E.T.S. DE INGENIERÍA INFORMÁTICA

INGENIERÍAS TÉCNICAS EN INFORMÁTICA DE SISTEMAS Y GESTIÓN

Apuntes de

CÁLCULO NUMÉRICO

para el curso 2001-2002

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.

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 $k - 1$.

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

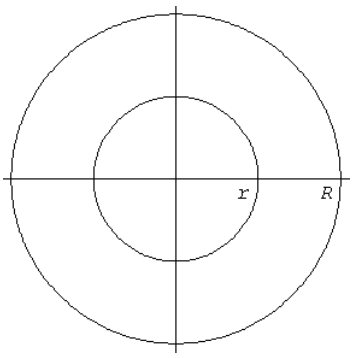
por lo que

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

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 encontrarlas, parece apropiado que un primer paso consista en detectar las posibles situaciones en éstas. 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 1.1 *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 1.1 [Regla de Laguerre] *Consideremos la ecuación*

$$P(x) = a_0 x^n + a_1 x^{n-1} + \cdots + a_{n-1} x + a_n = 0$$

Sean $C(x) = b_0 x^{n-1} + \cdots + b_{n-2} x + b_{n-1}$ *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.

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, aunque en la práctica sólo se utiliza en el caso de las ecuaciones algebraicas, por lo que lo veremos en la Sección 1.4.

1.1 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 $\alpha_0 = \dfrac{a+b}{2}$ y $\varepsilon = \dfrac{b-a}{2}$.

b) Si $f(\alpha_0) = 0$ entonces FIN. $\alpha = \alpha_0$ es la raíz exacta.

Si $f(\alpha_0) > 0$ entonces hacemos $b = \alpha_0$.

Si $f(\alpha_0) < 0$ entonces hacemos $a = \alpha_0$.

Se repite el paso 1, es decir, hacemos $\alpha_0 = \dfrac{a+b}{2}$ y $\varepsilon = \dfrac{b-a}{2}$.

c) Si $\varepsilon < 10^{-k}$ (error prefijado), entonces FIN. El valor de α_0 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 $\varepsilon_n = \dfrac{b-a}{2^{n+1}}$, por lo que si $b - a = 1$ y $n = 9$ se tiene que $\varepsilon_9 < \dfrac{1}{2^{10}} < 10^{-3}$, es decir, en 9 iteraciones obtenemos tres cifras decimales exactas.

1.1.1 Algoritmo

Para $i = 0, 1, 2, \ldots, n, \ldots$, $I_i = [a_i, b_i]$ y $m_i = \dfrac{a_i + b_i}{2}$ (punto medio del intervalo I_i) con

$$I_0 = [a, b] \quad \text{y} \quad I_{i+1} = \begin{cases} [a_i, m_i] & \text{si } \operatorname{sig}(f(a_i)) \neq \operatorname{sig}(f(m_i)) \\ [m_i, b_i] & \text{si } \operatorname{sig}(f(b_i)) \neq \operatorname{sig}(f(m_i)) \end{cases}$$

El proceso debe repetirse hasta que $\begin{cases} f(m_i) = 0 \\ \quad \text{o bien} \\ b_i - a_i < \varepsilon \ \text{ con } \ \varepsilon > 0 \ \text{ prefijado.} \end{cases}$

Se tiene, por tanto:

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

while $(b-a)/2 > \varepsilon$
  $m \leftarrow a + (b-a)/2$
  if $f(m) = 0$
    $a \leftarrow m$
    $b \leftarrow m$
  end if
  if $\operatorname{sign}(f(a)) = \operatorname{sign}(f(m))$
    $a \leftarrow m$
  end if
  if $\operatorname{sign}(f(b)) = \operatorname{sign}(f(m))$
    $b \leftarrow m$
  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

$$\frac{2-1}{2^{n+1}} < 10^{-14} \Longrightarrow 2^{n+1} > 10^{14} \Longrightarrow n \geq 46$$

es decir, tendríamos que detenernos en m_{46} para poder garantizar la precisión exigida. □

Vamos a ver a continuación otros métodos que reducen, de forma considerable, el número de operaciones.

1.2 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 $x \in [a, b]$ 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

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

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 1.1 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).

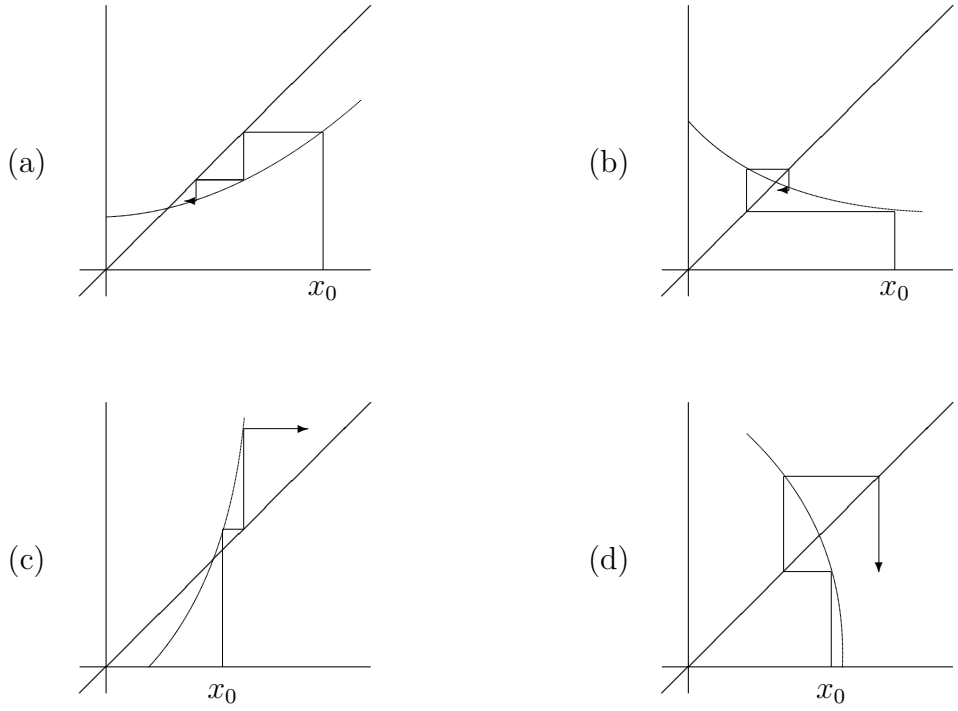


Figura 1.1: Esquema de la convergencia para el teorema del punto fijo.

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

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

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

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 46 iteraciones para calcular la raíz cuadrada de 3 (con 14 cifras decimales exactas) mediante el método de la bisección, mientras que ahora sólo hemos necesitado 26. Sin embargo vamos a ver a continuación cómo se puede reducir aún más el numero de iteraciones aplicando el método conocido como *método de Newton*.

1.3 Método de Newton: análisis de errores

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*

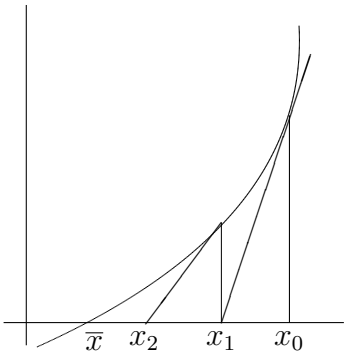
$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} \tag{1.1}$$

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 figura adjunta 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.

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

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

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

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.

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).

1.3.1 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 $e > \varepsilon$

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

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

end

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.

Puede observarse cómo la convergencia del método de Newton-Raphson es mucho más rápida que la del método de la bisección, ya que sólo hemos necesitado 5 iteraciones frente a las 46 que se necesitan en el método de la bisección.

De hecho, existen métodos para determinar el valor inicial x_0 que debe tomarse para que en la segunda iteración se disponga ya de 8 cifras decimales exactas.

1.3.2 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,

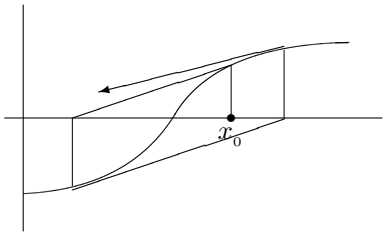
$$x_{n+1} = x_n - \frac{x_n^{10} - 1}{10x_n^9}.$$

Aplicando el algoritmo se obtienen los valores

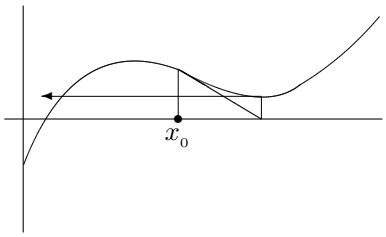
$$
\begin{array}{ll}
x_1 = 51.65 & x_{20} = 6.97714912329906 \\
x_2 = 46.485 & x_{30} = 2.43280139954230 \\
x_3 = 41.8365 & x_{40} = 1.00231602417741 \\
x_4 = 37.65285 & x_{41} = 1.00002393429084 \\
x_5 = 33.887565 & x_{42} = 1.00000000257760 \\
x_{10} = 20.01026825685012 & x_{43} = 1
\end{array}
$$

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. □

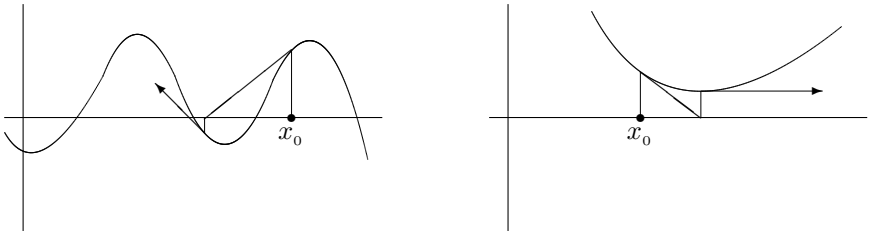
a) Si en las proximidades de la raíz existe un punto de inflexión, las iteraciones divergen progresivamente de la raíz.



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 1.2), 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.

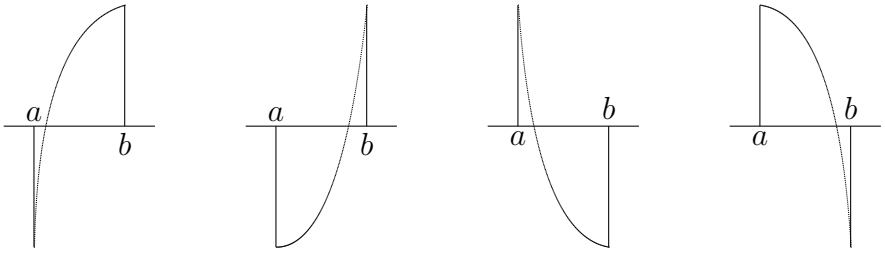
En estos casos, el método de Newton es convergente debiéndose tomar como valor inicial

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

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 1.2: 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.

1.3.3 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

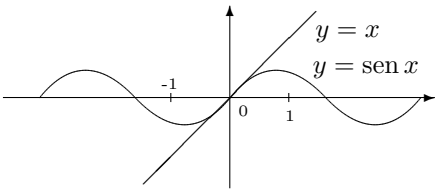
$$x_{n+1} = x_n - k\frac{f(x_n)}{f'(x_n)}$$

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 escribiéndola de la forma $\operatorname{sen} x = x$, por lo que las soluciones serán los puntos de intersección de la curva $y = \operatorname{sen} x$ con $y = x$.

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:

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

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:

$$x_0 = 1$$
$$x_1 = -0'034\ldots$$
$$x_2 = 0'000001376\ldots$$
$$x_3 = 0'00000000000009\ldots$$

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.

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

□

1.4 Cálculo de ceros de polinomios

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.

Si la función es un polinomio $P(x)$ hemos visto que el polinomio

$$Q(x) = \frac{P(x)}{\text{mcd}\,(P(x), P'(x))}$$

posee los mismos ceros que $P(x)$ pero todos simples. Con lo que el primer paso a la hora de calcular los ceros de un polinomio es *eliminar sus raíces múltiples*.

Vamos a estudiar, por tanto, un método que nos permite separar las raíces de una ecuación algebraica.

1.4.1 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 1.5 [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*

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

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

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

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

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

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

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

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

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

$f_4(x) = 1.$

	$-\infty$	-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.

1.4.2 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 1.6 [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 $(0 \le i \le n-1)$ 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

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

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		$x_0 b_0$	$x_0 b_1$	$\cdots$	$x_0 b_{n-2}$	$x_0 b_{n-1}$
	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		$x_0 b_0$	$x_0 b_1$	$\cdots$	$x_0 b_{n-3}$	$x_0 b_{n-2}$	$x_0 b_{n-1}$
	b_0	b_1	b_2	$\cdots$	b_{n-2}	b_{n-1}	$P(x_0)$
x_0		$x_0 c_0$	$x_0 c_1$	$\cdots$	$x_0 c_{n-3}$	$x_0 c_{n-2}$	
	c_0	c_1	c_2	$\cdots$	c_{n-2}	$P'(x_0)$	

Ejemplo 1.7 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	31
2		4	18	50	
	2	9	25	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.

1.5 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 el teorema del valor medio para varias variables al decir que

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

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.

1.5.1 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

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

de donde

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

y, por tanto

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

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

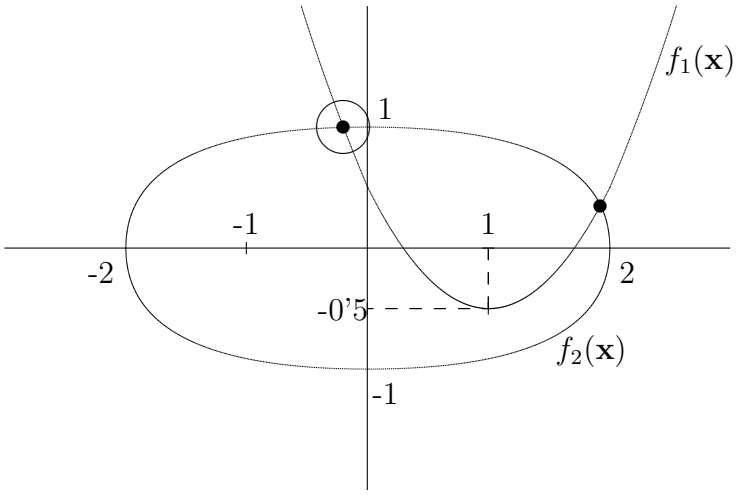
$$\mathbf{x}_{n+1} = \mathbf{x}_n - f'^{-1}(\mathbf{x}_n)f(\mathbf{x}_n)$$

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.

1.6 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 Se considera la ecuación real $2\cos(2x) + 4x - k = 0$.

a) Determinar el valor de k para que tenga una única raíz triple en el intervalo $[0, 1]$.

b) Para $k = 3$, probar que posee una única raíz simple en el intervalo $[0, 1]$, y calcularla con 6 cifras decimales exactas utilizando el método de Newton.

Ejercicio 1.3 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 1.4 Resolver, por los métodos de la bisección y Newton, la ecuación $\ln x - \operatorname{sen} x = 0$, acotando previamente sus raíces.

Ejercicio 1.5 Separar las raíces reales de la ecuación $xe^{-x} - x^2 + 1 = 0$, y obtenerlas con ocho cifras decimales exactas por el método de Newton, aplicando previamente la Regla de Fourier.

Ejercicio 1.6 Dada la ecuación $e^x - (x-2)^2 = 0$, probar que sólo posee una raíz real y obtenerla, por el método de Newton, con seis cifras decimales exactas.

Ejercicio 1.7 Dada la ecuación $e^x - (x+1)^2 = 0$, se pide:

a) Estudiar gráficamente sus raíces reales y acotarlas.

b) Obtener la mayor de ellas con dos cifras decimales exactas por el método de la bisección.

c) Obtenerla con seis cifras decimales exactas por el método de Newton.

Ejercicio 1.8 La ecuación $0'81(x-1) - \ln x = 0$, tiene dos raíces reales, una de las cuales es la unidad. Calcular la otra por el método de Newton, estudiando previamente el campo de convergencia.

Ejercicio 1.9 Se considera la ecuación $(x-1)\ln x^2 - 2x^2 + 7x - 7 = 0$. Separar sus raíces y obtener la mayor de ellas con seis cifras decimales exactas por el método de Newton aplicando, previamente, la regla de Fourier.

Ejercicio 1.10 Dada la ecuación $e^{-x^2} - \dfrac{x^2 - 7x + 7}{10 \cdot (x-1)^2} = 0$ se pide:

a) Determinar el número de raíces reales que posee y separarlas.

b) Demostrar que para cualquier $x > 1'6$ es $f'(x) < 0$ y $f''(x) > 0$.

c) Calcular la mayor de las raíces, con dos cifras decimales exactas, por el método de Newton.

Ejercicio 1.11 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 1.12 Dada la ecuación $8x^3 - 4x^2 - 18x + 9 = 0$, acotar y separar sus raíces reales.

Ejercicio 1.13 Dada la ecuación $x^3 - 6x^2 + 3x + 9 = 0$, acotar y separar sus raíces reales.

Ejercicio 1.14 Dada la ecuación $x^3 - 3ax - 2b = 0$ y basándose en el método de Sturm, discutir para qué valores de a y b, existe una única raíz real.

Ejercicio 1.15 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 1.16 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 1.17 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 $x = F(x) = \frac{1}{3}(x^3 - 6x^2 + 7)$ y, partiendo de un $x_0 \in (6, 7)$, realizamos el proceso $x_{n+1} = F(x_n)$, ¿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 1.18 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.

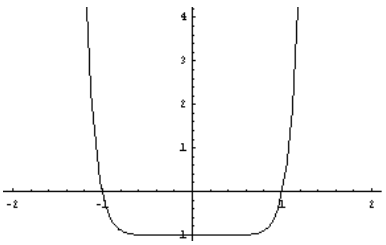


Fig. 1

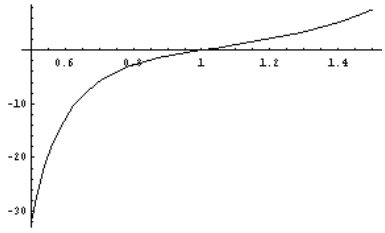


Fig. 2

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 1.19 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 1.20 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: $E_{n+1} = \frac{1}{k}(kE_1)^{2^n}$, con $k = \frac{\max|f''(x)|}{2\min|f'(x)|}$ en un intervalo adecuado.

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.- $\|x\| \geq 0 \quad \forall x \in E$ 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** $\|x\|_\infty = \max_i |x_i|$

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:

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

Si $\left(E, \| \ \|\right)$ 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:

  $d(x, y) = 0 \iff \|x - y\| = 0 \iff x - y = 0 \iff x = y.$
- $d(x, y) = \|x - y\| = \|-1(y - x)\| = |-1| \, \|y - x\| = \|y - x\| = d(y, x).$
- $d(x, y) = \|x - y\| = \|x - z + z - y\| \leq \|x - z\| + \|z - y\| =$
  $= d(x, z) + d(z, y).$

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.

- **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:

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

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

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

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

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

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

Demostración.

- Para la norma de Frobenius de matrices.
  $\|UA\|_F = \sqrt{\text{tr }[(UA)^*(UA)]} = \sqrt{\text{tr }(A^*U^*UA)} = \sqrt{\text{tr }(A^*A)} = \|A\|_F.$
  $\|A\,U\|_F = \sqrt{\text{tr }[(A\,U)(A\,U)^*]} = \sqrt{\text{tr }(A\,U\,U^*A^*)} = \sqrt{\text{tr }(A\,A^*)} = \|A\|_F.$
- Para la norma euclídea de vector.
  $\|Ux\|_2 = \sqrt{(Ux)^*(Ux)} = \sqrt{x^*U^*Ux} = \sqrt{x^*x} = \|x\|_2.$
  $\left\|x^TU\right\|_2 = \left\|(x^TU)^T\right\|_2 = \left\|U^Tx\right\|_2 = \sqrt{(U^Tx)^*(U^Tx)} = \sqrt{x^*(U^T)^*U^Tx} =$
  $= \sqrt{x^T(U^*)^TU^Tx} = \sqrt{x^*(UU^*)^Tx} = \sqrt{x^*I^Tx} = \sqrt{x^*x} = \|x\|_2.$ ■

Lema 2.2 *Las matrices A, A^*, AU y UA, donde U es una matriz unitaria, poseen los mismos valores singulares.*

Proposición 2.2 *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 2.3 *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 $\max_i |\lambda_i| \leq \|A\|$, es decir: $\rho(A) \leq \|A\|$. ■

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: $\begin{pmatrix} a_{11} & a_{12} & 0 & 0 \\ a_{21} & a_{22} & a_{23} & 0 \\ 0 & a_{32} & a_{33} & a_{34} \\ 0 & 0 & a_{43} & a_{44} \end{pmatrix}$
- Triangulares superiores: $\begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ 0 & a_{22} & a_{23} & a_{24} \\ 0 & 0 & a_{33} & a_{34} \\ 0 & 0 & 0 & a_{44} \end{pmatrix}$
- Triangulares inferiores: $\begin{pmatrix} a_{11} & 0 & 0 & 0 \\ a_{12} & a_{22} & 0 & 0 \\ a_{31} & a_{32} & a_{33} & 0 \\ a_{41} & a_{42} & a_{43} & a_{44} \end{pmatrix}$

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.

2.3 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:

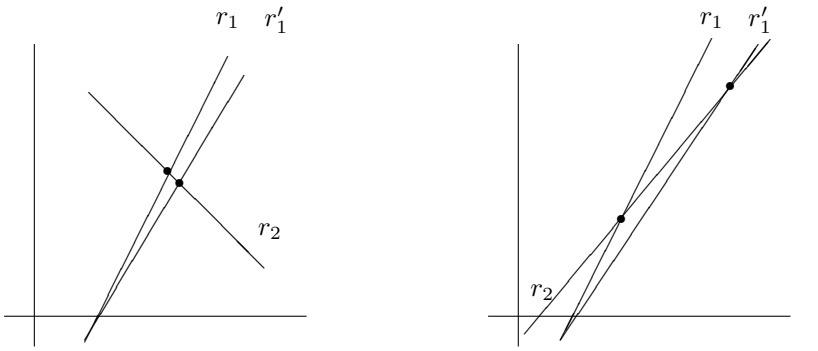
$$\begin{matrix} \left\|A - \overline{A}\right\| < \varepsilon \\ \left\|b - \overline{b}\right\| < \varepsilon \end{matrix} \Longrightarrow \begin{cases} \|x - \overline{x}\| \approx \varepsilon & \text{sistema bien condicionado} \\ \|x - \overline{x}\| \gg \varepsilon & \text{sistema mal condicionado} \end{cases}$$

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 2.1 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

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

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

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

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 2.2 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 $\|A\| = n \cdot \max_{i,j} |a_{ij}|$ se tiene que

$$\left.\begin{aligned} \|A\| &= 2 \cdot 4.00001 \\ \|A^{-1}\| &= 2 \cdot \frac{4.00001}{0.00003} \end{aligned}\right\} \Longrightarrow \kappa(A) \simeq \frac{64}{3} \cdot 10^5 > 2 \cdot 10^6$$

Se trata pues, de un sistema mal condicionado.

Si utilizamos la norma fila $\|A\| = \max_i \sum_{j=1}^{n} |a_{ij}|$ obtenemos:

$$\left.\begin{aligned} \|A\| &= 7.00001 \\ \|A^{-1}\| &= \frac{8.00001}{0.00003} \end{aligned}\right\} \Longrightarrow \kappa(A) \simeq \frac{56}{3} \cdot 10^5 > 1.9 \cdot 10^6$$

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

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

a) Como ya se ha visto anteriormente $\kappa(A) \geq 1$ 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:

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

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

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

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

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 y tiene los mismos autovalores que AA^*. Existe entonces una matriz R tal que

$$R^{-1}A^*AR = \begin{pmatrix} \sigma_1^2 & & & \\ & \sigma_2^2 & & \\ & & \ddots & \\ & & & \sigma_n^2 \end{pmatrix}$$

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

$$R^{-1}A^*AR = \sigma^2 I$$

Entonces

$$A^*A = R(\sigma^2 I)R^{-1} = \sigma^2(RIR^{-1}) = \sigma^2 I$$

Ahora bien

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

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

2.4 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 $F_kF_{k-1}\cdots F_1$ aplicadas a la matriz A. Llamando $L^{-1} = F_kF_{k-1}\cdots F_1$ (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 2.4 *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 2.1 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 *Hadamard* o de *diagonal dominante* :

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

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

Así, por ejemplo, la matriz $A = \begin{pmatrix} 3 & 1 & 1 \\ 0 & 2 & 1 \\ 2 & -1 & 5 \end{pmatrix}$ es de diagonal dominante por filas pero no por columnas.

Teorema 2.5 *Toda matriz de Hadamard es regular.*

Demostración. Supongamos que A es de diagonal 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 dominante por filas. Por tanto, toda matriz de Hadamard, es regular. ■

Teorema 2.6 *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 2.4, 2.6 y 2.5, podemos deducir el siguiente corolario.

Corolario 2.7 *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 2.4 garantiza la existencia de las matrices L y U.

2.5 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 2.8 [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 2.3 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 2.9 *Una matriz hermítica y regular A es definida positiva si, y sólo si, admite factorización de Cholesky.*

2.6 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 2.4 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 2.10 *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 2.11 *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:

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

Pasando al límite

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

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 2.12 *Si para alguna norma matricial subordinada 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

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

■

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 2.10 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 2.12, que el proceso será convergente si se verifica que $\|M^{-1}N\| < 1$ para alguna norma subordinada.

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

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

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

2.6.1 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 2.13 *Si A es una matriz de diagonal dominante, el método de Jacobi es convergente.*

2.6.2 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 2.14 *Si A es una matriz de diagonal estrictamente dominante, el método de Gauss-Seidel es convergente.*

2.6.3 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 2.15 *Una condición necesaria para que converja el método de relajación es que* $\omega \in (0, 2)$.

Teorema 2.16 *Si A es de diagonal dominante, el método de relajación es convergente cualquiera que sea* $\omega \in (0, 1]$.

Teorema 2.17 *Si A es simétrica y definida positiva, el método de relajación converge si, y sólo si,* $\omega \in (0, 2)$

2.7 Métodos del descenso más rápido y del gradiente conjugado

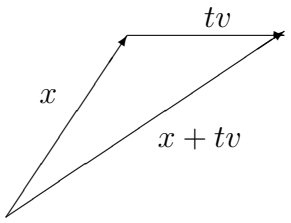
Los métodos que vamos a tratar a continuación son válidos para sistemas $Ax = b$ cuya matriz A es simétrica y definida positiva, es decir, para matrices tales que $A^T = A$ (simétrica) y $x^T Ax > 0$ cualquiera que sea el vector $x \neq 0$ (definida positiva).

Lema 2.18 *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 iterado 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

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

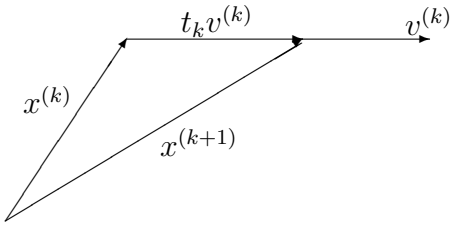
Estos vectores nos permitirán buscar una dirección apropiada $v^{(k)}$ y el siguiente punto de la sucesión vendrá dado por

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

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 $x^{(k+1)}$



2.7.1 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

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.

2.7.2 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.

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 $d^2 < \varepsilon$ **then stop**
  $v \leftarrow r + (d/c)v$
  $c \leftarrow d$
  output k, x, r
end

2.8 Factorizaciones ortogonales

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 $A = QR$.

Al resolver el sistema $Ax = b$, si realizamos la descomposición QR obtenemos $QRx = b$ o lo que es lo mismo, $Rx = Q^*b$ que se trata de un sistema triangular.

Podemos observar que con la descomposición LU o la de Cholesky, lo que conseguimos es transformar la resolución de nuestro sistema en la de *dos* sistemas triangulares. Así, por ejemplo, si descomponemos A como un producto LU, lo que nos queda es $LUx = b$. Haciendo $Ux = y$ podemos resolver el sistema triangular $Ly = b$ para obtener su solución $\overline{y}$. Nos queda ahora $Ux = \overline{y}$ que se trata de otro sistema triangular y, por tanto, de fácil resolución.

En la descomposición QR el proceso es otro. Dado que Q es unitaria se tiene, inmediatamente la matriz Q^*, por lo que se puede conocer de forma directa el valor de Q^*b, con lo que el sistema lo transformamos en $Rx = Q^*b$ que es un sistema triangular. Es decir, la ventaja de esta descomposición es que sólo debemos resolver *un*sistema triangular, frente a los dos que es necesario resolver en los casos LU ó Cholesky.

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. Por tanto, se nos plantean las mismas dificultades que en el referido método de ortonormalización.

Ello nos lleva a tratar de buscar un método por el que podamos realizar una descomposición QR sin necesidad de aplicar Gram-Schmidt. Este proceso lo lograremos mediante las denominadas *Transformaciones de Householder*.

2.9 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:

$$H = \begin{cases} I \in \mathbf{K}^{n\times n} & \text{si } v = 0 \\ I - \dfrac{2}{v^*v}vv^* & \text{si } v \neq 0 \end{cases}$$

Proposición 2.3 *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 ($H^*H = HH^* = I$).

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) $HH^* = HH = \left(I - \dfrac{2}{v^*v}vv^*\right)\left(I - \dfrac{2}{v^*v}vv^*\right) = I - \dfrac{4}{v^*v}vv^* + \left(\dfrac{2}{v^*v}\right)^2 vv^*vv^* =$

$$= I - \frac{4}{v^*v}vv^* + \frac{4}{(v^*v)^2}v(v^*v)v^* = I - \frac{4}{v^*v}vv^* + \frac{4}{(v^*v)^2}(v^*v)vv^* = I.$$

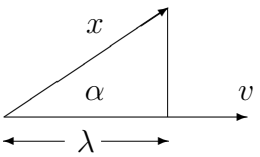
c) Basta observar que, por los apartados anteriores, $H^2 = HH = HH^* = I$.

2.9.1 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.

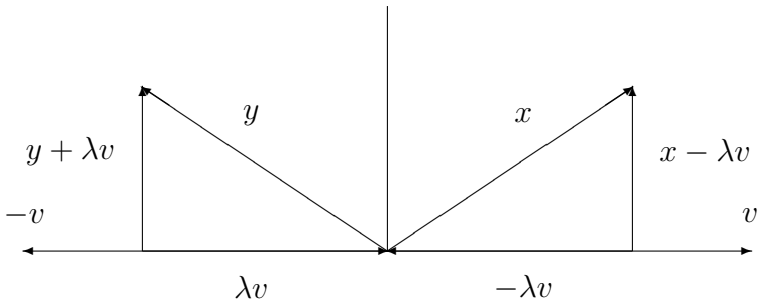
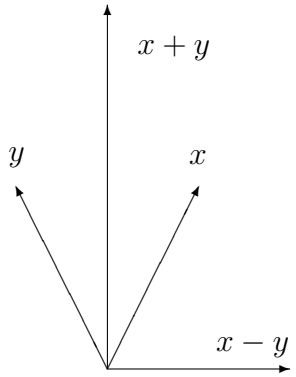


Figura 2.2: x y su transformado.

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.

2.9.2 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

$$\left.\begin{aligned} 2(\overline{v_1}x_1 + \overline{v_2}x_2 + \cdots + \overline{v_n}x_n) &= 1 \\ \overline{v_1}v_1 + \overline{v_2}v_2 + \cdots + \overline{v_n}v_n &= 1 \end{aligned}\right\} \Rightarrow \quad \left.\begin{aligned} 2(\overline{v_1}x_1 + |x_2|^2 + \cdots + |x_n|^2) &= 1 \\ |v_1|^2 + |x_2|^2 + \cdots + |x_n|^2 &= 1 \end{aligned}\right\} \Rightarrow$$

$$2(\overline{v_1}x_1 + \|x\|^2 - |x_1|^2) = 1 \tag{2.4}$$

$$|v_1|^2 + \|x\|^2 - |x_1|^2 = 1 \tag{2.5}$$

De la ecuación 2.4 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 2.4 y 2.5 se obtiene que

$$\left.\begin{aligned} 2(\lambda\,|x_1|^2 + \|x\|^2 - |x_1|^2) &= 1 \\ \lambda^2\,|x_1|^2 + \|x\|^2 - |x_1|^2 &= 1 \end{aligned}\right\} \Longrightarrow |x_1|^2\,(\lambda^2 - 2\lambda + 1) = \|x\|^2 \Longrightarrow$$

$$\lambda = 1 \pm \frac{\|x\|}{|x_1|} \quad \text{supuesto } x_1 \neq 0$$

(Si $x_1 = 0$ basta con tomar $v = (\|x\|, x_2, \ldots, x_n)$).

El vector que buscamos es, por tanto, $v = \begin{pmatrix} \left(1 \pm \frac{\|x\|}{|x_1|}\right) x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$ 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

$$H_v x = y = x - v = \begin{pmatrix} \mp \frac{x_1}{|x_1|} \|x\| \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

2.10 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 $v_1 = \frac{x_1 - y_1}{\|x_1 - y_1\|}$, 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_1A.

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

$$H_kH_{k-1}\cdots H_1A = R \iff Q^*A = R \quad \text{con} \quad Q^* = H_kH_{k-1}\cdots H_1$$

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.

$$H_kH_{k-1}\cdots H_1Ax = H_kH_{k-1}\cdots H_1b \iff Rx = b'$$

sistema triangular de fácil resolución.

Consideremos, por ejemplo, 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

$$v_2 = \frac{x_2 - y_2}{\|x_2 - y_2\|} = \frac{1}{\sqrt{10}} \begin{pmatrix} 0 \\ -1 \\ 3 \end{pmatrix}$$

$$H_2 = I - \frac{2}{v_2^* v_2} v_2 v_2^* = I - \frac{2}{10} \begin{pmatrix} 0 \\ -1 \\ 3 \end{pmatrix} \begin{pmatrix} 0 & -1 & 3 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \frac{4}{5} & \frac{3}{5} \\ 0 & \frac{3}{5} & -\frac{4}{5} \end{pmatrix}$$

$$H_2H_1A = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \frac{4}{5} & \frac{3}{5} \\ 0 & \frac{3}{5} & -\frac{4}{5} \end{pmatrix} \begin{pmatrix} 3 & -5 & -\frac{1}{3} \\ 0 & 4 & \frac{1}{3} \\ 0 & 3 & \frac{5}{3} \end{pmatrix} = \begin{pmatrix} 3 & -5 & -\frac{1}{3} \\ 0 & 5 & \frac{19}{15} \\ 0 & 0 & -\frac{17}{15} \end{pmatrix}$$

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

$$x_1 = a_1 = \begin{pmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{n1} \end{pmatrix}$$

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

$$H_1A = \begin{pmatrix} \alpha_{11} & a_{12}^{(1)} & \cdots & a_{1n}^{(1)} \\ 0 & a_{21}^{(1)} & \cdots & a_{2n}^{(1)} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & a_{n2}^{(1)} & \cdots & a_{nn}^{(1)} \end{pmatrix}$$

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 $\begin{pmatrix} \alpha_{11} \\ 0 \\ \vdots \\ 0 \end{pmatrix}$. 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:

$$H_2H_1A = H_2\left(\begin{array}{c|ccc} \alpha_{11} & a_{12}^{(1)} & \cdots & a_{1n}^{(1)} \\ \hline 0 & a_{21}^{(1)} & \cdots & a_{2n}^{(1)} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & a_{n2}^{(1)} & \cdots & a_{nn}^{(1)} \end{array}\right) =$$

$$= \left(\begin{array}{c|ccc} 1 & 0 & \cdots & 0 \\ \hline 0 & & & \\ \vdots & & \mathbf{H} & \\ 0 & & & \end{array}\right)\left(\begin{array}{c|ccc} 1 & a_{12}^{(1)} & \cdots & a_{1n}^{(1)} \\ \hline 0 & & & \\ \vdots & & \mathbf{A}^1 & \\ 0 & & & \end{array}\right) = \left(\begin{array}{c|ccc} 1 & a_{12}^{(1)} & \cdots & a_{1n}^{(1)} \\ \hline 0 & & & \\ \vdots & & \mathbf{HA}^1 & \\ 0 & & & \end{array}\right)$$

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.

2.11 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

$$\begin{array}{l} ma_1 + n = b_1 \\ ma_2 + n = b_2 \\ ma_3 + n = b_3 \end{array} \iff \begin{pmatrix} a_1 & 1 \\ a_2 & 1 \\ a_3 & 1 \end{pmatrix}\begin{pmatrix} m \\ n \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix}$$

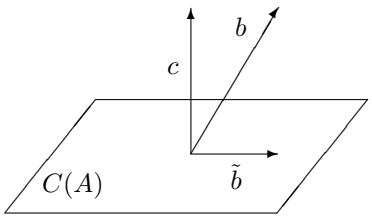
Decir que el sistema no posee solución equivale a decir que el vector $b = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix}$ no pertenece al espacio columna de la matriz $A = \begin{pmatrix} a_1 & 1 \\ a_2 & 1 \\ a_3 & 1 \end{pmatrix}$.

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 ecación por a_1, a_2, ..., a_n, obtenemos

$$\begin{aligned}
&\alpha_1\langle a_1, a_1\rangle + \cdots + \alpha_n\langle a_1, a_n\rangle = \langle \tilde{b}, a_1\rangle = \langle b, a_1\rangle \\
&\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\
&\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\
&\alpha_1\langle a_n, a_1\rangle + \cdots + \alpha_n\langle a_n, a_n\rangle = \langle \tilde{b}, a_n\rangle = \langle b, a_n\rangle
\end{aligned}$$

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 *seudosolución* del sistema superdeterminado $Ax = b$.

2.11.1 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 *seudosolución* que el sistema dado.

Obsérvese que para que los sistemas superdeterminados $Ax = b$ y $BAx = Bb$ posean la misma seudosolución, 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 transformaciones unitarias.

Dado que las transformaciones de Householder son unitarias, podemos resolver un sistema superdeterminado mediante transformaciones de Householder.

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 seudosolución 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 seudosolución 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 seudosolución, la norma del error está representada por la distancia $\left\| b - \tilde{b} \right\|$ que viene dada por

$$\left\| \left(\frac{T}{\Theta} \right) x - \begin{pmatrix} b'_1 \\ \vdots \\ b'_n \\ \vdots \\ b'_m \end{pmatrix} \right\| = \left\| \begin{pmatrix} b'_1 \\ \vdots \\ b'_n \\ 0 \\ \vdots \\ 0 \end{pmatrix} - \begin{pmatrix} b'_1 \\ \vdots \\ b'_n \\ b'_{n+1} \\ \vdots \\ b'_m \end{pmatrix} \right\| = \left\| \begin{pmatrix} b'_{n+1} \\ \vdots \\ b'_m \end{pmatrix} \right\|$$

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 pseudosoluciones 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 peudosoluciones del sistema.

2.12 Descomposición en valores singulares y seudoinversa de Penrose

La descomposición en valores singulares es otra factorización matricial que tiene muchas aplicaciones.

Teorema 2.19 *Toda matriz compleja A, de orden $m \times n$ puede ser factorizada de la forma $A = PDQ$ donde P es una matriz unitaria $m \times m$, D una matriz diagonal $m \times n$ y Q 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 $\{u_1, u_2, \ldots, u_n\}$ un conjunto ortonormal de vectores propios de A^*A, dispuestos de forma que

$$A^*Au_i = \sigma_i^2 u_i$$

Se verifica entonces que

$$\|Au_i\|_2^2 = u_i^*A^*Au_i = u_i^*\sigma_i^2 u_i = \sigma_i^2$$

Esto nos muestra que $Au_i = 0$ 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 Q de orden $n \times n$ cuyas filas son $u_1^*, u_2^*, \ldots, u_n^*$ y definamos

$$v_i = \sigma_i^{-1}Au_i \qquad 1 \leq i \leq r$$

Los vectores v_i constituyen un sistema ortonormal, ya que para $1 \leq i, j \leq r$ se tiene que

$$v_i^*v_j = \sigma_i^{-1}(Au_i)^*\sigma_j^{-1}(Au_j) = (\sigma_i\sigma_j)^{-1}(u_i^*A^*Au_j) = (\sigma_i\sigma_j)^{-1}(u_i^*\sigma_j^2 u_j) = \delta_{ij}$$

Eligiendo vectores adicionales $v_{n+1}, v_{n+2}, \ldots, v_m$ de tal forma que $\{v_1, \ldots, v_m\}$ constituya una base ortonormal de $\mathbf{C}^m$ y construyendo las matrices P de orden $m \times m$ cuyas columnas son los vectores v_i y la matriz D de orden $m \times n$ cuyos elementos diagonales $d_{ii} = \sigma_i$ y los restantes elementos nulos, se tiene que

$$A = PDQ$$

Para probarlo vamos a ver que $P^*AQ^* = D$. En efecto:

$$(P^*AQ^*)_{ij} = v_i^* A u_j = v_i^* \sigma_j v_j = \sigma_j v_i^* v_j = \sigma_j \delta_{ij} = D_{ij}$$

2.12.1 Seudoinversa 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 seudoinversa 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 seudoinversa A^+ a través de la factorización en valores singulares $A = PDQ$ de la forma $A^+ = Q^*D^+P^*$

La seudoinversa 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 2.20 [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 2.21 *La seudoinversa de una matriz tiene las cuatro propiedades de Penrose y, por tanto, es única.*

Demostración. Sea $A = PDQ$ la descomposición en valores singulares de una matriz A. Sabemos que $A^+ = Q^*D^+P^*$.

Si la matriz A es de orden $m \times n$ y tiene rango r, la matriz D es también del mismo orden y tiene la forma

$$D_{ij} = \begin{cases} \sigma_i & \text{si } i = j \leq r \\ 0 & \text{en caso contrario} \end{cases}$$

Se tiene entonces que

$$DD^*D = D$$

ya que

$$(DD^*D)_{ij} = \sum_{k=1}^{n} D_{ik} \sum_{l=1}^{m} D_{kl}^+ D_{lj}.$$

Los términos D_{ik} y D_{lj} hacen que el segundo miembro de la igualdad sea nulo excepto en los casos en que $i, j \leq r$ en cuyo caso

$$(DD^*D)_{ij} = \sum_{k=1}^{r} D_{ik} \sum_{l=1}^{r} D_{kl}^+ D_{lj} = \sigma_i \sum_{l=1}^{r} D_{il}^+ D_{lj} = \sigma_i \sigma_i^{-1} D_{ij} = D_{ij}.$$

Razonamientos análogos nos permiten probar que D^+ verifica las otras tres propiedades de Penrose.

Si nos fijamos ahora en A^+, se tiene que

$$AA^+A = PDQQ^*D^+P^*PDQ = PDD^+DQ = PDQ = A$$

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

$$x = (A^*A)^{-1}A^*b. \tag{2.6}$$

Por otra parte, si hubiésemos resuelto el sistema a través de la seudoinversa de Penrose habríamos obtenido

$$x = A^+b. \tag{2.7}$$

por lo que comparando las ecuaciones (2.6) y (2.7) obtenemos, teniendo en cuenta la unicidad de la seudoinversa, que "si A es de rango máximo", la seudoinversa viene dada por

$$A^+ = (A^*A)^{-1}A^*.$$

2.13 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 $b = (1 \quad 0 \quad 3)^t$

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: $\dfrac{||\delta\mathrm{x}||_\infty}{||\mathrm{x}||_\infty} \leq \kappa_\infty(A) \dfrac{||\delta b||_\infty}{||b||_\infty}$.

Ejercicio 2.11 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 2.12 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 2.13 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 2.14 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 2.15 Hallar la recta de regresión de los puntos:

$(1'1, 5)$, $(1, 5'1)$, $(2, 7'3)$, $(1'8, 6'9)$, $(1'5, 6'1)$, $(3, 8'8)$, $(3'1, 9)$ y $(2'9, 9'1)$

Ejercicio 2.16 Hallar la parábola de regresión de los puntos:

$$(1, 0), \quad (0, 0), \quad (-1, 0), \quad (1, 2) \quad \text{y} \quad (2, 3)$$

Ejercicio 2.17 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 2.18 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 2.19 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 2.20 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 2.21 Dado sistema superdeterminado $Ax = b$ con

$$A = \begin{pmatrix} 1 & 5 & 5 \\ 1 & 2 & 3 \\ 1 & 1 & 3 \\ 1 & 2 & 1 \end{pmatrix} \quad \text{y} \quad b = \begin{pmatrix} 7 \\ 16 \\ -3 \\ 10 \end{pmatrix}$$

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 2.22 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 2.23 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 $H_u = I - \beta u u^T$ 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 2.24 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?

3. Interpolación

3.1 Introducción

Supongamos que se conocen los $n+1$ valores que toma una función $f(x)$, en los puntos del conjunto $\{x_0, x_1, \ldots, x_n\}$ denominado *soporte* , es decir

$$f(x_0) = y_0 \quad f(x_1) = y_1 \quad \cdots \quad f(x_n) = y_n$$

El problema de la interpolación consiste en encontrar una función $g(x)$ *con determinadas características* y tal que $g(x_i) = y_i$ para $i = 0, 1, \ldots, n$. En caso de existir, se dice que $g(x)$ *interpola* a $f(x)$ en $x_0, x_1, \ldots, x_n$.

Al decir *con determinadas características* nos referimos a que se exige que $g(x)$ sea, por ejemplo, un polinomio, un cociente de polinomios, una función trigonométrica, etc.

La finalidad de encontrar una función $g(x)$ que interpola a otra $f(x)$ en los puntos $x_0, x_1, \ldots, x_n$ es la de aproximar la función $f(x)$ en un punto x de tal forma que se pueda decir que $f(x) \approx g(x)$ una vez encontrada $g(x)$. (Otra cosa es la evaluación de $f(x) - g(x)$).

Si los valores de x se encuentran en el intervalo $[x_0, x_n]$ se dice que estamos *interpolando*. Si se encuentran fuera de dicho intervalo, se dice que estamos *extrapolando*.

Como aplicaciones más directas tenemos:

- Evaluación: (una aproximación) de una función complicada f, en un cierto punto x.
- Si $g(x)$ es cómoda de derivar o integrar, la sustitución, en cierta medida, de f' por g' o $\int_a^b f$ por $\int_a^b g$.

En este capítulo sólo trataremos la *interpolación polinomial* y la *interpolación polinomial a trozos* (*splines*).

Ejemplo 3.1 Dada la tabla de valores

x	0	1	2
y	1	3	7

a) Dado que los tres puntos no están alineados, no existe ninguna recta que interpole a dichos valores.

b) Si queremos calcular la parábola $y = ax^2 + bx + c$ que interpola a dichos valores, planteando el correspondiente sistema se obtiene, como solución única, $y = x^2 + x + 1$.

c) Si nuestra intención es buscar una parábola cúbica $y = ax^3 + bx^2 + cx + d$ nos encontramos con que existen infinitas soluciones que son de la forma

$$y = x^2 + x + 1 + \alpha\, x(x-1)(x-2) \qquad \forall \alpha \in \mathbf{R}.$$

d) Por último, para calcular la función polinómica de grado n que interpola a dichos valores obtenemos $y = x^2 + x + 1 + \alpha\, x^{n_1}(x-1)^{n_2}(x-2)^{n_3}$ para cualesquiera $n_1 + n_2 + n_3 = n$ y cualquier $\alpha \in \mathbf{R}$. □

3.2 Interpolación polinomial

Trataremos en esta sección los tres tipos más generalizados de interpolación polinomial, a saber: Lagrange, Newton y Hermite.

3.2.1 Interpolación de Lagrange

Como en cualquier problema de interpolación, consideremos la tabla

x	x_0	x_1	$\cdots$	x_n
y	y_0	y_1	$\cdots$	y_n

y construyamos el polinomio de grado n que interpola a dichos valores. Para ello, consideremos los denominados *polinomios de Lagrange*

$$L_0(x) = \frac{(x-x_1)(x-x_2)\cdots(x-x_n)}{(x_0-x_1)(x_0-x_2)\cdots(x_0-x_n)}$$

$$L_1(x) = \frac{(x-x_0)(x-x_2)\cdots(x-x_n)}{(x_1-x_0)(x_1-x_2)\cdots(x_1-x_n)}$$

$$\vdots$$

$$L_i(x) = \prod_{\substack{j=0 \\ j\neq i}}^{n} \frac{x-x_j}{x_i-x_j}$$

$$\vdots$$

$$L_n(x) = \frac{(x-x_0)(x-x_1)\cdots(x-x_{n-1})}{(x_n-x_0)(x_n-x_1)\cdots(x_n-x_{n-1})}$$

Teorema 3.1 *Los polinomios de Lagrange, definidos anteriormente, verifican:*

a) $L_i(x_j) = \begin{cases} 0 & si \quad i \neq j \\ 1 & si \quad i = j \end{cases}$

b) $\mathrm{gr}(L_i(x)) = n$ *cualquiera que sea* $0 \leq i \leq n$.

c) El polinomio $P_n(x) = y_0L_0(x) + y_1L_1(x) + \cdots + y_nL_n(x)$ *interpola los valores de la tabla*

x	x_0	x_1	$\cdots$	x_n
y	y_0	y_1	$\cdots$	y_n

con $x_0 < x_1 < \cdots < x_n$, *siendo* $\mathrm{gr}(P(x)) \leq n$.

Demostración. Los dos primeros apartados son triviales. Para el tercero es fácil observar que (ver el apartado 1) $P_n(x_i) = y_i$ y que $P_n(x)$, al ser una combinación lineal de polinomios de grado n (apartado 2), no puede ser de grado superior a n (aunque sí inferior). ■

Ejemplo 3.2 Para interpolar los valores de la tabla

x	1	2	3	4
y	0	-1	2	-5

los polinomios de Lagrange son

$$L_0(x) = \frac{(x-2)(x-3)(x-4)}{(1-2)(1-3)(1-4)} = -\frac{1}{6}(x^3 - 9x^2 + 26x - 24)$$

$$L_1(x) = \frac{(x-1)(x-3)(x-4)}{(2-1)(2-3)(2-4)} = \frac{1}{2}(x^3 - 8x^2 + 19x - 12)$$

$$L_2(x) = \frac{(x-1)(x-2)(x-4)}{(3-1)(3-2)(3-4)} = -\frac{1}{2}(x^3 - 7x^2 + 14x - 8)$$

$$L_3(x) = \frac{(x-1)(x-2)(x-3)}{(4-1)(4-2)(4-3)} = \frac{1}{6}(x^3 - 6x^2 + 11x - 6)$$

y como $P_3(x) = y_0 \cdot L_0(x) + y_1 \cdot L_1(x) + y_2 \cdot L_2(x) + y_3 \cdot L_3(x)$ obtenemos que

$$P_3(x) = -\frac{7}{3}x^3 + 16x^2 - \frac{98}{3}x + 19$$ □

El cálculo de los polinomios de Lagrange, puede verse con el Ejemplo 3.2, no es un proceso dinámico, en el sentido de que si ahora añadiéramos un nuevo punto al soporte, habría que comenzar de nuevo todo el proceso.

Teorema 3.2 *Dados los números reales $x_0 < x_1 < \cdots < x_n$ y los $n+1$ números reales cualesquiera $y_0, y_1, \ldots, y_n$, existe un único polinomio $P_n(x)$ de grado no superior a n tal que $P_n(x_i) = y_i$ para $i = 0, 1, \ldots, n$.*

Demostración. La existencia del polinomio es obvia por el Teorema 3.1. Para probar la unicidad supongamos que existiesen dos polinomios $P_n(x)$ y $Q_n(x)$ de grados no superiores a n y tales que $P_n(x_i) = Q_n(x_i) = y_i$ para $0 \leq i \leq n$.

Consideremos el polinomio $D_n(x) = P_n(x) - Q_n(x)$.

$$\begin{cases} D_n(x_0) = 0 \Longrightarrow D_n(x) = (x - x_0)D_{n-1}(x) \\ D_n(x_1) = 0 \Longrightarrow D_{n-1}(x) = (x - x_1)D_{n-2}(x) \\ \quad \vdots \\ D_n(x_{n-1}) = 0 \Longrightarrow D_1(x) = (x - x_{n-1})D_0(x) = (x - x_{n-1}) \cdot k \\ D_n(x_n) = 0 \Longrightarrow k = 0 \end{cases}$$

Se obtiene, por tanto, que

$D_n(x) = (x - x_0)(x - x_1)\cdots(x - x_{n-1}) \cdot 0 = 0 \Longrightarrow P_n(x) = Q_n(x).$ ■

Dada una función $f(x)$ de la que se conocen los transformados de $n+1$ puntos $x_0, x_1, \ldots, x_n$ y su polinomio de interpolación $P_n(x)$, sólo nos falta dar una medida del error que se comete al sustituir la función $f(x)$ por el polinomio $P_n(x)$.

Teorema 3.3 *Sean $x_0 < x_1 < \cdots < x_n$ y sea f una función $n+1$ veces derivable tal que la derivada $f^{(n+1)}(x)$ es continua.*
Sean $y_0 = f(x_0)$, $y_1 = f(x_1), \ldots,$ $y_n = f(x_n)$ y $P_n(x)$ el polinomio de interpolación de los valores de la tabla $\begin{array}{c|cccc} x & x_0 & x_1 & \cdots & x_n \\ \hline y & y_0 & y_1 & \cdots & y_n \end{array}$ *y sea x un número real cualquiera. Se verifica que*

$$f(x) - P_n(x) = \frac{f^{(n+1)}(c)}{(n+1)!}(x - x_0)\cdots(x - x_n)$$

donde el punto c se encuentra en el intervalo determinado por los puntos $x, x_0, x_1, \ldots, x_n$.

Demostración. Sean $\varepsilon(x) = f(x) - P_n(x)$, $z(x) = (x - x_0)\cdots(x - x_n)$ y x un punto cualquiera. Consideremos la función $g(t) = \varepsilon(t) - \varepsilon(x)\dfrac{z(t)}{z(x)}$

- $g(x_i) = 0$ cualquiera que sea $0 \leq i \leq n$ ya que

  $$g(x_i) = \varepsilon(x_i) - \varepsilon(x)\frac{z(x_i)}{z(x)} = 0$$

  por serlo $\varepsilon(x_i)$ y $z(x_i)$.

- $g(x) = 0$ ya que $g(x) = \varepsilon(x) - \varepsilon(x)\dfrac{z(x)}{z(x)} = 0$

Por tanto, la función $g(x)$ se anula en $x, x_0, \ldots, x_{n+1}$, es decir, en $n+2$ puntos. Si ordenamos estos puntos y los denotamos por $t_1, t_2, \ldots, t_{n+2}$, en cada intervalo (t_i, t_{i+1}) se tiene, por Rolle, que existe un punto c_i tal que $g'(x_i) = 0$ (ver Figura 3.1)

Razonando de igual forma obtenemos n puntos donde se anula la derivada segunda $g''(x)$, $n-1$ donde se anula la derivada tercera, y así sucesivamente hasta obtener un punto c en el que se anula la derivada de orden $n+1$, siendo c un punto del intervalo que determinan los puntos $x, x_0, x_1, \ldots, x_n$.

Como $g^{(n+1)}(t) = \varepsilon^{(n+1)}(t) - \dfrac{\varepsilon(x)}{z(x)}z^{(n+1)}(t)$, podemos particularizar en $t = c$ y obtenemos

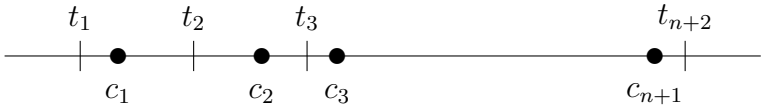


Figura 3.1: Los $n+1$ puntos en los que se anula $g'(x)$.

$$0 = \varepsilon^{(n+1)}(c) - \frac{\varepsilon(x)}{z(x)} z^{(n+1)}(c) = f^{(n+1)}(c) - P_n^{(n+1)}(c) - \frac{\varepsilon(x)}{z(x)}(n+1)!$$

por lo que $\dfrac{\varepsilon(x)}{z(x)}(n+1)! = f^{(n+1)}(c)$, o lo que es lo mismo,

$$\varepsilon(x) = \frac{f^{(n+1)}(c)}{(n+1)!} z(x) = \frac{f^{(n+1)}(c)}{(n+1)!}(x-x_0)\cdots(x-x_n)$$

■

3.2.2 Interpolación de Newton

1.- Diferencias divididas

Consideremos una función $f(x)$ y un soporte $\{x_0, x_1, \ldots, x_n\}$ de $n+1$ puntos. Denotemos por $f_i = f(x_i)$ y consideremos la tabla

x	x_0	x_1	$\cdots$	x_n
y	f_0	f_1	$\cdots$	f_n

Vamos a probar que el polinomio de grado no superior a n que interpola a estos valores es de la forma

$$P(x) = c_0 + c_1(x-x_0) + c_2(x-x_0)(x-x_1) + c_3(x-x_0)(x-x_1)(x-x_2) + \cdots + \\ + \cdots + c_n(x-x_0)(x-x_1)(x-x_2)\cdots(x-x_{n-1})$$

para después, calcular los valores de los coeficientes $c_0, c_1, \ldots, c_n$.

Proposición 3.1 *Los coeficientes $c_0, c_1, \ldots, c_n$, descritos más arriba, dependen de los valores $x_0, x_1, \ldots, x_n$ y $f_0, f_1, \ldots, f_n$.*

Demostración. Como el polinomio $P(x)$ interpola a la tabla de valores, se tiene que

$$P(x_0) = f_0 \Longrightarrow c_0 = f_0$$

$$P(x_1) = f_1 \Longrightarrow f_1 = f_0 + c_1(x_1 - x_0) \Longrightarrow c_1 = \frac{f_1 - f_0}{x_1 - x_0}$$

$$P(x_2) = f_2 \Longrightarrow f_2 = f_0 + \frac{f_1 - f_0}{x_1 - x_0}(x_2 - x_0) + c_2(x_2 - x_0)(x_2 - x_1) \Longrightarrow$$

$$c_2 \text{ depende de } x_0, x_1, x_2, f_0, f_1 \text{ y } f_2$$

Supuesto que c_{k-1} depende de $x_0, \ldots, x_{k-1}, f_0, \ldots, f_{k-1}$, se tiene que

$$P(x_k) = f_k = c_0 + c_1(x_k - x_0) + \cdots + c_k(x_k - x_0)\cdots(x_k - x_{k-1}) \text{ por lo que}$$

$$c_k = \frac{1}{(x_k - x_0)\cdots(x_k - x_{k-1})}\Big[f_k - c_0 - \cdots - c_{k-1}(x_k - x_0)\cdots(x_k - x_{k-2})\Big]$$

por lo que c_k depende de $x_0, x_1, \ldots, x_k, f_0, f_1, \ldots, x_k$. ■

En lo que sigue utilizaremos la notación $c_k = f[x_0, x_1, \ldots, x_k]$, con lo que el polinomio quedará de la forma

$$P(x) = f[x_0] + f[x_0, x_1](x - x_0) + \cdots + f[x_0, x_1, \ldots, x_n](x - x_0)(x - x_1)\cdots(x - x_n)$$

y quedará determinado una vez que se determinen los valores de los coeficientes $f[x_0, x_1, \ldots, x_k]$ con $k = 0, 1, \ldots, n$.

Proposición 3.2 *Sea $P(x)$ el polinomio de interpolación correspondiente a la tabla* $\begin{array}{c|cccc} x & x_0 & x_1 & \cdots & x_n \\ \hline y & f_0 & f_1 & \cdots & f_n \end{array}$. *Si $Q(x)$ y $R(x)$ son los polinomios que interpolan las tablas* $\begin{array}{c|cccc} x & x_0 & x_1 & \cdots & x_{n-1} \\ \hline y & f_0 & f_1 & \cdots & f_{n-1} \end{array}$ *y* $\begin{array}{c|cccc} x & x_1 & x_2 & \cdots & x_n \\ \hline y & f_1 & f_2 & \cdots & f_n \end{array}$ *respectivamente, se verifica que:*

$$P(x) = Q(x) + \frac{x - x_0}{x_n - x_0}\Big(R(x) - Q(x)\Big)$$

Demostración. Consideremos el polinomio

$$T(x) = Q(x) + \frac{x - x_0}{x_n - x_0}\Big(R(x) - Q(x)\Big).$$

Si probamos que $T(x)$ es de grado no superior a n y que interpola la tabla $\begin{array}{c|cccc} x & x_0 & x_1 & \cdots & x_n \\ \hline y & f_0 & f_1 & \cdots & f_n \end{array}$, dado que dicho polinomio es único, se habrá probado que $T(x) \equiv P(x)$ y, por tanto, nuestra proposición.

a) Es obvio que como los grados de $Q(x)$ y $R(x)$ no son superiores a $n-1$, $T(x)$ no puede tener grado superior a n.

b)
- $T(x_0) = Q(x_0) = f_0$
- Si $x_i \in \{x_1, x_2, \ldots, x_{n-1}\}$ se tiene que
$$T(x_i) = Q(x_i) + \frac{x_i - x_0}{x_n - x_0}\Big(R(x_i) - Q(x_i)\Big) = f_i + \frac{x_i - x_0}{x_n - x_0}(f_i - f_i) = f_i.$$
- $T(x_n) = Q(x_n) + \dfrac{x_n - x_0}{x_n - x_0}\Big(R(x_n) - Q(x_n)\Big) = R(x_n) = f_n.$ ■

Proposición 3.3 *Para cualquiera que sea $k = 0, 1, \ldots, n$ se verifica que*

$$f[x_0, x_1, \ldots, x_k] = \frac{f[x_1, \ldots, x_k] - f[x_0, \ldots, x_{k-1}]}{x_k - x_0}$$

siendo $f[x_i] = f_i$ para $0 \leq i \leq n$.

Demostración. $f[x_0, x_1, \ldots, x_k]$ es el coeficiente c_k del término de mayor grado del polinomio de interpolación de la tabla $\begin{array}{c|cccc} x & x_0 & x_1 & \cdots & x_k \\ \hline y & f_0 & f_1 & \cdots & f_k \end{array}$.

Basta ir haciendo tomar a k los valores de 0 a n para que la Proposición 3.2 nos justifique el resultado. ■

Ejemplo 3.3 Calculemos, por el método de las diferencias divididas de Newton, el polinomio de interpolación de la tabla

x	1	3	4	5	7
y	0	1	-1	2	3

x_i	f_i	$f[x_i, x_{i+1}]$	$f[x_i, x_{i+1}, x_{i+2}]$	$f[x_i, x_{i+1}, x_{i+2}, x_{i+3}]$	$f[x_i, x_{i+1}, x_{i+2}, x_{i+3}, x_{i+4}]$
1	$\boxed{0}$				
		$\boxed{1/2}$			
3	1		$\boxed{-5/6}$		
		$-2/1$		$\boxed{5/6}$	
4	-1		$5/2$		$\boxed{-5/18}$
		$3/1$		$-5/6$	
5	2		$-5/6$		
		$1/2$			
7	3				

Por lo que el polinomio de interpolación es

$$P(x) = \frac{1}{2}(x-1) - \frac{5}{6}(x-1)(x-3) + \frac{5}{6}(x-1)(x-3)(x-4) -$$
$$- \frac{5}{18}(x-1)(x-3)(x-4)(x-5) =$$
$$= -\frac{1}{18}(5x^4 - 80x^3 + 430x^2 - 889x + 534). \qquad \square$$

La ventaja de este método es que si ahora introducimos un nuevo dato, por ejemplo que $f(9) = 5$, es decir, $\begin{array}{c|c} x & 9 \\ \hline y & 5 \end{array}$ el polinomio que se obtiene es

$$Q(x) = P(x) + f[x_0, x_1, x_2, x_3, x_4, x_5](x-1)(x-3)(x-4)(x-5)(x-7)$$

y tan sólo habría que calcular el coeficiente $f[x_0, x_1, x_2, x_3, x_4, x_5]$ añadiendo una nueva línea a la tabla anterior.

Puede observarse que dada la tabla $\begin{array}{c|cccc} x & x_0 & \cdots & x_{n-1} & x_n \\ \hline y & y_0 & \cdots & y_{n-1} & y_n \end{array}$, el polinomio de interpolación es de la forma

$$P_n(x) = P_{n-1}(x) + f[x_0, \ldots, x_n](x - x_0) \cdots (x - x_{n-1})$$

Se tenía, también, que para dicho polinomio era

$$f(x) - P_{n-1}(x) = \frac{f^{(n}(\xi)}{n!}(x - x_0) \cdots (x - x_{n-1})$$

Sustituyendo x por x_n tenemos:

$$f(x_n) - P_{n-1}(x_n) = \frac{f^{(n}(\xi)}{n!}(x_n - x_0) \cdots (x_n - x_{n-1})$$

y dado que $f(x_n) = f_n = P_n(x_n)$, se tiene que:

$$P_n(x_n) - P_{n-1}(x_n) = \frac{f^{(n}(\xi)}{n!}(x_n - x_0) \cdots (x_n - x_{n-1}).$$

Podemos, por tanto, enunciar la siguiente proposición.

Proposición 3.1 *Dada la tabla* $\begin{array}{c|cccc} x & x_0 & \cdots & x_{n-1} & x_n \\ \hline y & y_0 & \cdots & y_{n-1} & y_n \end{array}$, *con* $x_0 < \cdots < x_n$, *existe un punto* c *en el intervalo* $[x_0, x_n]$ *para el que*

$$f[x_0, \ldots, x_n] = \frac{f^{(n}(c)}{n!}$$

2.- Diferencias finitas

Consideremos la tabla $\begin{array}{c|cccc} x & x_0 & x_1 & \cdots & x_n \\ \hline f(x) & f_0 & f_1 & \cdots & f_n \end{array}$ en donde el soporte $x_0, \ldots, x_n$ es *regular*, es decir, en donde $x_{i+1} - x_i = \text{cte}$.

Definición 3.1 Dados $y_0, y_1, \ldots, y_n$, se definen las *diferencias finitas* $\Delta^k y_i$ como

$$\Delta y_i = y_{i+1} - y_i \qquad \Delta^k y_i = \Delta(\Delta^{k-1} y_i)$$

Así, por ejemplo, para y_0, y_1, y_2, y_3 se tendrían:

$$\begin{array}{lll} \Delta y_0 = y_1 - y_0 & & \\ & \Delta^2 y_0 = \Delta y_1 - \Delta y_0 & \\ \Delta y_1 = y_2 - y_1 & & \Delta^3 y_0 = \Delta^2 y_1 - \Delta^2 y_0 \\ & \Delta^2 y_1 = \Delta y_2 - \Delta y_1 & \\ \Delta y_2 = y_3 - y_2 & & \end{array}$$

Proposición 3.4 *Dada la tabla* $\begin{array}{c|cccc} x & x_0 & x_1 & \cdots & x_n \\ \hline f(x) & f_0 & f_1 & \cdots & f_n \end{array}$ *en la que* $x_0, \ldots, x_n$ *es un soporte regular con* $x_{i+1} - x_i = h$, *se verifica que, para cualquier valor de* $k = 1, 2, \ldots, n$, *es:*

$$f[x_0, \ldots, x_k] = \frac{\Delta^k f_0}{h^k \cdot k!}$$

Demostración. Haremos inducción en k. Para $k = 1$ sabemos que $f[x_0, x_1] = \dfrac{f_1 - f_0}{x_1 - x_0} = \dfrac{\Delta f_0}{h \cdot 1!}$. Supuesto cierto para k vamos a probarlo para $k+1$.

$$f[x_0, \ldots, x_{k+1}] = \frac{f[x_1, \ldots, x_{k+1}] - f[x_0, \ldots, x_k]}{x_{k+1} - x_0} = \frac{\Delta^k f_1 - \Delta^k f_0}{h^k \cdot k! \ \cdot \ k \cdot h} =$$

$$= \frac{\Delta^{k+1} f_0}{h^{k+1} \cdot (k+1)!}.$$

■

El polinomio de interpolación del soporte $x_0, x_1, \ldots, x_n$ es, por tanto:

$$P_n(x) = f_0 + \Delta f_0 \left(\frac{x - x_0}{h}\right) + \frac{\Delta^2 f_0}{2!} \left(\frac{x - x_0}{h}\right) \left(\frac{x - x_1}{h}\right) + \cdots$$

$$\cdots + \frac{\Delta^n f_0}{n!} \left(\frac{x - x_0}{h}\right) \cdots \left(\frac{x - x_{n-1}}{h}\right)$$

Teniendo en cuenta que $x - x_k = x - (x_0 + k \cdot h) = (x - x_0) - k \cdot h$, podemos poner

$$P_n(x) = f_0 + \Delta f_0 \left(\frac{x - x_0}{h}\right) + \frac{\Delta^2 f_0}{2!} \left(\frac{x - x_0}{h}\right) \left(\frac{x - x_0}{h} - 1\right) + \cdots$$
$$\cdots + \frac{\Delta^n f_0}{n!} \left(\frac{x - x_0}{h}\right) \cdots \left(\frac{x - x_0}{h} - k - 1\right)$$

por lo que, si denotamos por $t = \dfrac{x - x_0}{h}$, se tiene que

$$P_n(x) = f_0 + \frac{\Delta f_0}{1!} t + \frac{\Delta^2 f_0}{2!} t(t-1) + \cdots + \frac{\Delta^n f_0}{n!} t(t-1) \cdots (t - (n-1))$$

Es decir: $P_n(x) = f_0 + \binom{t}{1} \Delta f_0 + \binom{t}{2} \Delta^2 f_0 + \cdots + \binom{t}{n} \Delta^n f_0$ donde $\binom{t}{k} = \dfrac{t(t-1) \cdots (t - (k-1))}{k!}$.

Fenómeno de Runge

Dada una función continua en $[a, b]$, podría pensarse que la sucesión $P_n(x)$ con $n \in \mathbf{N}$ de polinomios de interpolación, obtenidos al aumentar el número de puntos del soporte, converge a la función $f(x)$ es decir, podríamos pensar que

$$\lim_{n \to \infty} |f(x) - P_n(x)| = 0 \qquad \forall\, x \in [a, b]$$

cosa que, sin embargo, no es cierta. En realidad, al aumentar el número de puntos del soporte se mejora la aproximación en la parte central del intervalo, pero la diferencia entre la función y el polinomio interpolador puede aumentar rápidamente en los extremos. Ello nos dice que no es bueno hacer demasiado extenso el soporte, ya que además de aumentar el número de operaciones con la consecuente acumulación de errores, podemos aumentar la pérdida de precisión en los extremos. Este fenómeno es conocido como *fenómeno de Runge*.

Ejemplo 3.4 Si aproximamos la función $f(x) = \dfrac{1}{1 + x^2}$ por un polinomio de segundo grado, en el soporte $\{-4, 0, 4\}$, obtenemos que $P_2(x) = 1 - x^2/17$. En la Figura 3.2 (Izda.) podemos ver ambas gráficas.

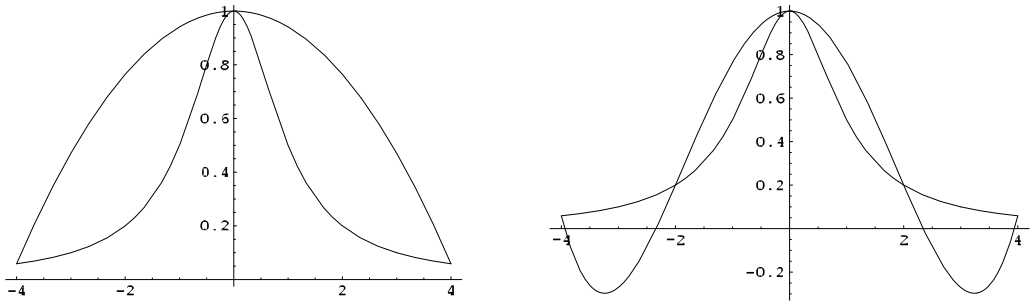


Figura 3.2: Las funciones $\{f(x),\ P_2(x)\}$ y $\{f(x),\ P_4(x)\}$

Si la aproximación la hacemos mediante un polinomio de grado 4 en el soporte $\{-4,-2,0,2,4\}$ obtenemos

$$P_4(x) = \frac{1}{85}(85 - 21x^2 + x^4)$$

que podemos ver representada junto a la función $f(x)$ en la Figura 3.2 (Dcha.).

Si afinamos aún más y aproximamos mediante un polinomio de grado 8 en el soporte $\{-4,-3,-2,-1,0,1,2,3,4\}$ obtenemos

$$P_8(x) = \frac{1}{1700}(1700 - 1124x^2 + 304x^4 - 31x^6 + x^8)$$

cuya gráfica podemos observar en la Figura 3.3.

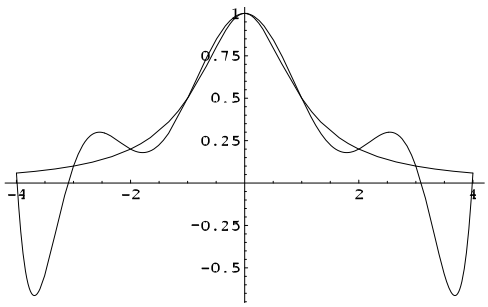


Figura 3.3: Las funciones $f(x)$ y $P_8(x)$

Puede verse el hecho comentado anteriormente del fenómeno de Runge . Vamos mejorando la aproximación en la parte central del intervalo, pero vamos empeorándola en los extremos. □

Ejemplo 3.5 Si construimos con MATHEMATICA el polinomio de interpolación de la función $Log(1+x)$ en el soporte $\{0,1,2,3,4\}$ y representamos el

resultado en el intervalo [0.5,1.5] obtenemos la gráfica de la Figura 3.4 (Izda.). Si utilizamos los nueve puntos del soporte $\{0, 0.5, 1, 1.5, 2, 2.5, 3, 3.5, 4\}$ y representamos el polinomio obtenido en el intervalo $[0.5, 1.5]$ obtenemos la gráfica de la Figura 3.4 (Dcha.).

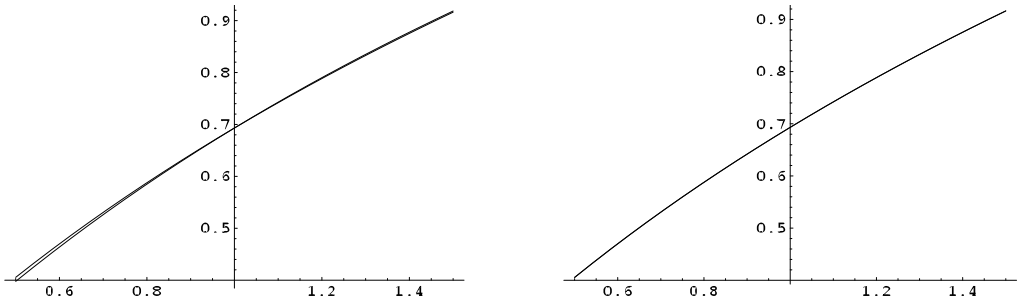


Figura 3.4: Las funciones $\{log(1+x),\ P_4(x)\}$ y $\{log(1+x),\ P_8(x)\}$

Intercalando los puntos medios para obtener ahora un soporte de 17 puntos y realizando la gráfica correspondiente obtenemos como resultado el que se muestra en la Figura 3.5 (Izda.). Sin embargo, si volvemos a utilizar los puntos medios para obtener un soporte de 33 puntos, podemos ver en la Figura 3.5 (Dcha.) que el fenómeno de Runge junto a la acumulación de errores hace que el polinomio obtenido deje de ser una buena aproximación de la función.

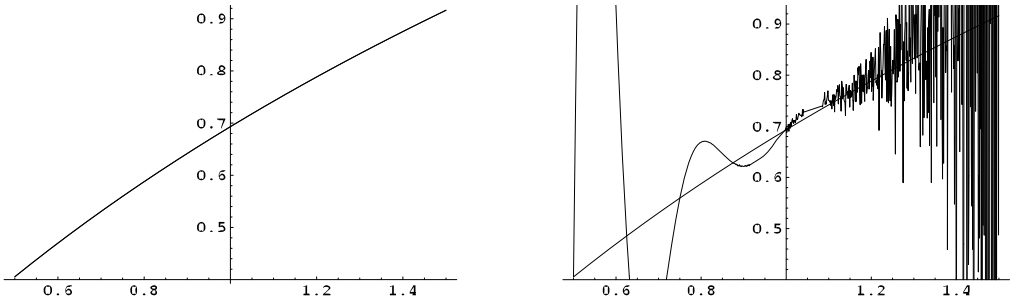


Figura 3.5: Las funciones $\{log(1+x),\ P_{16}(x)\}$ y $\{log(1+x),\ P_{32}(x)\}$

□

3.2.3 Interpolación de Hermite

Este método consiste en buscar un polinomio que interpole a una función $f(x)$ en el soporte $x_0, x_1, \ldots, x_n$ pero haciendo que coincidan, en los puntos del soporte, no sólo los valores de la función con los del polinomio, sino que también coincidan los valores de sus respectivas derivadas.

Consideremos, por tanto, la tabla

x	x_0	x_1	$\cdots$	x_n
$f(x)$	f_0	f_1	$\cdots$	f_n
$f'(x)$	f'_0	f'_1	$\cdots$	f'_n

donde $f_i = f(x_i)$ y $f'_i = f'(x_i)$ para $0 \leq i \leq n$.

Se tienen, en este caso, $2n+2$ condiciones, por lo que debemos buscar un polinomio de grado $2n+1$

$$P(x) = a_{2n+1}x^{2n+1} + a_{2n}x^{2n} + \cdots + a_1 x + a_0$$

que verifique las condiciones:

$$\begin{array}{ll} P(x_0) = f_0 & P'(x_0) = f'_0 \\ P(x_1) = f_1 & P'(x_1) = f'_1 \\ \vdots & \vdots \\ P(x_n) = f_n & P'(x_n) = f'_n \end{array}$$

Teorema 3.4 *Dada la tabla*

x	x_0	x_1	$\cdots$	x_n
$f(x)$	f_0	f_1	$\cdots$	f_n
$f'(x)$	f'_0	f'_1	$\cdots$	f'_n

sean $L_k(x)$ $(k = 0, 1, \ldots, n$ *los polinomios de Lagrange para el soporte dado. El polinomio* $P_{2n+1}(x) = \sum_{k=0}^{n} \Big[a_k + b_k(x - x_k)\Big] L_k^2(x)$ *en el que*

$$\begin{aligned} a_k &= f_k \\ b_k &= f'_k - 2f_k L'_k(x_k) \end{aligned}$$

verifica que

$$\left.\begin{array}{l} P_{2n+1}(x_k) = f_k \\ P'_{2n+1}(x_k) = f'_k \end{array}\right\} \quad k = 0, 1, \ldots, n$$

siendo, además, el único polinomio de grado $2n+1$ *que verifica las condiciones anteriores.*

La demostración del Teorema 3.4 debe realizarla el alumno a modo de ejercicio.

Los polinomios de Lagrange para el soporte $x_0, x_1, \ldots, x_n$ son

$$\begin{aligned}
L_0(x) &= \frac{(x-x_1)(x-x_2)\cdots(x-x_n)}{(x_0-x_1)(x_0-x_2)\cdots(x_0-x_n)} \\
L_1(x) &= \frac{(x-x_0)(x-x_2)\cdots(x-x_n)}{(x_1-x_0)(x_1-x_2)\cdots(x_1-x_n)} \\
&\vdots \\
L_i(x) &= \prod_{\substack{j=0 \\ j\neq i}}^{n} \frac{x-x_j}{x_i-x_j} \\
&\vdots \\
L_n(x) &= \frac{(x-x_0)(x-x_1)\cdots(x-x_{n-1})}{(x_n-x_0)(x_n-x_1)\cdots(x_n-x_{n-1})}
\end{aligned}$$

Llamando $z(x) = (x-x_0)(x-x_1)\cdots(x-x_n)$ se tiene que

$$z'(x) = 1\cdot(x-x_1)\cdots(x-x_n) + (x-x_0)\Big[(x-x_1)\cdots(x-x_n)\Big]'$$

por lo que

$$z'(x_0) = (x_0-x_1)\cdots(x_0-x_n)$$

y de manera análoga se obtiene que

$$z'(x_k) = (x_k-x_1)\cdots(x_k-x_{k-1})(x_k-x_{k+1})\cdots(x_k-x_n)$$

por lo que los polinomios de Lagrange pueden escribirse de la forma

$$L_k(x) = \frac{z(x)}{(x-x_k)z'(x_k)}$$

Teorema 3.5 *Sea* $f(x)$ *una función* $2n+2$ *veces derivable con derivada de orden* $2n+2$ *continua y sea* P_{2n+1} *el polinomio de Hermite que interpola a* $f(x)$ *en el soporte* $x_0, x_1, \ldots, x_n$*. Existe un punto* c *del intervalo que determinan los puntos* $x, x_0, \ldots, x_n$ *en el que se verifica que*

$$f(x) - P_{2n+1}(x) = \frac{f^{(2n+2)}(c)}{(2n+2)!}(x-x_0)^2\cdots(x-x_n)^2$$

Demostración. Sean $\varepsilon(x) = f(x) - P_{2n+1}(x)$ y $z(x) = (x - x_0)\cdots(x - x_n)$ y considérese la función

$$g(t) = \varepsilon(t) - \frac{\varepsilon(x)}{z^2(x)} z^2(t)$$

La demostración es similar a la del Teorema 3.3. ■

Ejemplo 3.6 Si aplicamos este método a la función del Ejercicio 3.4, en el soporte $\{-4, -2, 0, 2, 4\}$ obtenemos el polinomio de grado 8 (en realidad se busca de grado 9 pero al ser una función par, el término de grado 9 se anula)

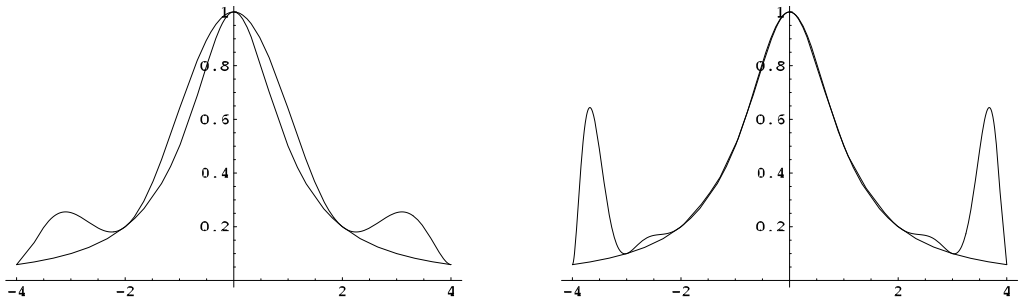


Figura 3.6: Las funciones $\{f(x),\ P_8(x)\}$ y $\{f(x),\ P_{16}(x)\}$

$$P_8(x) = \frac{1}{7225}(7225 - 3129x^2 + 569x^4 - 41x^6 + x^8)$$

cuya gráfica puede verse en la Figura 3.6 (Izda.).

Si lo hacemos en el soporte $\{-4, -3, -2, -1, 0, 1, 2, 3, 4\}$ obtenemos

$$P_{16}(x) = \frac{1}{2890000}(2890000 - 2558224x^2 + 1613584x^4 - 626688x^6 + 144408x^8 \\ -19527x^{10} + 1507x^{12} - 61x^{14} + x^{16})$$

que podemos ver en la Figura 3.6 (Dcha.).

Si comparamos con los resultados obtenidos en el Ejercicio 3.4, podemos observar la mejora que produce la imposición de que coincidan no sólo los valores de la función, sino que también lo hagan los de su derivada, en los puntos del soporte. Sin embargo, sigue manifestándose el fenómeno de Runge, es decir, se mejora el resultado en la parte central del intervalo, pero en los extremos, la diferencia entre el polinomio interpolador y la función es considerable. □

La manera de evitar el fenómeno de Runge es hacer una interpolación polinomial a trozos, es decir, lo que se conoce como una *interpolación por splines*.

3.3 Interpolación por splines

Consideremos una partición del intervalo $[a, b]$

$$\Delta = \{x_0 = a < x_1 < \cdots < x_{n-1} < x_n = b\}$$

en la que los puntos x_i reciben el nombre de *nodos*. Una interpolación por splines no es más que tomar un soporte en cada subintervalo $[x_{i-1}, x_i]$ y construir un polinomio de interpolación, de grado no superior a k (para un k prefijado) sobre dicho soporte, por lo que el método se conoce también como *interpolación polinomial a trozos*. Damos a continuación una definición formal de lo que denominaremos *función spline*.

Definición 3.2 Una *función spline de grado k* con *nodos* en $x_0, x_1, \ldots, x_n$ es una función $S(x)$ formada por varios polinomios, cada uno de ellos definido sobre un subintervalo y que se unen entre sí bajo ciertas condiciones de continuidad. Las condiciones que debe cumplir $S(x)$ son las siguientes:

a) En cada intervalo $[x_{i-1}, x_i)$, $S(x)$ es un polinomio de grado $\text{gr}[S(x)] \leq k$,

b) $S(x)$ admite derivada continua de orden $k-1$ en $[x_0, x_n]$.

En general, pueden crearse funciones spline de grado k, pero la interpolación más frecuente es a través de funciones spline de grado 3, es decir, de *splines cúbicos*.

3.3.1 Splines cúbicos

Dado que a partir de ahora vamos a trabajar con splines cúbicos, vamos a concretar la Definición 3.2 al caso de $k = 3$.

Definición 3.3 Dado el conjunto de puntos $\Delta = \{x_0, x_1, \ldots, x_n\}$, diremos que la función S_δ es un *spline cúbico* asociado a Δ si cumple las siguientes condiciones:

a) La restricción de S_Δ a cada intervalo $[x_{i-1}, x_i)$ para $i = 1, 2, \ldots n$ es un polinomio de grado no superior a tres. Es decir, $S_{\Delta|_{[x_{i-1}, x_i]}} \in \mathcal{P}_3[x]$, donde $\mathcal{P}_3[x]$ representa al conjunto de los polinomios de grado menor o igual a tres.

b) $S_\Delta \in \mathcal{C}^2[a, b]$, es decir, S_Δ es una función continua, dos veces derivable y con derivadas continuas en el intervalo $[a, b]$.

Definición 3.4 Diremos que $S_\Delta(x)$ es un *spline de interpolación* en x según la partición Δ, si

a) $S_\Delta(x)$ es un spline cúbico asociado a Δ.

b) $S_\Delta(x_i) = f(x_i) = y_i$ para $i = 0, 1, \ldots, n$, es decir, cumple las condiciones de interpolación.

Antes de construir un spline cúbico vamos a ver cuántas condiciones ha de cumplir y cuántas incógnitas van a hacernos falta. Si en cada intervalo de la partición intentamos construir un polinomio de grado tres que aproxime a la función, deberemos calcular cuatro incógnitas (los cuatro coeficientes del polinomio de grado tres) por intervalo, es decir, $4n$ incógnitas. Por otro lado, estos polinomios deben cumplir, en cada uno de los nodos, las condiciones:

$$\begin{aligned} S_{\Delta|_{[x_{i-1},x_i]}}(x_i) &= S_{\Delta|_{[x_i,x_{i+1}]}}(x_i) \\ S'_{\Delta|_{[x_{i-1},x_i]}}(x_i) &= S'_{\Delta|_{[x_i,x_{i+1}]}}(x_i) \qquad i = 1, 2, \ldots, n-1 \\ S''_{\Delta|_{[x_{i-1},x_i]}}(x_i) &= S''_{\Delta|_{[x_i,x_{i+1}]}}(x_i) \end{aligned} \tag{3.1}$$

Es decir, se deben cumplir un total de $3(n-1)$ condiciones además de las $n+1$ condiciones de interpolación

$$S_\Delta(x_i) = f(x_i) \qquad i = 0, 1, \ldots, n$$

Dado que tenemos un total de $4n$ incógnitas para $4n-2$ condiciones, debemos imponer dos nuevas condiciones para poder determinar los coeficientes de la función spline. Dependiendo de las condiciones que impongamos, obtendremos un tipo de spline u otro.

- Si exigimos que las derivadas segundas se anulen en los extremos, es decir, si
$$S''_\Delta(a) = S''_\Delta(b) = 0$$
diremos que $S_\Delta(x)$ es el *spline natural* asociado a la partición Δ.

- Si exigimos que
$$S'_\Delta(a) = S'_\Delta(b), \qquad S''_\Delta(a) = S''_\Delta(b)$$
diremos que se trata de un *spline periódico* .

3.3.2 Cálculo de los splines cúbicos de interpolación

Nos centraremos en el cálculo de los splines naturales y con al fin de simplificar la notación, llamaremos

$$\begin{aligned} h_i &= x_i - x_{i-1} \qquad & i = 1, 2, \ldots, n \\ M_i &= S''_\Delta(x_i) \qquad & i = 0, 1, \ldots, n \end{aligned}$$

Los valores M_i se denominan *momentos* y determinarán completamente los splines cúbicos.

Obsérvese, en primer lugar, que como en cada intervalo $[x_i, x_{i+1}]$ el spline S_Δ es un polinomio de grado tres, su segunda derivada es una recta (un polinomio de grado uno). En consecuencia, al imponer las condiciones (3.1) sobre la igualdad de las derivadas segundas en los nodos, obligamos a que la segunda derivada de la función spline $S''_\Delta(x)$ constituya un conjunto de rectas que se intersecan en los nodos de la partición elegida. Ahora bien, dado que cada recta queda determinado por dos puntos, podemos escribir el valor de las restricciones (3.1) sobre $S''_{\Delta|_{[x_i,x_{i+1}]}}$ como

$$S''_{\Delta|_{[x_i,x_{i+1}]}}(x) = M_i \frac{x_{i+1} - x}{h_{i+1}} + M_{i+1} \frac{x - x_i}{h_{i+1}}.$$

Integrando respecto a x obtenemos el valor de la primera derivada del spline en este intervalo

$$S'_{\Delta|_{[x_i,x_{i+1}]}}(x) = -\frac{M_i}{2} \frac{(x_{i+1} - x)^2}{h_{i+1}} + \frac{M_{i+1}}{2} \frac{(x - x_i)^2}{h_{i+1}} + A_i.$$

Volviendo a integrar respecto a x obtenemos

$$S_{\Delta|_{[x_i,x_{i+1}]}}(x) = \frac{M_i}{6} \frac{(x_{i+1} - x)^3}{h_{i+1}} + \frac{M_{i+1}}{6} \frac{(x - x_i)^3}{h_{i+1}} + A_i(x - x_i) + B_i.$$

Si imponemos ahora las condiciones de interpolación

$$S_\Delta(x_i) = y_i, \qquad S_\Delta(x_{i+1}) = y_{i+1}$$

obtenemos

$$\frac{M_i}{6} h_{i+i}^2 + B_i = y_i \Longrightarrow B_i = y_i - \frac{M_i}{6} h_{i+1}^2$$

$$\frac{M_{i+1}}{6} h_{i+i}^2 + A_i h_{i+1} + B_i = y_{i+1} \Longrightarrow A_i = \frac{y_{i+1} - y_i}{h_{i+1}} - \frac{h_{i+1}}{6}(M_{i+1} - M_i).$$

Podemos, por tanto, determinar los valores de las constantes A_i y B_i, que determinan el valor de $S_\Delta(x)$ en el intervalo $[x_i, x_{i+1}]$, en función de los momentos.

El problema se reduce, por tanto, a calcular los momentos para cada uno de los intervalos, para lo que utilizaremos la única condición de (3.1) que no hemos utilizado:

$$S'_{\Delta|_{[x_{i-1},x_i]}}(x_i) = S'_{\Delta|_{[x_i,x_{i+1}]}}(x_i).$$

Esta condición nos da, para cada $i = 1, 2, \ldots, n-1$, una ecuación:

$$\frac{h_i}{h_i + h_{i+1}} M_{i-1} + 2M_i + \frac{h_{i+1}}{h_i + h_{i+1}} M_{i+1} = \frac{6}{h_i + h_{i+1}} \left(\frac{y_{i+1} - y_i}{h_{i+1}} - \frac{y_i - y_{i-1}}{h_i} \right)$$

En el caso del spline natural tenemos que $M_0 = M - n = 0$, quedándonos el sistema tridiagonal de $n-1$ ecuaciones con $n-1$ incógnitas

$$\begin{pmatrix} 2 & \frac{h_2}{h_1+h_2} & & & \\ \frac{h_2}{h_2+h_3} & 2 & \frac{h_3}{h_2+h_3} & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & \frac{h_{n-1}}{h_{n-2}-h_{n-1}} \\ & & & \frac{h_{n-1}}{h_{n-1}+h_n} & 2 \end{pmatrix} \begin{pmatrix} M_1 \\ M_2 \\ \vdots \\ \vdots \\ M_{n-1} \end{pmatrix} =$$

$$= \begin{pmatrix} \frac{6}{h_1+h_2} \left(\frac{y_2 - y_1}{h_2} - \frac{y_1 - y_0}{h_1} \right) \\ \frac{6}{h_2+h_3} \left(\frac{y_3 - y_2}{h_3} - \frac{y_2 - y_1}{h_2} \right) \\ \vdots \\ \vdots \\ \frac{6}{h_{n-1}+h_n} \left(\frac{y_n - y_{n-1}}{h_n} - \frac{y_{n-1} - y_{n-2}}{h_{n-1}} \right) \end{pmatrix}$$

Este sistema puede resolverse por cualquiera de los métodos iterados estudiados en el Capítulo 2 ya que, al ser la matriz del sistema de diagonal dominante, todos ellos son convergentes.

Ejemplo 3.7 Si aplicamos le interpolación por splines cúbicos a la función del Ejemplo 3.4

$$f(x) = \frac{1}{1+x^2} \qquad \text{en la partición} \qquad \Delta = \{-4, -3, -2, -1, 0, 1, 2, 3, 4\}$$

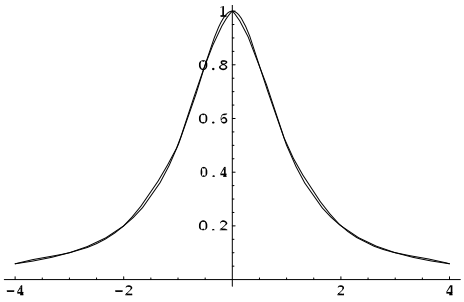


Figura 3.7: Las funciones $f(x)$ y $S_\Delta(x)$

obtenemos, utilizando MATHEMATICA, el resultado de la Figura 3.7, que puede verse que, independientemente de ser mejor que el que se obtuvo en la Figura 3.6 (Dcha.) con el método de Hermite, no aparece el fenómeno de Runge. □

3.4 Ejercicios

Ejercicio 3.1 Calcular los polinomios de Lagrange para el soporte canónico con $1 \leq n \leq 3$.

Ejercicio 3.2 Hallar el polinomio de interpolación de la función $f(x) = 2x^4$ en el soporte canónico $S = \{0, 1, 2, 3\}$. Obtener una expresión del error.

Ejercicio 3.3 Hallar el polinomio de interpolación de la función $f(x) = e^x$ en el soporte $\{0, 1\|\}$ y con él, aproximar $\sqrt{e}$ estimando el error cometido.

Ejercicio 3.4 Obtener el polinomio de interpolación de los puntos:

$$(0, -5),\ (1, -3),\ (2, 1),\ (3, 13)$$

a) Mediante resolución de un sistema de ecuaciones.

b) Mediante la fórmula de Lagrange

c) Mediante la fórmula de Newton para diferencias divididas.

d) Mediante la fórmula de Newton para diferencias finitas.

Ejercicio 3.5 Probar que $F(n) = 1^2 + 2^2 + 3^2 + \cdots + n^2$ es un polinomio en n y obtenerlo por interpolación.

Ejercicio 3.6 Obtener el polinomio de interpolación de Hermite de la función $f(x) = \ln x$ en el soporte $S = \{1, 2\}$ y, supuesto conocido $\ln 2$, aproximar el valor de $\ln 1'5$ acotando el error cometido.

Ejercicio 3.7 Dada la función $f(x) = e^x$, se pide:

a) Hallar el polinomio de interpolación en el soporte $\{-1, 0, 1\}$ y una cota del error en el intervalo $[-1, 1]$.
Calcular $P(0'01)$ y compararlo con el valor dado por la calculadora para $e^{0'01}$.

b) Hallar el polinomio de segundo grado, de mejor aproximación, en el intervalo $[-1, 1]$, respecto a la norma inducida por el producto escalar

$$< f, g >= \int_{-1}^{1} f(x)g(x)dx.$$

Calcular el error cometido en la aproximación de $e^{0'01}$ y comparar su valor con el que nos da la calculadora.

Ejercicio 3.8 Dada la tabla

x	0	1	2	3
y	-1	6	31	98

se pide:

a) Hallar su polinomio de interpolación por el método de los polinomios de Lagrange.

b) Determinar la forma general de todos los polinomios de cuarto grado que satisfacen dicha tabla, determinando aquel que verifica que para $x = 4$ es $y = 255$.

c) Determinar los polinomios anteriores (para los soportes $\{0, 1, 2, 3\}$ y $\{0, 1, 2, 3, 4\}$) por el método de las diferencias divididas de Newton.

4. Integración numérica

4.1 Introducción

Se pretende, en este tema, dar una aproximación numérica del valor de una integral $\int_a^b f(x)dx$ en los distintos problemas que se presentan en la práctica, como son:

a) Conocida una primitiva $F(x)$ de la función $f(x)$ sabemos que

$$\int_a^b f(x)dx = F(b) - F(a)$$

pero necesitamos aproximar el valor de $F(b) - F(a)$.

Así, por ejemplo,

$$\int_1^2 \frac{1}{x}dx = \Big[\log x\Big]_1^2 = \log 2 - \log 1 = \log 2$$

pero hay que aproximar el valor de $\log 2$.

b) Si se conoce la función $f(x)$, pero no se conoce ninguna primitiva suya, se busca otra función $g(x)$ que aproxime a la función $f(x)$ y de la cual sí se conozcan primitivas.

Así, por ejemplo, para calcular $\int_1^2 \frac{e^x}{x}dx$, se desarrolla en serie de potencias

$$f(x) = \frac{e^x}{x} = \frac{1 + x + \cdots + \frac{x^n}{n!}}{x} + \varepsilon(x) = \frac{1}{x} + 1 + \cdots + \frac{x^{n-1}}{n!} + \varepsilon(x) =$$
$$= g(x) + \varepsilon(x)$$

$$\int_1^2 f(x)dx = \int_1^2 g(x)dx + \int_1^2 \varepsilon(x)dx$$

en donde habrá que evaluar $\int_1^2 \varepsilon(x)dx$.

c) Sólo se conocen los valores de $f(x)$ en un soporte $\{x_0, x_1, \ldots, x_n\}$.

En éste caso, se interpola la función (por ejemplo mediante la interpolación polinómica).

$$\int_a^b f(x)dx = \int_a^b P_n(x)dx + \int_a^b \frac{f^{(n+1}(c)}{(n+1)!}(x-x_0)(x-x_1)\cdots(x-x_n)dx$$

$$\int_a^b f(x)dx = \int_a^b P_n(x)dx + \int_a^b \varepsilon(x)dx$$

4.2 Fórmulas de cuadratura

Si realizamos la interpolación de Lagrange, y llamamos

$$z(x) = (x-x_0)(x-x_1)\cdots(x-x_n),$$

el polinomio de interpolación es

$$P_n(x) = y_0L_0(x) + y_1L_1(x) + \cdots + y_nL_n(x)$$

en donde los polinomios de Lagrange $L_i(x)$ pueden expresarse de la forma

$$L_i(x) = \frac{z(x)}{(x-x_i)\ \ z'(x_i)}$$

Además,

$$\int_a^b P_n(x) = \sum_{i=0}^n \int_a^b y_iL_i(x)dx = \sum_{i=0}^n y_i \int_a^b L_i(x)dx = \sum_{i=0}^n a_iy_i$$

donde los coeficientes $a_i = \int_a^b L_i(x)dx$ no dependen de la función, sino sólo del soporte.

Además, si f es un polinomio de grado no superior a n, $\varepsilon(x) = 0$, por lo que *para polinomios* es

$$\int_a^b P(x)dx = \sum_{i=1}^n a_iP(x_i)$$

Por tanto:

$$\left.\begin{array}{ll}
P(x) = 1 & \Longrightarrow b-a = a_0 + a_1 + \cdots + a_n \\
P(x) = x & \Longrightarrow \dfrac{b^2-a^2}{2} = a_0x_0 + a_1x_1 + \cdots + a_nx_n \\
\cdots\cdots\cdots & \cdots\cdots\cdots\cdots\cdots\cdots \\
\cdots\cdots\cdots & \cdots\cdots\cdots\cdots\cdots\cdots \\
P(x) = x^n & \Longrightarrow \dfrac{b^{n+1}-a^{n+1}}{n+1} = a_0x_0^n + a_1x_1^n + \cdots + a_nx_n^n
\end{array}\right\} \qquad (4.1)$$

sistema que, en forma matricial es

$$\begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_0 & x_1 & \cdots & x_n \\ \vdots & \vdots & \ddots & \vdots \\ x_0^n & x_1^n & \cdots & x_n^n \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_n \end{pmatrix} = \begin{pmatrix} b-a \\ \frac{b^2-a^2}{2} \\ \vdots \\ \frac{a^{n+1}-b^{n+1}}{n+1} \end{pmatrix}$$

cuyo determinante es un Vandermonde.

Una vez calculados los coeficientes a_i se obtiene una fórmula de aproximación que sólo dependerá del soporte. Para cada soporte, las fórmulas reciben el nombre de *fórmulas de cuadratura* .

Ejemplo 4.1 Vamos a integrar una función $f(x)$ en $[0,1]$ considerando los soportes:

$$S_1 = \{0, \frac{1}{3}, \frac{1}{2}\} \qquad \text{y} \qquad S_2 = \{\frac{1}{4}, \frac{1}{2}, \frac{3}{4}\}.$$

a) En el soporte $S_1 = \{0, \frac{1}{3}, \frac{1}{2}\}$

$$\int_0^1 f(x)dx \simeq a_0 f(0) + a_1 f(\frac{1}{3}) + a_2 f(\frac{1}{2})$$

El sistema a resolver es, en este caso:

$$\begin{aligned} P(x) = 1 &\implies a_0 + a_1 + a_2 = 1 \\ P(x) = x &\implies 0 \cdot a_0 + \frac{1}{3}a_1 + \frac{1}{2}a_2 = \frac{1}{2} \\ P(x) = x^2 &\implies 0 \cdot a_0 + \frac{1}{9}a_1 + \frac{1}{4}a_2 = \frac{1}{3} \end{aligned}$$

cuya solución es

$$a_0 = \frac{1}{2} \qquad a_1 = -\frac{3}{2} \qquad a_2 = 2$$

luego

$$\int_0^1 f(x)dx \simeq \frac{1}{2}f(0) - \frac{3}{2}f(\frac{1}{3}) + 2f(\frac{1}{2})$$

b) En el soporte $S_2 = \{\frac{1}{4}, \frac{1}{2}, \frac{3}{4}\}$

$$\int_0^1 f(x)dx \simeq b_0 f(\frac{1}{4}) + b_1 f(\frac{1}{2}) + b_2 f(\frac{3}{4})$$

El sistema a resolver es, en este caso:

$$P(x) = 1 \implies b_0 + b_1 + b_2 = 1$$
$$P(x) = x \implies \frac{1}{4}b_0 + \frac{1}{2}b_1 + \frac{3}{4}b_2 = \frac{1}{2}$$
$$P(x) = x^2 \implies \frac{1}{16}b_0 + \frac{1}{4}b_1 + \frac{9}{16}b_2 = \frac{1}{3}$$

cuya solución es

$$b_0 = \frac{2}{3} \qquad b_1 = -\frac{1}{3} \qquad b_2 = \frac{2}{3}$$

luego

$$\int_0^1 f(x)dx \simeq \frac{2}{3}f(\frac{1}{4}) - \frac{1}{3}f(\frac{1}{2}) + \frac{2}{3}f(\frac{3}{4})$$ □

Las formas más generalizadas de aproximación de la integral de una función $f(x)$ se realizan mediante uno de los dos procesos siguientes:

a) Dando un soporte (generalmente regular) y los valores de la función en los puntos del soporte. *Fórmulas de Newton-Cötes* .

b) Dando diferentes soportes y buscando el polinomio $P(x)$ que hace más pequeña la integral $\int_a^b (f(x) - P(x))dx$. *Fórmulas de Gauss* que no se verán en este curso.

4.3 Fórmulas de Newton-Cötes

Partamos del soporte regular $\{x_0, x_1, \ldots, x_n\}$ con

$$x_0 = a \quad x_1 = a + h \quad \cdots \quad x_i = a + ih \quad \cdots \quad x_n = a + nh = b$$

Si llamamos $z(x) = (x - x_0)(x - x_1)\cdots(x - x_n)$ se tiene que los polinomios de Lagrange son

$$L_i(x) = \frac{(x - x_0)(x - x_1)\cdots(x - x_n)}{(x - x_i)(x_i - x_0)\cdots(x_i - x_{i-1})(x_i - x_{in1})\cdots(x_i - x_n)} =$$
$$= \frac{(x - a)(x - a - h)\cdots(x - a - nh)}{(x - a - ih)\, ih\, (i-1)h\cdots h\, (-h)\, (-2h)\cdots(-(n-i)h)} =$$

$$= \frac{(x-a)(x-a-h)\cdots(x-a-nh)}{(x-a-ih)\cdot i!\cdot(n-i)!\cdot h^{n-1}\cdot(-1)^{n-i}} =$$

$$= h\frac{\left(\frac{x-a}{h}\right)\left(\frac{x-a}{h}-1\right)\cdots\left(\frac{x-a}{h}-n\right)}{(x-a-ih)\cdot i!\cdot(n-1)!\cdot(-1)^{n-i}} =$$

$$= \frac{\left(\frac{x-a}{h}\right)\left(\frac{x-a}{h}-1\right)\cdots\left(\frac{x-a}{h}-n\right)}{\left(\frac{x-a}{h}-i\right)\cdot i!\cdot(n-1)!\cdot(-1)^{n-i}}$$

Por lo que haciendo $t = \dfrac{x-a}{h}$ se tiene que

$$L_i(x) = \frac{t(t-1)\cdots(t-n)}{(t-i)\cdot i!\cdot(n-i)!\cdot(-1)^{n-i}}$$

Por tanto:

$$a_i = \int_0^n L_i(x)dx = \int_0^n \frac{t(t-1)\cdots(t-n)}{(t-i)\cdot i!\cdot(n-i)!\cdot(-1)^{n-i}}hdt =$$

$$= \frac{(-1)^{n-i}h}{i!\cdot(n-i)!}\int_0^n \frac{t(t-1)\cdots(t-n)}{t-i}dt \implies$$

$$a_i = h(-1)^{n-i}\frac{\binom{n}{i}}{n!}\int_0^n \frac{z(t)}{t-i}dt$$

que son los *coeficientes de Newton-Cötes* .

Proposición 4.1 *Los coeficientes de Cötes verifican que* $a_k = a_{n-k}$.

Demostración.

$$a_{n-k} = h(-1)^k\frac{\binom{n}{n-k}}{n!}\int_0^n \frac{z(t)}{t-(n-k)}dt$$

Haciendo $\begin{cases} t-n=-u \\ dt=-du \end{cases}$ se tiene que

$$a_{n-k} = h(-1)^{k+1}\frac{\binom{n}{n-k}}{n!}\int_0^n \frac{(-u+n)(-u+n-1)\cdots(-u)}{-u+k}du =$$

$$= h(-1)^{k+1+n+1}\frac{\binom{n}{n-k}}{n!}\int_0^n \frac{u(u-1)\cdots(u-n)}{u-k}du =$$

$$= h(-1)^{n+k}\frac{\binom{n}{k}}{(-1)^{2k}n!}\int_0^n \frac{u(u-1)\cdots(u-n)}{u-k}du =$$

$$= h(-1)^{n-k}\frac{\binom{n}{k}}{n!}\int_0^n \frac{z(u)}{u-k}du = a_k$$

Teniendo en cuanta la Proposición 4.1, sólo hay que calcular la mitad de los coeficientes.

Las Fórmulas de Newton-Cötes en los casos $n = 1$ y $n = 2$ son conocidas como *Fórmula del trapecio* y *Fórmula de Simpson* respectivamente.

4.3.1 Fórmula del trapecio

La fórmula de Newton-Cötes en el caso $n = 1$ sólo tiene dos coeficientes. Como por la Proposición 4.1 es $a_0 = a_1$ y por las ecuaciones (4.1) es $a_o + a_1 = b - a$, se tiene que

$$a_0 = a_1 = \frac{b-a}{2}$$

por lo que

$$\int_a^b f(x)dx = \frac{b-a}{2}f(a) + \frac{b-a}{2}f(b) = (b-a)\Big(\frac{f(a)+f(b)}{2}\Big)$$

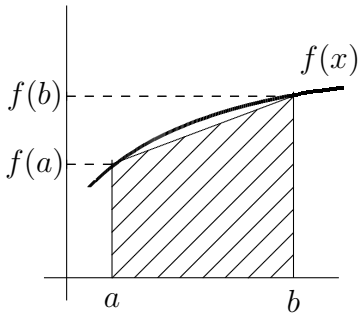


Figura 4.1: Método del trapecio

Es decir, el método del trapecio nos aproxima la integral por el área de la región plana limitada por las rectas $x = a$, $x = b$, $y = 0$ y la recta que pasa por los puntos $(a, f(a))$ y $(b, f(b))$, es decir, el área de un trapecio (ver Figura 4.1).

4.3.2 Fórmula de Simpson

Para el caso $n = 2$ tenemos que $x_0 = a$, $x_1 = \dfrac{a+b}{2}$ y $x_2 = b$.

$$a_0 = h(-1)^{n-0}\frac{\binom{n}{0}}{n!}\int_0^2 \frac{t(t-1)(t-2)}{t-0}dt = \frac{b-a}{4}\Big[\frac{t^3}{3} - 3\frac{t^2}{2} + 2t\Big]_0^2 = \frac{b-a}{6}$$

$$a_2 = a_0 = \frac{b-a}{6}$$

Dado que $a_0 + a_1 + a_2 = b - a$, se tiene que $a_1 = b - a - 2\dfrac{b-a}{6} = \dfrac{2(b-a)}{3}$.

Se tiene, por tanto, que

$$\int_a^b f(x)dx = \frac{b-a}{6}f(a) + \frac{2(b-a)}{3}f(\frac{a+b}{2}) + \frac{b-a}{6}f(b)$$

o, lo que es lo mismo:

$$\int_a^b f(x)dx = \frac{b-a}{6}\Big[f(a) + 4f(\frac{a+b}{2}) + f(b)\Big]$$

Teorema 4.1 *Al aplicar la fórmula de Newton-Cötes para un entero n, el error que se comete viene dado por:*

a) Si n es par:

$$\varepsilon_n = \frac{h^{n+3}f^{(n+2)}(c)}{(n+2)!}\int_0^n t \cdot t(t-1)\cdots(t-n)dt$$

b) Si n es impar:

$$\varepsilon_n = \frac{h^{n+2}f^{(n+1)}(c)}{(n+1)!}\int_0^n (t-1)\cdots(t-n)dt$$

Corolario 4.2 *El error cometido en la aproximación numérica de una integral es:*

a) Para la fórmula del trapecio:

$$\varepsilon = -\frac{h^3 \cdot f''(c)}{12}$$

b) Para el método de Simpson:

$$\varepsilon = -\frac{h^5 \cdot f^{(iv)}(c)}{90}$$

4.4 Fórmulas compuestas

4.4.1 Simpson para n par

Descomponiendo el soporte en $\{x_0, x_1, x_2\} \cup \{x_2, x_3, x_4\} \cup \cdots \cup \{x_{n-2}, x_{n-1}, x_n\}$ se obtiene que

$$\begin{aligned}\int_a^b f(x)dx \simeq \frac{x_2 - x_0}{6}\Big[f(x_0) + 4f(x_1) + f(x_2)\Big] + \\ + \frac{x_4 - x_2}{6}\Big[f(x_2) + 4f(x_3) + f(x_4)\Big] + \cdots + \\ + \cdots + \frac{x_n - x_{n-2}}{6}\Big[f(x_{n-2}) + 4f(x_{n-1}) + f(x_n)\Big] = \\ = \frac{h}{3}\Big[f(x_0) + 4f(x_1) + 2f(x_2) + 4f(x_3) + 2f(_4) + \cdots + \\ + \cdots + 2f(x_{n-2}) + 4f(x_{n-1}) + f(x_n)\Big] = \\ = \frac{h}{3}\Big[\Big(f(x_0) + f(x_n)\Big) + 2\Big(f(x_2) + \cdots + f(x_{n-2})\Big) + 4\Big(f(x_3) + \cdots + f(x_{n-1})\Big)\Big]\end{aligned}$$

El error viene dado por

$$|\varepsilon| = \frac{h^5}{90}\Big[f^{(iv}(c_0) + f^{(iv}(c_2) + \cdots + f^{(iv}(c_{\frac{n}{2}})\Big] \leq \frac{h^5}{90} \max_{x\in[x_0,x_n]} |f^{(iv}(x)| \cdot \frac{n}{2} \Longrightarrow$$

$$|\varepsilon| \leq \frac{(b-a)^5}{180\, n^4} \max_{x\in[x_0,x_n]} |f^{(iv}(x)|$$

4.4.2 Trapecios para n impar

Con un proceso análogo al anterior obtenemos que

$$\begin{aligned}\int_a^b f(x)dx \simeq \frac{h}{2}\Big[\Big(f(x_0) + f(x_1)\Big) + \Big(f(x_1) + f(x_2)\Big) + \cdots + \\ + \cdots + \Big(f(x_{n-1} + f(x_n)\Big)\Big] \Longrightarrow\end{aligned}$$

$$\int_a^b f(x)dx \simeq \frac{b-a}{2n}\Big[\Big(f(x_0) + f(x_n)\Big) + 2\Big(f(x_1) + f(x_2) + \cdots + f(x_{n-1})\Big)\Big]$$

El error que se comete viene dado por

$$|\varepsilon| \leq \frac{h^3}{12}\Big[f''(c_1) + f''(c_2) + \cdots + f''(c_n)\Big] \leq \frac{h^3}{12} \cdot n \cdot \max_{x\in[x_0,x_n]} |f''(x)| \implies$$

$$|\varepsilon| = \frac{(b-a)^3}{12\,n^2} \max_{x\in[x_0,x_n]} |f''(x)|$$

4.5 Ejercicios

Ejercicio 4.1 Probar que los coeficientes a_k de las fórmulas de Newton-Cötes verifican:

$$\sum_{k=0}^{n} \frac{(-1)^k a_k}{\binom{n}{k}} = 0.$$

Ejercicio 4.2 Dada la integral $\displaystyle\int_0^1 \frac{1-x^2}{1+x^2}\,dx$ se pide:

a) Calcularla exactamente.

b) Calcularla, aproximadamente, por la fórmula básica de Simpson.

c) Calcularla por la fórmula compuesta de Simpson de 11 sumandos.

d) Aplicar la siguiente fórmula:

$$\int_{-1}^{1} f(x)\,dx \simeq \frac{1}{9}\Big[5f(-\sqrt{3/5}) + 8f(0) + 5f(\sqrt{3/5})\Big]$$

comprobando que integra, exactamente, polinomios de grado menor o igual que 5.

Ejercicio 4.3 Se considera el soporte $\{-1, c, 1\}$ donde $c \in (-1, 1)$ es fijo. Sea $f(x) \in \mathcal{C}[-1, 1]$

a) Obtener el polinomio de interpolación de $f(x)$ y una expresión del error.

b) Determinar los coeficientes a_0, a_1, a_2 en la fórmula de cuadratura

$$\int_{-1}^{1} f(x)\,dx \simeq a_0 f(-1) + a_1 f(c) + a_2 f(1)$$

para que integre, exactamente, polinomios del mayor grado posible.

c) Dar una condición, necesaria y suficiente, para que dicha fórmula sea exacta para polinomios de tercer grado.

d) Aplicar la fórmula a $f(x) = \sqrt{\dfrac{5x+13}{2}}$ con $c = 0'1$ y comparar con el valor exacto.

Ejercicio 4.4 Calcular $\int_0^1 f(x) \ln x \, dx$ interpolando $f(x)$, por un polinomio de tercer grado, en el soporte $\{0, 1/3, 2/3, 1\}$ y aplicar el resultado al cálculo de $\int_0^1 \operatorname{sen} x \ln x \, dx$.
AYUDA: $\int_0^1 x^m \ln x \, dx = \dfrac{-1}{(m+1)^2}$ $(m \geq 0)$

Ejercicio 4.5 Se considera una fórmula de cuadratura del tipo:

$$\int_{-1}^{1} f(x) \, dx = a_0 f(-1/2) + a_1 f(0) + a_2 f(1/2) + E$$

a) Determinar los coeficientes a_0, a_1 y a_2, para que sea exacta para polinomios del mayor grado posible.

b) Sea $P_3(x)$ el polinomio de interpolación "mixta" tal que:

$P_3(-1/2) = f(-1/2)$, $P_3(0) = f(0)$, $P_3'(0) = f'(0)$ y $P_3(1/2) = f(1/2)$.

Razonar que la fórmula de cuadratura obtenida en el apartado a) es la misma que se obtendría integrando el polinomio $P_3(x)$.

c) Como consecuencia, deducir que el error de la fórmula puede expresarse de la forma:

$$E = \int_{-1}^{1} \frac{f^{(IV)}(c_x)}{4!} \, x^2 (x - \frac{1}{2})(x + \frac{1}{2}) \, dx.$$

Ejercicio 4.6 Probar que la fórmula compuesta de los trapecios para el intervalo $[0, 2\pi]$:

$$\int_0^{2\pi} f(x) \, dx = \frac{h}{2}[f(0) + 2f(h) + 2f(2h) + \cdots + 2f((n-1)h) + f(2\pi)] + E$$

$(h = 2\pi/n)$ integra, exactamente, las funciones:

$$1, \ \operatorname{sin} x, \ \cos x, \ \operatorname{sin} 2x, \ \cos 2x, \ldots, \ \operatorname{sin}(n-1)x, \ \cos(n-1)x.$$

Ejercicio 4.7 Determinar el número de sumandos necesarios, en las fórmulas compuestas de los trapecios y Simpson, para calcular, con seis cifras decimales exactas, las siguientes integrales:

$$a) \ I = \int_1^2 \ln x \, dx. \qquad b) \ I = \int_2^3 \frac{e^x}{x} \, dx.$$

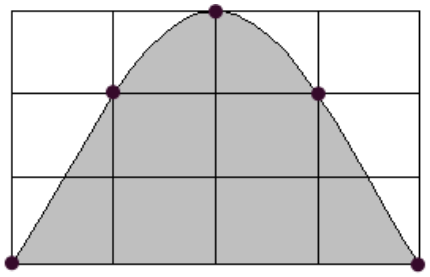
Ejercicio 4.8 Se considera la integral $\int_0^1 e^x(4-x)\,dx$:

a) Calcularla exactamente (se supone conocido el número e).

b) Determinar el número mínimo de sumandos necesarios, en la fórmula compuesta de Simpson, para que el error de discretización sea menor que 10^{-m} con $m = 2, 3, 4, 5$ y 6.

c) Calcular la integral, por la fórmula compuesta de Simpson, con cuatro cifras decimales exactas.

Ejercicio 4.9 El recinto de la figura adjunta, que se encuentra inmerso en una cuadrícula, está limitado por una recta y una curva de la que se conoce que se trata de un polinomio de cuarto grado.



a) Calcular el área exacta del recinto sin determinar el polinomio que la delimita.

b) Determinar, por el método de las diferencias divididas, el polinomio que la delimita y comprobar que el área calculada en el apartado anterior coincide con la que se obtiene por integración directa del polinomio.

Índice