H} & \\ 0 & & & \end{array}\right)$$

Aplicando a la matriz A ambas transformaciones, se tiene:

$$H_2 H_1 A = H_2 \left(\begin{array}{c|ccc} \alpha_{11} & a_{12}^{(1)} & \cdots & a_{1n}^{(1)} \\ \hline 0 & a_{21}^{(1)} & \cdots & a_{2n}^{(1)} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & a_{n2}^{(1)} & \cdots & a_{nn}^{(1)} \end{array}\right) =$$

$$= \left(\begin{array}{c|ccc} 1 & 0 & \cdots & 0 \\ \hline 0 & & & \\ \vdots & & \mathbf{H} & \\ 0 & & & \end{array}\right) \left(\begin{array}{c|ccc} 1 & a_{12}^{(1)} & \cdots & a_{1n}^{(1)} \\ \hline 0 & & & \\ \vdots & & \mathbf{A}^1 & \\ 0 & & & \end{array}\right) = \left(\begin{array}{c|ccc} 1 & a_{12}^{(1)} & \cdots & a_{1n}^{(1)} \\ \hline 0 & & & \\ \vdots & & \mathbf{HA}^1 & \\ 0 & & & \end{array}\right)$$

Es decir, se trata de realizar un proceso análogo al anterior sólo que ahora en $\mathbf{C}^{n-1}$. Posteriormente se realizará otro en $\mathbf{C}^{n-2}$ y así sucesivamente hasta haber triangularizado la matriz A.

3.5 Sistemas superdeterminados. Problema de los mínimos cuadrados

Dado un sistema de ecuaciones de la forma $Ax = b$ en el A es una matriz cuadrada de orden n, $A \in \mathbf{K}^{n\times n}$, y x y b son vectores de $\mathbf{K}^n$ sabemos, por el teorema de Rouche-Fröbenius , que tiene solución si, y sólo si, existen $x_1, x_2, \ldots, x_n \in \mathbf{K}^n$ tales que

$$\begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{pmatrix}\begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}=\begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix} \Leftrightarrow x_1 \begin{pmatrix} a_{11} \\ \vdots \\ a_{n1} \end{pmatrix}+\cdots+x_n \begin{pmatrix} a_{1n} \\ \vdots \\ a_{nn} \end{pmatrix}=\begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix}$$

En otras palabras, el vector b puede expresarse como una combinación lineal de las columnas de la matriz A, por lo que $b \in C(A)$ (espacio columna de A).

Sin embargo, existen problemas en los que no ocurre así. Supongamos que se tienen tres puntos en el plano y se desea calcular la recta que pasa por ellos. Evidentemente, y dado que una recta la determinan sólo dos puntos, el problema no tiene solución (salvo que los tres puntos estén alineados). Desde el punto de vista algebraico este problema se expresa de la siguiente forma: sean $P = (a_1, b_1)$, $Q = (a_2, b_2)$ y $R = (a_3, b_3)$. Si tratamos de hallar la ecuación de la recta $y = mx + n$ que pasa por ellos se obtiene

$$\begin{array}{l} ma_1 + n = b_1 \\ ma_2 + n = b_2 \\ ma_3 + n = b_3 \end{array} \iff \begin{pmatrix} a_1 & 1 \\ a_2 & 1 \\ a_3 & 1 \end{pmatrix}\begin{pmatrix} m \\ n \end{pmatrix}=\begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix}$$

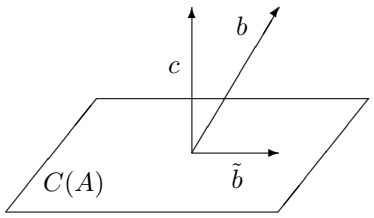
Decir que el sistema no posee solución equivale a decir que el vector $b = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix}$ no pertenece al espacio columna de la matriz $A = \begin{pmatrix} a_1 & 1 \\ a_2 & 1 \\ a_3 & 1 \end{pmatrix}$.

Se define un *sistema superdeterminado* como aquel sistema de ecuaciones lineales $Ax = b$ en el que $A \in \mathbf{K}^{m\times n}$, $x \in \mathbf{K}^n$ y $b \in \mathbf{K}^m$, donde $\mathbf{K} = \mathbf{R}$ ó $\mathbf{C}$.

Supongamos que se tiene un sistema superdeterminado

$$\begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \\ \vdots & & \vdots \\ a_{m1} & \cdots & a_{mn} \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} b_1 \\ \vdots \\ b_n \\ \vdots \\ b_m \end{pmatrix}$$

con $m > n$, en el que rg $A = n$ es decir, en el que la matriz del sistema tiene rango máximo, y denotemos por $a_1, a_2, \ldots, a_n$ las columnas de A.
Si el sistema es incompatible es debido a que el vector b no pertenece al espacio columna de A. Tomando cualquier vector $\overline{b} \in C(A)$ se sabe que el sistema $Ax = \overline{b}$ posee solución única.



De entre todos los vectores del espacio columna de A se trata de buscar aquel que minimiza su distancia al vector b, es decir, aquel vector $\tilde{b} \in C(A)$ tal que $\left\| b - \tilde{b} \right\|$ es mínima (problema de los mínimos cuadrados). Dicho vector sabemos que es la proyección ortogonal de b sobre el espacio $C(A)$ y, que respecto de la base formada por las columnas a_i $(1 \leq i \leq n)$ de la matriz A, tiene por coordenadas

$$\tilde{b} = (\langle\, b, a_1 \,\rangle, \langle\, b, a_2 \,\rangle, \ldots, \langle\, b, a_n \,\rangle\}$$

Dado que $b \notin C(A)$ y $\tilde{b} \in C(A)$ ($\tilde{b}$ proyección ortogonal de b sobre $C(A)$), podemos expresar b como suma de $\tilde{b}$ más otro vector c de la variedad ortogonal a $C(A)$ y, además, de forma única. Entonces:

$$\langle\, b, a_i \,\rangle = \langle\, \tilde{b} + c, a_i \,\rangle = \langle\, \tilde{b}, a_i \,\rangle + \langle\, c, a_i \,\rangle = \langle\, \tilde{b}, a_i \,\rangle \quad 1 \leq i \leq n$$

El sistema $Ax = \tilde{b}$ posee solución única es decir, existen $(\alpha_1, \alpha_2, \ldots, \alpha_n)$ únicos, tales que

$$\alpha_1 a_1 + \alpha_2 a_2 + \cdots + \alpha_n a_n = \tilde{b} \qquad \Longleftrightarrow \qquad A \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{pmatrix} = \tilde{b}$$

Multiplicando esta ecuación por a_1, a_2, ..., a_n, obtenemos

$$\alpha_1\langle\, a_1, a_1\,\rangle + \cdots + \alpha_n\langle\, a_1, a_n\,\rangle = \langle\, \tilde{b}, a_1\,\rangle = \langle\, b, a_1\,\rangle$$

$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$

$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$

$$\alpha_1\langle\, a_n, a_1\,\rangle + \cdots + \alpha_n\langle\, a_n, a_n\,\rangle = \langle\, \tilde{b}, a_n\,\rangle = \langle\, b, a_n\,\rangle$$

que equivale a

$$\begin{pmatrix} a_1^*a_1 & \cdots & a_1^*a_n \\ \vdots & \ddots & \vdots \\ a_n^*a_1 & \cdots & a_n^*a_n \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{pmatrix} = \begin{pmatrix} a_1^*b \\ \vdots \\ a_n^*b \end{pmatrix}$$

es decir

$$\begin{pmatrix} a_1^* \\ \vdots \\ a_n^* \end{pmatrix} \begin{pmatrix} a_1 & \cdots & a_n \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{pmatrix} = \begin{pmatrix} a_1^* \\ \vdots \\ a_n^* \end{pmatrix} b$$

o, lo que es lo mismo:

$$A^*A \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{pmatrix} = A^*b$$

Así pues, la solución del sistema $A^*Ax = A^*b$ nos proporciona las coordenadas, respecto de la base formada por las columnas de la matriz A, del vector $\tilde{b}$ proyección ortogonal de b sobre el espacio columna $C(A)$. Estas coordenadas constituyen lo que denominaremos *solución(es) en mínimos cuadrados* del sistema superdeterminado $Ax = b$ y de todas las posibles soluciones en mínimos cuadrados, la de menor norma recibe el nombre de *pseudosolución* del sistema.

Obsérvese que si la matriz A no tiene rango máximo, lo único que se dispone del espacio columna de A es de un sistema generador y no de una base, por lo que las coordenadas del vector proyección respecto de dicho sistema generador no son únicas obteniéndose infinitas soluciones en mínimos cuadrados del sistema.

3.5.1 Transformaciones en sistemas superdeterminados

Sabemos que dado un sistema compatible $Ax = b$ y mediante transformaciones elementales puede obtenerse otro sistema $BAx = Bb$ equivalente al anterior, es decir, obtenemos un sistema que posee la misma (o las mismas) soluciones que el sistema dado.

Si partimos de un sistema superdeterminado y realizamos, al igual que antes, transformaciones elementales, puede que el sistema obtenido no posea la misma *pseudosolución* que el sistema dado.

Obsérvese que para que los sistemas superdeterminados $Ax = b$ y $BAx = Bb$ posean la misma pseudosolución, han de tener igual solución (han de ser equivalentes) los sistemas $A^*Ax = A^*b$ y $(BA)^*BAx = (BA)^*Bb$. Dado que este último puede expresarse de la forma $A^*(B^*B)Ax = A^*(B^*B)b$, sólo podremos garantizar que ambos sistemas son equivalentes si $B^*B = I$ ya que, en dicho caso, ambos sistemas son el mismo. Es decir, las únicas transformaciones que podemos garantizar que no alterarán la solución de un sistema superdeterminado son las unitarias.

Dado que las transformaciones de Householder son unitarias, podemos utilizarlas para resolver sistemas superdeterminados.

Consideremos el sistema superdeterminado $Ax = b$ (en el que suponemos A de rango máximo). Mediante transformaciones de Householder $H_1, H_2, \ldots, H_n$ podemos transformar la matriz A en otra de la forma

$$HA = H_n \cdots H_1 A = \begin{pmatrix} t_{11} & t_{12} & \cdots & t_{1n} \\ 0 & t_{22} & \cdots & t_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & t_{nn} \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 0 \end{pmatrix} = \left(\begin{array}{c} T \\ \hline \Theta \end{array} \right)$$

La pseudosolución de este sistema superdeterminado es la solución del sistema

$$(HA)^*(HA)x = (HA)^*Hb \quad \Longleftrightarrow \quad \left(\begin{array}{c|c} T^* & \Theta \end{array} \right) \left(\begin{array}{c} T \\ \hline \Theta \end{array} \right) x = \left(\begin{array}{c|c} T^* & \Theta \end{array} \right) Hb$$

o llamando $Hb = b' = \begin{pmatrix} b'_1 \\ \vdots \\ b'_n \\ \vdots \\ b'_m \end{pmatrix}$, $\left(\begin{array}{c|c} T^* & \Theta \end{array} \right) \left(\begin{array}{c} T \\ \hline \Theta \end{array} \right) x = \left(\begin{array}{c|c} T^* & \Theta \end{array} \right) b'$.

Es fácil comprobar que $\left(\begin{array}{c|c} T^* & \Theta \end{array} \right) \begin{pmatrix} b'_1 \\ \vdots \\ b'_n \\ \vdots \\ b'_m \end{pmatrix} = T^* \begin{pmatrix} b'_1 \\ \vdots \\ b'_n \end{pmatrix}$, por lo que el cálculo de la pseudosolución del sistema superdeterminado $Ax = b$ se hace resolviendo el sistema

$$T^*Tx = T^* \begin{pmatrix} b'_1 \\ \vdots \\ b'_n \end{pmatrix}$$

y dado que estamos suponiendo que A tiene rango máximo, la matriz T posee inversa y por tanto T^*, por lo que la solución es la misma que la del sistema triangular

$$Tx = \begin{pmatrix} b'_1 \\ \vdots \\ b'_n \end{pmatrix} \qquad \Longleftrightarrow \qquad Tx = \tilde{b}$$

Una vez calculada la pseudosolución, la norma del error está representada por la distancia $\left\| b - \tilde{b} \right\|$ que viene dada por

$$\left\| \begin{pmatrix} T \\ \hline \Theta \end{pmatrix} x - \begin{pmatrix} b'_1 \\ \vdots \\ b'_n \\ \vdots \\ b'_m \end{pmatrix} \right\| = \left\| \begin{pmatrix} b'_1 \\ \vdots \\ b'_n \\ 0 \\ \vdots \\ 0 \end{pmatrix} - \begin{pmatrix} b'_1 \\ \vdots \\ b'_n \\ b'_{n+1} \\ \vdots \\ b'_m \end{pmatrix} \right\| = \left\| \begin{pmatrix} b'_{n+1} \\ \vdots \\ b'_m \end{pmatrix} \right\|$$

Por último, si la matriz A no tiene rango máximo, sus columnas no son linealmente independientes, por lo que sólo constituyen un sistema generador (no una base) del espacio columna $C(A)$. Ello nos lleva a la existencia de infinitas n-uplas $(\alpha_1, \alpha_2, \ldots, \alpha_n)$ soluciones del sistema $Ax = \tilde{b}$ y, por tanto, a infinitas soluciones en mínimos cuadrados del sistema superdeterminado, pero teniendo en cuenta que al ser única la proyección ortogonal $\tilde{b}$ de b sobre el espacio columna $C(A)$, todas ellas representan diferentes coordenadas del vector $\tilde{b}$ respecto del sistema generador de $C(A)$ dado por las columnas de A. Sin embargo, el error cometido $\left\|b - \tilde{b}\right\|$ es el mismo para todas las soluciones en mínimos cuadrados del sistema. De entre todas ellas, la de menor norma euclídea es la pseudosolución del sistema.

3.6 Descomposición en valores singulares y pseudoinversa de Penrose

La descomposición en valores singulares es otra factorización matricial que tiene muchas aplicaciones.

Teorema 3.2 *Toda matriz compleja A, de orden $m \times n$ puede ser factorizada de la forma $A = U\Sigma V^*$ donde U es una matriz unitaria $m \times m$, Σ una matriz diagonal $m \times n$ y V una unitaria de orden $n \times n$.*

Demostración. La matriz A^*A es hermítica de orden $n \times n$ y semidefinida positiva, ya que

$$x^*(A^*A)x = (Ax)^*(Ax) \geq 0$$

Resulta, de ello, que sus autovalores son reales no negativos $\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2$ (pudiendo estar repetidos, pero ordenados de forma que los r primeros son no nulos y los $n - r$ últimos son nulos). Los valores $\sigma_1, \sigma_2, \ldots, \sigma_n$ son los valores singulares de la matriz A.

Sea $\{v_1, v_2, \ldots, v_n\}$ un conjunto ortonormal de vectores propios de A^*A, dispuestos de forma que

$$A^*Av_i = \sigma_i^2 v_i$$

Se verifica entonces que

$$\|Av_i\|_2^2 = v_i^* A^* A v_i = v_i^* \sigma_i^2 v_i = \sigma_i^2$$

Esto nos muestra que $Av_i = 0$ si $i \geq r + 1$. Obsérvese que

$$r = \text{rg}\,(A^*A) \leq \text{mín}\{\text{rg}\,(A^*), \text{rg}\,(A)\} \leq \text{mín}\{m, n\}$$

Construyamos la matriz V de orden $n \times n$ cuyas filas son $v_1^*, v_2^*, \ldots, v_n^*$ y definamos

$$u_i = \sigma_i^{-1} A v_i \qquad 1 \leq i \leq r$$

Los vectores u_i constituyen un sistema ortonormal, ya que para $1 \leq i, j \leq r$ se tiene que

$$u_i^* u_j = \sigma_i^{-1}(Av_i)^* \sigma_j^{-1}(Av_j) = (\sigma_i\sigma_j)^{-1}(v_i^* A^* A v_j) = (\sigma_i\sigma_j)^{-1}(v_i^* \sigma_j^2 v_j) = \delta_{ij}$$

Eligiendo vectores adicionales $u_{n+1}, u_{n+2}, \ldots, u_m$ de tal forma que $\{u_1, \ldots, u_m\}$ constituya una base ortonormal de $\mathbf{C}^m$ y construyendo las matrices P de orden $m \times m$ cuyas columnas son los vectores u_i y la matriz Σ de orden $m \times n$ cuyos elementos diagonales $\Sigma_{ii} = \sigma_i$ y los restantes elementos nulos, se tiene que

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

Para probarlo vamos a ver que $U^* AV = \Sigma$. En efecto:

$$(U^* AV)_{ij} = u_i^* A v_j = u_i^* \sigma_j u_j = \sigma_j u_i^* u_j = \sigma_j \delta_{ij} = \Sigma_{ij}$$

3.6.1 Pseudoinversa de Penrose

Para las matrices D de orden $m \times n$ tales que $d_{ij} = 0$ si $i \neq j$ y $d_{ii} > 0$, se define la pseudoinversa como la matriz $n \times m$ D^+ cuyos elementos diagonales son los inversos de los elementos diagonales de D y el resto de los elementos son nulos.

En el caso general de una matriz A de orden $m \times n$ se define la pseudoinversa A^+ a través de la factorización en valores singulares $A = U\Sigma V^*$ de la forma $A^+ = V\Sigma^+ U^*$

La pseudoinversa comparte algunas propiedades con la inversa, pero sólo algunas ya que, por ejemplo, si A es de orden $m \times n$ con $m > n$ A^+ es de orden $n \times m$ y la matriz AA^+ no puede ser la matriz unidad, ya que AA^+ es de orden $m \times m$, por lo que si fuese la matriz unidad sería rg $(AA^*) = m$ cosa que no es posible por ser $n < m$ el máximo rango que pueden tener las matrices A y A^+ (recuérdese que el rango de la matriz producto nunca es superior al rango de las matrices que se multiplican).

Teorema 3.3 [Propiedades de Penrose] *Para cada matriz A existe, a lo más, una matriz X que verifica las siguientes propiedades:*

a) $AXA = A$

b) $XAX = X$

c) $(AX)^* = AX$

d) $(XA)^* = XA$

Demostración. Sean X e Y dos matrices que verifique las cuatro propiedades. Se tiene entonces que

$$\begin{aligned}
X &= XAX && \text{(b)}\\
&= XAY\,AX && \text{(a)}\\
&= XAY\,AY\,AY\,AX && \text{(a)}\\
&= (XA)^*(YA)^*Y(AY)^*(AX)^* && \text{(d) y (c)}\\
&= A^*X^*A^*Y^*YY^*A^*X^*A^*\\
&= (AXA)^*Y^*YY^*(AXA)^*\\
&= A^*Y^*YY^*A^* && \text{(a)}\\
&= (YA)^*Y(AY)^*\\
&= Y\,AY\,AY && \text{(d) y (c)}\\
&= Y\,AY && \text{(b)}\\
&= Y && \text{(b)}
\end{aligned}$$

■

Teorema 3.4 *La pseudoinversa de una matriz tiene las cuatro propiedades de Penrose y, por tanto, es única.*

Demostración. Sea $A = U\Sigma V^*$ la descomposición en valores singulares de una matriz A. Sabemos que $A^+ = V\Sigma^+U^*$.

Si la matriz A es de orden $m \times n$ y tiene rango r, la matriz Σ es también del mismo orden y tiene la forma

$$\Sigma_{ij} = \begin{cases} \sigma_i & \text{si } i = j \leq r \\ 0 & \text{en caso contrario} \end{cases}$$

Se tiene entonces que

$$\Sigma\Sigma^*\Sigma = \Sigma$$

ya que

$$(\Sigma\Sigma^*\Sigma)_{ij} = \sum_{k=1}^{n} \Sigma_{ik} \sum_{l=1}^{m} \Sigma_{kl}^{+} \Sigma_{lj}.$$

Los términos Σ_{ik} y Σ_{lj} hacen que el segundo miembro de la igualdad sea nulo excepto en los casos en que $i, j \leq r$ en cuyo caso

$$(\Sigma\Sigma^*\Sigma)_{ij} = \sum_{k=1}^{r} \Sigma_{ik} \sum_{l=1}^{r} \Sigma_{kl}^{+} \Sigma_{lj} = \sigma_i \sum_{l=1}^{r} \Sigma_{il}^{+} \Sigma_{lj} = \sigma_i \sigma_i^{-1} \Sigma_{ij} = \Sigma_{ij}.$$

Razonamientos análogos nos permiten probar que Σ^+ verifica las otras tres propiedades de Penrose.

Si nos fijamos ahora en A^+, se tiene que

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

y razonamientos similares nos permiten probar las otras tres propiedades. ■

Si la matriz A es de rango máximo, la matriz A^*A es invertible. Las ecuaciones normales del sistema $Ax = b$ vienen dadas por $A^*Ax = A^*b$ y dado que A^*A es invertible se tiene que

$$x = (A^*A)^{-1}A^*b. \tag{3.3}$$

Por otra parte, si hubiésemos resuelto el sistema a través de la pseudoinversa de Penrose habríamos obtenido

$$x = A^+b. \tag{3.4}$$

por lo que comparando las ecuaciones (3.3) y (3.4) obtenemos, teniendo en cuenta la unicidad de la pseudoinversa, que "si A es de rango máximo", la pseudoinversa viene dada por

$$A^+ = (A^*A)^{-1}A^*.$$

3.7 Ejercicios propuestos

Ejercicio 3.1 Dado el sistema:

$$\begin{aligned} 4x &+ 5y = 13 \\ 3x &+ 5y = 11 \end{aligned}$$

a) Realizar la factorización QR de la matriz, y resolverlo basándose en ella

   a.1) Mediante el método de Gram-Schmidt,

   a.2) Mediante transformaciones de Householder.

b) Calcular el número de condición euclídeo del sistema inicial y del transformado, comprobando que son iguales.

Ejercicio 3.2 Resolver por el método de Householder el sistema:

$$\begin{pmatrix} 1 & -1 & -1 \\ 2 & 0 & 1 \\ -2 & 7 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ 4 \\ -7 \end{pmatrix}$$

Ejercicio 3.3 Buscar la solución de mínimos cuadrados del sistema $Ax = b$, siendo:

$$A = \begin{pmatrix} 3 & -1 \\ 4 & 2 \\ 0 & 1 \end{pmatrix} \qquad \text{y} \qquad b = \begin{pmatrix} 0 \\ 2 \\ 1 \end{pmatrix}$$

a) A través de sus ecuaciones normales.

b) Por el método de Householder.

Ejercicio 3.4 Se considera el sistema de ecuaciones $Ax = b$ con

$$A = \begin{pmatrix} 1 & 2 \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \end{pmatrix} \quad \text{y} \quad b = \begin{pmatrix} 3 \\ 2 \\ 0 \\ 1 \end{pmatrix}.$$

Se pide:

a) Multiplicando el sistema por la traspuesta de A, calcular la pseudosolución utilizando el método de Choleski.

b) Sea $v = (-1, 1, 1, 1)^T$. Demostrar que la transformación de Householder asociada al vector v transforma la primera columna de la matriz A en el vector $(2, 0, 0, 0)^T$ dejando invariante la segunda columna de A así como al vector b.

c) Calcular la pseudosolución del sistema utilizando transformaciones de Householder, así como la norma del error.

d) Si la matriz A del sistema fuese cuadrada y su número de condición fuese mayor que 1, ¿qué ventajas e inconvenientes tendría el resolver el sistema multiplicando por la traspuesta de A y el resolverlo por transformaciones de Householder?

Ejercicio 3.5 Hallar la recta de regresión de los puntos:

$$(1'1, 5),\ (1, 5'1),\ (2, 7'3),\ (1'8, 6'9),\ (1'5, 6'1),\ (3, 8'8),\ (3'1, 9) \text{ y } (2'9, 9'1)$$

Ejercicio 3.6 Hallar la parábola de regresión de los puntos:

$$(1, 0),\ \ (0, 0),\ \ (-1, 0),\ \ (1, 2)\ \text{ y }\ (2, 3)$$

Ejercicio 3.7 Dado el sistema superdeterminado:

$$\begin{pmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 1 \\ 1 & 2 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \\ 0 \\ -1 \end{pmatrix}$$

calcular, mediante transformaciones de Householder, la solución en mínimos cuadrados (pseudosolución) así como la norma del error.

Ejercicio 3.8 Resolver el sistema

$$\begin{pmatrix} 2 & 1 \\ 2 & 0 \\ -1 & 2 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ -5 \end{pmatrix}$$

y obtener la norma del error:

a) Mediante sus ecuaciones normales.

b) Mediante transformaciones de Householder.

c) Hallando la inversa generalizada de la matriz del sistema.

Ejercicio 3.9 Se considera el sistema superdeterminado $Ax = b$ con

$$A = \begin{pmatrix} 1 & 7 & 15 \\ 1 & 4 & 8 \\ 1 & 0 & 1 \\ 1 & 3 & 6 \end{pmatrix} \quad \text{y} \quad b = \begin{pmatrix} 7 \\ 7 \\ -5 \\ -9 \end{pmatrix}$$

a) Resolverlo mediante transformaciones de Householder, dando la norma del vector error.

b) Hallar la inversa generalizada A^+ de la matriz A.

c) Utilizar la inversa generalizada para resolver el sistema y hallar la norma del vector error.

Ejercicio 3.10 Resolver el sistema superdeterminado

$$\begin{pmatrix} -3 & 1 & 1 \\ 1 & -3 & 1 \\ 1 & 1 & -3 \\ 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 8 \\ 4 \\ 0 \\ 4 \end{pmatrix}$$

calculando la inversa generalizada de la matriz A.

Ejercicio 3.11 Dado sistema superdeterminado $Ax = b$ con

$$A = \begin{pmatrix} 1 & 5 & 5 \\ 1 & 2 & 3 \\ 1 & 1 & 3 \\ 1 & 2 & 1 \end{pmatrix} \qquad \text{y} \qquad b = \begin{pmatrix} 7 \\ 16 \\ -3 \\ 10 \end{pmatrix}$$

se pide:

a) Resolverlo mediante transformaciones de Householder, dando la norma del vector error.

b) Teniendo en cuenta el rango de la matriz A, hallar su inversa generalizada.

c) Utilizar la inversa generalizada obtenida en el apartado anterior para calcular la pseudosolución del sistema y hallar la norma del vector error.

Ejercicio 3.12 Consideremos el sistema de ecuaciones $A\mathrm{x} = b$, con

$$A = \begin{pmatrix} 2 & -2 \\ 1 & -1 \\ -2 & 2 \end{pmatrix}, \quad \mathrm{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \quad \text{y} \quad b = \begin{pmatrix} 6 \\ 3 \\ 3 \end{pmatrix},$$

y un vector unitario u. Se pide:

a) Demostrar que si $H = I - 2uu^T$ es la matriz de Householder, asociada al vector u, entonces: H es ortogonal, $H^2 = I$ y $\|H\mathbf{a}\|_2 = \|\mathbf{a}\|_2$ cualquiera que sea el vector $\mathbf{a}$.

b) Obtener la matriz de Householder que transforma el vector $(2, 1, -2)^T$ en otro de la forma $(\alpha, 0, 0)^T$, con $\alpha > 0$.

c) Aplicando el método de Householder, probar que el sistema $A\mathrm{x} = b$ posee infinitas soluciones en cuadrados mínimos y que el error cometido, al considerar cualquiera de ellas, es el mismo.

d) Obtener la pseudosolución del sistema $A\mathrm{x} = b$. Es decir, la solución en cuadrados mínimos, de entre las obtenidas en el apartado anterior, que tenga menor norma euclídea.

Ejercicio 3.13 Sea el sistema $A\mathrm{x} = b$, donde

$$A = \begin{pmatrix} 0 & 3 \\ -3 & 5 \\ 4 & 0 \end{pmatrix}, \quad \mathrm{x} = \begin{pmatrix} x \\ y \end{pmatrix} \quad \text{y} \quad b = \begin{pmatrix} -10 \\ 6 \\ -8 \end{pmatrix}$$

a) Probar que la matriz $A^T \cdot A$ es definida positiva, obteniendo la factorización de Choleski.

b) Plantear la iteración $X_{n+1} = L_1 \cdot X_n + c$ que se obtiene de aplicar el método de Gauss-Seidel a las ecuaciones normales del sistema $A\mathrm{x} = b$. ¿Será convergente el proceso iterativo a la pseudosolución?

c) Hallar la matriz $H_u = I - \beta uu^T$ de la reflexión que transforma el vector $a = (0, -3, 4)^T$ en el vector $r = (-5, 0, 0)$.

d) Obtener la solución en mínimos cuadrados del sistema $A\mathrm{x} = b$, utilizando el método de Householder, y determinar la norma del error.

e) Sin haber resuelto el apartado anterior, ¿podrían predecirse $H_u A$ y $H_u b$ de las relaciones geométricas entre $L =< u >$, $L^{\perp}$ y los vectores columnas implicados?

Ejercicio 3.14 Se considera el sistema superdeterminado $Ax = b$ con

$$A = \begin{pmatrix} 3 & 2 \\ 4 & 5 \\ 12 & 0 \end{pmatrix} \quad \text{y} \quad b = \begin{pmatrix} 3 \\ 1 \\ 13 \end{pmatrix}$$

a) Calcular la pseudosolución (solución de mínimos cuadrados) así como la norma del error utilizando transformaciones de Householder.

b) Sea $T = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1/12 \end{pmatrix}$ la matriz asociada a la transformación elemental que divide por 12 la tercera de las ecuaciones del sistema:

$$TAx = Tb \iff \begin{pmatrix} 3 & 2 \\ 4 & 5 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 3 \\ 1 \\ 13/12 \end{pmatrix}$$

Calcular su pseudosolución haciendo uso de las ecuaciones normales. Determinar la norma del error.

c) ¿A qué se debe que no coincidan las pseudosoluciones obtenidas en los dos apartados anteriores? ¿Qué habría ocurrido si la matriz T hubiese sido unitaria?

Ejercicio 3.15 Sea el sistema $A\mathrm{x} = b$, donde

$$A = \begin{pmatrix} 3 & -2 \\ 0 & 3 \\ 4 & 4 \end{pmatrix}, \quad \mathrm{x} = \begin{pmatrix} x \\ y \end{pmatrix} \quad \text{y} \quad b = \begin{pmatrix} 2 \\ 0 \\ 1 \end{pmatrix}.$$

a) Probar que la matriz $B = A^T \cdot A$ es definida positiva, obteniendo la factorización de Cholesky $B = G^T \cdot G$.

b) Hacer uso de la factorización obtenida en el apartado anterior para hallar la pseudosolución mediante las ecuaciones normales del sistema. Calcular el número de condición, $\kappa_\infty(B)$, de la matriz B para la norma $||\ ||_\infty$. ¿Hasta que punto se podría considerar fiable la pseudosolución obtenida con aritmética de ordenador?

c) Hallar la matriz de la reflexión (matriz de Householder) H_u que transforma el vector $a = (3, 0, 4)^T$ en el vector $r = (-5, 0, 0)^T$. Una vez determinado el vector u, justificar que se pueden conocer $H_u A$ y $H_u b$ sin necesidad de efectuar los productos.

d) Obtener la solución en mínimos cuadrados del sistema $A\mathrm{x} = b$, utilizando el método de Householder y determinar la norma del error. Operando con el ordenador, ¿puede obtenerse una pseudosolución distinta de la obtenida en el apartado b? Si así ocurriera, ¿puede ser mayor el error?

Ejercicio 3.16 Sea el sistema $A\mathrm{x} = b$, donde

$$A = \begin{pmatrix} 1 & -1 & 2 \\ 0 & 3 & -3 \\ 0 & -4 & 4 \end{pmatrix},\ \mathrm{x} = \begin{pmatrix} x \\ y \\ z \end{pmatrix}\ \text{y}\ \ b = \begin{pmatrix} 0 \\ 1 \\ 2 \end{pmatrix}.$$

a) Hallar $\|A\|_\infty$. ¿Qué se puede decir sobre el número de condición de la matriz A para la norma infinito? ¿Qué estimación daría MATLAB para el número de condición espectral obtenido con el comando `cond(A)`?

b) Utilizar la descomposición LU de la matriz $A^T A$ para resolver el sistema $A^T A\mathrm{x} = A^T b$. ¿Qué propiedad caracteriza a las soluciones en relación al sistema $A\mathrm{x} = b$? Interpreta geométricamente el resultado.

c) Encontrar una matriz ortogonal Q que transforme el vector $a = (0, 3, -4)^T$ en el vector $r = (0, 5, 0)^T$. Obtener la norma del error para las soluciones en mínimos cuadrados del sistema $QA\mathrm{x} = Qb$.

d) ¿Qué relación hay entre las soluciones obtenidas en los apartados anteriores?

Si se obtienen las soluciones en mínimos cuadrados del sistema $A\mathrm{x} = b$, escalonando previamente la matriz A, ¿se debe obtener mismo resultado que en alguno de los apartados anteriores?

e) Probar que la matriz $P = \begin{pmatrix} \frac{2}{3} & \frac{3}{25} & -\frac{4}{25} \\ \frac{1}{3} & \frac{3}{25} & -\frac{4}{25} \\ \frac{1}{3} & 0 & 0 \end{pmatrix}$ es la pseudoinversa de A, verificando las propiedades de Penrose. (Hacer la comprobación sólo con dos de ellas).

De entre todas las soluciones en mínimos cuadrados del sistema $A\mathrm{x} = b$, hallar la de menor norma euclídea.

Ejercicio 3.17

a) En lo que sigue, H_v denota la transformación de Householder asociada al vector v. Sean x, y, v, z vectores no nulos, con $H_v x = y$ y $z \perp v$. Probar que $H_v v = -v$ y $H_v z = z$. Determinar razonadamente todos los vectores w tales que $H_w x = y$.

b) Se considera el sistema de ecuaciones dado por

$$\begin{pmatrix} -\frac{1}{2} & 1 & 0 \\ 1 & 2 & 1 \\ 1 & 0 & -1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 2 \\ -1 \\ -1 \end{pmatrix}$$

b.1) Estudiar el condicionamiento del sistema, utilizando la norma 1.

b.2) Resolver el sistema por medio de transformaciones de Householder.

b.3) Desde un punto de vista numérico, ¿sería razonable resolver el sistema escalonando por Gauss? Razonar la respuesta.

c) Demostrar que el vector $c = (-\frac{4}{3}, \frac{1}{2}, -\frac{4a}{3} - 1)^T$ y la matriz

$$G = \begin{pmatrix} 0 & -\frac{2}{3} & 0 \\ 0 & 0 & -\frac{1}{2} \\ 0 & -\frac{2a}{3} & 0 \end{pmatrix}$$

son los propios del método de Gauss-Seidel asociado al sistema

$$\begin{pmatrix} \frac{3}{2} & 1 & 0 \\ 0 & 2 & 1 \\ a & 0 & -1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} -2 \\ 1 \\ 1 \end{pmatrix}$$

d) Estudiar, en función del parámetro a, el carácter diagonal dominante por filas de la matriz de coeficientes del sistema dado, así como el radio espectral de G. ¿Para qué valores de a es convergente el método anterior?

e) Para $a = 0$ el método resulta convergente. Utilizando aritmética exacta, y toamndo como vector inicial $x_0 = (0, 0, 0)^T$, realizar dos iteraciones, acotando el error cometido. Razonar qué ocurre cuando se itera por tercera vez. ¿Hubiera ocurrido otro tanto al trabajar con aritmética de ordenador?

Ejercicio 3.18 Sea el sistema $A\mathrm{x} = b$, donde

$$A = \begin{pmatrix} 1 & 1 \\ \alpha & 0 \\ -2 & 2 \end{pmatrix}, \quad \mathrm{x} = \begin{pmatrix} x \\ y \end{pmatrix}, \quad b = \begin{pmatrix} 2 \\ \beta \\ \gamma \end{pmatrix}, \text{ con } \alpha > 0 \text{ y } \beta, \gamma \in \mathbb{R}.$$

a) Hallar α sabiendo que que existe una matriz de Householder, H_v, que transforma la primera columna de la matriz A en el vector $r = (3, 0, 0)^T$. ¿Quién es H_v?

b) Determinar el conjunto de vectores b para los que se verifica $H_v b = b$, siendo H_v la matriz del apartado anterior. Encontrar, entre ellos, el que tiene menor norma euclídea.

c) Hallar la pseudosolución del sistema $A\mathrm{x} = b_m$, para $\alpha = 2$ y $b_m = (2, 1, -1)^T$, utilizando transformaciones ortogonales para determinar el error.

d) Probar que si una matriz real B tiene sus columnas linealmente independientes, entonces $B^T B$ es definida positiva.

e) Sea el sistema $A^T A\,\mathrm{x} = A^T b_m$, con α y b_m como en el apartado (c).

e.1) ¿Sería posible utilizar una descomposición $A^T A = GG^T$, con G triangular inferior, para resolver el sistema?

e.2) Utilizando la norma $\| \ \|_\infty$ para medir el condicionamiento, ¿es un sistema mal condicionado para utilizar aritmética de ordenador en su resolución?

e.3) Sea $(s_0, s_1, s_2, \ldots)$ la sucesión que se obtiene al aplicar el método de Gauss-Seidel al sistema, con $s_0 = (0, 0)^T$. Probar que, operando en aritmética exacta, la sucesión (s_n) es convergente y obtener su límite s.

Ejercicio 3.19 Se considera el sistema $A\mathrm{x} = b$ con

$$A = \begin{pmatrix} 0 & 5 \\ 3 & 0 \\ 4 & 0 \end{pmatrix}, \quad \mathrm{x} = \begin{pmatrix} x \\ y \end{pmatrix} \quad \mathrm{y} \quad b = \begin{pmatrix} 5 \\ 2 \\ 11 \end{pmatrix}$$

a) ¿Existe alguna transformación de Householder que permute las columnas de la matriz A? Justificar la respuesta.

b) Calcular la pseudosolución del sistema mediante transformaciones de Householder dando la norma del vector error.

c) Calcular la inversa generalizada A^+ de la matriz A a través de su descomposición en valores singulares y hacer uso de ella para encontrar la pseudosolución del sistema $A\mathrm{x} = b$ dando la norma del vector error.

d) ¿Hubiésemos podido, en éste caso, calcular la inversa generalizada sin necesidad de realizar su descomposición en valores singulares?

Ejercicio 3.20 Se considera el sistema $A\mathrm{x} = b$ con

$$A = \begin{pmatrix} 1 & 2 \\ 4 & 8 \\ -1 & -2 \end{pmatrix}, \quad \mathrm{x} = \begin{pmatrix} x \\ y \end{pmatrix} \quad \mathrm{y} \quad b = \begin{pmatrix} 1 \\ 5 \\ 3 \end{pmatrix}$$

Determinar la pseudosolución del sistema dando la norma del error:

a) Mediante transformaciones de Householder.

b) A través de la inversa generalizada de la matriz A

Ejercicio 3.21 Hallar la pseudosolución del sistema $A\mathrm{x} = b$ en el que

$$A = \begin{pmatrix} 3 & -4 \\ 4 & 3 \\ 0 & 12 \end{pmatrix} \quad \mathrm{y} \quad b = \begin{pmatrix} 65 \\ -65 \\ 0 \end{pmatrix}$$

así como la norma del error a través de la pseudoinversa de la matriz A calculada mediante la descomposición en valores singulares.

Ejercicio 3.22 Se considera el sistema superdeterminado $A\mathrm{x} = b$ con

$$A = \begin{pmatrix} 2 & 1 \\ 2 & 0 \\ 1 & -2 \\ 0 & 2 \end{pmatrix} \quad \mathrm{x} = \begin{pmatrix} x \\ y \end{pmatrix} \quad \text{y} \quad b = \begin{pmatrix} 3 \\ 6 \\ 0 \\ 3 \end{pmatrix}$$

a) Encontrar una transformación de Householder que transforme la primera columna de la matriz A en el vector $r = (3, 0, 0, 0)^T$.

b) Probar que el producto de dos matrices de Householder es una matriz unitaria.

   Hallar una matriz ortogonal Q tal que $A = QR$ siendo R una matriz triangular superior de las mismas dimensiones que A.

c) Probar que si Q es ortogonal, los sistemas $A\mathrm{x} = b$ y $Q^T A\mathrm{x} = Q^T b$ tienen las mismas soluciones en mínimos cuadrados.

   Hallar el error cometido al obtener la pseudosolución del sistema $A\mathrm{x} = b$, utilizando transformaciones ortogonales.

d) Teniendo en cuenta el rango de la matriz A, calcular el vector $s = A^+ b$ donde A^+ representa la pseudoinversa de la matriz A.

e) Sea $x_{n+1} = L_1 x_n + c$ la sucesión resultante de aplicar el método de Gauss-Seidel a la resolución de las ecuaciones normales del sistema $A\mathrm{x} = b$. ¿Cuántas iteraciones son necesarias para la convergencia del método? Determina la pseudosolución así como la norma del error.

Ejercicio 3.23 El equipo *Astronomía para aficionados*, adquirido por el profesor Dana este verano, permitía determinar el plano $\Pi \equiv \alpha x + \beta y + \gamma z = 1$ donde se encuentra la trayectoria de Marte alrededor del Sol. En las instrucciones indicaba introducir en el "calculador mágico" una serie de coordenadas locales (x_i, y_i, z_i), obtenidas con el "telescopio marciano", y automáticamente proporcionaría los coeficientes α, β, γ. Entre otras cosas, sugería introducir entre 5 y 10 coordenadas para que el ajuste obtenido **en el sentido de los mínimos cuadrados** promediara "científicamente" los errores de observación...

a) Plantear el sistema superdeterminado, $A\boldsymbol{\alpha}= b$, con $\boldsymbol{\alpha}=(\alpha,\beta,\gamma)^T$, para determinar el plano Π, cuando las coordenadas locales son $(2,1,0)$, $(-1,2,1)$, $(0,1,2)$, $(-1,0,1)$, $(0,1,0)$. ¿Puede ser nulo el error cometido para la pseudosolución del sistema?

b) Poniendo $A=[a_1\ a_2\ a_3]$, donde a_i indica la correspondiente columna de A, razonar si es posible encontrar una transformación de Householder que transforme a_1 en a_2. Hallar una matriz unitaria, Q, de modo que $Qa_1 = a_3$.

c) Obtener las ecuaciones normales, $B\boldsymbol{\alpha}= c$, del sistema inicial $A\boldsymbol{\alpha}= b$. ¿Está la matriz B mal condicionada para la norma $||\ ||_\infty$?

d) Probar que los métodos iterados de Jacobi y Gauss-Seidel aplicados al sistema $B\boldsymbol{\alpha}= c$ son convergentes. ¿Cuál de ellos converge más rápido?

e) Partiendo de $\boldsymbol{\alpha_0}= (0,0,0)^T$, obtener la aproximación $\boldsymbol{\alpha_3}$, al aplicar 3 pasos del método de Gauss-Seidel al sistema $B\boldsymbol{\alpha}= c$, operando con dos cifras decimales. ¿Cuál es el error obtenido al tomar $\boldsymbol{\alpha_3}$ como la solución en mínimos cuadrados de $A\boldsymbol{\alpha}= b$?

Ejercicio 3.24 Dada la matriz $A = \begin{pmatrix} 1 & 5 & 5 \\ 1 & 2 & 1 \\ 1 & 2 & 3 \end{pmatrix}$, se pide:

a) Estudiar si admite factorizaciones LU y/o de Cholesky.

b) Utilizar dichas factorizaciones (en caso de existir) para resolver el sistema $A\mathrm{x} = b$ con $\mathrm{x} = \begin{pmatrix} x \\ y \\ z \end{pmatrix}$ y $b = \begin{pmatrix} 3 \\ 2 \\ 1 \end{pmatrix}$.

c) Resolver, mediante transformaciones de Householder el sistema superdeterminado resultante de añadir a nuestro sistema la ecuación $x+y+3z = \alpha$. Hallar la norma del error.

d) ¿Se puede calcular el valor de α que minimiza la norma del error sin resolver el sistema anterior? Justifíquese la respuesta.

e) Comenzando por el vector $v_0 = (-1,\ -1,\ 2)^T$, realizar dos iteraciones del método de la potencia simple y utilizar el resultado para obtener una aproximación del autovalor dominante de la matriz A.

f) Suponiendo que la aproximación obtenida en el apartado anterior es una buena aproximación, razonar si el método iterado definido por

$$\mathrm{x}_n = A\,\mathrm{x}_{n-1} + c$$

donde c es un vector de $\mathbf{R}^3$, resultaría convergente.

4. Autovalores y autovectores

4.1 Conceptos básicos

Una variedad lineal V de $\mathbf{K}^n$ se dice que es *invariante* mediante una aplicación A si cualquier vector $x \in V$ se transforma, mediante A, en otro vector de V, es decir, si $Ax \in V$ para cualquier vector x de V.

Las variedades invariantes más simples son aquellas que tienen dimensión 1, por lo que una base está constituida por un único vector $v \neq 0$. Podemos observar que en ese caso cualquier vector $x \in V$ puede expresarse de la forma $x = \alpha v$ con $\alpha \neq 0$ si $x \neq 0$ y que, por tratarse de una variedad invariante, ha de ser $Ax = \beta v$, pero entonces:

$$Ax = A\alpha v = \alpha Av = \beta v \Longrightarrow Av = \frac{\beta}{\alpha} v = \lambda v$$

Las variedades invariantes de dimensión 1 vienen determinadas por un vector $v \neq 0$ tal que $Av = \lambda v$. Estos vectores reciben en nombre de *autovectores* o *vectores propios* de la matriz A y los correspondientes valores de λ reciben el nombre de *autovalores* o *valores propios* de A.

Es obvio que si A posee un autovector es porque existe un vector $v \neq 0$ tal que $Av = \lambda v$, o lo que es lo mismo, tal que $(\lambda I - A)v = 0$. Por tanto, el sistema $(\lambda I - A)x = 0$ es compatible y además indeterminado, ya que si $v \neq 0$ es solución del sistema, cualquier vector proporcional a él también lo es. Se verifica entonces que rg $(\lambda I - A) < n$ (donde n representa el orden de la matriz) y, por tanto $\det(\lambda I - A) = 0$.

Al polinomio $p(\lambda) = \det(\lambda I - A)$ se le denomina *polinomio característico* de la matriz A y a la ecuación $p(\lambda) = 0$ *ecuación característica* .

Nótese que si λ es una raíz de la ecuación característica de la matriz A, existe un autovector v asociado al autovalor λ. Por tanto, desde el punto de vista

teórico, el problema del cálculo de los autovectores de una matriz se reduce a la resolución de los sistemas $(\lambda_i I - A)x = 0$ obtenidos para las diferentes raíces λ_i de la ecuación característica.

Sea A una matriz cuadrada de orden n y supongamos que existen $V_1, V_2, \ldots, V_k$ subespacios invariantes tales que $\mathbf{K}^n = V_1 \oplus V_2 \oplus \cdots \oplus V_k$.

Si $\{x_{i1}, x_{i2}, \ldots, x_{in_i}\}$ es una base de V_i se tiene que

$$\mathcal{B} = \{x_{11}, \ldots, x_{1n_1}, x_{21}, \ldots, x_{2n_2}, \ldots, x_{k1}, \ldots, x_{kn_k}\}$$

constituye una base de $\mathbf{K}^n$. Si P es la matriz del cambio de base de la base canónica a la base $\mathcal{B}$, se tiene que

$$P^{-1}AP = \begin{pmatrix} J_1 & \Theta & \cdots & \Theta \\ \Theta & J_2 & \cdots & \Theta \\ \vdots & \vdots & \ddots & \vdots \\ \Theta & \Theta & \cdots & J_k \end{pmatrix}$$

donde cada J_i es una caja cuadrada de dimensión $r_i = \dim(V_i)$.

La justificación es que $Ax_{ij} = (0, \ldots, 0, \alpha_{i1}, \ldots, \alpha_{in_i}, 0, \ldots, 0)_B^T$ con $1 \leq i \leq k$, $1 \leq j \leq n_i$ y, por tanto, puede verse fácilmente que

$$AP = P \begin{pmatrix} J_1 & \Theta & \cdots & \Theta \\ \Theta & J_2 & \cdots & \Theta \\ \vdots & \vdots & \ddots & \vdots \\ \Theta & \Theta & \cdots & J_k \end{pmatrix}$$

Proposición 4.1 *Autovectores asociados a autovalores diferentes son linealmente independientes*

Demostración. Sean u y v dos autovectores de la matriz A asociados a los autovalores λ y μ respectivamente con $\lambda \neq \mu$.

Si u y v fuesen linealmente dependientes se verificaría que $u = \alpha v$ con $\alpha \neq 0$. Entonces:

$$\lambda u = Au = A(\alpha v) = \alpha A v = \alpha \mu v = \mu(\alpha v) = \mu u$$

y por tanto $(\lambda - \mu)u = 0$, pero dado que $u \neq 0$ se tiene que $\lambda = \mu$ en contra de la hipótesis de que $\lambda \neq \mu$, lo que prueba el resultado. ■

4.2 Método interpolatorio para la obtención del polinomio característico

El problema del cálculo de los autovectores de una matriz, una vez calculados los autovalores, se reduce a la resolución de un sistema de ecuaciones por cada autovalor.

Trataremos, por tanto y, en primer lugar, calcular sus autovalores a partir de la propiedad de ser las raíces del polinomio característico de la matriz. Es decir, dada una matriz cuadrada A de orden n pretendemos calcular su polinomio característico $P(\lambda) = \det(\lambda I - A)$ para, posteriormente, hallar sus raíces mediante alguno de los métodos de resolución de ecuaciones estudiados.

El método interpolatorio consiste en calcular el polinomio característico de la matriz dando n valores a λ, para calcular n determinantes y, posteriormente, resolver el sistema de n ecuaciones con n incógnitas resultante. Veámoslo con detenimiento:

Sea $P(\lambda) = \det(\lambda I - A) = \lambda^n + a_1\lambda^{n-1} + \cdots + a_{n-2}\lambda^2 + a_{n-1}\lambda + a_n$. Dando a λ n valores $\lambda_1, \lambda_2, \ldots, \lambda_n$ se tiene que:

$$\begin{aligned}
\lambda = \lambda_1 \quad &\Longrightarrow \quad \lambda_1^n + a_1\lambda_1^{n-1} + \cdots + a_{n-2}\lambda_1^2 + a_{n-1}\lambda_1 + a_n = \det(\lambda_1 I - A)\\
\lambda = \lambda_2 \quad &\Longrightarrow \quad \lambda_2^n + a_1\lambda_2^{n-1} + \cdots + a_{n-2}\lambda_2^2 + a_{n-1}\lambda_2 + a_n = \det(\lambda_2 I - A)\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\
\lambda = \lambda_n \quad &\Longrightarrow \quad \lambda_n^n + a_1\lambda_n^{n-1} + \cdots + a_{n-2}\lambda_n^2 + a_{n-1}\lambda_n + a_n = \det(\lambda_n I - A)
\end{aligned}$$

Ejemplo 4.1 Dada la matriz $A = \begin{pmatrix} 1 & 3 & 2 & 5 \\ 4 & 2 & -1 & 0 \\ 0 & 1 & 0 & 1 \\ 2 & -2 & -1 & 1 \end{pmatrix}$, su polinomio característico es de grado cuatro $P(\lambda) = \lambda^4 + a_1\lambda^3 + a_2\lambda^2 + a_3\lambda + a_4$, por lo que vamos a dar a λ cuatro valores:

$$\begin{aligned}
\lambda = 0 \quad &\Longrightarrow \quad a_4 = \det(-A) = 41\\
\lambda = 1 \quad &\Longrightarrow \quad 1 + a_1 + a_2 + a_3 + a_4 = \det(I - A) = 74\\
\lambda = -1 \quad &\Longrightarrow \quad 1 - a_1 + a_2 - a_3 + a_4 = \det(-I - A) = -20\\
\lambda = 2 \quad &\Longrightarrow \quad 16 + 8a_1 + 4a_2 + 2a_3 + a_4 = \det(2I - A) = 67
\end{aligned}$$

dado que $a_4 = 41$ el sistema se reduce a

$$\left.\begin{array}{rrrr} a_1 + & a_2 + & a_3 = & 32 \\ -a_1 + & a_2 - & a_3 = & -62 \\ 8a_1 + & 4a_2 + & 2a_3 = & 10 \end{array}\right\} \Longrightarrow \; a_1 = -4,\; a_2 = -15 \text{ y } a_3 = 51$$

por lo que su polinomio característico es

$$P(\lambda) = \lambda^4 - 4\lambda^3 - 15\lambda^2 + 51\lambda + 41$$

□

4.3 Sensibilidad de los autovalores a las transformaciones de semejanza

Consideremos la matriz $A = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & \cdots & 0 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 1 \\ 0 & 0 & 0 & \cdots & 0 & 0 \end{pmatrix}$, que es una caja de Jordan y, por tanto, no diagonalizable, cuyo polinomio característico es λ^n.

Si introducimos en la matriz una perturbación y en vez de la matriz A tomamos la matriz $B = A + E$ con $E = \begin{pmatrix} 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 0 \\ \varepsilon & 0 & 0 & \cdots & 0 & 0 \end{pmatrix}$ siendo $\varepsilon > 0$ muy pequeño (incluso más pequeño que la resolución del ordenador) se obtiene como polinomio característico $\lambda^n + (-1)^n\varepsilon$ que posee n raíces distintas (las raíces n-ésimas de $(-1)^{n-1}\varepsilon$), por lo que la matriz resulta ser diagonalizable.

Sea A una matriz diagonalizable y B una perturbación de dicha matriz ($B = A + E$). Sean λ_i los autovalores de la matriz A y sea μ uno cualquiera de los autovalores de la matriz B con $\lambda_i \neq \mu$.

Teorema 4.2 *En las condiciones anteriores se tiene que*

$$\min_i |\mu - \lambda_i| \leq \|P\| \left\|P^{-1}\right\| \|E\|$$

Demostración. Por ser A diagonalizable existe P tal que $P^{-1}AP = D$. Sea x un autovector de B asociado a μ. Entonces $Bx = \mu x$ o lo que es lo mismo:

$(A+E)x = \mu x \iff (\mu I - A)x = Ex \iff (\mu I - PDP^{-1})x = Ex \iff$

$P(\mu I - D)P^{-1}x = Ex \iff (\mu I - D)(P^{-1}x) = P^{-1}Ex = P^{-1}EPP^{-1}x$

Supuesto que $\mu \neq \lambda_i$ cualquiera que sea $i = 1, 2, \ldots, n$, la matriz $\mu I - D$ es regular, por lo que

$P^{-1}x = (\mu I - D)^{-1}(P^{-1}EP)P^{-1}x$, por lo que para cualquier norma multiplicativa se tiene

$$\|P^{-1}x\| \leq \|(\mu I - D)^{-1}\| \, \|P^{-1}EP\| \, \|P^{-1}x\| \iff$$

$$1 \leq \left\|(\mu I - D)^{-1}\right\| \left\|P^{-1}EP\right\|$$

Dado que $(\mu I - D)^{-1} = \operatorname{diag}\left(\dfrac{1}{\mu - \lambda_1}, \ldots, \dfrac{1}{\mu - \lambda_n}\right)$ tenemos que

$$1 \leq \max_i \left\{\left|\frac{1}{\mu - \lambda_i}\right|\right\} \left\|P^{-1}\right\| \|E\| \, \|P\|$$

por lo que

$$\min_i |\mu - \lambda_i| \leq \left\|P^{-1}\right\| \|P\| \, \|E\|$$

Obsérvese que la perturbación cometida en los autovalores depende de la matriz de paso P y dado que esta no es única, se trata de elegir, entre todas las posibles matrices de paso, aquella que verifique que $\|P^{-1}\| \, \|P\|$ sea mínima. Es decir, aquella cuyo número de condición sea lo menor posible.

Corolario 4.3 *Si A es unitariamente diagonalizable (diagonalizable mediante una matriz de paso unitaria), la perturbación producida en los autovalores es menor o igual que la perturbación producida en la matriz.*

$$\min_i |\mu - \lambda_i| \leq \|E\|$$

Por tanto, las mejores matrices para el cálculo efectivo de sus autovalores y autovectores son las diagonalizables unitariamente y éstas reciben el nombre de *matrices normales*.

Es necesario aclarar que si una matriz A no es normal y le calculamos sus autovalores, estos pueden reflejar, con la exactitud que se desee, sus verdaderos valores, pero estos no van a representar la solución del problema que generó

dicha matriz, pues los elementos de A pueden arrastrar errores (redondeo, medición, etc.) E y estos hacen que los autovalores de las matrices A y $A+E$ puedan diferir bastante.

Por otra parte, es evidente que no podemos estudiar si una matriz es normal calculando sus autovalores y autovectores para comprobar que se puede diagonalizar por semejanza mediante una matriz de paso unitaria (matriz constituida por una base ortonormal de autovectores). Es necesario, por tanto, encontrar otros métodos que detecten si una matriz es, o no es, normal.

Teorema 4.4 *Sea T una matriz triangular. Si $T^*T = TT^*$ entonces T es diagonal.*

Demostración. Probaremos que se trata de una matriz diagonal por inducción en el orden de la matriz.

Para $n = 1$ es obvio. Para $n = 2$ es $T = \begin{pmatrix} a & b \\ 0 & c \end{pmatrix}$ y $T^* = \begin{pmatrix} \overline{a} & 0 \\ \overline{b} & \overline{c} \end{pmatrix}$.

$$T^*T = \begin{pmatrix} \overline{a} & 0 \\ \overline{b} & \overline{c} \end{pmatrix} \begin{pmatrix} a & b \\ 0 & c \end{pmatrix} = \begin{pmatrix} |a|^2 & \overline{a}b \\ a\overline{b} & |b|^2 + |c|^2 \end{pmatrix}$$

$$TT^* = \begin{pmatrix} a & b \\ 0 & c \end{pmatrix} \begin{pmatrix} \overline{a} & 0 \\ \overline{b} & \overline{c} \end{pmatrix} = \begin{pmatrix} |a|^2 + |b|^2 & b\overline{c} \\ \overline{b}c & |c|^2 \end{pmatrix}$$

Dado que se trata de una matriz normal es $T^*T = TT^*$ por lo que igualando ambas matrices se obtiene que $|b|^2 = 0$ es decir $b = 0$ y, por tanto, T es diagonal.

Supongamos ahora que el teorema es cierto para cualquier matriz triangular y normal de orden n y vamos a probarlo para otra de orden $n+1$.

$$\text{Sean } T = \left(\begin{array}{c|ccc} a_1 & a_2 & \cdots & a_{n+1} \\ \hline 0 & & & \\ \vdots & & T_n & \\ 0 & & & \end{array}\right) \text{ y } T^* = \left(\begin{array}{c|ccc} \overline{a_1} & 0 & \cdots & 0 \\ \hline \overline{a_2} & & & \\ \vdots & & T_n^* & \\ \overline{a_n} & & & \end{array}\right)$$

$$TT^* = \left(\begin{array}{c|c} \displaystyle\sum_{i=1}^{n} |a_i|^2 & \\ \hline & T_nT_n^* \end{array}\right) \qquad T^*T = \left(\begin{array}{c|c} |a_1|^2 & \\ \hline & T_n^*T \end{array}\right)$$

De la igualación de ambas obtenemos que $|a_2|^2 + |a_3|^2 + \cdots + |a_{n+1}|^2 = 0$, por lo

que $a_2 = a_3 = \cdots = a_{n+1} = 0$. Como, además, es $T_n T_n^* = T_n^* T_n$, por hipótesis de inducción sabemos que T_n es diagonal y, por tanto, T es diagonal.

Teorema 4.5 Teorema de Schur
Cualquier matriz cuadrada A es unitariamente semejante a una triangular superior T. Es decir, existe una unitaria U ($U^*U = UU^* = I$) tal que

$$U^*AU = T.$$

Demostración. Por inducción en el orden de A. Si $n = 1$ e obvio. Supuesto cierto para n vamos a probarlo para una matriz de orden $n + 1$.

Sea $A \in \mathbf{K}^{(n+1)\times(n+1)}$. Sea λ un autovalor de A y x un autovector unitario ($\|x\| = 1$) asociado a λ.

Consideremos la matriz $P = (x\ e_2 \cdots e_n)$ en la que sus columnas $x, e_2, \ldots, e_n$ constituyen una base ortonormal de $\mathbf{K}^{n+1}$

$$P^*AP = \left(\begin{array}{c|c} \lambda & \alpha_{ij} \\ \hline \Theta & A_n \end{array}\right)$$

Sea $Q = \left(\begin{array}{c|c} 1 & \Theta \\ \hline \Theta & U_n \end{array}\right)$ en donde U_n es la matriz unitaria que, por hipótesis de inducción, verifica que $U_n^* A_n U_n =$ Triangular superior.

Si consideremos la matriz $U = PQ$ es fácil comprobar que $U^*AU = Q^*P^*APQ$ es una triangular superior.

Teorema 4.6 *Una matriz A es normal si, y sólo si, $AA^* = A^*A$.*

Demostración. Supongamos que $AA^* = A^*A$. Por el Teorema 4.5 sabemos que existe una matriz unitaria U tal que $U^*AU = T$ con T triangular superior. Entonces $A = UTU^*$, por lo que

$$\left.\begin{array}{l} AA^* = (UTU^*)(UTU^*)^* = UTU^*UT^*U^* = UTT^*U^* \\ A^*A = (UTU^*)^*(UTU^*) = UT^*U^*UTU^* = UT^*TU^* \end{array}\right\} \Longrightarrow$$

$$UTT^*U^* = UT^*TU^*$$

es decir, $TT^* = T^*T$ por lo que T es una matriz normal y triangular, por lo que el Teorema 4.4 nos asegura que es diagonal y, por tanto, A es diagonalizable unitariamente, es decir, es normal.

Recíprocamente, si A es unitariamente diagonalizable (normal) existe una matriz unitaria U tal que $U^*AU = D$ o lo que es lo mismo, $A = UDU^*$.

$$AA^* = UDU^*(UDU^*)^* = UDD^*U^* = U\,\mathrm{diag}(|d_1|^2, \ldots, |d_n|^2)U^*$$
$$A^*A = (UDU^*)^*UDU^* = UD^*DU^* = U\,\mathrm{diag}(|d_1|^2, \ldots, |d_n|^2)U^*$$

es decir, $AA^* = A^*A$.

Hemos visto que para que los autovalores de una matriz A reflejen la solución del problema que la generó, ésta ha de ser normal. Si no podemos plantear nuestro problema mediante una matriz normal lo que debemos hacer es tratar de plantearlo mediante una matriz "*lo más normal posible*". Vamos a estudiar entonces el *condicionamiento* del problema.

Por el Teorema 4.5 cualquier matriz cuadrada A es unitariamente semejante a una triangular superior T donde

$$T = \begin{pmatrix} t_{11} & t_{12} & \cdots & t_{1n} \\ 0 & t_{22} & \cdots & t_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & t_{nn} \end{pmatrix} = \begin{pmatrix} t_{11} & 0 & \cdots & 0 \\ 0 & t_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & t_{nn} \end{pmatrix} + \begin{pmatrix} 0 & t_{12} & \cdots & t_{1n} \\ 0 & 0 & \cdots & t_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{pmatrix}$$

Es decir, $T = D + M$ y, obviamente, A es normal si, y sólo si, $M = \Theta$.

En cierta manera (a $\|M\|$ se le llama *desviación de la normalidad de* A) la matriz M y más concretamente su norma, mide el condicionamiento del problema.

Se tiene, además, que $\|M\|_2$ es constante (no depende de U y de T sino sólo de A), ya que

$$\|M\|_2^2 = \|T\|_2^2 - \|D\|_2^2 = \|U^*AU\|_2^2 - \|D\|_2^2 = \|A\|_2^2 - \sum_{i=1}^{n} |\lambda_i|^2 \Longrightarrow$$

$$\|M\|_2 = \sqrt{\|A\|_2^2 - \sum_{i=1}^{n} |\lambda_i|^2}$$

Definición 4.1 Se denomina *matriz conmutatriz* de una matriz cuadrada A a la matriz $C(A) = A^*A - AA^*$.

Otra forma de medir el condicionamiento de una matriz es considerar la norma euclídea de la matriz conmutatriz

$$\|C(A)\|_2 = \|AA^* - A^*A\|_2$$

obteniéndose el siguiente resultado:

Teorema 4.7 *Dada una matriz $A \in \mathbf{K}^{n\times n}$ se verifica:*

a) $\dfrac{\|C(A)\|_2^2}{6\,\|A\|_2^2} \leq \|M\|_2^2$ *(Desigualdad de Eberlein).*

b) $\|M\|_2^2 \leq \sqrt{\dfrac{n^3-3}{12}}\,\|C(A)\|_2^2$ *(Desigualdad de Heurici).*

Los autovalores de una matriz son, en general, números complejos. Sin embargo, las matrices hermíticas (en el caso real, las simétricas) tienen todos sus autovalores reales como prueba el siguiente teorema.

Teorema 4.8 *Los autovalores de una matriz hermítica son todos reales y autovectores correspondientes a dos autovalores diferentes son ortogonales.*

Demostración. Sea A una matriz hermítica, es decir, una matriz tal que $A^* = A$ y sea λ un autovalor de A asociado al autovector x. Se verifica entonces que $Ax = \lambda x$.

Multiplicando la expresión anterior, por la izquierda por x^* obtenemos que

$$x^*Ax = x^*\lambda x = \lambda x^*x \tag{4.1}$$

Trasponiendo y conjugando la expresión 4.1 obtenemos

$$(x^*Ax)^* = (\lambda x^*x)^* \implies x^*A^*x = \bar{\lambda}x^*x \implies x^*Ax = \bar{\lambda}x^*x \tag{4.2}$$

Si comparamos las expresiones (4.1) y (4.2) obtenemos que

$$\lambda x^*x = \bar{\lambda}x^*x$$

y dado que x es un autovector (un vector no nulo) sabemos que $x^*x = \|x\|^2 \neq 0$, por lo que podemos dividir por x^*x para obtener que $\lambda = \bar{\lambda}$, es decir, $\lambda \in \mathbf{R}$.

Por otra parte, si x e y son autovectores asociados a dos autovalores $\lambda \neq \mu$ de una matriz hermítica A se verifica que

$$Ax = \lambda x \implies (Ax)^* = (\lambda x)^* \implies x^*A = \lambda x^*$$

$$x^*Ay = \lambda x^*y \implies x^*\mu y = \lambda x^*y \implies \mu x^*y = \lambda x^*y \implies (\lambda-\mu)x^*y = 0$$

y dado que $\lambda \neq \mu \implies \lambda - \mu \neq 0$ obtenemos que

$$x^*y = 0 \iff x \perp y$$

■

Teorema 4.9 [Teorema espectral para matrices hermíticas] *Sea A una matriz hermítica de orden n. Existe una base de $\mathbf{C}^n$ constituida por autovectores de A. Dicho de otra forma, toda matriz hermítica es diagonalizable por semejanza (es normal).*

4.4 Métodos iterados para la obtención de autovalores y autovectores

Los métodos que veremos a continuación tratan de encontrar una sucesión convergente cuyo límite nos permitirá conocer los autovalores y autovectores de una matriz dada.

El primero que estudiaremos consiste en buscar una sucesión de vectores cuyo límite es un autovector de la matriz dada, pero dado que en la práctica no podemos calcular el límite y debemos detenernos en un determinado término de la sucesión, lo único que podremos determinar es una aproximación de dicho autovector.

Vamos a comenzar, por ello, a estudiar cómo podemos aproximar el autovalor correspondiente a un determinado autovector cuando lo que conocemos es una aproximación de éste.

4.4.1 Cociente de Rayleigh

Si nos limitamos a calcular la sucesión (z_n) hasta obtener una aproximación x adecuada del autovector podemos obtener el autovalor resolviendo el sistema vectorial $\lambda\, x = Ax$. Este sistema resulta, en general, incompatible por no ser x exactamente un autovector sino sólo una aproximación, por lo que la solución que mejor se ajusta es la pseudosolución del sistema, que nos dará una aproximación del autovalor λ.

$$\lambda\, x^* x = x^* Ax \Longrightarrow \lambda = \frac{x^* Ax}{x^* x}$$

Al cociente $\dfrac{x^* Ax}{x^* x}$ se le denomina *cociente de Rayleigh* del vector x respecto de la matriz A.

Podemos, por tanto, obtener una aproximación del autovector por un método iterado para más tarde aproximar el autovalor mediante el cociente de Rayleigh.

4.4.2 Método de la potencia simple y variantes

Sea $A \in \mathbf{R}^{n\times n}$ una matriz diagonalizable y supongamos que sus autovalores verifican que:

$$|\lambda_1| > |\lambda_2| \geq \cdots \geq |\lambda_n|$$

Sea $\mathcal{B} = \{x_1, x_2, \cdots, x_n\}$ una base de $\mathbf{R}^n$ formada por autovectores asociados a $\lambda_1, \lambda_2, \ldots, \lambda_n$ respectivamente. Se verifica entonces que

$$A^2 x_i = A(Ax_i) = A(\lambda_i x_i) = \lambda_i A x_i = \lambda_i(\lambda_i x_i) = \lambda_i^2 x_i$$

por lo que es fácil probar, por inducción, que

$$A^k x_i = \lambda_i^k x_i \quad \text{para cualquier} \quad i = 1, 2, \ldots, n$$

Dado un vector $z_0 \in \mathbf{R}^n$ se define, a partir de él, la sucesión (z_n) con

$$z_n = Az_{n-1} = A^2 z_{n-2} = \cdots = A^n z_0.$$

Si las coordenadas del vector z_0 respecto de la base $\mathcal{B}$ son $(\alpha_1, \alpha_2, \ldots, \alpha_n)$ se tiene que $z_0 = \alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_n x_n$, por lo que

$$z_k = A^k z_0 = A^k(\alpha_1 x_1 + \cdots + \alpha_n x_n) = \alpha_1 A^k x_1 + \cdots + \alpha_n A^k x_n =$$
$$= \lambda_1^k \alpha_1 x_1 + \cdots + \lambda_n^k \alpha_k x_k = \lambda_1^k \left[\alpha_1 x_1 + \sum_{i=2}^{n} \left(\frac{\lambda_i}{\lambda_1}\right)^k \alpha_i x_i\right]$$

Dado que $|\lambda_1| > |\lambda_i|$ se tiene que $\lim_{k\to\infty} \left(\frac{\lambda_i}{\lambda_1}\right)^k = 0 \quad \forall i = 2, 3, \ldots, n.$

Se verifica entonces que $\lim_{k\to\infty} \frac{z_k}{\lambda_1^k} = \alpha_1 x_1$ que es un autovector de A asociado al autovalor λ_1.

Si k es suficientemente grande, se tiene

$$Az_k = z_{k+1} \approx \lambda_1^{k+1} \alpha_1 x_1 = \lambda_1(\lambda_1^k \alpha_1 x_1) = \lambda_1 z_k$$

por lo que la sucesión (z_n) nos proporciona un método para aproximar el autovalor λ_1.

Nota: Para evitar que las coordenadas de los vectores z_k sean demasiado grandes se considera la sucesión formada por los vectores $z_n = Aw_{n-1}$ donde $w_n = \dfrac{z_n}{\|z_n\|_\infty}$.

Ejemplo 4.2 Para calcular, por el método de la potencia simple, el autovalor de mayor valor absoluto de la matriz $A = \begin{pmatrix} 6 & 2 & 5 \\ 2 & 2 & 3 \\ 5 & 3 & 6 \end{pmatrix}$, partiendo del vector $z_0 = (1\ 1\ 1)^T$ obtenemos:

$$w_0 = \begin{pmatrix} 1'0000 \\ 1'0000 \\ 1'0000 \end{pmatrix} \quad z_1 = \begin{pmatrix} 13'0000 \\ 7'0000 \\ 14'0000 \end{pmatrix} \quad w_1 = \begin{pmatrix} 0'9286 \\ 0'5000 \\ 1'0000 \end{pmatrix} \quad z_2 = \begin{pmatrix} 11'5714 \\ 5'8571 \\ 12'1429 \end{pmatrix}$$

$$w_2 = \begin{pmatrix} 0'9529 \\ 0'4824 \\ 1'0000 \end{pmatrix} \quad z_3 = \begin{pmatrix} 11'6824 \\ 5'8706 \\ 12'2118 \end{pmatrix} \quad w_3 = \begin{pmatrix} 0'9566 \\ 0'4807 \\ 1'0000 \end{pmatrix} \quad z_4 = \begin{pmatrix} 11'7013 \\ 5'8748 \\ 12'2254 \end{pmatrix}$$

$$w_4 = \begin{pmatrix} 0'9571 \\ 0'4805 \\ 1'0000 \end{pmatrix} \quad z_5 = \begin{pmatrix} 11'7039 \\ 5'8753 \\ 12'2273 \end{pmatrix} \quad w_5 = \begin{pmatrix} 0'9572 \\ 0'4805 \\ 1'0000 \end{pmatrix} \quad z_6 = \begin{pmatrix} 11'7042 \\ 5'8754 \\ 12'2275 \end{pmatrix}$$

$$w_6 = \begin{pmatrix} 0'9572 \\ 0'4805 \\ 1'0000 \end{pmatrix} \quad z_7 = \begin{pmatrix} 11'7042 \\ 5'8754 \\ 12'2275 \end{pmatrix}$$

Tenemos, por tanto, que z_7 es una aproximación del autovector asociado al autovalor dominante (el de mayor valor absoluto).

El autovalor asociado se determinará mediante el cociente de Rayleigh

$$\lambda_1 = \frac{z_7^T A z_7}{z_7^T z_7} = 12.22753579693696.$$

□

Este método sólo nos permite calcular, en caso de existir, el autovalor dominante (mayor valor absoluto) de una matriz. Existen sin embargo otras variantes del método de la potencia simple que nos permiten calcular cualquiera de sus autovalores a partir de una aproximación de ellos. Debido a la existencia de dichas variantes es por lo que el método estudiado anteriormente es conocido como *método de la potencia simple* para distinguirlo de los que estudiaremos a continuación.

Conviene aquí recordar algunas propiedades de los autovalores de una matriz.

Teorema 4.10 *Si A una matriz regular y λ un autovalor suyo asociado al autovector v, se verifica que $1/\lambda$ es un autovalor de A^{-1} asociado al mismo autovector v.*

Demostración. Si λ es un autovalor de A asociado a v sabemos que $Av = \lambda v$. Al tratarse de una matriz invertible ($\det(A) \neq 0$) sus autovalores son todos no nulos, por lo que podemos dividir por λ y multiplicar por A^{-1} la última igualdad para obtener que $A^{-1}v = \dfrac{1}{\lambda}\, v$ es decir, $1/\lambda$ es un autovalor de A^{-1} asociado a v.

Teorema 4.11 *Si λ es un autovalor de una matriz A asociado a un autovector v y α una constante cualquiera se verifica que $\lambda-\alpha$ es un autovalor de la matriz $A - \alpha I$ asociado al mismo autovalor v.*

Demostración. Sabemos, por hipótesis, que $Av = \lambda v$, por lo que $Av - \alpha v = \lambda v - \alpha v$ y, por tanto,

$$(A - \alpha I)v = (\lambda - \alpha)v$$

es decir, $\lambda - \alpha$ es un autovector de $A - \alpha I$ asociado a v.

Teorema 4.12 *Si λ es un autovalor de una matriz A asociado a un autovector v y α (una constante cualquiera) no es autovalor de A entonces $1/(\lambda - \alpha)$ es un autovalor de la matriz $(A - \alpha I)^{-1}$ asociado al autovector v.*

La demostración se basa en los teoremas 4.10 y 4.11 y se deja al lector.

Sea $A \in \mathbf{R}^{n\times n}$ una matriz diagonalizable regular para la que sus autovalores verifican que:

$$0 < |\lambda_1| < |\lambda_2| \leq \cdots \leq |\lambda_n| \tag{4.3}$$

Los autovalores $\mu_1, \mu_2, \ldots, \mu_n$ de la matriz A^{-1} son los inversos de los autovalores de A,

$$\mu_i = \frac{1}{\lambda_i} \quad \text{para} \quad i = 1, 2, \ldots, n.$$

Invirtiendo la expresión (4.3) nos queda

$$\frac{1}{|\lambda_1|} > \frac{1}{|\lambda_2|} \geq \cdots \geq \frac{1}{|\lambda_n|}$$

o lo que es lo mismo:

$$|\mu_1| > |\mu_2| \geq \cdots \geq |\mu_n|$$

es decir, el autovalor de menor valor absoluto de la matriz A se corresponde con el de mayor valor absoluto de A^{-1}, por lo que aplicando el método de la

potencia simple a la matriz A^{-1} obtenemos el inverso ($\mu_1 = \dfrac{1}{\lambda_1}$) del autovalor de menor valor absoluto de la matriz A y, por tanto, dicho autovalor.

Obsérvese que debemos ir calculando, en cada paso, los vectores $z_n = A^{-1}\omega_{n-1}$ y $\omega_n = \dfrac{z_n}{\|z_n\|_\infty}$. Pues bien, al cálculo de z_n se realiza resolviendo, por alguno de los métodos estudiados, el sistema $Az_n = \omega_{n-1}$ lo que nos evita calcular A^{-1} y arrastrar los errores que se cometan a lo largo de todo el proceso.

Este método es conocido como *método de la potencia inversa* y nos permite calcular, en caso de existir, el autovalor de menor valor absoluto de una matriz invertible A.

Ejemplo 4.3 Para calcular, por el método de la potencia inversa, el autovalor de menor valor absoluto de la matriz del Ejemplo 4.2 $A = \begin{pmatrix} 6 & 2 & 5 \\ 2 & 2 & 3 \\ 5 & 3 & 6 \end{pmatrix}$ buscamos el de mayor valor absoluto de su inversa $A^{-1} = \begin{pmatrix} 0'75 & 0'75 & -1 \\ 0'75 & 2'75 & -2 \\ -1 & -2 & 2 \end{pmatrix}$ por el método de la potencia simple. Partiendo del vector $z_0 = (1\ 1\ 1)^T$ obtenemos:

$$w_0 = \begin{pmatrix} 1'0000 \\ 1'0000 \\ 1'0000 \end{pmatrix} \quad z_1 = \begin{pmatrix} 0'5000 \\ 1'5000 \\ -1'0000 \end{pmatrix} \quad w_1 = \begin{pmatrix} 0'3333 \\ 1'0000 \\ -0'6667 \end{pmatrix} \quad z_2 = \begin{pmatrix} 1'6667 \\ 4'3333 \\ -3'6667 \end{pmatrix}$$

$$w_2 = \begin{pmatrix} 0'3846 \\ 1'0000 \\ -0'8462 \end{pmatrix} \quad z_3 = \begin{pmatrix} 1'8846 \\ 4'7308 \\ -4'0769 \end{pmatrix} \quad w_3 = \begin{pmatrix} 0'3984 \\ 1'0000 \\ -0'8618 \end{pmatrix} \quad z_4 = \begin{pmatrix} 1'9106 \\ 4'7724 \\ -4'1220 \end{pmatrix}$$

$$w_4 = \begin{pmatrix} 0'4003 \\ 1'0000 \\ -0'8637 \end{pmatrix} \quad z_5 = \begin{pmatrix} 1'9140 \\ 4'7777 \\ -4'1278 \end{pmatrix} \quad w_5 = \begin{pmatrix} 0'4006 \\ 1'0000 \\ -0'8640 \end{pmatrix} \quad z_6 = \begin{pmatrix} 1'9144 \\ 4'7784 \\ -4'1285 \end{pmatrix}$$

$$w_6 = \begin{pmatrix} 0'4006 \\ 1'0000 \\ -0'8640 \end{pmatrix} \quad z_7 = \begin{pmatrix} 1'9145 \\ 4'7785 \\ -4'1286 \end{pmatrix} \quad w_7 = \begin{pmatrix} 0'4006 \\ 1'0000 \\ -0'8640 \end{pmatrix} \quad z_8 = \begin{pmatrix} 1'9145 \\ 4'7785 \\ -4'1287 \end{pmatrix}$$

$$w_8 = \begin{pmatrix} 0'4006 \\ 1'0000 \\ -0'8640 \end{pmatrix} \quad z_9 = \begin{pmatrix} 1'9145 \\ 4'7785 \\ -4'1287 \end{pmatrix}$$

Tenemos, por tanto, que z_9 es una aproximación al autovector asociado al autovalor de menor valor absoluto de la matriz A, por lo que el cociente de Rayleigh nos proporcionará una aproximación de éste.

$$\lambda_3 = \frac{z_9^T A z_9}{z_9^T z_9} = 0.20927063325837.$$

Obsérvese que al tratarse de una matriz de orden 3 podemos calcular $z_{n+1} = A^{-1}w_n$ en vez de resolver el sistema $Az_{n+1} = w_n$ sin que los errores que se acumulan sean demasiado grandes, pero si se tratara de una matriz de un orden muy grande, el cálculo de los vectores z_n sería absolutamente necesario realizarlo resolviendo, por alguno de los sistemas iterados estudiados, los sistemas $Az_{n+1} = w_n$. □

Veamos cómo una nueva variante de este último método nos permite calcular un autovalor cualquiera de una matriz regular A a partir de una aproximación suya. Sin embargo, hay que tener en cuenta que si el autovalor que vamos a aproximar es múltiple o existen otros con el mismo módulo aparecen dificultades que sólo podremos solventar utilizando otros métodos.

Consideremos una matriz A regular tal que todos sus autovalores tengan diferente módulo y supongamos conocida una aproximación α de un determinado autovalor λ_k. Si la aproximación es buena se verificará que $|\lambda_k - \alpha| < |\lambda_i - \alpha|$ para cualquier $i = 1, 2, \ldots n$ con $i \neq k$.

Dado que $1/(\lambda_i - \alpha)$ son los autovalores de $A - \alpha I$ y ésta posee un autovalor $(\lambda_k - \alpha)$ de menor valor absoluto que los demás, el método de la potencia inversa nos proporcionará el valor $\mu_k = \dfrac{1}{\lambda_k - \alpha}$ pudiéndose, a partir de éste último, hallar el valor de $\lambda_k = \dfrac{1}{\mu_k} + \alpha$.

Está variante es conocida como *método de la potencia inversa con desplazamiento.*

Ejemplo 4.4 Supongamos ahora que sabemos que el otro autovalor de la matriz A del Ejemplo 4.2 es aproximadamente 1'5.

Para calcular, por el método de la potencia inversa con desplazamiento, dicho autovalor, comenzamos por calcular el de mayor valor absoluto de la matriz

$$(A - 1'5I)^{-1} = \begin{pmatrix} 7'7143 & -6,8571 & -4'0000 \\ -6'8571 & 5'4286 & 4'0000 \\ -4'0000 & 4'0000 & 2'0000 \end{pmatrix}$$

por el método de la potencia simple. Partiendo del vector $z_0 = (1\ 1\ 1)^T$

obtenemos:

$$w_0 = \begin{pmatrix} 1'0000 \\ 1'0000 \\ 1'0000 \end{pmatrix} \quad z_1 = \begin{pmatrix} -3'1429 \\ 2'5714 \\ 2'0000 \end{pmatrix} \quad w_1 = \begin{pmatrix} -1'0000 \\ 0'8182 \\ 0'6364 \end{pmatrix} \quad z_2 = \begin{pmatrix} -15'8701 \\ 13'8442 \\ 8'5455 \end{pmatrix}$$

$$w_2 = \begin{pmatrix} -1'0000 \\ 0'8723 \\ 0'5385 \end{pmatrix} \quad z_3 = \begin{pmatrix} -15'8499 \\ 13'7466 \\ 8'5663 \end{pmatrix} \quad w_3 = \begin{pmatrix} -1'0000 \\ 0'8673 \\ 0'5405 \end{pmatrix} \quad z_4 = \begin{pmatrix} -15'8233 \\ 13'7272 \\ 8'5501 \end{pmatrix}$$

$$w_4 = \begin{pmatrix} -1'0000 \\ 0'8675 \\ 0'5403 \end{pmatrix} \quad z_5 = \begin{pmatrix} -15'8244 \\ 13'7280 \\ 8'5508 \end{pmatrix} \quad w_5 = \begin{pmatrix} -1'0000 \\ 0'8675 \\ 0'5404 \end{pmatrix} \quad z_6 = \begin{pmatrix} -15'8244 \\ 13'7279 \\ 8'5508 \end{pmatrix}$$

$$w_6 = \begin{pmatrix} -1'0000 \\ 0'8675 \\ 0'5404 \end{pmatrix} \quad z_7 = \begin{pmatrix} -15'8244 \\ 13'7279 \\ 8'5508 \end{pmatrix}$$

Tenemos, por tanto, que z_7 es una aproximación al autovector asociado al autovalor buscado, por lo que

$$\lambda_2 = \frac{z_7^T A z_7}{z_7^T z_7} = 1.56319356980456.$$

□

4.4.3 Algoritmo QR de Francis

Dada una matriz $A \in \mathbf{K}^{n\times n}$, se pretende encontrar una sucesión de matrices $(A_n)_{n\in\mathbf{N}}$ convergente a la matriz triangular de Schur T (en cuya diagonal aparecen los autovalores de la matriz A).

La construcción de dicha sucesión se hace de la siguiente manera: $A_0 = A$.

Supuesta encontrada A_n, se realiza la factorización $A_n = Q_n R_n$ mediante matrices de Householder y $A_{n+1} = R_n Q_n$. En otras palabras:

$$\begin{cases} A_0 = \quad A \quad = Q_0 R_0 \\ A_1 = R_0 Q_0 = Q_1 R_1 \\ A_2 = R_1 Q_1 = Q_2 R_2 \\ \cdots\cdots\cdots\cdots\cdots \\ \cdots\cdots\cdots\cdots\cdots \end{cases}$$

Teniendo en cuenta que

$$A = Q_0R_0 \qquad \Rightarrow A_1 = R_0Q_0 = Q_0^*AQ_0$$

$$R_0Q_0 = Q_1R_1 \qquad \Rightarrow \quad A_2 = R_1Q_1 = Q_1^*R_0Q_0Q_1 = Q_1^*Q_0^*AQ_0Q_1 = \\ = (Q_0Q_1)^*A(Q_0Q_1)$$

. . . . . .

. . . . . .

$$R_{n-1}Q_{n-1} = Q_nR_n \Rightarrow A_n = R_nQ_n = (Q_0Q_1\cdots Q_{n-1})^*A(Q_0Q_1\cdots Q_{n-1})$$

. . . . . .

. . . . . .

En resumen, a partir de A se calcula Q_0 y con ella $A_1 = Q_0^*AQ_0$. Con A_1 calculamos Q_1 y, a continuación, Q_0Q_1 que nos permite calcular $A_2 = (Q_0Q_1)^*A(Q_0Q_1)$. Con A_2 calculamos Q_2 y, a continuación $Q_0Q_1Q_2$ (obsérvese que Q_0Q_1 está almacenada en la memoria) que nos permite calcular ahora $A_3 = (Q_0Q_1Q_2)^*A(Q_0Q_1Q_2)$ y así, sucesivamente.

Ejemplo 4.5 Para el cálculo de los autovalores de la matriz $A = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 2 & 3 \\ 3 & 3 & 3 \end{pmatrix}$ se tiene:

$$A_0 = A = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 2 & 3 \\ 3 & 3 & 3 \end{pmatrix} \qquad A_1 = \begin{pmatrix} 7 & -1'9415 & 0'6671 \\ -1'9415 & -0'5385 & 0'1850 \\ 0'6671 & 0'1850 & -0'4615 \end{pmatrix}$$

$$A_2 = \begin{pmatrix} 7'5034 & 0'3365 & 0'0344 \\ 0'3365 & -1'1474 & -0'1174 \\ 0'0344 & -0'1174 & -0'3554 \end{pmatrix} \quad A_3 = \begin{pmatrix} 7'5162 & -0'0530 & 0'0016 \\ -0'0530 & -1'1758 & 0'0348 \\ 0'0016 & 0'0348 & -0'3404 \end{pmatrix}$$

$$A_4 = \begin{pmatrix} 7'5165 & 0'0083 & 0'0001 \\ 0'0083 & -1'1775 & -0'0100 \\ 0'0001 & -0'0100 & -0'3390 \end{pmatrix} \quad A_5 = \begin{pmatrix} 7'5165 & -0'0013 & 0'0000 \\ -0'0013 & -1'1776 & 0'0029 \\ 0'0000 & 0'0029 & -0'3389 \end{pmatrix}$$

$$A_6 = \begin{pmatrix} 7'5165 & 0'0002 & 0 \\ 0'0002 & -1'1776 & -0'0008 \\ 0 & -0'0008 & -0'3389 \end{pmatrix} \quad A_7 = \begin{pmatrix} 7'5165 & 0 & 0 \\ 0 & -1'1776 & 0 \\ 0 & 0 & -0'3389 \end{pmatrix}$$

Por lo que los autovalores de la matriz A son:

$$-1'1776 \quad -0'3389 \quad \text{y} \quad 7'5165 \qquad \square$$

Nota: Si $\lambda_1, \lambda_2, \ldots, \lambda_n$ son los autovalores de una matriz A tales que $|\lambda_1| > |\lambda_2| > \cdots > |\lambda_n| > 0$, el algoritmo converge. En general se tiene que, si existen

autovalores de igual módulo, la sucesión converge a una matriz triangular por cajas en la que cada caja contiene a los autovalores del mismo módulo.

En otras palabras, la sucesión converge, en general, a una matriz triangular por cajas de tal manera que

a) Si todos los autovalores tienen distinto módulo el límite de la sucesión es una matriz triangular superior en la que los elementos de su diagonal son los autovalores de la matriz.

b) Si existen autovalores de igual módulo la matriz límite es una matriz triangular superior **por cajas** en la que cada caja de orden k de su diagonal es una matriz cuyos autovalores son todos los de la matriz A de igual módulo.

Así, por ejemplo, si una matriz A de orden 3 tiene un autovalor real y dos autovalores complejos conjugados (por tanto de igual módulo) de distinto módulo que el autovalor real se llegaría a una matriz del tipo

$$\left(\begin{array}{c|cc} a_{11} & a_{12} & a_{13} \\ \hline 0 & a_{22} & a_{23} \\ 0 & a_{32} & a_{33} \end{array}\right)$$

donde a_{11} es el autovalor real y los autovalores de la matriz $\begin{pmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{pmatrix}$ son los dos autovalores complejos de la matriz A.

De forma más general, si A es una matriz cualquiera (normal o no), al aplicarle el algoritmo, la sucesión converge (en la mayoría de los casos) a su forma de Schur.

Definición 4.2 Se dice que una matriz $A \in \mathbf{K}^{n\times n}$ es una *matriz (superior) de Hessenberg* si $a_{ij} = 0$ para $i > j+1$, es decir:

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1\,n-1} & a_{1n} \\ a_{21} & a_{22} & a_{23} & \cdots & a_{2\,n-1} & a_{2n} \\ 0 & a_{32} & a_{33} & \cdots & a_{3\,n-1} & a_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & a_{n-1\,n-1} & a_{n-1\,n} \\ 0 & 0 & 0 & \cdots & a_{n\,n-1} & a_{nn} \end{pmatrix}$$

Teorema 4.13 *Dada una matriz $A \in \mathbf{K}^{n\times n}$ existen $n-2$ transformaciones de Householder $H_1, H_2, \ldots, H_{n-2}$ tales que la matriz*

$$H_{n-2}H_{n-3}\cdots H_1 A H_1 \cdots H_{n-3}H_{n-2}$$

es una matriz de Hessenberg.

Demostración. La demostración se hará por inducción en el orden n de la matriz.

Si A es de orden 1 ó 2, el resultado es obvio. Supuesto cierto para matrices de orden n vamos a probarlo para matrices de orden $n+1$.

Dada la matriz $A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1\,n+1} \\ a_{21} & a_{22} & \cdots & a_{2\,n+1} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n+1\,1} & a_{n+1\,2} & \cdots & a_{n+1\,n+1} \end{pmatrix}$ sea $v = \begin{pmatrix} 0 \\ v_2 \\ \vdots \\ v_{n+1} \end{pmatrix}$ el vector cuya transformación de Householder asociada H_1 es tal que

$$H_1 \begin{pmatrix} a_{11} \\ a_{21} \\ a_{31} \\ \vdots \\ a_{n+1\,1} \end{pmatrix} = \begin{pmatrix} a_{11} \\ \tilde{a}_{21} \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

Dicha transformación es de la forma $H_1 = \left(\begin{array}{c|ccc} 1 & 0 & \cdots & 0 \\ \hline 0 & & & \\ \vdots & & H_1' & \\ 0 & & & \end{array}\right)$. Entonces:

$$H_1 A H_1 = \begin{pmatrix} a_{11} & * & \cdots & * \\ \tilde{a}_{21} & * & \cdots & * \\ 0 & * & \cdots & * \\ \vdots & \vdots & \ddots & \vdots \\ 0 & * & \cdots & * \end{pmatrix} \left(\begin{array}{c|ccc} 1 & 0 & \cdots & 0 \\ \hline 0 & & & \\ \vdots & & H_1' & \\ 0 & & & \end{array}\right) = \left(\begin{array}{c|ccc} a_{11} & * & \cdots & * \\ \hline \tilde{a}_{21} & & & \\ 0 & & A_n & \\ \vdots & & & \\ 0 & & & \end{array}\right)$$

Es fácil comprobar que si para $k = 2, 3, \ldots, n-1$ hacemos

$$H_k = \left(\begin{array}{c|ccc} 1 & 0 & \cdots & 0 \\ \hline 0 & & & \\ \vdots & & H_k' & \\ 0 & & & \end{array}\right)$$

siendo $H'_{n-1}, H'_{n-2}, \ldots, H'_2$ las $n-2$ transformaciones de Householder tales que

$$H'_{n-1}H'_{n-2}\cdots H'_2 A_n H'_2 \cdots H'_{n-2}H'_{n-1}$$

es una matriz de Hessenberg, la matriz

$$H_{n-1}H_{n-2}\cdots H_2 H_1 A_{n+1} H_1 H_2 \cdots H_{n-2}H_{n-1}$$

también es de Hessenberg.

4.4.4 Método de Jacobi para matrices simétricas reales

Comenzaremos viendo el método para dimensión dos y, más tarde, generalizaremos para una dimensión cualquiera.

Dada una matriz simétrica y real $A = \begin{pmatrix} a & b \\ b & c \end{pmatrix}$ el método trata de buscar un valor para α de tal forma que la rotación $P = \begin{pmatrix} \cos\alpha & \operatorname{sen}\alpha \\ -\operatorname{sen}\alpha & \cos\alpha \end{pmatrix}$ haga que la matriz $P^{-1}AP$ sea diagonal.

$$P^{-1}AP = \begin{pmatrix} \cos\alpha & -\operatorname{sen}\alpha \\ \operatorname{sen}\alpha & \cos\alpha \end{pmatrix} \begin{pmatrix} a & b \\ b & c \end{pmatrix} \begin{pmatrix} \cos\alpha & \operatorname{sen}\alpha \\ -\operatorname{sen}\alpha & \cos\alpha \end{pmatrix} = \begin{pmatrix} * & 0 \\ 0 & * \end{pmatrix}$$

Se debe verifica, por tanto, que

$$(a-c)\cos\alpha\operatorname{sen}\alpha + b(\cos^2\alpha - \operatorname{sen}^2\alpha) = 0$$

es decir:

$$b(\cos^2\alpha - \operatorname{sen}^2\alpha) = (c-a)\cos\alpha\operatorname{sen}\alpha$$

Si $c = a$ basta tomar $\alpha = \dfrac{\pi}{4}$.

Si $c \neq a$ podemos dividir por $c-a$ y obtenemos que

$$\frac{2b}{c-a} = \frac{2\cos\alpha\operatorname{sen}\alpha}{\cos^2\alpha - \operatorname{sen}^2\alpha} = \frac{\operatorname{sen}2\alpha}{\cos 2\alpha} = \operatorname{tg}\ 2\alpha$$

Llamando $m = \dfrac{2b}{c-a}$ se tiene que $\operatorname{tg}\ 2\alpha = m$, por lo que

$$t = \operatorname{tg}\ \alpha = \frac{-1 \pm \sqrt{1+m^2}}{m}$$

y, a partir del valor de t obtenido, se pueden calcular

$$\cos\alpha = \frac{1}{\sqrt{1+t^2}} \quad \text{y} \quad \operatorname{sen}\alpha = \frac{t}{\sqrt{1+t^2}}$$

valores, estos últimos, que nos permiten determinar la matriz P.

Algoritmo de cálculo

a) Dada $A = \begin{pmatrix} a & b \\ b & c \end{pmatrix}$

a.1) Si $b = 0$ FIN.

a.2) Si $b \neq 0$ y $a = c$

$$P = \begin{pmatrix} \sqrt{2}/2 & \sqrt{2}/2 \\ -\sqrt{2}/2 & \sqrt{2}/2 \end{pmatrix} \quad \text{y} \quad P^{-1}AP = \begin{pmatrix} * & 0 \\ 0 & * \end{pmatrix}. \quad \text{FIN.}$$

b) $m = \dfrac{2b}{c-a} \quad t = \dfrac{-1 \pm \sqrt{1+m^2}}{m}.$

c) $\cos\alpha = \dfrac{1}{\sqrt{1+t^2}} \quad \operatorname{sen}\alpha = \dfrac{t}{\sqrt{1+t^2}}.$

$$P = \begin{pmatrix} \cos\alpha & \operatorname{sen}\alpha \\ -\operatorname{sen}\alpha & \cos\alpha \end{pmatrix} \quad \text{y} \quad P^{-1}AP = \begin{pmatrix} * & 0 \\ 0 & * \end{pmatrix}. \quad \text{FIN.}$$

Para dimensiones superiores sea P la matriz diagonal por bloques

$$P = \begin{pmatrix} I_{p-1} & & \\ & T_{q-p+1} & \\ & & I_{n-q} \end{pmatrix}$$

en la que

$$T_{q-p+1} = \begin{pmatrix} \cos\alpha & & \operatorname{sen}\alpha \\ & I_{q-p-1} & \\ -\operatorname{sen}\alpha & & \cos\alpha \end{pmatrix}$$

y todos los demás elementos nulos.

- Si $a_{pq} = 0$, hacemos $\cos\alpha = 1$ y $\operatorname{sen}\alpha = 0$.
- Si $a_{pp} = a_{qq}$, hacemos $\cos\alpha = \operatorname{sen}\alpha = \dfrac{\sqrt{2}}{2}$.
- Si $a_{pp} \neq a_{qq}$ llamando $m = \dfrac{2a_{pq}}{a_{qq} - a_{pp}}$ y $t = \dfrac{-1 \pm \sqrt{1+m^2}}{m}$ hacemos

$$\cos\alpha = \frac{1}{\sqrt{1+t^2}} \quad \text{y} \quad \operatorname{sen}\alpha = \frac{t}{\sqrt{1+t^2}}$$

Entonces, los elementos que ocupan los lugares pq y qp de la matriz $P^{-1}AP$ son nulos.

El método de Jacobi consiste en buscar el elemento a_{pq} con $p \neq q$ de mayor módulo, anularlo mediante una matriz P_1, buscar el siguiente elemento de mayor módulo y anularlo mediante otra matriz P_2 y así sucesivamente hasta diagonalizar la matriz A.

Dado que las matrices P_k, o matrices de Jacobi, son ortogonales, se verifica que $P_k^{-1} = P_k^T$ por lo que se trata, en definitiva, de buscar $P_1, P_2, \ldots, P_k$ de la forma descrita anteriormente, para obtener

$$P_k^T \cdots P_1^T A P_1 \cdots P_K = D$$

Nota: Lo más importante de este método es que el cálculo de las matrices P_i puede hacerse simultáneamente ya que no se requiere el conocimiento de ninguna de ellas en el cálculo de cualquier otra. Es decir, el método es paralelizable.

Ejemplo 4.6 Consideremos la matriz $A = \begin{pmatrix} 1 & -2 & 3 & 1 \\ -2 & 0 & 4 & 2 \\ 3 & 4 & 1 & 1 \\ 1 & 2 & 1 & 0 \end{pmatrix}$ que es simétrica y real.

El elemento extradiagonal de mayor valor absoluto es $a_{23} = 4$, por lo que comenzamos haciendo

$$m = \frac{2a_{23}}{a_{33} - a_{22}} \quad t = \frac{-1 + \sqrt{1 + m^2}}{m} \quad cos = \frac{1}{\sqrt{1 + t^2}} \quad sen = \frac{t}{\sqrt{1 + t^2}}$$

$$P_{23} = I_4 \quad p_{22} = p_{33} = cos \quad p_{23} = -p_{32} = sen$$

obteniéndose

$$P_{23} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0.7497 & 0.6618 & 0 \\ 0 & -0.6618 & 0.7497 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

por lo que

$$P_{23}^T A P_{23} = \begin{pmatrix} 1 & -3.4848 & 0.9254 & 1 \\ -3.4848 & -3.5311 & 0 & 0.8376 \\ 0.9254 & 0 & 4.5311 & 2.0733 \\ 1 & 0.8376 & 2.0733 & 0 \end{pmatrix}$$

Obsérvese que la transformación de semejanza sólo ha afectado a las filas y columnas 2 y 3 dejando invariantes a los demás elementos. Si queremos aplicar el método a la matriz $P_{23}^T A P_{23}$ obtenida, utilizando como referencia el elemento $a_{14} = 1$, podemos observar que todos los cálculos necesarios para obtener la nueva matriz P_{14} podíamos haberlos realizados al mismo tiempo que los realizados para P_{23} ya que sólo necesitamos los valores de los elementos $(1,1)$, $(1,4)$, $(4,1)$ y $(4,4)$ de la matriz $P_{23}^T A P_{23}$, que por haber quedado invariantes son los mismos que los de la matriz A.

Realizando los cálculos necesarios de forma análoga al caso anterior obtenemos

$$P_{14} = \begin{pmatrix} 0.8507 & 0 & 0 & -0.5257 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0.5257 & 0 & 0 & 0.8507 \end{pmatrix}$$

que aplicaríamos a $P_{23}^T A P_{23}$ para obtener $P_{14}^T P_{23}^T A P_{23} P_{14} = P^T A P$ donde la matriz P viene dada por $P = P_{23} P_{14}$.

Ahora bien, la matriz

$$P = P_{23} P_{14} = \begin{pmatrix} 0.8507 & 0 & 0 & -0.5257 \\ 0 & 0.7497 & 0.6618 & 0 \\ 0 & -0.6618 & 0.7497 & 0 \\ 0.5272 & 0 & 0 & 0.8507 \end{pmatrix}$$

puede ser escrita directamente sin necesidad de construir previamente P_{23} y P_{14} y multiplicarlas.

En resumen, se pueden enviar a un ordenador los valores de a_{22}, $a_{23} = a_{32}$ y a_{33} para que calcule el ángulo de giro correspondiente y nos devuelva los valores de $p_{22} = p_{33}$ y $p_{23} = -p_{32}$, mientras que de forma simultánea enviamos a otro ordenador los valores de a_{11}, a_{14} y a_{44} y nos devolverá los de $p_{11} = p_{44}$ y $p_{14} = -p_{41}$.

En el ordenador central construimos la matriz P con los datos recibidos y calculamos

$$P^T A P = \begin{pmatrix} 1.6180 & -2.5240 & 1.8772 & 0 \\ -2.5240 & -3.5311 & 0 & 2.5445 \\ 1.8772 & 0 & 4.5311 & 1.2771 \\ 0 & 2.5445 & 1.2771 & -0.6180 \end{pmatrix}$$

que volveremos a renombrar con A.

A la vista de la matriz enviaríamos al **Ordenador 1** los datos de los elementos a_{22}, $a_{24} = a_{42}$ y a_{44} mientras que enviaríamos al **Ordenador 2** los de a_{11},

$a_{13} = a_{31}$ y a_{33} que nos devolverían los elementos necesarios para construir una nueva matriz

$$P = \begin{pmatrix} 0.8981 & 0 & 0.4399 & 0 \\ 0 & 0.8661 & 0 & 0.5016 \\ -0.4399 & 0 & 0.8981 & 0 \\ 0 & -0.5016 & 0 & 0.8651 \end{pmatrix}$$

y a partir de ella

$$P^T AP = \begin{pmatrix} 0.6986 & -1.6791 & 0 & -1.6230 \\ -1.6791 & -5.0065 & -1.5358 & 0 \\ 0 & -1.5358 & 5.4506 & 0.4353 \\ -1.6230 & 0 & 0.4353 & 0.8573 \end{pmatrix}$$

Reiterando el proceso llegamos a la matriz

$$\begin{pmatrix} 2.5892 & 0 & 0 & 0 \\ 0 & -5.6823 & 0 & 0 \\ 0 & 0 & 5.7118 & 0 \\ 0 & 0 & 0 & -0.6188 \end{pmatrix}$$

cuyos elementos diagonales son los autovalores de la matriz original. □

Los autovalores de una matriz normal cualquiera pueden ser reales o complejos mientras que los de una matriz hermítica son todos reales, por lo que si pudiéramos transformar el cálculo de los autovalores de una matriz normal en los de otra hermítica habríamos simplificado el problema. Es más, si pudiéramos transformarlo en el cálculo de los autovalores de una matriz simétrica real, podríamos trabajar en el campo real en vez de hacerlo en el campo complejo y, entre otras cosas, podríamos utilizar el método de Jacobi a partir de una matriz normal cualquiera previa transformación en un problema de autovalores de una matriz simétrica real.

En las siguientes secciones estudiaremos cómo pueden realizarse dichas transformaciones.

4.5 Reducción del problema a matrices hermíticas

Proposición 4.14 *Dada cualquier matriz cuadrada A, existen dos matrices hermíticas H_1 y H_2 tales que $A = H_1 + iH_2$.*

Demostración. Si se quiere que $A = H_1 + iH_2$ con H_1 y H_2 hermíticas, se tendrá que $A^* = H_1^* - iH_2^* = H_1 - iH_2$, por lo que resolviendo el sistema resultante se tiene que

$$H_1 = \frac{A + A^*}{2} \qquad\qquad H_2 = \frac{A - A^*}{2i}$$

además, dado que el sistema es compatible determinado, la solución es única.

Teorema 4.15 *Sea $A = H_1 + iH_2$ una matriz normal con H_1 y H_2 hermíticas. Si $x \neq 0$ es un autovector de A asociado al autovalor $\alpha + i\beta$ ($\alpha, \beta \in \mathbf{R}$) se verifica que*

$$H_1 x = \alpha x \qquad y \qquad H_2 x = \beta x$$

Demostración. Por ser A normal, existe una matriz unitaria U tal que $U^*AU = D$

$$D = U^*AU = \begin{pmatrix} \alpha + i\beta & 0 & \cdots & 0 \\ 0 & \alpha_2 + i\beta_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \alpha_n + i\beta_n \end{pmatrix} \Longrightarrow$$

$$A = U \begin{pmatrix} \alpha & 0 & \cdots & 0 \\ 0 & \alpha_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \alpha_n \end{pmatrix} U^* + iU \begin{pmatrix} \beta & 0 & \cdots & 0 \\ 0 & \beta_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \beta_n \end{pmatrix} U^* \Longrightarrow$$

$$H_1 = U \begin{pmatrix} \alpha & 0 & \cdots & 0 \\ 0 & \alpha_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \alpha_n \end{pmatrix} U^* \qquad H_2 = U \begin{pmatrix} \beta & 0 & \cdots & 0 \\ 0 & \beta_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \beta_n \end{pmatrix} U^*$$

donde H_1 y H_2 son hermíticas. Se tiene entonces que α es un autovalor de H_1 asociado a x (primera columna de U) y β un autovalor de H_2 asociado también a x.

Las componentes hermíticas H_1 y H_2 de una matriz cuadrada A nos proporcionan otro método para estudiar si la matriz A es, o no es, normal.

Teorema 4.16 *Una matriz cuadrada A es normal si, y sólo si, sus componentes hermíticas conmutan.*

$$A \text{ es normal} \iff H_1H_2 = H_2H_1$$

Demostración.

$$H_1H_2 = \frac{A+A^*}{2}\cdot\frac{A-A^*}{2i} = \frac{A^2 - AA^* + A^*A + (A^*)^2}{4i}$$

$$H_2H_1 = \frac{A-A^*}{2i}\cdot\frac{A+A^*}{2} = \frac{A^2 + AA^* - A^*A + (A^*)^2}{4i}$$

a) Si A es normal se verifica que $AA^* = A^*A$ por lo que

$$\left.\begin{aligned} H_1H_2 &= \frac{A+A^*}{2}\cdot\frac{A-A^*}{2i} = \frac{A^2+(A^*)^2}{4i} \\ H_2H_1 &= \frac{A-A^*}{2i}\cdot\frac{A+A^*}{2} = \frac{A^2+(A^*)^2}{4i} \end{aligned}\right\} \Longrightarrow\ H_1H_2 = H_2H_1$$

b) Si $H_1H_2 = H_2H_1$ se verifica que

$$A^2 - AA^* + A^*A + (A^*)^2 = A^2 + AA^* - A^*A + (A^*)^2$$

por lo que

$$-AA^* + A^*A = A^2 + AA^* - A^*A \iff 2A^*A = 2AA^* \iff A^*A = AA^*$$

y, por tanto, A es normal.

4.6 Reducción del problema a matrices simétricas reales

Sea H una matriz hermítica y sean $h_{ij} = a_{ij} + ib_{ij}$ con $1 \leq i, j \leq n$ sus elementos. Podemos descomponer la matriz H de la forma $H = A + iB$ donde A y B son matrices reales con $A = (a_{ij})$ y $B = (b_{ij})$. Por ser H hermítica se verifica que

$$A+iB = H = H^* = A^* - iB^* = A^T - iB^T \text{ (por ser } A \text{ y } B \text{ reales)} \Longrightarrow \begin{cases} A = A^T \\ B = -B^T \end{cases}$$

Sea $x = a + ib$ un autovector de H asociado al autovalor α (recuérdese que $\alpha \in \mathbf{R}$ por ser H una matriz hermítica). Se verifica entonces que

$$\left.\begin{aligned} Hx &= (A+iB)(a+ib) = (Aa - Bb) + i(Ab + Ba) \\ \alpha x &= \alpha(a+ib) = \alpha a + i\alpha b \end{aligned}\right\}$$

$$Hx = \alpha x \Rightarrow \begin{cases} Aa - Bb = \alpha a \\ Ab + Ba = \alpha b \end{cases}$$

por lo que

$$\begin{pmatrix} A & -B \\ B & A \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \alpha \begin{pmatrix} a \\ b \end{pmatrix}$$

Obsérvese además que la matriz real $\begin{pmatrix} A & -B \\ B & A \end{pmatrix}$ es simétrica, ya que

$$\begin{pmatrix} A & -B \\ B & A \end{pmatrix}^T = \begin{pmatrix} A^T & B^T \\ -B^T & A^T \end{pmatrix} = \begin{pmatrix} A & -B \\ B & A \end{pmatrix}$$

Los autovalores de la matriz $\begin{pmatrix} A & -B \\ B & A \end{pmatrix}$ son, por tanto, los mismos que los de H y si un autovector de $\begin{pmatrix} A & -B \\ B & A \end{pmatrix}$ es $(x_1, \ldots, x_n, x_{n+1}, \ldots, x_{2n})$, el correspondiente autovector de la matriz H es $\Big((x_1 + ix_{n+1}), \ldots, (x_n + ix_{2n})\Big)$.

Dada una matriz normal calcularemos sus autovalores de la siguiente forma:

Algoritmo de cálculo de los autovalores de una matriz normal

Paso 1 Calcular las componentes hermíticas

$$H_1 = \frac{A + A^*}{2} \quad \text{y} \quad H_2 = \frac{A - A^*}{2i}$$

de la matriz A.

Paso 2 Calcular los autovalores λ_i $(1 \leq i \leq n)$ de la primera componente hermítica H_1 y sus autovectores asociados v_i reduciendo previamente el problema al caso de matrices reales y simétricas.

Paso 3 Calcular los cocientes de Rayleigh $\mu_i = \dfrac{v_i^* H_2 v_i}{v_i^* v_i}$.

Como los vectores v_i $(1 \leq i \leq n)$ son también autovectores de H_2, los μ_i obtenidos son sus autovalores asociados.

Paso 4 Los autovectores de A son los mismos vectores v_i y sus autovalores asociados vienen dados por $\lambda_i + i\mu_i$.

4.7 Aplicación al cálculo de las raíces de un polinomio

Consideremos el polinomio

$$P(x) = a_0x^n + a_1x^{n-1} + a_2x^{n-2} + \cdots + a_{n-1}x + a_n$$

Dado que el polinomio característico de la matriz

$$A = \left(\begin{array}{cccc|c} -a_1/a_0 & -a_2/a_0 & \cdots & -a_{n-1}/a_0 & -a_n/a_0 \\ & & & & 0 \\ & & I_{n-1} & & \vdots \\ & & & & 0 \end{array}\right)$$

(donde I_{n-1} representa la matriz unidad de orden $n-1$) es precisamente $P(x)$, la raíces de dicho polinomio coinciden con los autovalores de la matriz A, por lo que dichas raíces pueden ser obtenidas, en bloque, mediante un método iterado de cálculo de autovalores.

Así, por ejemplo, MATLAB en su comando **roots(P)** para calcular las raíces de un polinomio P construye, en primer lugar, la matriz A definida anteriormente y se limita luego a calcular sus autovalores aplicando internamente el comando **eig(A)** de cálculo de los autovalores de una matriz. Este comando lo que hace, en primer lugar, es aplicar un algoritmo para llevar la matriz A a una matriz de Hessenberg y posteriormente aplicarle el algoritmo QR mediante transformaciones de Householder.

4.8 Ejercicios propuestos

Ejercicio 4.1 Realizar la descomposición de Schur de la matriz

$$A = \begin{pmatrix} -6 & 9 & 3 \\ -2 & 3 & 1 \\ -4 & 6 & 2 \end{pmatrix}.$$

Ejercicio 4.2 Comprobar que la matriz $U = \begin{pmatrix} 0'6 & 0'8 \\ -0'8 & 0'6 \end{pmatrix}$ es unitaria (ortogonal) y obtener, basándose en ella, una matriz normal A que tenga por autovalores 2 y $3i$. Calcular la conmutatriz de A y comprobar que sus componentes hermíticas conmutan.

Ejercicio 4.3 Probar, basándose en el teorema de Gerchgorin, que la matriz:

$$A = \begin{pmatrix} 9 & 1 & -2 & 1 \\ 0 & 8 & 1 & 1 \\ -1 & 0 & 7 & 0 \\ 1 & 0 & 0 & 1 \end{pmatrix}$$

tiene, al menos, dos autovalores reales.

Ejercicio 4.4 Dada la matriz $A = \begin{pmatrix} 2+3i & 1+2i \\ 1+2i & 2+3i \end{pmatrix}$ se pide:

a) Comprobar que es normal sin calcular la matriz conmutatriz.

b) Calcular sus autovalores a partir de los de sus componentes hermíticas.

c) Comprobar que estos autovalores están en el dominio de Gerchgorin.

Ejercicio 4.5 Dada la matriz $A = \begin{pmatrix} a & 1 & 1 \\ 1 & a & 1 \\ 1 & 1 & a \end{pmatrix}$ donde a es un número complejo cualquiera, se pide:

a) Obtener su polinomio característico.

b) Probar que tiene por autovalores: $\lambda = a - 1$ doble y $\lambda = a + 2$ simple.

c) Calcular los autovectores y comprobar que no dependen de a.

Ejercicio 4.6 Dada la matriz $A = \begin{pmatrix} 2-i & 0 & -2+4i \\ 0 & 4-5i & 2-4i \\ -2+4i & 2-4i & 3-3i \end{pmatrix}$, se pide:

a) Probar que es normal.

b) Obtener su primera componente hermítica H_1 y calcular el polinomio característico de dicha componente.

c) Calcular los autovalores de H_1 y, por el mismo método anterior, sus autovectores.

d) Teniendo en cuenta que estos autovectores también lo son de la matriz H_2 (segunda componente hermítica), calcular sus autovalores.

e) Obtener, a partir de los resultados anteriores, los autovalores y autovectores de A, así como la matriz de paso unitaria U tal que $U^*AU = D$.

Ejercicio 4.7 Dada la matriz $A = \begin{pmatrix} 2 & 1 \\ 1 & 0 \end{pmatrix}$ se pide:

a) Calcular su polinomio característico por el método interpolatorio.

b) Tomar una aproximación, con dos cifras decimales exactas, del mayor de los autovalores y afinarla con el cociente de Rayleigh.

Ejercicio 4.8 Dada la matriz $A = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 2 & 3 \\ 3 & 3 & 3 \end{pmatrix}$

a) Hallar sus autovalores mediante el algoritmo QR.

b) Hallar el autovalor de mayor valor absoluto, por el método de la potencia, partiendo del vector $(10, 11, 1)$

Ejercicio 4.9 Dada la matriz $A = \begin{pmatrix} 1+i & -2+i \\ 2-i & 1+i \end{pmatrix}$ se pide:

a) Comprobar que es normal y que $v_1 = \begin{pmatrix} -i \\ 1 \end{pmatrix}$ es un autovector de su primera componente hermítica H_1 asociado al autovalor $\lambda_1 = 0$.

b) Calcular el otro autovalor (y un autovector asociado) de la matriz H_1 aplicando el método de la potencia.

c) Considérese la matriz real y simétrica $S = \begin{pmatrix} x & y \\ y & x \end{pmatrix}$. Probar que la transformación de Jacobi Q^tSQ con $Q = \begin{pmatrix} \cos\alpha & \operatorname{sen}\alpha \\ -\operatorname{sen}\alpha & \cos\alpha \end{pmatrix}$ y $\alpha = \pi/4$ nos anula los elementos extradiagonales.

d) Transformar el problema del cálculo de los autovalores de la matriz H_2 (segunda componente hermítica de la matriz A) al del cálculo de los autovalores de una matriz C simétrica real y comprobar que son suficientes dos transformaciones de Jacobi Q_1 y Q_2, del tipo de las del apartado anterior, para diagonalizar dicha matriz C y obtener los autovalores de H_2.

e) Obtener, a partir de las columnas de la matriz $Q = Q_1Q_2$, los autovectores de la matriz H_2. ¿Cuáles son los autovalores de la matriz A?

Ejercicio 4.10 Sea $\lambda = \alpha + i\beta$, $\alpha, \beta \in \mathbf{R}$, autovalor de la matriz $A = \begin{pmatrix} 2i & -2 \\ 2 & 2i \end{pmatrix}$.

a) Utilizar el teorema de Gerschgorin para probar que el único autovalor real de la matriz A sólo puede ser $\lambda = 0$.

b) Probar que $A^\star = -A$ y deducir, a partir de ello, que A es una matriz normal. ¿Puede ser **NO** diagonalizable una matriz compleja que verifique esa relación?

c) Utilizar la descomposición hermítica de la matriz, $A = H_1 + iH_2$, para deducir que la parte real de los autovalores de A tiene que ser $\alpha = 0$.

d) Hallar el autovalor dominante (de máximo módulo) de la componente hermítica H_2 aplicando el método de la potencia. ¿Quién es el autovalor dominante de A?

Sugerencia: Iniciar el método con el vector $v_1 = (1, 0)^T$.

e) Si se perturba la matriz A en la matriz

$$A + \delta A = \begin{pmatrix} (2 - 10^{-3})\, i & -2 \\ 2 & (2 + 10^{-2})\, i \end{pmatrix},$$

hallar la norma euclídea de la matriz δA. ¿Puedes encontrar una cota del error $E = |\mu - \lambda|$, transmitido al autovalor dominante?

Indicación: $|\mu - \lambda| \leq ||P||\, ||P^{-1}||\, ||\delta A||$, siendo $P^{-1}AP$ diagonal.

Ejercicio 4.11

a) Probar que las raíces del polinomio $p(\lambda) = a + b\lambda + c\lambda^2 + \lambda^3$ son los autovalores de la matriz $A(p) = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -a & -b & -c \end{pmatrix}$.

b) Si el método de la potencia aplicado a la matriz $A(p)$ converge a un vector v, ¿qué relación tiene v con las raíces del polinomio $p(\lambda)$?

c) Si el algoritmo QR aplicado a la matriz $A(p)$ converge a una matriz triangular T, ¿qué relación tiene T con las raíces del polinomio $p(\lambda)$?

d) Si se puede obtener la factorización LU de la matriz $PA(p)$, siendo P una matriz de permutación, ¿quiénes tienen que ser P, L y U?

Ejercicio 4.12 Sean el polinomio $p(x) = x^3+2x^2-2x-4$ y su correspondiente matriz $A = A(p)$, definido en el ejercicio 4.11

a) Utilizar una sucesión de Sturm para probar que el polinomio $p(x)$ tiene sus raíces reales y que sólo una de ellas, que denotaremos α, es positiva.

b) Utilizar el método de Newton para obtener una aproximación de la raíz α, garantizando 5 cifras decimales exactas.

c) Obtener la pseudosolución, β, del sistema $(A^2v)x = A^3v$, determinando la norma del error, para $v = (1, 1, 1)^T$. ¿Debería ser β una aproximación de α?

d) Obtener la matriz de Householder que transforma el vector $a = (0, 0, 4)^T$ en el vector $b = (4, 0, 0)^T$. ¿Se podía haber predicho el resultado?

e) Obtener la factorización QR de la matriz A, utilizando el método de Householder. (Sugerencia: ¡el apartado anterior!)

f) Dar el primer paso del algoritmo QR aplicado a la matriz A. Indicar cómo podría el método de Gram-Schmidt utilizarse para los sucesivos pasos del algoritmo y si esto sería una buena decisión para obtener las raíces del polinomio $p(x)$.

Ejercicio 4.13

a) ¿Qué pasos se dan para calcular los autovalores de una matriz cuadrada A mediante el algoritmo QR? y ¿que forma tiene la matriz a la que converge el algoritmo en los siguientes casos?

   a.1) Si todos sus autovalores tienen distinto módulo.

   a.2) Si existen autovalores de igual módulo.

b) El polinomio característico de la matriz $A = \begin{pmatrix} -4 & 2 & -4 & 3 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}$ es precisamente el polinomio $P(x)$ del Ejercicio 1. Calculando sus autovalores mediante el algoritmo QR el proceso converge a la matriz

$$\begin{pmatrix} -4.64575131106459 & 4.07664693269566 & 1.32820441231845 & -2.21143157264058 \\ 0 & -0.24888977635522 & -0.86635866374600 & 0.58988079050108 \\ 0 & 1.22575806673700 & 0.24888977635522 & 0.03848978825890 \\ 0 & 0 & 0 & 0.64575131106459 \end{pmatrix}$$

Calcular, a partir de dicha matriz, las raíces del polinomio (autovalores de A).

c) Al aplicar el método de la potencia y comenzando el proceso con el vector $x = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}$ se obtiene en la cuarta iteración el vector $\begin{pmatrix} 1 \\ -0.2152 \\ 0.0463 \\ -0.0100 \end{pmatrix}$. Determinar una aproximación de la raíz de mayor valor absoluto del polinomio $P(x)$ (autovalor correspondiente) utilizando el cociente de Rayleigh.

Ejercicio 4.14 Sean las matrices A, A_n y B definidas como:

$$A = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 3 & 1 & 0 \end{pmatrix},\ A_n = \begin{pmatrix} 1'671 & 0'242 & 2'164 \\ 0'00 & -0'50 & 1'47 \\ 0'00 & -0'81 & -1'16 \end{pmatrix} \text{ y } B = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 3 & 1 & 0'1 \end{pmatrix}.$$

a) Aplicando el algoritmo QR real a la matriz A se obtiene (iterando suficientemente), como aproximación "aceptable" del método, la matriz A_n. ¿Por qué las matrices A y A_n deben tener los mismos autovalores?

Hallar las aproximaciones de los autovalores de la matriz A que se obtienen de A_n.

b) Tomando v_0 aleatoriamente, ¿se debe esperar convergencia o divergencia en el método de la potencia aplicado a la matriz A?

Empezar en $v_0 = (1, 1, 1)^T$ y determinar los tres primeros vectores v_1, v_2 y v_3 que proporciona el método. Hallar la mejor aproximación del autovalor dominante de A, en norma $||\ ||_2$, que se obtiene con v_3.

c) Estudiar si A es una matriz normal. Si se perturba A en la matriz $B = A + \delta A$, hallar la medida de la perturbación $||\delta A||_2$.

¿Se podría asegurar que los autovalores dominantes de las matrices A y B difieren, a lo más, en 0'1?

Ejercicio 4.15 Sean $A = \begin{pmatrix} 1 & 1 & 2 \\ 0 & 1 & 1 \\ 1 & -1 & \varepsilon \end{pmatrix}$ con $0 < \varepsilon \leq 1$, $b = \begin{pmatrix} 0 \\ -1 \\ 2 \end{pmatrix}$ y $J = \begin{pmatrix} 0 & -1 & -2 \\ 0 & 0 & -1 \\ -1 & 1 & 0 \end{pmatrix}$.

a) Obtener la factorización $A = LU$. Utilizar la factorización obtenida para resolver el sistema $A\mathrm{x} = b$.

b) Hallar el número de condición $\kappa_\infty(A)$ de la matriz A para la norma $||\ ||_\infty$. Razonar si el resultado del apartado anterior, obtenido con aritmética de ordenador, podría ser considerado fiable para ε próximo a cero.

c) Para $\varepsilon = 1$, comprobar que J es la matriz de la iteración $\mathrm{x}_{n+1} = J \cdot \mathrm{x}_n + c$ que se obtiene al aplicar el método de Jacobi al sistema $A\mathrm{x} = b$. Determinar c y, empezando en $\mathrm{x}_1 = (1, 0, 0)^T$, hallar el vector x_3.

d) Hallar la aproximación λ_3 del autovalor dominante λ de la matriz J utilizando el método de la potencia, con $v_0 = (1, 0, 0)^T$, y el cociente de Rayleigh para determinar λ_3 con el valor obtenido para v_3.
Sabiendo que λ_3 tiene una cota de error estimada en $e < 0'5$. ¿Es suficiente dicha aproximación para analizar la convergencia de la sucesión (x_n) del método de Jacobi?

e) Para $\varepsilon = 0$, hallar la solución en mínimos cuadrados del sistema $A'\mathrm{x} = b$ que se obtiene al suprimir la primera columna de A, utilizando las ecuaciones normales. Determinar el error y justificar el resultado obtenido.

f) Analizar si es posible encontrar la matriz H de Householder que transforma la segunda columna de A' en el vector b. En caso afirmativo, ¿es normal la matriz H resultante?

Ejercicio 4.16 Considérese la matriz $A = \begin{pmatrix} 3 & 0 & -1 \\ 1 & -2 & 2 \\ -1 & -1 & 8 \end{pmatrix}$.

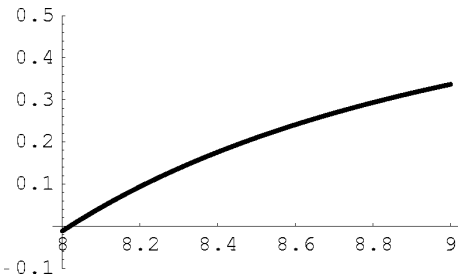
a) Hacer uso de los círculos de Gerschgorin para estudiar el número de autovalores reales que posee.

Obtener su polinomio característico $P(\lambda)$ y un intervalo de amplitud 1 que contenga a su autovalor dominante.

b) Comprobar que la fórmula de Newton-Raphson asociada a dicho polinomio es

$$\lambda_{n+1} = \varphi(\lambda_n) = \frac{2\lambda_n^3 - 9\lambda_n^2 - 39}{3\lambda_n^2 - 18\lambda_n + 3}.$$

Sabiendo que la gráfica de la función $y = \varphi'(\lambda)$ en el intervalo $[8, 9]$ viene dada por la figura adjunta, ¿podemos garantizar la convergencia del método de Newton partiendo de **cualquier** punto $\lambda_0 \in [8, 9]$?



Nota: $\varphi(\lambda)$ es la función que aparece en la fórmula de Newton-Raphson.

c) Si tomamos $\lambda_0 = 8$ ¿con qué error se obtiene la aproximación λ_1?

d) ¿Existe algún vector v_0 para el que podamos garantizar la convergencia del método de la potencia simple aplicado a la matriz A? ¿Qué aproximación se obtiene para el autovalor dominante aplicando el cociente de Rayleigh al vector v_1 si partimos de $v_0 = (1\ 0\ 0)^T$?

e) Aplicando el método QR para el cálculo de los autovalores de A, con una aritmética de ordenador con una precisión de cuatro decimales, el método se estabiliza en la matriz A_n en la que hemos omitido dos de sus elementos x e y

$$A_n = \begin{pmatrix} 8'0195 & -0'5134 & 2'7121 \\ 0 & 2'7493 & 1'5431 \\ 0 & x & y \end{pmatrix}$$

¿Puede ser nulo el elemento x?, ¿se pueden determinar los elementos que faltan sin necesidad de volver a aplicar el algoritmo QR? ¿Sabrías decir cuál es la aproximación obtenida para los otros dos autovalores de la matriz A?

Ejercicio 4.17 Se considera la matriz $A = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -0.25 & -0.125 & 1 \end{pmatrix}$.

a) Demostrar que las raíces del polinomio $P(x) = 2+x-8x^2+8x^3$ coinciden con los autovalores de A. Acotar y separar las raíces de $P(x)$, indicando cuántas raíces reales y complejas tiene. Comparar los resultados con la información que se desprende del estudio de los círculos de Gerschgorin.

b) Determinar un intervalo de amplitud 0.5 con un extremo entero que contenga a la raíz negativa de $P(x)$. Razonar si se verifican en dicho intervalo las condiciones de Fourier. Aproximar por el método de Newton-Raphson dicha raíz con 2 cifras decimales exactas.

c) Tomando como vector inicial $z_0 = (0, 1, 0)^T$, realizar dos iteraciones del método de la potencia inversa. Por medio del cociente de Rayleigh asociado al vector hallado, determinar una aproximación del autovalor de A correspondiente. ¿Qué relación existe entre éste valor y la aproximación hallada en el apartado anterior? ¿Puede haber autovalores de la matriz A en el círculo de centro 0 y radio $\frac{1}{4}$? Razonar las respuestas.

d) Al aplicar el algoritmo QR a la matriz A se obtiene como salida la matriz $T = Q^*AQ$, para cierta matriz unitaria Q. ¿Puede ser T una matriz triangular superior? Justificar la respuesta.

Ejercicio 4.18

a) Utilizar el método interpolatorio para determinar el polinomio característico $p(\lambda)$ de la matriz

$$A = \begin{pmatrix} 2 & -1 & 0 \\ 0 & -2 & 1 \\ 1 & 0 & 5 \end{pmatrix}$$

b) A la vista de los círculos de Gerschgorin, ¿se puede garantizar que el algoritmo QR aplicado a la matriz A convergerá a una matriz triangular con sus autovalores en la diagonal? ¿Se puede garantizar la convergencia del método de la potencia simple si comenzamos a iterar con el vector $v = (1 \;\; 1 \;\; 1)^T$?

c) Haciendo uso de los círculos de Gerschgorin, determinar cuántos autovalores reales posee y calcular un intervalo de amplitud 1 y extremos enteros que contenga al autovalor dominante.

d) Comprobar que, en dicho intervalo, se verifican TODAS las hipótesis de Fourier para la convergencia del método de Newton. ¿En qué extremo deberíamos comenzar a iterar?

e) Tomando $x_0 = 5$ y aplicando el método de Newton, ¿con cuántas cifras exactas se obtiene x_1?

Ejercicio 4.19 Se considera la matriz $A = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -0.25 & -0.125 & 1 \end{pmatrix}$.

a) Demostrar que las raíces del polinomio $P(x) = 2+x-8x^2+8x^3$ coinciden con los autovalores de A. Acotar y separar las raíces de $P(x)$, indicando cuántas raíces reales y complejas tiene. Comparar los resultados con la información que se desprende del estudio de los círculos de Gerschgorin.

b) Determinar un intervalo de amplitud 0.5 con un extremo entero que contenga a la raíz negativa de $P(x)$. Razonar si se verifican en dicho intervalo las condiciones de Fourier. Aproximar por el método de Newton-Raphson dicha raíz con 2 cifras decimales exactas.

c) Tomando como vector inicial $z_0 = (0, 1, 0)^T$, realizar dos iteraciones del método de la potencia inversa. Por medio del cociente de Rayleigh asociado al vector hallado, determinar una aproximación del autovalor de A correspondiente. ¿Qué relación existe entre éste valor y la aproximación hallada en el apartado anterior? ¿Puede haber autovalores de la matriz A en el círculo de centro 0 y radio $\frac{1}{4}$? Razonar las respuestas.

d) Al aplicar el algoritmo QR a la matriz A se obtiene como salida la matriz $T = Q^*AQ$, para cierta matriz unitaria Q. ¿Puede ser T una matriz triangular superior? Justificar la respuesta.

Ejercicio 4.20

a) Utilizar el método interpolatorio para determinar el polinomio característico $p(\lambda)$ de la matriz

$$A = \begin{pmatrix} 2 & -1 & 0 \\ 0 & -2 & 1 \\ 1 & 0 & 5 \end{pmatrix}$$

b) A la vista de los círculos de Gerschgorin, ¿se puede garantizar que el algoritmo QR aplicado a la matriz A convergerá a una matriz triangular con sus autovalores en la diagonal? ¿Se puede garantizar la convergencia del método de la potencia simple si comenzamos a iterar con el vector $v = (1 \;\; 1 \;\; 1)^T$?

c) Haciendo uso de los círculos de Gerschgorin, determinar cuántos autovalores reales posee y calcular un intervalo de amplitud 1 y extremos enteros que contenga al autovalor dominante.

d) Comprobar que, en dicho intervalo, se verifican TODAS las hipótesis de Fourier para la convergencia del método de Newton. ¿En qué extremo deberíamos comenzar a iterar?

e) Tomando $x_0 = 5$ y aplicando el método de Newton, ¿con cuántas cifras exactas se obtiene x_1?

Ejercicio 4.21 Dado el polinomio $P(x) = x^3 - 3x^2 + 3x + 5$

a) Probar, mediante una *sucesión de Sturm*, que sólo tiene una raíz real y determinar $\alpha \in \mathbf{Z}$ para que dicha raíz esté contenida en el intervalo $[\alpha, \alpha + 1]$.

b) Comprobar, mediante las *condiciones de Fourier*, que el método de Newton converge tomando como valor inicial $x = \alpha$.

c) Si tomamos como aproximación de la raíz el valor $\overline{x} = -0.50$ ¿se tiene garantizada alguna cifra decimal exacta?

d) Utilizar el *método interpolatorio* para comprobar que el polinomio característico de la matriz $A = \begin{pmatrix} 3 & -3 & -5 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$ es el polinomio $P(x)$ dado.

e) Para resolver el sistema $(A + 0'5 \cdot I_3)\mathrm{x} = (1, -1, 1)^T$ observamos que resulta más cómodo llevar la primera ecuación al último lugar, es decir, multiplicar el sistema por la matriz $P = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}$. ¿Se altera de esta forma el condicionamiento del sistema? Comprueba que la solución es el vector $\overline{\mathrm{x}} = \frac{1}{21}(-50, 58, -74)^T$.

f) Tomando $-0'50$ como una primera aproximación de su autovalor real, partiendo del vector $z_0 = (1, -1, 1)^T$ y trabajando sólo con dos cifras decimales, realizar una iteración del método de la *potencia inversa con desplazamiento* para calcular, mediante el *cociente de Rayleigh*, una nueva aproximación de dicho autovalor. ¿Se puede garantizar ahora alguna cifra decimal exacta?

Ejercicio 4.22 Sean las matrices $A, B, C \in \mathcal{M}_{2\times 2}(\mathbb{C})$, definidas por

$$A = \begin{pmatrix} -1-i & 3-3i \\ -3+3i & -1-i \end{pmatrix},\ B = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & i \\ i & 1 \end{pmatrix},\ C = \begin{pmatrix} 2+2i & 0 \\ 0 & -4-4i \end{pmatrix}.$$

a) Probar que $A^* = -iA$ y que $B^* = B^{-1}$. ¿Es normal la matriz $B \cdot C \cdot B^*$?

b) Comprobar que se verifica la igualdad $A = BCB^*$. Hallar los autovalores y autovectores de las componentes hermíticas de la matriz A.

c) Probar que si una matriz M verifica la propiedad $M^* = -iM$, entonces es normal y sus autovalores son de la forma $\alpha + i\alpha$, con $\alpha \in \mathbb{R}$. ¿Puede la matriz M tener únicamente autovalores reales?

Indicación: De $M = QDQ^*$, deducir $D^* = -iD$.

d) Se perturba la matriz A en $A+\delta A$, de modo que $||\delta A||_2 \leq 10^{-6}$. Razonar si los autovalores de $A + \delta A$, obtenidos con MATLAB, pueden ser muy diferentes de los de la matriz A. ¿Sucedería lo mismo para una matriz M, con $M^* = -iM$ y dimensión elevada?

e) Hallar la aproximación del autovalor dominante de A que se obtiene con un paso del método de la potencia, partiendo de $q_0 = (1, 0)^T$, y utilizando el cociente de Rayleigh.

Explicar cómo se aplicaría el algoritmo QR para obtener aproximaciones de los autovalores y autovectores de la matriz A. ¿Cuál sería la forma de Schur de A?

Índice

Bibliografía

[1] Burden, R.L. y Faires, J.D. *Análisis Numérico (Sexta edición)*. Internacional Thomson Ed. 1998.

[2] Golub, G.H. y Van Loan, C.F. *Matrix Computations (Third edition)*. Johns Hopkins University Press

[3] Hager, W. *Applied Numerical Linear Algebra*. Ed. Prentice-Hall International. 1988.

[4] Kincaid, D. y Cheney, W. *Análisis Numérico*. Ed. Addison-Wesley Iberoamericana. 1994.

[5] Noble, D. y Daniel, J.W. *Álgebra Lineal Aplicada*. Ed. Prentice-Hall. 1989.

[6] Watkins, D.S. *Fundamentals of MATRIX Computations*. John Wiley & Sons. 1991.