1. Métodos para resolver sistemas lineales de ecuaciones

Una fuente importante de problemas que requieren resolver sistemas lineales son las ecuaciones diferenciales, y generalmente dichos sistemas poseen una estructura *dispersa* (la cantidad de entradas nulas es considerablemente mayor que las entradas no nulas) que puede ser aprovechada. Se presenta un ejemplo sencillo, pero ilustrativo, de un problema con condiciones de frontera cuyas soluciones se aproximan resolviendo un sistema lineal.

Considere el problema unidimensional con condiciones de frontera

$$-u''(x) = f(x), \quad x \in (0,1) \tag{1}$$
$$u(0) = u(1) = 0 \tag{2}$$

No se pretende buscar la solución exacta de la ecuación diferencial, sino una aproximación de los valores de la solución en un conjunto de puntos dado. Con este objetivo, discretizamos uniformemente el intervalo $[0,1]$ tomando $n+2$ puntos equiespaciados

$$x_i = ih, \qquad i = 0, \ldots, n+1$$

donde $h = 1/(n+1)$ es la anchura de la malla. Para aproximar la u'' usamos los desarrollos de Taylor (asumiendo que $u \in C^4(0,1)$)

$$u(x+h) = u(x) + hu'(x) + \frac{h^2}{2}u''(x) + \frac{h^3}{6}u'''(x) + \frac{h^4}{24}u^{(4)}(\xi_1)$$
$$u(x-h) = u(x) - hu'(x) + \frac{h^2}{2}u''(x) - \frac{h^3}{6}u'''(x) + \frac{h^4}{24}u^{(4)}(\xi_2)$$

$\xi_1 \in [x, x+h]$ y $\xi_2 \in [x-h, x]$. Sumando miembro las dos ecuaciones y teniendo en cuenta que por el teorema del valor intermedio, existe ξ entre ξ_1 y ξ_2 tal que $u^{(4)}(\xi) = \frac{1}{2}(u^{(4)}(\xi_1) + u^{(4)}(\xi_2))$, resulta

$$u''(x) = \frac{u(x+h) - 2u(x) + u(x-h)}{h^2} - \frac{h^2}{12}u^{(4)}(\xi).$$

Luego,

$$u''(x) \approx \frac{u(x+h) - 2u(x) + u(x-h)}{h^2},$$

donde la aproximación es de orden dos. Si denotamos por u_i una aproximación para $u(x_i)$ y $f_i = f(x_i)$, las incógnitas u_i, u_{i-1} y u_{i+1} satisfacen la relación

$$\frac{-u_{i-1} + 2u_i - u_{i+1}}{h^2} = f_i.$$

Note que para $i = 1$ y $i = n$ las ecuaciones involucran a u_0 y u_{n+1}, los cuales son cantidades conocidas, iguales a cero en este caso. Se obtiene así el sistema lineal

$$A\boldsymbol{u} = \boldsymbol{f},$$

donde

$$A = \frac{1}{h^2}\begin{pmatrix} 2 & -1 & & & & \\ -1 & 2 & -1 & & & \\ & -1 & 2 & -1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & -1 & 2 & -1 \\ & & & & -1 & 2 \end{pmatrix}, \quad \boldsymbol{u} = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix} \quad \text{y} \quad \boldsymbol{f} = \begin{pmatrix} f_1 \\ f_2 \\ \vdots \\ f_n \end{pmatrix}.$$

Considere una viga de sección rectangular uniforme y longitud L empotrada en ambos extremos y sujeta a una distribución de carga $f(x)$ que varía a lo largo de la coordenada x. Bajo el supuesto que se presentan pequeños desplazamientos, la flexión transversal de la viga es gobernada por la ecuación diferencial de cuarto orden

$$u^{(4)}(x) = f(x), \quad x \in (0, L) \tag{3}$$

donde $u(x)$ denota el dezplazamiento vertical. Puesto que la viga está sujeta en ambos extremos, las condiciones de frontera son

$$u(0) = u(L) = 0 \quad \text{y} \quad u'(0) = u'(L) = 0. \tag{4}$$

Para resolver numéricamente el problema, se usa el método de diferencias finitas. Nuevamente hacemos una partición uniforme del intervalo $[0, L]$ tomando $n + 2$ puntos equiespaciados

$$x_i = ih, \qquad i = 0, \ldots, n + 1$$

donde $h = L/(n + 1)$ es la anchura de la malla. Al usar desarrollos de Taylor se obtiene la siguiente aproximación de segundo orden de $u^{(4)}$

$$u^{(4)}(x) \approx \frac{u(x - 2h) - 4u(x - h) + 6u(x) - 4u(x + h) + u(x + 2h)}{h^4}$$

Si denotamos por u_i la aproximación para $u(x_i)$ y $f_i = f(x_i)$, se sigue que, para los nodos interiores $1 \leq i \leq n$, la derivada de cuarto orden es aproximada por el esquema de diferencias finitas de segundo orden:

$$u^{(4)}(x_i) \approx \frac{u_{i-2} - 4u_{i-1} + 6u_i - 4u_{i+1} + u_{i+2}}{h^4}.$$

Luego, la discretización de la ecuación diferencial es

$$\frac{1}{h^4}\left(u_{i-2} - 4u_{i-1} + 6u_i - 4u_{i+1} + u_{i+2}\right) = f_i. \tag{5}$$

Para $i = 2$ y $i = n-1$ se imponen las condiciones de frontera $u_0 = u(0) = 0$ y $u_{n+1} = u(L) = 0$. Así, las incógnitas son $u_1, u_2, \ldots, u_n$. Para $i = 1$ y $i = n$, los puntos u_{-1} y u_{n+2} están fuera del dominio computacional (reciben el nombre de *puntos ficticios*) y se eliminan usando un esquema central de aproximación para las condiciones de frontera sobre la derivada de u. Más exactamente, de las condiciones de frontera $u'(0) = u'(L) = 0$ y la aproximación

$$u'(x_i) \approx \frac{u_{i+1} - u_{i-1}}{2h},$$

resulta

$$0 = u_0' = \frac{u_1 - u_{-1}}{2h} \quad \text{y} \quad 0 = u_{n+1}' = \frac{u_{n+2} - u_n}{2h}.$$

Por lo tanto, $u_{-1} = u_1$ y $u_{n+2} = u_n$. Así, el sistema de ecuaciones lineales resultante es

$$A\boldsymbol{u} = h^4 \boldsymbol{f}, \tag{6}$$

donde

$$A = \begin{pmatrix} 7 & -4 & 1 & 0 & & \cdots & 0 \\ -4 & 6 & -4 & \ddots & \ddots & & \vdots \\ 1 & -4 & \ddots & \ddots & & \ddots & \\ 0 & \ddots & \ddots & & \ddots & \ddots & 0 \\ & \ddots & & & \ddots & -4 & 1 \\ \vdots & & \ddots & \ddots & -4 & 6 & -4 \\ 0 & \cdots & & 0 & 1 & -4 & 7 \end{pmatrix}, \quad \boldsymbol{u} = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix} \quad \text{y} \quad \boldsymbol{f} = \begin{pmatrix} f_1 \\ f_2 \\ \vdots \\ f_n \end{pmatrix}.$$

La matriz A es pentadiagonal y admite la descomposición (ver [5])

$$A = T^2 + MM^\top, \tag{7}$$

siendo T la matriz tridiagonal

$$T = \begin{pmatrix} 2 & -1 & & & & \\ -1 & 2 & -1 & & & \\ & -1 & 2 & -1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & -1 & 2 & -1 \\ & & & & -1 & 2 \end{pmatrix} \quad \text{y} \quad M = \begin{pmatrix} \sqrt{2} & 0 \\ 0 & 0 \\ \vdots & \vdots \\ 0 & \sqrt{2} \end{pmatrix} \in \mathbb{R}^{n\times 2}.$$

1. Normas vectoriales y matriciales

Los efectos de los errores de redondeo que se pueden presentar en la solución de sistemas lineales requieren ser medidos para analizarlos y buscar minimizarlos. Para este objetivo se usa el concepto de norma.

Definición 1.1. Sea $\mathcal{V}$ espacio vectorial sobre un campo $\mathbb{K}$ con ($\mathbb{K} = \mathbb{R}$ o $\mathbb{K} = \mathbb{C}$). Se dice que la función $||\cdot||$ de $\mathcal{V}$ en $\mathbb{R}$ es una norma si satisface los siguientes axiomas:

N1. $||\boldsymbol{v}|| \geq 0$ $\forall \boldsymbol{v} \in \mathcal{V}$ y $||\boldsymbol{v}|| = 0$ si y solo si $\boldsymbol{v} = \boldsymbol{0}$;

N2. $||\alpha \boldsymbol{v}|| = |\alpha|\,||\boldsymbol{v}||$ $\forall \boldsymbol{v} \in \mathcal{V}$ y $\forall \alpha \in \mathbb{K}$ (Propiedad de homogeneidad);

N3. $||\boldsymbol{v} + \boldsymbol{w}|| \leq ||\boldsymbol{v}|| + ||\boldsymbol{w}||$ $\forall \boldsymbol{v}, \boldsymbol{w} \in \mathcal{V}$ (Desigualdad triangular).

Aquí $|\alpha|$ denota el valor absoluto de α si $\mathbb{K} = \mathbb{R}$ o el modulo de α si $\mathbb{K} = \mathbb{C}$.

El par $(\mathcal{V}, ||\cdot||)$ recibe el nombre de *espacio vectorial normado*. Ahora, si $(\mathcal{V}, ||\cdot||)$ es un espacio vectorial normado, es posible definir en $\mathcal{V}$ una métrica a partir de la norma. Más concretamente, la función

$$\begin{aligned} d : \mathcal{V} \times \mathcal{V} &\longrightarrow \mathbb{R} \\ (\boldsymbol{v}, \boldsymbol{w}) &\longrightarrow d(\boldsymbol{v}, \boldsymbol{w}) = ||\boldsymbol{v} - \boldsymbol{w}|| \end{aligned}$$

es una métrica, lo que puede verificarse de manera rutinaria.

Proposición 1.1. Sea $(\mathcal{V}, ||\cdot||)$ un espacio vectorial normado y $(\mathbb{R}, d)$ el espacio métrico real con la métrica usual. Entonces la función

$$\begin{aligned} ||\cdot|| : \mathcal{V} &\longrightarrow \mathbb{R} \\ \boldsymbol{v} &\longrightarrow ||\boldsymbol{v}|| \end{aligned}$$

es continua.

Demostración. Consecuencia directa de la observación previa (mediante la cual se dota a $\mathcal{V}$ con la métrica definida por la norma), la desigualdad

$$|\,||\boldsymbol{v}|| - ||\boldsymbol{w}||\,| \leq ||\boldsymbol{v} - \boldsymbol{w}||$$

y la definición de continuidad. □

Cualquier norma en los espacios $\mathcal{V} = \mathbb{R}^n$ o $\mathcal{V} = \mathbb{C}^n$ se llamará *norma vectorial*. Un ejemplo de norma en $\mathbb{R}^n$ es la norma p o norma de Holder, la

cual se define de la siguiente manera: para un vector $\boldsymbol{v} = (v_1, v_2, \ldots, v_n)^T$

$$||\boldsymbol{v}||_p = \left(\sum_{i=1}^{n} |v_i|^p \right)^{1/p}, \qquad \text{para } 1 \leq p < \infty \tag{8}$$

Cuando $p = 2$, se obtiene la *norma Euclidiana*. Para $p = 1$ es sencillo verificar que (8) satisface los axiomas de una norma. Para $p > 1$, también es fácil verificar los dos primeros axiomas. Para probar la desigualdad triangular se hará uso de la desigualdad de Holder, la cual a su vez puede probarse usando la desigualdad de Young en cuyo enunciado usaremos la siguiente definición: se dice que $p, q > 1$ son *exponentes conjugados* si

$$\frac{1}{p} + \frac{1}{p} = 1.$$

Teorema 1.1 (Desigualdad de Young)**.** Si $p, q > 1$ son exponentes conjugados y a, b son números reales no negativos, entonces

$$ab \leq \frac{a^p}{p} + \frac{b^q}{q} \tag{9}$$

Demostración. Si $a = 0$ o $b = 0$, el resultado se verifica trivialmente. Supongamos entonces que $a > 0$ y $b > 0$ y considere la función

$$g(t) = tb - \frac{t^p}{p} \qquad \text{con } t > 0.$$

Un simple cálculo muestra que g tiene un máximo en $t = b^{\frac{1}{p-1}}$. Luego,

$$g(a) \leq g(b^{\frac{1}{p-1}}), \tag{10}$$

Ahora, usando el hecho de que p y q son exponentes conjugado se tiene que

$$g(b^{\frac{1}{p-1}}) = b^{\frac{p}{p-1}} - \frac{b^{\frac{p}{p-1}}}{p} = \frac{b^q}{q} \qquad \text{y} \qquad g(a) = ab - \frac{a^p}{p}.$$

Reemplazando en (10) se obtiene el resultado. □

Teorema 1.2 (Desigualdad de Holder)**.** Sean $p, q > 1$ exponentes conjugados. Entonces para cualesquiera vectores $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^n$, se tiene que

$$\left| \sum_{i=1}^{n} u_i v_i \right| \leq ||\boldsymbol{u}||_p ||\boldsymbol{v}||_q. \tag{11}$$

Demostración. Si $\boldsymbol{u} = \mathbf{0}$ o $\boldsymbol{v} = \mathbf{0}$, la desigualdad se tiene trivialmente. Suponga ahora que $\boldsymbol{u} \neq \mathbf{0}$ y $\boldsymbol{v} \neq \mathbf{0}$. Usando la desigualdad de Young, se obtiene

$$\begin{aligned}
\frac{1}{||\boldsymbol{u}||_p||\boldsymbol{v}||_q}\left|\sum_{i=1}^{n} u_i v_i\right| &\leq \frac{1}{||\boldsymbol{u}||_p||\boldsymbol{v}||_q}\sum_{i=1}^{n}|u_i v_i| \\
&= \sum_{i=1}^{n}\left|\frac{u_i}{||\boldsymbol{u}||_p}\right|\left|\frac{v_i}{||\boldsymbol{v}||_q}\right| \\
&\leq \sum_{i=1}^{n}\left(\frac{|u_i|^p}{p||\boldsymbol{u}||_p^p} + \frac{|v_i|^q}{q||\boldsymbol{v}||_q^q}\right) \\
&= \frac{1}{p} + \frac{1}{q} \\
&= 1
\end{aligned}$$

de donde se sigue el resultado. □

Teorema 1.3 (Desigualdad de Minkowski)**.** Si $1 \leq p < \infty$ y $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^n$, entonces

$$||\boldsymbol{u} + \boldsymbol{v}||_p \leq ||\boldsymbol{u}||_p + ||\boldsymbol{v}||_p$$

Demostración. Como se mencionó antes, la prueba de la desigualdad para $p = 1$ es fácil, en consecuencia, se desarrolla la prueba para $1 < p < \infty$. Si $\boldsymbol{u} = \mathbf{0}$ o $\boldsymbol{v} = \mathbf{0}$, la desigualdad se verifica trivialmente. En el caso no trivial $\boldsymbol{u} \neq \mathbf{0}$ y $\boldsymbol{v} \neq \mathbf{0}$, sea q es exponente conjugado de p, usando que $(p-1)q = p$ y la desigualdad de Holder, resulta

$$\begin{aligned}
||\boldsymbol{u} + \boldsymbol{v}||_p^p &= \sum_{i=1}^{n}|u_i + v_i|^p \\
&\leq \sum_{i=1}^{n}|u_i + v_i|^{p-1}(|u_i| + |v_i|) \\
&= \sum_{i=1}^{n}|u_i||u_i + v_i|^{p-1} + \sum_{i=1}^{n}|v_i||u_i + v_i|^{p-1} \\
&\leq ||\boldsymbol{u}||_p\left(\sum_{i=1}^{n}|u_i + v_i|^{(p-1)q}\right)^{1/q} + ||\boldsymbol{v}||_p\left(\sum_{i=1}^{n}|u_i + v_i|^{(p-1)q}\right)^{1/q} \\
&= ||\boldsymbol{u}||_p\left(\sum_{i=1}^{n}|u_i + v_i|^{p}\right)^{1/q} + ||\boldsymbol{v}||_p\left(\sum_{i=1}^{n}|u_i + v_i|^{p}\right)^{1/q}
\end{aligned}$$

$$= (||\boldsymbol{u}||_p + ||\boldsymbol{v}||_p)||\boldsymbol{u}+\boldsymbol{v}||_p^{p/q},$$

y el resultado se obtiene dividiendo por $||\boldsymbol{u}+\boldsymbol{v}||_p^{p/q}$ y teniendo en cuenta que $p - p/q = 1$. □

Otro ejemplo de norma vectorial es la norma del máximo, definida por

$$||\boldsymbol{u}||_\infty = \max_{1\le i\le n} |u_i|.$$

Si para un vector no nulo cualquiera $\boldsymbol{u}$, se define $\hat{\boldsymbol{u}} = \boldsymbol{u}/||\boldsymbol{u}||_\infty$, se tiene que

$$1 \le ||\hat{\boldsymbol{u}}||_p \le n^{1/p},$$

y en consecuencia, $\lim_{p\to\infty} ||\hat{\boldsymbol{u}}||_p = 1$. Por lo tanto,

$$\lim_{p\to\infty} ||\boldsymbol{u}||_p = ||\boldsymbol{u}||_\infty.$$

Esta identidad justifica el uso de la notación $||\cdot||_\infty$ para representar la norma del máximo.

Recordamos algunos ideas relacionadas con el *producto escalar Euclidiano* en $\mathbb{C}^n$ (o en $\mathbb{R}^n$), el cual se define por

$$\langle \boldsymbol{u}, \boldsymbol{v}\rangle = \sum_{i=1}^n u_i\overline{v}_i = \boldsymbol{v}^*\boldsymbol{u},$$

donde $\boldsymbol{v}^*$ es el transpuesto conjugado de $\boldsymbol{v}$. Note que $\langle \boldsymbol{u}, \boldsymbol{u}\rangle = ||\boldsymbol{u}||_2^2$.

Sean $A = (a_{ij}) \in \mathbb{C}^{n\times n}$ y $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{C}^n$, entonces

$$\langle A\boldsymbol{u}, \boldsymbol{v}\rangle = \sum_{i=1}^n \left(\sum_{j=1}^n a_{ij}u_j\right)\overline{v}_i.$$

Por otra parte, si $A^* = (b_{kl})$ es la transpuesta conjugada de A, esto es, $b_{kl} = \overline{a}_{lk}$, entonces

$$\begin{aligned}\langle \boldsymbol{u}, A^*\boldsymbol{v}\rangle &= \sum_{k=1}^n u_k \overline{\left(\sum_{l=1}^n b_{kl}v_l\right)} \\ &= \sum_{k=1}^n u_k \left(\sum_{l=1}^n \overline{b}_{kl}\overline{v}_l\right) \\ &= \sum_{k=1}^n u_k \left(\sum_{l=1}^n a_{lk}\overline{v}_l\right)\end{aligned}$$

$$= \sum_{l=1}^{n} \left(\sum_{k=1}^{n} a_{lk} u_k \right) \overline{v}_l$$
$$= \langle A\boldsymbol{u}, \boldsymbol{v} \rangle$$

Se ha obtenido la relación $\langle A\boldsymbol{u}, \boldsymbol{v} \rangle = \langle \boldsymbol{u}, A^*\boldsymbol{v} \rangle$. En particular, si A es una matriz unitaria, entonces

$$\langle A\boldsymbol{u}, A\boldsymbol{v} \rangle = \langle \boldsymbol{u}, A^*A\boldsymbol{v} \rangle = \langle \boldsymbol{u}, \boldsymbol{v} \rangle$$

Consideramos ahora normas en el espacio vectorial $\mathcal{V} = R^{m\times n}$ de las matrices con m filas y n columnas y entradas reales. Tales normas reciben el nombre de *normas matriciales*. Aunque las definiciones y resultados que siguen a continuación se presentan para matrices reales, éstos pueden extenderse también a matrices con entradas complejas, es decir, tomando $\mathcal{V} = \mathbb{C}^{m\times n}$.

Para caracterizar mejor las normas matriciales y seleccionar aquellas de interés práctico, se introducen a continuación dos definiciones adicionales.

Definición 1.2. Se dice que una norma matricial $||\cdot||$ es *compatible* con la norma vectorial $||\cdot||$ si

$$||A\boldsymbol{v}|| \leq ||A||\,||\boldsymbol{v}||, \qquad \forall \boldsymbol{v} \in \mathbb{R}^n$$

Definición 1.3. Se dice que una norma matricial $||\cdot||$ es *sub-multiplicativa* si

$$||AB|| \leq ||A||\,||B||, \qquad \forall A \in \mathbb{R}^{m\times n},\ \forall B \in \mathbb{R}^{n\times p}$$

Si una norma matricial es submultiplicativa, entonces $||A^2|| \leq ||A||\,||A|| = ||A||^2$. Así, para cualquier matriz A tal que $A^2 = A$, se tiene que $||A|| \geq 1$. En particular, $||I|| \geq 1$.

Un ejemplo relevante de normas matriciales son las normas inducidas por normas vectoriales, según se precisa en el siguiente teorema.

Teorema 1.4. Sea $||\cdot||$ una norma vectorial. La función

$$||A|| = \sup_{||\boldsymbol{v}||=1} ||A\boldsymbol{v}|| = \sup_{\boldsymbol{v}\neq\boldsymbol{0}} \frac{||A\boldsymbol{v}||}{||\boldsymbol{v}||} \tag{12}$$

es una norma matricial llamada *norma matricial inducida* (por la norma vectorial) o *norma matricial natural*.

Demostración. Nótese inicialmente que como la función norma es continua y el conjunto $\{\boldsymbol{v} \in R^n : ||\boldsymbol{v}|| = 1\}$ es compacto, el supremo corresponde a

un máximo. Para verificar la segunda igualdad, sea $\beta = \sup_{||\boldsymbol{v}||=1} ||A\boldsymbol{v}||$ y $\gamma = \sup_{\boldsymbol{v}\neq\boldsymbol{0}} \frac{||A\boldsymbol{v}||}{||\boldsymbol{v}||}$. Como

$$\{\boldsymbol{v} \in R^n : ||\boldsymbol{v}|| = 1\} \subset \{\boldsymbol{v} \in R^n : \boldsymbol{v} \neq \boldsymbol{0}\},$$

entonces por propiedades del supremo, se tiene que $\beta \leq \gamma$. Para la otra desigualdad, sea $\boldsymbol{v} \neq \boldsymbol{0}$ cualquiera y considere el vector unitario $\boldsymbol{u} = \boldsymbol{v}/||\boldsymbol{v}||$. Entonces,

$$\frac{||A\boldsymbol{v}||}{||\boldsymbol{v}||} = \left|\left|A\left(\frac{\boldsymbol{v}}{||\boldsymbol{v}||}\right)\right|\right| = ||A\boldsymbol{u}|| \leq \beta.$$

Así, por definición de supremo, $\gamma \leq \beta$. Luego $\beta = \gamma$. Resta verificar que la función (12) satisface los axiomas de norma.

N1. Como la norma vectorial satisface $||A\boldsymbol{v}|| \geq 0$, para todo $\boldsymbol{v}$ entonces $||A|| = \sup_{||\boldsymbol{v}||=1} ||A\boldsymbol{v}|| \geq 0$. Además,

$$||A|| = \sup_{\boldsymbol{v}\neq\boldsymbol{0}} \frac{||A\boldsymbol{v}||}{||\boldsymbol{v}||} = 0 \Longleftrightarrow ||A\boldsymbol{v}|| = 0, \quad \forall \boldsymbol{v} \neq \boldsymbol{0},$$

pero $A\boldsymbol{v} = \boldsymbol{0}$ $\forall \boldsymbol{v} \neq \boldsymbol{0}$ si y solo si $A = 0$; de donde, $||A|| = 0$ si y solo si $A = 0$.

N2. Dado un escalar α, por propiedades del supremo, se tiene que

$$||\alpha A|| = \sup_{||\boldsymbol{v}||=1} ||(\alpha A)\boldsymbol{v}|| = \sup_{||\boldsymbol{v}||=1} ||\alpha(A\boldsymbol{v})|| = |\alpha| \sup_{||\boldsymbol{v}||=1} ||A\boldsymbol{v}|| = |\alpha|\,||A||.$$

N3. Si $\boldsymbol{v} \neq \boldsymbol{0}$, por definición del supremo

$$\frac{||A\boldsymbol{v}||}{||\boldsymbol{v}||} \leq ||A||,$$

lo cual implica que

$$||A\boldsymbol{v}|| \leq ||A||\,||\boldsymbol{v}||, \tag{13}$$

así, para cualquier vector unitario $\boldsymbol{v}$, usando la desigualdad triangular para la norma vectorial, resulta

$$||(A + B)\boldsymbol{v}|| = ||A\boldsymbol{v} + B\boldsymbol{v}|| \leq ||A\boldsymbol{v}|| + ||B\boldsymbol{v}|| \leq ||A|| + ||B||,$$

de donde, por definición del supremo, se obtiene

$$||A + B|| = \sup_{||\boldsymbol{v}||=1} ||(A + B)\boldsymbol{v}|| \leq ||A|| + ||B||,$$

lo cual da la desigualdad triangular.

□

Observación 1.1. De (13), se tiene que la norma matricial inducida es compatible con la norma vectorial que la induce. Usando nuevamente (13), vemos que

$$||(AB)\boldsymbol{v}|| = ||A(B\boldsymbol{v})|| \leq ||A||\,||B\boldsymbol{v}|| \leq ||A||\,||B||\,||\boldsymbol{v}||.$$

Por lo tanto,

$$||AB|| \leq ||A||\,||B||,$$

es decir, la norma matricial inducida también es sub-multiplicativa.

Note también que si I es la matriz identidad de orden n y $||\cdot||$ es la norma matricial inducida, entonces $||I|| = 1$.

Ejemplos importantes de normas matriciales son las normas inducidas por la normas p, definidas por

$$||A||_p = \sup_{\boldsymbol{v}\neq\boldsymbol{0}} \frac{||A\boldsymbol{v}||_p}{||\boldsymbol{v}||_p}$$

Nótese que si $D = diag(d_1, \ldots, d_n)$, entonces $||D||_p = \text{máx}_{1\leq i\leq n}\,|d_i|$. En efecto, defina $d = \text{máx}_{1\leq i\leq n}\,|d_i|$ y sea $\boldsymbol{v} = (v_1, \ldots, v_n)^T$ un vector no nulo. Entonces,

$$||D\boldsymbol{v}||_p^p = \sum_{i=1}^{n} |d_i v_i|^p \leq d^p \sum_{i=1}^{n} |v_i|^p = d^p ||\boldsymbol{v}||_p^p.$$

Así, $||D\boldsymbol{v}||_p \leq d||\boldsymbol{v}||_p$ y la igualdad se alcanza si $\boldsymbol{v}$ es un vector unitario coordenado. Por lo tanto,

$$||D||_p = \sup_{\boldsymbol{v}\neq\boldsymbol{0}} \frac{||D\boldsymbol{v}||_p}{||\boldsymbol{v}||_p} = d.$$

En general, calcular la norma de una matriz usando (12) puede resultar una tarea no tan sencilla pues implicaría un problema de optimización. Sin embargo, para los casos $p = 1$ y $p = \infty$ el cálculo es sencillo con base en las siguientes fórmulas

$$||A||_\infty = \max_{i=1,\ldots,m} \sum_{j=1}^{n} |a_{ij}| \tag{14}$$

$$||A||_1 = \max_{j=1,\ldots,n} \sum_{i=1}^{m} |a_{ij}| \tag{15}$$

donde $A = (a_{ij})$ es una matriz $m \times n$.

Particular atención merece la norma matricial inducida por la norma Euclidiana. El cálculo de dicha norma se abordará más adelante después de probar algunas factorizaciones matriciales.

2. Métodos directos para resolver sistemas de ecuaciones lineales y algunas factorizaciones matriciales

Los métodos directos permiten obtener, teóricamente, una solución exacta del sistema en un número finito de pasos. Decimos teóricamente porque en la práctica se presentan errores de redondeo en los cálculos. El más conocido de este tipo de métodos es la eliminación Gaussiana. Por otra parte, los métodos iterativos permiten obtener una aproximación de la solución del sistema a partir de una sucesión que converge a dicha solución. La conveniencia de usar uno u otro método depende del tamaño y la estructura de la matriz de coeficientes del sistema.

Método de eliminación de Gauss

Esta técnica para resolver sistemas de ecuaciones lineales fue desarrollada por Carl Friedrich Gauss en su tratado *Theoria motus corporum coelestium in sectionibus conicis solem ambientum* (1809) y la idea básica consiste en transformar el sistema dado en uno equivalente que sea triangular el cual es fácil de resolver.

Comenzamos recordando algunos conceptos familiares de álgebra lineal.

Definición 1.4. Una matriz $A = (a_{ij})$ es *triangular superior* si $a_{ij} = 0$ para $i > j$, es decir, si las entradas que están debajo de la diagonal principal son ceros.

Definición 1.5. Una matriz $A = (a_{ij})$ es *triangular inferior* si $a_{ij} = 0$ para $i < j$, es decir, si las entradas que están encima de la diagonal principal son ceros.

Se puede probar fácilmente (ejercicio) que el producto de dos matrices triangulares superiores de orden n es una también una matriz triangular superior y el producto de dos matrices triangulares inferiores de orden n es una también una matriz triangular inferior.

Usaremos la notación $E_{ij}(k)$ para referirnos a la matriz elemental que se obtiene de la matriz identidad I_n de orden n, al multiplicar la i-ésima fila

por k y sumarle el resultado a la j-ésima fila (esta operación la denotamos $kF_i + F_j$). Recordamos que toda matriz elemental cuadrada es invertible y su inversa es una matriz elemental del mismo tipo. En particular,

$$E_{ij}^{-1}(k) = E_{ij}(-k).$$

También recordamos que el resultado de realizar una operación elemental (con filas) sobre una matriz A de orden n es equivalente a multiplicar A por la izquierda con la matriz elemental se obtiene de I_n realizado la misma operación elemental. Por ejemplo,

$$A = \begin{pmatrix} 2 & -4 \\ -6 & 3 \end{pmatrix} \xrightarrow{3F_1 + F_2} \begin{pmatrix} 2 & -4 \\ 0 & -9 \end{pmatrix}$$

La matriz resultante de la operación elemental $3F_1+F_2$ sobre A es equivalente al producto $E_{12}(3)A$.

Se presenta ahora un ejemplo que será ilustrativo para las ideas generales del método de Gauss y la factorización LU.

Ejemplo 1.1. Considere el sistema lineal

$$\begin{aligned} 2x_1 + x_2 - x_3 &= 3 \\ -4x_1 - 3x_2 + 5x_3 &= 0 \\ 2x_1 + 3x_2 + 2x_3 &= 1 \end{aligned}$$

En forma matricial el sistema puede escribirse como

$$\underbrace{\begin{pmatrix} 2 & 1 & -1 \\ -4 & -3 & 5 \\ 2 & 3 & 2 \end{pmatrix}}_{=A} \underbrace{\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}}_{=\boldsymbol{x}} = \underbrace{\begin{pmatrix} 3 \\ 0 \\ 1 \end{pmatrix}}_{=\boldsymbol{b}}$$

$$\xrightarrow{2F_1 + F_2| - F_1 + F_3} \underbrace{\begin{pmatrix} 2 & 1 & -1 \\ 0 & -1 & 3 \\ 0 & 2 & 3 \end{pmatrix}}_{=E_{13}(-1)E_{12}(2)A} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 3 \\ 6 \\ -2 \end{pmatrix}$$

$$\xrightarrow{2F_2 + F_3} \underbrace{\begin{pmatrix} 2 & 1 & -1 \\ 0 & -1 & 3 \\ 0 & 0 & 9 \end{pmatrix}}_{=E_{23}(2)E_{13}(-1)E_{12}(2)A} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 3 \\ 6 \\ 10 \end{pmatrix}$$

Note que el último sistema tiene matriz de coeficientes triangular superior, la llamremos U, y se puede resolver fácilmente en forma regresiva, esto es, de la última ecuación se obtiene el valor de $x_3 = 10/9$, el cual se usa para hallar el valor de $x_2 = -8/3$ en la segunda ecuación y con estos dos valores se encuentra $x_1 = 61/18$ de la primera ecuación. Por otra parte, como

$$E_{23}(2)E_{13}(-1)E_{12}(2)A = U,$$

entonces

$$A = E_{12}^{-1}(2)E_{13}^{-1}(-1)E_{23}^{-1}(2)U = E_{12}(-2)E_{13}(1)E_{23}(-2)U = LU,$$

donde $L = E_{12}(-2)E_{13}(1)E_{23}(-2)$ es triangular inferior por ser producto de matrices triangulares inferiores. Más adelante veremos condiciones para la existencia de este tipo de factorizaciones LU.

Se procede ahora a la descripción del método de Gauss para resolver el sistema lineal de ecuaciones

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + \cdots + a_{1n}x_n &= b_1 \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 + \cdots + a_{2n}x_n &= b_2 \\ a_{31}x_1 + a_{32}x_2 + a_{33}x_3 + \cdots + a_{3n}x_n &= b_3 \\ &\vdots \\ a_{n1}x_1 + a_{n2}x_2 + a_{n3}x_3 + \cdots + a_{nn}x_n &= b_n, \end{aligned} \tag{16}$$

el cual puede escribirse en forma matricial como

$$A\boldsymbol{x} = \boldsymbol{b},$$

donde $A = (a_{ij})$ es una matriz $n \times n$ con entradas reales (o complejas) llamada matriz de coeficientes, $\boldsymbol{b} = (b_1, b_2, \ldots, b_n)^{\mathrm{T}}$ y $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)^{\mathrm{T}}$ es el vector de incógnitas. Suponemos que la matriz de coeficientes es no singular y por tanto, el sistema (16) tiene solución única. Inicialmente se usa la primera ecuación para eliminar la primera incógnita x_1 en las $n-1$ ecuaciones restantes. Más concretamente, para eliminar x_1 en la i-ésima ecuación con $i = 2, \ldots, n$, se multiplica la primera ecuación por $-a_{i1}/a_{11}$ y se le suma a la i-ésima ecuación. En este punto se requiere que $a_{11} \neq 0$, si este no fuese el caso, dado que la matriz de coeficientes es no singular, al menos uno de los elementos de la primera columna es no nulo, así se pueden reordenar las

ecuaciones de manera que el coeficiente en la fila 1 y columna 1 sea no nulo. Este primer paso transforma el sistema original en el sistema equivalente

$$\begin{aligned} a_{11}^{(1)}x_1 + a_{12}^{(1)}x_2 + a_{13}^{(1)}x_3 + \cdots + a_{1n}^{(1)}x_n &= b_1^{(1)} \\ a_{22}^{(2)}x_2 + a_{23}^{(2)}x_3 + \cdots + a_{2n}^{(2)}x_n &= b_2^{(2)} \\ a_{32}^{(2)}x_2 + a_{33}^{(2)}x_3 + \cdots + a_{3n}^{(2)}x_n &= b_3^{(2)} \\ &\vdots \\ a_{n2}^{(2)}x_2 + a_{n3}^{(2)}x_3 + \cdots + a_{nn}^{(2)}x_n &= b_n^{(2)} \end{aligned} \tag{17}$$

donde los nuevos coeficientes están dados por

$$\begin{aligned} a_{1j}^{(1)} &:= a_{1j}, \qquad j = 1, \ldots, n \\ a_{ij}^{(2)} &:= a_{ij}^{(1)} - \frac{a_{i1}^{(1)} a_{1j}^{(1)}}{a_{11}^{(1)}}, \qquad i, j = 2, \ldots, n, \end{aligned}$$

y los nuevos términos del lado derecho están dados por

$$\begin{aligned} b_1^{(1)} &:= b_1, \\ b_i^{(2)} &:= b_i^{(1)} - \frac{a_{i1}^{(1)} b_1^{(1)}}{a_{11}^{(1)}}, \qquad i = 2, \ldots, n. \end{aligned}$$

Obsérvese que el determinante de la matriz de coeficientes del sistema (17) es diferente de cero pues es igual al determinante de la matriz de coeficientes del sistema (16) (con la excepción de un posible cambio de signo en caso de que haya cambiado el orden de las filas). Luego, al calcular el determinante del sistema (17) por cofactores (usando la primera columna) se concluye que el determinante de la submatriz de coeficientes que resulta del sistema (17) al eliminar la primera fila y la primera columna es también diferente de cero. El paso siguiente es usar segunda ecuación del sistema (17) para eliminar la incógnita x_2 en las $n-2$ ecuaciones que están debajo de dicha ecuación. Esto se logra multiplicando la segunda ecuación por $-a_{i2}^{(2)}/a_{22}^{(2)}$ y se le suma a la i-ésima ecuación para cada $i = 3, \cdots, n$. Aquí se ha asumido que $a_{22}^{(2)} \neq 0$, si este no fuese el caso, se realiza un intercambio de filas teniendo en cuenta que al menos uno de coeficientes $a_{i2}^{(2)}$ con $i \in \{2, \cdots n\}$ es no nulo por la

observación previa. Este segundo paso transforma el sistema (17) en el sistema equivalente

$$\begin{aligned} a_{11}^{(1)}x_1 + a_{12}^{(1)}x_2 + a_{13}^{(1)}x_3 + \cdots + a_{1n}^{(1)}x_n &= b_1^{(1)} \\ a_{22}^{(2)}x_2 + a_{23}^{(2)}x_3 + \cdots + a_{2n}^{(2)}x_n &= b_2^{(2)} \\ a_{33}^{(3)}x_3 + \cdots + a_{3n}^{(3)}x_n &= b_3^{(3)} \\ &\vdots \\ a_{n3}^{(3)}x_3 + \cdots + a_{nn}^{(3)}x_n &= b_n^{(3)} \end{aligned} \tag{18}$$

con

$$\begin{aligned} a_{ij}^{(3)} &:= a_{ij}^{(2)} - \frac{a_{i2}^{(2)}a_{2j}^{(2)}}{a_{22}^{(2)}}, \qquad i,j = 3,\ldots,n, \\ b_{i}^{(3)} &:= b_{i}^{(2)} - \frac{a_{i2}^{(2)}b_{2}^{(2)}}{a_{22}^{(2)}}, \qquad i = 3,\ldots,n. \end{aligned}$$

Repitiendo este proceso, se logra transformar el sistema original en el sistema (triangular) equivalente

$$\begin{aligned} a_{11}^{(1)}x_1 + a_{12}^{(1)}x_2 + a_{13}^{(1)}x_3 + \cdots + a_{1n}^{(1)}x_n &= b_1^{(1)} \\ a_{22}^{(2)}x_2 + a_{23}^{(2)}x_3 + \cdots + a_{2n}^{(2)}x_n &= b_2^{(2)} \\ a_{33}^{(3)}x_3 + \cdots + a_{3n}^{(3)}x_n &= b_3^{(3)} \\ &\vdots \\ a_{nn}^{(n)}x_n &= b_n^{(n)} \end{aligned} \tag{19}$$

donde

$$\begin{aligned} a_{ij}^{(m+1)} &:= a_{ij}^{(m)} - \frac{a_{im}^{(m)}a_{mj}^{(m)}}{a_{mm}^{(m)}}, \qquad i,j = m+1,\ldots,n, \\ b_{i}^{(m+1)} &:= b_{i}^{(m)} - \frac{a_{im}^{(m)}b_{m}^{(m)}}{a_{mm}^{(m)}}, \qquad i = m+1,\ldots,n, \end{aligned}$$

para $m = 1,\ldots,n-1$. Las entradas $a_{mm}^{(m)}$ reciben el nombre de *pivotes* y su elección merece particular atención para evitar pérdida de precisión en los

cálculos computacionales. Más concretamente, para controlar la influencia de errores de redondeo, se debe buscar que el cociente $a_{im}^{(m)}/a_{mm}^{(m)}$ sea lo más pequeño posible, lo cual se logra reordenando las filas y columnas de manera que el elemento pivote $a_{mm}^{(m)}$ sea la entrada de mayor valor absoluto en la $(n-m+1)\times(n-m+1)$ submatriz resultante del m-ésimo paso de eliminación. Este proceso se conoce como *pivotado completo*; cuando la entrada de mayor valor absoluto se busca intercambiando solo filas (ecuaciones) o solo columnas, el procedimiento recibe el nombre de *pivotado parcial*. Veamos un ejemplo que ilustra los problemas de inestabilidad que se pueden presentar en una máquina con el proceso de eliminación Gaussiana cuando no se realiza el proceso de pivotado.

Ejemplo 1.2. Considere el sistema

$$\begin{pmatrix} 10^{-10} & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

Suponga que el problema se va a resolver con una máquina cuyos cálculos se desarrollan con una aritmética de punto flotante de 10^{-8}.

Si se multiplica la primera fila por -10^{10} y se le suma a la segunda, obtenemos

$$\begin{pmatrix} 10^{-10} & 1 \\ 0 & 1-10^{10} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 \\ -10^{10} \end{pmatrix}$$

pero, el número $1-10^{10}$ no será representado en forma exacta; este será redondeado a -10^{10}. En consecuencia, el sistema queda

$$\begin{pmatrix} 10^{-10} & 1 \\ 0 & -10^{10} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 \\ -10^{10} \end{pmatrix}$$

cuya solución es $(0,1)^{\mathrm{T}}$, pero la solución correcta del sistema original es $(-1,1)^{\mathrm{T}}$.

Veamos ahora que ocurre si hacemos el proceso de eliminación de Gauss pero intercambiando filas, resulta el sistema equivalente

$$\begin{pmatrix} 1 & 1 \\ 10^{-10} & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

Multiplicando la primera fila por -10^{-10} y sumándosela a la segunda, obtenemos

$$\begin{pmatrix} 1 & 1 \\ 0 & 1-10^{-10} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

y con el redondeo queda

$$\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

cuya solución es $(-1, 1)^{\mathrm{T}}$, que corresponde a solución correcta del sistema original.

En este ejemplo, la primera forma de resolver el sistema era un algoritmo inestable, en el sentido de que amplificaba los errores conduciendo a resultados incorrectos, sin embargo, esto se pudo resolver con el proceso de pivotado en el segundo algoritmo.

El sistema triangular (19) se resuelve fácilmente despejando x_n de la última ecuación, luego se reemplaza dicho valor en la penúltima ecuación para encontrar x_{n-1} y así sucesivamente procediendo en forma regresiva hasta hallar x_1 de la primera ecuación. El método de eliminación Gaussiana sin pivotar se describe en el siguiente algoritmo [4].

Algoritmo (Eliminacion Gaussiana sin pivotar)**.**

1. Eliminación progresiva

for $m = 1, ..., n-1$
  for $i = m+1, ..., n$
    for $j = m+1, ..., n$
      $a_{ij} := a_{ij} - \dfrac{a_{im}a_{mj}}{a_{mm}}$
    end
    $b_i := b_i - \dfrac{a_{im}b_m}{a_{mm}}$
  end
end

2. Sustitución regresiva

for $m = n, n-1, \ldots, 1$
  $x_m := b_m$
  for $j = m+1, ..., n$
    $x_m := x_m - a_{mj}x_j$
  end
  $x_m := \dfrac{x_m}{a_{mm}}$
end

Factorización *LU*

Se dice que una matriz A tiene una factorización o descomposición LU si existen matrices $L = (l_{ii})$ triangular inferior y $U = (u_{ii})$ matriz triangular superior tales que $A = LU$.

Definición 1.6. Dada una matriz $A \in \mathbb{R}^{n\times n}$, la *submatriz principal* de orden m, para $1 \leq m \leq n$, es la matriz formada por la primeras m filas y m columnas de A.

Definición 1.7. Dada una matriz $A \in \mathbb{R}^{n\times n}$, la *menor principal* de orden m, para $1 \leq m \leq n$, es el determinante de la submatriz principal de order m.

El siguiente resultado da condiciones suficientes y necesarias para la existencia y unicidad de una factorización LU para una matriz cuadrada.

Teorema 1.5. Sea $A \in \mathbb{R}^{n\times n}$. La factorización LU de A con $l_{ii} = 1$ para $i = 1, \ldots, n$ existe y es única si y solo si las submatrices principales A_i de A de orden $i = 1, \ldots, n-1$ son no singulares.

Demostración.

Suponga que las submatrices principales A_i de A de orden $i = 1, \ldots, n-1$ son no singulares y se probará, por inducción sobre n (el orden de A), que la factorización LU de A con $l_{ii} = 1$ para $i = 1, \ldots, n$ existe y es única. Para $n = 1$ el resultado es obvio tomando $L = (1)$ y $U = (a_{11})$. Suponga que el resultado vale para matrices de tamaño $n-1$ y sea $A_n = (a_{ij}) \in \mathbb{R}^{n\times n}$ una matriz cuyas submatrices principales de orden $i = 1, \ldots, n-1$ son no singulares. A_n puede escribirse en bloques como

$$A_n = \begin{pmatrix} A_{n-1} & \boldsymbol{c} \\ \boldsymbol{d}^{\mathrm{T}} & a_{nn} \end{pmatrix}, \tag{20}$$

donde $\boldsymbol{c} = (a_{1n}, a_{2n}, \ldots, a_{(n-1)n})^{\mathrm{T}}$, $\boldsymbol{d}^{\mathrm{T}} = (a_{n1}, a_{n2}, \ldots, a_{n(n-1)})$ y A_{n-1} es la submatriz principal que se obtiene de A_n eliminando la enésima fila y la enésima columna. Puesto que por hipótesis las submatrices principales A_i de A de orden $i = 1, \ldots, n-1$ son no singulares, entonces las submatrices principales de su submatriz A_{n-1} son no singulares. Luego, por hipótesis de inducción sobre A_{n-1}, existe una matriz triangular inferior L_{n-1} con unos en la diagonal principal y una matriz triangular superior U_{n-1} tal que

$$A_{n-1} = L_{n-1}U_{n-1},$$

y esta descomposición es única. Como A_{n-1} es no singular, entonces los factores L_{n-1} y U_{n-1} también los son. Ahora, la expresión de A_n en bloques dada por (20) sugiere que en la descomposición de A_n como un producto L_nU_n, los factores deben elegirse de manera que las primeras $(n-1)$ filas y columnas de L_n y U_n sean las matrices L_{n-1} y U_{n-1} respectivamente. Así, se busca una factorización única de A_n de la forma

$$A_n = L_nU_n = \begin{pmatrix} L_{n-1} & \mathbf{0} \\ \boldsymbol{v}^{\mathrm{T}} & 1 \end{pmatrix} \begin{pmatrix} U_{n-1} & \boldsymbol{w} \\ \mathbf{0}^{\mathrm{T}} & u_{nn} \end{pmatrix} \tag{21}$$

donde el vector columna $\mathbf{0}$, el vector fila $\boldsymbol{v}^{\mathrm{T}}$, vector fila $\mathbf{0}^{\mathrm{T}}$ y el vector columna $\boldsymbol{w}$ tienen todos $(n-1)$ componentes y u_{nn} es el elemento en la última fila y última columna de U_n. El próximo paso es encontrar los vectores $\boldsymbol{v}^{\mathrm{T}}$ y $\boldsymbol{w}$, y el escalar u_{nn} para los cuales la factorización (21) es válida. Con este objetivo en mente, se procede a realizar la multiplicación en bloques (recuerde que este proceso de multiplicación obedece la mismas reglas de la multiplicación matricial ordinaria tomando los bloques como si fueran escalares y que los bloques tengan las dimensiones adecuadas para realizar los productos entre ellos). De esta forma se obtiene

$$A_n = \begin{pmatrix} L_{n-1}U_{n-1} & L_{n-1}\boldsymbol{w} \\ \boldsymbol{v}^{\mathrm{T}}U_{n-1} & \boldsymbol{v}^{\mathrm{T}}\boldsymbol{w} + u_{nn} \end{pmatrix} \tag{22}$$

Igualando los bloques correspondientes en los lados derechos de las ecuaciones (20) y (21), resulta $A_{n-1} = L_{n-1}U_{n-1}$ que era una ecuación ya conocida y también resultan las ecuaciones

$$L_{n-1}\boldsymbol{w} = \boldsymbol{c} \tag{23}$$

$$\boldsymbol{v}^{\mathrm{T}}U_{n-1} = \boldsymbol{d}^{\mathrm{T}} \tag{24}$$

$$\boldsymbol{v}^{\mathrm{T}}\boldsymbol{w} + u_{nn} = a_{nn}. \tag{25}$$

Puesto que L_{n-1} es no singular, se sigue que el sistema lineal (23) determina un valor único para el vector $\boldsymbol{w}$, y como U_{n-1} también es no singular, el vector $\boldsymbol{v}^{\mathrm{T}}$ está univocamente determinado por (24). En consecuencia, de (25), se tiene que u_{nn} está dado en forma única por $u_{nn} = a_{nn} - \boldsymbol{v}^{\mathrm{T}}\boldsymbol{w}$. Esto completa la prueba por inducción.

Resta probar la recíproca, esto es, que si la factorización LU de A con unos en la diagonal principal de L, existe y es única, entonces las submatrices principales A_i de A de orden $i = 1, \ldots, n-1$ son no singulares. Para esto,

se consideran dos casos, cuando A es no singular y cuando A es singular. Veamos el primer caso. Si A es no singular entonces

$$0 \neq \det(A) = \det(L)\det(U) = \det(U) = u_{11}u_{22}\cdots u_{nn}$$

Ahora, para cada $i = 1, \ldots, n$, la submatriz principal A_i de A de orden i puede escribirse como

$$A_i = L_i U_i = \begin{pmatrix} L_{i-1} & \mathbf{0} \\ \boldsymbol{l}^{\mathrm{T}} & 1 \end{pmatrix} \begin{pmatrix} U_{i-1} & \boldsymbol{u} \\ \mathbf{0}^{\mathrm{T}} & u_{ii} \end{pmatrix} \tag{26}$$

con $\boldsymbol{l}^{\mathrm{T}} = (l_{i1}, l_{i2}, \ldots, l_{i(i-1)})$ y $\boldsymbol{u} = (u_{1i}, u_{2i}, \ldots, u_{(i-1)i})^{\mathrm{T}}$. Luego,

$$det(A_i) = \det(L_i)\det(U_i) = \det(U_i) = u_{11}u_{22}\cdots u_{ii}$$

y como $u_{11}u_{22}\cdots u_{ii}\cdots u_{nn} \neq 0$, se sigue que $det(A_i) = u_{11}u_{22}\cdots u_{ii} \neq 0$, es decir, cada A_i es no singular.

Ahora suponga que A es singular. Entonces,

$$0 = \det(A) = u_{11}u_{22}\cdots u_{nn},$$

lo cual implica que al menos una de las entradas diagonales de U es cero. Sea $k \in \{1, \ldots, n\}$ el menor índice para el cual $u_{kk} = 0$. Si $k < n$, de la descomposición (26), se deduce que la factorización se puede desarrollar sin problema hasta el paso $(k+1)$. A partir de este paso y debido a que U_k es singular, no se tiene la existencia y unicidad del vector $\boldsymbol{l}^{\mathrm{T}}$, y por ende, tampoco se tendría la existencia y unicidad de la factorización lo que contradice la hipótesis. Luego, $k = n$ y así $u_{kk} \neq 0$ para cada $k \in \{1, \ldots, n-1\}$. Por lo tanto, para cada $k \in \{1, \ldots, n-1\}$ $det(A_k) = \det(L_k)\det(U_k) = \det(U_k) = u_{11}u_{22}\cdots u_{kk} \neq 0$, es decir, todas las submatrices principales A_k son no singulares para $k = 1, \ldots, n-1$.

□

Del anterior teorema se concluye que, si alguna submatriz principal A_i de A con $i \in \{1, \ldots, n-1\}$ es singular, entonces o la factorización LU de A no existe o no es única. Considere, por ejemplo, la matriz (no singular)

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

Claramente no es posible encontrar un valor no nulo de u_{11} tal que

$$\begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ l_{21} & 1 \end{bmatrix} \begin{bmatrix} u_{11} & u_{12} \\ 0 & u_{22} \end{bmatrix}.$$

Si se busca resolver el sistema $A\boldsymbol{x} = \boldsymbol{b}$, donde A es una matriz no singular cuya factorización LU existe, el sistema se puede escribir como

$$LU\boldsymbol{x} = \boldsymbol{b}.$$

Tomando $\boldsymbol{y} = U\boldsymbol{x}$, el sistema original se ha reducido a resolver dos sistemas lineales cuyas matrices de coeficientes son triangulares y este tipo de sistemas son más fáciles de resolver. Primero se resuelve para $\boldsymbol{y}$ el sistema triangular inferior

$$L\boldsymbol{y} = \boldsymbol{b},$$

usando sustitución progresiva, es decir, si

$$\begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ l_{21} & 1 & 0 & \cdots & 0 \\ l_{31} & l_{32} & 1 & \cdots & \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ l_{n1} & l_{n2} & l_{n3} & & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ \vdots \\ b_n \end{bmatrix},$$

entonces

$$y_1 = b_1 \qquad \text{y} \qquad y_i = b_i - \sum_{k=1}^{i-1} l_{ik} y_k \qquad \text{para} \qquad i = 1, \ldots, n. \tag{27}$$

Como $\boldsymbol{y}$ ya es conocido, el sistema triangular superior $U\boldsymbol{x} = \boldsymbol{y}$, o bien

$$\begin{bmatrix} u_{11} & u_{12} & u_{13} & \cdots & u_{1n} \\ 0 & u_{22} & u_{23} & \cdots & u_{2n} \\ 0 & 0 & u_{33} & \cdots & u_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & u_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix},$$

se resuelve por sustitución regresiva

$$x_n = y_n/u_{nn} \qquad \text{y} \qquad x_i = \frac{1}{u_{ii}} \left(y_i - \sum_{k=i+1}^{n} u_{ik} x_k \right), \tag{28}$$

para $i = n-1, n-2, \ldots, 1$.

Cuando se dispone de la factorización LU de la matriz de coeficientes del sistema $A\boldsymbol{x} = \boldsymbol{b}$ y se usan las fórmulas (27)-(28) para resolver los dos sistemas triangulares $L\boldsymbol{y} = \boldsymbol{b}$ y $U\boldsymbol{x} = \boldsymbol{y}$ se requieren n^2 multiplicaciones/divisiones y $n^2 - n$ sumas/sustracciones, lo cual resulta relativamente más económico que la eliminanación Gaussiana en la cual se realizan aproximadamente $n^3/3$ multiplicaciones/divisiones y $n^3/3$ sumas/sustracciones.

Factorización LDU

En la descomposión LU de la matriz A descrita en el Teorema (1.5) la matriz L tiene unos en la diagonal principal mientras que para la matriz U eso no se tiene. Esto puede remediarse reescribiendo U de la siguiente manera

$$\begin{bmatrix} u_{11} & u_{12} & \cdots & u_{1n} \\ 0 & u_{22} & \cdots & u_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \vdots & u_{nn} \end{bmatrix} = \begin{bmatrix} u_{11} & 0 & \cdots & 0 \\ 0 & u_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \vdots & u_{nn} \end{bmatrix} \begin{bmatrix} 1 & u_{12}/u_{11} & \cdots & u_{1n}/u_{11} \\ 0 & 1 & \cdots & u_{2n}/u_{22} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \vdots & 1 \end{bmatrix}$$

Así, se obtiene la factorización $A = LDU$, donde L y U son matrices triangulares inferior y superior respectivamente, con unos en la diagonal principal y $D = diag(u_{11}, u_{22}, \ldots, u_{nn})$ es la matriz diagonal de pivotes.

Factorización de Cholesky

Para el caso A simétrica y definida positiva, se puede obtener una factorización LU con $U = L^{\mathrm{T}}$ teniendo L entradas diagonales positivas como lo indica el siguiente resultado debido a André-Louis Cholesky (1875-1918). La prueba se puede obtener directamente del Teorema de factorización LU teniendo en cuenta que si una matriz A es simétrica y definida positiva, entonces las submatrices principales de A son definidas positivas y por tanto, no singulares. Sin embargo, optaremos por imitar la prueba del teorema anterior para obtener también por inducción las entradas de L.

Teorema 1.6 (Factorización de Cholesky)**.** Si $A \in \mathbb{R}^{n\times n}$ es simétrica y definida positiva, entonces existe una única matriz triangular inferior L con entradas diagonales positivas tal que

$$A = LL^{\mathrm{T}} \tag{29}$$

Las entradas l_{ij} de L vienen dadas por

$$l_{11} = \sqrt{a_{11}},$$

y para $i = 2, \ldots, n$

$$l_{ij} = \frac{a_{ij} - \sum_{m=1}^{j-1} l_{im} l_{jm}}{l_{jj}}, \qquad j = 1, \ldots, i-1 \tag{30}$$

$$l_{ii} = \left(a_{ii} - \sum_{m=1}^{i-1} l_{im}^2 \right)^{1/2} \tag{31}$$

Demostración.
Nótese inicialmente que el hecho de que A sea definida positiva implica que sus entradas diagonales son positivas. En efecto, para cada $i = 1, \ldots, n$, $a_{ii} = \boldsymbol{e}_i^{\mathrm{T}} A \boldsymbol{e}_i > 0$, donde $\boldsymbol{e}_i$ es el vector cuya $i-$ésima entrada es uno y el resto de las entradas son nulas. Con esta observación se procede a la prueba del teorema por inducción sobre n.

Para $n = 1$ el resultado es obvio tomando $L = (\sqrt{a_{11}})$. Suponga que el resultado vale para matrices de tamaño $n-1$ y sea $A = A_n \in \mathbb{R}^{n\times n}$ simétrica y definida positiva. A_n puede escribirse como

$$A_n = \begin{pmatrix} A_{n-1} & \boldsymbol{v} \\ \boldsymbol{v}^{\mathrm{T}} & a_{nn} \end{pmatrix}, \tag{32}$$

donde A_{n-1} es la submatriz principal que se obtiene de A_n eliminando la enésima fila y la enésima columna, $\boldsymbol{v} = (a_{1n}, a_{2n}, \ldots, a_{(n-1)n})^{\mathrm{T}}$. Puesto que A_n es simétrica y definida positiva, la submatriz principal A_{n-1} también lo es. En efecto, la simetría es clara. Sea $\boldsymbol{w} = (w_1, w_2, \ldots, w_{n-1})^{\mathrm{T}}$ un elemento cualquiera no nulo de $\mathbb{R}^{n-1}$ y tome $\boldsymbol{u} = (w_1, w_2, \ldots, w_{n-1}, 0)^{\mathrm{T}}$. Luego, $\boldsymbol{w}^{\mathrm{T}} A_{n-1} \boldsymbol{w} = \boldsymbol{u}^{\mathrm{T}} A_n \boldsymbol{u} > 0$, lo cual implica que A_{n-1} es definida positiva. Aplicando la hipótesis de inducción a A_{n-1}, existe una matriz triangular inferior L_{n-1} con entradas diagonales positivas tal que

$$A_{n-1} = L_{n-1} L_{n-1}^{\mathrm{T}},$$

y las entradas l_{ij} de L_{n-1}, $i, j = 1, 2 \ldots, n-1$ vienen dadas por

$$l_{11} = \sqrt{a_{11}}.$$

Para $i = 2, \ldots, n-1$

$$l_{ij} = \frac{a_{ij} - \sum_{m=1}^{j-1} l_{im} l_{jm}}{l_{jj}}, \qquad j = 1, \ldots, i-1 \tag{33}$$

$$l_{ii} = \left(a_{ii} - \sum_{m=1}^{i-1} l_{im}^2 \right)^{1/2}. \tag{34}$$

Se busca una factorización única de A_n de la forma

$$A_n = \begin{pmatrix} L_{n-1} & \boldsymbol{O}_{(n-1)\times 1} \\ \boldsymbol{y}^{\mathrm{T}} & \beta \end{pmatrix} \begin{pmatrix} L_{n-1}^{\mathrm{T}} & \boldsymbol{y} \\ \boldsymbol{O}_{1\times(n-1)} & \beta \end{pmatrix} \tag{35}$$

Haciendo el producto de las matrices por bloques e igualando con A_n (ver (32)), resulta

$$L_{n-1}\boldsymbol{y} = \boldsymbol{v} \qquad \text{y} \qquad \boldsymbol{y}^{\mathrm{T}}\boldsymbol{y} + \beta^2 = a_{nn}. \tag{36}$$

Puesto que L_{n-1} es no singular, se sigue que $\boldsymbol{y}$ está unívocamente determinada. Además,

$$0 < \det(A_n) = \beta \det(L_{n-1})\beta \det(L_{n-1}^{\mathrm{T}}) = \beta^2(\det(L_{n-1}))^2,$$

y como $det(L_{n-1}) > 0$, se sigue que β es un real no nulo, de hecho, $\beta = \sqrt{a_{nn} - \boldsymbol{y}^{\mathrm{T}}\boldsymbol{y}}$. Tomando

$$L_n := \begin{pmatrix} L_{n-1} & \boldsymbol{O}_{(n-1)\times 1} \\ \boldsymbol{y}^{\mathrm{T}} & \beta \end{pmatrix}$$

se tiene la primera parte del teorema y las fórmulas (30)-(31) para $i = 1, \ldots, n-1$. Las fórmulas completas hasta $i = n$ resultan de (36):

$$\begin{pmatrix} l_{11} & 0 & 0 & \cdots & 0 \\ l_{21} & l_{22} & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ l_{(n-1)1} & l_{(n-1)2} & l_{(n-1)3} & \cdots & l_{(n-1)(n-1)} \end{pmatrix} \begin{pmatrix} l_{n1} \\ l_{n2} \\ \vdots \\ l_{n(n-1)} \end{pmatrix} = \begin{pmatrix} a_{1n} \\ a_{2n} \\ \vdots \\ a_{(n-1)n} \end{pmatrix},$$

y

$$\beta = \sqrt{a_{nn} - \boldsymbol{y}^{\mathrm{T}}\boldsymbol{y}} = \sqrt{a_{nn} - \sum_{m=1}^{n-1} l_{nm}^2}.$$

□

La factorización de Cholesky es muy conveniente para matrices simétricas y definidas positivas, entre otras, por las siguientes razones:

- Para n grande, el número de operaciones requeridas es aproximadamente $n^3/6$, esto es la mitad del número de operaciones requeridas en la factorización LU de una matriz no simétrica.
- No se requiere pivoteo para la estabilidad numérica.

Ejemplo 1.3. Use el método de factorización de Cholesky para resolver el sistema

$$\begin{aligned} x_1 - 2x_2 + 2x_3 &= 4 \\ -2x_1 + 5x_2 - 3x_3 &= -7 \\ 2x_1 - 3x_2 + 6x_3 &= 10 \end{aligned}$$

Solución.

En forma matricial el sistema puede escribirse como

$$\underbrace{\begin{pmatrix} 1 & -2 & 2 \\ -2 & 5 & -3 \\ 2 & -3 & 6 \end{pmatrix}}_{=A} \underbrace{\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}}_{=\boldsymbol{x}} = \underbrace{\begin{pmatrix} 4 \\ -7 \\ 10 \end{pmatrix}}_{=\boldsymbol{b}}$$

La matriz de coeficientes A es simétrica y definida positiva (verifíquelo). Luego, por el Teorema de Descomposición de Cholesky, existe una única matriz triangular inferior L con entradas diagonales positivas tal que

$$A = LL^{\mathrm{T}} \tag{37}$$

Las entradas l_{ij} de L vienen dadas por (ver (34))

$$l_{11} = \sqrt{a_{11}} = 1.$$

Para $i = 2$:

$$l_{21} = \frac{a_{21} - \sum_{m=1}^{0} l_{2m} l_{1m}}{l_{11}} = \frac{-2 - 0}{1} = -2$$

$$l_{22} = \left(a_{22} - \sum_{m=1}^{1} l_{2m}^2 \right)^{1/2} = \left(5 - (-2)^2\right)^{1/2} = 1.$$

Para $i = 3$:

$$l_{31} = \frac{a_{31} - \sum_{m=1}^{0} l_{3m} l_{1m}}{l_{11}} = \frac{2 - 0}{1} = 2$$

$$l_{32} = \frac{a_{32} - \sum_{m=1}^{1} l_{3m} l_{2m}}{l_{22}} = \frac{-3 - (2)(-2)}{1} = 1$$

$$l_{33} = \left(a_{33} - \sum_{m=1}^{2} l_{3m}^2 \right)^{1/2} = \left(6 - (2)^2 - (1)^2\right)^{1/2} = 1.$$

Así,

$$L = \begin{pmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ 2 & 1 & 1 \end{pmatrix} \qquad \text{y} \qquad L^T = \begin{pmatrix} 1 & -2 & 2 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix}$$

Luego, el sistema puede escribirse como

$$\begin{pmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ 2 & 1 & 1 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} 4 \\ -7 \\ 10 \end{pmatrix} \tag{38}$$

con

$$\boldsymbol{y} = L^T \boldsymbol{x}. \tag{39}$$

El sistema triangular (38) se puede resolver muy fácilmente en forma progresiva, obteniéndose $\boldsymbol{y} = (4, 1, 1)^{\mathrm{T}}$. Finalmente, con el vector $\boldsymbol{y}$ encontrado, se resuelve el sistema (39):

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

cuya solución es $\boldsymbol{x} = (2, 0, 1)^{\mathrm{T}}$.

Ejercicio 1.1.

1. Considere la matriz tridiagonal de tamaño

$$T = \begin{pmatrix} \beta_1 & \gamma_1 & 0 & 0 & \cdots & 0 \\ \alpha_1 & \beta_2 & \gamma_2 & 0 & \cdots & 0 \\ 0 & \alpha_2 & \beta_3 & \gamma_3 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \alpha_{n-2} & \beta_{n-1} & \gamma_{n-1} \\ 0 & 0 & 0 & 0 & \alpha_{n-1} & \beta_n \end{pmatrix}$$

a) Si T tiene una factorización LU, verifique que estos factores están dados por

$$L = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ \alpha_1/\pi_1 & 1 & 0 & \cdots & 0 \\ 0 & \alpha_2/\pi_2 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \alpha_{n-1}/\pi_{n-1} & 1 \end{pmatrix}, \quad U = \begin{pmatrix} \pi_1 & \gamma_1 & 0 & \cdots & 0 \\ 0 & \pi_2 & \gamma_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \pi_{n-1} & \gamma_{n-1} \\ 0 & 0 & 0 & 0 & \pi_n \end{pmatrix}$$

donde los valores π's están dados por la fórmula de recursión

$$\pi_1 = \beta_1 \qquad \text{and} \qquad \pi_{i+1} = \beta_{i+1} - \frac{\alpha_i \gamma_i}{\pi_i}.$$

b) Escriba un código para la descomposición descrita en el punto anterior.

2. Sea $A \in \mathbb{R}^{n\times n}$. Pruebe que si $A = LL^T$, donde L es una matriz triangular inferior con entradas diagonales positivas (*descomposición de Cholesky*), entonces A es simétrica y definida positiva.

3. Considere la matriz

$$A = \begin{pmatrix} 4 & -1 & 0 \\ -1 & 4 & -1 \\ 0 & -1 & 4 \end{pmatrix}.$$

Verifique que A es simétrica y definida positiva. Luego, halle la descomposición de Cholesky de A.

4. Considere el sistema $A\boldsymbol{u} = \boldsymbol{b}$, dado en (6) tomando una distribución uniforme de cargas, esto es, $\boldsymbol{b} = (1, 1, \ldots, 1)^\top \in \mathbb{R}^n$.

a) Use la descomposición (7) para probar que la matriz A es simétrica y definida positiva.

b) Tome $n = 1000$ y resuelva el sistema usando eliminación Gaussiana y el método de Cholesky. Compare los tiempo de cómputo y las soluciones obtenidas.

3. Factorización QR

Para ciertas factorizaciones matriciales es útil construir, a partir de una base cualquiera, una base ortonormal de un subespacio vectorial dotado de un producto interior. Una manera de obtener bases ortonormales es el proceso de *Gram-Schmidt*, el cual se describe en el siguiente teorema.

Teorema 1.7. Sea $\{\boldsymbol{a}_1, \boldsymbol{a}_2, \dots\}$ un conjunto finito o contable de vectores linealmente independientes en un espacio con producto interior, entonces existe un sistema ortonormal $\{\boldsymbol{q}_1, \boldsymbol{q}_2, \dots\}$ con la propiedad

$$span\{\boldsymbol{a}_1, \dots, \boldsymbol{a}_j\} = span\{\boldsymbol{q}_1, \dots, \boldsymbol{q}_j\}, \qquad \forall j \geq 1. \tag{40}$$

Demostración.

La prueba se realiza inductivamente con el proceso de *Gram-Schmidt*. Para $j = 1$, defina

$$\boldsymbol{q}_1 = \frac{\boldsymbol{a}_1}{||\boldsymbol{a}_1||}.$$

Así, $||\boldsymbol{q}_1|| = 1$ y $span\{\boldsymbol{a}_1\} = span\{\boldsymbol{q}_1\}$. Para $j \geq 2$, suponga que se ha construido $\{\boldsymbol{q}_1, \dots, \boldsymbol{q}_{j-1}\}$ tal que

$$\langle \boldsymbol{q}_i, \boldsymbol{q}_l \rangle = \delta_{il}, \qquad 1 \leq i, l \leq j-1, \quad \text{y}$$

$$span\{\boldsymbol{a}_1, \dots, \boldsymbol{a}_{j-1}\} = span\{\boldsymbol{q}_1, \dots, \boldsymbol{q}_{j-1}\}. \tag{41}$$

Ahora se construye $\boldsymbol{q}_j$ de manera que sea ortonormal a $\{\boldsymbol{q}_1, \dots, \boldsymbol{q}_{j-1}\}$. Inicialmente se busca un vector $\tilde{\boldsymbol{q}}_j$ ortogonal a $\{\boldsymbol{q}_1, \dots, \boldsymbol{q}_{j-1}\}$ tal que

$$span\{\boldsymbol{a}_1, \dots, \boldsymbol{a}_{j-1}, \boldsymbol{a}_j\} = span\{\boldsymbol{q}_1, \dots, \boldsymbol{q}_{j-1}, \tilde{\boldsymbol{q}}_j\}, \qquad \forall j \geq 1.$$

Esto último junto con (41) sugiere la elección

$$\tilde{\boldsymbol{q}}_j := \boldsymbol{a}_j - \sum_{i=1}^{j-1} r_{ij} \boldsymbol{q}_i.$$

Para obtener la ortogonalidad, los coeficientes $r_{1j}, r_{2j}, \dots, r_{(j-1)j}$ se elijen de manera que

$$\langle \tilde{\boldsymbol{q}}_j, \boldsymbol{q}_i \rangle = 0 \qquad 1 \leq i \leq j-1.$$

Esto implica que

$$r_{ij} = \langle \boldsymbol{a}_j, \boldsymbol{q}_i \rangle, \qquad 1 \leq i \leq j-1.$$

Note que $\tilde{\boldsymbol{q}}_j$ no es el vector nulo, pues en caso contrario, el conjunto $\{\boldsymbol{a}_1, \dots, \boldsymbol{a}_{j-1}, \boldsymbol{a}_j\}$ no sería linealmente independiente. Finalmente, se normaliza el vector obtenido tomando

$$\boldsymbol{q}_j = \frac{\tilde{\boldsymbol{q}}_j}{||\tilde{\boldsymbol{q}}_j||}.$$

Por lo tanto, el conjunto $\{\boldsymbol{q}_1, \dots, \boldsymbol{q}_{j-1}, \boldsymbol{q}_j\}$ satisface

$$\langle \boldsymbol{q}_i, \boldsymbol{q}_l \rangle = \delta_{il}, \qquad 1 \leq i, l \leq j,$$

y se verifica (40). □

Del teorema anterior, se obtiene el siguiente algoritmo para obtener una base ortonormal $\{\boldsymbol{q}_1, \boldsymbol{q}_2, \ldots, \boldsymbol{q}_n\}$ de $span\{\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_n\}$.

Algoritmo (Gram-Schmidt)**.**

$r_{11} := ||\boldsymbol{a}_1||$, $\boldsymbol{q}_1 := \boldsymbol{a}_1/r_{11}$
for $j = 2, ..., n$
  for $i = 1, ..., j-1$
    $r_{ij} := \langle \boldsymbol{a}_j, \boldsymbol{q}_i \rangle$
  end
  $\tilde{\boldsymbol{q}} := \boldsymbol{a}_j - \sum_{i=1}^{j-1} r_{ij}\boldsymbol{q}_i$
  $r_{jj} := ||\tilde{\boldsymbol{q}}||$,
  $\boldsymbol{q}_j := \tilde{\boldsymbol{q}}/r_{jj}$
end

En cada paso del algoritmo anterior, se obtiene la relación

$$\boldsymbol{a}_j = \sum_{i=1}^{j} r_{ij}\boldsymbol{q}_i. \tag{42}$$

Si definimos $A := [\boldsymbol{a}_1|\boldsymbol{a}_2|\cdots|\boldsymbol{a}_n]$, $\widetilde{Q} := [\boldsymbol{q}_1|\boldsymbol{q}_2|\cdots|\boldsymbol{q}_n]$, y denotamos por $\widetilde{R}$ la matriz triangular superior cuyas entradas no nulas son los valores r_{ij} definidos en el algoritmo de Gram-Schmidt, entonces de la relación (42) se obtiene

$$A = \widetilde{Q}\widetilde{R}.$$

Esta descomposición recibe el nombre de *factorización QR reducida* de la $m \times n$ matriz A, donde $\widetilde{Q}$ es una matriz $m \times n$ con columnas ortonormales y $\widetilde{R}$ es una matriz $n \times n$ triangular superior.

Una *factorización QR completa* de $A \in \mathbb{R}^{m\times n}$ con $m \geq n$ se obtiene añadiendo $m - n$ columnas ortonormales a $\widetilde{Q}$ obteniéndose de esta manera una matriz ortogonal Q. En el proceso, filas de ceros se añaden a la matriz $\widetilde{R}$ de manera que resulta una matriz R de tamaño $m \times n$, la cual sigue siendo triangular superior (aunque ya no es cuadrada, en la literatura a veces la llama matriz *trapezoidal superior*). La relación entre las factorizaciones reducida y completa se ilustra en la Figura 1

El resultado siguiente está relacionado con la existencia de la factorización QR y, bajo restricciones adecuadas, con la unicidad en el caso reducido . El teorema vale también para matrices con entradas complejas.

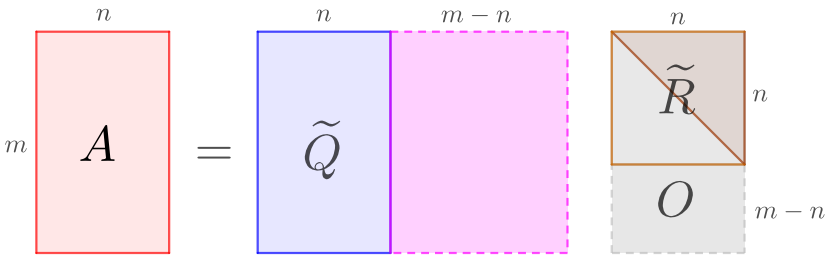

Figura 1: Factorización QR reducida (lineas continuas) y completa (líneas discontinuas).

Teorema 1.8. Si $A \in \mathbb{R}^{m\times n}$ con $m \geq n$, entonces existe una matriz $Q \in \mathbb{R}^{m\times n}$ con columnas ortonormales y una matriz triangular superior $R \in \mathbb{R}^{n\times n}$ tal que $A = QR$. Si $m = n$, Q es una matriz ortogonal; si además A es no singular, R se puede elegir de manera que todas sus entradas diagonales sean positivas, y en este caso la factorización QR es única.

El ejemplo que se presenta a continuación ilustra el proceso de Gram-Schmidt para obtener una factorización QR reducida

Ejemplo 1.4. Considere la matriz

$$A = \begin{pmatrix} 1 & 0 & -1 \\ 1 & 2 & 1 \\ 1 & 1 & -3 \\ 0 & 1 & 1 \end{pmatrix}$$

Denotemos por $\boldsymbol{a}_1$, $\boldsymbol{a}_2$ y $\boldsymbol{a}_3$ las columnas de A, las cuales son linealmente independientes (por qué?).

$$j = 1 : r_{11} = ||\boldsymbol{a}_1||_2 = \sqrt{3} \Rightarrow \boldsymbol{q}_1 = \frac{\boldsymbol{a}_1}{r_{11}} = \frac{1}{\sqrt{3}} \begin{pmatrix} 1 \\ 1 \\ 1 \\ 0 \end{pmatrix}$$

$$j=2: r_{12}=\langle \boldsymbol{a}_2, \boldsymbol{q}_1\rangle = \sqrt{3} \Rightarrow \tilde{\boldsymbol{q}}_2 = \boldsymbol{a}_2 - r_{12}\boldsymbol{q}_1 = \begin{pmatrix} -1 \\ 1 \\ 0 \\ 1 \end{pmatrix}$$

$$\Rightarrow r_{22} = ||\tilde{\boldsymbol{q}}_2||_2 = \sqrt{3} \quad \text{y} \quad \boldsymbol{q}_2 = \frac{\tilde{\boldsymbol{q}}_2}{r_{22}} = \frac{1}{\sqrt{3}} \begin{pmatrix} -1 \\ 1 \\ 0 \\ 1 \end{pmatrix}$$

$$j=3: r_{13}=\langle \boldsymbol{a}_3, \boldsymbol{q}_1\rangle = -\sqrt{3} \quad \text{y} \quad r_{23}=\langle \boldsymbol{a}_3, \boldsymbol{q}_2\rangle = \sqrt{3}$$

$$\Rightarrow \tilde{\boldsymbol{q}}_3 = \boldsymbol{a}_3 - r_{13}\boldsymbol{q}_1 - r_{23}\boldsymbol{q}_2 = \begin{pmatrix} 1 \\ 1 \\ -2 \\ 0 \end{pmatrix}$$

$$\Rightarrow r_{33} = ||\tilde{\boldsymbol{q}}_3||_2 = \sqrt{6} \quad \text{y} \quad \boldsymbol{q}_3 = \frac{\tilde{\boldsymbol{q}}_3}{r_{33}} = \frac{1}{\sqrt{6}} \begin{pmatrix} 1 \\ 1 \\ -2 \\ 0 \end{pmatrix}$$

Por lo tanto,

$$A = \begin{pmatrix} 1 & 0 & -1 \\ 1 & 2 & 1 \\ 1 & 1 & -3 \\ 0 & 1 & 1 \end{pmatrix} = \underbrace{\begin{pmatrix} \frac{1}{\sqrt{3}} & \frac{-1}{\sqrt{3}} & \frac{1}{\sqrt{6}} \\ \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{6}} \\ \frac{1}{\sqrt{3}} & 0 & \frac{-2}{\sqrt{6}} \\ 0 & \frac{1}{\sqrt{3}} & 0 \end{pmatrix}}_{\tilde{Q}} \underbrace{\begin{pmatrix} \sqrt{3} & \sqrt{3} & -\sqrt{3} \\ 0 & \sqrt{3} & \sqrt{3} \\ 0 & 0 & \sqrt{6} \end{pmatrix}}_{\tilde{R}}.$$

Observación 1.2. La versión clásica del algoritmo de Gram-Schmidt presentada arriba no es usada en la práctica debido a que los errores de redondeo pueden causar la pérdida de independencia lineal en los vectores generados. De hecho, en aritmética de punto flotante, el algoritmo puede generar valores muy pequeños de $||\tilde{\boldsymbol{q}}||_2$ y r_{jj}, lo cual genera inestabilidad numérica y pérdida de ortogonalidad. Para ilustrar esto, considere el siguiente ejemplo tomado de [7].

Ejercicio 1.2. La *matriz de Hilbert* $\mathcal{H}_n = (h_{ij})_{n\times n}$ tiene entradas definidas por

$$h_{ij} = \frac{1}{i+j-1}, \qquad i,j = 1,\ldots,n \tag{43}$$

Considere la matriz de Hilbert de orden cuatro $\mathcal{H}_4$ y aplique el algoritmo de Gram-Schmitd a las columnas de dicha matriz. Forme la matriz $\widetilde{Q}$ y calcule $I - \widetilde{Q}^T\widetilde{Q}$. Qué ocurre si se trabaja con aritmética de diez digitos, es decir, hasta el orden de 10^{-10}? Haga el mismo procedimiento con la siguiente versión del algoritmo de Gram-Schmidt y compare los resultados.

La siguiente es una versión más estable desde el punto de vista numérico del algoritmo de Gram-Schimdt

Algoritmo (Gram-Schmidt modificado)**.**

$r_{11} := ||\boldsymbol{a}_1||$, $\boldsymbol{q}_1 := \boldsymbol{a}_1/r_{11}$

for $j = 2,...,n$
  $\tilde{\boldsymbol{q}} := \boldsymbol{a}_j$
  for $i = 1,...,j-1$
    $r_{ij} := \langle \tilde{\boldsymbol{q}}, \boldsymbol{q}_i \rangle$
    $\tilde{\boldsymbol{q}} := \tilde{\boldsymbol{q}} - r_{ij}\boldsymbol{q}_i$
  end
  $r_{jj} := ||\tilde{\boldsymbol{q}}||$,
  $\boldsymbol{q}_j := \tilde{\boldsymbol{q}}/r_{jj}$
end

Ejercicios 1.1. 1. Sea $\mathcal{C}([-1,1])$ el espacio de todas las funciones de valor real que son continuas en el intervalo $[-1,1]$. En $\mathcal{C}([-1,1])$ considere el producto interior definido por

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

Use el proceso de Gram-Schmidt para hallar un conjunto ortonormal de polinomios $\{q_1, q_2, q_3\}$ tal que

$$span\{1, x, x^2\} = span\{q_1, q_2, q_3\}.$$

Los polinomios q_1, q_2, q_3 son los llamados tres primeros *polinomios de Legendre* . Verifique que q_n satisface la ecuación diferencial de Legendre

$$(1-x^2)y'' - 2xy' + n(n-1)y = 0 \tag{44}$$

para $n = 1, 2, 3$.

Una forma alternativa al método de Gram-Schmidt para ortogonalizar una sucesión de vectores es utilizando las matrices o transformaciones de Householder.

Definición 1.8. Sea $\boldsymbol{v} \in \mathbb{C}^n$ un vector normalizado por $\boldsymbol{v}^*\boldsymbol{v} = 1$. Una matriz de la forma

$$P_{\boldsymbol{v}} = I_n - 2\boldsymbol{v}\boldsymbol{v}^*$$

recibe el mombre de *matriz de Householder*. Cuando del contexto sea claro cual es el vector $\boldsymbol{v}$, se escribe simplemente P en vez de $P_{\boldsymbol{v}}$.

Las matrices de Householder son unitarias y Hermitianas. En efecto,

$$P^* = (I_n - 2\boldsymbol{v}\boldsymbol{v}^*)^* = I_n^* - 2(\boldsymbol{v}\boldsymbol{v}^*)^* = I_n - 2\boldsymbol{v}\boldsymbol{v}^* = P.$$

Teniendo en cuenta que P es Hermitiana y $\boldsymbol{v}$ está normalizado, resulta

$$PP^* = P^*P = (I_n - 2\boldsymbol{v}\boldsymbol{v}^*)(I_n - 2\boldsymbol{v}\boldsymbol{v}^*) = I_n - 4\boldsymbol{v}\boldsymbol{v}^* + 4\boldsymbol{v}\boldsymbol{v}^*\boldsymbol{v}\boldsymbol{v}^* = I_n.$$

Si $\boldsymbol{w} \in span\{\boldsymbol{v}\}^\perp$, entonces

$$P\boldsymbol{w} = I_n\boldsymbol{w} - 2\boldsymbol{v}\boldsymbol{v}^*\boldsymbol{w} = \boldsymbol{w},$$

es decir, P actúa como la identidad en el hiperplano $span\{\boldsymbol{v}\}^\perp$. Además, $P\boldsymbol{v} = -\boldsymbol{v}$. Más aun, dado un vector $\boldsymbol{x}$ cualquiera, $P\boldsymbol{x}$ es una reflexión de $\boldsymbol{x}$ con respecto al hiperplano $span\{\boldsymbol{v}\}^\perp$. Para ver esto, exprese $\boldsymbol{x}$ como la suma de dos componentes, una en la dirección de $\boldsymbol{v}$ y la otra (denotada por $\boldsymbol{y}$) ortogonal a $\boldsymbol{v}$:

$$\boldsymbol{x} = \boldsymbol{v}\boldsymbol{v}^*\boldsymbol{x} + \boldsymbol{y}.$$

Así (ver Figura 2),

$$P\boldsymbol{x} = -\boldsymbol{v}\boldsymbol{v}^*\boldsymbol{x} + \boldsymbol{y}.$$

Debido a esta última propiedad, las matrices de Householder también reciben el nombre de *reflectores de Householder*.

Siguiendo la referencia [8], se describirá el proceso de ortonormalización de Householder para matrices con entradas reales, formulando el problema

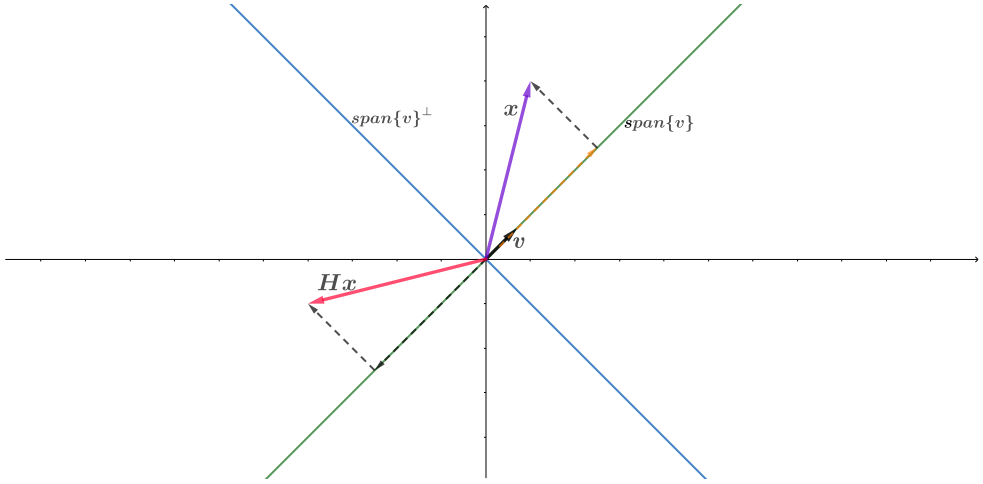

Figura 2: Ilustración del reflector de Householder en el plano.

como una factorización QR de una matriz $A \in \mathbb{R}^{n\times m}$, es decir, descomponer A como

$$A = QR,$$

donde Q es una matriz con columnas ortonormales y R es una matriz triangular superior. Denotemos por $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_m$ las columnas de A y por $\boldsymbol{e}_j$ el j-ésimo vector unitario coordenado. Se inicia seleccionando el vector $\boldsymbol{v}_1$ de manera que

$$P_1\boldsymbol{a}_1 = \boldsymbol{a}_1 - 2\boldsymbol{v}_1\boldsymbol{v}_1^T\boldsymbol{a}_1 = \alpha\boldsymbol{e}_1, \tag{45}$$

donde $P_1 = P_{\boldsymbol{v}_1}$ y α es un escalar por determinar. Note que en este primer paso, se busca que, excepto por la primera entrada, las componentes de $P_1\boldsymbol{a}_1$ sean ceros. De (45),

$$2\boldsymbol{v}_1\boldsymbol{v}_1^T\boldsymbol{a}_1 = \boldsymbol{a}_1 - \alpha\boldsymbol{e}_1, \tag{46}$$

lo cual indica que el vector buscado $\boldsymbol{v}_1$ es un múltiplo del vector $\boldsymbol{a}_1 - \alpha\boldsymbol{e}_1$, lo cual justifica la elección

$$\boldsymbol{v}_1 = \pm\frac{\boldsymbol{a}_1 - \alpha\boldsymbol{e}_1}{||\boldsymbol{a}_1 - \alpha\boldsymbol{e}_1||_2}.$$

Reemplazando este $\boldsymbol{v}_1$ en (46) y multiplicando a la izquierda por $(\boldsymbol{a}_1 - \alpha\boldsymbol{e}_1)^T$, resulta

$$2(\boldsymbol{a}_1 - \alpha\boldsymbol{e}_1)^T\boldsymbol{a}_1 = ||\boldsymbol{a}_1 - \alpha\boldsymbol{e}_1||_2^2,$$

de donde se obtiene $2(||\boldsymbol{a}_1||_2^2 - \alpha\xi_1) = ||\boldsymbol{a}_1||_2^2 - 2\alpha\xi_1 + \alpha^2$, donde $\xi_1 := \boldsymbol{e}_1^T\boldsymbol{a}_1$ es la primera componente del vector $\boldsymbol{a}_1$. Por lo tanto, es necesario que $\alpha = \pm||\boldsymbol{a}_1||_2$.

Para evitar que el vector resultante tenga norma pequeña (lo que puede generar overflow) se toma

$$\alpha = -sign(\xi_1)||\boldsymbol{a}_1||_2,$$

lo cual da

$$\boldsymbol{v}_1 = \frac{\boldsymbol{a}_1 + sign(\xi_1)||\boldsymbol{a}_1||_2\boldsymbol{e}_1}{||\boldsymbol{a}_1 + sign(\xi_1)||\boldsymbol{a}_1||_2\boldsymbol{e}_1||_2}. \tag{47}$$

En síntesis, la primera columna de A se puede transformar en un múltiplo del vector $\boldsymbol{e}_1$, multiplicando A a la izquierda por la matriz de Householder $P_1 = P_{\boldsymbol{v}_1}$, más exactamente

$$P_{\boldsymbol{v}_1}\boldsymbol{a}_1 = (-sign(\xi_1)||\boldsymbol{a}_1||_2, 0, \ldots, 0)^T, \qquad \xi_1 = \boldsymbol{e}_1^T\boldsymbol{a}_1$$

Como ilustración, sea

$$A = \begin{bmatrix} -1 & -5/2 & 2 \\ 2 & 4 & 0 \\ -2 & -3 & 4 \end{bmatrix}. \tag{48}$$

Entonces,

$$\boldsymbol{v}_1 = \frac{\boldsymbol{a}_1 + sign(\boldsymbol{e}_1^T\boldsymbol{a}_1)||\boldsymbol{a}_1||_2\boldsymbol{e}_1}{||\boldsymbol{a}_1 + sign(\boldsymbol{e}_1^T\boldsymbol{a}_1)||\boldsymbol{a}_1||_2\boldsymbol{e}_1||_2} = \frac{(-4,2,-2)^T}{||(-4,2,-2)^T||_2} = \frac{1}{2\sqrt{6}}(-4,2,-2).$$

Luego,

$$P_1 = I_3 - 2\boldsymbol{v}_1\boldsymbol{v}_1^T = \frac{1}{3}\begin{bmatrix} -1 & 2 & -2 \\ 2 & 2 & 1 \\ -2 & 1 & 2 \end{bmatrix} \Rightarrow P_1\boldsymbol{a}_1 = P_{\boldsymbol{v}_1}\boldsymbol{a}_1 = \begin{bmatrix} 3 \\ 0 \\ 0 \end{bmatrix}.$$

Si se define $X_1 := P_1A$, entonces $X_1\boldsymbol{e}_1 = \alpha\boldsymbol{e}_1$. Ahora, suponga que inductivamente después de $k-1$ pasos sucesivos, la matriz A se ha transformado en la matriz

$$X_k = P_{k-1}P_{k-2}\cdots P_1A = \begin{bmatrix} x_{11} & x_{12} & x_{13} & \cdots & \cdots & \cdots & x_{1m} \\ & x_{22} & x_{23} & \cdots & \cdots & \cdots & x_{2m} \\ & & x_{33} & \cdots & \cdots & \cdots & x_{3m} \\ & & & \ddots & \cdots & \cdots & \vdots \\ & & & & x_{kk} & \cdots & \vdots \\ & & & & x_{(k+1)k} & \cdots & x_{(k+1)m} \\ & & & & \vdots & \vdots & \vdots \\ & & & & x_{nk} & \cdots & x_{nm} \end{bmatrix} \tag{49}$$

Esta matriz es triangular superior hasta la columna $k-1$. Para avanzar un paso más, hay que transformarla en triangular superior hasta la columna k sin alterar la forma de las columnas previas. Esto se logra seleccionando un vector $\boldsymbol{v}_k$ que tenga ceros en las primeras $k-1$ componentes. Así, la siguiente matriz de Householder $P_k := P_{\boldsymbol{v}_k}$ se define como

$$P_k = I - 2\boldsymbol{v}_k\boldsymbol{v}_k^T, \qquad \boldsymbol{v}_k = \frac{\boldsymbol{z}}{||\boldsymbol{z}||_2}, \tag{50}$$

donde las componentes del vector $\boldsymbol{z}$ están dadas por

$$z_i = \begin{cases} 0, & \text{if } i < k, \\ \beta + x_{ii}, & \text{if } i = k, \\ x_{ik}, & \text{if } i > k, \end{cases} \tag{51}$$

con

$$\beta = sign(x_{kk})\sqrt{\sum_{i=k}^{n} x_{ik}^2}. \tag{52}$$

Ahora suponga que se han realizado $m-1$ transformaciones de Householder a la matriz A hasta reducirla a la forma triangular superior

$$X_m := P_{m-1}P_{m-2}\cdots P_1 A = \begin{bmatrix} x_{11} & x_{12} & x_{13} & \cdots & x_{1m} \\ & x_{22} & x_{23} & \cdots & x_{2m} \\ & & & \ddots & \vdots \\ & & & & x_{mm} \\ & & & & 0 \\ & & & & \vdots \\ & & & & 0 \end{bmatrix} \tag{53}$$

Para construir las matrices Q y R, sea $P := P_{m-1}P_{m-2}\cdots P_1$, note que P es ortogonal por ser producto de matrices ortogonales. Así (53) puede escribirse como

$$PA = \begin{bmatrix} R \\ O \end{bmatrix} = E_m R, \tag{54}$$

donde R es una matriz triangular superior $m \times m$, O es una matriz nula $(n-m) \times m$ y E_m es la matriz de tamaño $n \times m$ formada por las primeras m columnas de la matriz identidad $n \times n$. Puesto que P es ortogonal, de (54) vemos que

$$A = P^T E_m R.$$

Defina $Q := P^T E_m$, entonces

$$Q^T Q = E_m^T P P^T E_m = I,$$

es decir, las columnas de Q son ortonormales. Note que Q es de tamaño $n \times m$ y

$$Q\boldsymbol{e}_j = P^T E_m \boldsymbol{e}_j = P_1 P_2 \cdots P_{m-1} \boldsymbol{e}_j. \tag{55}$$

En síntesis,

$$A = QR,$$

donde R es una matriz triangular obtenida de la reducción de Householder de A (ver (53) y (54)) y

$$Q\boldsymbol{e}_j = P^T E_m = P_1 P_2 \cdots P_{m-1} \boldsymbol{e}_j.$$

Además, si $m = n$, entonces Q es ortogonal.

Ilustremos el procedimiento descrito con la matriz A dada por (48). Se tiene que

$$X_1 := P_1 A = \frac{1}{3} \begin{bmatrix} 9 & 33/2 & -10 \\ 0 & 0 & 8 \\ 0 & 3 & 4 \end{bmatrix}.$$

Usando las expresiones (50)-(51) y (52) con $k = 2$, se obtiene que

$$\boldsymbol{v}_2 = \frac{1}{\sqrt{2}} \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix} \qquad \text{y} \qquad P_2 = I - 2\boldsymbol{v}_2 \boldsymbol{v}_2^T = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & -1 & 0 \end{bmatrix}.$$

Así,

$$X_3 := \underbrace{P_2 P_1}_{=:P} A = \frac{1}{3} \begin{bmatrix} -1 & 2 & -2 \\ 2 & -1 & -2 \\ -2 & -2 & -1 \end{bmatrix} \begin{bmatrix} -1 & 5/2 & 2 \\ 2 & 4 & 0 \\ -2 & -3 & 4 \end{bmatrix} = \frac{1}{3} \begin{bmatrix} 9 & 33/2 & -10 \\ 0 & -3 & -4 \\ 0 & 0 & -8 \end{bmatrix} =: R$$

Por lo tanto,

$$A = P^T R = \begin{bmatrix} -1/3 & 2/3 & -2/3 \\ 2/3 & -1/3 & -2/3 \\ -2/3 & -2/3 & -1/3 \end{bmatrix} \begin{bmatrix} 3 & 11/2 & -10/3 \\ 0 & -1 & -4/3 \\ 0 & 0 & -8/3 \end{bmatrix}.$$

Observación 1.3.

(i) Puesto que

$$(I - 2\boldsymbol{v}\boldsymbol{v}^T)A = A - \boldsymbol{v}\boldsymbol{w}^T, \qquad \boldsymbol{w} = 2A^T\boldsymbol{v},$$

para multiplicar a la izquierda por reflector de Householder no es necesario calcular explícitamente las matrices de Householder.

(ii) Aunque en el ejemplo presentado arriba, se usó el reflector de Householder para anular los elementos que están debajo de la primera componente en $\boldsymbol{a}_1$, las matrices de Householder permiten también anular bloques de componentes de un vector dado $\boldsymbol{x} \in \mathbb{R}^n$. En particular, si se desea anular todas las componentes de $\boldsymbol{x}$, excepto la $m-$ésima componente, se toma

$$\boldsymbol{v} = \frac{\boldsymbol{x} + sign(\xi)||\boldsymbol{x}||_2\boldsymbol{e}_m}{||\boldsymbol{x} + sign(\xi)||\boldsymbol{x}||_2\boldsymbol{e}_m||_2}, \tag{56}$$

donde $\boldsymbol{e}_m$ es el $m-$ésimo vector unitario coordenado de $\mathbb{R}^n$ y $\xi = \boldsymbol{e}_m^T\boldsymbol{x}$. Así,

$$P_{\boldsymbol{v}}\boldsymbol{x} = (0, \ldots, 0, \underbrace{-sign(\xi)||\boldsymbol{x}||_2}_{m-\text{ésima componente}}, 0, \ldots, 0)^T. \tag{57}$$

Por ejemplo, si $\boldsymbol{x} = (-1, 2, -2)^T$ y $m = 2$, entonces

$$\boldsymbol{v} = \frac{\boldsymbol{x} + sign(\boldsymbol{e}_2^T\boldsymbol{x})||\boldsymbol{x}||_2\boldsymbol{e}_2}{||\boldsymbol{x} + sign(\boldsymbol{e}_2^T\boldsymbol{x})||\boldsymbol{x}||_2\boldsymbol{e}_2||_2} = \frac{(-1, 5, -2)^T}{||(-1, 5, -2)^T||} = \frac{1}{\sqrt{30}}(-1, 5, -2).$$

Luego,

$$P_{\boldsymbol{v}} = I_3 - 2\boldsymbol{v}\boldsymbol{v}^T = \frac{1}{15}\begin{pmatrix} 14 & 5 & -2 \\ 5 & -10 & 10 \\ -2 & 10 & 11 \end{pmatrix}, \qquad P_{\boldsymbol{v}}\boldsymbol{x} = \begin{pmatrix} 0 \\ -3 \\ 0 \end{pmatrix}.$$

Algoritmo (Ortogonalización de Householder)**.**

Defina $A := [\boldsymbol{a}_1|\boldsymbol{a}_2|\cdots|\boldsymbol{a}_m]$

$\boldsymbol{v}_1 = \dfrac{\boldsymbol{a}_1 + sign(\xi_1)||\boldsymbol{a}_1||_2\boldsymbol{e}_1}{||\boldsymbol{a}_1 + sign(\xi_1)||\boldsymbol{a}_1||_2\boldsymbol{e}_1||_2}$

$\boldsymbol{r}_1 := P_1\boldsymbol{a}_1$ con $P_1 = I - 2\boldsymbol{v}_1\boldsymbol{v}_1^*$

$\boldsymbol{q}_1 := P_1\boldsymbol{e}_1$

for $k = 2, ..., m$

$\boldsymbol{r}_k := P_{k-1}P_{k-2}\cdots P_1\boldsymbol{a}_k$
$\boldsymbol{v}_k = \dfrac{\boldsymbol{z}}{||\boldsymbol{z}||_2}$, con $\boldsymbol{z}$ dado por (51) y (52)
$\boldsymbol{r}_k := P_k\boldsymbol{r}_k$ con $P_k = I - 2\boldsymbol{v}_k\boldsymbol{v}_k^*$
$\boldsymbol{q}_k := P_1P_2\cdots P_k\boldsymbol{e}_k$
end

Dado un sistema lineal $A\boldsymbol{x} = \boldsymbol{b}$, si $A \in \mathbb{R}^{n\times n}$ es no singular y $A = QR$ es una factorización QR, el sistema puede expresarse como $QR\boldsymbol{x} = \boldsymbol{b}$, o bien

$$R\boldsymbol{x} = Q^*\boldsymbol{b}$$

Como la matriz de coeficientes del sistema equivalente es triangular, el sistema se puede resolver fácilmente, pero se requiere primero calcular R y Q. El método de eliminación por factorización QR usando matrices de Householder, es una alternativa al método de eliminación Gaussiana. Sin embargo, el método por factorización QR requiere $2n^3/3 + \mathcal{O}(n^2)$ multiplicaciones, lo que duplica el número de operaciones realizadas en la eliminación Gaussiana.

Reducción a la forma de Hessenberg usando matrices de Householder

En algunas aplicaciones de interés se requiere reducir, con un bajo costo computacional, una matriz dada a una forma casi triangular mediante transformaciones de similaridad de manera que la factorización QR de la matriz obtenida se pueda realizar de una manera eficiente. En particular, esto se requiere en el método de iteración QR para aproximar valores propios en el cual se hace uso intensivo de dicha factorización. El tipo de matrices casi triangulares de interés son las matrices de Hessenberg, las cuales se caracterizan por tener ceros debajo de la primera subdiagonal. Más exactamente, $H = (h_{ij})$ es una *matriz de Hessenberg* (superior) si $h_{ij} = 0$ para $i > j + 1$. Cualquier matriz se puede reducir a la forma de Hessenberg en un número finito de pasos usando transformaciones de similaridad ortogonales y esto lo haremos usando matrices de Householder. El procedimiento es el siguiente: dada la k-ésima columna $\boldsymbol{x}$ de una matriz $n \times n$, se pueden anular las componentes de $\boldsymbol{x}$ desde posición $k+2$ hasta la posición n sin cambiar las primeras k componentes mediante la matriz de Householder P_k dada por

$$\widetilde{P}_k = \begin{bmatrix} I_k & \boldsymbol{O}_{k\times(n-k)} \\ \boldsymbol{O}_{(n-k)\times k} & P_{n-k} \end{bmatrix}, \qquad \text{con} \qquad P_{n-k} = I_{n-k} - 2\boldsymbol{w}_k\boldsymbol{w}_k^T, \tag{58}$$

donde I_k es la matriz identidad $k \times k$, $\boldsymbol{O}$ denota la matriz nula y P_{n-k} es el reflector de Householder $(n-k) \times (n-k)$ asociado con la reflexión respecto al hiperplano ortogonal al vector $\boldsymbol{w}_k \in \mathbb{R}^{n-k}$. Usando (56), el vector $\boldsymbol{w}_k$ está dado por

$$\boldsymbol{w}_k = \frac{\boldsymbol{x}_{n-k} \pm ||\boldsymbol{x}_{n-k}||_2 \boldsymbol{e}_1}{||\boldsymbol{x}_{n-k} \pm ||\boldsymbol{x}_{n-k}||_2 \boldsymbol{e}_1||_2}, \tag{59}$$

siendo $\boldsymbol{x}_{n-k}$ el vector de $\mathbb{R}^{n-k}$ formado por las últimas $n-k$ componentes de $\boldsymbol{x}$ y $\boldsymbol{e}_1$ el primer vector unitario coordenado de $\mathbb{R}^{n-k}$. Además, las componentes $y_1, y_2, \ldots, y_n$ del vector $\boldsymbol{y} := \widetilde{P}_k \boldsymbol{x}$ vienen dadas por

$$y_i = \begin{cases} x_i, & \text{si } 1 \leq i \leq k, \\ 0, & \text{si } i = k+1. \\ \pm||\boldsymbol{x}_{n-k}||_2, & \text{si } k+2 \leq i \leq n \end{cases}$$

Como un ejemplo simple, pero ilustrativo, consideremos nuevamente la matriz

$$A = \begin{bmatrix} -1 & -5/2 & 2 \\ 2 & 4 & 0 \\ -2 & -3 & 4 \end{bmatrix}. \tag{60}$$

Para llevar la matriz A a la forma de Hessenberg, basta con anular la tercera componente de la columna $k = 1$. Así, usando (58), la matriz de Householder $\widetilde{P}_1$ es

$$\widetilde{P}_1 = \begin{bmatrix} 1 & \boldsymbol{O}_{1\times 2} \\ \boldsymbol{O}_{2\times 1} & P_2 \end{bmatrix}, \qquad \text{con} \qquad P_2 = I_2 - 2\boldsymbol{w}_1 \boldsymbol{w}_1^T.$$

Como

$$\boldsymbol{w}_1 = \frac{\boldsymbol{x}_2 - ||\boldsymbol{x}_2||_2 \boldsymbol{e}_1}{||\boldsymbol{x}_2 - ||\boldsymbol{x}_2||_2 \boldsymbol{e}_1||_2} = \frac{(2,-2)^T - ||(2,-2)^T||_2 (1,0)^T}{||(2,-2)^T - ||(2,-2)^T||_2 (1,0)^T||_2} = \frac{(1-\sqrt{2},-1)^T}{\sqrt{4-2\sqrt{2}}},$$

entonces

$$P_2 = I_2 - 2\boldsymbol{w}_1 \boldsymbol{w}_1^T = \begin{bmatrix} \sqrt{2}/2 & -\sqrt{2}/2 \\ -\sqrt{2}/2 & -\sqrt{2}/2 \end{bmatrix}, \qquad \widetilde{P}_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \sqrt{2}/2 & -\sqrt{2}/2 \\ 0 & -\sqrt{2}/2 & -\sqrt{2}/2 \end{bmatrix}$$

La matriz $\widetilde{P}_1$ es ortogonal y

$$\widetilde{P}_1^T A = \begin{bmatrix} -1 & -5/2 & 2 \\ 2\sqrt{2} & 7\sqrt{2}/2 & -2\sqrt{2} \\ 0 & -\sqrt{2}/2 & -2\sqrt{2}, \end{bmatrix}, \quad \widetilde{P}_1^T A \widetilde{P}_1 = \begin{bmatrix} -1 & -9\sqrt{2}/4 & \sqrt{2}/4 \\ 2\sqrt{2} & 11/2 & -3/2 \\ 0 & 3/2 & 5/2, \end{bmatrix} =: H$$

En general, para una matriz $n \times n$, si definimos $A_0 = A$, el k-ésimo paso consiste en multiplicar por la matriz $\widetilde{P}_k$ que anula las posiciones $k+2, k+3, \ldots, n$ de la k-ésima columna de A_{k-1} para $k = 1, 2, \ldots, n-2$. De esta manera, se genera una sucesión de matrices $\{A_k\}$ que son ortogonalmente similares a A y

$$A_k = \widetilde{P}_k^T A_{k-1} \widetilde{P}_k = \cdots = \widetilde{P}_k^T \cdots \widetilde{P}_1^T A \widetilde{P}_1 \cdots \widetilde{P}_k = \widetilde{Q}_k^T A \widetilde{Q}_k, \qquad k \geq 1,$$

donde $\widetilde{Q}_k := \widetilde{P}_1 \cdots \widetilde{P}_k$ es ortogonal. Después de $n-2$ pasos, el producto $Q = \widetilde{P}_1 \cdots \widetilde{P}_{n-2}$ es una matriz ortogonal con $Q^T A Q = H$ siendo H una matriz de Hessenberg superior. Si A es simétrica, entonces $H^T = (Q^T A Q)^T = Q^T A Q = H$, y en consecuencia, la matriz de Hesserberg H es tridiagonal.

4. Mínimos cuadrados lineales

Hasta ahora se han considerado sistemas lineales de ecuaciones en los cuales la matriz de coeficientes A es cuadrada. Sin embargo, en algunas aplicaciones surgen sistemas cuyas matrices de coeficientes no son cuadradas. El caso más común es cuando $A \in \mathbb{R}^{m\times n}$ con $m > n$, es decir, cuando el sistema lineal tiene más ecuaciones que incógnitas, el cual es llamado sistema *sobredeterminado*. Por otra parte, cuando el sistema tiene más incógnitas que ecuaciones se llama *subdeterminado*. En el caso de un sistema cuadrado, la invertibilidad de la matriz de coeficientes garantizaba la existencia y unicidad de la solución, pero para un sistema lineal $A\boldsymbol{x} = \boldsymbol{b}$, con $A \in \mathbb{R}^{m\times n}$ y $m \neq n$ es posible que no tenga solución o no haya solución única.

En estadística aparece de manera usual el problema de ajuste de datos, el cual consiste en que dado un conjunto puntos (datos) (t_i, y_i), $i = 1, \ldots, m$, se desea hallar un vector $\boldsymbol{x} \in \mathbb{R}^n$ de parámetros que proporcione el 'mejor ajuste' a la función modelo $f(t, \boldsymbol{x})$, siendo f una función definida en $\mathbb{R}^{n+1}$ y con valores reales. Por mejor ajuste queremos decir que la suma de los cuadrados de las distancias entre los datos y el modelo (estas distancias corresponden a los errores) sea mínima, lo que se conoce como el problema de *mínimos cuadrados*:

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} \sum_{i=1}^{m} (y_i - f(t_i, \boldsymbol{x}))^2.$$

Cuando la función f es lineal en las componentes del vector parámetro $\boldsymbol{x}$, es decir, f tiene la forma

$$f(t, \boldsymbol{x}) = \phi_1(t)x_1 + \phi_2(t)x_2 + \cdots + \phi_n(t)x_n,$$

entonces se dice que el problema de mínimos cuadrados es *lineal*.

Tomando $a_{ij} = \phi_j(t_i)$ y $b_i = y_i$, con $i = 1, \ldots, m$ y $j = 1, \ldots, n$, se tiene que el problema lineal de miñimos cuadrados se puede escribir en forma matricial como

$$A\boldsymbol{x} \approx \boldsymbol{b}.$$

Observe que se escribe $A\boldsymbol{x} \approx \boldsymbol{b}$ en vez de $A\boldsymbol{x} = \boldsymbol{b}$ porque la ecuación no se satisface de manera exacta. Como ilustración consideremos el problema de ajuste de los m datos (t_1, y_1), (t_2, y_2),. . . ,(t_m, y_m) con un polinomio lineal $f(t, \boldsymbol{x}) = x_1 + x_2 t$, el cual recibe en estadística el nombre de *recta de regresión*

de mínimos cuadrados. En este caso, el problema tiene la forma

$$A\boldsymbol{x} = \begin{pmatrix} 1 & t_1 \\ 1 & t_2 \\ \vdots & \vdots \\ 1 & t_m \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \approx \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix} = \boldsymbol{b}.$$

Ejemplo 1.5. Veamos un sencillo ejemplo de mínimos cuadrados ajustando un polinomio lineal $f(t) = x_1 + x_2 t$ a los siguientes datos

t	-1.0	0.0	0.5	1.0
y	-1.5	0.0	0.5	2.0

El sistema sobredeterminado 4×2 queda así

$$A\boldsymbol{x} = \begin{pmatrix} 1 & -1 \\ 1 & 0 \\ 1 & 0.5 \\ 1 & 1.0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \approx \begin{pmatrix} -1.5 \\ 0.0 \\ 0.5 \\ 2.0 \end{pmatrix} = \boldsymbol{b}.$$

La solución de este sistema, la cual veremos como calcular más adelante, es $\boldsymbol{x} = (0.0428, 1.6572)^T$. Luego el polinomio que ajusta los datos es (ver Figura 3)

$$f(t) = 0.0428 + 1.6572t.$$

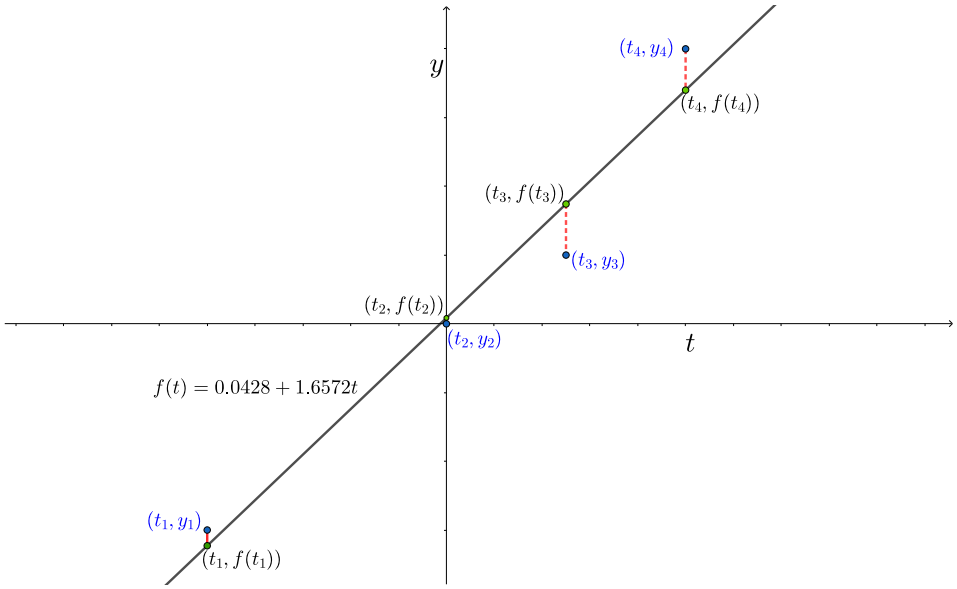

Figura 3: Ajuste de mínimos cuadrados de un polinomio lineal para los datos del Ejemplo 1.5.

El método clásico para resolver problemas de mínimos cuadrados lineales fue desarrollado por Gauss al estudiar las órbitas de algunos cuerpos celestes y se puede obtener de distintas maneras. Una de ellas se describe a continuación usando cálculo de varias variables. Denotemos por $\boldsymbol{r}$ el *vector residual*

$$\boldsymbol{r} = \boldsymbol{b} - A\boldsymbol{x},$$

entonces, en notación matricial, el problema de mínimos cuadrados puede expresarse como el problema de minimizar la norma Euclidiana

$$||\boldsymbol{r}||_2^2 = \langle \boldsymbol{r}, \boldsymbol{r} \rangle = \boldsymbol{r}^T \boldsymbol{r}.$$

Ahora, para minimizar

$$\Phi(\boldsymbol{x}) := ||\boldsymbol{r}||_2^2 = \boldsymbol{r}^T\boldsymbol{r} = (\boldsymbol{b} - A\boldsymbol{x})^T(\boldsymbol{b} - A\boldsymbol{x}) = \boldsymbol{b}^T\boldsymbol{b} - 2\boldsymbol{x}^T A^T\boldsymbol{b} + \boldsymbol{x}^T A^T A\boldsymbol{x},$$

se iguala el gradiente (con respecto a $\boldsymbol{x}$) de $\Phi(\boldsymbol{x})$ a cero. Como para $i = 1, \ldots, n$

$$\begin{aligned}
\frac{\partial(\boldsymbol{x}^T A^T\boldsymbol{b})}{\partial x_i} &= \frac{\partial(\boldsymbol{x}^T)}{\partial x_i} A^T\boldsymbol{b} = \boldsymbol{e}_i^T A^T\boldsymbol{b} = (A^T\boldsymbol{b})_i, \\
\frac{\partial(\boldsymbol{x}^T A^T A\boldsymbol{x})}{\partial x_i} &= \frac{\partial(\boldsymbol{x}^T)}{\partial x_i} A^T A\boldsymbol{x} + \boldsymbol{x}^T A^T A\frac{\partial \boldsymbol{x}}{\partial x_i} = \boldsymbol{e}_i^T A^T A\boldsymbol{x} + \boldsymbol{x}_i^T A^T A\boldsymbol{e}_i \\
&= (2A^T A\boldsymbol{x})_i,
\end{aligned}$$

se sigue que

$$2A^T A\boldsymbol{x} - 2A^T\boldsymbol{b} = \boldsymbol{0}_{\mathbb{R}^n},$$

lo cual se reduce al sistema lineal cuadrado $n \times n$

$$A^T A\boldsymbol{x} = A^T\boldsymbol{b}.$$

Este último sistema es comúnmente conocido como el sistema de *ecuaciones normales*. Nótese que la matriz $A^T A$ es simétrica, y si A tiene rango completo ($= n$), es decir, las columnas de A son linealmente independientes, entonces la matriz $A^T A$ es no singular (pues $rank(A^T A) = rank(A)$), y en consecuencia, el sistema de ecuaciones normales tiene solución única, la cual es también la única solución del problema original de mínimos cuadrados.

Usando cálculo se ha verificado que el mínimo de Φ ocurre en alguna solución del sistema de ecuaciones normales. Veamos ahora que cada solución del sistema de ecuaciones normales es una solución del problema de mínimos

cuadrados, es decir, que la función Φ alcanza su valor mínimo en cada solución del sistema de ecuaciones normales. En efecto, si $\boldsymbol{z}$ es una solución del sistema de ecuaciones normales, entonces

$$\Phi(\boldsymbol{z}) = \boldsymbol{b}^T\boldsymbol{b} - 2\boldsymbol{z}^T A^T\boldsymbol{b} + \boldsymbol{z}^T A^T A\boldsymbol{z} = \boldsymbol{b}^T\boldsymbol{b} - 2\boldsymbol{z}^T A^T\boldsymbol{b} + \boldsymbol{z}^T A^T\boldsymbol{b} = \boldsymbol{b}^T\boldsymbol{b} - \boldsymbol{z}^T A^T\boldsymbol{b}.$$

Ahora, para cualquier $\boldsymbol{w} \in \mathbb{R}^n$, sea $\boldsymbol{u} = \boldsymbol{w} - \boldsymbol{z}$. Entonces

$$\begin{aligned}\Phi(\boldsymbol{w}) = \Phi(\boldsymbol{z}+\boldsymbol{u}) &= \boldsymbol{b}^T\boldsymbol{b} - 2(\boldsymbol{z}+\boldsymbol{u})^T A^T A\boldsymbol{b} + (\boldsymbol{z}+\boldsymbol{u})^T A^T A(\boldsymbol{z}+\boldsymbol{u}) \\ &= \boldsymbol{b}^T\boldsymbol{b} - \boldsymbol{z}^T A^T\boldsymbol{b} + \boldsymbol{u}^T A^T A\boldsymbol{u} \\ &= \Phi(\boldsymbol{z}) + ||A\boldsymbol{u}||_2^2 \\ &\geq \Phi(\boldsymbol{z}).\end{aligned}$$

Luego, $\Phi(\boldsymbol{w}) \geq \Phi(\boldsymbol{z})$ para todo $\boldsymbol{w} \in \mathbb{R}^n$, i.e., Φ alcanza su valor mínimo en cada solución de las ecuaciones normales.

Observación 1.4.

El sistema de ecuaciones normales $A^T A\boldsymbol{x} = A^T\boldsymbol{b}$ es siempre consistente, independientemente de si el sistema $A\boldsymbol{x} = \boldsymbol{b}$ lo es. En efecto, se conoce del álgebra lineal que

$$R(A^T) := \{A^T\boldsymbol{v} : \ \boldsymbol{v} \in \mathbb{R}^n\} = R(A^T A),$$

lo cual implica que $A^T\boldsymbol{b} \in R(A^T A)$, es decir, el lado derecho del sistema de ecuaciones normales está en el espacio imagen (espacio generado por las columnas) de la matriz de coeficientes, en consecuencia, se tiene la consistencia.

Por otra parte, si $A\boldsymbol{x} = \boldsymbol{b}$ es consistente y tiene solución única, entonces lo mismo se tiene para el sistema de ecuaciones normales, y la única solución (común) para ambos sistemas es

$$\boldsymbol{x} = (A^T A)^{-1} A^T\boldsymbol{b}. \tag{61}$$

Esto es cierto porque una solución única para ambos sistemas existe si y solo si $\{\boldsymbol{0}\} = ker(A) = ker(A^T A)$, y esto asegura que $A^T A$ debe ser no singular, así su única solución está dada por (61). Aquí, $ker(A)$ es el núcleo o espacio nulo de A:

$$ker(A) = \{\boldsymbol{v} \in \mathbb{R}^n : \ A\boldsymbol{v} = \boldsymbol{0}\}.$$

Veamos ahora dos alternativas para resolver el sistema de ecuaciones normales.

Métodos que usan la factorización LU

Si A tiene rango completo, entonces la matriz A^TA es no singular. Más aun, A^TA es simétrica y definida positiva y en consecuencia, tiene un factorización de Cholesky

$$A^TA = LL^T.$$

Entonces la solución del sistema de ecuaciones normales $A^TA\boldsymbol{x} = A^T\boldsymbol{b}$ se puede calcular resolviendo los sistemas triangulares $L\boldsymbol{y} = A^T\boldsymbol{b}$ y $L^T\boldsymbol{x} = \boldsymbol{y}$. Ilustremos este método con el sencillo ejemplo 1.5:

$$A^TA = \begin{pmatrix} 1 & 1 & 1 & 1 \\ -1 & 0 & 0.5 & 1 \end{pmatrix} \begin{pmatrix} 1 & -1 \\ 1 & 0 \\ 1 & 0.5 \\ 1 & 1 \end{pmatrix} = \begin{pmatrix} 4 & 0.5 \\ 0.5 & 2.25 \end{pmatrix},$$

$$A^T\boldsymbol{b} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ -1 & 0 & 0.5 & 1 \end{pmatrix} \begin{pmatrix} -1.5 \\ 0 \\ 0.5 \\ 2.0 \end{pmatrix} = \begin{pmatrix} 1 \\ 3.75 \end{pmatrix}.$$

La factorización de Cholesky de A^TA es

$$A^TA = \begin{pmatrix} 4 & 0.5 \\ 0.5 & 2.25 \end{pmatrix} = \begin{pmatrix} 2 & 0 \\ 0.25 & 1.4790 \end{pmatrix} \begin{pmatrix} 2 & 0.25 \\ 0 & 1.4790 \end{pmatrix} = LL^T.$$

Resolviendo el sistema triangular $L\boldsymbol{y} = A^T\boldsymbol{b}$ por sustitución progresiva, se obtiene $\boldsymbol{y} = (0.5, 2.4510)^T$. Finalmente, al resolver el sistema triangular $L^T\boldsymbol{x} = \boldsymbol{y}$, resulta $\boldsymbol{x} = (0.0428, 1.6572)^T$.

Teóricamente la factorización LU luce atractiva, sin embargo, en general no es recomendable desde el punto de vista numérico por varias razones. Una de ellas es que el cálculo de A^TA con aritmética de punto flotante puede ocasionar pérdida de información significativa como lo ilustra el siguiente ejemplo. Suponga que se trabaja con una máquina con aritmética de punto flotante igual a 10^{-8} y tome

$$A = \begin{pmatrix} 1 & 1 \\ 10^{-5} & 0 \\ 0 & 10^{-5} \end{pmatrix}$$

Esta matriz tiene rango completo. Por otra parte,

$$A^TA = \begin{pmatrix} 1 + 10^{-10} & 1 \\ 1 & 1 + 10^{-10} \end{pmatrix}$$

Esta matriz en aritmética de punto flotante es aproximada a

$$A^TA = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix},$$

la cual es singular. Otra desventaja de este enfoque es que si A está mal condicionada, entonces A^TA está mucho peor en este sentido. Por ejemplo, si

$$A = \begin{pmatrix} 10^{-2} & 0 \\ 0 & 1 \end{pmatrix},$$

entonces $\kappa_2(A) = 100$, mientras que $\kappa_2(A^TA) = 10000$.

Métodos que usan la factorización QR

Un método alternativo que soslaya las potenciales dificultades numéricas de la factorización LU, consiste en aprovechar la factorización QR de la matriz A. Suponga nuevamente que A tiene rango completo (recuerde que esto implica que existe una única solución de mínimos cuadrados), y considere la factorización QR reducida de A (ver Teorema 1.8), esto es, $A = QR$ donde $R \in \mathbb{R}^{n\times n}$ es una matriz triangular superior con entradas diagonales positivas y las columnas de $Q \in \mathbb{R}^{m\times n}$ forman un conjunto ortonormal, es decir, $Q^TQ = I_n$. Luego,

$$A^TA = (QR)^TQR = R^TQ^TQR = R^TR.$$

En consecuencia, el sistema de ecuaciones normales se puede reescribir como

$$R^TR\boldsymbol{x} = R^TQ^T\boldsymbol{b}.$$

Como R^T es no singular (pues es triangular con entradas diagonales positivas), este sistema se reduce al sistema triangular

$$R\boldsymbol{x} = Q^T\boldsymbol{b},$$

el cual se resuelve de manera eficiente por sustitución regresiva. Note que

$$\boldsymbol{x} = R^{-1}Q^T\boldsymbol{b} = (A^TA)^{-1}A^T\boldsymbol{b}$$

es la solución de $A\boldsymbol{x} = \boldsymbol{b}$ cuando este sistema es consistente y también es la solución de mínimos cuadrados cuando el sistema es inconsistente (ver Observación 61). Por lo tanto, el método para resolver el problema de mínimos cuadrados con la factorización QR aplica independientemente de si el sistema $A\boldsymbol{x} = \boldsymbol{b}$ es o no consistente.

Nótese también que, al resolver el sistema de ecuaciones normales usando factorización QR, se evita el cálculo del producto A^TA. Pero aún en caso de que requiera el cálculo de dicho producto, al usar la factorización (única) de Cholesky se tiene

$$A^TA = LL^T = A^TA = R^TQ^TQR = R^TR,$$

de donde se deduce que R^T es el factor de Cholesky L de A^TA.

Ejercicio 1.3.

1. Pruebe que la recta de regresión de mínimos cuadrados para los datos (t_1, y_1), (t_2, y_2),...,(t_m, y_m) está dada por $f(t, \boldsymbol{x}) = x_1 + x_2t$, donde

$$x_2 = \frac{m\sum_{i=1}^m t_iy_i - \sum_{i=1}^m t_i\sum_{i=1}^m y_i}{m\sum_{i=1}^m t_i^2 - \left(\sum_{i=1}^m t_i\right)^2} \qquad \text{y} \qquad x_1 = \frac{1}{m}\left(\sum_{i=1}^m y_i - x_2\sum_{i=1}^m t_i\right) \tag{62}$$

2. Muestre que la pendiente de la recta que pasa por el origen del plano y que ajusta, en el sentido de mínimos cuadrados, los puntos (t_1, y_1), (t_2, y_2),...,(t_m, y_m), está dada por

$$x_2 = \frac{\sum_{i=1}^m t_iy_i}{\sum_{i=1}^m t_i^2}.$$

Descomposición en Valores singulares (SVD)

Sea $A \in \mathbb{C}^{m\times n}$ y considere la matriz Hermitiana A^*A. Si λ es un valor propio de A^*A con vector propio asociado $\boldsymbol{v} \neq \mathbf{0}$, esto es, $A^*A\boldsymbol{v} = \lambda\boldsymbol{v}$, entonces

$$||A\boldsymbol{v}||_2^2 = (A\boldsymbol{v})^*(A\boldsymbol{v}) = \boldsymbol{v}^*A^*A\boldsymbol{v} = \boldsymbol{v}^*(\lambda\boldsymbol{v}) = \lambda\boldsymbol{v}^*\boldsymbol{v} = \lambda||\boldsymbol{v}||_2^2.$$

Luego, los valores propios de A^*A son reales no negativos.

Definición 1.9 (Valores singulares de una matriz)**.** Los *valores singulares* de $A \in \mathbb{C}^{m\times n}$ se definen como las raíces cuadradas no negativas de los valores propios de la matriz Hermitiana A^*A.

Los valores singulares tienen diversas aplicaciones, entre las que mencionamos, procesamiento de imágenes, compresión de datos, para el cálculo del rango de una matriz y en problemas relacionados con aproximaciones de mínimos cuadrados, justamente esta última aplicación se presentará con detalle.

Teorema 1.9 (Descomposición en valores singulares)**.** Sea $A \in \mathbb{C}^{m\times n}$. Entonces existen dos matrices unitarias $U \in \mathbb{C}^{m\times m}$ y $V \in \mathbb{C}^{n\times n}$ tales que

$$U^*AV = \Sigma = diag(\sigma_1, \ldots, \sigma_p) \in \mathbb{R}^{m\times n} \qquad \text{con } p = min\{m, n\}, \tag{63}$$

donde $\sigma_1 \geq \sigma_2 \geq \cdots \sigma_p \geq 0$ son los valores singulares de A.

Si $A \in \mathbb{R}^{m\times n}$, las matrices U y V en la descomposición en valores singulares también tienen entradas reales y en la fórmula simplemente se reemplaza U^* por U^T.

Si $\sigma_1 \geq \sigma_2 \geq \cdots \sigma_r > \sigma_{r+1} = \cdots = \sigma_p = 0$, entonces $rank(A) = r$. De hecho, el rango de una matriz es igual al nmero de valores singulares no nulos que tiene dicha matriz. Como $A = U\Sigma V^T$, entonces al realizar este producto matricial resulta

$$A = \sum_{i=1}^{r} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^T, \tag{64}$$

donde $\boldsymbol{u}_i$ y $\boldsymbol{v}_i$ son los vectores columnas de U y V respectivamente, y reciben el nombre de vectores singulares asociados al valor singular σ_i.

Definición 1.10 (Pseudoinversa de Moore-Penrose)**.** Sea $A \in \mathbb{C}^{n\times n}$ tal que $rank(A) = r$ y que admite la descomposición en valores singulares $U^*AV =$

$diag(\sigma_1, \ldots, \sigma_p)$ con $\sigma_1 \geq \sigma_2 \geq \cdots \sigma_r > \sigma_{r+1} = \cdots = \sigma_p = 0$. La matriz $A^\dagger$ dada por

$$A^\dagger = V\Sigma^\dagger U^* \qquad \text{con } \Sigma^\dagger = diag\left(\frac{1}{\sigma_1}, \ldots, \frac{1}{\sigma_r}, 0, \ldots, 0\right) \tag{65}$$

es llamada la pseudoinversa de Moore-Penrose o inversa generalizada de A.

Si $rank(A) = n < m$, se tiene que $A^\dagger = (A^T A)^{-1} A^T$, y si $rank(A) = m = n$, entonces $A^\dagger = A^{-1}$.

Ahora usaremos la descomposición en valores singulares para abordar el problema de mínimos cuadrados cuando la matriz A no tiene rango completo. En este caso, fallan las técnicas desarrolladas arriba para minimizar

$$\Phi(\boldsymbol{x}) = ||\boldsymbol{b} - A\boldsymbol{x}||_2^2$$

debido a que si $\widehat{\boldsymbol{x}}$ es una solución y $\boldsymbol{w} \in ker(A) \neq \emptyset$, entonces $\widehat{\boldsymbol{x}} + \boldsymbol{w}$ también es solución. Por lo tanto, es necesario introducir una restricción adicional para tener unicidad de la solución de mínimos cuadrados. Una manera de hacerlo es requerir que $\widehat{\boldsymbol{x}}$ tenga norma Euclidiana mínima.

Teorema 1.10. Sea $A \in \mathbb{R}^{m \times n}$ con descomposición en valores singulares dada por $A = U\Sigma V^T$. Entonces problema de hallar el vector de $\mathbb{R}^n$ con norma Euclidiana mínima tal que minimice

$$\Phi(\boldsymbol{x}) = ||\boldsymbol{b} - A\boldsymbol{x}||_2^2, \tag{66}$$

tiene como única solución

$$\widehat{\boldsymbol{x}} = A^\dagger \boldsymbol{b} = \sum_{i=1}^{r} \frac{\boldsymbol{u}_i^T \boldsymbol{b}}{\sigma_i} \boldsymbol{v}_i, \tag{67}$$

donde $A^\dagger$ es la pseudoinversa de Moore-Penrose dada por (65).

Demostración.

Teniendo en cuenta que las matrices ortogonales preservan la norma Euclidiana, se tiene que

$$||\boldsymbol{b} - A\boldsymbol{x}||_2^2 = ||\boldsymbol{b} - U\Sigma V^T \boldsymbol{x}||_2^2 = ||U^T(\boldsymbol{b} - U\Sigma V^T \boldsymbol{x})||_2^2 = ||U^T \boldsymbol{b} - \Sigma V^T \boldsymbol{x}||_2^2.$$

Además, $||\boldsymbol{x}||_2 = ||V^T \boldsymbol{x}||_2$. Luego, el problema original es equivalente a encontrar $\boldsymbol{w} = V^T \boldsymbol{x}$ tal que $\boldsymbol{w}$ tenga norma Euclidiana mínima y

$$||\Sigma \boldsymbol{w} - U^T \boldsymbol{b}||_2^2 \leq ||\Sigma \boldsymbol{z} - U^T \boldsymbol{b}||_2^2, \qquad \forall \boldsymbol{z} \in \mathbb{R}^n. \tag{68}$$

Si $\sigma_1, \sigma_2, \ldots, \sigma_r$ son los valores singulares no nulos de A, entonces

$$||\Sigma \boldsymbol{w} - U^T \boldsymbol{b}||_2^2 = \sum_{i=1}^{r} (\sigma_i w_i - (U^T \boldsymbol{b})_i)^2 + \sum_{i=r+1}^{m} ((U^T \boldsymbol{b})_i)^2.$$

Pero esta última expresión es mínima si $w_i = (U^T \boldsymbol{b})_i / \sigma_i = \boldsymbol{u}_i^T \boldsymbol{b} / \sigma_i$ para $i = 1, \ldots, r$. Así, la solución buscada para (68) es

$$\widehat{\boldsymbol{w}} = \left(\frac{(U^T \boldsymbol{b})_1}{\sigma_1}, \frac{(U^T \boldsymbol{b})_2}{\sigma_2}, \ldots, \frac{(U^T \boldsymbol{b})_r}{\sigma_r}, 0, \ldots, 0 \right) = \Sigma^\dagger U^T \boldsymbol{b}.$$

Claramente, si $\boldsymbol{w}$ es cualquier otro vector que tiene las mismas primeras r componentes que $\widehat{\boldsymbol{w}}$, entonces $||\widehat{\boldsymbol{w}}||_2 \leq ||\boldsymbol{w}||_2$. Finalmente, de $\widehat{\boldsymbol{w}} = \Sigma^\dagger U^T \boldsymbol{b}$, resulta que el vector solución (único) del problema original es

$$\widehat{\boldsymbol{x}} = V \Sigma^\dagger U^T \boldsymbol{b} = A^\dagger \boldsymbol{b} = \sum_{i=1}^{r} \frac{\boldsymbol{u}_i^T \boldsymbol{b}}{\sigma_i} \boldsymbol{v}_i.$$

□

A continuación se presenta uno de los teoremas de reducción más útiles en la teoría elemental de matrices.

Teorema 1.11 (Teorema de Shur)**.** Para cualquier matriz cuadrada A, existe una matriz unitaria U tal que

$$U^*AU = \begin{bmatrix} \lambda_1 & b_{12} & b_{13} & \cdots & b_{1n} \\ & \lambda_2 & b_{23} & \cdots & b_{2n} \\ & & \lambda_3 & \cdots & b_{3n} \\ & & & \ddots & \vdots \\ & & & & \lambda_n \end{bmatrix}, \tag{69}$$

donde λ_i, $i = 1, 2, \ldots, n$ son los valores propios de A en cualquier orden prescrito. Es decir, cualquier matriz cuadrada es unitariamente similar a una matriz triangular superior.

Demostración.

Se procede por inducción sobre la dimensión n, comenzando por el caso $n = 2$ (el caso $n = 1$ es trivial) el cual dará la ideas para el caso más general. Sea A una matriz 2×2, y λ_1 un valor propio de A con vector propio asociado $\boldsymbol{w}_1$ normalizado por la condición $\boldsymbol{w}_1^*\boldsymbol{w}_1 = 1$. El vector no nulo $\boldsymbol{w}_1$ se puede extender a una base $\{\boldsymbol{w}_1, \boldsymbol{z}_2\}$ de $\mathbb{C}^2$. Aplicando el proceso de ortonormalización de Gram-Schmidt a esta base, se obtiene una base ortonormal $\{\boldsymbol{w}_1, \boldsymbol{q}_2\}$ de $\mathbb{C}^2$. La matriz $U := [\boldsymbol{w}_1|\boldsymbol{q}_2]$ es unitaria por construcción y

$$AU = [A\boldsymbol{w}_1|A\boldsymbol{q}_2] = [\lambda_1\boldsymbol{w}_1|A\boldsymbol{q}_2].$$

Luego,

$$U^*(AU) = \begin{bmatrix} \boldsymbol{w}_1^* \\ \boldsymbol{q}_2^* \end{bmatrix} [\lambda\boldsymbol{w}_1|A\boldsymbol{q}_2] = \begin{bmatrix} \lambda_1 & \boldsymbol{w}_1^*A\boldsymbol{q}_2 \\ 0 & \boldsymbol{q}_2^*A\boldsymbol{q}_2 \end{bmatrix}.$$

En consecuencia, los valores propios de $U^*AU = U^{-1}AU$ son λ_1 y $\boldsymbol{q}_2^*A\boldsymbol{q}_2$, pero $U^{-1}AU$ tiene los mismos valores propios de A por ser matrices similares. De donde $\lambda_2 = \boldsymbol{q}_2^*A\boldsymbol{q}_2$.

Suponga ahora que el resultado vale para $n - 1$ y a partir de esto, veamos que vale para n. Sea $A \in \mathbb{C}^{n\times n}$. Como antes, sea λ_1 un valor propio de A con vector propio normalizado asociado $\boldsymbol{w}_1$. Extendamos $\boldsymbol{w}_1$ a una base $\{\boldsymbol{w}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_n\}$ de $\mathbb{C}^n$. Aplíquese el proceso de Gram-Schmidt para obtener una base ortonormal $\{\boldsymbol{w}_1, \boldsymbol{q}_2, \ldots, \boldsymbol{q}_n\}$ de $\mathbb{C}^n$. Entonces, la matriz

$S = [\boldsymbol{w}_1|\boldsymbol{q}_2|\dots|\boldsymbol{q}_n]$ es unitaria y como se hizo para $n = 2$, se tiene que

$$S^*(AS) = \begin{bmatrix} \lambda_1 & c_{12} & c_{13} & \cdots & c_{1n} \\ 0 & & & & \\ \vdots & & & B_{n-1} & \\ 0 & & & & \end{bmatrix}, \tag{70}$$

donde B_{n-1} es una matriz $(n-1) \times (n-1)$.

Como la ecuación característica del segundo miembro de (70) es

$$(\lambda - \lambda_1) \det(\lambda I - B_{n-1}) = 0,$$

se sigue que los valores propios de B_{n-1} son $\lambda_2, \lambda_3, \dots, \lambda_n$, i.e., los restantes valores propios de A. Por la hipótesis de inducción, existe una matriz unitaria V_{n-1} de tamaño $(n-1) \times (n-1)$ tal que

$$V_{n-1}^* B_{n-1} V_{n-1} = \begin{bmatrix} \lambda_2 & d_{12} & \cdots & d_{1(n-1)} \\ & \lambda_3 & \cdots & c_{2(n-1)} \\ & & \ddots & \vdots \\ & & & \lambda_n \end{bmatrix},$$

Sea F la matriz unitaria $n \times n$ formada de la siguiente manera

$$F = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & & & \\ \vdots & & V_{n-1} & \\ 0 & & & \end{bmatrix}$$

Defina $U := SF$, entonces U es unitaria y además,

$$\begin{aligned} U^*AU &= (SF)^* A (SF) \\ &= F^*(S^*AS)F \\ &= \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & & & \\ \vdots & & V_{n-1}^* & \\ 0 & & & \end{bmatrix} \begin{bmatrix} \lambda_1 & c_{12} & c_{13} & \cdots & c_{1n} \\ 0 & & & & \\ \vdots & & & B_{n-1} & \\ 0 & & & & \end{bmatrix} \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & & & \\ \vdots & & V_{n-1} & \\ 0 & & & \end{bmatrix} \\ &= \begin{bmatrix} \lambda_1 & b_{12} & \cdots & b_{1n} \\ & \lambda_2 & \cdots & b_{2n} \\ & & \ddots & \vdots \\ & & & \lambda_n \end{bmatrix} \end{aligned}$$

□

Teorema 1.12. Si $A \in \mathbb{C}^{n\times n}$ es Hermitiana, entonces existe una matriz unitaria U tal que

$$U^*AU = diag(\lambda_1, \ldots, \lambda_n),$$

siendo λ_i, $i = 1, \ldots, n$ los valores propios de A. Además, dichos valores propios son reales y los vectores propios de A forman una base ortonormal de $\mathbb{C}^n$.

Demostración.

Por el Teorema de Shur, existe U unitaria tal que

$$U^*AU = \begin{bmatrix} \lambda_1 & b_{12} & b_{13} & \cdots & b_{1n} \\ & \lambda_2 & b_{23} & \cdots & b_{2n} \\ & & \lambda_3 & \cdots & b_{3n} \\ & & & \ddots & \vdots \\ & & & & \lambda_n \end{bmatrix} =: B,$$

La hipótesis A Hermitiana implica que la matriz triangular B satisface

$$B^* = U^*A^*U^{**} = U^*AU = B,$$

de donde se concluye que los valores propios de A son reales y además que la matriz B es diagonal. Así,

$$U^*AU = diag(\lambda_1, \ldots, \lambda_n). \tag{71}$$

De esta última relación, se obtiene que

$$AU = Udiag(\lambda_1, \ldots, \lambda_n).$$

Luego, para cada $j = 1, \ldots, n$

$$AU^{(j)} = \lambda_j U^{(j)},$$

donde $U^{(j)}$ es la j-ésima columna de U. Así, $U^{(1)}, \ldots, U^{(n)}$ son los vectores propios asociados a los valores propios $\lambda_1, \ldots, \lambda_n$, respectivamente. Finalmente, $U^{(1)}, \ldots, U^{(n)}$ forman una base ortonormal de $\mathbb{C}^n$ pues U es unitaria. □

El resultado anterior indica que toda matriz Hermitiana es unitariamente similar a una matriz diagonal. Esto vale también para matrices normales.

Teorema 1.13. Si A es una matriz normal, existe una matriz unitaria U tal que la matriz U^*AU es diagonal.

Demostración.

Por el teorema de Schur, existe una matriz unitaria U tal que

$$U^*AU = \begin{bmatrix} \lambda_1 & b_{12} & b_{13} & \cdots & b_{1n} \\ & \lambda_2 & b_{23} & \cdots & b_{2n} \\ & & \lambda_3 & \cdots & b_{3n} \\ & & & \ddots & \vdots \\ & & & & \lambda_n \end{bmatrix} =: B,$$

o bien,

$$A = U \begin{bmatrix} \lambda_1 & b_{12} & b_{13} & \cdots & b_{1n} \\ & \lambda_2 & b_{23} & \cdots & b_{2n} \\ & & \lambda_3 & \cdots & b_{3n} \\ & & & \ddots & \vdots \\ & & & & \lambda_n \end{bmatrix} U^* = UBU^*$$

Por ser U unitaria, se tiene que

$$A^* = U \begin{bmatrix} \overline{\lambda}_1 & & & & \\ \overline{b}_{12} & \overline{\lambda}_2 & & & \\ \overline{b}_{13} & \overline{b}_{23} & \overline{\lambda}_3 & & \\ \vdots & \vdots & \vdots & \ddots & \\ \overline{b}_{1n} & \overline{b}_{2n} & \overline{b}_{3n} & & \overline{\lambda}_n \end{bmatrix} U^* = UB^*U^*$$

Luego, $AA^* = (UBU^*)(UB^*U^*) = UBB^*U^* = A^*A = UB^*BU^*$, y en consecuencia,

$$BB^* = B^*B \tag{72}$$

Comparando el primer elemento diagonal de la matriz del lado derecho de (72) con el correspondiente elemento de la matriz del lado derecho, resulta

$$|\lambda_1|^2 = |\lambda_1|^2 + \sum_{j=2}^{n} |b_{1j}|^2,$$

lo cual muestra que los elementos de la primera columna de B son ceros excepto el elemento de la diagonal principal. El mismo argumento se usa para la segunda fila teniendo en cuanta lo que ya se conoce de la primera y se continúa de esta manera hasta llegar a la ultima fila para obtener que $b_{ij} = 0$ para $i \neq j$. □

Observación 1.5. A partir de la última parte de la prueba del teorema anterior, se deduce que si una matriz normal es triangular, entonces es una matriz diagonal.

En este aparte se presenta una expresión para el cálculo de la norma matricial inducida por la norma vectorial Euclidiana, la cual recibe el nombre de *norma espectral*.

Teorema 1.14. Sea $A \in \mathbb{C}^{n\times n}$ y denotemos por σ_1 el mayor de los valores singulares de A, esto es, $\sigma_1 := \sqrt{\rho(A^*A)}$. Entonces,

$$||A||_2 = \sigma_1. \tag{73}$$

En particular, si A es Hermitiana,

$$||A||_2 = \rho(A),$$

y si A es unitaria, $||A||_2 = 1$.

Demostración.

Sean $\lambda_j(A^*A)$, $j = 1, \ldots, n$ los valores propios de A^*A, los cuales son no negativos como ya se ha visto. Tómese un vector cualquiera $\boldsymbol{v} \in \mathbb{C}^n$. Por el colorario (1.12), los vectores propios de A^*A, los cuales denotamos por $\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_n$, forman una base ortonormal de $\mathbb{C}^n$. Luego, $\boldsymbol{v}$ puede escribirse como

$$\boldsymbol{v} = \sum_{j=1}^{n} \alpha_j \boldsymbol{u}_j, \qquad \alpha_j \in \mathbb{C}.$$

De donde,

$$||\boldsymbol{v}||_2^2 = \langle \boldsymbol{v}, \boldsymbol{v}\rangle = \left\langle \sum_{j=1}^{n} \alpha_j \boldsymbol{u}_j, \sum_{m=1}^{n} \alpha_m \boldsymbol{u}_m \right\rangle = \sum_{j=1}^{n}\sum_{m=1}^{n} \alpha_j \overline{\alpha}_m \langle \boldsymbol{u}_j, \boldsymbol{u}_m\rangle = \sum_{j=1}^{n} |\alpha_j|^2.$$

Luego,

$$\begin{aligned} ||A\boldsymbol{v}||_2^2 = \langle A\boldsymbol{v}, A\boldsymbol{v}\rangle &= \langle \boldsymbol{v}, A^*A\boldsymbol{v}\rangle \\ &= \left\langle \sum_{j=1}^{n} \alpha_j \boldsymbol{u}_j, \sum_{m=1}^{n} \lambda_m(A^*A)\alpha_m \boldsymbol{u}_m \right\rangle \\ &= \sum_{j=1}^{n} \lambda_j(A^*A)|\alpha_j|^2 \end{aligned}$$

$$\leq \rho(A^*A)\sum_{j=1}^{n}|\alpha_j|^2 = \rho(A^*A)||\boldsymbol{v}||_2^2.$$

En consecuencia,

$$||A||_2 \leq \sqrt{\rho(A^*A)}.$$

Para probar la otra desigualdad, sea j_0 el índice para el cual $\rho(A^*A) = \lambda_{j_0}(A^*A) \geq 0$. Entonces,

$$\begin{aligned}
||A||_2^2 &= \left(\sup_{||\boldsymbol{v}||_2=1} ||A\boldsymbol{v}||_2\right)^2 \\
&\geq ||A\boldsymbol{u}_{j_0}||_2^2 \\
&= \langle A\boldsymbol{u}_{j_0}, A\boldsymbol{u}_{j_0}\rangle \\
&= \langle \boldsymbol{u}_{j_0}, A^*A\boldsymbol{u}_{j_0}\rangle \\
&= \langle \boldsymbol{u}_{j_0}, \lambda_{j_0}(A^*A)\boldsymbol{u}_{j_0}\rangle \\
&= \lambda_{j_0}(A^*A) = \rho(A^*A).
\end{aligned}$$

Así,

$$||A||_2 \geq \sqrt{\rho(A^*A)}.$$

Si A es Hermitiana, se tiene que

$$||A||_2 = \sqrt{\rho(A^*A)} = \sqrt{\rho(A^2)} = \sqrt{(\rho(A))^2} = \rho(A).$$

Finalmente, si A es unitaria,

$$||A\boldsymbol{v}||_2^2 = \langle A\boldsymbol{v}, A\boldsymbol{v}\rangle = \langle \boldsymbol{v}, A^*A\boldsymbol{v}\rangle = \langle \boldsymbol{v}, \boldsymbol{v}\rangle = ||\boldsymbol{v}||_2^2,$$

de donde, $||A||_2 = 1$. □

El siguiente teorema se usará más adelante para probar la convergencia de algunos métodos iterativos.

Teorema 1.15. Para cada norma vectorial en $\mathbb{C}^n$ y cada matriz A de tamaño $n \times n$ se cumple que

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

donde la norma matricial es la inducida por la norma vectorial dada. También se cumple que para cada matriz A y cada $\epsilon > 0$, existe una norma en $\mathbb{C}^n$ tal que

$$||A|| \leq \rho(A) + \epsilon.$$

Demostración.

Para probar la primera parte del teorema, sea λ un valor propio cualquiera de A con vector propio asociado $\boldsymbol{v}$ tal que $||\boldsymbol{v}|| = 1$. Entonces,

$$||A|| = \sup_{||\boldsymbol{v}||=1} ||A\boldsymbol{u}|| \geq ||A\boldsymbol{v}|| = ||\lambda\boldsymbol{v}|| = |\lambda|.$$

Luego, $\rho(A) \leq ||A||$.

Para la segunda parte, sean A una matriz $n \times n$ y $\epsilon > 0$ cualesquiera. Por el Teorema de descomposición de Shur, existe una matriz unitaria U tal que

$$B := U^*AU = \begin{bmatrix} \lambda_1 & b_{12} & b_{13} & \cdots & b_{1n} \\ & \lambda_2 & b_{23} & \cdots & b_{2n} \\ & & \lambda_3 & \cdots & b_{3n} \\ & & & \ddots & \vdots \\ & & & & \lambda_n \end{bmatrix},$$

donde $b_{ii} = \lambda_i$, $i = 1, 2, \ldots, n$ son los valores propios de A. Defina

$$\gamma := \max_{1 \leq i \leq j \leq n} |b_{ij}|$$

y

$$\delta := \min \left\{ 1, \frac{\epsilon}{\gamma(n-1)} \right\}.$$

Considere la matriz diagonal $R := diag(1, \delta, \delta^2, ..., \delta^{n-1})$ cuya inversa es $R^{-1} := diag(1, \delta^{-1}, \delta^{-2}, ..., \delta^{-(n-1)})$. Definiendo $S := R^{-1}BR$, se tiene de manera directa que

$$S = \begin{bmatrix} \lambda_1 & \delta b_{12} & \delta^2 b_{13} & \cdots & \delta^{n-1} b_{1n} \\ & \lambda_2 & \delta b_{23} & \cdots & \delta^{n-2} b_{2n} \\ & & \lambda_3 & \cdots & \delta^{n-3} b_{3n} \\ & & & \ddots & \vdots \\ & & & & \lambda_n \end{bmatrix}$$

Teniendo en cuenta que $\delta \leq 1$ y la definicion de γ y δ, se obtiene

$$\begin{aligned} ||S||_\infty &= \max \left\{ \sum_{j=1}^{n} |s_{1j}|, \sum_{j=1}^{n} |s_{2j}|, \ldots, \sum_{j=1}^{n} |s_{nj}| \right\} \\ &= \max \left\{ |\lambda_1| + \sum_{j=2}^{n} \delta^{j-1} |b_{1j}|, |\lambda_2| + \sum_{j=3}^{n} \delta^{j-2} |b_{2j}|, \ldots, |\lambda_n| \right\} \end{aligned}$$

$$\leq \max_{i=1,\ldots,n} |\lambda_i| + (n-1)\delta\gamma$$
$$\leq \rho(A) + \epsilon.$$

Ahora se define una norma vectorial en $\mathbb{C}^n$ por $||x|| := ||V^{-1}x||_\infty$, donde $V := UR$. Luego, usando la relación $SV^{-1} = R^{-1}U^*A = V^{-1}A$, se obtiene que para todo $x \in \mathbb{C}^n$

$$||A\boldsymbol{x}|| = ||V^{-1}A\boldsymbol{x}||_\infty = ||SV^{-1}\boldsymbol{x}||_\infty \leq ||S||_\infty ||V^{-1}\boldsymbol{x}||_\infty = ||S||_\infty ||\boldsymbol{x}||.$$

Por lo tanto,

$$||A|| \leq ||S||_\infty \leq \rho(A) + \epsilon,$$

lo cual completa la prueba. □

2. Número de condición de una matriz y sistemas mal condicionados

Ejemplo 2.1. Considere el sistema lineal

$$\begin{aligned} x_1 + 1.001x_2 &= -0.999 \\ 0.999x_1 + x_2 &= -0.998 \end{aligned}$$

cuya solución exacta es $x_1 = -2$, $x_2 = 1$. Si se realiza una pequeña perturbación del sistema cambiando la primera componente del vector del lado derecho por $b_1 = -0.998$, la solución del nuevo sistema cambia dramáticamente pues resulta $x_1 = 998$, $x_2 = -998$. Por lo tanto, el sistema considerado es muy sensible a pequeñas perturbaciones y tal sensibilidad no se debe a procedimiento numérico alguno, sino que es intrínseca del sistema. Ahora, al buscar las soluciones del sistema usando un algoritmo implementado en una máquina, se producen errores de redondeo, los cuales en la práctica pueden producir perturbaciones del sistema lineal y, si este es muy sensible (como en el ejemplo) terminamos obteniendo un resultado que no corresponde a la solución correcta. Por lo tanto, se hace necesario estudiar como se comporta un sistema lineal cuando se realizan pequeñas perturbaciones.

Decimos que un sistema de ecuaciones lineales está *mal condicionado* si pequeñas perturbaciones en el sistema producen cambios relativamente grandes en la solución exacta. En caso contrario, se dice que el sistema está *bien condicionado*.

Con el objetivo de medir cuán bien o mal condicionado está un sistema lineal, se presenta la definición de número de condición de una matriz.

Definición 2.1. El *número de condición* de una matriz A con respecto a una norma matricial inducida $||\cdot||$, se define como

$$\kappa(A) = \begin{cases} ||A||\,||A^{-1}||, & \text{si } A \text{ es no singular} \\ \infty, & \text{si } A \text{ es singular.} \end{cases}.$$

Nótese que el número de condición depende de la norma elegida. Cuando $||\cdot||_p$ sea la norma matricial inducida por una p norma vectorial, se usará la notación $\kappa_p(A)$, es decir, $\kappa_p(A) = ||A||_p\,||A^{-1}||_p$.

Es claro que el número de condición de una matriz no singular es igual al número de condición de su inversa y que el número de condición de la matriz

identidad es 1. Además, si A es no singular, entonces

$$1 \leq ||I_n|| = ||AA^{-1}|| \leq ||A||\,||A^{-1}|| = \kappa(A).$$

Así, el número de condición de una matriz cualquiera es mayor o igual que uno.

Para cualquier matriz A y α un escalar arbitrario no nulo, se tiene que $\kappa(\alpha A) = \kappa(A)$.

Volviendo al ejemplo 2.1, el número de condición de la matriz de coeficientes A con respecto a la norma 1 es

$$\kappa_1(A) = ||A||_1\,||A^{-1}||_1 \approx 4 \times 10^6.$$

Esta matriz tiene un número de condición grande y en este punto, tal vez se habrá percatado el lector de que la matriz está 'cerca' de la matriz singular

$$B = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}.$$

En general, el número de condición de una matriz proporciona una medida de cuán cerca (relativamente) está dicha matriz de ser singular. Más formalmente, se tiene el siguiente resultado (Gastinel) para el cual se presenta parte de la demostración debido a que uno de los argumentos usa un importante resultado del análisis funcional conocido como el Teorema de Hanh-Banach (ver también Corolario 5.5.15 de la referencia [3]).

Teorema 2.1. Sea A una matriz $n \times n$ no singular y $||\cdot||$ una norma matricial inducida cualquiera. Entonces

$$\frac{1}{\kappa(A)} = \text{mín}\left\{ \frac{||A - B||}{||A||} : \ B \text{ es singular} \right\} \tag{1}$$

Demostración.

Sea B una matriz singular cualquiera $n \times n$ y sea $\boldsymbol{x}$ un vector no nulo con n componentes tal que $B\boldsymbol{x} = \mathbf{0}$. Por la consistencia de la norma matricial inducida se tiene que

$$||\boldsymbol{x}|| = ||A^{-1}(A - B)\boldsymbol{x}| \leq ||A^{-1}||\,||A - B||\,||\boldsymbol{x}||.$$

Así, $1 \leq ||A^{-1}||\,||A - B||$. En consecuencia,

$$\frac{1}{\kappa(A)} = \frac{1}{||A||\,||A^{-1}||} \leq \frac{||A^{-1}||\,||A - B||}{||A||\,||A^{-1}||} = \frac{||A - B||}{||A||} \qquad \forall\, B \text{ singular.}$$

Luego,

$$\frac{1}{\kappa(A)} \leq \text{mín}\left\{\frac{||A-B||}{||A||} : \ B \text{ es singular}\right\}.$$

Para tener la prueba completa, basta probar que existe una matriz singular $\widetilde{B}$ tal que

$$\frac{1}{\kappa(A)} = \frac{||A-\widetilde{B}||}{||A||}. \tag{2}$$

En efecto, sea $\widetilde{\boldsymbol{y}}$ tal que

$$||A^{-1}|| = \frac{||A^{-1}\widetilde{\boldsymbol{y}}||}{||\widetilde{\boldsymbol{y}}||}.$$

Defina

$$\boldsymbol{y} := \frac{1}{||\widetilde{\boldsymbol{y}}||\,||A^{-1}||}\widetilde{\boldsymbol{y}} \qquad \text{y} \qquad \boldsymbol{x} := A^{-1}\boldsymbol{y}.$$

Note que $||\boldsymbol{x}|| = 1$. El teorema de Hahn-Banach garantiza la existencia de un vector $\boldsymbol{z}$ tal que

$$\boldsymbol{z}^T\boldsymbol{x} = 1 \qquad \text{y} \qquad \boldsymbol{z}^T\boldsymbol{u} \leq 1$$

para todo vector $\boldsymbol{u}$ de norma (la que induce la norma matricial) uno. Observe que si $\boldsymbol{u}$ tiene norma uno, entonces $-\boldsymbol{u}$ también es unitario, en consecuencia, $|\boldsymbol{z}^T\boldsymbol{u}| \leq 1$. La matriz $\widetilde{B}$ buscada se define así

$$\widetilde{B} := A - \boldsymbol{y}\boldsymbol{z}^T$$

Como $\widetilde{B}\boldsymbol{x} = A\boldsymbol{x} - \boldsymbol{y}\boldsymbol{z}^T\boldsymbol{x} = \boldsymbol{y} - \boldsymbol{y} = \boldsymbol{0}$, se deduce que la matriz $\widetilde{B}$ es singular, pues en caso contrario se tendría que $\boldsymbol{x} = \boldsymbol{0}$, lo cual contradice que $\boldsymbol{x}$ sea unitario. Luego, usando la definicíon del vector $\boldsymbol{y}$ y la desigualdad $|\boldsymbol{z}^T\boldsymbol{u}| \leq 1$, se obtiene

$$||A-\widetilde{B}|| = \sup_{||\boldsymbol{u}||=1} ||(A-\widetilde{B})\boldsymbol{u}|| = \sup_{||\boldsymbol{u}||=1} ||\boldsymbol{y}\boldsymbol{z}^T\boldsymbol{u}|| = \sup_{||\boldsymbol{u}||=1} ||\boldsymbol{y}||\,|\boldsymbol{z}^T\boldsymbol{u}| \leq \frac{1}{||A^{-1}||}.$$

Así,

$$\frac{||A-\widetilde{B}||}{||A||} \leq \frac{1}{\kappa(A)}.$$

De esta desigualdad y la primera parte de la prueba, resulta (2). □

Hay que aclarar que, a pesar de que las matrices singulares se caracterizan por su determinante nulo, el determinante no es un buen indicador de cuan cerca está dicha matriz de ser singular en el sentido de que un determinante muy grande o muy pequeño no permite concluir algo sobre qué tan

cerca, según el teorema previo, está la matriz de ser singular. Por ejemplo, el determinante $\det(\alpha I_n) = \alpha^n$ se puede hacer arbitrariamente pequeño tomando $0 < |\alpha| < 1$ y sin embargo, $\kappa(\alpha I_n) = 1$, por ende, αI_n es una matriz perfectamente condicionada. Por otra parte, podemos tener matrices con determinante grande pero, mal condicionadas. Por ejemplo, para $< 0\epsilon < 1$,sea $A = diag(1, 1/\epsilon)$. Como $\det(A) = 1/\epsilon$, este puede hacerse tan grande como se quiera tomando ϵ lo suficientemente pequeño pero, para $p = 1, 2, \infty$, $\kappa_p(A) = 1/\epsilon$. Nótese que la matriz A también es simétrica y definida positiva, por lo tanto, esta propiedad no implica que, en general, una matriz esté bien condicionada. Este hecho puede apreciarse mejor teniendo en cuenta que, si $A \in \mathbb{R}^{n\times n}$ es simétrica y definida positiva con valores propios $0 < \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$, entonces

$$\kappa_2(A) = ||A||_2 \, ||A^{-1}||_2 = \frac{\lambda_n}{\lambda_1} = \rho(A)\rho(A^{-1}).$$

El siguiente teorema proporciona una medida del cambio relativo en la solución de un sistema cuando se perturban datos.

Teorema 2.2. Sean $A \in \mathbb{R}^{n\times n}$ una matriz no singular, $\boldsymbol{b}\,, \Delta\boldsymbol{b} \in \mathbb{R}^n$ y $\boldsymbol{x}$, $\boldsymbol{x} + \Delta\boldsymbol{x}$ soluciones de los sistemas lineales

$$\begin{aligned} A\boldsymbol{x} &= \boldsymbol{b} \\ A(\boldsymbol{x} + \Delta\boldsymbol{x}) &= \boldsymbol{b} + \Delta\boldsymbol{b}, \end{aligned} \tag{3}$$

respectivamente. Entonces, para $\boldsymbol{b} \neq \boldsymbol{0}$,

$$\frac{1}{\kappa(A)} \frac{||\Delta\boldsymbol{b}||}{||\boldsymbol{b}||} \leq \frac{||\Delta\boldsymbol{x}||}{||\boldsymbol{x}||} \leq \kappa(A) \frac{||\Delta\boldsymbol{b}||}{||\boldsymbol{b}||}, \tag{4}$$

donde $||\cdot||$ es una norma matricial inducida. Además, la igualdad se alcanza para alguna elección de $\boldsymbol{b}$ y $\Delta\boldsymbol{b}$.

Demostración.

Probaremos inicialmente la segunda desigualdad. De las ecuaciones (3) y la invertibilidad de A, se tiene que $\Delta\boldsymbol{x} = A^{-1}\Delta\boldsymbol{b}$. Luego, teniendo en cuenta que la norma inducida es compatible, resulta

$$\begin{aligned} \frac{||\Delta\boldsymbol{x}||}{||\boldsymbol{x}||} = \frac{||A^{-1}\Delta\boldsymbol{b}||}{||\boldsymbol{x}||} \leq \frac{||A^{-1}||\,||\Delta\boldsymbol{b}||}{||\boldsymbol{x}||} &= \frac{||A^{-1}||\,||\Delta\boldsymbol{b}||\,||A\boldsymbol{x}||}{||\boldsymbol{b}||\,||\boldsymbol{x}||} \\ &\leq \frac{||A^{-1}||\,||\Delta\boldsymbol{b}||\,||A||}{||\boldsymbol{b}||} = \kappa(A)\frac{||\Delta\boldsymbol{b}||}{||\boldsymbol{b}||}. \end{aligned}$$

Para probar que la igualdad se alcanza, se tiene en cuenta que, por la definición de la norma matricial inducida, existen $\widetilde{\boldsymbol{x}}$ y $\widetilde{\Delta \boldsymbol{b}}$ tales que

$$||A^{-1}|| = \frac{||A^{-1}\widetilde{\Delta \boldsymbol{b}}||}{||\widetilde{\Delta \boldsymbol{b}}||} \qquad \text{y} \qquad ||A|| = \frac{||A\widetilde{\boldsymbol{x}}||}{||\widetilde{\boldsymbol{x}}||}. \tag{5}$$

Tomando $\widetilde{\boldsymbol{b}} = A\widetilde{\boldsymbol{x}}$ y procediendo como en la primera parte, pero cambiando las desigualdades por las igualdades dadas por (5), se obtiene que

$$\frac{||\Delta\widetilde{\boldsymbol{x}}||}{||\widetilde{\boldsymbol{x}}||} = \kappa(A)\frac{||\widetilde{\Delta \boldsymbol{b}}||}{||\widetilde{\boldsymbol{b}}||}.$$

Veamos ahora la primera desigualdad en (4).

$$||\boldsymbol{x}||\,||\Delta \boldsymbol{b}|| = ||A^{-1}\boldsymbol{b}||\,||A(\Delta \boldsymbol{x})|| \leq ||A^{-1}||\,||\boldsymbol{b}||\,||A||\,||\Delta \boldsymbol{x}|| = \kappa(A)||\boldsymbol{b}||\,||\Delta \boldsymbol{x}||,$$

lo cual da la desigualdad deseada. Que la igualdad se alcanza en este caso, se obtiene usando un argumento similar al que se usó para probar la igualdad en la primera parte de la demostración y se deja como ejercicio. □

Ejercicios 2.1.

1. Pruebe que $\kappa(AB) \leq \kappa(A)\kappa(B)$.

2. Halle una expresión que permita calcular de manera sencilla $\kappa_p(D)$ si D es una matriz diagonal.

3. Si U es una matriz unitaria, calcular $\kappa_2(A)$.

4. Pruebe que si $U \in \mathbb{C}^{n\times n}$ es una matriz unitaria y $A \in \mathbb{C}^{n\times n}$, entonces $\kappa_2(UA) = \kappa_2(A)$.

5. Sean $\boldsymbol{x}$ y $\boldsymbol{x} + \Delta\boldsymbol{x}$ las soluciones de los sistemas $A\boldsymbol{x} = \boldsymbol{b}$ y $A(\boldsymbol{x} + \Delta\boldsymbol{x}) = \boldsymbol{b} + \Delta\boldsymbol{b}$ respectivamente, donde

$$A = \begin{pmatrix} 3 & 0 & 0 \\ 0 & 5/4 & 3/4 \\ 0 & 3/4 & 5/4 \end{pmatrix}$$

Si se sabe que $\frac{||\Delta\boldsymbol{b}||_2}{||\boldsymbol{b}||_2} < 10^{-4}$, hallar una estimación del error relativo $\frac{||\Delta\boldsymbol{x}||_2}{||\boldsymbol{x}||_2}$

1. Métodos iterativos para resolver sistemas de ecuaciones

El número de operaciones computacionales en el método de eliminación es $n^3/3+\mathcal{O}(n^2)$, en consecuencia, al aumentar el tamaño del sistema, el número de operaciones crece dramaticamente, lo que hace necesario considerar métodos alternativos a los métodos directos, a saber, los métodos iterativos. Como también se verá, la precisión obtenida en un método iterativo convergente depende del número de iteraciones realizadas, por ende, se puede disminuir el costo (tiempo) computacional si se requiere menos precisión en un problema particular. Esto no es cierto para los métodos de eliminación. Algunos métodos iterativos para resolver sistemas lineales fueron originalmente propuestos por Gauss (1823), Liouville (1837) y Jacobi (1845).

En este capítulo se presentan algunos métodos iterativos para aproximar soluciones de un sistema de n ecuaciones con n incógnitas expresado en la forma

$$\begin{aligned} f_1(x_1, x_2, \ldots, x_n) &= 0 \\ f_2(x_1, x_2, \ldots, x_n) &= 0 \\ &\vdots \\ f_n(x_1, x_2, \ldots, x_n) &= 0 \end{aligned} \tag{6}$$

En forma compacta, el sistema (6) puede escribirse como

$$\boldsymbol{f}(\boldsymbol{x}) = \mathbf{0}, \tag{7}$$

donde $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)^{\mathrm{T}}$, $\boldsymbol{f}(\boldsymbol{x}) = (f_1(\boldsymbol{x}), \ldots, f_n(\boldsymbol{x}))^{\mathrm{T}}$ y $\mathbf{0}$ simboliza el vector nulo de $\mathbb{R}^n$. Un caso particular de (6) es el sistema lineal

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= b_2 \\ &\vdots \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n &= b_n \end{aligned} \tag{8}$$

o bien, en forma matricial $A\boldsymbol{x} = \boldsymbol{b}$, con $A = (a_{ij})$ y $\boldsymbol{b} = (b_1, b_2, \ldots, b_n)^{\mathrm{T}}$ En la notación de (6),

$$f_i(x_1, x_2, \ldots, x_n) = a_{i1}x_1 + a_{i2}x_2 + \cdots + a_{in}x_n - b_i, \qquad i = 1, \ldots, n.$$

En términos generales, la idea de los métodos iterativos para aproximar una solución $\boldsymbol{x}$ de (6) (más adelante nos ocuparemos de discutir condiciones de

existencia y unicidad del problema) es generar una sucesión $\{\boldsymbol{x}_k\}$ partiendo de un dato inicial $\boldsymbol{x}_0$ de tal forma que $\lim_{k\to\infty} \boldsymbol{x}_k = \boldsymbol{x}$. Una herramienta de gran utilidad, no solo desde el punto de vista teórico, sino también práctico es el teorema del punto fijo de Banach. Esto porque para introducir los métodos iterativos, se busca reformular el sistema como un problema de punto fijo, i.e., se considera un sistema de la forma

$$\boldsymbol{g}(\boldsymbol{x}) = \boldsymbol{x}, \tag{9}$$

Recordemos un teorema de punto fijo con su respectiva prueba.

Teorema 2.3 (Teorema del punto fijo de Banach)**.** Sean W un subconjunto cerrado de un espacio de Banach V, $T : W \to W$ una contracción con constante de contractilidad $\alpha \in [0, 1)$, i.e.,

$$\| T(\boldsymbol{u}) - T(\boldsymbol{v})|| \leq \alpha||\boldsymbol{u} - \boldsymbol{v}||, \qquad \forall \boldsymbol{u}, \boldsymbol{v} \in W.$$

Entonces

(i) La ecuación $T(\boldsymbol{x}) = \boldsymbol{x}$ tiene una solución única $\boldsymbol{x} \in W$, y la sucesión de aproximaciones

$$\boldsymbol{x}_{k+1} = T(\boldsymbol{x}_k), \qquad k = 0, 1, 2, \ldots,$$

con $\boldsymbol{x}_0 \in W$ arbitrario, converge a esta solución.

(ii) Se tienen las siguientes estimaciones del error

$$||\boldsymbol{x}_k - \boldsymbol{x}|| \leq \frac{\alpha^k}{1-\alpha}||\boldsymbol{x}_0 - \boldsymbol{x}_1|| \tag{10}$$

$$||\boldsymbol{x}_k - \boldsymbol{x}|| \leq \frac{\alpha}{1-\alpha}||\boldsymbol{x}_{k-1} - \boldsymbol{x}_k|| \tag{11}$$

$$||\boldsymbol{x}_k - \boldsymbol{x}|| \leq \alpha||\boldsymbol{x}_{k-1} - \boldsymbol{x}|| \tag{12}$$

Demostración.

Puesto que $T : W \to W$, la sucesión $\{\boldsymbol{x}_k\}$ está bien definida. Por otra parte, dado que T es contractiva se tiene que

$$||\boldsymbol{x}_{k+1} - \boldsymbol{x}_k|| = ||T(\boldsymbol{x}_k) - T(\boldsymbol{x}_{k-1})|| \leq \alpha||\boldsymbol{x}_k - \boldsymbol{x}_{k-1}|| \leq \cdots \leq \alpha^k||\boldsymbol{x}_1 - \boldsymbol{x}_0||.$$

Ahora usamos esta última desigualdad para probar que la sucesión $\{\boldsymbol{x}_k\}$ es de Cauchy. En efecto, para cualesquiera $m \geq k \geq 1$,

$$||\boldsymbol{x}_k - \boldsymbol{x}_m|| \leq ||\boldsymbol{x}_k - \boldsymbol{x}_{k+1}|| + ||\boldsymbol{x}_{k+1} - \boldsymbol{x}_{k+2}|| + \cdots + ||\boldsymbol{x}_{m-1} - \boldsymbol{x}_m||$$

$$\begin{aligned}
&\leq \alpha^k||\boldsymbol{x}_1 - \boldsymbol{x}_0|| + \alpha^{k+1}||\boldsymbol{x}_1 - \boldsymbol{x}_0|| + \cdots + \alpha^{m-1}||\boldsymbol{x}_1 - \boldsymbol{x}_0|| \\
&= \alpha^k(1 + \alpha + \alpha^2 + \cdots + \alpha^{m-k-1})||\boldsymbol{x}_1 - \boldsymbol{x}_0|| \\
&= \alpha^k \frac{1-\alpha^{m-k}}{1-\alpha}||\boldsymbol{x}_1 - \boldsymbol{x}_0|| \\
&\leq \frac{\alpha^k}{1-\alpha}||\boldsymbol{x}_1 - \boldsymbol{x}_0|| \qquad (13)
\end{aligned}$$

Puesto que $\alpha \in [0,1)$, $||\boldsymbol{x}_k - \boldsymbol{x}_m|| \to 0$ cuando $m, k \to \infty$. Así, $\{\boldsymbol{x}_k\}$ es una sucesión de Cauchy, y como W es subconjunto cerrado del espacio de Banach W, se sigue que $\{\boldsymbol{x}_k\}$ tiene un límite $\boldsymbol{x} \in W$. Teniendo en cuenta que T es continua (por ser contractiva), resulta

$$\boldsymbol{x} = \lim_{k\to\infty} \boldsymbol{x}_{k+1} = \lim_{k\to\infty} T(\boldsymbol{x}_k) = T\left(\lim_{k\to\infty} \boldsymbol{x}_k\right) = T(\boldsymbol{x}),$$

esto es, $\boldsymbol{x}$ es un punto fijo de T. Si $\boldsymbol{y}$ fuese otro punto fijo de T, entonces

$$||\boldsymbol{x} - \boldsymbol{y}|| = ||T(\boldsymbol{x}) - T(\boldsymbol{y})|| \leq \alpha||\boldsymbol{x} - \boldsymbol{y}||,$$

lo cual implica que $||\boldsymbol{x} - \boldsymbol{y}|| = 0$ (pues $\alpha \in [0,1)$) y por ende, $\boldsymbol{x} = \boldsymbol{y}$, obteniéndose de esta manera la unicidad del punto fijo. Veamos finalmente las estimaciones del error. La desigualdad (10), se obtiene tomando $m \to \infty$ en (13). Por otra parte,

$$||\boldsymbol{x}_k - \boldsymbol{x}|| = ||T(\boldsymbol{x}_{k-1}) - T(\boldsymbol{x})|| \leq \alpha||\boldsymbol{x}_{k-1} - \boldsymbol{x}||,$$

esto es la desigualdad (12). De la estimación (12) junto con la desigualdad triangular, se obtiene

$$||\boldsymbol{x}_k - \boldsymbol{x}|| \leq \alpha||\boldsymbol{x}_{k-1} - \boldsymbol{x}|| \leq \alpha(||\boldsymbol{x}_{k-1} - \boldsymbol{x}_k|| + ||\boldsymbol{x}_k - \boldsymbol{x}||),$$

lo cual implica (11). □

El anterior teorema (y su demostración) sugieren el siguiente algoritmo para para aproximar la solución del problema (9).

Algoritmo (Algoritmo 1)**.** Sea $\boldsymbol{x}_0 = (x_{0,1}, x_{0,2}, \ldots, x_{0,n})$ una primera aproximación a la solución del sistema $\boldsymbol{g}(\boldsymbol{x}) = \boldsymbol{x}$. Se genera recursivamente la sucesión $\{\boldsymbol{x}_k\}$ por medio de la relación

$$\boldsymbol{x}_{k+1} = g(\boldsymbol{x}_k), \qquad k = 0, 1, 2, \ldots$$

En lo que sigue nos ocupamos del caso de sistemas lineales de n ecuaciones y n incógnitas en la forma $g(\boldsymbol{x}) = \boldsymbol{x}$. En notación matricial, tal sistema lineal puede escribirse como

$$\boldsymbol{x} = B\boldsymbol{x} + \boldsymbol{c}, \tag{14}$$

donde $B = (b_{ij})_{n\times n}$ y $\boldsymbol{c} = (c_i)_{n\times 1}$. De manera natural, a partir del Algoritmo 1 y de la forma (14), se obtiene el siguiente método iterativo para resolver el sistema $A\boldsymbol{x} = \boldsymbol{b}$

$$\boldsymbol{x}_{k+1} = B\boldsymbol{x}_k + \boldsymbol{c}, \qquad k = 0, 1, 2, \ldots \tag{15}$$

La matriz B recibe el nombre de *matriz de iteración*. Se dice que el método es *estacionario* cuando B y $\boldsymbol{c}$ son constantes en todas las iteraciones.

Observación 2.1. Si $||\cdot||$ es una norma matricial inducida por alguna norma vectorial y se toma $T(\boldsymbol{x}) = B\boldsymbol{x} + \boldsymbol{c}$, entonces

$$||T(\boldsymbol{x}) - T(\boldsymbol{y})|| = ||B(\boldsymbol{x} - \boldsymbol{y})|| \leq ||B||||\boldsymbol{x} - \boldsymbol{y}||.$$

Luego, el método iterativo (15) converge si $||B|| < 1$.

Teorema 2.4. Sea B una matriz cuadrada. Entonces

$$\lim_{k\to\infty} B^k = \mathbf{0} \quad \text{si y solo si} \quad \rho(B) < 1. \tag{16}$$

Más aun, la serie (geométrica) $\sum_{k=0}^{\infty} B^k$ es convergente, si y solo si $\rho(B) < 1$. En tal caso, la matriz $I - B$ es invertible y

$$(I - B)^{-1} = \sum_{k=0}^{\infty} B^k.$$

Demostración.

Veamos primero la prueba de (16). Suponga que $\lim_{k\to\infty} B^k = 0$, y sea λ un valor propio cualquiera de B con vector propio (a derecha) asociado $\boldsymbol{v} \neq \mathbf{0}$, esto es, $B\boldsymbol{v} = \lambda\boldsymbol{v}$. Esto implica que $B^k\boldsymbol{v} = \lambda^k\boldsymbol{v}$ para $k \in \mathbb{Z}^+$. Luego, de la hipótesis $B^k \to 0$ y teniendo en cuenta que $\boldsymbol{v} \neq \mathbf{0}$, resulta $\lim_{k\to\infty} \lambda^k = 0$ y así, $|\lambda| < 1$. Como λ se tomó arbitrario, se deduce que $\rho(B) < 1$.

Para la recíproca, suponga que $\rho(B) < 1$. Entonces existe $\epsilon > 0$ tal que $\rho(B) < 1 - \epsilon$. Por el Teorema (1.15), existe una norma matricial inducida $||\cdot||$ tal que $||B|| \leq \rho(B) + \epsilon < (1 - \epsilon) + \epsilon$. Dado que $0 \leq ||B^k|| \leq ||B||^k$ y $||B|| < 1$, se sigue que $||B^k|| \to 0$, i.e., $\lim_{k\to\infty} B^k = 0$.

Para la segunda parte, sean $\lambda_1, \lambda_2, \ldots, \lambda_n$ los valores propios de B, entonces $1 - \lambda_i$, $i = 1, 2, \ldots, n$ son los valores propios $I - B$. Luego, si $\rho(B) < 1$,

$\det(I-B)=\prod_{i=1}^{n}(1-\lambda_i)\neq 0$. En consecuencia, $I-B$ es invertible. Por otra parte, al tomar el límite en la identidad

$$(I-B)(I+B+B^2+\cdots+B^k)=(I-B^{k+1}),$$

y teniendo en cuenta que $\lim_{k\to\infty}B^k=\mathbf{0}$, se obtiene

$$(I-B)\sum_{k=0}^{\infty}B^k=I-\lim_{k\to\infty}B^{k+1}=I.$$

Por lo tanto, $\sum_{k=0}^{\infty}B^k$ converge y

$$(I-B)^{-1}=\sum_{k=0}^{\infty}B^k.$$

Recíprocamente, si $\sum_{k=0}^{\infty}B^k$ converge, entonces $B^k\to\mathbf{0}$ y se concluye de la primera parte que $\rho(B)<1$. □

Teorema 2.5. Sea B una matriz cuadrada. La sucesión $\{\boldsymbol{x}_k\}$ definida por el método iterativo

$$\boldsymbol{x}_{k+1}=B\boldsymbol{x}_k+\boldsymbol{c}$$

converge para cada $\boldsymbol{c}$ y cada $\boldsymbol{x}_0$ a la solución única de $\boldsymbol{x}=B\boldsymbol{x}+\boldsymbol{c}$ si y solo si $\rho(B)<1$.

Demostración.

Suponga que $\rho(B)<1$, entonces existe $\epsilon>0$ tal que $\rho(B)+\epsilon<1$. Luego, se sigue del Teorema (1.15) que existe una norma vectorial en $\mathbb{C}^n$, la cual induce una norma matricial tal que $||B||<1$. Defina $T(\boldsymbol{x}):=B\boldsymbol{x}+\boldsymbol{c}$. Puesto que (ver observación (2.1))

$$||T(\boldsymbol{x})-T(\boldsymbol{y})||=||B(\boldsymbol{x}-\boldsymbol{y})||\leq||B||||\boldsymbol{x}-\boldsymbol{y}||,$$

se deduce que T es una contracción con constante de contractilidad $\alpha=||B||$. Luego, en virtud del Teorema del punto fijo de Banach, se concluye que la sucesión $\{\boldsymbol{x}_k\}$ con $\boldsymbol{x}_{k+1}=T(\boldsymbol{x}_k)=B\boldsymbol{x}_k+\boldsymbol{c}$ converge a la única solución de $\boldsymbol{x}=T(\boldsymbol{x})=B\boldsymbol{x}+\boldsymbol{c}$ y se tienen las estimaciones

$$||\boldsymbol{x}_k-\boldsymbol{x}||\leq\frac{||B||^k}{1-||B||}||\boldsymbol{x}_0-\boldsymbol{x}_1||$$

$$||\boldsymbol{x}_k-\boldsymbol{x}||\leq\frac{||B||}{1-||B||}||\boldsymbol{x}_{k-1}-\boldsymbol{x}_k||$$

$$||\boldsymbol{x}_k - \boldsymbol{x}|| \leq ||B||||\boldsymbol{x}_{k-1} - \boldsymbol{x}||.$$

Además, dado que en $\mathbb{C}^n$ todas las normas son equivalentes, se obtiene que la sucesión es convergente con cualquier norma de $\mathbb{C}^n$.

Para la recíproca, se procede por contradicción suponiendo que la sucesión de iteraciones $\{\boldsymbol{x}_k\}$ converge para cualesquiera $\boldsymbol{c}$ y $\boldsymbol{x}_0$, y que $\rho(B) \geq 1$. Esto último implica existe λ valor propio de B tal que $|\lambda| \geq 1$. Sea $\boldsymbol{v} \neq \boldsymbol{0}$ el vector propio asociado con λ, entonces tomando $\boldsymbol{x}_0 = \boldsymbol{c} = \boldsymbol{v}$ resulta

$$\boldsymbol{x}_{k+1} = \left(\sum_{m=0}^{k+1} \lambda^m\right) \boldsymbol{v}.$$

Puesto que $|\lambda| \geq 1$, $\boldsymbol{x}_{k+1} = \left(\sum_{m=0}^{k+1} \lambda^m\right) \boldsymbol{v}$ diverge, lo cual contradice la hipótesis. □

Definición 2.2. Se dice que el método iterativo (15) es *consistente* con el sistema $A\boldsymbol{x} = \boldsymbol{b}$ si $\boldsymbol{x} = B\boldsymbol{x} + \boldsymbol{c}$.

Esta definición junto con el teorema 2.5 nos da el siguiente resultado que nos permite determinar cuando converge el método iterativo (15) a la solución del sistema $A\boldsymbol{x} = \boldsymbol{b}$.

Teorema 2.6. Suponga que el método iterativo (15) es consistente con el sistema $A\boldsymbol{x} = \boldsymbol{b}$ con A no singular. Entonces, el método iterativo converje a la solución única del sistema si y solo si $\rho(B) < 1$.

Acto seguido se definen algunos tipos de matrices tales que, al fungir como matrices de iteración, se obtiene la convergencia de algunos métodos iterativos.

Definición 2.3. Se dice que la matriz $A = (a_{ij})$ es *estrictamente diagonal dominante por filas* si

$$\sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}| < |a_{ii}|, \qquad \forall i = 1, \ldots, n.$$

El siguiente resultado, llamado Teorema de Gershgorin, aparte proveer una estimación de la localización de los valores propios de una matriz, permite concluir la no singularidad de una matriz estrictamente diagonal dominante por filas.

Teorema 2.7 (Gershgorin)**.** Sea $A \in \mathbb{C}^{n\times n}$. Cualquier valor propio λ de $A = (a_{ij})$ está localizado en uno de los discos cerrados del plano complejo centrados en a_{ii} y de radio

$$r_i := \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}|.$$

En otras palabras, para todo $\lambda \in \sigma(A)$, existe $i \in \{1, \ldots, n\}$ tal que

$$|\lambda - a_{ii}| \leq r_i = \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}|.$$

Demostración.

Sea λ un valor propio de A y $\boldsymbol{v}$ un vector propio asociado con $||\boldsymbol{v}||_\infty = 1$. Tómese i como el índice para el cual $||\boldsymbol{v}||_\infty = |v_i|$. Entonces $|v_i| = 1$ y $|v_j| \leq 1$ para $j \neq i$. Puesto que $A\boldsymbol{v} = \lambda\boldsymbol{v}$, se tiene que $\sum_{j=1}^{n} a_{ij}v_j = \lambda v_i$, o bien

$$(\lambda - a_{ii})v_i = \sum_{\substack{j=1 \\ j\neq i}}^{n} a_{ij}v_j,$$

de donde

$$|\lambda - a_{ii}| \leq \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}||v_j| \leq \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}| =: r_i.$$

□

Corolario 2.1. Si $A \in \mathbb{C}^{n\times n}$ es estrictamente diagonal dominante por filas, entonces A es no singular.

Demostración.

Procedemos por contradicción suponiendo que A es singular, lo cual implica que $\det(A) = 0$. Puesto que el determinante de A es igual al producto de sus valores propios, al menos uno de ellos será igual a cero. Luego, por el Teorema de Gershgorin, existe i tal que

$$|0 - a_{ii}| \leq \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}|,$$

lo cual contradice la hipótesis que A sea estrictamente diagonal dominante por filas. □

Algunos métodos iterativos para resolver el sistema lineal $A\boldsymbol{x} = \boldsymbol{b}$ se diferencian por la forma en que se descompone la matriz de coeficientes como

$$A = D - E - F,$$

donde $D = diag(a_{11}, a_{22}, \ldots, a_{nn})$ es la matriz diagonal cuyas entradas son las diagonales de A, $E = (e_{ij})$ es la matriz triangular inferior con $e_{ij} = -a_{ij}$ si $i > j$, $e_{ij} = 0$ si $i \leq j$, y $F = (f_{ij})$ es la matriz triangular superior con $f_{ij} = -a_{ij}$ si $j > i$, $f_{ij} = 0$ si $j \leq i$. Se supondrá que las entradas diagonales de A son diferentes de cero y así la matriz D es invertible.

2. Método de Jacobi

En el método atribuido a Jacobi, el sistema $A\boldsymbol{x} = \boldsymbol{b}$ se transforma en la forma de punto fijo equivalente

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

resultando la fórmula de iteración

$$\boldsymbol{x}_{k+1} = D^{-1}(E + F)\boldsymbol{x}_k + D^{-1}\boldsymbol{b}, \qquad k = 0, 1, 2, \ldots,$$

con un dato inicial $\boldsymbol{x}_0$ cualquiera. Luego, la matriz de iteración para el método de Jacobi es

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

y $\boldsymbol{c} := D^{-1}\boldsymbol{b}$.

En términos de las componentes, un paso en el método de Jacobi queda así

$$x_{k+1,i} = \frac{1}{a_{ii}} \left(b_i - \sum_{\substack{j=1 \\ j \neq i}}^{n} a_{ij} x_{k,j} \right), \qquad i = 1, \ldots, n. \tag{17}$$

Teorema 2.8. Si A es estrictamente diagonal dominante por filas, el método de Jacobi converge (en cualquier norma de $\mathbb{C}^n$) para cada $\boldsymbol{b}$ a la única solución del sistema $A\boldsymbol{x} = \boldsymbol{b}$.

Demostración.

Obsérvese que la componente ij de la matriz de iteración $B_J = D^{-1}(E + F) = I - D^{-1}A$ del método de Jacobi está dada por

$$b_{ij} = \begin{cases} -\frac{a_{ij}}{a_{ii}}, & \text{if } i \neq j, \\ 0, & \text{if } i = j. \end{cases}$$

Debido a que A es estrictamente diagonal dominante por filas, se tiene que

$$||B_J||_\infty = \max_{1\leq i\leq n} \sum_{j=1}^{n} |b_{ij}| = \max_{1\leq i\leq n} \sum_{\substack{j=1\\ j\neq i}}^{n} \frac{|a_{ij}|}{|a_{ii}|} < 1.$$

Luego, en virtud del Teorema 1.15, $\rho(B_J) < 1$. Esto implica, por el Teorema 2.5, que el método es convergente. Como el vector $\boldsymbol{x}$ al cual converge el método verifica

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

o bien

$$D\boldsymbol{x} = (E+F)\boldsymbol{x} + \boldsymbol{b},$$

es decir, $A\boldsymbol{x} = \boldsymbol{b}$, lo cual da la consistencia del método. □

Observación 2.2. Puesto que

$$||B_J||_1 = \max_{1\leq j\leq n} \sum_{i=1}^{n} |b_{ij}| = \max_{1\leq j\leq n} \sum_{\substack{i=1\\ i\neq j}}^{n} \frac{|a_{ij}|}{|a_{ii}|},$$

el método de Jacobi también converge si

$$\max_{1\leq j\leq n} \sum_{\substack{i=1\\ i\neq j}}^{n} \frac{|a_{ij}|}{|a_{ii}|} < 1.$$

3. Método de Gauss-Seidel

En este método se usa la descomposición $A = D - E - F$, siendo E, D y F son las mismas matrices del método de Jacobi, para escribir el sistema $A\boldsymbol{x} = \boldsymbol{b}$ en la forma

$$(D-E)\boldsymbol{x} = F\boldsymbol{x} + \boldsymbol{b}$$

o bien,

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

Tenga en cuenta que $D-E$ es no singular puesto que es triangular inferior y sus entradas diagonales son las diagonal de A y se ha supuesto las entradas diagonales de A son no nulas. Por lo tanto, la matriz de iteración para el método de Gauss-Seidel es

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

con $\boldsymbol{c} = (D - E)^{-1}\boldsymbol{b}$.

En términos de las componentes, un paso del método de Gauss-Seidel se puede expresar como

$$x_{k+1,i} = \frac{1}{a_{ii}}\left(b_i - \sum_{j=1}^{i-1} a_{ij}x_{k+1,j} - \sum_{j=i+1}^{n} a_{ij}x_{k,j}\right), \qquad i = 1, \ldots, n. \tag{18}$$

Nótese que en el método de Gauss-Seidel, cada nueva componente obtenida de $\boldsymbol{x}_{k+1}$ es usada inmediatamente para calcular la próxima componente, es decir, para calcular la i-ésima componente de $\boldsymbol{x}_{k+1}$, se usan usan las componentes $x_{k+1,1}$, $x_{k+1,2}$,..., $x_{k+1,i-1}$. Esto es conveniente desde el punto de vista computacional porque se reducen los requerimientos de almacenamiento.

A continuación un resultado que da condiciones suficientes para la convergencia del método de Gauss-Seidel.

Teorema 2.9. Suponga que la matriz de coeficientes $A = (a_{ij})$ satisface el criterio de Sassenfeld

$$p := \max_{1 \le i \le n} p_i < 1,$$

donde los números p_i se definen recursivamente por

$$p_1 := \frac{1}{|a_{11}|}\sum_{j=2}^{n} |a_{1j}|, \qquad p_i := \frac{1}{|a_{ii}|}\sum_{j=1}^{i-1} |a_{ij}|p_j + \frac{1}{|a_{ii}|}\sum_{j=i+1}^{n} |a_{ij}|, \quad i = 2, \ldots, n.$$

Entonces el método de Gauss-Seidel converge (en cualquier norma de $\mathbb{C}^n$) para cada $\boldsymbol{b}$ a la única solución del sistema $A\boldsymbol{x} = \boldsymbol{b}$.

Demostración.

Considere la ecuación

$$(D - E)\boldsymbol{x} = F\boldsymbol{z}$$

para $\boldsymbol{z} \in \mathbb{C}^n$ con $||\boldsymbol{z}||_\infty = 1$, esto es,

$$x_i = -\frac{1}{a_{ii}}\sum_{j=1}^{i-1} a_{ij}x_j - \frac{1}{a_{ii}}\sum_{j=i+1}^{n} a_{ij}z_j, \qquad i = 1, \ldots, n.$$

esto implica por inducción que $|x_i| \le p_i$ para cada $j = 1, \ldots, n$, y en consecuencia, $||\boldsymbol{x}||_\infty \le p$. Luego,

$$||B_{GS}||_\infty = ||(D - E)^{-1}F||_\infty \le p < 1.$$

Por el Teorema 1.15 se obtiene que $\rho(B_{GS}) < 1$, de donde resulta la convergencia en virtud del Teorema 2.5.

Finalmente, como el vector $\boldsymbol{x}$ al cual converge el método verifica

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

o bien

$$(D-E)\boldsymbol{x} = F\boldsymbol{x} + \boldsymbol{b},$$

de donde resulta la consistencia del método. □

Note que si A es estrictamente diagonal dominante por filas, se satisface el criterio de Sassenfeld

Corolario 2.2. Si A es estrictamente diagonal dominante por filas entonces el método de Gauss-Seidel converge.

Ejemplo 2.2. Considere la matriz tridiagonal

$$A = \text{tridiag}(-1, 2, -1) = \begin{pmatrix} 2 & -1 & & & & \\ -1 & 2 & -1 & & & \\ & -1 & 2 & -1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & -1 & 2 & -1 \\ & & & & -1 & 2 \end{pmatrix}$$

Esta matriz no es estrictamente diagonal dominante por filas, sin embargo, satisface el criterio de Sassenfeld. En efecto, $p_1 = 1/2$, $p_n = p_{n-1}/2$ y para $i = 2, \ldots, n-1$ los únicos términos no nulos de la i-ésima fila son $a_{ii} = 2$, $a_{i(i-1)} = a_{i(i+1)} = -1$. Luego,

$$p_i = \frac{1}{|a_{ii}|}(|a_{i(i-1)}|p_{i-1} + |a_{i(i+1)}|) = \frac{1}{2}p_{i-1} + \frac{1}{2}, \qquad i = 2, \ldots, n-1.$$

Por lo anterior, se obtiene fácilmente por inducción que

$$p_i = 1 - \frac{1}{2^i}, \qquad i = 1, 2, \ldots, n-1 \qquad \text{y} \qquad p_n = \frac{1}{2} - \frac{1}{2^n}.$$

Por lo tanto, $p = \text{máx}_{1 \le i \le n}\, p_i < 1$, y así el método de Gauss-Seidel converge.

Teorema 2.10. Si $A \in \mathbb{C}^{n\times n}$ es Hermitiana y definida positiva, entonces el método de Gauss-Seidel es convergente.

Demostración.

Puesto que A es definida positiva, sus entradas diagonales son positivas, y así la matriz diagonal D (en la descomposición $A = D - E - F$) es real. Además, como A es Hermitiana, se tiene

$$A = D - E - F = D - E - E^*.$$

Se debe probar que el radio espectral de

$$B_{GS} = (D-E)^{-1}F = (D-E)^{-1}E^* = I - (D-E)^{-1}A$$

es menor que uno. Para este objetivo, tómese λ valor propio cualquiera de B_{GS} y sea $\boldsymbol{v} \neq \mathbf{0}$ un vector propio asociado a λ, es decir,

$$(I - (D-E)^{-1}A)\boldsymbol{v} = \lambda \boldsymbol{v},$$

o bien

$$A\boldsymbol{v} = (1-\lambda)(D-E)\boldsymbol{v}.$$

Como $A\boldsymbol{v} \neq \mathbf{0}$ (pues $\boldsymbol{v} \neq 0$ y A es invertible), se deduce que $\lambda \neq 1$. Luego multiplicando a la izquierda por $\boldsymbol{v}^*$, se obtiene

$$\frac{1}{1-\lambda} = \frac{\boldsymbol{v}^*(D-E)\boldsymbol{v}}{\boldsymbol{v}^*A\boldsymbol{v}}. \tag{19}$$

Ahora, tomando el complejo conjugado y teniendo en cuenta que A es Hermitiana, D es una matriz real, $\boldsymbol{v}^*A\boldsymbol{v} \in \mathbb{R}$, y $\overline{\boldsymbol{v}^*C\boldsymbol{v}} = \boldsymbol{v}^*C^*\boldsymbol{v}$ para cualquier matriz C, resulta

$$\frac{1}{1-\overline{\lambda}} = \frac{\overline{\boldsymbol{v}^*(D-E)\boldsymbol{v}}}{\boldsymbol{v}^*A\boldsymbol{v}} = \frac{\boldsymbol{v}^*(D-E)^*\boldsymbol{v}}{\boldsymbol{v}^*A\boldsymbol{v}} = \frac{\boldsymbol{v}^*(D-E^*)\boldsymbol{v}}{\boldsymbol{v}^*A\boldsymbol{v}}. \tag{20}$$

Sumando (19) y (20), y teniendo en cuenta que $A = D - E - E^*$, se obtiene

$$2Re\left(\frac{1}{1-\lambda}\right) = \frac{\boldsymbol{v}^*(D-E-E^*)\boldsymbol{v} + \boldsymbol{v}^*D\boldsymbol{v}}{\boldsymbol{v}^*A\boldsymbol{v}} = 1 + \frac{\boldsymbol{v}^*D\boldsymbol{v}}{\boldsymbol{v}^*A\boldsymbol{v}} > 1. \tag{21}$$

Esta última desigualdad es válida porque A y D son definidas positivas. Por otra parte, si $\lambda = \alpha + i\beta$, se tiene

$$\frac{1}{1-\lambda} = \frac{1}{1-\alpha-i\beta}\frac{1-\alpha+i\beta}{1-\alpha+i\beta} = \frac{1-\alpha+i\beta}{(1-\alpha)^2+\beta^2}.$$

Así,

$$Re\left(\frac{1}{1-\lambda}\right) = \frac{1-\alpha}{(1-\alpha)^2+\beta^2}.$$

Esta igualdad junto con (21) da

$$\frac{1-\alpha}{(1-\alpha)^2+\beta^2} > \frac{1}{2},$$

de donde $|\lambda| = \alpha^2 + \beta^2 < 1$. Puesto que el valor propio λ de B_{GS} se eligió arbitrario, se deduce que $\rho(B_{GS}) < 1$, y por lo tanto, el método de Gauss-Seidel converge. □

Ejercicio 2.1.

1. Para la solución del sistema lineal $A\mathbf{x} = \mathbf{b}$ con

$$A = \begin{pmatrix} 1 & 2 \\ 2 & 3 \end{pmatrix}, \qquad \mathbf{b} = \begin{pmatrix} 3 \\ 5 \end{pmatrix},$$

considere el siguiente método iterativo:

$$\mathbf{x}^{k+1} = B(\theta)\mathbf{x}^k + \mathbf{g}(\theta), \qquad \mathbf{x}^0 \in \mathbb{R}^2, \tag{22}$$

donde $\theta \in \mathbb{R}$ y

$$B(\theta) = \frac{1}{4}\begin{pmatrix} 2\theta^2+2\theta+1 & -2\theta^2+2\theta+1 \\ -2\theta^2+2\theta+1 & 2\theta^2+2\theta+1 \end{pmatrix}, \qquad \mathbf{g}(\theta) = \begin{pmatrix} \frac{1}{2}-\theta \\ \frac{1}{2}-\theta \end{pmatrix}$$

Determine los valores de θ para los cuales el método (22) es convergente.

2. Consider la matriz

$$A = \begin{pmatrix} 1 & 2 & -2 \\ 1 & 1 & 1 \\ 2 & 2 & 1 \end{pmatrix}$$

Analice la convergencia de los métodos de Jacobi y Gauss-Seidel para el sistema $A\mathbf{x} = \mathbf{b}$, siendo $\mathbf{b}$ cualquier vector de $\mathbb{R}^{3\times 1}$.

3. Considere el siguiente método iterativo simple para aproximar la solución del sistema $A\boldsymbol{x} = \boldsymbol{b}$ conocido como *iteración de Richardson*

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \theta(\boldsymbol{b} - A\boldsymbol{x}_k),$$

donde θ es un escalar no negativo. Demuestre que si los valores propios de A son reales positivos, entonces el método converge para cualquier dato inicial $\boldsymbol{x}_0$ si y solo si

$$0 < \theta < \frac{2}{\rho(A)}. \tag{23}$$

Determine el valor óptimo de θ para el cual el radio espectral de la matriz de iteración es mínimo.

Solución El método puede reescribirse como

$$\boldsymbol{x}_{k+1} = (I - \theta A)\boldsymbol{x}_k + \theta \boldsymbol{b}.$$

Así, la matriz de iteración es $B_\theta = I - \theta A$. Sean λ_i, $i = 1, \ldots, n$ los valores propios de A, entonces los valores propios de B_θ son $\sigma_i := 1 - \theta\lambda_i$, $i = 1, \ldots, n$. Si denotamos por $\lambda_{\text{mín}}$ y $\lambda_{\text{máx}}$ el menor y mayor de los valores propios de A respectivamente, se tiene que

$$1 - \theta\lambda_{\text{máx}} \leq \sigma_i \leq 1 - \theta\lambda_{\text{mín}} \tag{24}$$

Note que, como los valores propios de A son positivos, $\lambda_{\text{máx}} = \rho(A)$. Suponga que se satisface , entonces $1 - \theta\lambda_{\text{máx}} > -1$ y $1 - \theta\lambda_{\text{mín}} < 1$, esto junto con (24) implica que $\rho(B_\theta) < 1$ y así el método converge. Recíprocamente, suponga que el método converge, esto es, $\rho(B_\theta) < 1$. Entonces

$$\begin{aligned} 1 - \theta\lambda_{\text{mín}} &< 1 \\ 1 - \theta\lambda_{\text{máx}} &> -1 \end{aligned}$$

De la primera desigualdad junto con el hecho de que $\lambda_{\text{mín}} > 0$, se obtiene que $\theta > 0$. De la segunda desigualdad resulta $\theta < 2/\lambda_{\text{máx}}$, de donde se obtiene (3).

Finalmente, para hallar el valor óptimo de θ, nótese que

$$\rho(B_\theta) = \text{máx}\{|1 - \theta\lambda_{\text{mín}}|, |1 - \theta\lambda_{\text{máx}}|\}.$$

Graficando $|1 - \theta\lambda_{\text{mín}}|$ y $|1 - \theta\lambda_{\text{máx}}|$ se observa que el valor mínimo de de $\rho(B_\theta)$ se alcanza cuando $|1 - \theta\lambda_{\text{mín}}| = |1 - \theta\lambda_{\text{máx}}|$, i.e.

$$-1 + \theta\lambda_{\text{máx}} = 1 - \theta\lambda_{\text{mín}}.$$

Luego,

$$\theta_{opt} = \frac{2}{\lambda_{\text{mín}} + \lambda_{\text{máx}}},$$

lo cual da

$$\rho_{opt} = \frac{\lambda_{\text{máx}} - \lambda_{\text{mín}}}{\lambda_{\text{mín}} + \lambda_{\text{máx}}}.$$

4. Métodos de relajación

En vista del teorema (2.5), es razonable esperar que el radio espectral de la matriz de iteración sea una indicador de la rapidez de convergencia del método iterativo. En consecuencia, tiene sentido desarrollar método iterativos o modificar los existentes de manera que el radio espectral de la matriz de iteración sea pequeño. Esta es, en gran medida, la idea detrás de los *métodos de relajación.*

Una generalización del método de Jacobi es el *método de sobrerelajación* (JOR), el cual se obtiene introduciendo un parámetro de relajación ω de la siguiente manera

$$x_{k+1,i} = \frac{\omega}{a_{ii}} \left(b_i - \sum_{\substack{j=1 \\ j \neq i}}^{n} a_{ij} x_{k,j} \right) + (1-\omega) x_{k,i}, \qquad i = 1, \ldots, n. \tag{25}$$

La correspondiente matriz de iteración es

$$B_{J_\omega} = \omega B_J + (1-\omega) I = I - \omega D^{-1} A, \tag{26}$$

y $\boldsymbol{c} := \omega D^{-1} \boldsymbol{b}$. Note que para $\omega \neq 0$, si el método de JOR converge, entonces converge a la solución de $A\boldsymbol{x} = \boldsymbol{b}$.

Teorema 2.11. Si el método de Jacobi es convergente, entonces el método JOR converge si $0 < \omega \leq 1$.

Demostración.

Puesto que la matriz de iteración del método JOR es $B_{J_\omega} = \omega B_J + (1-\omega) I$, entonces sus valores propios son

$$\mu_j = \omega \lambda_j + 1 - \omega, \qquad j = 1, \ldots, n,$$

donde λ_j son los valores propios de B_J. Ahora, en la forma polar $\lambda_j = |\lambda_j| e^{i\theta_j}$. Luego, si $0 < \omega \leq 1$ y $\rho(B_J) < 1$, se tiene

$$|\mu_j|^2 = \omega^2 |\lambda_j|^2 + 2\omega |\lambda_j| \cos(\theta_j)(1-\omega) + (1-\omega)^2 \leq (\omega |\lambda_j| + 1 - \omega)^2 < 1.$$

Por lo tanto, $\rho(B_{J_\omega}) < 1$. □

A partir del método de Gauss-Seidel se introduce ahora el método de *sobrerelajación sucesiva* (SOR) el cual está basado en la partición

$$\omega A = \omega(D - E - F) = (D - \omega E) - (\omega F + (1-\omega) D),$$

de donde se obtiene la iteración

$$(D - \omega E)\boldsymbol{x}_{k+1} = (\omega F + (1-\omega)D)\boldsymbol{x}_k + \omega \boldsymbol{b},$$

o bien (multiplicando a la izquierda por D^{-1})

$$(I - \omega D^{-1}E)\boldsymbol{x}_{k+1} = (\omega D^{-1}F + (1-\omega)I)\boldsymbol{x}_k + \omega D^{-1}\boldsymbol{b}.$$

Luego, la matriz de iteración del método SOR es

$$B(\omega) = (D - \omega E)^{-1}(\omega F + (1-\omega)D) \tag{27}$$

o

$$B(\omega) = (I - \omega D^{-1}E)^{-1}(\omega D^{-1}F + (1-\omega)I). \tag{28}$$

Note que $B(1) = B_{GS}$.

En términos de las componentes, un paso del método SOR se puede expresar como

$$x_{k+1,i} = \frac{\omega}{a_{ii}}\left(b_i - \sum_{j=1}^{i-1} a_{ij}x_{k+1,j} - \sum_{j=i+1}^{n} a_{ij}x_{k,j}\right) + (1-\omega)x_{k,i}, \qquad i = 1, \ldots, n. \tag{29}$$

El siguiente resultado provee una condición necesaria para que el método SOR sea convergente.

Teorema 2.12 (Kahan)**.**

Si el método SOR converge, entonces $\omega \in]0, 2[$.

Demostración.

Puesto que E y $D^{-1}F$ son matrices triangulares con ceros en la diagonal principal, se tiene que $\det(D^{-1}) = \det((D-\omega E)^{-1})$ y $\det(\omega D^{-1}F + (1-\omega)I) = \det((1-\omega)I)$. Luego

$$\begin{aligned}
\det(B(w)) &= \det((D - \omega E)^{-1})\det((\omega F + (1-\omega)D)) \\
&= \det(D^{-1})\det(\omega F + (1-\omega)D) \\
&= \det(\omega D^{-1}F + (1-\omega)I) \\
&= \det((1-\omega)I) \\
&= (1-\omega)^n.
\end{aligned}$$

Por lo tanto, si λ_i, $i = 1, \ldots, n$ son los valores propios de la matriz de iteración del método SOR, se tiene

$$\left|\prod_{i=1}^{n} \lambda_i\right| = |\det(B(\omega))| = |1-\omega|^n,$$

De donde, $\rho(B(\omega)) \geq |1-\omega|$. Esto junto con el hecho de que $\rho(B(\omega)) < 1$ (pues el método es convergente por hipótesis) implica que $\omega \in]0,2[$. □

Observación 2.3. Cuando A es simétrica y definida positiva, la condición $\omega \in]0,2[$, además de ser necesaria, también es suficiente para la convergencia del método SOR a la solución del sistema $A\boldsymbol{x} = \boldsymbol{b}$.

Ejercicios 2.2.

1. Si A es una matriz simétrica definida positiva y tridiagonal, entonces $\rho(B_{GS}) = [\rho(B_J)]^2 < 1$, y la elección óptima del parámero de relajación ω para el método SOR es

$$\omega = \frac{2}{1+\sqrt{1-[\rho(B_J)]^2}}.$$

   Verifique que la matriz

$$A = \begin{pmatrix} 4 & 3 & 0 \\ 3 & 4 & -1 \\ 0 & -1 & 4 \end{pmatrix}$$

   satisface las hipótesis del resultado anterior y luego calcule el valor óptimo de ω para el método SOR si la matriz de coeficientes del sistema lineal es A.

2. Verifique que los valores propios de la matriz $n \times n$ tridiagonal $A = \text{tridiag}(-1, 2, -1)$ están dados por

$$\lambda_j = 2 - 2\cos(j\theta), \quad j = 1, \ldots, n, \tag{30}$$

   donde $\theta = \pi/(n+1)$ y los correspondientes vectores propios son

$$\boldsymbol{v}_j = (\sin(j\theta), \sin(2j\theta), \ldots, \sin(nj\theta))^\top, \quad j = 1, \ldots, n. \tag{31}$$

   Use (30) para concluir que A es definida positiva. Para $n = 5$ calcule el valor óptimo de ω para el método SOR dado que la matriz de coeficientes del sistema lineal es A.

5. Métodos del gradiente y del gradiente conjugado

Una limitante del método SOR y de otros métodos asociados, radica en la dificultad de calcular, en general, el valor óptimo del parámetro de relajación ω debido a que este depende del mayor y menor valor propio de la matriz de iteración. En esta sección, se presenta un método que soslaya la anterior dificultad y que aplica para matrices simétricas y definidas positivas. La idea es resolver un problema de optimización equivalente a resolver el sistema lineal. Más concretamente, considere el sistema lineal

$$A\boldsymbol{x} = \boldsymbol{b},$$

donde la matriz A es simétrica y definida positiva. Considere también la forma cuadrática

$$\Phi(\boldsymbol{y}) = \frac{1}{2}\langle A\boldsymbol{y}, \boldsymbol{y}\rangle - \langle \boldsymbol{b}, \boldsymbol{y}\rangle = \frac{1}{2}\boldsymbol{y}^T A\boldsymbol{y} - \boldsymbol{y}^T\boldsymbol{b}. \tag{32}$$

Si $\boldsymbol{x}$ es solución del sistema lineal, entonces teniendo en cuenta que A es simétrica y definida positiva, resulta que para todo $\boldsymbol{y} \neq \boldsymbol{x}$

$$\Phi(\boldsymbol{y}) = \Phi(\boldsymbol{x} + (\boldsymbol{y} - \boldsymbol{x})) = \Phi(\boldsymbol{x}) + \frac{1}{2}(\boldsymbol{y} - \boldsymbol{x})^T A(\boldsymbol{y} - \boldsymbol{x}) > \Phi(\boldsymbol{x})$$

Luego, la función Φ tiene único mínimo en $\boldsymbol{x}$. Recíprocamente, debido a que

$$\nabla\Phi(\boldsymbol{y}) = \frac{1}{2}(A + A^T)\boldsymbol{y} - \boldsymbol{b} = A\boldsymbol{y} - \boldsymbol{b},$$

se obtiene que si $\nabla\Phi(\boldsymbol{y}) = \boldsymbol{0}$, entonces $\boldsymbol{y} = \boldsymbol{x}$ es solución del sistema lineal. Se busca entonces el vector $\boldsymbol{x}$ para el cual la función Φ alcanza su valor mínimo partiendo de un punto $\boldsymbol{x}_0$. Puesto que el vector gradiente de una función da localmente la dirección de mas rápido ascenso o de máximo crecimiento, una buena idea para disminuir el valor de la función es moverse en la dirección opuesta a la del vector gradiente, es decir, en la dirección del vector

$$-\nabla\Phi(\boldsymbol{y}) = \boldsymbol{b} - A\boldsymbol{y} := \boldsymbol{r},$$

donde $\boldsymbol{r}$ es llamado *vector residual*. Así, partiendo del punto $\boldsymbol{x}_0$, en la k-ésima etapa $\boldsymbol{x}_{k+1}$ está dado por

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \alpha_k \boldsymbol{r}_k, \tag{33}$$

donde

$$\boldsymbol{r}_k = \boldsymbol{b} - A\boldsymbol{x}_k$$

y α_k es un parámetro cuyo valor se determina de manera que $\Phi(\boldsymbol{x}_{k+1})$ sea mínimo. Este método recibe el nombre de *método del gradiente* o del *más rápido descenso.* Para calcular explícitamente el parámetro α_k, note que

$$\begin{aligned}\Phi(\boldsymbol{x}_{k+1}) &= \Phi(\boldsymbol{x}_k + \alpha_k \boldsymbol{r}_k) \\ &= \frac{1}{2}\langle A(\boldsymbol{x}_k + \alpha_k \boldsymbol{r}_k), \boldsymbol{x}_k + \alpha_k \boldsymbol{r}_k\rangle - \langle \boldsymbol{b}, \boldsymbol{x}_k + \alpha_k \boldsymbol{r}_k\rangle \\ &= \frac{1}{2}\langle A\boldsymbol{x}_k, \boldsymbol{x}_k\rangle + \alpha_k\langle A\boldsymbol{x}_k, \boldsymbol{r}_k\rangle + \frac{1}{2}\alpha_k^2\langle A\boldsymbol{r}_k, \boldsymbol{r}_k\rangle - \langle \boldsymbol{b}, \boldsymbol{x}_k\rangle - \alpha_k\langle \boldsymbol{b}, \boldsymbol{r}_k\rangle \\ &= \Phi(\boldsymbol{x}_k) - \alpha_k\langle \boldsymbol{r}_k, \boldsymbol{r}_k\rangle + \frac{1}{2}\alpha_k^2\langle A\boldsymbol{r}_k, \boldsymbol{r}_k\rangle.\end{aligned}$$

Así, $\Phi(\boldsymbol{x}_{k+1})$ es una función cuadrática en α_k con coeficiente principal positivo, en consecuencia, tiene un valor mínimo, el cual se calcula fácilmente derivando con respecto a α_k e igualando a cero, se obtiene que

$$\alpha_k = \frac{\langle \boldsymbol{r}_k, \boldsymbol{r}_k\rangle}{\langle A\boldsymbol{r}_k, \boldsymbol{r}_k\rangle} = \frac{||\boldsymbol{r}_k||_2^2}{\langle A\boldsymbol{r}_k, \boldsymbol{r}_k\rangle} \tag{34}$$

es el valor para el cual $\Phi(\boldsymbol{x}_{k+1})$ alcanza su valor mínimo dado por

$$\Phi(\boldsymbol{x}_{k+1}) = \Phi(\boldsymbol{x}_k) - \frac{1}{2}\frac{||\boldsymbol{r}_k||_2^4}{\langle A\boldsymbol{r}_k, \boldsymbol{r}_k\rangle},$$

lo que muestra que $\Phi(\boldsymbol{x}_k)$ decrece cuando k aumenta hasta que el vector residual sea cero. Ahora, de (33)

$$\boldsymbol{r}_{k+1} = \boldsymbol{b} - A\boldsymbol{x}_{k+1} = \boldsymbol{b} - A(\boldsymbol{x}_k + \alpha_k \boldsymbol{r}_k) = \boldsymbol{r}_k - \alpha_k A\boldsymbol{r}_k,$$

y así de (34),

$$\langle \boldsymbol{r}_{k+1}, \boldsymbol{r}_k\rangle = \langle \boldsymbol{r}_k, \boldsymbol{r}_k\rangle - \alpha_k\langle A\boldsymbol{r}_k, \boldsymbol{r}_k\rangle = 0,$$

es decir, dos vectores residuales consecutivos son ortogonales.

Resumiendo, el método del gradiente se describe como sigue:

Algoritmo (Método del Gradiente o descenso más rápido)**.**

Dado $\boldsymbol{x}_0$, sea $\boldsymbol{r}_0 = \boldsymbol{b} - A\boldsymbol{x}_0$

for $k = 0, 1, 2, ...$ hasta convergencia

$\alpha_k = \frac{\langle \boldsymbol{r}_k, \boldsymbol{r}_k\rangle}{\langle A\boldsymbol{r}_k, \boldsymbol{r}_k\rangle}$

$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \alpha_k \boldsymbol{r}_k$

$\boldsymbol{r}_{k+1} = \boldsymbol{r}_k - \alpha_k A\boldsymbol{r}_k$

end

Se probará la convergencia del método del más rápido descenso a partir del la siguiente desigualdad.

Lema 2.1 (Desigualdad de Kantorovich)**.** Si $A \in \mathbb{R}^{n\times n}$ es simétrica y definida positiva, entonces para cualquier vector $\boldsymbol{x} \neq \mathbf{0}$

$$\frac{\langle A\boldsymbol{x}, \boldsymbol{x}\rangle\langle A^{-1}\boldsymbol{x}, \boldsymbol{x}\rangle}{\langle \boldsymbol{x}, \boldsymbol{x}\rangle^2} \leq \frac{(\lambda_{\text{máx}} + \lambda_{\text{mín}})^2}{4\lambda_{\text{máx}}\lambda_{\text{mín}}}. \tag{35}$$

Aquí $\lambda_{\text{máx}}$ es el mayor valor propio de A y $\lambda_{\text{mín}}$ el más pequeño.

Demostración.

Claramente, es suficiente probar que la desigualdad se cumple para cualquier vector unitario $\boldsymbol{x}$. Tómese $\boldsymbol{x} \in \mathbb{R}^n$ con $||\boldsymbol{x}||_2 = 1$. Puesto que A es simétrica, entonces es unitariamente similar a una matriz diagonal cuyas entradas son los valores propios de A (ver Corolario 1.12), esto es, existe Q unitaria tal que

$$A = Q^T DQ,$$

donde $D = diag(\lambda_1, \ldots, \lambda_n)$ con $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$. Definiendo $\boldsymbol{y} := Q\boldsymbol{x} = (y_1, \ldots, y_n)^T$, vemos que

$$\begin{aligned}
\langle A\boldsymbol{x}, \boldsymbol{x}\rangle\langle A^{-1}\boldsymbol{x}, \boldsymbol{x}\rangle &= \langle Q^T DQ\boldsymbol{x}, \boldsymbol{x}\rangle\langle Q^T D^{-1}Q\boldsymbol{x}, \boldsymbol{x}\rangle \\
&= \langle DQ\boldsymbol{x}, Q\boldsymbol{x}\rangle\langle D^{-1}Q\boldsymbol{x}, Q\boldsymbol{x}\rangle \\
&= \langle D\boldsymbol{y}, \boldsymbol{y}\rangle\langle D^{-1}\boldsymbol{y}, \boldsymbol{y}\rangle \\
&= \left(\sum_{i=1}^n y_i^2\lambda_i\right)\left(\sum_{i=1}^n y_i^2\frac{1}{\lambda_i}\right)
\end{aligned} \tag{36}$$

Como Q es unitaria, $\sum_{i=1}^n y_i^2 = ||\boldsymbol{y}||_2^2 = ||Q\boldsymbol{x}||_2^2 = ||\boldsymbol{x}||_2^2 = 1$, por ende,

$$\hat{t} := \sum_{i=1}^n y_i^2\lambda_i$$

es una combinación convexa de los valores propios de A. Considere la función $f(t) = 1/t$ con $t \in [\lambda_1, \lambda_n]$ y la recta que pasa los puntos $(\lambda_1, 1/\lambda_1)$ y $(\lambda_n, 1/\lambda_n)$ cuya ecuación es

$$h(t) = \frac{1}{\lambda_1} + \frac{1}{\lambda_n} - \frac{t}{\lambda_1\lambda_n}.$$

Como f es convexa, $f(t) \leq h(t)$ para todo $t \in [\lambda_1, \lambda_n]$. De este hecho y la igualdad $\sum_{i=1}^n y_i^2 = 1$, se deduce

$$\begin{aligned}\sum_{i=1}^n y_i^2 \frac{1}{\lambda_i} = \sum_{i=1}^n y_i^2 f(\lambda_i) &\leq \sum_{i=1}^n y_i^2 h(\lambda_i) \\ &= \sum_{i=1}^n y_i^2 \left(\frac{1}{\lambda_1} + \frac{1}{\lambda_n} - \frac{\lambda_i}{\lambda_1 \lambda_n}\right) \\ &= \frac{1}{\lambda_1} + \frac{1}{\lambda_n} - \frac{\sum_{i=1}^n y_i^2 \lambda_i}{\lambda_1 \lambda_n} \\ &= h(\hat{t}).\end{aligned}$$

Combinando la última igualdad con (36), se obtiene

$$\langle A\boldsymbol{x}, \boldsymbol{x}\rangle \langle A^{-1}\boldsymbol{x}, \boldsymbol{x}\rangle \leq \hat{t} h(\hat{t}),$$

y como el máximo de la función $th(t)$ se alcanza en el punto $t = (\lambda_1 + \lambda_n)/2$, resulta

$$\langle A\boldsymbol{x}, \boldsymbol{x}\rangle \langle A^{-1}\boldsymbol{x}, \boldsymbol{x}\rangle \leq \frac{\lambda_1 + \lambda_n}{2} h\left(\frac{\lambda_1 + \lambda_n}{2}\right) = \frac{(\lambda_1 + \lambda_n)^2}{4\lambda_1 \lambda_n},$$

lo cual da el resultado deseado. □

Para $A \in \mathbb{R}^{n\times n}$ simétrica y definida positiva, la *A-norma* o *norma de energía*, notada por $||\cdot||_A$ se define así

$$||\boldsymbol{x}||_A = \langle A\boldsymbol{x}, \boldsymbol{x}\rangle^{1/2}.$$

La propiedades N1 y N2 de la definición de norma son fáciles de verificar. Para probar la desigualdad triangular, se usa la desigualdad de Cauchy-Schwarz para matrices simétricas y definidas positivas, la cual afirma que para cualesquiera $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$:

$$\langle A\boldsymbol{x}, \boldsymbol{y}\rangle \leq \sqrt{\langle A\boldsymbol{x}, \boldsymbol{x}\rangle}\sqrt{\langle A\boldsymbol{y}, \boldsymbol{y}\rangle}. \tag{37}$$

En efecto, como A es definida positiva, entonces para cualesquiera $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$,

$$0 \leq \langle A(\boldsymbol{x} + \alpha\boldsymbol{y}), \boldsymbol{x} + \alpha\boldsymbol{y}\rangle = \langle A\boldsymbol{x}, \boldsymbol{x}\rangle + 2\alpha\langle A\boldsymbol{x}, \boldsymbol{y}\rangle + \alpha^2\langle A\boldsymbol{y}, \boldsymbol{y}\rangle := p(\alpha)$$

Así, el polinomio p no cambia de signo y por lo tanto, su discriminante es no positivo, esto es,

$$4\langle A\boldsymbol{x}, \boldsymbol{y}\rangle^2 - 4\langle A\boldsymbol{x}, \boldsymbol{x}\rangle\langle A\boldsymbol{y}, \boldsymbol{y}\rangle \leq 0,$$

de donde resulta (37). La desigualdad (37) también puede expresarse como

$$\langle A\boldsymbol{x}, \boldsymbol{y}\rangle \leq ||\boldsymbol{x}||_A ||\boldsymbol{y}||_A. \tag{38}$$

Verificamos ahora la desigualdad triangular. Para $\boldsymbol{x}$, $\boldsymbol{y}$ cualesquiera,

$$\begin{aligned} ||\boldsymbol{x}+\boldsymbol{y}||_A^2 &= \langle A(\boldsymbol{x}+\boldsymbol{y}), \boldsymbol{x}+\boldsymbol{y}\rangle \\ &= \langle A\boldsymbol{x}, \boldsymbol{x}\rangle + 2\langle A\boldsymbol{x}, \boldsymbol{y}\rangle + \langle A\boldsymbol{y}, \boldsymbol{y}\rangle \\ &\leq \langle A\boldsymbol{x}, \boldsymbol{x}\rangle + 2\sqrt{\langle A\boldsymbol{x}, \boldsymbol{x}\rangle}\sqrt{\langle A\boldsymbol{y}, \boldsymbol{y}\rangle} + \langle A\boldsymbol{y}, \boldsymbol{y}\rangle \\ &= \left(\sqrt{\langle A\boldsymbol{x}, \boldsymbol{x}\rangle} + \sqrt{\langle A\boldsymbol{y}, \boldsymbol{y}\rangle}\right)^2 \\ &= (||\boldsymbol{x}||_A + ||\boldsymbol{y}||_A)^2, \end{aligned}$$

y tomando la raz se obtiene la desigualdad triangular.

Ejercicio 2.2. Pruebe que si A es simétrica y definida positiva con valores propios $\lambda_1, \ldots, \lambda_n$, y q es un polinomio, entonces

$$||q(A)||_A = \max_{1\leq j\leq n} |q(\lambda_j)|.$$

Aquí $||q(A)||_A$ es la norma matricial de $q(A)$ inducida por la A-norma vectorial. **Sugerencia.** Use el hecho de que si λ es un valor propio de A con vector propio asociado $\boldsymbol{v}$, entonces $q(\lambda)$ es un valor propio de la matriz $q(A)$ con vector propio asociado $\boldsymbol{v}$. Pruebe también esta afirmación.

Demostración. Sean λ_j, $j = 1, \ldots, n$ los valores propios de A, los cuales son positivos, pues A es simétrica y definida positiva. Por el colorario (1.12), los vectores propios de A, los cuales denotamos por $\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_n$, forman una base ortonormal de $\mathbb{R}^n$. Luego, cualquier vector $\boldsymbol{v}$ puede escribirse como

$$\boldsymbol{v} = \sum_{j=1}^{n} \alpha_j \boldsymbol{u}_j.$$

De donde,

$$\begin{aligned} ||\boldsymbol{v}||_A^2 = \boldsymbol{v}^\top A\boldsymbol{v} &= \langle A\boldsymbol{v}, \boldsymbol{v}\rangle \\ &= \left\langle \sum_{j=1}^{n} \alpha_j A\boldsymbol{u}_j, \sum_{m=1}^{n} \alpha_m \boldsymbol{u}_m \right\rangle \\ &= \left\langle \sum_{j=1}^{n} \alpha_j \lambda_j \boldsymbol{u}_j, \sum_{m=1}^{n} \alpha_m \boldsymbol{u}_m \right\rangle \end{aligned}$$

$$= \sum_{j=1}^{n} \lambda_j |\alpha_j|^2.$$

Por otra parte, teniendo en cuenta que $q(A)\boldsymbol{u}_j = q(\lambda_j)\boldsymbol{u}_j$, $j = 1, \ldots, n$, resulta

$$\begin{aligned}
||q(A)\boldsymbol{v}||_A^2 &= \langle Aq(A)\boldsymbol{v}, q(A)\boldsymbol{v}\rangle \\
&= \left\langle \sum_{j=1}^{n} \alpha_j Aq(A)\boldsymbol{u}_j, \sum_{m=1}^{n} \alpha_m q(A)\boldsymbol{u}_m \right\rangle \\
&= \left\langle \sum_{j=1}^{n} \alpha_j q(\lambda_j)A\boldsymbol{u}_j, \sum_{m=1}^{n} \alpha_m q(\lambda_j)\boldsymbol{u}_m \right\rangle \\
&= \left\langle \sum_{j=1}^{n} \alpha_j q(\lambda_j)\lambda_j\boldsymbol{u}_j, \sum_{m=1}^{n} \alpha_m q(\lambda_j)\boldsymbol{u}_m \right\rangle \\
&= \sum_{j=1}^{n} |q(\lambda_j)|^2 \lambda_j |\alpha_j|^2 \\
&\leq (\max_{1\leq j\leq n} |q(\lambda_j)|)^2 \sum_{j=1}^{n} \lambda_j |\alpha_j|^2 \\
&= (\max_{1\leq j\leq n} |q(\lambda_j)|)^2 ||\boldsymbol{v}||_A^2.
\end{aligned}$$

En consecuencia,

$$||q(A)||_A \leq \max_{1\leq j\leq n} |q(\lambda_j)|.$$

Veamos ahora la otra desigualdad, tómese λ un valor propio cualquiera de A con vector propio asociado $\boldsymbol{w} \neq \boldsymbol{0}$. Entonces

$$||q(A)||_A = \max_{\boldsymbol{v}\neq\boldsymbol{0}} \frac{||q(A)\boldsymbol{v}||_A}{||\boldsymbol{v}||_A} \geq \frac{||q(A)\boldsymbol{w}||_A}{||\boldsymbol{w}||_A} = \frac{||q(\lambda)\boldsymbol{w}||_A}{||\boldsymbol{w}||_A} = |q(\lambda)|.$$

De donde,

$$||q(A)||_A \geq \max_{1\leq j\leq n} |q(\lambda_j)|.$$

□

Ahora se establece el resultado concerniente a la convergencia del método del gradiente.

Teorema 2.13. Si A es simétrica y definida positiva, entonces la A-norma del vector error $\boldsymbol{e}_k = \boldsymbol{x} - \boldsymbol{x}_k$ generado por el método del gradiente satisface la desigualdad

$$||\boldsymbol{e}_{k+1}||_A \leq \frac{\lambda_{\text{máx}} - \lambda_{\text{mín}}}{\lambda_{\text{máx}} + \lambda_{\text{mín}}} ||\boldsymbol{e}_k||_A, \tag{39}$$

Luego,

$$||\boldsymbol{e}_{k+1}||_A \leq \left(\frac{\lambda_{\text{máx}} - \lambda_{\text{mín}}}{\lambda_{\text{máx}} + \lambda_{\text{mín}}} \right)^{k+1} ||\boldsymbol{e}_0||_A, \tag{40}$$

y en consecuencia, el método del gradiente converge para cualquier dato inicial $\boldsymbol{x}_0$.

Demostración.

Teniendo en cuenta que $A\boldsymbol{e}_{k+1} = \boldsymbol{r}_{k+1}$, $||\boldsymbol{e}_k||_A^2 = \langle \boldsymbol{r}_k, A^{-1}\boldsymbol{r}_k \rangle$, y que dos vectores residuales consecutivos son ortogonales, resulta

$$\begin{aligned}
||\boldsymbol{e}_{k+1}||_A^2 &= \langle A\boldsymbol{e}_{k+1}, \boldsymbol{e}_{k+1} \rangle \\
&= \langle \boldsymbol{r}_{k+1}, \boldsymbol{e}_{k+1} \rangle \\
&= \langle \boldsymbol{r}_{k+1}, \boldsymbol{e}_k - \alpha_k \boldsymbol{r}_k \rangle \\
&= \langle \boldsymbol{r}_{k+1}, \boldsymbol{e}_k \rangle \\
&= \langle \boldsymbol{r}_k - \alpha_k A\boldsymbol{r}_k, A^{-1}\boldsymbol{r}_k \rangle \\
&= \langle \boldsymbol{r}_k, A^{-1}\boldsymbol{r}_k \rangle - \alpha_k \langle \boldsymbol{r}_k, \boldsymbol{r}_k \rangle \\
&= ||\boldsymbol{e}_k||_A^2 \left(1 - \frac{\langle \boldsymbol{r}_k, \boldsymbol{r}_k \rangle}{\langle A\boldsymbol{r}_k, \boldsymbol{r}_k \rangle} \frac{\langle \boldsymbol{r}_k, \boldsymbol{r}_k \rangle}{\langle \boldsymbol{r}_k, A^{-1}\boldsymbol{r}_k \rangle} \right) \\
&\leq ||\boldsymbol{e}_k||_A^2 \left(1 - \frac{4\lambda_{\text{máx}}\lambda_{\text{mín}}}{(\lambda_{\text{máx}} + \lambda_{\text{mín}})^2} \right) \qquad \text{(por la Des. de Kantorovich)}
\end{aligned}$$

de donde se sigue (39). □

Recordemos que si A es simétrica y definida positiva, se tiene que $\kappa_2(A) = \lambda_{\text{máx}}/\lambda_{\text{mín}}$. Así, la desigualdad (40) puede reescribirse como

$$||\boldsymbol{e}_{k+1}||_A \leq \left(\frac{\kappa_2(A) - 1}{\kappa_2(A) + 1} \right)^{k+1} ||\boldsymbol{e}_0||_A. \tag{41}$$

Por lo tanto, aunque el teorema previo garantiza la convergencia del método del gradiente, la razón de convergencia puede ser muy lenta si A está mal condicionada, es decir, si $\lambda_{\text{máx}}$ es muy grande con respecto a $\lambda_{\text{mín}}$. Desde el

punto de vista geométrico, se tiene una descripción interesante que, por sencillez, se ilustra para $n = 2$. Suponga que $A = diag(\lambda_{\text{máx}}, \lambda_{\text{mín}})$, como A es definida positiva, las curvas de nivel $\{(x_1, x_2) : \Phi(\boldsymbol{x}) = c\}$ con $\boldsymbol{x} = (x_1, x_2)$ y $c \in \mathbb{R}^+$ son elipses. Además, la recta que pasa por los puntos $\boldsymbol{x}_k$ y $\boldsymbol{x}_{k+1}$ es tangente a la curva de nivel de ecuación $\Phi(\boldsymbol{x}) = \Phi(\boldsymbol{x}_{k+1})$. Luego, si las elipses son muy elongadas, el método converge más lentamente describiendo una trayectoria en zig-zag (ver Figura (4)). Si $\lambda_{\text{máx}} = \lambda_{\text{mín}}$, el método converge en una sola iteración.

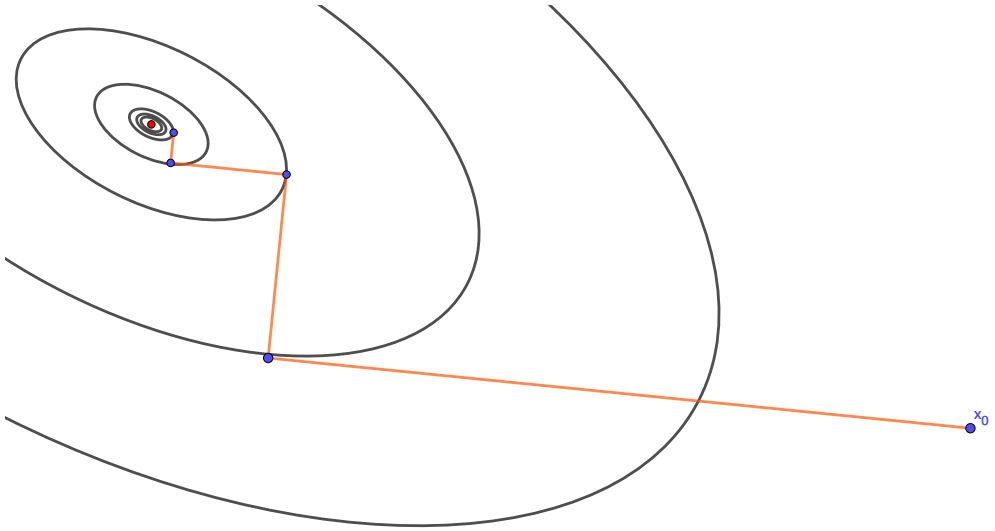

Figura 4: Ilustración del método del gradiente en dos dimensiones.

Para subsanar el problema de la eventual convergencia lenta del método del gradiente, se considera el *método del gradiente conjugado*, el cual se puede ver como una aceleración del primero. La idea es reemplazar la dirección $\boldsymbol{r}_k$ del método del gradiente (en la actualización $k+1$) por una nueva dirección, dada por el vector $\boldsymbol{p}_k$ (llamada *dirección de búsqueda*) que no sea paralela al vector gradiente. Se comienza expresando el nuevo cambio en la posición $\boldsymbol{p}_k$ como una combinación lineal de la dirección del vector gradiente y el cambio previo en la posición. Más concretamente, $\boldsymbol{x}_{k+1}$ se calcula ahora con la fórmula

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \alpha_k \boldsymbol{p}_k, \tag{42}$$

donde

$$\boldsymbol{p}_k = \boldsymbol{r}_k + \gamma_k(\boldsymbol{x}_k - \boldsymbol{x}_{k-1})$$

$$\begin{aligned} &= \boldsymbol{r}_k + \gamma_k \alpha_{k-1} \boldsymbol{p}_{k-1} \\ &= \boldsymbol{r}_k + \beta_{k-1} \boldsymbol{p}_{k-1} \end{aligned} \tag{43}$$

El error residual se puede expresar como

$$\boldsymbol{r}_{k+1} = \boldsymbol{b} - A\boldsymbol{x}_{k+1} = \boldsymbol{r}_k - \alpha_k A \boldsymbol{p}_k. \tag{44}$$

El próximo paso es determinar los valores de los parámetros α_k y β_k, y de la dirección de búsqueda inicial $\boldsymbol{p}_0$ para que que la sucesión $\{\boldsymbol{x}_k\}$ generada por (42) converja rápidamente. Tal como se procedió con el método del más rápido descenso, se desea elegir $\boldsymbol{x}_{k+1}$ de manera que $\Phi(\boldsymbol{x}_{k+1})$ sea mínimo. Se tiene que

$$\begin{aligned} \Phi(\boldsymbol{x}_{k+1}) &= \Phi(\boldsymbol{x}_k + \alpha_k \boldsymbol{p}_k) \\ &= \frac{1}{2}\langle A(\boldsymbol{x}_k + \alpha_k \boldsymbol{p}_k), \boldsymbol{x}_k + \alpha_k \boldsymbol{p}_k \rangle - \langle \boldsymbol{b}, \boldsymbol{x}_k + \alpha_k \boldsymbol{r}_k \rangle \\ &= \frac{1}{2}\langle A\boldsymbol{x}_k, \boldsymbol{x}_k \rangle + \alpha_k \langle A\boldsymbol{x}_k, \boldsymbol{p}_k \rangle + \frac{1}{2}\alpha_k^2 \langle A\boldsymbol{p}_k, \boldsymbol{p}_k \rangle - \langle \boldsymbol{b}, \boldsymbol{x}_k \rangle - \alpha_k \langle \boldsymbol{b}, \boldsymbol{p}_k \rangle \\ &= \Phi(\boldsymbol{x}_k) - \alpha_k \langle \boldsymbol{r}_k, \boldsymbol{p}_k \rangle + \frac{1}{2}\alpha_k^2 \langle A\boldsymbol{p}_k, \boldsymbol{p}_k \rangle. \end{aligned}$$

El mínimo de esta función ocurre en

$$\alpha_k = \frac{\langle \boldsymbol{r}_k, \boldsymbol{p}_k \rangle}{\langle A\boldsymbol{p}_k, \boldsymbol{p}_k \rangle}, \qquad k \geq 0 \tag{45}$$

y el valor mínimo es

$$\Phi(\boldsymbol{x}_{k+1}) = \Phi(\boldsymbol{x}_k) - \frac{1}{2}\frac{\langle \boldsymbol{r}_k, \boldsymbol{p}_k \rangle^2}{\langle A\boldsymbol{p}_k, \boldsymbol{p}_k \rangle}.$$

Considerando el caso $k = 0$, tómese $\boldsymbol{p}_0 = \boldsymbol{r}_0$, lo cual obviamente implica que $\Phi(\boldsymbol{x}_1)$ es menor que $\Phi(\boldsymbol{x}_0)$, otras justificaciones de esta elección se presentarán más adelante. Por el momento, note que en virtud del valor encontrado de α_k y (44), se tiene que

$$\langle \boldsymbol{r}_{k+1}, \boldsymbol{p}_k \rangle = 0.$$

Luego, teniendo en cuenta la relación (43), resulta que para $k \geq 0$,

$$\langle \boldsymbol{r}_{k+1}, \boldsymbol{p}_{k+1} \rangle = \langle \boldsymbol{r}_{k+1}, \boldsymbol{r}_{k+1} \rangle + \beta_k \langle \boldsymbol{r}_{k+1}, \boldsymbol{p}_k \rangle = ||\boldsymbol{r}_{k+1}||_2^2.$$

Por la elección de $\boldsymbol{p}_0$, se obtiene

$$\langle \boldsymbol{r}_k, \boldsymbol{p}_k \rangle = ||\boldsymbol{r}_k||_2^2, \qquad \forall k \geq 0,$$

lo cual permite escribir α_k de manera conveniente como

$$\alpha_k = \frac{||\boldsymbol{r}_k||_2^2}{\langle A\boldsymbol{p}_k, \boldsymbol{p}_k\rangle}. \tag{46}$$

Así,

$$\Phi(\boldsymbol{x}_{k+1}) = \Phi(\boldsymbol{x}_k) - \frac{1}{2}\frac{||\boldsymbol{r}_k||_2^4}{\langle A\boldsymbol{p}_k, \boldsymbol{p}_k\rangle}.$$

Esta última fórmula indica que para minimizar $\Phi(\boldsymbol{x}_{k+1})$, $\boldsymbol{p}_k$ se debe elegir de manera que $\langle A\boldsymbol{p}_k, \boldsymbol{p}_k\rangle$ sea mínimo. Usando relación (43), se tiene que

$$\begin{aligned}\langle A\boldsymbol{p}_k, \boldsymbol{p}_k\rangle &= \langle A(\boldsymbol{r}_k + \beta_{k-1}\boldsymbol{p}_{k-1}), \boldsymbol{r}_k + \beta_{k-1}\boldsymbol{p}_{k-1}\rangle \\ &= \langle A\boldsymbol{r}_k, \boldsymbol{r}_k\rangle + 2\beta_{k-1}\langle A\boldsymbol{p}_{k-1}, \boldsymbol{r}_k\rangle + \beta_{k-1}^2\langle A\boldsymbol{p}_{k-1}, \boldsymbol{p}_{k-1}\rangle.\end{aligned}$$

Dado $\boldsymbol{p}_{k-1}$, el valor mínimo de esta función cuadrática (en la variable β_{k-1}) ocurre en

$$\beta_{k-1} = -\frac{\langle A\boldsymbol{p}_{k-1}, \boldsymbol{r}_k\rangle}{\langle A\boldsymbol{p}_{k-1}, \boldsymbol{p}_{k-1}\rangle} \qquad k \geq 1,$$

o bien,

$$\beta_k = -\frac{\langle A\boldsymbol{p}_k, \boldsymbol{r}_{k+1}\rangle}{\langle A\boldsymbol{p}_k, \boldsymbol{p}_k\rangle} \qquad k \geq 0. \tag{47}$$

Usando este valor de β_k junto con la relación $\boldsymbol{p}_{k+1} = \boldsymbol{r}_{k+1} + \beta_k\boldsymbol{p}_k$ (ver (43)), se obtiene

$$\langle A\boldsymbol{p}_k, \boldsymbol{p}_{k+1}\rangle = 0 \qquad k \geq 0, \tag{48}$$

es decir, dos direcciones de búsqueda consecutivas son *conjugadas ortogonales* o *A-ortogonales*. En virtud de este hecho y (43), también se obtiene la ortogonalidad de dos vectores residuales consecutivos. En efecto,

$$\begin{aligned}\langle \boldsymbol{r}_k, \boldsymbol{r}_{k+1}\rangle &= \langle \boldsymbol{r}_k, \boldsymbol{r}_k\rangle - \alpha_k\langle \boldsymbol{r}_k, A\boldsymbol{p}_k\rangle \\ &= \langle \boldsymbol{r}_k, \boldsymbol{r}_k\rangle - \alpha_k\left(\langle \boldsymbol{r}_k, A\boldsymbol{p}_k\rangle + \beta_{k-1}\langle \boldsymbol{p}_{k-1}, A\boldsymbol{p}_k\rangle\right) \\ &= \langle \boldsymbol{r}_k, \boldsymbol{r}_k\rangle - \alpha_k\left(\langle \boldsymbol{r}_k + \beta_{k-1}\boldsymbol{p}_{k-1}, A\boldsymbol{p}_k\rangle\right) \\ &= \langle \boldsymbol{r}_k, \boldsymbol{r}_k\rangle - \alpha_k\langle \boldsymbol{p}_k, A\boldsymbol{p}_k\rangle \\ &= 0.\end{aligned}$$

Más generalmente, se tiene que

$$\langle \boldsymbol{r}_\nu, \boldsymbol{r}_k\rangle = \langle A\boldsymbol{p}_\nu, \boldsymbol{p}_k\rangle = 0 \qquad \text{para } k \neq \nu. \tag{49}$$

Este resultado se prueba por inducción. Puesto que dos vectores residuales son ortogonales y dos direcciones de búsqueda son $A-$ortogonales, se tiene que

$$\langle \boldsymbol{r}_0, \boldsymbol{r}_1 \rangle = 0 \qquad \text{y} \qquad \langle A\boldsymbol{p}_0, \boldsymbol{p}_1 \rangle = 0.$$

Ahora suponga que

$$\langle \boldsymbol{r}_\nu, \boldsymbol{r}_l \rangle = \langle A\boldsymbol{p}_\nu, \boldsymbol{p}_l \rangle = 0 \qquad \text{para } 0 \leq \nu < l \leq k \text{ (hipótesis de inducción)}$$

Hay que probar que esto vale para $0 \leq \nu < l \leq k+1$. Para el caso $\nu = k$ y $l = k+1$, el resultado se sigue de la ortogonalidad de dos vectores residuales consecutivos e igualmente para dos direcciones de búsqueda consecutivas. Tomando $1 \leq \nu < k$ y $l = k+1$, de las relaciones (43)-(44) y la hipótesis de inducción, resulta

$$\begin{aligned}
\langle \boldsymbol{r}_\nu, \boldsymbol{r}_{k+1} \rangle &= \langle \boldsymbol{r}_\nu, \boldsymbol{r}_k \rangle - \alpha_k \langle \boldsymbol{r}_\nu, A\boldsymbol{p}_k \rangle \\
&= -\alpha_k \langle \boldsymbol{p}_\nu - \beta_{\nu-1}\boldsymbol{p}_{\nu-1}, A\boldsymbol{p}_k \rangle \\
&= -\alpha_k \langle \boldsymbol{p}_\nu, A\boldsymbol{p}_k \rangle + \alpha_k \beta_{\nu-1} \langle \boldsymbol{p}_{\nu-1}, A\boldsymbol{p}_k \rangle \\
&= 0.
\end{aligned}$$

A partir de este resultado y la hipótesis de inducción también se obtiene que

$$\begin{aligned}
\langle A\boldsymbol{p}_\nu, \boldsymbol{p}_{k+1} \rangle &= \langle A\boldsymbol{p}_\nu, \boldsymbol{r}_{k+1} + \beta_k \boldsymbol{p}_k \rangle \\
&= \langle A\boldsymbol{p}_\nu, \boldsymbol{r}_{k+1} \rangle \\
&= \frac{1}{\alpha_\nu} \langle \boldsymbol{r}_\nu - \boldsymbol{r}_{\nu+1}, \boldsymbol{r}_{k+1} \rangle \\
&= \frac{1}{\alpha_\nu} \langle \boldsymbol{r}_\nu, \boldsymbol{r}_{k+1} \rangle - \frac{1}{\alpha_\nu} \langle \boldsymbol{r}_{\nu+1}, \boldsymbol{r}_{k+1} \rangle \\
&= 0,
\end{aligned}$$

cuando $\alpha_\nu \neq 0$. Si $\alpha_\nu = 0$, entonces $\boldsymbol{r}_\nu = \boldsymbol{p}_\nu = \boldsymbol{0}$, lo cual completa la prueba.

Una consecuencia inmediata de la propiedad de ortogonalidad (49) es el siguiente resultado.

Teorema 2.14. Si $A \in \mathbb{R}^{n\times n}$ es simétrica y definida positiva, el método del gradiente conjugado converge a lo sumo en n pasos.

Demostración. Considere los vectores residuales $\boldsymbol{r}_0, \boldsymbol{r}_1, \ldots, \boldsymbol{r}_{n-1}$, si alguno de ellos es nulo, el resultado se obtiene de inmediato. Si ninguno de tales vectores es nulo, de la propiedad (49), estos n vectores residuales son ortogonales

y por ende, son linealmente independientes, lo que permite concluir que forman una base ortogonal de $\mathbb{R}^n$. Por la misma propiedad, $\boldsymbol{r}_n$ es ortogonal $span\{\boldsymbol{r}_0, \boldsymbol{r}_1, \ldots, \boldsymbol{r}_{n-1}\} = \mathbb{R}^n$ y así $\boldsymbol{r}_n = \boldsymbol{0}$, lo cual implica que $\boldsymbol{x}_n = \boldsymbol{x}$. □

Por la propiedad de ortogonalidad también se tiene que

$$\langle \boldsymbol{r}_{k+1}, \boldsymbol{r}_{k+1} \rangle = \langle \boldsymbol{r}_k - \alpha_k A\boldsymbol{p}_k, \boldsymbol{r}_{k+1} \rangle = -\alpha_k \langle A\boldsymbol{p}_k, \boldsymbol{r}_{k+1} \rangle.$$

Luego, remplazando esta igualdad en la fórmula (47), resulta

$$\beta_k = -\frac{\alpha_k^{-1} \langle \boldsymbol{r}_{k+1}, \boldsymbol{r}_{k+1} \rangle}{\langle A\boldsymbol{p}_k, \boldsymbol{p}_k \rangle} = \frac{\langle \boldsymbol{r}_{k+1}, \boldsymbol{r}_{k+1} \rangle}{\langle \boldsymbol{r}_k, \boldsymbol{r}_k \rangle} = \frac{||\boldsymbol{r}_{k+1}||_2^2}{||\boldsymbol{r}_k||_2^2}. \tag{50}$$

Se presenta ahora algoritmo del método del gradiente conjugado

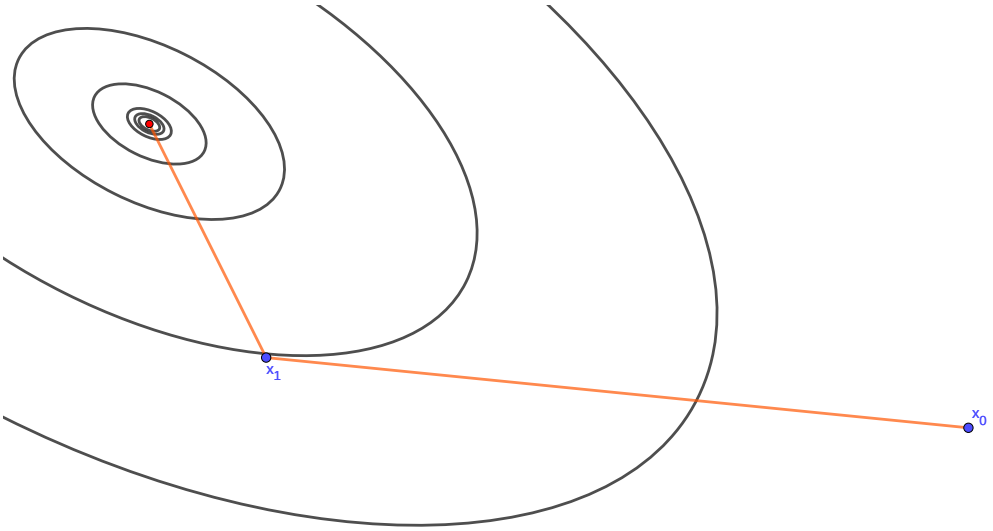

Figura 5: Ilustración del método del gradiente conjugado en dos dimensiones.

Algoritmo (Método del Gradiente Conjugado)**.**

Dado $\boldsymbol{x}_0$, sean $\boldsymbol{r}_0 = \boldsymbol{b} - A\boldsymbol{x}_0$, $\boldsymbol{p}_0 = \boldsymbol{r}_0$

for $k = 0, 1, 2, \ldots$ hasta convergencia

$\alpha_k = \frac{\langle \boldsymbol{r}_k, \boldsymbol{r}_k \rangle}{\langle A\boldsymbol{p}_k, \boldsymbol{r}_k \rangle}$

$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \alpha_k \boldsymbol{p}_k$

$\boldsymbol{r}_{k+1} = \boldsymbol{r}_k - \alpha_k A\boldsymbol{p}_k$

$\beta_k = \frac{\langle \boldsymbol{r}_{k+1}, \boldsymbol{r}_{k+1} \rangle}{\langle \boldsymbol{r}_k, \boldsymbol{r}_k \rangle}$

$\boldsymbol{p}_{k+1} = \boldsymbol{r}_{k+1} + \beta_k \boldsymbol{p}_k$

end

Observación 2.4. Como $\boldsymbol{p}_0 = \boldsymbol{r}_0$ y $\boldsymbol{p}_{k+1} = \boldsymbol{r}_{k+1} + \beta_k \boldsymbol{p}_k$, es claro que para $k \geq 0$,

$$span\{\boldsymbol{p}_0, \boldsymbol{p}_1, \ldots, \boldsymbol{p}_k\} = span\{\boldsymbol{r}_0, \boldsymbol{r}_1, \ldots, \boldsymbol{r}_k\} \tag{51}$$

Se afirma también que para $k \geq 0$,

$$span\{\boldsymbol{p}_0, \boldsymbol{p}_1, \ldots, \boldsymbol{p}_k\} = span\{\boldsymbol{r}_0, A\boldsymbol{r}_0, \ldots, A^k\boldsymbol{r}_0\} =: \mathcal{K}_{k+1}(\boldsymbol{r}_0). \tag{52}$$

En efecto, para $k = 0$ la afirmación es claramente válida porque $\boldsymbol{p}_0 = \boldsymbol{r}_0$. Suponga que

$$span\{\boldsymbol{p}_0, \boldsymbol{p}_1, \ldots, \boldsymbol{p}_k\} = \mathcal{K}_{k+1}(\boldsymbol{r}_0)$$

y probemos que

$$span\{\boldsymbol{p}_0, \boldsymbol{p}_1, \ldots, \boldsymbol{p}_k, \boldsymbol{p}_{k+1}\} = \mathcal{K}_{k+2}(\boldsymbol{r}_0).$$

Puesto que por hipótesis de inducción $\boldsymbol{p}_k \in \mathcal{K}_{k+1}(\boldsymbol{r}_0)$, se sigue que $A\boldsymbol{p}_k \in \mathcal{K}_{k+2}(\boldsymbol{r}_0)$, y así $\boldsymbol{r}_{k+1} = \boldsymbol{r}_k - \alpha_k A\boldsymbol{p}_k \in \mathcal{K}_{k+2}(\boldsymbol{r}_0)$, lo que a su vez implica que $\boldsymbol{p}_{k+1} = \boldsymbol{r}_{k+1} + \beta_k A\boldsymbol{p}_k \in \mathcal{K}_{k+2}(\boldsymbol{r}_0)$. En consecuencia,

$$span\{\boldsymbol{p}_0, \boldsymbol{p}_1, \ldots, \boldsymbol{p}_k, \boldsymbol{p}_{k+1}\} \subseteq \mathcal{K}_{k+2}(\boldsymbol{r}_0).$$

Por otra parte, de la hipótesis de inducción, $A^k\boldsymbol{r}_0 \in span\{\boldsymbol{p}_0, \boldsymbol{p}_1, \ldots, \boldsymbol{p}_k\}$ y consecuentemente, $A^{k+1}\boldsymbol{r}_0 \in span\{A\boldsymbol{p}_0, A\boldsymbol{p}_1, \ldots, A\boldsymbol{p}_k\}$. Luego, de la fórmula $\boldsymbol{r}_{k+1} = \boldsymbol{r}_k + \alpha_k A\boldsymbol{p}_k$, se deduce que $A^{k+1}\boldsymbol{r}_0 \in span\{\boldsymbol{r}_0, \boldsymbol{r}_1, \ldots, \boldsymbol{r}_{k+1}\}$. Esto, junto con (51), implica la otra contenencia.

A continuación se presenta un resultado sobre polinomios de Tchevyshev que será de gran utilidad para demostrar una estimación de la tasa de convergencia del método del gradiente conjugado.

Teorema 2.15. Sea $[a, b] \subset \mathbb{R}$ un intervalo que no contiene al cero. Entonces

$$\min_{h_k \in \mathbb{P}_k, h_k(0)=1} \max_{t \in [a,b]} |h_k(t)| = \frac{T_k\left(\frac{b+a-2t}{b-a}\right)}{T_k\left(\frac{b+a}{b-a}\right)},$$

donde $T_k(\mu)$ es el polinomio de Tchebyshev de grado k dado por

$$T_k(\mu) = \begin{cases} \cos(k \cos^{-1} \mu), & \text{if } |\mu| \leq 1, \\ \cosh(k \cosh^{-1} \mu), & \text{if } |\mu| \geq 1. \end{cases}$$

Observación 2.5. Si $a > 0$, entonces para $t \in [a, b]$,

$$\left| \frac{b + a - 2t}{b - a} \right| \leq 1.$$

Luego, del teorema previo, si se toma

$$h_k(t) = \frac{T_k\left(\frac{b+a-2t}{b-a}\right)}{T_k\left(\frac{b+a}{b-a}\right)},$$

se obtiene

$$\max_{t\in[a,b]} |h_k(t)| \leq \frac{1}{T_k\left(\frac{b+a}{b-a}\right)} = \left[\cosh\left(k \cosh^{-1}\left(\frac{b+a}{b-a}\right)\right)\right]^{-1}. \tag{53}$$

Note que la expresión entre corchetes crece a medida k crece, lo cual implica que $||\boldsymbol{e}_k||_A$ decrece si k crece. Para obtener una estimación más útil, defina

$$\eta = \cosh^{-1}\left(\frac{b+a}{b-a}\right).$$

Luego,

$$\exp(\eta) = \frac{\sqrt{b} + \sqrt{a}}{\sqrt{b} - \sqrt{a}}.$$

Por lo tanto,

$$\cosh(k\eta) = \frac{\exp(k\eta) + \exp(-k\eta)}{2} \geq \frac{\exp(k\eta)}{2} = \frac{1}{2}\left(\frac{\sqrt{b} + \sqrt{a}}{\sqrt{b} - \sqrt{a}}\right)^k.$$

Esto combinado con (53), da como resultado

$$\max_{t\in[a,b]} |h_k(t)| \leq 2\left(\frac{\sqrt{b} - \sqrt{a}}{\sqrt{b} + \sqrt{a}}\right)^k. \tag{54}$$

Teorema 2.16. Si A es una matriz simétrica y definida positiva cuyos valores propios están en el intervalo $[a, b]$, con $a > 0$, el vector error $\boldsymbol{e}_k = \boldsymbol{x} - \boldsymbol{x}_k$ para el método del gradiente conjugado, satisface la desigualdad

$$||\boldsymbol{e}_k||_A \leq 2\left(\frac{\sqrt{b} - \sqrt{a}}{\sqrt{b} + \sqrt{a}}\right)^k ||\boldsymbol{e}_0||_A. \tag{55}$$

Demostración. Por la Observación (2.4), el vector residual $\boldsymbol{r}_k$ pertenece a $span\{\boldsymbol{r}_0, A\boldsymbol{r}_0, \ldots, A^k\boldsymbol{r}_0\}$. Así,

$$\boldsymbol{r}_k = q_k(A)\boldsymbol{r}_0, \tag{56}$$

donde $q_k(t)$ es un polinomio de grado k en la variable t. Note que si $A = \mathbf{0}$, entonces por (44), $\boldsymbol{r}_k = \boldsymbol{r}_0$ y así $q_k(0) = 1$. Ahora, como $\boldsymbol{r}_k = A\boldsymbol{e}_k$, de (56) y el hecho de que $q_k(A)$ y A conmutan, resulta

$$0 = A\boldsymbol{e}_k - q_k(A)A\boldsymbol{e}_0 = A\boldsymbol{e}_k - Aq_k(A)\boldsymbol{e}_0 = A\left(\boldsymbol{e}_k - q_k(A)\boldsymbol{e}_0\right),$$

y como A es no singular, se obtiene

$$\boldsymbol{e}_k = q_k(A)\boldsymbol{e}_0. \tag{57}$$

Se afirma que para cualquier polinomio $h_k(t)$ de grado k tal que $h_k(0) = 1$ se cumple lo siguiente

$$\langle A\boldsymbol{e}_k, \boldsymbol{e}_k\rangle = \langle A\boldsymbol{e}_k, h_k(A)\boldsymbol{e}_0\rangle. \tag{58}$$

En efecto, sea $h_k(t)$ un polinomio de grado k con $h_k(0) = 1$ y elija $\gamma_{k-1}, \ldots, \gamma_1, \gamma_0$ de manera que

$$h_k(t) = q_k(t) + \sum_{j=0}^{k-1} \gamma_j t q_j(t).$$

Luego, usando (57) y la ortogonalidad de los vectores residuales, se tiene

$$\begin{aligned}
\langle A\boldsymbol{e}_k, h_k(A)\boldsymbol{e}_0\rangle &= \left\langle A\boldsymbol{e}_k, \left(q_k(A) + \sum_{j=0}^{k-1} \gamma_j A q_j(A)\right)\boldsymbol{e}_0 \right\rangle \\
&= \left\langle A\boldsymbol{e}_k, q_k(A)\boldsymbol{e}_0 + \sum_{j=0}^{k-1} \gamma_j A q_j(A)\boldsymbol{e}_0 \right\rangle \\
&= \left\langle A\boldsymbol{e}_k, \boldsymbol{e}_k + \sum_{j=0}^{k-1} \gamma_j A\boldsymbol{e}_j \right\rangle \\
&= \left\langle A\boldsymbol{e}_k, \boldsymbol{e}_k + \sum_{j=0}^{k-1} \gamma_j \boldsymbol{r}_j \right\rangle \\
&= \langle A\boldsymbol{e}_k, \boldsymbol{e}_k\rangle + \left\langle \boldsymbol{r}_k, \sum_{j=0}^{k-1} \gamma_j \boldsymbol{r}_j \right\rangle \\
&= \langle A\boldsymbol{e}_k, \boldsymbol{e}_k\rangle,
\end{aligned}$$

como se deseaba.

Por otra parte, teniendo en cuenta de desigualdad de Cauchy-Schwarz para matrices simétricas y definidas positivas, resulta

$$||\boldsymbol{e}_k||_A^2 = \langle A\boldsymbol{e}_k, \boldsymbol{e}_k\rangle = \langle A\boldsymbol{e}_k, h_k(A)\boldsymbol{e}_0\rangle \leq ||\boldsymbol{e}_k||_A||h_k(A)\boldsymbol{e}_0||_A,$$

o bien

$$||\boldsymbol{e}_k||_A \leq ||h_k(A)\boldsymbol{e}_0||_A. \tag{59}$$

La idea ahora es estimar el mínimo del lado derecho en esta desigualdad. Usando el ejercicio (2.2) se obtiene

$$||h_k(A)\boldsymbol{e}_0||_A \leq ||h_k(A)||_A||\boldsymbol{e}_0||_A = \max_{1\leq j\leq n} |h_k(\lambda_j)|||\boldsymbol{e}_0||_A \leq \max_{a\leq t\leq b} |h_k(t)|||\boldsymbol{e}_0||_A.$$

En la última desigualdad se ha tenido en cuenta que los valores propios de A están en el intervalo $[a, b]$. Se toma

$$h_k(t) = \frac{T_k\left(\frac{b+a-2t}{b-a}\right)}{T_k\left(\frac{b+a}{b-a}\right)},$$

y en virtud de la estimación (54), resulta

$$||\boldsymbol{e}_k||_A \leq \max_{a\leq t\leq b} |h_k(t)|||\boldsymbol{e}_0||_A \leq 2\left(\frac{\sqrt{b}-\sqrt{a}}{\sqrt{b}+\sqrt{a}}\right)^k ||\boldsymbol{e}_0||_A. \tag{60}$$

□

6. Métodos basados en subespacios de Krylov [8]

En el método del gradiente conjugado, se mostró que el vector residual $\boldsymbol{r}_k$ en el paso k pertenece al subespacio $span\{\boldsymbol{r}_0, A\boldsymbol{r}_0, \ldots, A^k\boldsymbol{r}_0\}$, es decir, podemos escribir $\boldsymbol{r}_k$ como

$$\boldsymbol{r}_k = q_k(A)\boldsymbol{r}_0,$$

donde q_k es un polinomio de grado k con $q_k(0) = 1$. Por otra parte (ver (52)),

$$span\{\boldsymbol{p}_0, \boldsymbol{p}_1, \ldots, \boldsymbol{p}_k\} = span\{\boldsymbol{r}_0, A\boldsymbol{r}_0, \ldots, A^k\boldsymbol{r}_0\}.$$

Así, para $k \geq 0$ tenemos que $\boldsymbol{p}_k = g_k(A)\boldsymbol{r}_0$, donde g_k es un polinomio de grado k. Luego, la iteración en el paso k del método del gradiente conjugado satisface

$$\begin{aligned}
\boldsymbol{x}_k &= \boldsymbol{x}_{k-1} + \alpha_{k-1}\boldsymbol{p}_{k-1} \\
&= \boldsymbol{x}_{k-2} + \alpha_{k-2}\boldsymbol{p}_{k-2} + \alpha_{k-1}g_{k-1}(A)\boldsymbol{r}_0 \\
&= \boldsymbol{x}_{k-2} + \alpha_{k-2}g_{k-2}(A)\boldsymbol{r}_0 + \alpha_{k-1}g_{k-1}(A)\boldsymbol{r}_0 \\
&\quad \vdots \\
&= \boldsymbol{x}_0 + \alpha_0\boldsymbol{p}_0 + \alpha_1 g_1(A)\boldsymbol{r}_0 + \cdots + \alpha_{k-1}g_{k-1}(A)\boldsymbol{r}_0 \\
&= \boldsymbol{x}_0 + \sum_{m=0}^{k-1} \alpha_m g_m(A)\boldsymbol{r}_0.
\end{aligned}$$

Por lo tanto, $\boldsymbol{x}_k - \boldsymbol{x}_0 \in span\{\boldsymbol{r}_0, A\boldsymbol{r}_0, \ldots, A^{k-1}\boldsymbol{r}_0\}$. Esto significa que buscamos la aproximación $\boldsymbol{x}_k$ en el subespacio afín

$$\boldsymbol{x}_0 + span\{\boldsymbol{r}_0, A\boldsymbol{r}_0, \ldots, A^{k-1}\boldsymbol{r}_0\}$$

Este hecho revela la importancia de los subespacios de la forma

$$\mathcal{K}_m(A, \boldsymbol{v}) := span\{\boldsymbol{v}, A\boldsymbol{v}, A^2\boldsymbol{v}, \ldots, A^{m-1}\boldsymbol{v}\}, \tag{61}$$

los cuales reciben el nombre de *subespacios de Krylov*[1] *de orden* m. Cuando no haya lugar a confusión se usará la notación $\mathcal{K}_m$ en vez de $\mathcal{K}_m(A, \boldsymbol{v})$. En general, un *método basado en subespacios de Krylov* es aquel que busca una aproximación $\boldsymbol{x}_m$ de la solución del sistema $A\boldsymbol{x} = \boldsymbol{b}$ en el subespacio afín $\boldsymbol{x}_0 + \mathcal{K}_m(A, \boldsymbol{r}_0)$. Nótese que si $\boldsymbol{x} \in \mathcal{K}_m(A, \boldsymbol{v})$, entonces $\boldsymbol{x} = p(A)\boldsymbol{v}$, donde p es un polinomio de grado menor o igual que $m - 1$. Antes de presentar otras propiedades de los subespacios de Krylov, recordamos que el *polinomio*

[1] Aleksei Nikolaevich Krylov (1863-1945)

mínimo de un vector $\boldsymbol{v} \in \mathbb{R}^n$ con respecto a la matriz $A \in \mathbb{R}^{n\times n}$, es el polinomio mónico $p(t)$ de menor grado tal que $p(A)\boldsymbol{v} = \boldsymbol{0}$. Por el teorema de Cayley-Hamilton, la matriz A es un cero de su polinomio característico, en consecuencia, el grado del polinomio mínimo no puede ser mayor que n. Con ayuda del polinomio mínimo se puede ver también por qué los subespacios de Krylov aparecen de manera natural en la solución del sistemas de ecuaciones lineales $A\boldsymbol{x} = \boldsymbol{b}$. En efecto, para $\boldsymbol{b}$ distinto de cero y A no singular, sea $p(t) = t^m + \alpha_{m-1}t^{m-1} + \cdots + \alpha_1 t + \alpha_0$ el polinomio mínimo de $\boldsymbol{b}$ con respecto a A. Entonces

$$A^m\boldsymbol{b} + \alpha_{m-1}A^{m-1}\boldsymbol{b} + \cdots + \alpha_1 A\boldsymbol{b} + \alpha_0\boldsymbol{b} = \boldsymbol{0}.$$

Puesto que $\alpha_0 \neq 0$ (en caso contrario $q(A)\boldsymbol{v} = \boldsymbol{0}$ para el polinomio $q(t) = p(t)/t$ cuyo grado es menor que m, lo cual contradice que p sea mínimo), de la última ecuación se obtiene

$$A\left(c_{m-1}A^{m-1}\boldsymbol{b} + c_{m-2}A^{m-2}\boldsymbol{b} + \cdots + c_1\boldsymbol{b}\right) = \boldsymbol{b},$$

donde $c_i = -\alpha_i/\alpha_0$ para $i = 1, \ldots, m-1$. Por ende, la solución del sistema está en el subespacio de Krylov $\mathcal{K}_m(A, \boldsymbol{b})$.

Se presentan ahora algunas propiedades básicas de los subespacios de Krylov, la primera de ellas es la invariancia bajo A.

Proposición 2.1. Sea m el grado del polinomio mínimo de $\boldsymbol{v}$ con respecto a A. Entonces $A\mathcal{K}_m(A, \boldsymbol{v}) \subseteq \mathcal{K}_m(A, \boldsymbol{v})$, es decir, $\mathcal{K}_m(A, \boldsymbol{v})$ es invariante bajo A. Además,

$$\mathcal{K}_l(A, \boldsymbol{v}) = \mathcal{K}_m(A, \boldsymbol{v}) \qquad \forall\, l \geq m. \tag{62}$$

Demostración.

Sea $p(t) = t^m + \alpha_{m-1}t^{m-1} + \cdots + \alpha_1 t + \alpha_0$ el polinomio mínimo de $\boldsymbol{v}$ con respecto a A. Sea $\boldsymbol{w} \in A\mathcal{K}_m(A, \boldsymbol{v})$, entonces

$$\boldsymbol{w} = \sum_{i=1}^{m-1} \beta_i A^i\boldsymbol{v} + \beta_m A^m\boldsymbol{v}.$$

Como $\sum_{i=1}^{m-1} \beta_i A^i\boldsymbol{v} \in \mathcal{K}_m(A, \boldsymbol{v})$, será suficiente probar que $A^m\boldsymbol{v} \in \mathcal{K}_m(A, \boldsymbol{v})$. En efecto, de la igualdad $p(A)\boldsymbol{v} = \boldsymbol{0}$ se sigue que $A^m\boldsymbol{v} + \sum_{i=0}^{m-1} \alpha_i A^i\boldsymbol{v} = \boldsymbol{0} \in \mathcal{K}_m$ o bien, $A^m\boldsymbol{v} = -\sum_{i=0}^{m-1} \alpha_i A^i\boldsymbol{v} \in \mathcal{K}_m(A, \boldsymbol{v})$, como se deseaba. A partir de lo obtenido en la primera parte de la prueba, se deduce fácilmente que $A^k\mathcal{K}_m \subseteq \mathcal{K}_m(A, \boldsymbol{v})$ para todo $k \geq 1$. Luego, para $l \geq m$, $A^m\boldsymbol{v}, A^{m+1}\boldsymbol{v}, \ldots, A^{l-1}\boldsymbol{v} \in \mathcal{K}_m(A, \boldsymbol{v})$, lo cual da (62). □

El siguiente teorema indica como determinar la dimensión de un subespacio de Krylov.

Teorema 2.17. Sea μ el grado del polinomio mínimo de $\boldsymbol{v}$ con respecto a la matriz A. Entonces

$$\dim \mathcal{K}_m(A, \boldsymbol{v}) = \text{mín}\{m,\, \mu\}$$

Demostración.

Suponga que $\mu \geq m$ y sean $\beta_1, \beta_2, \ldots, \beta_{m-1}$ escalares tales que $\sum_{i=0}^{m-1} \beta_i A^i \boldsymbol{v} = \boldsymbol{0}$. Entonces $\beta_1 = \beta_2 = \cdots = \beta_{m-1} = 0$, pues en caso contrario, se tendría que $q(A)\boldsymbol{v} = \boldsymbol{0}$ para el polinomio no nulo $q(t) = \sum_{i=0}^{m-1} \beta_i t^i$ lo cual contradice que el grado de $\boldsymbol{v}$ con respecto a A sea μ. Así, el conjunto $\{\boldsymbol{v}, A\boldsymbol{v}, \ldots, A^{m-1}\boldsymbol{v}\}$ es linealmente independiente y además, por definición, genera a $\mathcal{K}_m(A, \boldsymbol{v})$. Por lo tanto, $\dim \mathcal{K}_m(A, \boldsymbol{v}) = m$.

Ahora, si $\mu \leq m$, por la proposición anterior se tiene que $\mathcal{K}_m(A, \boldsymbol{v}) = \mathcal{K}_\mu(A, \boldsymbol{v})$. Además, por la definición de polinomio mínimo, el conjunto $\{\boldsymbol{v}, A\boldsymbol{v}, \ldots, A^{\mu-1}\boldsymbol{v}\}$ es linealmente independiente y claramente genera a $\mathcal{K}_\mu(A, \boldsymbol{v})$. Luego, $\dim \mathcal{K}_m(A, \boldsymbol{v}) = \dim \mathcal{K}_\mu(A, \boldsymbol{v}) = \mu$. En cualquier caso, se tiene que $\dim \mathcal{K}_m(A, \boldsymbol{v}) = \text{mín}\{m, \mu\}$. □

El procedimiento que se presenta a continuación, llamado algoritmo de Arnoldi[2], permite construir una base ortonormal para los subespacios de Krylov asumiendo aritmética exacta. Compárese dicho algoritmo con el de Gram-Schmidt.

Algoritmo (Algoritmo de ortogonalización de Arnoldi)**.**

Inicie con un vector no nulo $\boldsymbol{z} \in \mathbb{R}^n$ y defina $\boldsymbol{q}_1 = \boldsymbol{z}/||\boldsymbol{z}||_2$

for $j = 1, 2, ..., m$
  for $i = 1, 2, ..., j$
    $h_{ij} = \langle A\boldsymbol{q}_j, \boldsymbol{q}_i \rangle$
  end
  $\boldsymbol{w}_j = A\boldsymbol{q}_j - \sum_{i=1}^{j} h_{ij}\boldsymbol{q}_i$
  $h_{j+1,j} = ||\boldsymbol{w}_j||_2$
  if $h_{j+1,j} = 0$**, then stop**
  $\boldsymbol{q}_{j+1} = \boldsymbol{w}_j / h_{j+1,j}$
end

[2]Walter Edwin Arnoldi (1917-1995)

Si el algoritmo no se detiene antes del paso m, entonces resulta

$$A\boldsymbol{q}_j = \sum_{i=1}^{j+1} h_{ij}\boldsymbol{q}_i, \quad j = 1, 2, \ldots, m, \tag{63}$$

o equivalentemente, en forma matricial

$$\begin{aligned} AQ_m &= Q_{m+1} \begin{pmatrix} h_{11} & h_{12} & h_{13} & \cdots & h_{1,m-1} & h_{1m} \\ h_{21} & h_{22} & h_{23} & \cdots & h_{2,m-1} & h_{2m} \\ 0 & h_{32} & h_{33} & \cdots & h_{3,m-1} & h_{3m} \\ 0 & 0 & h_{43} & \ddots & h_{4,m-1} & h_{4m} \\ \vdots & \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & h_{m,m-1} & h_{mm} \\ 0 & 0 & 0 & \cdots & 0 & h_{m+1,m} \end{pmatrix} \\ &= Q_{m+1} \begin{pmatrix} H_m \\ h_{m+1,m}\boldsymbol{e}_m^T \end{pmatrix}, \qquad (64) \\ &= \left(Q_m | \boldsymbol{q}_{m+1}\right) \begin{pmatrix} H_m \\ h_{m+1,m}\boldsymbol{e}_m^T \end{pmatrix} \\ &= Q_m H_m + h_{m+1,m}\boldsymbol{q}_{m+1}\boldsymbol{e}_m^T, \qquad (65) \end{aligned}$$

donde Q_m es la matriz $n \times m$ cuyas columnas son los vectores ortonormales $\boldsymbol{q}_1, \boldsymbol{q}_2, \ldots, \boldsymbol{q}_m$ y H_m es la llamada *matriz superior de Hessenberg* $m \times m$, la cual se se caracteriza porque sus entradas debajo de la primera subdiagonal son nulas.

Ahora, pre-multiplicado ambos lados de (65) por la matriz Q_m^T y usando el hecho de que los vectores $\boldsymbol{q}_1, \boldsymbol{q}_2, \ldots, \boldsymbol{q}_m, \boldsymbol{q}_{m+1}$ son ortogonales, resulta

$$Q_m^T A Q_m = H_m. \tag{66}$$

Veamos que el conjunto $\{\boldsymbol{q}_1, \boldsymbol{q}_2, \ldots, \boldsymbol{q}_m\}$ forma una base de $\mathcal{K}_m(A, \boldsymbol{q}_1)$. Puesto que dicho conjunto es ortonormal, entonces es linealmente independiente. Ahora, para cada $j = 1, 2 \ldots, m$ se afirma que $\boldsymbol{q}_j = \theta_{j-1}(A)\boldsymbol{q}_1$, siendo θ_{j-1} un polinomio de grado $j-1$. Probemos esto último por inducción. Para $j = 1$, es claro que $\boldsymbol{q}_1 = \theta_0(A)\boldsymbol{q}_1$ con $\theta_0(t) \equiv 1$. Suponga que el resultado se tiene para los enteros positivos menores o iguales que j. Del algoritmo de Arnoldi y la hipótesis de inducción, obtenemos

$$\boldsymbol{q}_{j+1} = \frac{1}{h_{j+1,j}}\boldsymbol{w}_j = \frac{1}{h_{j+1,j}} \left(A\boldsymbol{q}_j - \sum_{i=1}^{j} h_{ij}\boldsymbol{q}_i \right)$$

$$= \frac{1}{h_{j+1,j}} \left(A\theta_{j-1}(A)\boldsymbol{q}_1 - \sum_{i=1}^{j} h_{ij}\theta_{i-1}(A)\boldsymbol{q}_1 \right)$$
$$= \theta_j(A)\boldsymbol{q}_1,$$

con

$$\theta_j(t) = \frac{t\theta_{j-1}(t)\boldsymbol{q}_1 - \sum_{i=1}^{j} h_{ij}\theta_{i-1}(t)\boldsymbol{q}_1}{h_{j+1,j}}.$$

El hecho de que $\boldsymbol{q}_j = \theta_{j-1}(A)\boldsymbol{q}_1$ implica que el conjunto $\{\boldsymbol{q}_1, \boldsymbol{q}_2, \ldots, \boldsymbol{q}_m\}$ genera a $\mathcal{K}_m(A, \boldsymbol{q}_1)$. En efecto, como $\boldsymbol{q}_2 = \theta_1(A)\boldsymbol{q}_1 = \gamma_{0,2}\boldsymbol{q}_1 + \gamma_{1,2}A\boldsymbol{q}_1$ y $\gamma_{1,2} \neq 0$ (pues $\boldsymbol{q}_1$ y $\boldsymbol{q}_2$ son linealmente independientes), entonces $A\boldsymbol{q}_1 = (\boldsymbol{q}_2 - \gamma_{0,2}\boldsymbol{q}_1)/\gamma_{1,2}$. En forma similar, como $\boldsymbol{q}_3 = \theta_2(A)\boldsymbol{q}_1 = \gamma_{0,3}\boldsymbol{q}_1 + \gamma_{1,3}A\boldsymbol{q}_1 + \gamma_{2,3}A^2\boldsymbol{q}_1$ y $\gamma_{2,3} \neq 0$, entonces

$$A^2\boldsymbol{q}_1 = \frac{1}{\gamma_{2,3}} \left(\boldsymbol{q}_3 - \gamma_{0,3}\boldsymbol{q}_1 - \frac{\gamma_{1,3}}{\gamma_{1,2}} (\boldsymbol{q}_2 - \gamma_{0,2}\boldsymbol{q}_1) \right).$$

Procediendo de esta forma, se deduce que para cada $i = 1, 2, \ldots, m-1$, $A^i\boldsymbol{q}_1 \in span\{\boldsymbol{q}_1, \boldsymbol{q}_2, \ldots, \boldsymbol{q}_m\}$. Por lo tanto, $\{\boldsymbol{q}_1, \boldsymbol{q}_2, \ldots, \boldsymbol{q}_m\}$ genera a $\mathcal{K}_m(A, \boldsymbol{q}_1)$. En síntesis, si el algoritmo de Arnoldi no se detiene antes del paso m, entonces el conjunto $\{\boldsymbol{q}_1, \boldsymbol{q}_2, \ldots, \boldsymbol{q}_m\}$ forma una base ortonormal del subespacio de Krylov

$$\mathcal{K}_m(A, \boldsymbol{q}_1) = span\{\boldsymbol{q}_1, A\boldsymbol{q}_1, \ldots, A^{m-1}\boldsymbol{q}_1\}.$$

Usando el grado del polinomio mínimo de $\boldsymbol{q}_1$ con respecto a A, se puede determinar en cual paso se detiene el agoritmo de Arnoldi.

Proposición 2.2. El algoritmo de Arnoldi se detiene en el paso j, es decir, $h_{j+1,j} = 0$ si y solo si el grado del polinomio mínimo de $\boldsymbol{q}_1$ es j.

Demostración. Denotemos por μ el grado del polinomio mínimo de $\boldsymbol{q}_1$ con respecto a A y suponga que $\mu = j$. Se probará por contradicción que $h_{j+1,j} = 0$. Si $h_{j+1,j} \neq 0$, entonces se puede definir $\boldsymbol{q}_{j+1}$. Luego, $\{\boldsymbol{q}_1, \boldsymbol{q}_2, \ldots, \boldsymbol{q}_{j+1}\}$ forma una base de $\mathcal{K}_{j+1}(A, \boldsymbol{q}_1)$. Así, $\dim \mathcal{K}_{j+1}(A, \boldsymbol{q}_1) = j + 1$. Pero como $\mu = j$, por el teorema 2.17, $\dim \mathcal{K}_{j+1}(A, \boldsymbol{q}_1) = \text{mín}\{\mu, j + 1\} = j$, lo cual da una contradicción.

Suponga ahora que $h_{j+1,j} = 0$ o equivalentemente, $\boldsymbol{w}_j = \boldsymbol{0}$. Usando nuevamente el teorema 2.17, se deduce que μ no puede ser mayor que j pues, en tal caso se tendría que $\dim \mathcal{K}_{j+1}(A, \boldsymbol{q}_1) = \text{mín}\{\mu, j + 1\} = j + 1$, lo cual es una contradicción. Tampoco es posible que μ sea menor que j, pues eso implicaría

(por la primera parte de la prueba) que el algoritmo se detiene en el paso μ anterior al paso j. En consecuencia, $\mu = j$, lo cual completa la prueba.

□

Para evitar errores de redondeo y obtener un algoritmo más estable es conveniente considerar la siguiente modificación del algoritmo de Arnoldi (compárela con el algoritmo modificado de Gram-Schmidt):

Algoritmo (Algoritmo de Arnoldi modificado)**.**

Inicie con un vector no nulo $\boldsymbol{z} \in \mathbb{R}^n$ y defina $\boldsymbol{q}_1 = \boldsymbol{z}/||\boldsymbol{z}||_2$

for $j = 1, 2, ..., m$
  $\boldsymbol{w}_j = A\boldsymbol{q}_j$
  for $i = 1, 2, ..., j$
    $h_{ij} = \langle \boldsymbol{w}_j, \boldsymbol{q}_i \rangle$
    $\boldsymbol{w}_j = \boldsymbol{w}_j - h_{ij}\boldsymbol{q}_i$
  end
  $h_{j+1,j} = ||\boldsymbol{w}_j||_2$
  if $h_{j+1,j} = 0$**, then stop**
  $\boldsymbol{q}_{j+1} = \boldsymbol{w}_j/h_{j+1,j}$
end

Cabe mencionar que existen otras alternativas para mejorar el algoritmo de Arnoldi, por ejemplo, considerando los reflectores de Householder (para más detalles se remite a la referencia [8]).

Veamos ahora cómo usar la base ortonormal generada por el algoritmo de Arnoldi para buscar una aproximación $\boldsymbol{x}_m$ de la solución del sistema $A\boldsymbol{x} = \boldsymbol{b}$ en el subespacio afín $\boldsymbol{x}_0 + \mathcal{K}_m(A, \boldsymbol{r}_0)$ siendo $\boldsymbol{r}_0 = \boldsymbol{b} - A\boldsymbol{x}_0$ y $\boldsymbol{x}_0$ un dato inicial. Concretamente, suponga que se han realizado m pasos del Algoritmo de Arnoldi de manera que se dispone de la base ortonormal $\{\boldsymbol{q}_1, \boldsymbol{q}_2, \ldots, \boldsymbol{q}_m\}$ de $\mathcal{K}_m(A, \boldsymbol{r}_0)$ con $\boldsymbol{q}_1 = \boldsymbol{r}_0/||\boldsymbol{r}_0||_2$. Así, cualquier elemento de $\mathcal{K}(A, \boldsymbol{r}_0)$ puede escribirse como $\sum_{l=1}^{m} \alpha_l \boldsymbol{q}_l = Q_m \boldsymbol{y}_m$, donde $\boldsymbol{y}_m = (\alpha_1, \alpha_2, \ldots, \alpha_m)^T$. Luego, cada nueva iteración $\boldsymbol{x}_m$ se puede escribir como

$$\boldsymbol{x}_m = \boldsymbol{x}_0 + Q_m \boldsymbol{y}_m. \tag{67}$$

Para hallar $\boldsymbol{y}_m$ se impone la condición de ortogonalidad o condición de Petrov-Galerkin (ver fig. 6)

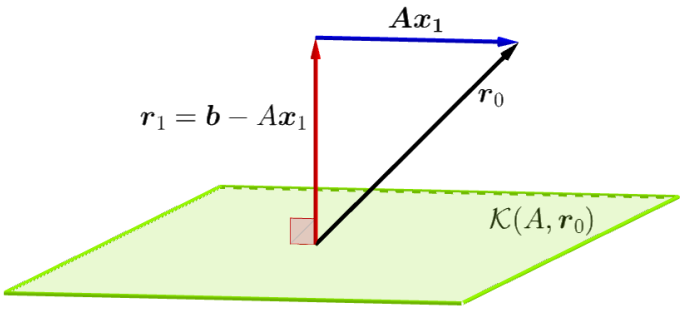

Figura 6: Ilustración de la condición de ortogonalidad.

$$\boldsymbol{r}_m = \boldsymbol{b} - A\boldsymbol{x}_m \perp \mathcal{K}_m(A, \boldsymbol{r}_0).$$

Así, usando la relación (66), se obtiene

$$\begin{aligned} \boldsymbol{0} = Q_m^T \boldsymbol{r}_m &= Q_m^T(\boldsymbol{b} - A\boldsymbol{x}_m) \\ &= Q_m^T(\boldsymbol{b} - A\boldsymbol{x}_0 - AQ_m\boldsymbol{y}_m) \\ &= Q_m^T(\boldsymbol{r}_0 - AQ_m\boldsymbol{y}_m) \\ &= ||\boldsymbol{r}_0||_2 Q_m^T\boldsymbol{q}_1 - Q_m^T AQ_m\boldsymbol{y}_m \\ &= ||\boldsymbol{r}_0||_2\boldsymbol{e}_1 - H_m\boldsymbol{y}_m. \end{aligned}$$

Por lo tanto,

$$H_m\boldsymbol{y}_m = ||\boldsymbol{r}_0||_2\boldsymbol{e}_1. \tag{68}$$

Puesto que H_m es una matriz de Hessenberg, el sistema lineal (68) se puede resolver fácilmente y una vez se tenga $\boldsymbol{y}_m$, se calcula $\boldsymbol{x}_m$ usando (67).

El método descrito recibe el nombre de Método de Ortogonalización Completa y se describe en el siguiente algoritmo.

Algoritmo (Método de Ortogonalización Completa)**.**

$\boldsymbol{r}_0 = \boldsymbol{b} - A\boldsymbol{x}_0$, $\boldsymbol{q}_1 = \boldsymbol{r}_0/||\boldsymbol{r}_0||_2$

for $j = 1, 2, ..., m$

$\boldsymbol{w}_j = A\boldsymbol{q}_j$

for $i = 1, 2, ..., j$
$h_{ij} = \langle \boldsymbol{w}_j, \boldsymbol{q}_i \rangle$
$\boldsymbol{w}_j = \boldsymbol{w}_j - h_{ij}\boldsymbol{q}_i$
end
$h_{j+1,j} = ||\boldsymbol{w}_j||_2$
if $h_{j+1,j} = 0$**, set** $m = j$ **and Goto line 12**
$\boldsymbol{q}_{j+1} = \boldsymbol{w}_j / h_{j+1,j}$
end

12. $\boldsymbol{y}_m = H_m^{-1}(||\boldsymbol{r}_0||_2 \boldsymbol{e}_1), \quad \boldsymbol{x}_m = \boldsymbol{x}_0 + Q_m \boldsymbol{y}_m$

La siguiente proposición permite calcular de una forma económica, desde el punto de vista computacional, la norma del vector residual sin necesidad de calcular explícitamente la solución $\boldsymbol{x}_m$. Esta información es útil para establecer un criterio de detención del algoritmo.

Proposición 2.3. El vector residual $\boldsymbol{r}_m = \boldsymbol{b} - A\boldsymbol{x}_m$ calculado con el método de ortogonalización completa satisface

$$||\boldsymbol{b} - A\boldsymbol{x}_m||_2 = h_{m+1,m}|\boldsymbol{e}_1^T \boldsymbol{y}_m|. \tag{69}$$

Demostración.

Usando las igualdades (65)-(68) y teniendo en cuenta que $Q_m \boldsymbol{e}_1 = \boldsymbol{q}_1$, resulta

$$\begin{aligned} \boldsymbol{b} - A\boldsymbol{x}_m &= \boldsymbol{b} - A(\boldsymbol{x}_0 + Q_m \boldsymbol{y}_m) \\ &= \boldsymbol{r}_0 - AQ_m \boldsymbol{y}_m \\ &= \boldsymbol{r}_0 - Q_m H_m \boldsymbol{y}_m - h_{m+1,m}\boldsymbol{q}_{m+1}\boldsymbol{e}_m^T \boldsymbol{y}_m \\ &= ||\boldsymbol{r}_0||_2 \boldsymbol{q}_1 - Q_m ||\boldsymbol{r}_0||_2 \boldsymbol{e}_1 - h_{m+1,m}\boldsymbol{q}_{m+1}\boldsymbol{e}_m^T \boldsymbol{y}_m \\ &= -h_{m+1,m}\boldsymbol{q}_{m+1}\boldsymbol{e}_m^T \boldsymbol{y}_m. \end{aligned}$$

Tomando la norma y teniendo presente que $||\boldsymbol{q}_{m+1}||_2 = 1$, se obtiene el resultado deseado. □

De acuerdo a la proposición anterior, dada una tolerencia $\epsilon > 0$, podemos decidir detener el algoritmo cuando

$$\frac{h_{m+1,m}|\boldsymbol{e}_1^T \boldsymbol{y}_m|}{||\boldsymbol{r}_0||_2} \leq \epsilon.$$

Es importante señalar que el método de ortogonalización completa es viable para valores pequeños de m debido a sus requerimientos de memoria y a la acumulación de errores de redondeo. Sin embargo, existen modificaciones del método que permiten soslayar estas dificultades, una descripción completa de dichas alternativas se puede encontrar en [8].

7. Método de residuos mínimos generalizado (GMRES)

Este método se caracteriza por la selección de la solución aproximada $\boldsymbol{x}_m$ de manera que se minimice la norma Euclidiana del vector residual en el paso m. Es decir, se busca $\boldsymbol{x}_m \in \boldsymbol{x}_0 + \mathcal{K}_m(A, \boldsymbol{r}_0)$ de manera que

$$||\boldsymbol{r}_m||_2 = ||\boldsymbol{b} - A\boldsymbol{x}_m||_2 \leq ||\boldsymbol{b} - A\boldsymbol{x}||_2 \quad \forall\, \boldsymbol{x} \in \boldsymbol{x}_0 + \mathcal{K}_m(A, \boldsymbol{r}_0).$$

Definiendo

$$\widetilde{H}_m = \begin{pmatrix} H_m \\ h_{m+1,m}\boldsymbol{e}_m^T \end{pmatrix}$$

y usando las relaciones (64)-(67) obtenemos

$$\begin{aligned} ||\boldsymbol{b} - A\boldsymbol{x}||_2 &= ||\boldsymbol{b} - A(\boldsymbol{x}_0 + Q_m\boldsymbol{y})||_2 \\ &= ||\boldsymbol{r}_0 - AQ_m\boldsymbol{y}||_2 \\ &= ||\,||\boldsymbol{r}_0||_2\boldsymbol{q}_1 - Q_{m+1}\widetilde{H}_m\boldsymbol{y}||_2 \\ &= ||\,||\boldsymbol{r}_0||_2 Q_{m+1}\boldsymbol{e}_1 - Q_{m+1}\widetilde{H}_m\boldsymbol{y}||_2 \\ &= ||Q_{m+1}(||\boldsymbol{r}_0||_2\boldsymbol{e}_1 - \widetilde{H}_m\boldsymbol{y})||_2 \\ &= ||\,||\boldsymbol{r}_0||_2\boldsymbol{e}_1 - \widetilde{H}_m\boldsymbol{y}||_2, \end{aligned}$$

donde en el último paso se usa el hecho de que las columnas de Q_{m+1} son ortonormales. Luego, la aproximación con el método GMRES es el único vector $\boldsymbol{x}_m = \boldsymbol{x}_0 + Q_m\boldsymbol{y}_m$, donde $\boldsymbol{y}_m \in \mathbb{R}^m$ es el vector que minimiza

$$||\,||\boldsymbol{r}_0||_2\boldsymbol{e}_1 - \widetilde{H}_m\boldsymbol{y}||_2.$$

Este problema de minimización se puede resolver usando la factorización QR. De hecho, la factorización QR se puede calcular eficientemente puesto que H_m es una matriz de Hessenberg.

Algoritmo (GMRES Básico)**.**

$\boldsymbol{r}_0 = \boldsymbol{b} - A\boldsymbol{x}_0$, $\boldsymbol{q}_1 = \boldsymbol{r}_0/||\boldsymbol{r}_0||_2$

for $j = 1, 2, ..., m$

$\boldsymbol{w}_j = A\boldsymbol{q}_j$

for $i = 1, 2, ..., j$

$h_{ij} = \langle \boldsymbol{w}_j, \boldsymbol{q}_i \rangle$

$\boldsymbol{w}_j = \boldsymbol{w}_j - h_{ij}\boldsymbol{q}_i$

end

$h_{j+1,j} = ||\boldsymbol{w}_j||_2$

if $h_{j+1,j} = 0$**, set** $m = j$ **and Goto line 12**

$\boldsymbol{q}_{j+1} = \boldsymbol{w}_j / h_{j+1,j}$

end

12. Defina $\widetilde{H}_m = \{h_{ij}\}_{1 \leq 1 \leq m+1, 1 \leq j \leq m}$ y calcule el minimizador $\boldsymbol{y}_m$ de $||||\boldsymbol{r}_0||_2\boldsymbol{e}_1 - \widetilde{H}_m\boldsymbol{y}||_2$ y $\boldsymbol{x}_m = \boldsymbol{x}_0 + Q_m\boldsymbol{y}_m$.

Bibliografía

[1] K. Atkinson and W. Han. *Theoretical Numerical Analysis*, Springer, New York, Third Edition, 2009.

[2] R. Bellman. *Introducción al Análisis Matricial*, Editorial Reverté, 1965.

[3] R. Horn and C. Johnson. *Matrix Analysis*, Cambridge University Press, 1985.

[4] R. Kress. *Numerical Analysis*, Springer, New York, 1998.

[5] : Kurmanbek, Y. Erlangga, Y. Amanbek. *Explicit inverse of near Toeplitz pentadiagonal matrices related to higher order difference operators.* Results in Applied Mathematics, vol. 11 (2021).

[6] D. Moursund and C. Duris *Elementary Theory and Application of Numerical Analysis*, McGraw-Hill, New York, 1967.

[7] A. Quarteroni, R. Sacco and F. Saleri. *Numerical Mathematics*, Springer, Second Edition,2007.

[8] Y. Saad. *iterative Methods for Sparse Linear Systems*, SIAM, Second Edition, 2003.

[9] J. C. Strikwerda *Finite Difference Schemes and Partial differential Equations*, SIAM, Second Edition, 2004.