Análisis Numérico II

Apuntes
Curso Código 525441
Primer Semestre 2011

Raimund Bürger & Rommel Bustinza
Centro de Investigación en Ingeniería Matemática (CI^{2}MA)
& Departamento de Ingeniería Matemática
Facultad de Ciencias Físicas y Matemáticas
Universidad de Concepción
Casilla 160-C
Concepción, Chile

23 de diciembre de 2011

Índice general

Capítulo 1

Conceptos básicos

Designaremos por $\mathbb{C}^{n\times n}$ el espacio de las matrices cuadradas de orden n en el cuerpo $\mathbb{C}$, mientras que cuando los coeficientes pertenezcan al cuerpo $\mathbb{R}$, usaremos la notación $\mathbb{R}^{n\times n}$.

Definición 1.1. *Para* $\mathbf{A} \in \mathbb{C}^{n\times n}$ *se definen la* matriz transpuesta *como la matriz* $\mathbf{B}$ *de elementos* $b_{ij} := a_{ji}$*, y la* matriz conjugada transpuesta *de* $\mathbf{A}$ *como la matriz* $\mathbf{C}$ *de elementos* $c_{ij} := \bar{a}_{ji}$*. Notación:* $\mathbf{A}^{\mathrm{T}} := \mathbf{B}$ *y* $\mathbf{A}^* := \mathbf{C}$*.*

Definición 1.2. *Una matriz* $\mathbf{A} \in \mathbb{C}^{n\times n}$ *se dice* simétrica *si* $\mathbf{A} = \mathbf{A}^{\mathrm{T}}$*,* hermitiana *si* $\mathbf{A} = \mathbf{A}^*$*,* ortogonal *si*

$$\mathbf{A}\mathbf{A}^{\mathrm{T}} = \mathbf{A}^{\mathrm{T}}\mathbf{A} = \mathbf{I}$$

y unitaria *si*

$$\mathbf{A}\mathbf{A}^* = \mathbf{A}^*\mathbf{A} = \mathbf{I}.$$

Una manera de caracterizar la ortogonalidad, respectivamente la unitariedad, de una matriz $\mathbf{A}$ es a través de las igualdades $\mathbf{A}^{-1} = \mathbf{A}^{\mathrm{T}}$ y $\mathbf{A}^{-1} = \mathbf{A}^*$, respectivamente.

Definición 1.3. *Un escalar* $\lambda \in \mathbb{C}$ *se dice un* valor propio *de una matriz* $\mathbf{A} \in \mathbb{C}^{n\times n}$ *si existe* $\mathbf{x} \in \mathbb{C}^n$*,* $\mathbf{x} \neq 0$ *tal que*

$$\mathbf{A}\mathbf{x} = \lambda\mathbf{x}. \tag{1.1}$$

En tal caso, el vector $\mathbf{x}$ *se llama* vector propio *de* $\mathbf{A}$ *asociado a* λ*.*

Definición 1.4. *Sea* $\lambda \in \mathbb{C}$ *un valor propio de* $\mathbf{A}$*. Se llama* espacio propio *asociado a* λ *al conjunto*

$$L(\lambda) := \{\mathbf{x} \in \mathbb{C}^n \mid \mathbf{A}\mathbf{x} = \lambda\mathbf{x}\}.$$

Note que $L(\lambda)$ contiene, además del vector nulo, a todos los vectores propios asociados a λ. Se puede demostrar que $L(\lambda)$ es un subespacio vectorial de $\mathbb{C}^n$ con *dimensión* $\varrho(\lambda)$ dada por

$$\varrho(\lambda) = n - \operatorname{rango}(\mathbf{A} - \lambda\mathbf{I}). \tag{1.2}$$

El número $\varrho(\lambda)$ se llama también *multiplicidad geométrica* de λ.

Lema 1.1. *Sea* $\mathbf{A} \in \mathbb{C}^{n\times n}$*. Un escalar* $\lambda \in \mathbb{C}$ *es un valor propio de* $\mathbf{A}$ *si y sólo si*

$$\det(\mathbf{A} - \lambda\mathbf{I}) = 0.$$

Demostración. De acuerdo a la Definición 1.3, un escalar $\lambda \in \mathbb{C}$ es un valor propio de $\mathbf{A}$ si y sólo si existe $\mathbf{x} \in \mathbb{C}^n$, $\mathbf{x} \neq 0$ tal que $\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$, equivalentemente, si y sólo si $(\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = 0$ con $\mathbf{x} \neq 0$. Esta última relación es un sistema lineal de ecuaciones lineales homogéneo de n ecuaciones y n incógnitas. Para no obtener únicamente la solución trivial $\mathbf{x} = 0$, que no nos interesa, imponemos la condición necesaria y suficiente $\det(\mathbf{A} - \lambda\mathbf{I}) = 0$. ■

La expresión $f_{\mathbf{A}}(\lambda) := \det(\mathbf{A} - \lambda\mathbf{I})$ es un polinomio de grado n que se llama *polinomio característico* de $\mathbf{A}$, y tiene la forma

$$f_{\mathbf{A}}(\lambda) = (-1)^n(\lambda^n + \alpha_{n-1}\lambda^{n-1} + \cdots + \alpha_1\lambda + \alpha_0). \tag{1.3}$$

Si $\lambda_1, \lambda_2, \ldots, \lambda_k$ son los ceros del polinomio característico, entonces $f_{\mathbf{A}}(\lambda)$ puede factorizarse como

$$f_{\mathbf{A}}(\lambda) = (-1)^n(\lambda - \lambda_1)^{\beta_1}(\lambda - \lambda_2)^{\beta_2} \cdot \ldots \cdot (\lambda - \lambda_k)^{\beta_k}, \tag{1.4}$$

donde $\beta_1, \ldots, \beta_k$ son números naturales tales que

$$\beta_1 + \cdots + \beta_k = n.$$

El número β_i, $i = 1, \ldots, k$ de veces que se repite el factor $(\lambda - \lambda_i)$ se llama *multiplicidad algebraica* de λ_i. Al valor propio λ_i pueden corresponder a lo más β_i vectores propios linealmente independientes. El número de vectores propios de $\mathbf{A}$ asociados al valor propio λ_i, y que son linealmente independientes, es igual a $\varrho(\lambda_i)$. En otras palabras, se tiene que

$$\varrho(\lambda_i) \leqslant \beta_i, \quad i = 1, \ldots, k. \tag{1.5}$$

Ejemplo 1.1. *La matriz diagonal de orden n, $\mathbf{D} := \mu\mathbf{I}$ con $\mu \in \mathbb{C}$, tiene el polinomio característico*

$$f_{\mathbf{D}}(\lambda) = (\mu - \lambda)^n.$$

Luego $\lambda = \mu$, único valor propio de $\mathbf{D}$, tiene multiplicidad algebraica n y multiplicidad geométrica n. Esto indica que $L(\mu) = \mathbb{C}^n$, es decir, todo vector $\mathbf{x} \in \mathbb{C}^n$, $\mathbf{x} \neq 0$ es vector propio de $\mathbf{D}$ asociado a μ.

Ejemplo 1.2. *Consideremos la matriz $\mathbf{A} \in \mathbb{R}^{n\times n}$ dada por*

$$\mathbf{A} = \begin{bmatrix} \mu & 1 & 0 & \ldots & 0 \\ 0 & \mu & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & \mu & 1 \\ 0 & \ldots & \ldots & 0 & \mu \end{bmatrix},$$

con $\mu \in \mathbb{C}$, que tiene el mismo polinmomio característico que la matriz del ejemplo anterior, en efecto,

$$f_{\mathbf{A}}(\lambda) = \begin{vmatrix} \mu-\lambda & 1 & 0 & \cdots & 0 \\ 0 & \mu-\lambda & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & \mu-\lambda & 1 \\ 0 & \cdots & \cdots & 0 & \mu-\lambda \end{vmatrix} = (\mu-\lambda)^n.$$

En este caso, el único valor propio de $\mathbf{A}$, $\lambda = \mu$, tiene multiplicidad algebraica n, mientras que su multiplicidad geométrica es $\varrho(\mu) = 1$. En efecto, de la Definición 1.4 tenemos

$$\begin{aligned} L(\mu) &= \left\{\mathbf{x} \in \mathbb{C}^n \mid (\mathbf{A}-\mu\mathbf{I})\mathbf{x} = 0\right\} \\ &= \left\{ \mathbf{x} \in \mathbb{C}^n \;\middle|\; \begin{bmatrix} 0 & 1 & \dots & & 0 \\ & 0 & 1 & & \\ & & \ddots & \ddots & \\ & & & 0 & 1 \\ 0 & & & & 0 \end{bmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_{n-1} \\ x_n \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix} \right\}, \end{aligned}$$

esto es,

$$L(\mu) = \{\mathbf{x} \in \mathbb{C}^n \mid x_2 = 0, \dots, x_n = 0\} = \left\{\mathbf{x} = (x_1, 0, \dots, 0)^{\mathrm{T}} \mid x_1 \in \mathbb{C}\right\}.$$

Lo anterior muestra que $\varrho(\mu) = \dim L(\mu) = 1$.

Definición 1.5. *Sean $\mathbf{A}$ y $\mathbf{B} \in \mathbb{C}^{n\times n}$. Las matrices $\mathbf{A}$ y $\mathbf{B}$ se dicen* similares *si existe una matriz $\mathbf{P} \in \mathbb{C}^{n\times n}$ invertible tal que*

$$\mathbf{B} = \mathbf{P}^{-1}\mathbf{A}\mathbf{P}. \tag{1.6}$$

Lema 1.2. *Sean $\mathbf{A}$ y $\mathbf{B} \in \mathbb{C}^{n\times n}$. Si $\mathbf{A}$ y $\mathbf{B}$ son similares, entonces ellas tienen los mismos n valores propios, contando su multiplicidad algebraica. Además, si $\mathbf{x}$ es un vector propio de $\mathbf{A}$, entonces $\mathbf{P}^{-1}\mathbf{x}$ es vector propio de $\mathbf{B}$, con $\mathbf{P}$ que satisface* (1.6).

Demostración. Como $\mathbf{A}$ y $\mathbf{B}$ son similares, existe $\mathbf{P}$ invertible tal que

$$\mathbf{B} = \mathbf{P}^{-1}\mathbf{A}\mathbf{P}.$$

De lo anterior se deduce que

$$\mathbf{A} = \mathbf{P}\mathbf{B}\mathbf{P}^{-1}$$

y luego

$$f_{\mathbf{A}}(\lambda) = \det(\mathbf{A}-\lambda\mathbf{I}) = \det(\mathbf{P}\mathbf{B}\mathbf{P}^{-1} - \lambda\mathbf{P}\mathbf{P}^{-1}) = \det\big(\mathbf{P}(\mathbf{B}-\lambda\mathbf{I})\mathbf{P}^{-1}\big).$$

Puesto que

$$\forall \mathbf{A}, \mathbf{B} \in \mathbb{C}^{n\times n}: \quad \det(\mathbf{A}\mathbf{B}) = \det(\mathbf{A})\det(\mathbf{B}),$$

entonces

$$\det(\mathbf{P})\det(\mathbf{P}^{-1}) = \det(\mathbf{P}\mathbf{P}^{-1}) = \det(\mathbf{I}) = 1,$$

y en consecuencia,

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

Eso muestra que $\mathbf{A}$ y $\mathbf{B}$ tienen el mismo polinomio característico y por lo tanto los mismos n valores propios, contando su multiplicidad algebraica.

Consideremos ahora un valor propio λ de $\mathbf{A}$ y un vector propio $\mathbf{x}$ asociado a λ. Multiplicando a la izquierda por $\mathbf{P}^{-1}$ la ecuación $\mathbf{Ax} = \lambda\mathbf{x}$ obtenemos

$$\mathbf{P}^{-1}\mathbf{Ax} = \lambda(\mathbf{P}^{-1}\mathbf{x}). \tag{1.7}$$

Por otra parte,

$$\mathbf{P}^{-1}\mathbf{Ax} = \mathbf{P}^{-1}\mathbf{A}(\mathbf{P}\mathbf{P}^{-1})\mathbf{x} = (\mathbf{P}^{-1}\mathbf{AP})(\mathbf{P}^{-1}\mathbf{x}),$$

lo cual, en virtud de la igualdad $\mathbf{B} = \mathbf{P}^{-1}\mathbf{AP}$, conduce a

$$\mathbf{P}^{-1}\mathbf{Ax} = \mathbf{B}(\mathbf{P}^{-1}\mathbf{x}). \tag{1.8}$$

Se sigue de las igualdades (1.7) y (1.8) que

$$\mathbf{B}(\mathbf{P}^{-1}\mathbf{x}) = \lambda(\mathbf{P}^{-1}\mathbf{x}).$$

Notando que $\mathbf{P}^{-1}\mathbf{x} \neq 0$, concluimos que $\mathbf{P}^{-1}\mathbf{x}$ es un vector propio de $\mathbf{B}$ asociado al valor propio λ. ■

Definición 1.6. *Sea $\mathcal{B} = \{\mathbf{u}_1, \ldots, \mathbf{u}_n\} \subset \mathbb{C}^n$. Se dice que $\mathcal{B}$ es una* base ortonormal *de $\mathbb{C}^n$ si*

$$\mathbf{u}_i^*\mathbf{u}_j = \delta_{ij} = \begin{cases} 1 & \text{para } i = j, \\ 0 & \text{para } i \neq j. \end{cases}$$

Teorema 1.1 (Forma normal de Schur)**.** *Sea $\mathbf{A} \in \mathbb{C}^{n\times n}$. Entonces existen matrices $\mathbf{U}, \mathbf{T} \in \mathbb{C}^{n\times n}$, $\mathbf{U}$ unitaria y $\mathbf{T}$ triangular superior, tales que*

$$\mathbf{T} = \mathbf{U}^*\mathbf{AU} = \mathbf{U}^{-1}\mathbf{AU}. \tag{1.9}$$

Es decir, $\mathbf{A}$ es unitariamente similar a una matriz triangular superior.

Demostración. Procedemos por inducción sobre el orden n de la matriz $\mathbf{A}$. Para $n = 1$, es trivial porque basta elegir $\mathbf{U} = \begin{bmatrix}1\end{bmatrix}$ y $\mathbf{T} = \mathbf{A}$. Supongamos que el resultado es válido para todas las matrices de orden $k-1$. Probemos que es cierto para todas las matrices de orden k. Sea $\mathbf{A} \in \mathbb{C}^{k\times k}$ y consideremos un valor propio λ_1 de $\mathbf{A}$ y $\mathbf{u}^{(1)}$ un vector propio asociado elegido de manera que

$$\left\|\mathbf{u}^{(1)}\right\|_2^2 = (\mathbf{u}^{(1)})^*\mathbf{u}^{(1)} = 1.$$

Ahora, de acuerdo al Teorema de Completación de Base, podemos elegir una base ortonormal de $\mathbb{C}^k$ que contenga a $\mathbf{u}^{(1)}$, digamos $\mathcal{B} = \{\mathbf{u}^{(1)}, \ldots, \mathbf{u}^{(k)}\}$, y definir la matriz unitaria $\mathbf{P}_1 \in \mathbb{C}^{k\times k}$ como

$$\mathbf{P}_1 := \begin{bmatrix}\mathbf{u}^{(1)} & \mathbf{u}^{(2)} & \cdots & \mathbf{u}^{(k)}\end{bmatrix}.$$

A continuación consideremos la matriz $\mathbf{B}_1 = \mathbf{P}_1^*\mathbf{AP}_1$. Notemos primero que

$$\mathbf{AP}_1 = \begin{bmatrix}\mathbf{Au}^{(1)} & \mathbf{Au}^{(2)} & \cdots & \mathbf{Au}^{(k)}\end{bmatrix} = \begin{bmatrix}\lambda_1\mathbf{u}^{(1)} & \mathbf{v}^{(2)} & \cdots & \mathbf{v}^{(k)}\end{bmatrix}, \tag{1.10}$$

donde $\mathbf{v}^{(j)} := \mathbf{A}\mathbf{u}^{(j)}$ para $j = 2, \ldots, k$. Al multiplicar por la izquierda (1.10) por $\mathbf{P}_1^*$ se obtiene

$$\mathbf{B}_1 = \mathbf{P}_1^* \begin{bmatrix} \lambda_1 \mathbf{u}^{(1)} & \mathbf{v}^{(2)} & \cdots & \mathbf{v}^{(k)} \end{bmatrix} = \begin{bmatrix} \lambda_1 \mathbf{P}_1^* \mathbf{u}^{(1)} & \mathbf{P}_1^* \mathbf{v}^{(2)} & \cdots & \mathbf{P}_1^* \mathbf{v}^{(k)} \end{bmatrix}.$$

Como $\mathbf{P}_1^* \mathbf{P}_1 = \mathbf{I}$ y dado que $\mathbf{u}^{(1)}$ es la primera columna de $\mathbf{P}_1$, entonces

$$\mathbf{P}_1^* \mathbf{u}^{(1)} = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix},$$

y por lo tanto

$$\mathbf{B}_1 = \begin{bmatrix} \lambda_1 & \alpha_2 & \cdots & \alpha_k \\ 0 & & & \\ \vdots & & \mathbf{A}_2 & \\ 0 & & & \end{bmatrix},$$

donde $\mathbf{A}_2 \in \mathbb{C}^{(k-1)\times(k-1)}$ y $\alpha_2, \ldots, \alpha_k$ son escalares en $\mathbb{C}$. Aplicamos la hipótesis de inducción para concluir que existe $\hat{\mathbf{P}}_2 \in \mathbb{C}^{(k-1)\times(k-1)}$, $\hat{\mathbf{P}}_2$ unitaria, tal que

$$\hat{\mathbf{P}}_2^* \mathbf{A}_2 \hat{\mathbf{P}}_2 = \hat{\mathbf{T}}, \tag{1.11}$$

con $\hat{\mathbf{T}}$ triangular superior. Entonces, al definir la matriz $\mathbf{P}_2 \in \mathbb{C}^{k\times k}$ por

$$\mathbf{P}_2 := \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & & & \\ \vdots & & \hat{\mathbf{P}}_2 & \\ 0 & & & \end{bmatrix},$$

obtenemos

$$\mathbf{P}_2^* \mathbf{P}_2 = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & & & \\ \vdots & & \hat{\mathbf{P}}_2^* \hat{\mathbf{P}}_2 & \\ 0 & & & \end{bmatrix} = \mathbf{I}.$$

Así, $\mathbf{P}_2$ es unitaria y además satisface

$$\begin{aligned} \mathbf{P}_2^* \mathbf{B}_1 \mathbf{P}_2 &= \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & & & \\ \vdots & & \hat{\mathbf{P}}_2^* \hat{\mathbf{P}}_2 & \\ 0 & & & \end{bmatrix} \begin{bmatrix} \lambda_1 & \alpha_2 & \cdots & \alpha_k \\ 0 & & & \\ \vdots & & \mathbf{A}_2 & \\ 0 & & & \end{bmatrix} \mathbf{P}_2 \\ &= \begin{bmatrix} \lambda_1 & \alpha_2 & \cdots & \alpha_k \\ 0 & & & \\ \vdots & & \hat{\mathbf{P}}_2^* \mathbf{A}_2 & \\ 0 & & & \end{bmatrix} \mathbf{P}_2 \end{aligned}$$

$$= \begin{bmatrix} \lambda_1 & \alpha_2 & \cdots & \alpha_k \\ 0 & & & \\ \vdots & & \hat{\mathbf{P}}_2^* \mathbf{A}_2 & \\ 0 & & & \end{bmatrix} \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & & & \\ \vdots & & \hat{\mathbf{P}}_2 & \\ 0 & & & \end{bmatrix}.$$

Al realizar la multiplicación de matrices indicada y usando (1.11) llegamos a

$$\mathbf{P}_2^* \mathbf{B}_1 \mathbf{P}_2 = \begin{bmatrix} \lambda_1 & \omega_2 & \cdots & \omega_k \\ 0 & & & \\ \vdots & & \hat{\mathbf{P}}_2^* \mathbf{A}_2 \hat{\mathbf{P}}_2 & \\ 0 & & & \end{bmatrix} = \begin{bmatrix} \lambda_1 & \omega_2 & \cdots & \omega_k \\ 0 & & & \\ \vdots & & \hat{\mathbf{T}} & \\ 0 & & & \end{bmatrix} =: \mathbf{T},$$

donde $\mathbf{T}$ es una matriz triangular superior y los ω_j, $j = 2, \ldots, k$, están dados por

$$\omega_j = (\alpha_2, \ldots, \alpha_k) \hat{\mathbf{P}}_2^{(j)},$$

donde $\hat{\mathbf{P}}_2^{(j)}$ es la columna j de $\hat{\mathbf{P}}_2$.

Puesto que

$$\mathbf{T} = \mathbf{P}_2^* \mathbf{B}_1 \mathbf{P}_2 = \mathbf{P}_2^* (\mathbf{P}_1^* \mathbf{A} \mathbf{P}_1) \mathbf{P}_2 = (\mathbf{P}_1 \mathbf{P}_2)^* \mathbf{A} (\mathbf{P}_1 \mathbf{P}_2),$$

podemos elegir $\mathbf{U}$ como la matriz unitaria $\mathbf{U} = \mathbf{P}_1 \mathbf{P}_2$ con lo cual

$$\mathbf{T} = \mathbf{U}^* \mathbf{A} \mathbf{U}.$$

Del principio de inducción se concluye la validez del teorema. ■

Capítulo 2

Métodos directos para la solución de sistemas lineales (Parte I)

En este capítulo se consideran métodos directos para la solución de sistemas lineales

$$\mathbf{A}\mathbf{x} = \mathbf{b}, \quad \mathbf{A} \in \mathbb{K}^{n\times n}, \quad \mathbf{b} \in \mathbb{K}^n, \quad \mathbb{K} = \mathbb{R} \text{ o } \mathbb{K} = \mathbb{C}, \tag{2.1}$$

donde se supone que $\det(\mathbf{A}) \neq 0$. (Un *método directo* entrega la solución *exacta* del problema en un número finito de pasos, al contrario de los *métodos iterativos*, que se estudiarán más adelante.)

Teóricamente, la solución de (2.1) está dada por $\mathbf{x} = (\xi_1, \ldots, \xi_n)^{\mathrm{T}}$ con

$$\xi_i = \frac{\det\left(\begin{bmatrix}\mathbf{a}_1 & \cdots & \mathbf{a}_{i-1} & \mathbf{b} & \mathbf{a}_{i+1} & \cdots & \mathbf{a}_n\end{bmatrix}\right)}{\det(\mathbf{A})}, \quad i = 1, \ldots, n,$$

donde

$$\mathbf{A} = \begin{bmatrix}\mathbf{a}_1 & \cdots & \mathbf{a}_n\end{bmatrix}.$$

Esta regla es conocida como *regla de Cramer*. Practicamente, sólo en el caso $n = 2$ o para matrices $\mathbf{A}$ especiales, la fórmula es util por razones de esfuerzo computacional y la acumulación de errores de redondeo.

El problema (2.1) nos lleva al problema más general

$$\mathbf{A}\mathbf{X} = \mathbf{B}, \quad \mathbf{A} \in \mathbb{K}^{n\times n}, \quad \mathbf{X} \in \mathbb{K}^{n\times p}, \quad \mathbf{B} \in \mathbb{K}^{n\times p}, \tag{2.2}$$

el cual incluye el problema de

$$\mathbf{A}\mathbf{X} = \mathbf{I}$$

de encontrar la inversa de $\mathbf{A}$. Para resolver (2.2), tomamos en cuenta que este problema representa la solución simultánea de p problemas del tipo (2.1), dado que para

$$\mathbf{X} = \begin{bmatrix}\mathbf{x}_1 & \cdots & \mathbf{x}_p\end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix}\mathbf{b}_1 & \cdots & \mathbf{b}_p\end{bmatrix},$$

tenemos que

$$\mathbf{A}\mathbf{X} = \mathbf{B} \Longleftrightarrow \mathbf{A}\mathbf{x}_i = \mathbf{b}_i, \quad i = 1, \ldots, p;$$

el problema de encontrar $\mathbf{A}^{-1}$ es equivalente a n problemas del tipo (2.1), dado que

$$\mathbf{A}\mathbf{X} = \mathbf{I} \Longleftrightarrow \mathbf{A}\mathbf{x}_i = \mathbf{e}_i, \quad i = 1, \ldots, n,$$

con $\mathbf{X} = \begin{bmatrix} \mathbf{x}_1 & \cdots & \mathbf{x}_n \end{bmatrix}$ y

$$\mathbf{e}_i := \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \leftarrow i. \tag{2.3}$$

2.1. Sistemas lineales escalonados. Matrices triangulares y su inversión

Definición 2.1. *Sea* $\mathbf{A} = (\alpha_{ij}) \in \mathbb{K}^{n\times n}$*. Si* $\alpha_{ij} = 0$ *para* $j < i$*, entonces* $\mathbf{A}$ *se llama* matriz triangular superior*; si* $\alpha_{ij} = 0$ *para* $i < j$*, entonces* $\mathbf{A}$ *se llama* matriz triangular inferior*. Un sistema lineal con una matriz triangular se llama* escalonado*.*

Los sistemas escalonados juegan un rol importante, dado que sus soluciones pueden ser determinadas fácilmente, por ejemplo en el caso de una matriz $\mathbf{A}$ triangular superior:

$$\begin{aligned} \alpha_{11}x_1 + \alpha_{12}x_2 + \cdots + \alpha_{1n}x_n &= b_1, \\ \alpha_{22}x_2 + \cdots + \alpha_{2n}x_n &= b_2, \\ &\vdots \\ \alpha_{n-1,n-1}x_{n-1} + \alpha_{n-1,n}x_n &= b_{n-1}, \\ \alpha_{nn}x_n &= b_n. \end{aligned} \tag{2.4}$$

Se usa la última ecuación para calcular $x_n = b_n/\alpha_{nn}$, luego se remplaza x_n en la penúltima ecuación para determinar x_{n-1}, etcétera. Recordamos que para una matriz $\mathbf{A}$ triangular,

$$\det(\mathbf{A}) = \prod_{i=1}^{n} \alpha_{ii} \neq 0 \Longleftrightarrow \forall i = 1, \ldots, n: \quad \alpha_{ii} \neq 0.$$

Una matriz triangular puede ser invertida fácilmente resolviendo los n sistemas lineales con las n columnas unitarias. Dado que la inversa nuevamente es una matriz triangular del mismo tipo, resultan simplificaciones significativas. Considerando el sistema $\mathbf{Rx} = \mathbf{e}_i$, nos damos cuenta que $\mathbf{x}$ no depende de las columnas $i+1, \ldots, n$ de $\mathbf{R}$. Entonces, si particionamos la matriz $\mathbf{R}$ como

$$\mathbf{R} = \begin{bmatrix} \mathbf{R}_{11} & \mathbf{r} \\ 0 & \varrho \end{bmatrix}, \quad \mathbf{R}_{11} \in \mathbb{K}^{(n-1)\times(n-1)}, \quad \mathbf{r} \in \mathbb{K}^{n-1}, \quad \varrho \in \mathbb{K},$$

esta observación se refleja en la fórmula

$$\mathbf{R}^{-1} = \begin{bmatrix} \mathbf{R}_{11}^{-1} & -\varrho^{-1}\mathbf{R}_{11}^{-1}\mathbf{r} \\ 0 & \varrho^{-1} \end{bmatrix}.$$

Eso significa que para la implementación de la inversión de una matriz triangular superior, podemos sucesivamente remplazar las columnas $n, n-1, \ldots, 2, 1$ de $\mathbf{R}$ por las columnas de $\mathbf{R}^{-1}$.

2.2. El método de eliminación de Gauss

La idea del método de eliminación de Gauss consiste en transformar un sistema arbitrario con una matriz $n \times n$ regular en un sistema con una matriz triangular superior, usando a lo más $n-1$ pasos de transformación de equivalencia:

$$\begin{bmatrix} * & \cdots & \cdots & \cdots & * \\ \vdots & & & & \vdots \\ \vdots & & & & \vdots \\ \vdots & & & & \vdots \\ * & \cdots & \cdots & \cdots & * \end{bmatrix} \begin{pmatrix} * \\ \vdots \\ \vdots \\ \vdots \\ * \end{pmatrix} = \begin{pmatrix} * \\ \vdots \\ \vdots \\ \vdots \\ * \end{pmatrix} \Longleftrightarrow \begin{bmatrix} * & * & \cdots & \cdots & * \\ 0 & \vdots & & & \vdots \\ \vdots & \vdots & & & \vdots \\ \vdots & \vdots & & & \vdots \\ 0 & * & \cdots & \cdots & * \end{bmatrix} \begin{pmatrix} * \\ \vdots \\ \vdots \\ \vdots \\ * \end{pmatrix} = \begin{pmatrix} * \\ \vdots \\ \vdots \\ \vdots \\ * \end{pmatrix}$$

$$\Longleftrightarrow \begin{bmatrix} * & * & * & \cdots & * \\ 0 & * & * & & \vdots \\ \vdots & 0 & \vdots & & \vdots \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & * & \cdots & * \end{bmatrix} \begin{pmatrix} * \\ \vdots \\ \vdots \\ \vdots \\ * \end{pmatrix} = \begin{pmatrix} * \\ \vdots \\ \vdots \\ \vdots \\ * \end{pmatrix} \Longleftrightarrow \ldots \Longleftrightarrow \begin{bmatrix} * & * & \cdots & \cdots & * \\ 0 & * & & & \vdots \\ \vdots & 0 & \ddots & & \vdots \\ \vdots & & \ddots & * & * \\ 0 & 0 & \cdots & 0 & * \end{bmatrix} \begin{pmatrix} * \\ \vdots \\ \vdots \\ \vdots \\ * \end{pmatrix} = \begin{pmatrix} * \\ \vdots \\ \vdots \\ \vdots \\ * \end{pmatrix}.$$

En esta representación esquemática del algoritmo, el símbolo "$*$" representa un elemento que puede asumir un valor diferente de cero, mientras que por "0" se marca cualquier elemento que debe asumir el valor cero debido a la definición del algoritmo.

En el i-ésimo paso, $i = 1, \ldots, n-1$, usamos las siguientes transformaciones:

a) Si es necesario, intercambiamos la fila i con una de las filas $i+1, \ldots, n$ del sistema.
b) Si así se desea, intercambiamos la columna i con alguna de las columnas $i+1, \ldots, n$ del sistema. Tal medida sirve para reducir el efecto de acumulación de errores de redondeo.
c) Sustracción de múltiplos apropiados de la fila i de las filas $i+1, \ldots, n$.

Para la administración de los pasos, usaremos el siguiente esquema, que tambien incluye los números de filas y columnas. Sean

$$\alpha_{ij}^{(1)} := \alpha_{ij}, \quad i,j = 1, \ldots, n; \quad \beta_i^{(1)} := \beta_i, \quad i = 1, \ldots, n. \tag{2.5}$$

Al iniciarse la computación, el esquema está dado por

	1	2	$\cdots$	n	
1	$\alpha_{11}^{(1)}$	$\alpha_{12}^{(1)}$	$\cdots$	$\alpha_{1n}^{(1)}$	$\beta_1^{(1)}$
$\vdots$	$\vdots$			$\vdots$	$\vdots$
n	$\alpha_{n1}^{(1)}$	$\alpha_{n2}^{(1)}$	$\cdots$	$\alpha_{nn}^{(1)}$	$\beta_n^{(1)}$

.

Después de $i-1$ pasos, el esquema asume la siguiente forma:

$$\begin{array}{c|cccccccc|c}
 & \tilde\sigma_1^{(1)} & \cdots & \cdots & \cdots & \tilde\sigma_{i-1}^{(i-1)} & \sigma_i^{(i)} & \cdots & \sigma_n^{(i)} & \\ \hline
\tilde\pi_1^{(1)} & \tilde\alpha_{11}^{(1)} & * & \cdots & * & \tilde\alpha_{1,i-1}^{(i-1)} & \tilde\alpha_{1,i}^{(i)} & \cdots & \tilde\alpha_{1,n}^{(i)} & \tilde\beta_1^{(1)} \\
\vdots & 0 & * & & & * & * & & * & \vdots \\
\vdots & \vdots & 0 & \ddots & & \vdots & \vdots & & \vdots & \vdots \\
\vdots & \vdots & & \ddots & * & * & * & & * & \vdots \\
\tilde\pi_{i-1}^{(i-1)} & 0 & \cdots & \cdots & 0 & \tilde\alpha_{i-1,i-1}^{(i-1)} & \tilde\alpha_{i-1,i}^{(i)} & \cdots & \tilde\alpha_{i-1,n}^{(i)} & \tilde\beta_{i-1}^{(i-1)} \\
\pi_i^{(i)} & 0 & \cdots & \cdots & 0 & 0 & \alpha_{ii}^{(i)} & \cdots & \alpha_{in}^{(i)} & \beta_i^{(i)} \\
\vdots & \vdots & & & \vdots & \vdots & \vdots & & \vdots & \vdots \\
\pi_n^{(i)} & 0 & \cdots & \cdots & 0 & 0 & \alpha_{ni}^{(i)} & \cdots & \alpha_{nn}^{(i)} & \beta_n^{(i)}
\end{array}.$$

Aquí

$$\left(\tilde\pi_1^{(1)},\dots,\tilde\pi_{i-1}^{(i-1)},\pi_i^{(i)},\dots,\pi_n^{(i)}\right) \quad \text{y} \quad \left(\tilde\sigma_1^{(1)},\dots,\tilde\sigma_{i-1}^{(i-1)},\sigma_i^{(i)},\dots,\sigma_n^{(i)}\right)$$

son permutaciones del vector de índices $(1,\dots,n)$,

$$\tilde\alpha_{kj}^{(k)}, \quad \tilde\beta_k^{(k)}, \quad k=1,\dots,i-1, \quad j=k,\dots,i-1$$

son elementos "listos" del sistema final, y

$$\alpha_{kj}^{(i)}, \quad \beta_k^{(i)}, \quad j,k=1,\dots,n$$

son elementos del sistema restante antes de los intercambios.

El i-ésimo paso de transformación consiste en primer lugar en un intercambio de filas (columnas) entre la fila i y una fila $j>i$ y posiblemente entre la columna i y una columna $k>i$, de tal forma que $\alpha_{jk}^{(i)}$ se cambia a la posición (i,i). Los elementos intercambiados los llamamos

$$\tilde\sigma_k^{(i)}, \quad \tilde\alpha_{jk}^{(i)}, \quad j=1,\dots,n, \quad k=i,,\dots,n; \quad \tilde\beta_j^{(i)}, \quad \pi_j^{(i)}, \quad j=i,\dots,n.$$

Ahora supongamos que $\tilde\alpha_{ii}^{(i)}\neq 0$ (mas adelante demostraremos que esto siempre se puede lograr). La i-ésima fila ya no se modifica. Los ceros en las posiciones $i+1,\dots,n$ de la columna i se generan de la siguiente forma:

$$\begin{aligned}
&\alpha_{jk}^{(i+1)} := \tilde\alpha_{jk}^{(i)} - \frac{\tilde\alpha_{ji}^{(i)}}{\tilde\alpha_{ii}^{(i)}}\tilde\alpha_{ik}^{(i)}, \quad \beta_j^{(i+1)} := \tilde\beta_j^{(i)} - \frac{\tilde\alpha_{ji}^{(i)}}{\tilde\alpha_{ii}^{(i)}}\tilde\beta_i^{(i)}, \quad i+1\leqslant j,k\leqslant n,\\
&\pi_j^{(i+1)} := \tilde\pi_j^{(i)}, \quad i+1\leqslant j\leqslant n,\\
&\sigma_k^{(i+1)} := \tilde\sigma_k^{(i)}, \quad i+1\leqslant k\leqslant n.
\end{aligned} \tag{2.6}$$

El cuociente $\tilde\alpha_{ji}^{(i)}/\tilde\alpha_{ii}^{(i)}$ se llama *multiplicador* para la fila j.

Después de $n-1$ pasos ponemos por unificación formal

$$\tilde\alpha_{nn}^{(n)} := \alpha_{nn}^{(n)}, \quad \tilde\beta_n^{(n)} := \beta_n^{(n)}, \quad \tilde\pi_n^{(n)} := \pi_n^{(n)}, \quad \tilde\sigma_n^{(n)} := \sigma_n^{(n)}.$$

Al final, llegamos al esquema

$$\begin{array}{c|ccc|c} & \tilde{\sigma}_1^{(1)} & \cdots & \tilde{\sigma}_n^{(n)} & \\ \hline \tilde{\pi}_1^{(1)} & \tilde{\alpha}_{11}^{(1)} & \cdots & \tilde{\alpha}_{1n}^{(n)} & \tilde{\beta}_1^{(1)} \\ \vdots & & \ddots & \vdots & \vdots \\ \tilde{\pi}_n^{(n)} & & & \tilde{\alpha}_{nn}^{(n)} & \tilde{\beta}_n^{(n)} \end{array}, \tag{2.7}$$

el cual puede ser escrito en forma

$$\mathbf{R}\mathbf{y} = \mathbf{c}$$

con una matriz triangular superior $\mathbf{R}$. La solución de este sistema es $\mathbf{y} = (\eta_1, \ldots, \eta_n)^{\mathrm{T}}$, y podemos recuperar la solución $\mathbf{x} = (\xi_1, \ldots, \xi_n)^{\mathrm{T}}$ del sistema original a través de

$$\xi_{\tilde{\sigma}_i^{(i)}} = \eta_i, \quad i = 1, \ldots, n, \tag{2.8}$$

usando la información de la primera fila del diagrama (2.7), que indica los índices de las componentes de $\mathbf{x}$ correspondientes.

La fórmula esencial para la conversión del sistema restante depués de los cambios es la siguiente:

$$(j,k)_{\text{nuevo}} = (j,k)_{\text{antiguo}} - \frac{(j,i)_{\text{antiguo}}}{(i,i)_{\text{antiguo}}}(i,k)_{\text{antiguo}}, \quad i+1 \leqslant j \leqslant n, \quad i+1 \leqslant k \leqslant n. \tag{2.9}$$

El divisor de los multiplicadores, el elemento $(i,i)_{\text{antiguo}}$, se llama *elemento pivote*. Es un elemento diagonal de la matriz triangular correspondiente.

La parte derecha se transforma a través del mismo esquema. Resulta útil guardar los multiplicadores da cada paso de transformación; se pueden almacenar en las posiciones ocupadas por cero, y tambien se intercambian.

Ejemplo 2.1. *Para ilustrar el algoritmo, consideramos el sistema* $\mathbf{A}\mathbf{x} = \mathbf{b}$ *con*

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & -3 \\ 1 & 1 & 3 \\ 1 & -1 & 3 \end{bmatrix}, \quad \mathbf{b} = \begin{pmatrix} 3 \\ -4 \\ 5 \end{pmatrix}.$$

La aplicación del algoritmo genera la siguiente sucesión de esquemas. Partimos de

$$\begin{array}{c|ccc|c} & 1 = \sigma_1^{(1)} & 2 = \sigma_2^{(1)} & 3 = \sigma_3^{(1)} & \\ \hline \pi_1^{(1)} = 1 & 0 = \alpha_{11}^{(1)} & 1 = \alpha_{12}^{(1)} & -3 = \alpha_{13}^{(1)} & 3 = \beta_1^{(1)} \\ \pi_2^{(1)} = 2 & 1 = \alpha_{21}^{(1)} & 1 = \alpha_{22}^{(1)} & 3 = \alpha_{23}^{(1)} & -4 = \beta_2^{(1)} \\ \pi_3^{(1)} = 3 & 1 = \alpha_{31}^{(1)} & -1 = \alpha_{32}^{(1)} & 3 = \alpha_{33}^{(1)} & 5 = \beta_3^{(1)} \end{array}.$$

Intercambiamos filas y columnas para que el elemento $(2,3)$ *asume la posición diagonal* $(1,1)$. *Es decir, este elemento lo consideramos como pivote:*

$$\begin{array}{c|ccc|c} & 3=\tilde{\sigma}_1^{(1)} & 2=\tilde{\sigma}_2^{(1)} & 1=\tilde{\sigma}_3^{(1)} & \\ \hline \tilde{\pi}_1^{(1)}=2 & 3=\tilde{\alpha}_{11}^{(1)} & 1=\tilde{\alpha}_{12}^{(1)} & 1=\tilde{\alpha}_{13}^{(1)} & -4=\tilde{\beta}_1^{(1)} \\ \tilde{\pi}_2^{(1)}=1 & -3=\tilde{\alpha}_{21}^{(1)} & 1=\tilde{\alpha}_{22}^{(1)} & 0=\tilde{\alpha}_{23}^{(1)} & 3=\tilde{\beta}_2^{(1)} \\ \tilde{\pi}_3^{(1)}=3 & 3=\tilde{\alpha}_{31}^{(1)} & -1=\tilde{\alpha}_{32}^{(1)} & 1=\tilde{\alpha}_{33}^{(1)} & 5=\tilde{\beta}_3^{(1)} \end{array}.$$

Ahora calculamos los multiplicadores

$$\lambda_{21} := \frac{\tilde{\alpha}_{21}^{(1)}}{\tilde{\alpha}_{11}^{(1)}} = -1,$$

el cual corresponde a la sustracción de la fila 1, multiplicada por (-1), *de la fila 2, y*

$$\lambda_{31} := \frac{\tilde{\alpha}_{31}^{(1)}}{\tilde{\alpha}_{11}^{(1)}} = 1,$$

que corresponde a la sustracción de la fila 1 de la fila 3. El resultado de estas operaciones es

$$\begin{array}{c|ccc|c} & 3=\tilde{\sigma}_1^{(1)} & 2=\sigma_2^{(2)} & 1=\sigma_3^{(2)} & \\ \hline \tilde{\pi}_1^{(1)}=2 & 3=\tilde{\alpha}_{11}^{(1)} & 1=\tilde{\alpha}_{12}^{(1)} & 1=\tilde{\alpha}_{13}^{(1)} & -4=\tilde{\beta}_1^{(1)} \\ \pi_2^{(2)}=1 & -1=\lambda_{21} & 2=\alpha_{22}^{(2)} & 1=\alpha_{23}^{(2)} & -1=\beta_2^{(2)} \\ \pi_3^{(2)}=3 & 1=\lambda_{31} & -2=\alpha_{32}^{(2)} & 0=\alpha_{33}^{(2)} & 9=\beta_3^{(2)} \end{array}.$$

Intercambiamos las filas 2 y 3 por motivo de ilustración:

$$\begin{array}{c|ccc|c} & 3=\tilde{\sigma}_1^{(1)} & 2=\tilde{\sigma}_2^{(2)} & 1=\tilde{\sigma}_3^{(2)} & \\ \hline \tilde{\pi}_1^{(1)}=2 & 3=\tilde{\alpha}_{11}^{(1)} & 1=\tilde{\alpha}_{12}^{(2)} & 1=\tilde{\alpha}_{13}^{(2)} & -4=\tilde{\beta}_1^{(1)} \\ \tilde{\pi}_2^{(2)}=3 & 1=\lambda_{21} & -2=\tilde{\alpha}_{22}^{(2)} & 0=\tilde{\alpha}_{23}^{(2)} & 9=\tilde{\beta}_2^{(2)} \\ \tilde{\pi}_3^{(2)}=1 & -1=\lambda_{31} & 2=\tilde{\alpha}_{32}^{(2)} & 1=\tilde{\alpha}_{33}^{(2)} & -1=\tilde{\beta}_3^{(2)} \end{array},$$

donde los multiplicadores fueron intercambiados con las filas y luego renombrados. Ahora calculamos que

$$\lambda_{32} := \frac{\tilde{\alpha}_{32}^{(2)}}{\tilde{\alpha}_{22}^{(2)}} = -1,$$

el cual corresponde a la sustracción de la fila 2, multiplicada por (-1)*, de la fila 3. Así finalmente llegamos al esquema*

	$3 = \tilde{\sigma}_1^{(1)}$	$2 = \tilde{\sigma}_2^{(2)}$	$1 = \tilde{\sigma}_3^{(3)}$	
$\tilde{\pi}_1^{(1)} = 2$	$3 = \tilde{\alpha}_{11}^{(1)}$	$1 = \tilde{\alpha}_{12}^{(2)}$	$1 = \tilde{\alpha}_{13}^{(3)}$	$-4 = \tilde{\beta}_1^{(1)}$
$\tilde{\pi}_2^{(2)} = 3$	$1 = \lambda_{21}$	$-2 = \tilde{\alpha}_{22}^{(2)}$	$0 = \tilde{\alpha}_{23}^{(3)}$	$9 = \tilde{\beta}_2^{(2)}$
$\pi_3^{(3)} = 1$	$-1 = \lambda_{31}$	$-1 = \lambda_{32}$	$1 = \alpha_{33}^{(3)}$	$8 = \beta_3^{(3)}$
$\tilde{\pi}_3^{(3)}$			$\tilde{\alpha}_{33}^{(3)}$	$\tilde{\beta}_3^{(3)}$

Entonces obtenemos la matriz triangular superior $\mathbf{R}$ *y la parte derecha transformada* $\mathbf{c}$ *dadas por*

$$\mathbf{R} = \begin{bmatrix} 3 & 1 & 1 \\ 0 & -2 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad \mathbf{c} = \begin{pmatrix} -4 \\ 9 \\ 8 \end{pmatrix}.$$

La solución del sistema $\mathbf{Ry} = \mathbf{c}$ *entrega*

$$\eta_3 = 8 (= \xi_1), \quad \eta_2 = -\frac{9}{2} (= \xi_2), \quad \eta_1 = \frac{1}{3}\left(-4 + \frac{9}{2} - 8\right) = -\frac{5}{2} (= \xi_3).$$

Hasta ahora siempre se ha presumido que cuando la matriz $\mathbf{A}$ es no singular, el intercambio de filas (y columnas) siempre nos permite lograr que

$$\tilde{\alpha}_{ii}^{(i)} \neq 0, \quad i = 1, \ldots, n.$$

Ahora vemos que este enunciado realmente es válido. Usaremos las siguientes *estrategias de pivote*: la *búsqueda del pivote en la columna*, donde en el k-ésimo paso determinamos el índice k tal que

$$\left|\tilde{\alpha}_{kk}^{(k)}\right| = \max_{i \geqslant k} \left|\alpha_{ik}^{(k)}\right| \tag{2.10}$$

y sólo se intercambain filas, o la búsqueda del pivote en la matriz restante, donde determinamos el índice k tal que

$$\left|\tilde{\alpha}_{kk}^{(k)}\right| = \max_{i,j \geqslant k} \left|\alpha_{ij}^{(k)}\right|, \tag{2.11}$$

la cual implica el intercambio de filas y columnas. En ambos casos, los multiplicadores satisfacen

$$\left|\tilde{\alpha}_{ji}^{(i)} / \tilde{\alpha}_{ii}^{(i)}\right| \leqslant 1,$$

lo que causa un comportamento favorable del error de redondeo.

Ejemplo 2.2 (Tarea 17, Curso 2006)**.** *Se considera el sistema lineal* $\mathbf{Ax} = \mathbf{b}$ *dado por*

$$\mathbf{A} = \begin{bmatrix} 7 & 1 & 1 \\ 10 & 1 & 1 \\ 1000 & 0 & 1 \end{bmatrix}, \quad \mathbf{b} = \begin{pmatrix} 10 \\ 13 \\ 1001 \end{pmatrix}.$$

La solución exacta del sistema es $\mathbf{x} = (1, 2, 1)^{\mathrm{T}}$. *Resolvemos ahora el sistema usando una aritmética con cuatro dígitos significativos, usando el algoritmo de Gauss*

a) *sin pivoteo,*
b) *con búsqueda del pivote en la columna.*
c) *Interpretar los resultados.*

Usaremos la representación científica de los números, por ejemplo

$$\begin{aligned} 1234{,}567 &\longrightarrow 1{,}234567 \times 10^{3} \longrightarrow \texttt{1,235E+3} \\ 3{,}141759 &\longrightarrow 3{,}141759 \times 10^{0} \longrightarrow \texttt{3,142E+0} \\ 0{,}000654321 &\longrightarrow 6{,}5432 \times 10^{-4} \longrightarrow \texttt{6,543E-4.} \end{aligned}$$

Transformamos cada resultado intermedio a esta forma y redondeamos hasta el último dígito. Ojo: Internamente, las calculadoras usan una exactitud mayor que la desplegada en la pantalla.

a) *Sin pivoteo obtenemos:*

$$\widetilde{\mathbf{A}} = \left[\begin{array}{ccc|c} 7{,}000 \times 10^{0} & 1{,}000 \times 10^{0} & 1{,}000 \times 10^{0} & 1{,}000 \times 10^{1} \\ 1{,}000 \times 10^{1} & 1{,}000 \times 10^{0} & 1{,}000 \times 10^{0} & 1{,}300 \times 10^{1} \\ 1{,}000 \times 10^{3} & 0{,}000 \times 10^{0} & 1{,}000 \times 10^{0} & 1{,}001 \times 10^{3} \end{array}\right]$$

Fila $2_{\text{nueva}} =$ Fila $2_{\text{antigua}} - 1{,}419 \times 10^{0}$Fila 1_{antigua} *y*
Fila $3_{\text{nueva}} =$ Fila $3_{\text{antigua}} - 1{,}429 \times 10^{2}$Fila 1_{antigua}*:*

$$\widetilde{\mathbf{A}}^{(1)} = \left[\begin{array}{ccc|c} 7{,}000 \times 10^{0} & 1{,}000 \times 10^{0} & 1{,}000 \times 10^{0} & 1{,}000 \times 10^{1} \\ & -4{,}290 \times 10^{-1} & -4{,}290 \times 10^{-1} & 1{,}290 \times 10^{0} \\ & 1{,}429 \times 10^{2} & -1{,}419 \times 10^{2} & -4{,}280 \times 10^{2} \end{array}\right].$$

Ahora calculamos Fila $3_{\text{nueva}} =$ Fila $3_{\text{antigua}} - 3{,}331 \times 10^{2}$Fila 2_{antigua}*:*

$$\widetilde{\mathbf{A}}^{(2)} = \left[\begin{array}{ccc|c} 7{,}000 \times 10^{0} & 1{,}000 \times 10^{0} & 1{,}000 \times 10^{0} & 1{,}000 \times 10^{1} \\ & -4{,}290 \times 10^{-1} & -4{,}290 \times 10^{-1} & 1{,}290 \times 10^{0} \\ & & -1{,}000 \times 10^{2} & 1{,}700 \times 10^{0} \end{array}\right].$$

La resubstitución entrega

$$x_3 = 1{,}700 \times 10^{0}, \quad x_2 = 1{,}307 \times 10^{0}, \quad x_1 = 0{,}999 \times 10^{0}.$$

b) *Con pivoteo obtenemos*

$$\widetilde{\mathbf{A}} = \left[\begin{array}{ccc|c} 7{,}000 \times 10^{0} & 1{,}000 \times 10^{0} & 1{,}000 \times 10^{0} & 1{,}000 \times 10^{1} \\ 1{,}000 \times 10^{1} & 1{,}000 \times 10^{0} & 1{,}000 \times 10^{0} & 1{,}300 \times 10^{1} \\ 1{,}000 \times 10^{3} & 0{,}000 \times 10^{0} & 1{,}000 \times 10^{0} & 1{,}001 \times 10^{3} \end{array}\right].$$

Intercambiamos la primera y la tercera fila:

$$\left[\begin{array}{ccc|c} 1{,}000 \times 10^{3} & 0{,}000 \times 10^{0} & 1{,}000 \times 10^{0} & 1{,}001 \times 10^{3} \\ 1{,}000 \times 10^{1} & 1{,}000 \times 10^{0} & 1{,}000 \times 10^{0} & 1{,}300 \times 10^{1} \\ 1{,}000 \times 10^{0} & 1{,}000 \times 10^{0} & 1{,}000 \times 10^{0} & 1{,}000 \times 10^{1} \end{array}\right].$$

Fila $2_{\text{nueva}} = \text{Fila } 2_{\text{antigua}} - 1{,}000 \times 10^{-2}\text{Fila } 1_{\text{antigua}}$ *y*
Fila $3_{\text{nueva}} = \text{Fila } 3_{\text{antigua}} - 7{,}000 \times 10^{-3}\text{Fila } 1_{\text{antigua}}$*:*

$$\widetilde{\mathbf{A}}^{(1)} = \left[\begin{array}{ccc|c} 1{,}000 \times 10^{3} & 0{,}000 \times 10^{0} & 1{,}000 \times 10^{0} & 1{,}001 \times 10^{3} \\ & 1{,}000 \times 10^{0} & 9{,}900 \times 10^{-1} & 2{,}990 \times 10^{0} \\ & 1{,}000 \times 10^{0} & 9{,}930 \times 10^{-1} & 2{,}993 \times 10^{0} \end{array}\right].$$

Fila $3_{\text{nueva}} = \text{Fila } 3_{\text{antigua}} - 1{,}000 \times 10^{0}\text{Fila } 2_{\text{antigua}}$*:*

$$\widetilde{\mathbf{A}}^{(1)} = \left[\begin{array}{ccc|c} 1{,}000 \times 10^{3} & 0{,}000 \times 10^{0} & 1{,}000 \times 10^{0} & 1{,}001 \times 10^{3} \\ & 1{,}000 \times 10^{0} & 9{,}900 \times 10^{-1} & 2{,}990 \times 10^{0} \\ & & 3{,}000 \times 10^{-3} & 3{,}000 \times 10^{-3} \end{array}\right].$$

La resubstitución entrega

$$x_1 = 1{,}000 \times 10^{0}, \quad x_2 = 2{,}000 \times 10^{0}, \quad x_3 = 1{,}000 \times 10^{0}.$$

c) *Con pivoteo, no hay errores de redondeo en este ejemplo, mientras que sin pivoteo, el error en la segunda componente es de aprox. 35 % y en la tercera de aprox. 70 %.*

2.3. Descripción matricial del algoritmo de Gauss y el teorema LR

La transformación descrita en la sección anterior, $\mathbf{Ax} = \mathbf{b} \Leftrightarrow \mathbf{Ry} = \mathbf{c}$, será descrita ahora como operación matricial. Recordamos que en el algoritmo aparecen las sigiuentes operaciones: intercambios de filas y combinaciones lineales de filas, que son operaciones matriciales "de la izquierda", y el intercambio de columnas, lo cual es una operación matricial "de la derecha".

El intercambio de la fila i con una fila $k > i$ es efectuado por multiplicación de la izquierda con una matriz de permutación

$$\mathbf{P} = \begin{bmatrix} \mathbf{e}_1^{\mathrm{T}} \\ \vdots \\ \mathbf{e}_{i-1}^{\mathrm{T}} \\ \mathbf{e}_k^{\mathrm{T}} \\ \mathbf{e}_{i+1}^{\mathrm{T}} \\ \vdots \\ \mathbf{e}_{k-1}^{\mathrm{T}} \\ \mathbf{e}_i^{\mathrm{T}} \\ \mathbf{e}_{k+1}^{\mathrm{T}} \\ \vdots \\ \mathbf{e}_n^{\mathrm{T}} \end{bmatrix} = \begin{bmatrix} 1 & & & & & & & & & & \\ & \ddots & & & & & & & & & \\ & & 1 & & & & & & & & \\ & & & 0 & 0 & \cdots & 0 & 1 & & & \\ & & & \vdots & 1 & & & 0 & & & \\ & & & \vdots & & \ddots & & \vdots & & & \\ & & & 0 & & & 1 & \vdots & & & \\ & & & 1 & 0 & \cdots & \cdots & 0 & & & \\ & & & & & & & & 1 & & \\ & & & & & & & & & \ddots & \\ & & & & & & & & & & 1 \end{bmatrix}.$$

Análogamente, las columnas i y $j > i$ se intercambian a través de la multiplicación de la derecha por

$$\mathbf{Q} = \begin{bmatrix} \mathbf{e}_1 & \cdots & \mathbf{e}_{i-1} & \mathbf{e}_j & \mathbf{e}_{i+1} & \cdots & \mathbf{e}_{j-1} & \mathbf{e}_i & \mathbf{e}_{j+1} & \cdots & \mathbf{e}_n \end{bmatrix}.$$

Nos damos cuenta que $\mathbf{P} = \mathbf{P}^{\mathrm{T}}$, $\mathbf{Q} = \mathbf{Q}^{\mathrm{T}}$ y $\mathbf{P}^2 = \mathbf{Q}^2 = \mathbf{I}$. Finalmente, los ceros debajo del elemento $\tilde{\alpha}_{ii}^{(i)}$ se generan por multiplicación de la izquierda por la matriz

$$\mathbf{T}_i = \begin{bmatrix} 1 & & & & & \\ & \ddots & & & & \\ & & 1 & & & \\ & & -\dfrac{\tilde{\alpha}_{i+1,i}^{(i)}}{\tilde{\alpha}_{ii}^{(i)}} & 1 & & \\ & & \vdots & & \ddots & \\ & & -\dfrac{\tilde{\alpha}_{n,i}^{(i)}}{\tilde{\alpha}_{ii}^{(i)}} & & & 1 \end{bmatrix} = \mathbf{I} - \mathbf{q}_i \mathbf{e}_i^{\mathrm{T}}, \quad \mathbf{q}_i := \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \dfrac{\tilde{\alpha}_{i+1,i}^{(i)}}{\tilde{\alpha}_{ii}^{(i)}} \\ \vdots \\ \dfrac{\tilde{\alpha}_{n,i}^{(i)}}{\tilde{\alpha}_{ii}^{(i)}} \end{pmatrix}.$$

Aquí $\mathbf{q}_i$ es el vector compuesto por i ceros y los multiplicadores del i-ésimo paso. Para explicarlo, sea

$$\mathbf{g} = \begin{pmatrix} \gamma_1 \\ \vdots \\ \gamma_l \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \quad l < i$$

algún vector (que corresponde a una de las columnas 1 a $i-1$ de la matriz transformada en el i-ésimo paso). Entonces

$$\mathbf{T}_i \mathbf{g} = (\mathbf{I} - \mathbf{q}_i \mathbf{e}_i^{\mathrm{T}})\mathbf{g} = \mathbf{g} - \mathbf{q}_i \underbrace{(\mathbf{e}_i^{\mathrm{T}} \mathbf{g})}_{=0} = \mathbf{g},$$

o sea, las $i-1$ primeras columnas quedan sin cambiar. La columna j, $j \geqslant i$, es de la siguiente forma:

$$\mathbf{a}_j^{(i)} = \begin{pmatrix} \tilde{\alpha}_{1j}^{(i)} \\ \vdots \\ \tilde{\alpha}_{i-1,j}^{(i)} \\ \tilde{\alpha}_{ij}^{(i)} \\ \vdots \\ \tilde{\alpha}_{n,j}^{(i)} \end{pmatrix} = \mathbf{g}_j^{(i)} + \mathbf{h}_j^{(i)}, \quad \mathbf{g}_j^{(i)} := \begin{pmatrix} \tilde{\alpha}_{1j}^{(i)} \\ \vdots \\ \tilde{\alpha}_{i-1,j}^{(i)} \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \quad \mathbf{h}_j^{(i)} := \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \tilde{\alpha}_{ij}^{(i)} \\ \vdots \\ \tilde{\alpha}_{n,j}^{(i)} \end{pmatrix}.$$

Entonces

$$\mathbf{T}_i \mathbf{a}_j^{(i)} = \mathbf{T}_i \big(\mathbf{g}_j^{(i)} + \mathbf{h}_j^{(i)}\big) = \mathbf{T}_i \mathbf{g}_j^{(i)} + \mathbf{T}_i \mathbf{h}_j^{(i)} = \mathbf{g}_j^{(i)} + \mathbf{h}_j^{(i)} - \mathbf{q}_i \underbrace{\mathbf{e}_i^{\mathrm{T}} \mathbf{h}_j^{(i)}}_{=\tilde{\alpha}_{ij}^{(i)}}$$

$$= \begin{pmatrix} \tilde{\alpha}_{1j}^{(i)} \\ \vdots \\ \tilde{\alpha}_{i-1,j}^{(i)} \\ 0 \\ \vdots \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \tilde{\alpha}_{ij}^{(i)} \\ \vdots \\ \tilde{\alpha}_{n,j}^{(i)} \end{pmatrix} - \tilde{\alpha}_{ij}^{(i)} \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \dfrac{\tilde{\alpha}_{i+1,i}^{(i)}}{\tilde{\alpha}_{ii}^{(i)}} \\ \vdots \\ \dfrac{\tilde{\alpha}_{n,i}^{(i)}}{\tilde{\alpha}_{ii}^{(i)}} \end{pmatrix} = \begin{pmatrix} \tilde{\alpha}_{1j}^{(i)} \\ \vdots \\ \tilde{\alpha}_{ij}^{(i)} \\ \tilde{\alpha}_{i+1,j}^{(i+1)} \\ \vdots \\ \tilde{\alpha}_{n,j}^{(i+1)} \end{pmatrix}, \quad j = 1, \ldots, n,$$

donde $\alpha_{ki}^{(i+1)} = 0$ para $k \geqslant i+1$. Escrita en forma de matrices, la transformación equivalente efectuada por el algoritmo de Gauss es la siguiente:

$$\begin{aligned} & \mathbf{A}\mathbf{x} = \mathbf{b} \\ \iff \quad & \mathbf{T}_1\mathbf{P}_1\mathbf{A}\mathbf{Q}_1\mathbf{Q}_1\mathbf{x} = \mathbf{T}_1\mathbf{P}_1\mathbf{b} \\ \iff \quad & \mathbf{T}_2\mathbf{P}_2\mathbf{T}_1\mathbf{P}_1\mathbf{A}\mathbf{Q}_1\mathbf{Q}_2\mathbf{Q}_2\mathbf{Q}_1\mathbf{x} = \mathbf{T}_2\mathbf{P}_2\mathbf{T}_1\mathbf{P}_1\mathbf{b} \\ & \qquad\qquad \vdots \\ \iff \quad & \underbrace{\mathbf{T}_{n-1}\mathbf{P}_{n-1} \cdot \ldots \cdot \mathbf{T}_1\mathbf{P}_1\mathbf{A}\mathbf{Q}_1 \cdot \ldots \cdot \mathbf{Q}_{n-1}}_{=\mathbf{R}} \underbrace{\mathbf{Q}_{n-1} \cdot \ldots \cdot \mathbf{Q}_1\mathbf{x}}_{=\mathbf{y}} = \underbrace{\mathbf{T}_{n-1}\mathbf{P}_{n-1} \cdot \ldots \cdot \mathbf{T}_1\mathbf{P}_1\mathbf{b}}_{=\mathbf{c}}. \end{aligned}$$

Sea $\mathbf{Q} := \mathbf{Q}_1\mathbf{Q}_2 \cdot \ldots \cdot \mathbf{Q}_{n-1}$. La matriz $\mathbf{Q}$ describe el efecto combinado de todas las permutaciones de columnas. Eso significa que con la notación definitiva $\tilde{\sigma}_1^{(1)}, \ldots, \tilde{\sigma}_n^{(n)}$, tenemos

$$\mathbf{Q} = \left[\mathbf{e}_{\tilde{\sigma}_1^{(1)}} \quad \mathbf{e}_{\tilde{\sigma}_2^{(2)}} \quad \cdots \quad \mathbf{e}_{\tilde{\sigma}_n^{(n)}} \right].$$

Entonces $\mathbf{Q}^{\mathrm{T}}\mathbf{x} = \mathbf{y}$, o sea

$$\xi_{\tilde{\sigma}_i^{(i)}} = \eta_i, \quad i = 1, \ldots, n,$$

identidad que ya usamos más arriba. Ahora podemos escribir

$$\begin{aligned} \mathbf{R} =& \mathbf{T}_{n-1}\mathbf{P}_{n-1}\mathbf{T}_{n-2}\mathbf{P}_{n-2} \cdot \ldots \cdot \mathbf{T}_1\mathbf{P}_1\mathbf{A}\mathbf{Q} \\ =& \underbrace{\mathbf{T}_{n-1}}_{=:\hat{\mathbf{T}}_{n-1}} (\underbrace{\mathbf{P}_{n-1}\mathbf{T}_{n-2}\mathbf{P}_{n-1}}_{=:\hat{\mathbf{T}}_{n-2}}) (\underbrace{\mathbf{P}_{n-1}\mathbf{P}_{n-2}\mathbf{T}_{n-3}\mathbf{P}_{n-2}\mathbf{P}_{n-1}}_{=:\hat{\mathbf{T}}_{n-2}}) \cdot \ldots \cdot \\ & \cdot \ldots \cdot (\underbrace{\mathbf{P}_{n-1}\mathbf{P}_{n-2} \cdot \ldots \cdot \mathbf{P}_2\mathbf{T}_1\mathbf{P}_2 \cdot \ldots \cdot \mathbf{P}_{n-2}\mathbf{P}_{n-1}}_{=:\hat{\mathbf{T}}_1}) (\underbrace{\mathbf{P}_{n-1}\mathbf{P}_{n-2} \cdot \ldots \cdot \mathbf{P}_1}_{=:\mathbf{P}} \mathbf{A}\mathbf{Q}), \end{aligned}$$

es decir, definiendo

$$\begin{aligned} \mathbf{P} &:= \mathbf{P}_{n-1}\mathbf{P}_{n-2} \cdot \ldots \cdot \mathbf{P}_1, \\ \hat{\mathbf{T}}_{n-1} &:= \mathbf{T}_{n-1}, \\ \hat{\mathbf{T}}_i &:= \mathbf{P}_{n-1} \cdot \ldots \cdot \mathbf{P}_{i+1}\mathbf{T}_i\mathbf{P}_{i+1} \cdot \ldots \cdot \mathbf{P}_{n-1}, \quad i = 1, \ldots, n-2, \end{aligned}$$

obtenemos la fórmula

$$\mathbf{R} = \hat{\mathbf{T}}_{n-1}\hat{\mathbf{T}}_{n-2} \cdot \ldots \cdot \hat{\mathbf{T}}_1\mathbf{P}\mathbf{A}\mathbf{Q}.$$

Podemos aprovechar $\mathbf{P}_j^2 = \mathbf{I}$ para concluir que

$$\begin{aligned}\hat{\mathbf{T}}_i &= \mathbf{P}_{n-1} \cdot \ldots \cdot \mathbf{P}_{i+1}(\mathbf{I} - \mathbf{q}_i \mathbf{e}_i^{\mathrm{T}})\mathbf{P}_{i+1} \cdot \ldots \cdot \mathbf{P}_{n-1} \\ &= \mathbf{I} - \underbrace{\mathbf{P}_{n-1} \cdot \ldots \cdot \mathbf{P}_{i+1}\mathbf{q}_i}_{=:\hat{\mathbf{q}}_i} \mathbf{e}_i^{\mathrm{T}} \\ &= \mathbf{I} - \hat{\mathbf{q}}_i \mathbf{e}_i^{\mathrm{T}}, \quad i = 1, \ldots, n-1,\end{aligned}$$

puesto que las matrices $\mathbf{P}_{i+1}, \ldots, \mathbf{P}_n$ describen intercambios de elementos con índice $\geqslant i+1$, es decir, no afectan a $\mathbf{e}_i^{\mathrm{T}}$. Según nuestra construcción, el vector $\mathbf{q}_i$ es el vector de los multiplicadores del i-ésimo paso de eliminación, los cuales estan sujetos a los mismos intercambios de filas que el sistema restante en los pasos de eliminación $i+1, \ldots, n-1$.

En virtud de lo anterior, $\hat{\mathbf{T}}_i$ es una matriz triangular inferior con diagonal $(1, \ldots, 1)$, tanto como las matrices $\hat{\mathbf{T}}_i^{-1}$ y el producto $\hat{\mathbf{T}}_1^{-1} \cdot \ldots \cdot \hat{\mathbf{T}}_{n-1}^{-1}$. Entonces tenemos que

$$\mathbf{PAQ} = \underbrace{\hat{\mathbf{T}}_1^{-1} \cdot \ldots \cdot \hat{\mathbf{T}}_{n-1}^{-1}}_{=:\mathbf{L}} \mathbf{R} = \mathbf{LR},$$

donde $\mathbf{R}$ es una matriz triangular superior y $\mathbf{L}$ es una matriz triangular inferior con diagonal $(1, \ldots, 1)$. Además, sabemos que

$$\hat{\mathbf{T}}_i^{-1} = \mathbf{I} + \hat{\mathbf{q}}_i \mathbf{e}_i^{\mathrm{T}}, \quad i = 1, \ldots, n-1,$$

lo que implica que

$$\mathbf{L} = (\mathbf{I} + \hat{\mathbf{q}}_1 \mathbf{e}_1^{\mathrm{T}})(\mathbf{I} + \hat{\mathbf{q}}_2 \mathbf{e}_2^{\mathrm{T}}) \cdot \ldots \cdot (\mathbf{I} + \hat{\mathbf{q}}_{n-1} \mathbf{e}_{n-1}^{\mathrm{T}}).$$

Dado que $\mathbf{e}_j^{\mathrm{T}} \hat{\mathbf{q}}_k = 0$ para $j \leqslant k$, podemos escribir

$$\mathbf{L} = \mathbf{I} + \sum_{k=1}^{n-1} \hat{\mathbf{q}}_k \mathbf{e}_k^{\mathrm{T}},$$

es decir, los elementos de $\mathbf{L}$ debajo de la diagonal son los multiplicadores (intercambiados).

Ahora queda para demostrar que el algoritmo nunca termina para una matriz $\mathbf{A}$ regular, o sea, que aplicando intercambios apropiados siempre podemos lograr que $\tilde{\alpha}_{ii}^{(i)} \neq 0$ para $i = 1, \ldots, n$. Eso incluso es válido si no usamos intercambios de columnas (sólamente de filas). Si no fuera así, existiría un índice k tal que

$$\alpha_{ik}^{(k)} = 0, \quad i = k, \ldots, n,$$

o sea

$$\underbrace{\mathbf{T}_{k-1}\mathbf{P}_{k-1}\mathbf{T}_{k-2} \cdot \ldots \cdot \mathbf{T}_1\mathbf{P}_1}_{\det(\ldots)\neq 0} \mathbf{A} = \begin{bmatrix} * & * & \ldots & * & * & * & \ldots & * \\ & * & & \vdots & \vdots & \vdots & & \vdots \\ & & \ddots & \vdots & \vdots & \vdots & & \vdots \\ & & & * & * & \vdots & & \vdots \\ & & & & 0 & \vdots & & \vdots \\ & & & & \vdots & \vdots & & \vdots \\ & & & & 0 & * & \ldots & * \end{bmatrix} \Longrightarrow \det(\mathbf{A}) = 0,$$

una contradicción. La consideraciones anteriores pueden ser resumidas en el siguiente teorema.

Teorema 2.1. *Sea* $\mathbf{A} \in \mathbb{K}^{n\times n}$ *una matriz regular. Entonces existen una matriz de permutación* $\mathbf{P}$*, una matriz triangular inferior* $\mathbf{L}$ *con diagonal* $(1,\dots,1)$ *y una matriz triangular superior* $\mathbf{R}$*, todas pertenecientes a* $\mathbb{K}^{n\times n}$*, tales que* $\mathbf{PA} = \mathbf{LR}$*.*

Si el algoritmo de Gauss es aplicado a un sistema lineal $\mathbf{Ax} = \mathbf{b}$, la matriz $\mathbf{P}$ es la matriz de permutación que describe el efecto de todos los intercambios de filas, $\mathbf{Q}$ es la matriz de permutación que describe el efecto de todos los intercambios de columnas, $\mathbf{R}$ es la matriz triangular superior que resulta y $\mathbf{L}$ es la matriz triangular inferior con diagonal $(1,\dots,1)$ y los multiplicadores (adecuadamente intercambiados), entonces tenemos que $\mathbf{PAQ} = \mathbf{LR}$.

Ejemplo 2.3. *Para la matriz del Ejemplo 2.1, obtenemos que*

$$\mathbf{P} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}, \quad \mathbf{Q} = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}, \quad \mathbf{L} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ -1 & -1 & 1 \end{bmatrix}, \quad \mathbf{R} = \begin{bmatrix} 3 & 1 & 1 \\ 0 & -2 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Ejemplo 2.4.

a) *Nos interesa calcular una descomposición triangular* $\mathbf{PAQ} = \mathbf{LR}$*, con búsqueda de pivote en la matriz restante, de la matriz*

$$\mathbf{A} = \begin{bmatrix} 1 & 3 & -2 \\ 2 & 1 & -4 \\ 2 & 2 & 8 \end{bmatrix}. \tag{2.12}$$

Indicar explícitamente las matrices $\mathbf{P}$*,* $\mathbf{Q}$*,* $\mathbf{L}$ *y* $\mathbf{R}$

b) *Utilizando la descomposición de (a), queremos calcular* $\mathbf{A}^{-1}$*.*

Solución sugerida.

a) *Las etapas consecutivas del algoritmo de Gauss son las siguientes:*

$$\begin{array}{c|ccc} & 1 & 2 & 3 \\ \hline 1 & 1 & 3 & -2 \\ 2 & 2 & 1 & -4 \\ 3 & 2 & 2 & 8 \end{array} \rightarrow \begin{array}{c|ccc} & 3 & 2 & 1 \\ \hline 3 & 8 & 2 & 2 \\ 2 & -4 & 1 & 2 \\ 1 & -2 & 3 & 1 \end{array} \rightarrow \begin{array}{c|ccc} & 3 & 2 & 1 \\ \hline 3 & 8 & 2 & 2 \\ 2 & -\frac{1}{2} & 2 & 3 \\ 1 & -\frac{1}{4} & \frac{7}{2} & \frac{3}{2} \end{array} \rightarrow \begin{array}{c|ccc} & 3 & 2 & 1 \\ \hline 3 & 8 & 2 & 2 \\ 1 & -\frac{1}{4} & \frac{7}{2} & \frac{3}{2} \\ 2 & -\frac{1}{2} & \frac{4}{7} & \frac{15}{7} \end{array}.$$

Entonces

$$\mathbf{L} = \begin{bmatrix} 1 & 0 & 0 \\ -\frac{1}{4} & 1 & 0 \\ -\frac{1}{2} & \frac{4}{7} & 1 \end{bmatrix}, \ \mathbf{R} = \begin{bmatrix} 8 & 2 & 2 \\ 0 & \frac{7}{2} & \frac{3}{2} \\ 0 & 0 & \frac{15}{7} \end{bmatrix}, \ \mathbf{P} = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}, \ \mathbf{Q} = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}.$$

b) *Usando*

$$\mathbf{L}^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ \frac{1}{4} & 1 & 0 \\ \frac{5}{14} & -\frac{4}{7} & 1 \end{bmatrix}, \ \mathbf{R}^{-1} = \begin{bmatrix} \frac{1}{8} & -\frac{1}{14} & -\frac{1}{15} \\ 0 & \frac{2}{7} & -\frac{1}{5} \\ 0 & 0 & \frac{7}{15} \end{bmatrix},$$

obtenemos

$$\mathbf{A}^{-1} = \mathbf{Q}\mathbf{R}^{-1}\mathbf{L}^{-1}\mathbf{P} = \mathbf{Q} \begin{bmatrix} \frac{1}{12} & -\frac{1}{30} & -\frac{1}{15} \\ 0 & \frac{2}{5} & -\frac{1}{5} \\ \frac{1}{6} & -\frac{4}{15} & \frac{7}{15} \end{bmatrix} \mathbf{P} = \begin{bmatrix} -\frac{4}{15} & \frac{7}{15} & \frac{1}{6} \\ \frac{2}{5} & -\frac{1}{5} & 0 \\ -\frac{1}{30} & -\frac{1}{15} & \frac{1}{12} \end{bmatrix}.$$

Ejemplo 2.5 (Certamen 1, Curso 2006)**.** *Se consideran la matriz* $\mathbf{A}$ *y el vector* $\mathbf{b}$ *dados por*

$$\mathbf{A} = \begin{bmatrix} 6 & 3 & 1 \\ 8 & 5 & 2 \\ 9 & 7 & 4 \end{bmatrix}, \quad \mathbf{b} = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}.$$

a) *Usando el algoritmo de Gauss con búsqueda del pivote en la columna, determinar una matriz* $\mathbf{P}$ *de permutación, una matriz* $\mathbf{L} = (\lambda_{ij})$ *triangular inferior con* $\lambda_{11} = \lambda_{22} = \lambda_{33} = 1$ *y una matriz* $\mathbf{R}$ *triangular superior tales que* $\mathbf{PA} = \mathbf{LR}$.
b) *Resolver el sistema* $\mathbf{Ax} = \mathbf{b}$.
c) *Usando el algoritmo de Gauss con búsqueda del pivote en la matriz restante, determinar matrices* $\mathbf{P}$, $\mathbf{Q}$ *de permutación, una matriz* $\mathbf{L} = (\lambda_{ij})$ *triangular inferior con* $\lambda_{11} = \lambda_{22} = \lambda_{33} = 1$ *y una matriz* $\mathbf{R}$ *triangular superior tales que* $\mathbf{PAQ} = \mathbf{LR}$.

Solución sugerida.

a) *Salimos del esquema*

$$\widetilde{\mathbf{A}} = \begin{array}{c|ccc|c} & & & & \\ \hline 1 & 6 & 3 & 1 & 1 \\ 2 & 8 & 5 & 2 & 2 \\ 3 & \mathbf{9} & 7 & 4 & 3 \end{array}$$

donde $\mathbf{9}$ *es el pivote. Intercambiando Fila 1 con Fila 3, obtenemos*

$$\begin{array}{c|ccc|c} & & & & \\ \hline 3 & 9 & 7 & 4 & 3 \\ 2 & 8 & 5 & 2 & 2 \\ 1 & 6 & 3 & 1 & 1 \end{array}$$

Ahora,

$$\text{Fila } 2_{\text{nueva}} = \text{Fila } 2_{\text{antigua}} - \frac{8}{9}\text{Fila } 1_{\text{antigua}}$$

y $$\text{Fila } 3_{\text{nueva}} = \text{Fila } 3_{\text{antigua}} - \frac{2}{3}\text{Fila } 1_{\text{antigua}} :$$

$$\begin{array}{c|cccc} & & & & \\ \hline 3 & 9 & 7 & 4 & 3 \\ 2 & \frac{8}{9} & -\frac{11}{9} & -\frac{14}{9} & -\frac{2}{3} \\ 1 & \frac{2}{3} & -\mathbf{\frac{5}{3}} & -\frac{5}{3} & -1 \end{array},$$

donde $\frac{8}{9}$ *y* $\frac{2}{3}$ *son multiplicadores almacenados y* $-\mathbf{\frac{5}{3}}$ *es el nuevo pivote. Luego, intercambiando Fila 2 con Fila 3,*

$$\begin{array}{c|cccc} & & & & \\ \hline 3 & 9 & 7 & 4 & 3 \\ 1 & \frac{2}{3} & -\frac{5}{3} & -\frac{5}{3} & -1 \\ 2 & \frac{8}{9} & -\frac{11}{9} & -\frac{14}{9} & -\frac{2}{3} \end{array},$$

Ahora, calculamos $\text{Fila } 3_{\text{nueva}} = \text{Fila } 3_{\text{antigua}} - \frac{11}{15}\text{Fila } 2_{\text{antigua}}$ *y almacenamos el multiplicador* $\frac{11}{15}$ *para obtener*

$$\begin{array}{c|cccc} & & & & \\ \hline 3 & 9 & 7 & 4 & 3 \\ 1 & \frac{2}{3} & -\frac{5}{3} & -\frac{5}{3} & -1 \\ 2 & \frac{8}{9} & \frac{11}{15} & -\frac{1}{3} & \frac{1}{15} \end{array}$$

El último esquema implica que

$$\mathbf{P} = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}, \quad \mathbf{L} = \begin{bmatrix} 1 & 0 & 0 \\ \frac{2}{3} & 1 & 0 \\ \frac{8}{9} & \frac{11}{15} & 1 \end{bmatrix}, \quad \mathbf{R} = \begin{bmatrix} 9 & 7 & 4 \\ 0 & -\frac{5}{3} & -\frac{5}{3} \\ 0 & 0 & -\frac{1}{3} \end{bmatrix}.$$

b) *Aplicando una resubstitución a la parte derecha co-transformada, obtenemos*

$$x_3 = \frac{\frac{1}{15}}{-\frac{1}{3}} = -\frac{1}{5}, \quad x_2 = \frac{-1 - \left(-\frac{5}{3}\right)x_3}{-\frac{5}{3}} = \frac{4}{5}, \quad x_1 = \frac{3 - 4x_3 - 7x_2}{9} = -\frac{1}{5}.$$

c) *Una inspección del primer paso de (a) muestra que "9" seria escogido como pivote también por la búsqueda en la matriz restante, asi que este primer paso es igual al primer paso del método con búsqueda en la matriz restante. Asimismo,* $-\frac{5}{3}$ *también seria el pivote en el segundo paso. Concluimos que la búsqueda del pivote en la matriz restante genera el mismo resultado que (a), asi que las matrices* $\mathbf{P}$, $\mathbf{A}$ *y* $\mathbf{R}$ *son las especificadas arriba, y* $\mathbf{Q} = \mathbf{I}$.

El significado de los elementos pivotes se aclara en el siguiente teorema, formulado para una descomposición triangular sin intercambios. Para el caso general, hay que remplazar $\mathbf{A}$ por $\mathbf{PAQ}$.

Teorema 2.2. *Si el algoritmo de Gauss se ejecuta hasta el paso k, $1 \leqslant k \leqslant n-1$, sin intercambios de filas o columnas, entonces tenemos que*

$$\prod_{j=1}^{k} \alpha_{jj}^{(j)} = \begin{vmatrix} \alpha_{11} & \dots & \alpha_{1k} \\ \vdots & & \vdots \\ \alpha_{k1} & \dots & \alpha_{kk} \end{vmatrix} = k\text{-ésimo subdeterminante principal de } \mathbf{A}.$$

Entonces, si todos los subdeterminantes principales de $\mathbf{A}$ son diferentes de cero, se puede ejecutar el algoritmo de Gauss sin intercambio de filas ni de columnas. En este caso, finalmente obtenemos que

$$\det(\mathbf{A}) = \prod_{j=1}^{n} \alpha_{jj}^{(j)}.$$

Demostración. Primero notamos que

$$\begin{bmatrix} \alpha_{11} & \dots & \alpha_{1k} \\ \vdots & & \vdots \\ \alpha_{k1} & \dots & \alpha_{kk} \end{bmatrix} = \begin{bmatrix} \mathbf{e}_1 & \cdots & \mathbf{e}_k \end{bmatrix}^{\mathrm{T}} \mathbf{A} \begin{bmatrix} \mathbf{e}_1 & \cdots & \mathbf{e}_k \end{bmatrix}.$$

Luego tomamos en cuenta que según nuestra construcción,

$$\mathbf{T}_k \cdot \ldots \cdot \mathbf{T}_1 \mathbf{A} = \begin{bmatrix} \alpha_{11}^{(1)} & \cdots & * & * & \cdots & * \\ 0 & \ddots & \vdots & \vdots & & \vdots \\ \vdots & \ddots & \alpha_{kk}^{(k)} & \vdots & & \vdots \\ \vdots & & 0 & \vdots & & \vdots \\ \vdots & & \vdots & \vdots & & \vdots \\ 0 & \cdots & 0 & * & \cdots & * \end{bmatrix} = \begin{bmatrix} \mathbf{R}_k & \tilde{\mathbf{A}}_k \\ 0 & * \end{bmatrix},$$

o sea, dado que

$$\mathbf{T}_1^{-1} \cdot \ldots \cdot \mathbf{T}_k^{-1} = \begin{bmatrix} 1 & & & & & & \\ \lambda_{21} & \ddots & & & & & \\ \vdots & \ddots & 1 & & & & \\ \vdots & & \lambda_{k+1,k} & 1 & & & \\ \vdots & & \vdots & 0 & \ddots & & \\ \vdots & & \vdots & \vdots & \ddots & \ddots & \\ \lambda_{n1} & \cdots & \lambda_{nk} & 0 & \cdots & 0 & 1 \end{bmatrix},$$

donde λ_{ji} son los "multiplicadores", obtenemos que

$$\begin{aligned}
\begin{bmatrix} \alpha_{11} & \dots & \alpha_{1k} \\ \vdots & & \vdots \\ \alpha_{k1} & \dots & \alpha_{kk} \end{bmatrix} &= \begin{bmatrix} \mathbf{e}_1^{\mathrm{T}} \\ \vdots \\ \mathbf{e}_k^{\mathrm{T}} \end{bmatrix} \mathbf{T}_1^{-1} \cdot \ldots \cdot \mathbf{T}_k^{-1} \begin{bmatrix} \mathbf{R}_k & \tilde{\mathbf{A}}_k \\ 0 & * \end{bmatrix} \begin{bmatrix} \mathbf{e}_1 & \cdots & \mathbf{e}_k \end{bmatrix} \\
&= \begin{bmatrix} \mathbf{e}_1^{\mathrm{T}} \\ \vdots \\ \mathbf{e}_k^{\mathrm{T}} \end{bmatrix} \mathbf{T}_1^{-1} \cdot \ldots \cdot \mathbf{T}_k^{-1} \begin{bmatrix} \mathbf{R}_k \\ 0 \end{bmatrix} \\
&= \begin{bmatrix} 1 & & & & 0 & \cdots & 0 \\ \lambda_{21} & \ddots & & & \vdots & & \vdots \\ \vdots & \ddots & \ddots & & \vdots & & \vdots \\ \lambda_{k1} & \cdots & \lambda_{k,k-1} & 1 & 0 & \cdots & 0 \end{bmatrix} \begin{bmatrix} \mathbf{R}_k \\ 0 \end{bmatrix} = \mathbf{L}_k \mathbf{R}_k.
\end{aligned}$$

Puesto que $\det(\mathbf{L}_k) = 1$, concluimos que

$$\begin{vmatrix} \alpha_{11} & \dots & \alpha_{1k} \\ \vdots & & \vdots \\ \alpha_{k1} & \dots & \alpha_{kk} \end{vmatrix} = \det(\mathbf{L}_k \mathbf{R}_k) = \det(\mathbf{R}_k) = \prod_{j=1}^{k} \alpha_{jj}^{(j)}.$$

Para $k = n-1$, de $\mathbf{T}_{n-1} \cdot \ldots \cdot \mathbf{T}_1 \mathbf{A} = \mathbf{R}$ resulta

$$\underbrace{\det(\mathbf{T}_{n-1})}_{=1} \underbrace{\det(\mathbf{T}_{n-2})}_{=1} \cdot \ldots \cdot \underbrace{\det(\mathbf{T}_1)}_{=1} \det \mathbf{A} = \det \mathbf{R} = \prod_{j=1}^{n} \alpha_{jj}^{(j)}.$$

■

Las hipótesis del Teorema 2.2 son satisfechas para matrices *estrictamente diagonal dominantes* y *matrices definidas positivas*.

Definición 2.2. *Una matriz* $\mathbf{A} \in \mathbb{K}^{n\times n}$ *se llama* estrictamente diagonal dominante *si*

$$\forall i = 1, \dots, n: \quad |\alpha_{ii}| > \sum_{\substack{j=1 \\ j\neq i}}^{n} |\alpha_{ij}|.$$

Una matriz $\mathbf{A} = \mathbf{A}^*$ *se llama* definida positiva *si*

$$\forall \mathbf{x} \in \mathbb{C}^n, \mathbf{x} \neq 0: \quad \mathbf{x}^* \mathbf{A} \mathbf{x} > 0.$$

Teorema 2.3. *Todos los subdeterminantes principales de una matriz* $\mathbf{A} \in \mathbb{C}^{n\times n}$ *estrictamente diagonal dominante son diferentes de cero.*

Demostración. Para demostrar el teorema, es suficiente demostrar que una matriz $\mathbf{A} = (a_{ij}) \in \mathbb{C}^{n\times n}$ estrictamente diagonal dominante es no singular, dado que cada submatriz principal de una matriz estrictamente diagonal dominante es estrictamente diagonal dominante. Para tal efecto, supongamos que $\mathbf{A}$ es una matriz estrictamente diagonal dominante del tamaño $n \times n$, pero que existe un vector $\mathbf{x} = (x_1, \dots, x_n)^{\mathrm{T}} \neq 0$ tal que

$$\mathbf{A}\mathbf{x} = 0. \tag{2.13}$$

Dado que $\mathbf{x} \neq 0$, existe un índice $m \in \{1, \ldots, n\}$ tal que

$$|x_m| = \max\{|x_1|, \ldots, |x_n|\} > 0. \tag{2.14}$$

Evaluando la m-ésima componente de (2.13), tenemos

$$a_{mm}x_m + \sum_{\substack{j \neq m \\ j=1}}^{n} a_{mj}x_j = 0,$$

lo que podemos reescribir como

$$a_{mm}x_m = -\sum_{\substack{j \neq m \\ j=1}}^{n} a_{mj}x_j.$$

Tomando valores absolutos, tenemos

$$|a_{mm}||x_m| \leqslant \sum_{\substack{j \neq m \\ j=1}}^{n} |a_{mj}||x_j|,$$

y usando (2.14), llegamos a

$$|a_{mm}||x_m| \leqslant |x_m| \sum_{\substack{j \neq m \\ j=1}}^{n} |a_{mj}|.$$

Dividiendo por $|x_m|$, obtenemos

$$|a_{mm}| \leqslant \sum_{\substack{j \neq m \\ j=1}}^{n} |a_{mj}|,$$

una contradicción a la diagonaldominancia estricta de $\mathbf{A}$. ■

Teorema 2.4. *Todas las submatrices principales de una matriz definida positiva y hermitiana son definidas positivas y tienen determinante positivo. Todos los valores propios de una matriz definida positiva son positivos.*

Demostración. Sea $\mathbf{A}$ hermitiana, entonces existe una matriz $\mathbf{U}$ unitaria tal que

$$\mathbf{U}^*\mathbf{A}\mathbf{U} = \text{diag } (\lambda_1, \ldots, \lambda_n),$$

donde $\lambda_1, \ldots, \lambda_n$ son los valores propios de $\mathbf{A}$. Sean $\mathbf{y}_1, \ldots, \mathbf{y}_n$ las columnas de $\mathbf{U}$. Ahora, sea $\mathbf{x} := \mathbf{y}_i$ para $i = 1, \ldots, n$. Entonces tenemos

$$0 < \mathbf{x}^*\mathbf{A}\mathbf{x} = \mathbf{y}_i^* \begin{bmatrix} \mathbf{y}_1 & \cdots & \mathbf{y}_n \end{bmatrix} \text{diag}(\lambda_1, \ldots, \lambda_n) \begin{bmatrix} \mathbf{y}_1^* \\ \vdots \\ \mathbf{y}_n^* \end{bmatrix} \mathbf{y}_i = \mathbf{e}_i^{\mathrm{T}} \text{diag } (\lambda_1, \ldots, \lambda_n)\mathbf{e}_i = \lambda_i.$$

Dado que $\det \mathbf{A} = \lambda_1 \cdot \lambda_2 \cdot \ldots \cdot \lambda_n$, resulta $\det \mathbf{A} > 0$. Ahora sea $\mathbf{A}_k$ una submatriz principal de $\mathbf{A}$, es decir,

$$\mathbf{A}_k = \begin{bmatrix} a_{i_1 i_1} & \cdots & a_{i_1 i_k} \\ \vdots & & \vdots \\ a_{i_k i_1} & \cdots & a_{i_k i_k} \end{bmatrix}, \quad \text{y sea} \quad \mathbf{x}_k = \begin{pmatrix} \xi_{i_1} \\ \vdots \\ \xi_{i_k} \end{pmatrix} \neq 0.$$

Ahora sea $\xi_i = 0$ si $i \in \{1, \ldots, n\} \backslash \{i_1, \ldots, i_k\}$ y $\mathbf{x} = (\xi_1, \ldots, \xi_n)^{\mathrm{T}}$. Entonces

$$\mathbf{x}^* \mathbf{A} \mathbf{x} = \mathbf{x}_k^* \mathbf{A}_k \mathbf{x}_k > 0.$$

Ademas, la matriz $\mathbf{A}$ es hermitiana, por lo tanto podemos aplicar la misma conclusión que para la matriz $\mathbf{A}$ a la matriz $\mathbf{A}_k$. ■

2.4. La descomposición de Cholesky

Según el Teorema 2.4, en el caso de una matriz hermitiana definida positiva no es necesario intercambiar columnas y filas durante la ejecución del algoritmo de Gauss, es decir, al calcular la factorización en matrices triangulares. Puesto que

$$\begin{vmatrix} \alpha_{11} & \ldots & \alpha_{1k} \\ \vdots & & \vdots \\ \alpha_{k1} & \ldots & \alpha_{kk} \end{vmatrix} > 0, \quad k = 1, \ldots, n \quad \text{y} \quad \alpha_{ii}^{(i)} = \frac{\begin{vmatrix} \alpha_{11} & \ldots & \alpha_{1i} \\ \vdots & & \vdots \\ \alpha_{i1} & \ldots & \alpha_{ii} \end{vmatrix}}{\begin{vmatrix} \alpha_{11} & \ldots & \alpha_{1,i-1} \\ \vdots & & \vdots \\ \alpha_{i-1,1} & \ldots & \alpha_{i-1,i-1} \end{vmatrix}},$$

tenemos que $\alpha_{ii}^{(i)} > 0$ para $i = 1, \ldots, n$. Finalmente, resulta que todas las matrices "restantes"

$$\left(\alpha_{ij}^{(k+1)}\right), \quad k+1 \leqslant i, j \leqslant n, \quad k = 1, \ldots, n-1$$

son hermitianas, o sea llegamos a $\mathbf{A} = \mathbf{L}\mathbf{R}$ con

$$\mathbf{R} = \begin{bmatrix} \alpha_{11}^{(1)} & \cdots & \cdots & \alpha_{1n}^{(n)} \\ & \alpha_{22}^{(2)} & & \vdots \\ & & \ddots & \vdots \\ & & & \alpha_{nn}^{(n)} \end{bmatrix}, \quad \alpha_{ii}^{(i)} > 0, \quad \mathbf{L} = \begin{bmatrix} 1 & & & \\ \dfrac{\bar{\alpha}_{12}^{(1)}}{\alpha_{11}^{(1)}} & 1 & & \\ \vdots & \ddots & \ddots & \\ \dfrac{\bar{\alpha}_{1n}^{(1)}}{\alpha_{11}^{(1)}} & \cdots & \dfrac{\bar{\alpha}_{n-1,n}^{(n-1)}}{\alpha_{n-1,n-1}^{(n-1)}} & 1 \end{bmatrix}.$$

Entonces, definiendo

$$\mathbf{D} := \operatorname{diag}\left(\frac{1}{\sqrt{\alpha_{11}^{(1)}}}, \ldots, \frac{1}{\sqrt{\alpha_{nn}^{(n)}}}\right),$$

llegamos a

$$\mathbf{A} = \mathbf{L}\mathbf{R} = \mathbf{L}\mathbf{D}^{-1}\mathbf{D}\mathbf{R} = \tilde{\mathbf{L}}\tilde{\mathbf{L}}^*,$$

donde definimos $\tilde{\mathbf{L}} := \mathbf{L}\mathbf{D}^{-1}$. Esa forma simétrica de descomposición en matrices triangulares se llama *descomposición de Cholesky*. Existe solamente para matrices hermitianas definidas positivas. Los elementos $\tilde{\lambda}_{ij}$ de la matriz triangular inferior $\tilde{\mathbf{L}}$ pueden ser calculados sucesivamente por columnas, aprovechando que α_{ij} es el producto escalar de la fila j de $\tilde{\mathbf{L}}$ con la columna i de $\tilde{\mathbf{L}}^*$. El cálculo es único cuando exigimos que $\tilde{\lambda}_{ii} > 0$ para todo i. En este caso, la identidad

$$\alpha_{ii} = \sum_{k=1}^{i} \tilde{\lambda}_{ik}\bar{\tilde{\lambda}}_{ik}$$

nos lleva a la identidad

$$|\lambda_{ii}|^2 = \alpha_{ii} - \sum_{k=1}^{i-1} \left|\tilde{\lambda}_{ik}\right|^2 > 0,$$

de la cual podemos despejar $\tilde{\lambda}_{ii}$ de forma única de la siguiente forma:

$$\tilde{\lambda}_{ii} := \sqrt{\alpha_{ii} - \sum_{k=1}^{i-1} \left|\tilde{\lambda}_{ik}\right|^2}. \tag{2.15}$$

Ahora, para $j > i$ sabemos que

$$\alpha_{ji} = \sum_{k=1}^{i} \tilde{\lambda}_{jk}\bar{\tilde{\lambda}}_{ik},$$

por lo tanto,

$$\tilde{\lambda}_{ji} = \frac{1}{\bar{\tilde{\lambda}}_{ii}} \left(\alpha_{ji} - \sum_{k=1}^{i-1} \tilde{\lambda}_{jk}\bar{\tilde{\lambda}}_{ik} \right), \quad i = 1, \ldots, n. \tag{2.16}$$

Las ecuaciones (2.15) y (2.16) implican que

$$\left|\tilde{\lambda}_{jk}\right| \leqslant \sqrt{\alpha_{jj}}, \quad k = 1, \ldots, j, \quad j = 1, \ldots, n. \tag{2.17}$$

Eso significa que ninguna componente del factor $\tilde{\mathbf{L}}$ de Cholesky es "grande" comparado con los elementos de $\mathbf{A}$, lo que significa que el algoritmo no es muy sensible con respecto a errores de redondeo.

Ejemplo 2.6. *Aplicando las fórmulas* (2.15) *y* (2.16)*, calculamos sucesivamente para la matriz*

$$\mathbf{A} = \begin{bmatrix} 60 & 30 & 20 \\ 30 & 20 & 15 \\ 20 & 15 & 12 \end{bmatrix}$$

los siguientes elementos de $\tilde{\mathbf{L}}$*:*

$$\tilde{\lambda}_{11} = \sqrt{60} = 2\sqrt{15},$$
$$\tilde{\lambda}_{21} = \frac{30}{\sqrt{60}} = \sqrt{15},$$

$$\tilde{\lambda}_{31} = \frac{20}{\sqrt{60}} = 2\sqrt{\frac{5}{3}},$$
$$\tilde{\lambda}_{22} = \sqrt{20 - (\sqrt{15})^2} = \sqrt{5},$$
$$\tilde{\lambda}_{32} = \frac{1}{\sqrt{5}}\left(15 - \sqrt{15}\cdot 2\sqrt{\frac{5}{3}}\right) = \sqrt{5},$$
$$\tilde{\lambda}_{33} = \sqrt{12 - \left(2\sqrt{\frac{5}{3}}\right)^2 - (\sqrt{5})^2} = \frac{1}{\sqrt{3}}.$$

El siguiente teorema es una consecuencia inmediata del Teorema 2.4 y de la Definición 2.2.

Teorema 2.5. *La matriz* $\mathbf{A}$ *es hermitiana y definida positiva si y sólo si ella posee una descomposición* $\mathbf{A} = \mathbf{L}\mathbf{L}^*$*, donde* $\mathbf{L}$ *es una matriz triangular inferior invertible.*

Ejemplo 2.7 (Tarea 4, Curso 2006)**.** *Queremos determinar una matriz triangular inferior* $\mathbf{L}$ *tal que* $\mathbf{A} = \mathbf{L}\mathbf{L}^{\mathrm{T}}$*, donde*

$$\mathbf{A} \in \mathbb{R}^{n\times n}, \quad \mathbf{A} = \begin{bmatrix} 2 & -1 & & \\ -1 & \ddots & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 2 \end{bmatrix}, \tag{2.18}$$

donde el resultado debe ser general con respecto a n*. Despúes de calcular a mano algunos casos con* n *pequeño, una solución razonable es*

$$\mathbf{L} = \begin{bmatrix} \sqrt{2} & & & & \\ -\sqrt{\frac{1}{2}} & \sqrt{\frac{3}{2}} & & & \\ & -\sqrt{\frac{2}{3}} & \sqrt{\frac{4}{3}} & & \\ & & \ddots & \ddots & \\ & & & -\sqrt{\frac{n-1}{n}} & \sqrt{\frac{n+1}{n}} \end{bmatrix}. \tag{2.19}$$

Para verificar que (2.19) *realmente es la solución deseada, definimos los vectores*

$$\mathbf{l}_i := \Bigg(0, \ldots, 0, -\sqrt{\frac{i-1}{i}}, \underbrace{\sqrt{\frac{i+1}{i}}}_{i}, 0, \ldots, 0\Bigg) = -\sqrt{\frac{i-1}{i}}\mathbf{e}_{i-1}^{\mathrm{T}} + \sqrt{\frac{i+1}{i}}\mathbf{e}_i^{\mathrm{T}}.$$

Entonces tenemos

$$\langle \mathbf{l}_i, \mathbf{l}_i\rangle = 2, \quad \langle \mathbf{l}_{i-1}, \mathbf{l}_i\rangle = \langle \mathbf{l}_i, \mathbf{l}_{i+1}\rangle = -1, \quad y\ \langle \mathbf{l}_i, \mathbf{l}_j\rangle = 0\ si\ |i-j| \geqslant 2.$$

Ejemplo 2.8 (Certamen 1, Curso 2006)**.** *Se considera la matriz*

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 0 \\ 2 & 13 & 12 \\ 0 & 12 & 41 \end{bmatrix}.$$

a) *Econtrar una matriz triangular inferior* $\mathbf{L}$ *tal que* $\mathbf{A} = \mathbf{L}\mathbf{L}^{\mathrm{T}}$ *(descomposición de Cholesky).*

b) *Determinar la ultima columna de* $\mathbf{A}^{-1}$*, usando la descomposición de Cholesky.*

c) *Las matrices* $\mathbf{B}$ *y* $\mathbf{L}$ *sean dadas por*

$$\mathbf{B} = \begin{bmatrix} 1 & 3 & 2 & 1 & 2 \\ 3 & 10 & 10 & 6 & 11 \\ 2 & 10 & 24 & 18 & 26 \\ 1 & 6 & 18 & 39 & 24 \\ 2 & 11 & 26 & 24 & 32 \end{bmatrix}, \quad \mathbf{L} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ \alpha & 1 & 0 & 0 & 0 \\ 2 & 4 & \beta & 0 & 0 \\ 1 & 7 & 2 & 5 & 0 \\ 2 & 5 & 1 & \gamma & 1 \end{bmatrix}$$

¿Se pueden encontrar números α*,* β *y* γ *tales que* $\mathbf{A} = \mathbf{L}\mathbf{L}^{\mathrm{T}}$*?*

Solución sugerida.

a) *Calculando sucesivamente los elementos de* $\mathbf{L}$*, resulta*

$$\mathbf{L} = \begin{bmatrix} 1 & 0 & 0 \\ 2 & 3 & 0 \\ 0 & 4 & 5 \end{bmatrix}.$$

b) *Sean*

$$\mathbf{A}^{-1} = \begin{bmatrix} \mathbf{x} & \mathbf{y} & \mathbf{z} \end{bmatrix}, \quad \mathbf{I} = \begin{bmatrix} \mathbf{e}_1 & \mathbf{e}_2 & \mathbf{e}_3 \end{bmatrix}.$$

De la identidad $\mathbf{A}\mathbf{A}^{-1} = \mathbf{I}$ *sacamos que el vector* $\mathbf{z}$ *deseado satisface el sistema lineal* $\mathbf{A}\mathbf{z} = \mathbf{e}_3$*. Para aprovechar la descomposición de Cholesky,* $\mathbf{L}\mathbf{L}^{\mathrm{T}}\mathbf{z} = \mathbf{e}_3$*, determinamos primero un vector* $\mathbf{w}$ *tal que* $\mathbf{L}\mathbf{w} = \mathbf{e}_3$*, luego determinamos* $\mathbf{z}$ *de* $\mathbf{L}^{\mathrm{T}}\mathbf{z} = \mathbf{w}$*. Este procedimiento entrega*

$$w_1 = 0, \quad w_2 = 0, \quad w_3 = \frac{1}{5}; \quad z_3 = \frac{1}{25}; \quad z_2 = -\frac{4}{75}; \quad z_1 = \frac{8}{75}.$$

Entonces, el vector deseado es

$$\mathbf{z} = \frac{1}{75}(3, -4, 8)^{\mathrm{T}}.$$

c) *En clase demostramos que los elementos en la* j*-ésima fila de* $\mathbf{L}$ *son menores o iguales en valor absolutos que la raiz del* j*-ésimo elemento diagonal de* $\mathbf{A}$*. En la fila* 4*, aparece el elemento* $\lambda_{42} = 7$*. Pero* $7^2 = 49 < 39$*, lo cual es el elemento diagonal de* $\mathbf{A}$*, independiente de* α*,* β *y* γ*. Entonces nunca se pueden determinar tales que* $\mathbf{A} = \mathbf{L}\mathbf{L}^{\mathrm{T}}$*.*

Ejemplo 2.9 (Tarea 7, Curso 2006)**.** *Sean*

$$\mathbf{A} = \begin{bmatrix} -1 & 2 & 1 \\ 2 & 0 & 2 \\ 1 & 2 & -1 \end{bmatrix}, \quad \mathbf{I} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}. \tag{2.20}$$

Usando el algoritmo de la descomposición de Cholesky, calcular hasta un decimal

$$t_0 := \min\{t \in \mathbb{R} : \mathbf{A} + t\mathbf{I} \text{ es definida positiva}\}. \tag{2.21}$$

Solución sugerida. *Según el Teorema 2.5, la matriz* $\mathbf{A}$ *es definida positiva si y solo si el algoritmo de Cholesky puede ser ejecutado. Tratamos de hacerlo para la matriz* $\mathbf{A} + t\mathbf{I}$ *e identificamos las restricciones para* t *que aparecen. Recordamos que los elementos diagonales de* $\mathbf{L}$ *deben ser reales. Para la primera columna de* $\mathbf{L} = (\lambda_{ij})$ *obtenemos*

$$\lambda_{11} = \sqrt{t-1}, \tag{2.22}$$

$$\lambda_{21} = \frac{2}{\sqrt{t-1}}, \tag{2.23}$$

$$\lambda_{31} = \frac{1}{\sqrt{t-1}}. \tag{2.24}$$

Obviamente, de (2.22) *obtenemos el requerimiento*

$$t > 1. \tag{2.25}$$

Para los elementos de la segunda columna tenemos

$$\lambda_{22} = \sqrt{t - \frac{4}{t-1}} = \sqrt{\frac{t^2-t-4}{t-1}}, \tag{2.26}$$

$$\lambda_{32} = \sqrt{\frac{t-1}{t^2-t-4}}(2 - \lambda_{21}\bar{\lambda}_{21}) = \sqrt{\frac{t-1}{t^2-t-4}}\frac{2t-4}{t-1} = \frac{2(t-2)}{\sqrt{(t-1)(t^2-t-4)}}. \tag{2.27}$$

La solución de $t^2 - t - 4 = 0$ *es*

$$t = \frac{1}{2} \pm \sqrt{\frac{17}{4}};$$

usando (2.25) *concluimos que*

$$t > \frac{1}{2} + \sqrt{\frac{17}{4}} = 2{,}56155\ldots. \tag{2.28}$$

Finalmente, para el último elemento de $\mathbf{L}$ *tenemos*

$$\lambda_{33} = \sqrt{t - 1 - \frac{1}{t-1} - \frac{4(t-2)^2}{(t-1)(t^2-t-4)}} = \sqrt{\frac{\varphi(t)}{(t-1)(t^2-t-4)}},$$

donde la función

$$\varphi(t) = (t^2-2t)(t^2-t-4) - 4(t-2)^2 = t^4 - 3t^3 - 6t^2 + 24t - 16$$

debe ser positiva. Ahora, tratando $t = 3$*, obtenemos* $\varphi(t) = 2 > 0$*. Usando que para cualquier matriz* $\mathbf{B}$ *definida positiva, tambien* $\mathbf{B} + t\mathbf{I}$ *es definida positiva para* $t > 0$*, tenemos que buscar*

$$t_0 \in \left(\frac{1}{2} + \sqrt{\frac{17}{4}}, 3\right)$$

Usando $\varphi(2{,}75) = -0{,}5742\ldots$, $\varphi(2{,}8) = -0{,}2304\ldots$ *y* $\varphi(2{,}9) = 0{,}7011$, *la respuesta es* $t_0 = 2{,}8\ldots$. *(De hecho,* $\varphi(2\sqrt{2}) = 0$.*)*

Ejemplo 2.10 (Certamen 1, Curso 2010)**.** *Se considera la matriz*

$$\mathbf{A} = \begin{bmatrix} 1 & -2 & 1 & 1 \\ -2 & 8 & -4 & -2 \\ 1 & -4 & 18 & -3 \\ 1 & -2 & -3 & 11 \end{bmatrix}.$$

a) *Econtrar una matriz triangular inferior* $\mathbf{L}$ *tal que* $\mathbf{A} = \mathbf{L}\mathbf{L}^{\mathrm{T}}$ *(descomposición de Cholesky).*
b) *Determinar la tercera columna de* $\mathbf{A}^{-1}$*, usando la descomposición de Cholesky.*
c) *¿Se puede aplicar la descomposición de Cholesky a la siguiente matriz?*

$$\mathbf{B} = \begin{bmatrix} 5 & 4 & 2 & 1 & 1 & 2 \\ 4 & 3 & 1 & -1 & 1 & 1 \\ 2 & 1 & 2 & 0 & 1 & 4 \\ 1 & -1 & 0 & 3 & 2 & -1 \\ 1 & 1 & 1 & 2 & 0 & 1 \\ 2 & 1 & 4 & -1 & 1 & 2 \end{bmatrix}$$

Solución sugerida.

a) *Sea*

$$\mathbf{L} = \begin{bmatrix} l_1 & 0 & 0 & 0 \\ l_2 & l_3 & 0 & 0 \\ l_4 & l_5 & l_6 & 0 \\ l_7 & l_8 & l_9 & l_{10} \end{bmatrix}.$$

Entonces, comparando los elementos de $\mathbf{L}\mathbf{L}^{\mathrm{T}}$*, obtenemos sucesivamente*

$$\begin{aligned}
&l_1^2 = 1 \Rightarrow l_1 = 1, \\
&l_1 l_2 = -2 \Rightarrow l_2 = -2, \\
&l_2^2 + l_3^2 = 8 \Rightarrow l_3 = 2, \\
&l_1 l_4 = 1 \Rightarrow l_4 = 1, \\
&l_2 l_4 + l_3 l_5 = -4 \Rightarrow l_5 = \frac{1}{2}(-4 - (-2)) = -1, \\
&l_4^2 + l_h^2 + l_6^2 = 18 \Rightarrow l_6 = \sqrt{18 - 1 - 1} = 4, \\
&l_1 l_7 = 1 \Rightarrow l_7 = 1, \\
&l_2 l_7 + l_3 l_8 = -2 \Rightarrow l_8 = \frac{1}{2}(-2 - (-2) \cdot 1) = 0, \\
&l_4 l_7 + l_5 l_8 + l_6 l_9 = -3 \Rightarrow l_9 = \frac{1}{4}(-3 - 1 \cdot 1) = -1, \\
&l_7^2 + l_8^2 + l_9^2 + l_{10}^2 = 11 \Rightarrow l_{10} = \sqrt{11 - 1 - 1} = 3,
\end{aligned}$$

es decir

$$\mathbf{L} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -2 & 2 & 0 & 0 \\ 1 & -1 & 4 & 0 \\ 1 & 0 & -1 & 3 \end{bmatrix}.$$

b) *Sea* $\mathbf{z}$ *la tercera columna de* $\mathbf{A}^{-1}$*, entonces* $\mathbf{z}$ *es la solución del sistema lineal* $\mathbf{Az} = \mathbf{e}_3 = (0,0,1,0)^{\mathrm{T}}$*. Utilizando la descomposición de Cholesky, podemos determinar* $\mathbf{Z}$ *resolviendo primeramente el sistema* $\mathbf{Ly} = \mathbf{e}_3$ *y luego* $\mathbf{L}^{\mathrm{T}}\mathbf{z} = \mathbf{y}$*. Así obtenemos*

$$\mathbf{y} = \frac{1}{12}\begin{pmatrix} 0 \\ 0 \\ 3 \\ 1 \end{pmatrix}; \qquad \mathbf{z} = \frac{1}{144}\begin{pmatrix} -4 \\ 5 \\ 10 \\ 4 \end{pmatrix}.$$

c) *No. La matriz contiene la submatriz principal*

$$\boldsymbol{\Xi} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} = \begin{bmatrix} 5 & 4 \\ 4 & 3 \end{bmatrix}$$

con $\det \boldsymbol{\Xi} = -1 < 0$*; según el Teorema 2.4,* $\mathbf{B}$ *no es definida positiva y no se puede aplicar la descomposición de Cholesky.*

2.5. Aplicaciones de la descomposición triangular y casos especiales

Cuando conocemos una descomposición $\mathbf{PAQ} = \mathbf{LR}$ de una matriz $\mathbf{A}$ donde $\mathbf{P}$ y $\mathbf{Q}$ son matrices de permutación, podemos fácilmente resolver el sistema lineal $\mathbf{Ax} = \mathbf{b}$ para una parte $\mathbf{b}$ derecha arbitraria: en virtud de $\mathbf{A} = \mathbf{P}^{\mathrm{T}}\mathbf{LRQ}^{\mathrm{T}}$, tenemos que

$$\mathbf{Ax} = \mathbf{b} \Longleftrightarrow \mathbf{P}^{\mathrm{T}}\mathbf{LRQ}^{\mathrm{T}}\mathbf{x} = \mathbf{b} \Longleftrightarrow \mathbf{LRQ}^{\mathrm{T}}\mathbf{x} = \mathbf{Pb}.$$

Con la notación $\mathbf{d} := \mathbf{Pb}$, $\mathbf{y} := \mathbf{Q}^{\mathrm{T}}\mathbf{x}$ y $\mathbf{z} := \mathbf{Ry}$ procedemos de la siguiente forma:

1. Definimos
$$\delta_i := \beta_{\pi_i}, \quad i = 1, \ldots, n, \quad \text{donde } \mathbf{d} = \begin{pmatrix} \delta_1 \\ \vdots \\ \delta_n \end{pmatrix}, \ \mathbf{b} = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_n \end{pmatrix}, \ \mathbf{P} = \begin{bmatrix} \mathbf{e}_{\pi_1}^{\mathrm{T}} \\ \vdots \\ \mathbf{e}_{\pi_n}^{\mathrm{T}} \end{bmatrix}.$$
2. Resolver $\mathbf{Lz} = \mathbf{d}$ para determinar $\mathbf{z}$.
3. Resolver $\mathbf{Ry} = \mathbf{z}$ para determinar $\mathbf{y}$.
4. Obtenemos
$$\xi_{\sigma_i} = \eta_i, \quad i = 1, \ldots, n, \quad \text{donde } \mathbf{y} = \begin{pmatrix} \eta_1 \\ \vdots \\ \eta_n \end{pmatrix}, \ \mathbf{x} = \begin{pmatrix} \xi_1 \\ \vdots \\ \xi_n \end{pmatrix}, \ \mathbf{Q} = \begin{bmatrix} \mathbf{e}_{\sigma_1} & \cdots & \mathbf{e}_{\sigma_n} \end{bmatrix}.$$

Eso significa que es posible tratar primero la matriz $\mathbf{A}$ por el algoritmo de Gauss para determinar su descomposición triangular, y luego resolver el sistema $\mathbf{Ax} = \mathbf{b}$ siguiendo los pasos 1 a 4. En comparación con el algoritmo original, este procedimiento no significa ningún aumento del tiempo computacional ni del espacio de almacenaje. La descomposición

triangular también puede ser usada para invertir la matriz $\mathbf{A}$, aunque esta tarea se presenta solo rara vez.

Sea $\mathbf{PAQ} = \mathbf{LR}$, donde $\mathbf{P}$ y $\mathbf{Q}$ son matrices de permutación, $\mathbf{L}$ es triangular inferior y $\mathbf{R}$ es triangular superior. Entonces sabemos que

$$\mathbf{A} = \mathbf{P}^{\mathrm{T}}\mathbf{LRQ}^{\mathrm{T}}, \quad \mathbf{A}^{-1} = \mathbf{QR}^{-1}\mathbf{L}^{-1}\mathbf{P}. \tag{2.29}$$

Las matrices $\mathbf{R}$ y $\mathbf{L}$ pueden ser invertidas (sin espacio de almacenaje adicional) si formamos sucesivamente las columnas $n, n-1, \ldots, 1$ de $\mathbf{R}^{-1}$ y $1, 2, \ldots, n$ de $\mathbf{L}^{-1}$ (aprovechando que la diagonal $(1, \ldots, 1)$ de $\mathbf{L}$ es conocida). Al formar el producto $\mathbf{R}^{-1}\mathbf{L}^{-1}$ podemos usar la estructura especial de la matriz. Si λ'_{ij}, ϱ'_{ik} y α'_{ij} son los elementos de $\mathbf{L}^{-1}$, $\mathbf{R}^{-1}$ y $\mathbf{A}^{-1}$, respectivamente, sabemos que

$$\alpha'_{ij} = \sum_{k=\max\{i,j\}}^{n} \varrho'_{ik}\lambda'_{kj}, \quad i, j = 1, \ldots, n,$$

donde $\lambda'_{jj} = 1$. Finalmente, hay que aplicar las permutaciones decritas por $\mathbf{P}$ y $\mathbf{Q}$. dado que $\mathbf{P} = \mathbf{P}_{n-1} \cdot \ldots \cdot \mathbf{P}_1$ y $\mathbf{Q} = \mathbf{Q}_1 \cdot \ldots \cdot \mathbf{Q}_{n-1}$, (2.29) implica que los intercambios de filas aplicados durante la descomposición triangular deben ser aplicados en el orden revertido a las columnas del producto, y analogamente los intercambios de las columnas a las filas del producto. Se puede demostrar que el esfuerzo computacional es de $n^3 + \mathcal{O}(n^2)$ operaciones del tipo $\alpha := \beta + \gamma\delta$ o $\alpha := \beta + \gamma/\delta$.

Definición 2.3. *Una matriz* $\mathbf{A} \in \mathbb{K}^{n\times n}$ *se llama* matriz casi triangular *o* matriz de Hessenberg *si* $\alpha_{ij} = 0$ *para* $j < i - 1$.

Definición 2.4. *Una matriz* $\mathbf{A} \in \mathbb{K}^{n\times n}$ *se llama* (p, q)-matriz de banda *si* $\alpha_{ij} = 0$ *para* $j < i - p$ *y* $j > i + q$.

En las aplicaciones frecuentamente aparecen matrices *tridiagonales* con $p = q = 1$. Si no se usa el intercambio de columnas para una matriz de Hessenberg, no es necesario eliminar en cada paso la matriz restante entera, sino que solo una fila de ella. (Por ejemplo, la desconocida ξ_1 aparece solamente en la primera y la segunda ecuación.) Eso significa que abajo de la diagonal, la matriz $\mathbf{L}$ tiene a lo más un elemento diferente de cero en la primera subdiagonal. En el caso que no se necesita ningún intercambio, la matriz $\mathbf{L}$ es una matriz bidiagonal ($p = 1$, $q = 0$).

Si para una matriz de banda no se necesita ningún intercambio, la matriz $\mathbf{L}$ tiene $p + 1$ bandas y la matriz $\mathbf{R}$ tiene $q + 1$ bandas, o sea la información sobre la descomposición ocupa solo $n \cdot (p + q + 1)$ elementos de almacenaje.

Capítulo 3

Métodos directos para la solución de sistemas lineales (Parte II)

3.1. Normas de vectores y matrices

Definición 3.1. *Sea* $\mathbf{A} \in \mathbb{C}^{n\times n}$*. Se define el* espectro *de* $\mathbf{A}$*, denotado* $\sigma(\mathbf{A})$*, como el conjunto de todos los valores propios de* $\mathbf{A}$*. Además, se llama* radio espectral *de* $\mathbf{A}$ *a*

$$r_\sigma(\mathbf{A}) := \max_{\lambda\in\sigma(\mathbf{A})} |\lambda|.$$

Definición 3.2. *Sea* V *un espacio vectorial sobre el cuerpo* $\mathbb{C}$*. Se llama* norma de vector *a toda aplicación* $\|\cdot\| : V \to \mathbb{R}_0^+$ *tal que para todo* $\mathbf{x}, \mathbf{y} \in V$ *y* $\lambda \in \mathbb{C}$ *se verifica:*

1. $\|\mathbf{x}\| > 0$ *si* $\mathbf{x} \neq 0$ *y* $\|\mathbf{x}\| = 0$ *si y sólo si* $\mathbf{x} = 0$*.*
2. $\|\lambda\mathbf{x}\| = |\lambda|\|\mathbf{x}\|$*.*
3. $\|\mathbf{x} + \mathbf{y}\| \leqslant \|\mathbf{x}\| + \|\mathbf{y}\|$*.*

Damos a continuación algunos ejemplos de normas para el espacio $\mathbb{C}^n$:

$$\|\mathbf{x}\|_1 := \sum_{i=1}^{n} |x_i|, \tag{3.1}$$

$$\|\mathbf{x}\|_2 := (\mathbf{x}^*\mathbf{x})^{1/2} = \left(\sum_{i=1}^{n} |x_i|^2\right)^{1/2}, \tag{3.2}$$

$$\|\mathbf{x}\|_\infty := \max_{1\leqslant i\leqslant n} |x_i|, \tag{3.3}$$

a las que nos referimos como "norma 1", "norma 2" y "norma ∞", respectivamente; en general, para $p \in [1, \infty)$ definimos

$$\|\mathbf{x}\|_p := \left(\sum_{i=1}^{n} |x_i|^p\right)^{1/p} \tag{3.4}$$

como "norma p".

Teorema 3.1. *Una norma* $\|\cdot\| : \mathbb{K}^n \to \mathbb{R}_0^+$ *es una función continua.*

Demostración. Tomamos en cuenta que para $\mathbf{x} = (\xi_1, \ldots, \xi_n)^{\mathrm{T}}$, $\mathbf{y} = (\eta_1, \ldots, \eta_n)^{\mathrm{T}}$

$$\begin{aligned}\big|\|\mathbf{x}\| - \|\mathbf{y}\|\big| &= \big|\|\mathbf{x}\| - \|-\mathbf{y}\|\big| \\ &\leqslant \|\mathbf{x} - \mathbf{y}\| \\ &\leqslant \sum_{i=1}^{n} |\xi_i - \eta_i|\|\mathbf{e}_i\|\end{aligned}$$

$$\leqslant \max_{1\leqslant i\leqslant n} \|\mathbf{e}_i\| \max_{1\leqslant i\leqslant n} |\xi_i - \eta_i|,$$

y aplicamos la definición de la continuidad. ■

Teorema 3.2. *Si $\|\cdot\|$ y $\|\cdot\|^*$ son dos normas sobre $\mathbb{K}^n$, entonces existen constantes $m, M > 0$ tales que*

$$\forall \mathbf{x} \in \mathbb{K}^n : \quad m\|\mathbf{x}\| \leqslant \|\mathbf{x}\|^* \leqslant M\|\mathbf{x}\|. \tag{3.5}$$

Demostración. Sean $\|\cdot\| := \|\cdot\|_\infty$ y $\|\cdot\|^*$ arbitraria (el caso general sigue por transitividad). Sea

$$\mathcal{S} := \Big\{ \mathbf{x} \in \mathbb{K}^n \;\Big|\; \mathbf{x} = (\xi_1, \ldots, \xi_n)^{\mathrm{T}},\ \max_{1\leqslant i\leqslant n} |\xi_i| = 1 \Big\}.$$

El conjunto $\mathcal{S}$ es compacto. Puesto que $\|\cdot\|^*$ es continua, existen $\mathbf{x}_m, \mathbf{x}_M \in \mathcal{S}$ tales que

$$\|\mathbf{x}_m\|^* = \min_{\mathbf{x}\in\mathcal{S}} \|\mathbf{x}\|^* =: m, \quad \|\mathbf{x}_M\|^* = \max_{\mathbf{x}\in\mathcal{S}} \|\mathbf{x}\|^* =: M.$$

Entonces,

$$\forall \mathbf{x} \in \mathbb{K}^n, \mathbf{x} \neq 0 : \quad m \leqslant \left\| \frac{\mathbf{x}}{\|\mathbf{x}\|_\infty} \right\|^* \leqslant M,$$

lo que implica (3.5) en virtud de la homogeneidad (item 2. en la Definición 3.2). (Para $\mathbf{x} = 0$, (3.5) es trivial.) ■

Definición 3.3. *En el espacio $\mathbb{C}^{n\times n}$ se llama* norma de matriz *a toda aplicación $\|\cdot\| : \mathbb{C}^{n\times n} \to \mathbb{R}_0^+$ tal que para todas matrices $\mathbf{A}, \mathbf{B} \in \mathbb{C}^{n\times n}$ y todo $\lambda \in \mathbb{C}$ se verifica:*

1. $\|\mathbf{A}\| > 0$ *si* $\mathbf{A} \neq 0$ *y* $\|\mathbf{A}\| = 0$ *si y sólo si* $\mathbf{A} = 0$.
2. $\|\lambda\mathbf{A}\| = |\lambda|\|\mathbf{A}\|$.
3. $\|\mathbf{A} + \mathbf{B}\| \leqslant \|\mathbf{A}\| + \|\mathbf{B}\|$.
4. $\|\mathbf{AB}\| \leqslant \|\mathbf{A}\|\|\mathbf{B}\|$.

En correspondencia con cada norma vectorial $\|\cdot\|$ de $\mathbb{C}^n$, se define una norma para matrices $\mathbf{A} \in \mathbb{C}^{n\times n}$ por medio de la expresión

$$\|\mathbf{A}\| := \max_{\mathbf{x}\neq 0} \frac{\|\mathbf{Ax}\|}{\|\mathbf{x}\|} = \max_{\|\mathbf{x}\|=1} \|\mathbf{Ax}\|. \tag{3.6}$$

Esta norma matricial se dice *inducida* por la norma vectorial. En particular, las normas vectoriales 1, 2 e ∞ inducen las siguientes normas matriciales, a las cuales igualmente nos referimos como "norma 1", "norma 2" y "norma ∞", respectivamente:

$$\|\mathbf{A}\|_1 := \max_{\|\mathbf{x}\|_1=1} \|\mathbf{Ax}\|_1 := \max_{1\leqslant j\leqslant n} \sum_{i=1}^{n} |a_{ij}|, \tag{3.7}$$

$$\|\mathbf{A}\|_2 := \max_{\|\mathbf{x}\|_2=1} \|\mathbf{Ax}\|_2, \tag{3.8}$$

$$\|\mathbf{A}\|_\infty := \max_{\|\mathbf{x}\|_\infty=1} \|\mathbf{Ax}\|_\infty = \max_{1\leqslant i\leqslant n} \sum_{j=1}^{n} |a_{ij}|. \tag{3.9}$$

También se define sobre $\mathbb{C}^{n\times n}$ la siguiente norma, la cual *no* es inducida por una norma vectorial,

$$\|\mathbf{A}\|_{\mathrm{F}} := \left(\sum_{i,j=1}^{n} |a_{ij}|^2\right)^{1/2}, \tag{3.10}$$

y que se llama *norma de Frobenius.*

Definición 3.4. *Una norma de matriz se dice* compatible *con una norma de vector si, para cada* $\mathbf{A} \in \mathbb{C}^{n\times n}$ *y para cada* $\mathbf{x} \in \mathbb{C}^n$*, se tiene que*

$$\|\mathbf{Ax}\|_{\text{vector}} \leqslant \|\mathbf{A}\|_{\text{matriz}}\|\mathbf{x}\|_{\text{vector}}.$$

Note que de la definición (3.6) de desprende que cada norma matricial inducida por una norma vectorial es compatible con la norma vectorial que la induce. Así tenemos, en particular, que las normas matriciales 1, 2 e ∞ son compatibles con las correspondientes normas vectoriales 1, 2 e ∞. Por otra parte, la norma de Frobenius, que como indicamos no es inducida por norma vectorial alguna, es compatible con la norma vectorial 2.

Teorema 3.3. *Si* $\mathbf{A} \in \mathbb{C}^{n\times n}$*, entonces*

$$\|\mathbf{A}\|_2 = \sqrt{r_\sigma(\mathbf{A}^*\mathbf{A})}.$$

Demostración. Es claro que $\mathbf{A}^*\mathbf{A}$ es una matriz hermitiana. Por el Teorema del Eje Principal sabemos que $\mathbf{A}^*\mathbf{A}$ tiene n vectores propios que forman una base ortonormal de $\mathbb{C}^n$.

Veremos a continuación que los valores propios de $\mathbf{A}^*\mathbf{A}$ son además no negativos. En efecto, si λ es un valor propio de $\mathbf{A}^*\mathbf{A}$ y $\mathbf{v}$ es un correspondiente vector propio asociado, entonces

$$\mathbf{A}^*\mathbf{Av} = \lambda\mathbf{v}$$

y además,

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

Como $\|\mathbf{v}\| \neq 0$, de esta última relación deducimos que

$$\lambda = \frac{\|\mathbf{Av}\|_2^2}{\|\mathbf{v}\|_2^2} \geqslant 0. \tag{3.11}$$

Ahora sean $\lambda_1 \geqslant \lambda_2 \geqslant \ldots \geqslant \lambda_n$ son los valores propios de $\mathbf{A}^*\mathbf{A}$ y $\{\mathbf{v}^{(1)}, \mathbf{v}^{(2)}, \ldots, \mathbf{v}^{(n)}\}$ un conjunto de vectores propios asociados que forman una base ortonormal de $\mathbb{C}^n$. Entonces, para $\mathbf{x} \in \mathbb{C}^n\backslash\{0\}$ existen escalares $\alpha_1, \ldots, \alpha_n$ tales que

$$\mathbf{x} = \sum_{j=1}^{n} \alpha_j \mathbf{v}^{(j)}. \tag{3.12}$$

Por otro lado,

$$\|\mathbf{Ax}\|_2^2 = (\mathbf{Ax})^*\mathbf{Ax} = \mathbf{x}^*(\mathbf{A}^*\mathbf{A})\mathbf{x}. \tag{3.13}$$

Remplazando (3.12) en (3.13) y reordenando,

$$\|\mathbf{Ax}\|_2^2 = \left(\sum_{j=1}^{n} \bar{\alpha}_j \mathbf{v}^{(j)*}\right) (\mathbf{A}^*\mathbf{A}) \left(\sum_{i=1}^{n} \alpha_i \mathbf{v}^{(i)}\right) = \sum_{j=1}^{n}\sum_{i=1}^{n} \lambda_i \bar{\alpha}_j \alpha_i \mathbf{v}^{(j)*}\mathbf{v}^{(i)},$$

y como $\mathbf{v}^{(j)*}\mathbf{v}^{(i)} = \delta_{ij}$, entonces

$$\|\mathbf{Ax}\|_2^2 = \sum_{i=1}^{n} |\alpha_i|^2 \lambda_i \leqslant \lambda_1 \sum_{i=1}^{n} |\alpha_i|^2. \tag{3.14}$$

En forma análoga calculamos

$$\|\mathbf{x}\|_2^2 = \mathbf{x}^*\mathbf{x} = \sum_{j=1}^{n}\sum_{i=1}^{n} \bar{\alpha}_i \alpha_j \mathbf{v}^{(j)*}\mathbf{v}^{(i)} = \sum_{i=1}^{n} |\alpha_i|^2. \tag{3.15}$$

De (3.14) y (3.15) concluimos que

$$\forall \mathbf{x} \in \mathbb{C}^n \backslash \{0\} : \quad \frac{\|\mathbf{Ax}\|_2}{\|\mathbf{x}\|_2} \leqslant \sqrt{\lambda_1},$$

lo cual equivale a

$$\|\mathbf{A}\|_2 := \max_{\mathbf{x}\neq 0} \frac{\|\mathbf{Ax}\|_2}{\|\mathbf{x}\|_2} \leqslant \sqrt{\lambda_1}. \tag{3.16}$$

Para mostrar que la cota $\sqrt{\lambda_1}$ se alcanza, basta exhibir un vector no nulo para el cual la igualdad se cumpla en (3.16). Con esta finalidad sea $\mathbf{v}_1$ un vector propio asociado a λ_1. Entonces, de (3.11) obtenemos inmediatamente que

$$\sqrt{\lambda_1} = \frac{\|\mathbf{Av}_1\|_2}{\|\mathbf{v}_1\|_2},$$

esto es, el máximo de (3.16) se alcanza en $\mathbf{x} = \mathbf{v}_1$ y es igual a $\sqrt{\lambda_1}$. Notando que $\lambda_1 = r_\sigma(\mathbf{A}^*\mathbf{A})$ se concluye la demostración. ■

Corolario 3.1. *Si* $\mathbf{A}$ *es hermitiana, entonces* $\|\mathbf{A}\|_2 = r_\sigma(\mathbf{A})$.

Demostración. Puesto que $\|\mathbf{A}\|_2 = \sqrt{r_\sigma(\mathbf{A}^*\mathbf{A})}$ y $\mathbf{A}^* = \mathbf{A}$, entonces $\|\mathbf{A}\|_2 = \sqrt{r_\sigma(\mathbf{A}^2)}$. Como se tiene que

$$r_\sigma(\mathbf{A}^2) = (r_\sigma(\mathbf{A}))^2,$$

inmediatamente se llega a

$$\|\mathbf{A}\|_2 = \sqrt{(r_\sigma(\mathbf{A}))^2} = r_\sigma(\mathbf{A}),$$

que es lo que se quería demostrar. ■

Teorema 3.4. *Sea* $\|\cdot\|$ *alguna norma vectorial sobre* $\mathbb{C}^n$ *y* $\|\cdot\|$ *la norma matricial inducida. En este caso,*

$$\forall \mathbf{B} \in \mathbb{C}^{n\times n} : \quad r_\sigma(\mathbf{B}) \leqslant \|\mathbf{B}\|.$$

Demostración. Sea λ un valor propio de $\mathbf{B}$ con $|\lambda| = r_\sigma(\mathbf{B})$ y $\mathbf{x} \neq 0$ un vector propio asociado. Entonces

$$\|\lambda\mathbf{x}\| = |\lambda|\|\mathbf{x}\| = r_\sigma(\mathbf{B})\|\mathbf{x}\| = \|\mathbf{B}\mathbf{x}\| \leqslant \|\mathbf{B}\|\|\mathbf{x}\|,$$

es decir,

$$\|\mathbf{B}\| \geqslant \frac{\|\mathbf{B}\mathbf{x}\|}{\|\mathbf{x}\|} = r_\sigma(\mathbf{B}).$$

■

Teorema 3.5. *Sea $\mathbf{B} \in \mathbb{C}^{n\times n}$ una matriz arbitraria, con $\varepsilon > 0$ arbitrario dado. Entonces existe una norma vectorial $\|\cdot\|_{\mathbf{B}}$ sobre $\mathbb{C}^n$ tal que para la norma matricial asociada,*

$$\|\mathbf{B}\|_{\mathbf{B}} \leqslant r_\sigma(\mathbf{B}) + \varepsilon.$$

Demostración. Según el Teorema 1.1 (sobre la forma normal de Schur), existe una matriz $\mathbf{U}$ unitaria tal que

$$\mathbf{U}^*\mathbf{B}\mathbf{U} = \begin{bmatrix} \lambda_1 & * & \cdots & * \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & * \\ 0 & \cdots & 0 & \lambda_n \end{bmatrix} =: (\varrho_{ik}).$$

Para los elementos arriba de la diagonal (ϱ_{ik} con $k > i$) sabemos que

$$\varrho_{ik} = \sum_{s=1}^{n}\sum_{l=1}^{n} \bar{\nu}_{li}\beta_{ls}\nu_{sk}, \quad \mathbf{U} =: (\nu_{sk}).$$

Eso significa

$$|\varrho_{ik}| \leqslant n^2\beta, \quad \beta := \max_{1\leqslant l,s\leqslant n} |\beta_{ls}|.$$

Sean

$$\delta := \min\left\{\frac{\varepsilon}{n^3(\beta+1)}, 1\right\}, \quad \mathbf{D} := \operatorname{diag}(1, \delta, \ldots, \delta^{n-1}).$$

En este caso,

$$\mathbf{D}^{-1}\mathbf{U}^*\mathbf{B}\mathbf{U}\mathbf{D} = (\varrho_{ik}\delta^{k-i}),$$

entonces

$$\begin{aligned} \|\mathbf{D}^{-1}\mathbf{U}^*\mathbf{B}\mathbf{U}\mathbf{D}\|_\infty &= \max_{1\leqslant i\leqslant n} \sum_{k=i}^{n} |\varrho_{ik}\delta^{k-i}| \\ &\leqslant \max_{1\leqslant i\leqslant n} |\varrho_{ii}| + \delta \max_{1\leqslant i\leqslant n} \sum_{k=i+1}^{n} |\varrho_{ik}| \\ &\leqslant r_\sigma(\mathbf{B}) + \delta(n-1)n^2\beta \\ &\leqslant r_\sigma(\mathbf{B}) + \varepsilon\frac{n^2(n-1)\beta}{n^3(\beta+1)} < r_\sigma(\mathbf{B}) + \varepsilon. \end{aligned}$$

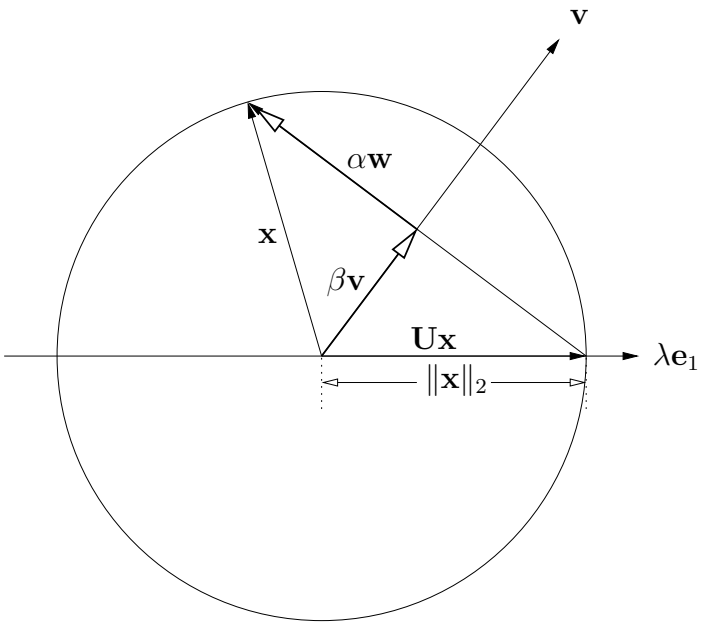


FIGURA 3.1. Ilustración de (3.17), (3.18) (demostración del Teorema 3.6).

Por otro lado,

$$\|\mathbf{D}^{-1}\mathbf{U}^*\mathbf{B}\mathbf{U}\mathbf{D}\|_\infty = \max_{\mathbf{x}\neq 0} \frac{\|\mathbf{D}^{-1}\mathbf{U}^*\mathbf{B}\mathbf{U}\mathbf{D}\mathbf{x}\|_\infty}{\|\mathbf{x}\|_\infty} = \max_{\mathbf{y}\neq 0} \frac{\|\mathbf{D}^{-1}\mathbf{U}^*\mathbf{B}\mathbf{y}\|_\infty}{\|\mathbf{D}^{-1}\mathbf{U}^*\mathbf{y}\|_\infty}.$$

Entonces la norma deseada es $\|\mathbf{x}\|_{\mathbf{B}} := \|\mathbf{D}^{-1}\mathbf{U}^*\mathbf{x}\|_\infty$. ■

Teorema 3.6. *Sea* $\mathbf{x} \in \mathbb{C}^n$, $\mathbf{x} \neq 0$. *Entonces existe una matriz* $\mathbf{U}$ *unitaria y hermitiana tal que*

$$\mathbf{U}\mathbf{x} = \exp(\mathrm{i}\delta)\|\mathbf{x}\|_2\mathbf{e}_1 \quad (\delta \in \mathbb{R} \text{ apropiado}).$$

Específicamente para $\mathbf{x} \in \mathbb{R}^n$, *podemos elegir*

$$\mathbf{U} = \mathbf{I} - 2\mathbf{u}\mathbf{u}^{\mathrm{T}}, \quad \text{donde } \mathbf{u}^{\mathrm{T}}\mathbf{u} = 1 \text{ (Matriz de Householder)}.$$

Demostración. Sea

$$\mathbf{x} \neq \exp(\mathrm{i}\alpha)\|\mathbf{x}\|_2\mathbf{e}_1,$$

sino ponemos

$$\mathbf{U} := \mathbf{I} - 2\mathbf{e}_1\mathbf{e}_1^{\mathrm{T}}.$$

Podemos ilustrar el problema de la siguiente forma: estamos buscando una transformación unitaria del vector $\mathbf{x}$ al primer eje de coordenadas. Cuando $\mathbf{x}$ ya es un múltiple de $\mathbf{e}_1$, podríamos elegir $\mathbf{U} = \mathbf{I}$. Pero, para lograr una fórmula única para todo vector $\mathbf{x}$, ponemos

$$\mathbf{U} := \mathbf{I} - 2\mathbf{e}_1\mathbf{e}_1^{\mathrm{T}}$$

(cambio de signo). Ahora, si $\mathbf{x}$ *no* es un múltiple de $\mathbf{e}_1$, podriamos transformar $\mathbf{x}$ a $\|\mathbf{x}\|_2\mathbf{e}_1$ a través de una rotación. Pero como exigimos que $\mathbf{U}^* = \mathbf{U}$, $\mathbf{U}^2 = \mathbf{I}$, la aplicación deseada debe ser involutiva, es decir, una reflexión (ver Figura 3.1). Ahora que conocemos el resultado de la aplicación a un vector $\mathbf{x}$, podemos elegir $\mathbf{U}$ como la reflexión de $\mathbf{x}$ en un hiperplano con vector normal $\mathbf{w}$, es decir, cuando

$$\mathbf{x} = \alpha\mathbf{w} + \beta\mathbf{v}, \quad \mathbf{w}^*\mathbf{v} = 0 \quad (\mathbf{v} \text{ arbitrario}), \tag{3.17}$$

debemos tener

$$\mathbf{U}\mathbf{x} = -\alpha\mathbf{w} + \beta\mathbf{v}. \tag{3.18}$$

Esto se satisface para $\mathbf{U}\mathbf{w} = -\mathbf{w}$ y $\mathbf{U}\mathbf{v} = \mathbf{v}$. Si $\{\mathbf{w}, \mathbf{v}_1, \ldots, \mathbf{v}_{n-1}\}$ es un sistema ortonormal completo de $\mathbb{C}^n$, entonces

$$\begin{aligned}\mathbf{U} &= \begin{bmatrix}-\mathbf{w} & \mathbf{v}_1 & \cdots & \mathbf{v}_{n-1}\end{bmatrix}\begin{bmatrix}\mathbf{w} & \mathbf{v}_1 & \cdots & \mathbf{v}_{n-1}\end{bmatrix}^* \\ &= \begin{bmatrix}\mathbf{w} & \mathbf{v}_1 & \cdots & \mathbf{v}_{n-1}\end{bmatrix}\begin{bmatrix}\mathbf{w} & \mathbf{v}_1 & \cdots & \mathbf{v}_{n-1}\end{bmatrix}^* - 2\begin{bmatrix}\mathbf{w} & 0 & \cdots & 0\end{bmatrix}\begin{bmatrix}\mathbf{w} & \mathbf{v}_1 & \cdots & \mathbf{v}_{n-1}\end{bmatrix}^* \\ &= \mathbf{I} - 2\mathbf{w}\mathbf{w}^*,\end{aligned}$$

con $\mathbf{w}^*\mathbf{w} = 1$. Ahora falta determinar el vector $\mathbf{w}$. Queremos que

$$\mathbf{U}\mathbf{x} = \mathbf{x} - 2(\mathbf{w}\mathbf{w}^*)\mathbf{x} = \exp(\mathrm{i}\alpha)\|\mathbf{x}\|_2\mathbf{e}_1.$$

Entonces, cuando $\mathbf{w} = (w_1, \ldots, w_n)^{\mathrm{T}}$ y $\tau := \mathbf{w}^*\mathbf{x}$, se debe cumplir

$$\mathbf{x} - 2\tau\mathbf{w} = \begin{pmatrix}\exp(\mathrm{i}\alpha)\|\mathbf{x}\|_2 \\ 0 \\ \vdots \\ 0\end{pmatrix},$$

lo que significa que

$$w_2 = \frac{\xi_2}{2\tau}, \ldots, w_n = \frac{\xi_n}{2\tau}.$$

Queda para determinar τ y w_1. Sabemos que

$$\xi_1 - 2\tau w_1 = \exp(\mathrm{i}\delta)\|\mathbf{x}\|_2 \quad (\delta \text{ apropiado}),$$

lo que es equivalente a

$$w_1 = \frac{\xi_1 - \exp(\mathrm{i}\delta)\|\mathbf{x}\|_2}{2\tau}.$$

Si $\xi_1 = \exp(\mathrm{i}\alpha)|\xi_1|$, sea $\delta := \alpha + \pi$, entonces $-\exp(\mathrm{i}\delta) = \exp(\mathrm{i}\alpha)$ y luego

$$w_1 = \frac{\exp(\mathrm{i}\alpha)(|\xi_1| + \|\mathbf{x}\|_2)}{2\tau}.$$

En virtud de $\mathbf{w}^*\mathbf{w} = 1$ tenemos que

$$\frac{1}{4|\tau|^2}\big(|\xi_1|^2 + 2|\xi_1|\|\mathbf{x}\|_2 + \|\mathbf{x}\|_2^2 + |\xi_2|^2 + \cdots + |\xi_n|^2\big) = 1,$$

lo que implica

$$4|\tau|^2 = 2\|\mathbf{x}\|_2\big(|\xi_1| + \|\mathbf{x}\|_2\big),$$

es decir, podemos elegir

$$2\tau = \sqrt{2\|\mathbf{x}\|_2\big(|\xi_1| + \|\mathbf{x}\|_2\big)}.$$

En este caso, $\mathbf{w}$ es determinado unicamente hasta un factor complejo de valor absoluto 1. Eso entrega

$$\mathbf{w} = \frac{1}{\sqrt{2\|\mathbf{x}\|_2(|\xi_1| + \|\mathbf{x}\|_2)}} \begin{pmatrix} \exp(\mathrm{i}\alpha)(|\xi_1| + \|\mathbf{x}\|_2) \\ \xi_2 \\ \vdots \\ \xi_n \end{pmatrix}$$

con $\xi_1 = \exp(\mathrm{i}\alpha)|\xi_1|$ y $\mathbf{U}\mathbf{x} = -\exp(\mathrm{i}\alpha)\|\mathbf{x}\|_2\mathbf{e}_1$. Especialmente para $\mathbf{x} \in \mathbb{R}^n$ tenemos que $\exp(\mathrm{i}\alpha) = \pm 1$, entonces $\mathbf{U} = \mathbf{I} - 2\mathbf{u}\mathbf{u}^{\mathrm{T}}$ con $\mathbf{u} \in \mathbb{R}$, $\mathbf{u}^{\mathrm{T}}\mathbf{u} = 1$. ■

Para la aplicación del Teorema 3.6 hay que tomar en cuenta que la matriz $\mathbf{U}$ puede ser escrita como

$$\mathbf{U} = \mathbf{I} - \beta\hat{\mathbf{w}}\hat{\mathbf{w}}^*, \quad \hat{\mathbf{w}} := \begin{pmatrix} \exp(\mathrm{i}\alpha)(|\xi_1| + \|\mathbf{x}\|_2) \\ \xi_2 \\ \vdots \\ \xi_n \end{pmatrix},$$

donde

$$\exp(\mathrm{i}\alpha) = \begin{cases} 1 & \text{si } \xi_1 = 0, \\ \xi_1/|\xi_1| & \text{sino,} \end{cases} \quad \beta = \frac{1}{\|\mathbf{x}\|_2(\xi_1 + \|\mathbf{x}\|_2)}.$$

Para aplicar $\mathbf{U}$ a algún vector $\mathbf{y}$, tomamos en cuenta que

$$\mathbf{U}\mathbf{y} = \mathbf{y} - (\beta\hat{\mathbf{w}}^*\mathbf{y})\hat{\mathbf{w}},$$

es decir, *nunca* hay que almacenar la $n \times n$-matriz $\mathbf{U}$, sino que sólo la información esencial, β y $\hat{\mathbf{w}}$.

3.2. El problema de la sensitividad para un sistema lineal

Consideremos el siguiente problema: está dado el sistema $\mathbf{A}\mathbf{x} = \mathbf{b}$, suponiendo que el problema tiene una única solución. Ahora, cuál es la relación entre $\mathbf{x}$ y la solución $\tilde{\mathbf{x}}$ del sistema $\tilde{\mathbf{A}}\tilde{\mathbf{x}} = \tilde{\mathbf{b}}$, cuando $\|\mathbf{A} - \tilde{\mathbf{A}}\|$ y $\|\mathbf{b} - \tilde{\mathbf{b}}\|$ son suficientemente pequeñas? Del mismo tipo es el problema de estimar la norma $\|\tilde{\mathbf{A}}^{-1} - \mathbf{A}^{-1}\|$. Empezamos con un caso simple, la pertubación de la matriz unitaria.

Teorema 3.7. *Sea $\|\cdot\|$ una norma vectorial sobre $\mathbb{C}^n$. Como norma matricial sobre $\mathbb{C}^{n\times n}$ se usa la norma matricial asociada. Si $\mathbf{H} \in \mathbb{C}^{n\times n}$ cumple $\|\mathbf{H}\| < 1$, entonces $\mathbf{I} + \mathbf{H}$ es regular y tenemos que*

$$\big\|(\mathbf{I} + \mathbf{H})^{-1}\big\| \leqslant \frac{1}{1 - \|\mathbf{H}\|}, \tag{3.19}$$

$$\big\|(\mathbf{I} + \mathbf{H})^{-1} - \mathbf{I}\big\| \leqslant \frac{\|\mathbf{H}\|}{1 - \|\mathbf{H}\|}. \tag{3.20}$$

Demostración. Dea $\mathbf{x} \neq 0$ un vector arbitrario. Primero hay que demostrar que $(\mathbf{I}+\mathbf{H})\mathbf{x} \neq 0$, lo que es equivalente a $\|(\mathbf{I}+\mathbf{H})\mathbf{x}\| \neq 0$. Pero

$$\begin{aligned}\big\|(\mathbf{I}+\mathbf{H})\mathbf{x}\big\| &= \|\mathbf{x}+\mathbf{H}\mathbf{x}\| \\ &\geqslant \|\mathbf{x}\| - \|\mathbf{H}\mathbf{x}\| \\ &\geqslant \|\mathbf{x}\| - \|\mathbf{H}\|\|\mathbf{x}\| \\ &= \big(1-\|\mathbf{H}\|\big)\|\mathbf{x}\| > 0.\end{aligned}$$

Además, tenemos que

$$\begin{aligned}1 = \|\mathbf{I}\| &= \big\|(\mathbf{I}+\mathbf{H})(\mathbf{I}+\mathbf{H})^{-1}\big\| \\ &\geqslant \big\|(\mathbf{I}+\mathbf{H})^{-1}\big\| - \big\|\mathbf{H}(\mathbf{I}+\mathbf{H})^{-1}\big\| \\ &\geqslant \big\|(\mathbf{I}+\mathbf{H})^{-1}\big\| - \|\mathbf{H}\|\big\|(\mathbf{I}+\mathbf{H})^{-1}\big\| \\ &= \big(1-\|\mathbf{H}\|\big)\big\|(\mathbf{I}+\mathbf{H})^{-1}\big\|,\end{aligned}$$

lo que implica (3.19). Por otro lado,

$$\begin{aligned}\big\|(\mathbf{I}+\mathbf{H})^{-1}-\mathbf{I}\big\| &= \big\|(\mathbf{I}+\mathbf{H})^{-1}-(\mathbf{I}+\mathbf{H})^{-1}(\mathbf{I}+\mathbf{H})\big\| \\ &= \big\|-(\mathbf{I}+\mathbf{H})^{-1}\mathbf{H}\big\| \\ &\leqslant \|\mathbf{H}\|\big\|(\mathbf{I}+\mathbf{H})^{-1}\big\|,\end{aligned}$$

lo que demuestra (3.20). ■

Corolario 3.2. *Si $r_\sigma(\mathbf{H}) < 1$, entonces $\mathbf{I}+\mathbf{H}$ es regular y también en este caso el Teorema 3.7 es válido.*

Demostración. Usar el Teorema 3.5. ■

Corolario 3.3. *Si $r_\sigma(\mathbf{H}) < 1$, entonces $\mathbf{I}+\mathbf{H}$ es regular y*

$$(\mathbf{I}+\mathbf{H})^{-1} = \sum_{k=0}^{\infty}(-1)^k\mathbf{H}^k \quad \text{(Series de Neumann).} \tag{3.21}$$

Demostración. Usamos el Corolario 3.2 y definimos

$$\mathbf{S}_n := \sum_{k=0}^{n}(-1)^k\mathbf{H}^k.$$

Entonces, para $m > n$,

$$\|\mathbf{S}_n - \mathbf{S}_m\| \leqslant \sum_{k=n+1}^{m}\|\mathbf{H}\|^k \leqslant \|\mathbf{H}\|^{n+1}\frac{1}{1-\|\mathbf{H}\|} < \varepsilon$$

para $n > N(\varepsilon)$, es decir existe el límite

$$\mathbf{S} := \lim_{n\to\infty}\mathbf{S}_n.$$

En virtud de

$$\mathbf{S}_n(\mathbf{I}+\mathbf{H}) = \mathbf{S}_n + \mathbf{S}_n\mathbf{H}$$

$$= \sum_{k=0}^{n} (-1)^k \mathbf{H}^k - \sum_{k=0}^{n} (-1)^{k+1} \mathbf{H}^{k+1}$$
$$= \mathbf{I} - (-1)^{n+1} \mathbf{H}^{n+1}$$

resulta

$$\big\|\mathbf{S}(\mathbf{I}+\mathbf{H}) - \mathbf{I}\big\| = \big\|\mathbf{S}_n(\mathbf{I}+\mathbf{H}) - \mathbf{I} + (\mathbf{S}-\mathbf{S}_n)(\mathbf{I}+\mathbf{H})\big\|$$
$$\leqslant \|\mathbf{H}\|^{n+1} + \|\mathbf{S}-\mathbf{S}_n\| \frac{1}{1-\|\mathbf{H}\|} < \varepsilon$$

para $n > N(\varepsilon)$, mientras que la parte izquierda no depende de ε, lo que concluye la demostración de (3.21). ■

Corolario 3.4. *Para una matriz* $\mathbf{A}$ *regular y otra matriz* $\tilde{\mathbf{A}}$ *con*

$$\|\mathbf{A}^{-1}\|\|\tilde{\mathbf{A}} - \mathbf{A}\| < 1,$$

$\tilde{\mathbf{A}}$ *es invertible, y*

$$\frac{\|\tilde{\mathbf{A}}^{-1} - \mathbf{A}^{-1}\|}{\|\mathbf{A}^{-1}\|} \leqslant \|\mathbf{A}^{-1}\|\|\tilde{\mathbf{A}} - \mathbf{A}\| \frac{1}{1 - \|\mathbf{A}^{-1}\|\|\tilde{\mathbf{A}} - \mathbf{A}\|}.$$

Demostración. Usamos

$$\tilde{\mathbf{A}} = \mathbf{A} + \tilde{\mathbf{A}} - \mathbf{A} = \mathbf{A}\big(\mathbf{I} + \mathbf{A}^{-1}(\tilde{\mathbf{A}} - \mathbf{A})\big)$$

y definimos

$$\mathbf{H} := \mathbf{A}^{-1}(\tilde{\mathbf{A}} - \mathbf{A}).$$

Entonces $\|\mathbf{H}\| < 1$, por lo tanto $\mathbf{I} + \mathbf{H}$ es invertible y $\tilde{\mathbf{A}}$ es regular. Luego obtenemos

$$\frac{\|\tilde{\mathbf{A}}^{-1} - \mathbf{A}^{-1}\|}{\|\mathbf{A}^{-1}\|} = \frac{\|(\mathbf{I}+\mathbf{H})^{-1}\mathbf{A}^{-1} - \mathbf{A}^{-1}\|}{\|\mathbf{A}^{-1}\|} \leqslant \|(\mathbf{I}+\mathbf{H})^{-1} - \mathbf{I}\|$$
$$\leqslant \frac{\|\mathbf{H}\|}{1-\|\mathbf{H}\|} \leqslant \frac{\|\mathbf{A}^{-1}\|\|\tilde{\mathbf{A}} - \mathbf{A}\|}{1 - \|\mathbf{A}^{-1}\|\|\tilde{\mathbf{A}} - \mathbf{A}\|}.$$

■

Definición 3.5. *Sea* $\mathbf{A}$ *regular. La cantidad*

$$\operatorname{cond}_{\|\cdot\|}(\mathbf{A}) := \|\mathbf{A}\|\|\mathbf{A}^{-1}\|$$

se llama número de condición de $\mathbf{A}$ *para la solución de un sistema lineal.*

Teorema 3.8. *Sea* $\mathbf{A} \in \mathbb{K}^{n\times n}$ *regular y* $0 \neq \mathbf{b} \in \mathbb{K}^n$, $\tilde{\mathbf{A}} \in \mathbb{K}^{n\times n}$, $\tilde{\mathbf{b}} \in \mathbb{K}^n$. *Sea* $\|\cdot\|$ *la norma matricial inducida por la norma vectorial* $\|\cdot\|$ *y*

$$\|\mathbf{A}^{-1}\|\|\tilde{\mathbf{A}} - \mathbf{A}\| < 1.$$

Además, sea $\mathbf{x} := \mathbf{A}^{-1}\mathbf{b}$. *Entonces la solución única de* $\tilde{\mathbf{A}}\tilde{\mathbf{x}} = \tilde{\mathbf{b}}$ *satisface*

$$\frac{\|\tilde{\mathbf{x}} - \mathbf{x}\|}{\|\mathbf{x}\|} \leqslant \mathrm{cond}_{\|\cdot\|}(\mathbf{A}) \left(\frac{\|\tilde{\mathbf{b}} - \mathbf{b}\|}{\|\mathbf{b}\|} + \frac{\|\tilde{\mathbf{A}} - \mathbf{A}\|}{\|\mathbf{A}\|} \right) \frac{1}{1 - \mathrm{cond}_{\|\cdot\|}(\mathbf{A}) \dfrac{\|\tilde{\mathbf{A}} - \mathbf{A}\|}{\|\mathbf{A}\|}}. \tag{3.22}$$

Demostración. La existencia de $\tilde{\mathbf{A}}^{-1}$ sigue del Corolario 3.4. Luego calculamos que

$$\tilde{\mathbf{x}} = \tilde{\mathbf{A}}^{-1}\tilde{\mathbf{b}} = \mathbf{A}^{-1}\mathbf{b} + \mathbf{A}^{-1}(\tilde{\mathbf{b}} - \mathbf{b}) + \big((\mathbf{I} + \mathbf{A}^{-1}(\tilde{\mathbf{A}} - \mathbf{A}))^{-1} - \mathbf{I}\big)\mathbf{A}^{-1}(\mathbf{b} + \tilde{\mathbf{b}} - \mathbf{b}),$$

es decir, definiendo $\mathbf{H} := \mathbf{A}^{-1}(\tilde{\mathbf{A}} - \mathbf{A})$,

$$\tilde{\mathbf{x}} = \mathbf{x} + \mathbf{A}^{-1}(\tilde{\mathbf{b}} - \mathbf{b}) + \big((\mathbf{I} + \mathbf{H})^{-1} - \mathbf{I}\big)\big(\mathbf{x} + \mathbf{A}^{-1}(\tilde{\mathbf{b}} - \mathbf{b})\big).$$

Entonces, aprovechando (3.20) y $\|\mathbf{H}\| \leqslant \|\mathbf{A}^{-1}\|\|\tilde{\mathbf{A}} - \mathbf{A}\| < 1$, llegamos a

$$\frac{\|\tilde{\mathbf{x}} - \mathbf{x}\|}{\|\mathbf{x}\|} \leqslant \|\mathbf{A}^{-1}\| \frac{\|\tilde{\mathbf{b}} - \mathbf{b}\|}{\|\mathbf{x}\|} + \frac{\|\mathbf{A}^{-1}\|\|\tilde{\mathbf{A}} - \mathbf{A}\|}{1 - \|\mathbf{A}^{-1}\|\|\tilde{\mathbf{A}} - \mathbf{A}\|} \left(1 + \|\mathbf{A}^{-1}\| \frac{\|\tilde{\mathbf{b}} - \mathbf{b}\|}{\|\mathbf{x}\|} \right).$$

Dado que

$$\|\mathbf{b}\| = \|\mathbf{A}\mathbf{x}\| \leqslant \|\mathbf{A}\|\|\mathbf{x}\| \Longrightarrow \frac{1}{\|\mathbf{x}\|} \leqslant \frac{\|\mathbf{A}\|}{\|\mathbf{b}\|},$$

y definiendo

$$\alpha := \|\mathbf{A}\|\|\mathbf{A}^{-1}\| \frac{\dfrac{\|\tilde{\mathbf{A}} - \mathbf{A}\|}{\|\mathbf{A}\|}}{1 - \|\mathbf{A}\|\|\mathbf{A}^{-1}\| \dfrac{\|\tilde{\mathbf{A}} - \mathbf{A}\|}{\|\mathbf{A}\|}},$$

resulta la desigualdad

$$\frac{\|\tilde{\mathbf{x}} - \mathbf{x}\|}{\|\mathbf{x}\|} \leqslant \alpha + \|\mathbf{A}\|\|\mathbf{A}^{-1}\| \frac{\|\tilde{\mathbf{b}} - \mathbf{b}\|}{\|\mathbf{b}\|}(1 + \alpha). \tag{3.23}$$

■

Obviamente, siempre se tiene que

$$\mathrm{cond}_{\|\cdot\|}(\mathbf{A}) \geqslant r_\sigma(\mathbf{A}) r_\sigma(\mathbf{A}^{-1}) \geqslant 1.$$

Cuando $\mathrm{cond}_{\|\cdot\|}(\mathbf{A}) \gg 1$, eso significa que la influencia de errores menores (por ejemplo, de errores en $\mathbf{A}$ o en errores de redondeo) pueden causar cambios fuertes en la solución del sistema lineal. Se dice entonces que el sistema es *mal acondicionado*. (Recordamos que los errores de redondeo pueden ser interpretados como una modificación la matriz $\mathbf{A}$ seguida por la solución exacta del sistema.) Este problema se ilustra en el siguiente ejemplo.

Ejemplo 3.1. *Consideramos los sistemas* $\mathbf{A}\mathbf{x} = \mathbf{b}$ *y* $\tilde{\mathbf{A}}\tilde{\mathbf{x}} = \tilde{\mathbf{b}}$ *dados por*

$$\mathbf{A} = \begin{bmatrix} \frac{1}{2} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{4} \end{bmatrix}, \quad \mathbf{b} = \begin{pmatrix} -\frac{1}{6} \\ -\frac{1}{6} \end{pmatrix}, \quad \tilde{\mathbf{A}} = \begin{bmatrix} 0{,}5 & 0{,}337 \\ 0{,}337 & 0{,}246 \end{bmatrix}, \quad \tilde{\mathbf{b}} = \begin{pmatrix} -0{,}165 \\ -0{,}165 \end{pmatrix}$$

con las respectivas soluciones

$$\mathbf{x} = \begin{pmatrix} 1 \\ -2 \end{pmatrix}, \quad \tilde{\mathbf{x}} = \begin{pmatrix} 1{,}5920898 \\ -2{,}8517654 \end{pmatrix}. \tag{3.24}$$

En este caso,

$$\frac{\|\tilde{\mathbf{A}} - \mathbf{A}\|_\infty}{\|\mathbf{A}\|_\infty} = \frac{0{,}007\bar{6}}{0{,}8\bar{3}} = 0{,}0092, \quad \frac{\|\tilde{\mathbf{b}} - \mathbf{b}\|_\infty}{\|\mathbf{b}\|_\infty} = 0{,}01, \quad \operatorname{cond}_{\|\cdot\|_\infty}(\mathbf{A}) = 50.$$

Insertando esto en (3.22)*, resulta la cota*

$$\frac{\|\tilde{\mathbf{x}} - \mathbf{x}\|_\infty}{\|\mathbf{x}\|_\infty} \leqslant \frac{16}{9} = 1.\bar{7},$$

mientras que usando las verdaderas soluciones (3.24)

$$\frac{\|\tilde{\mathbf{x}} - \mathbf{x}\|_\infty}{\|\mathbf{x}\|_\infty} = 0{,}42588,$$

es decir, en este caso la cota sobreestima el verdadero error relativo en un factor un poco más de 4*.*

Ejemplo 3.2 (Tarea 8, Curso 2006)**.** *Se considera la matriz*

$$\mathbf{A} := \begin{bmatrix} 10 & 5 & 1 \\ 8 & 9 & 1 \\ 0 & 1 & 3 \end{bmatrix}.$$

a) *Usando el Teorema 3.7, demostrar que* $\mathbf{A}$ *es invertible. Aviso: Usar* $\mathbf{A} = \mathbf{D}+\mathbf{B}$*, donde* $\mathbf{D} = \operatorname{diag}(a_{11}, a_{22}, a_{33})$*.*

b) *Determinar una cota superior para* $\operatorname{cond}_{\|\cdot\|}(\mathbf{A})$ *en una norma* $\|\cdot\|$ *apropiada* sin *invertir* $\mathbf{A}$ *o calcular* $\det \mathbf{A}$*.*

c) *Además consideramos*

$$\mathbf{b} = \begin{pmatrix} 10 \\ -10 \\ 10 \end{pmatrix}, \quad \tilde{\mathbf{b}} = \begin{pmatrix} 10{,}1 \\ -9{,}8 \\ 9{,}7 \end{pmatrix}, \quad \tilde{\mathbf{A}} = \begin{bmatrix} 10{,}05 & 5{,}1 & 1{,}05 \\ 8{,}1 & 9{,}1 & 0{,}95 \\ 0{,}05 & 1 & 3{,}1 \end{bmatrix}.$$

Los vectores $\mathbf{x}$ *y* $\tilde{\mathbf{x}}$ *sean la solución de* $\mathbf{Ax} = \mathbf{b}$ *y* $\tilde{\mathbf{A}}\tilde{\mathbf{x}} = \tilde{\mathbf{b}}$*, respectivamente. Determinar una cota superior (la mejor posible) para* $\|\mathbf{x} - \tilde{\mathbf{x}}\|/\|\mathbf{x}\|$ *sin calcular* $\mathbf{x}$ *o* $\tilde{\mathbf{x}}$*.*

Solución sugerida.

a) *Usamos* $\mathbf{A} = \mathbf{D} + \mathbf{B} = \mathbf{D}(\mathbf{I} + \mathbf{D}^{-1}\mathbf{B})$*, donde*

$$\mathbf{D}^{-1}\mathbf{B} = \begin{bmatrix} 0 & \frac{1}{2} & \frac{1}{10} \\ \frac{8}{9} & 0 & \frac{1}{9} \\ 0 & \frac{1}{3} & 0 \end{bmatrix}.$$

Dado que $\|\mathbf{D}^{-1}\mathbf{B}\|_1 = 8/9 < 1$*, la matriz* $\mathbf{A}$ *es invertible.*

b) *Obviamente* $\|\mathbf{A}\|_1 = 18$. *Usando la parte (a), tenemos*

$$\|\mathbf{A}^{-1}\|_1 = \|(\mathbf{I}+\mathbf{D}^{-1}\mathbf{B})^{-1}\mathbf{D}^{-1}\|_1 \leqslant \|(\mathbf{I}+\mathbf{D}^{-1}\mathbf{B})^{-1}\|_1\|\mathbf{D}^{-1}\|_1$$
$$\leqslant \frac{1}{1-\|\mathbf{D}^{-1}\mathbf{B}\|_1}\|\mathbf{D}^{-1}\|_1 = \frac{1}{1-\dfrac{8}{9}}\cdot\frac{1}{3} = 3,$$

entonces $\operatorname{cond}_{\|\cdot\|_1}(\mathbf{A}) \leqslant 54$.

c) *Obtenemos*

$$\frac{\|\tilde{\mathbf{b}}-\mathbf{b}\|_1}{\|\mathbf{b}\|_1} = 0{,}02, \quad \|\tilde{\mathbf{A}}-\mathbf{A}\|_1 = 0{,}2$$
$$\Longrightarrow \frac{\|\tilde{\mathbf{x}}-\mathbf{x}\|_1}{\|\mathbf{x}\|_1} \leqslant 54\left(0{,}02+\frac{0{,}2}{18}\right)\left(1-54\frac{0{,}2}{18}\right)^{-1} = 4{,}2.$$

Ejemplo 3.3 (Tarea 10, Curso 2006)**.** *Se desea resolver el sistema* $\mathbf{Ax}=\mathbf{b}$ *con*

$$\mathbf{A} = \begin{bmatrix} 1000 & 10 & 1 \\ -1000 & 0 & 0 \\ 1000 & 0 & 1 \end{bmatrix}, \quad \mathbf{b} = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix}, \quad 1 \leqslant b_1, b_2, b_3 \leqslant 10.$$

Los coeficientes de $\mathbf{A}$ *y* $\mathbf{b}$ *han sido pertubados por ciertos errores* $\delta\mathbf{A}$ *y* $\delta\mathbf{b}$.

a) *Determinar cotas para* α *y* β *con*

$$\alpha := \frac{\|\delta\mathbf{A}\|_\infty}{\|\mathbf{A}\|_\infty}, \quad \beta := \frac{\|\delta\mathbf{b}\|_\infty}{\|\mathbf{b}\|_\infty}$$

tales que puede ser garantizado que

$$\|\mathbf{x}-\tilde{\mathbf{x}}\|/\|\mathbf{x}\| < 0{,}01,$$

donde $\mathbf{Ax}=\mathbf{b}$ *y* $(\mathbf{A}+\delta\mathbf{A})\tilde{\mathbf{x}} = \mathbf{b}+\delta\mathbf{b}$.

b) *Supongamos que de la solución* $\mathbf{x}$ *nos interesa solamente la tercera componente. Indicar una transformación simple del sistema original que permite una cota significativamente mejor (que la de (a)) de* $|\tilde{x}_3 - x_3|/|x_3|$ *en dependencia de las perturbaciones de los coeficientes del sistema transformado.*

Solución sugerida.

a) *En lo siguiente, sea* $\|\cdot\| = \|\cdot\|_\infty$. *Tenemos con la notación indicada*

$$\frac{\|\tilde{\mathbf{x}}-\mathbf{x}\|}{\|\mathbf{x}\|} \leqslant \frac{(\alpha+\beta)\ \operatorname{cond}_{\|\cdot\|}(\mathbf{A})}{1-\alpha\ \operatorname{cond}_{\|\cdot\|}(\mathbf{A})}.$$

Dado que

$$\mathbf{A}^{-1} = \begin{bmatrix} 0 & -0{,}001 & 0 \\ 0{,}1 & 0 & -0{,}1 \\ 0 & 1 & 1 \end{bmatrix},$$

sabemos que $\|\mathbf{A}\| = 1011$, $\|\mathbf{A}^{-1}\| = 2$ *y por lo tanto* $\mathrm{cond}_{\|\cdot\|}(\mathbf{A}) = 2022$. *Sean* $\alpha, \beta \leqslant s$, *entonces*

$$\frac{\|\tilde{\mathbf{x}} - \mathbf{x}\|}{\|\mathbf{x}\|} \leqslant 4044 \frac{s}{1 - 2022s} < 0{,}01.$$

Dado que

$$\frac{2as}{1 - as} \leqslant b \Longrightarrow s \leqslant \frac{b}{a(2+b)},$$

resulta $s = 2{,}47 \times 10^{-6}$.

b) *Dividimos la primera columna por* 1000 *y la segunda por* 10, *lo que es equivalente a resolver un sistema para* $(\hat{x}_1, \hat{x}_2, x_3)$ *con* $1000\hat{x}_1 = x_1$ *y* $10\hat{x}_2 = x_2$. *En este caso,*

$$\hat{\mathbf{A}} = \begin{bmatrix} 1 & 1 & 1 \\ -1 & 0 & 0 \\ 1 & 0 & 1 \end{bmatrix} \Rightarrow \hat{\mathbf{A}}^{-1} = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & -1 \\ 0 & 1 & 1 \end{bmatrix} \Rightarrow \mathrm{cond}_{\|\cdot\|}(\hat{\mathbf{A}}) = 6,\ s = 8{,}3 \times 10^{-4}.$$

Ahora uno podría pensar que debido a la técnica de estimación usada en la demostración del Teorema 3.8, siempre se sobreestima bruscamente el error, es decir la cantidad $\|\tilde{\mathbf{x}} - \mathbf{x}\|/\|\mathbf{x}\|$. Demostraremos ahora que esto no es así. Para tal efecto, vamos a construir matrices para las cuales esta cantidad alcanza la cota establecida por el lado de derecho de (3.22) hasta un error arbitrariamente pequeño. Para construir tales matrices necesitamos el concepto de la *descomposición en valores singulares*.

Teorema 3.9 (Descomposición en valores singulares)**.** *Sea* $\mathbf{A} \in \mathbb{C}^{m\times n}$ *con* $m \geqslant n$. *Entonces existen matrices unitarias* $\mathbf{U} \in \mathbb{C}^{m\times m}$ *y* $\mathbf{V} \in \mathbb{C}^{n\times n}$ *y una matriz diagonal* $\mathbf{\Sigma} = \mathrm{diag}(\sigma_1, \ldots, \sigma_n)$ *con elementos diagonales* $\sigma_i \geqslant 0$, $i = 1, \ldots, n$, *tales que*

$$\mathbf{A} = \mathbf{U} \begin{bmatrix} \mathbf{\Sigma} \\ 0 \end{bmatrix} \mathbf{V}^* \tag{3.25}$$

Demostración. Las matrices $\mathbf{A}\mathbf{A}^*$ y $\mathbf{A}^*\mathbf{A}$ son ambas hermitianas y definidas semi-positivas, dado que

$$\mathbf{x}^*\mathbf{A}^*\mathbf{A}\mathbf{x} = \|\mathbf{A}\mathbf{x}\|_2^2 \geqslant 0.$$

Por lo tanto, existe una matriz unitaria $\mathbf{V}$ tal que

$$\mathbf{V}^*\mathbf{A}^*\mathbf{A}\mathbf{V} = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_n^2), \tag{3.26}$$

donde $\sigma_1^2, \ldots, \sigma_n^2$ son los valores propios (no negativos) de $\mathbf{A}^*\mathbf{A}$. Ahora, si

$$\mathbf{A}\mathbf{A}^*\mathbf{y} = \lambda\mathbf{y}, \quad \mathbf{y} \neq 0,$$

tenemos $\lambda = 0$ o $\mathbf{A}^*\mathbf{y} \neq 0$. Si $\mathbf{A}^*\mathbf{y} \neq 0$, entonces

$$\mathbf{A}^*\mathbf{A}\mathbf{A}^*\mathbf{y} = \lambda\mathbf{A}^*\mathbf{y},$$

o sea $\mathbf{A}^*\mathbf{y}$ es un vector propio de $\mathbf{A}^*\mathbf{A}$ y λ es el vector propio correspondiente, es decir, $\lambda \in \{\sigma_1^2, \ldots, \sigma_n^2\}$. Entonces $\mathbf{A}\mathbf{A}^*$ posee el valor propio 0 con multiplicidad $m - n$ ($\mathbf{A}^*\mathbf{y} = 0$ posee por lo menos $m - n$ soluciones linealmente independientes) y los valores propios

$\sigma_1^2, \ldots, \sigma_n^2$ (pueden ocurrir valores $\sigma_i^2 = 0$). En virtud de lo anterior, existe una matriz unitaria $\mathbf{U} \in \mathbb{C}^{m\times m}$ tal que

$$\mathbf{U}^*\mathbf{A}\mathbf{A}^*\mathbf{U} = \begin{bmatrix} \mathbf{\Sigma}^2 & 0 \\ 0 & 0 \end{bmatrix}.$$

La matriz $\mathbf{B} := \mathbf{A}^*\mathbf{U}$ satisface

$$\mathbf{B}^*\mathbf{B} = \begin{bmatrix} \mathbf{b}_1^* \\ \vdots \\ \mathbf{b}_m^* \end{bmatrix} \begin{bmatrix} \mathbf{b}_1 & \cdots & \mathbf{b}_m \end{bmatrix} = \operatorname{diag}(\sigma_1^2, \ldots, \sigma_n^2, 0, \ldots, 0),$$

o sea

$$\mathbf{B} = \begin{bmatrix} \mathbf{b}_1 & \cdots & \mathbf{b}_n & 0 & \cdots & 0 \end{bmatrix}, \quad \text{donde } \mathbf{b}_i^*\mathbf{b}_j = \sigma_i^2\delta_{ij},\ i,j = 1,\ldots,n.$$

Eso significa que

$$\mathbf{B} = \begin{bmatrix} \tilde{\mathbf{V}}\mathbf{\Sigma} & 0 \end{bmatrix} = \tilde{\mathbf{V}} \begin{bmatrix} \mathbf{\Sigma} & 0 \end{bmatrix}$$

con una matriz $\tilde{\mathbf{V}}$ unitaria. Dado que

$$\mathbf{B}\mathbf{B}^* = \mathbf{A}^*\mathbf{A} = \mathbf{V}\mathbf{\Sigma}^2\mathbf{V}^*,$$

esta matriz $\tilde{\mathbf{V}}$ es la misma matriz $\mathbf{V}$ que aparece en (3.26). Finalmente, resulta

$$\mathbf{A} = \mathbf{U}\mathbf{B}^* = \mathbf{U} \begin{bmatrix} \mathbf{\Sigma} \\ 0 \end{bmatrix} \mathbf{V}^*.$$

■

Ejemplo 3.4 (Tarea 11, Curso 2006)**.** *Sea*

$$\mathbf{A} = \begin{bmatrix} 5 & 2{,}5 \\ 5 & 2{,}5 \\ 1{,}4 & 7{,}7 \\ 0{,}2 & 1{,}1 \end{bmatrix}.$$

Determinar matrices unitarias $\mathbf{U} \in \mathbb{C}^{4\times 4}$ *y* $\mathbf{V} \in \mathbb{C}^{2\times 2}$ *y una matriz diagonal* $\mathbf{\Sigma} = \operatorname{diag}(\sigma_1, \sigma_2)$ *con* $\sigma_1, \sigma_2 \geqslant 0$ *tales que*

$$\mathbf{A} = \mathbf{U} \begin{bmatrix} \mathbf{\Sigma} \\ 0 \end{bmatrix} \mathbf{V}^*.$$

Aviso: calcular primero $\mathbf{V}$ *de* $\mathbf{V}^*\mathbf{A}^*\mathbf{A}\mathbf{V} = \mathbf{\Sigma}^2$*, y luego para la computación de* $\mathbf{U}$*, usar*

$$\mathbf{A}\mathbf{V} = \mathbf{U} \begin{bmatrix} \mathbf{\Sigma} \\ 0 \end{bmatrix}$$

y un método de ortonormalización para las otras columnas de $\mathbf{U}$*.*
Solución sugerida. *El polinomio característico de la matriz*

$$\mathbf{A}^*\mathbf{A} = \begin{bmatrix} 52 & 36 \\ 36 & 73 \end{bmatrix}$$

tiene los ceros

$$(52-\lambda)(73-\lambda) = 1296 \Longrightarrow \lambda_1 = 25 = \sigma_1^2,\ \lambda_2 = 100 = \sigma_2^2.$$

Los vectores propios correspondientes son

$$\mathbf{v}_1 = \begin{pmatrix} 0{,}8 \\ -0{,}6 \end{pmatrix},\ \mathbf{v}_2 = \begin{pmatrix} 0{,}6 \\ 0{,}8 \end{pmatrix} \Longrightarrow \mathbf{V} = \begin{bmatrix} 0{,}8 & 0{,}6 \\ -0{,}6 & 0{,}8 \end{bmatrix}.$$

Para determinar las primeras dos columnas de $\mathbf{U}$, $\mathbf{u}_1$ *y* $\mathbf{u}_2$, *usamos que*

$$\mathbf{AV} = \begin{bmatrix} 2{,}5 & 5 \\ 2{,}5 & 5 \\ -3{,}5 & 7 \\ -0{,}5 & 1 \end{bmatrix} = \begin{bmatrix} \sigma_1\mathbf{u}_1 & \sigma_2\mathbf{u}_2 \end{bmatrix},$$

donde $\sigma_1 = 5$ *y* $\sigma_2 = 10$. *Entonces*

$$\mathbf{u}_1 = \begin{pmatrix} 0{,}5 \\ 0{,}5 \\ -0{,}7 \\ -0{,}1 \end{pmatrix}, \quad \mathbf{u}_2 = \begin{pmatrix} 0{,}5 \\ 0{,}5 \\ 0{,}7 \\ 0{,}1 \end{pmatrix}.$$

Como dos vectores ortonormales adicionales podemos usar

$$\mathbf{u}_3 = \begin{pmatrix} -0{,}7 \\ 0{,}7 \\ 0{,}02 \\ -0{,}14 \end{pmatrix}, \quad \mathbf{u}_4 = \begin{pmatrix} -0{,}1 \\ 0{,}1 \\ -0{,}14 \\ 0{,}98 \end{pmatrix} \Longrightarrow \mathbf{U} = \begin{bmatrix} 0{,}5 & 0{,}5 & -0{,}7 & -0{,}1 \\ 0{,}5 & 0{,}5 & 0{,}7 & 0{,}1 \\ -0{,}7 & 0{,}7 & 0{,}02 & -0{,}14 \\ -0{,}1 & 0{,}1 & -0{,}14 & 0{,}98 \end{bmatrix}.$$

La descomposición en valores singulares sirve para la computación de la pseudo-inversa de Moore-Penrose de una matriz.

Definición 3.6. *Sea* $\mathbf{A} \in \mathbb{C}^{m,n}$. *La* pseudo-inversa de Moore-Penrose *es una matriz* $\mathbf{A}^+$ *con*

$$\mathbf{A}^+\mathbf{A} = (\mathbf{A}^+\mathbf{A})^*, \tag{3.27}$$

$$\mathbf{A}\mathbf{A}^+ = (\mathbf{A}\mathbf{A}^+)^*, \tag{3.28}$$

$$\mathbf{A}^+\mathbf{A}\mathbf{A}^+ = \mathbf{A}^+, \tag{3.29}$$

$$\mathbf{A}\mathbf{A}^+\mathbf{A} = \mathbf{A}. \tag{3.30}$$

Se puede demostrar que la matriz $\mathbf{A}^+$ *siempre existe y es única.*

Lema 3.1. *Sea* $\mathbf{A} \in \mathbb{C}^{m\times n}$, $m \geqslant n$, *y*

$$\mathbf{A} = \mathbf{U}\begin{bmatrix} \mathbf{\Sigma} \\ 0 \end{bmatrix}\mathbf{V}^*$$

la descomposición en valores singulares de $\mathbf{A}$. *Entonces*

$$\mathbf{A}^+ = \mathbf{V}\begin{bmatrix} \mathbf{\Sigma}^+ & 0 \end{bmatrix}\mathbf{U}^*, \quad \mathbf{\Sigma}^+ := \operatorname{diag}(\sigma_1^+, \ldots, \sigma_n^+), \quad \sigma_i^+ := \begin{cases} 1/\sigma_i & \text{si } \sigma_i > 0, \\ 0 & \text{sino.} \end{cases}$$

Demostración. Tarea. ■

Lema 3.2. *Sea* $\mathbf{A} \in \mathbb{C}^{n\times n}$ *una matriz regular con la descomposición en valores singulares* $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^*$ *con* $\sigma_1 \geqslant \sigma_2 \geqslant \ldots \geqslant \sigma_n > 0$, $\mathbf{b} := \mathbf{u}_1$ *(la primera columna de* $\mathbf{U}$*),* $\tilde{\mathbf{b}} := \mathbf{b} + \varepsilon\mathbf{u}_n$, $\varepsilon > 0$, $\tilde{\mathbf{A}} := \mathbf{A} - \varepsilon\mathbf{u}_n\mathbf{v}_1^*$, $\mathbf{A}\mathbf{x} = \mathbf{b}$ *y* $\tilde{\mathbf{A}}\tilde{\mathbf{x}} = \tilde{\mathbf{b}}$. *En este caso,*

$$\frac{\|\tilde{\mathbf{x}} - \mathbf{x}\|_2}{\|\mathbf{x}\|_2} = \frac{\sigma_1}{\sigma_n}\varepsilon\left(1 + \frac{1}{\sigma_1}\right),$$

mientras que la evaluación del lado derecho de la desigualdad (3.22) *entrega la cota*

$$\frac{\|\tilde{\mathbf{x}} - \mathbf{x}\|_2}{\|\mathbf{x}\|_2} \leqslant \frac{\sigma_1}{\sigma_n}\varepsilon\left(1 + \frac{1}{\sigma_1}\right)\frac{1}{1 - \dfrac{\varepsilon}{\sigma_n}},$$

es decir, para ε*, la cota establecida por el Teorema 3.8 se alcanza hasta* $\mathcal{O}(\varepsilon^2)$.

Demostración. Primero demostramos que para cada matriz $\mathbf{A} \in \mathbb{C}^{n\times n}$, cada vector $\mathbf{b} \in \mathbb{C}^n$ y cada matriz unitaria $\mathbf{B} \in \mathbb{C}^{n\times n}$ se cumple

$$\|\mathbf{A}\|_2 = \|\mathbf{B}\mathbf{A}\|_2, \quad \|\mathbf{b}\|_2 = \|\mathbf{B}\mathbf{b}\|_2. \tag{3.31}$$

Para demostrar (3.31), notamos que

$$\begin{aligned}\|\mathbf{b}\|_2^2 &= \mathbf{b}^*\mathbf{b} = \mathbf{b}^*\mathbf{B}^*\mathbf{B}\mathbf{b} = \|\mathbf{B}\mathbf{b}\|_2^2,\\ \|\mathbf{A}\|_2^2 &= r_\sigma(\mathbf{A}^*\mathbf{A}) = r_\sigma(\mathbf{A}^*\mathbf{B}^*\mathbf{B}\mathbf{A}) = r_\sigma\big((\mathbf{B}\mathbf{A})^*\mathbf{B}\mathbf{A}\big) = \|\mathbf{B}\mathbf{A}\|_2^2.\end{aligned}$$

Luego, usando $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^*$ y tomando en cuenta que $r_\sigma(\mathbf{V}\mathbf{V}^*) = 1$, tenemos

$$\|\mathbf{A}\|_2^2 = \|\mathbf{U}^*\mathbf{A}\|_2^2 = \|\mathbf{\Sigma}\mathbf{V}^*\|_2^2 \leqslant \|\mathbf{\Sigma}\|_2^2 r_\sigma(\mathbf{V}\mathbf{V}^*) = \max_{1\leqslant i\leqslant n}\sigma_i^2.$$

Dado que

$$\|\mathbf{A}\|_2 = \max_{\|\mathbf{x}\|_2=1}\|\mathbf{A}\mathbf{x}\|_2,$$

obtenemos (usando $\mathbf{x} = \mathbf{v}_1$):

$$\|\mathbf{A}\mathbf{x}\|_2^2 = \|\mathbf{U}\mathbf{\Sigma}\mathbf{V}^*\mathbf{v}_1\|_2^2 = \|\mathbf{\Sigma}\mathbf{e}_1\|_2^2 = \sigma_1^2 \Longrightarrow \|\mathbf{A}\|_2 = \sigma_1. \tag{3.32}$$

Luego derivamos una expresión para $\mathrm{cond}_{\|\cdot\|}(\mathbf{A})$. (En lo que sigue, usamos $\|\cdot\| = \|\cdot\|_2$.) Para tal efecto, notamos que

$$\mathbf{A}^{-1} = (\mathbf{U}\mathbf{\Sigma}\mathbf{V}^*)^{-1} = (\mathbf{V}^*)^{-1}\mathbf{\Sigma}^{-1}\mathbf{U}^{-1} = \mathbf{V}\mathbf{\Sigma}^{-1}\mathbf{U}^*.$$

Eso significa que $\mathbf{A}^{-1}$ posee la siguiente descomposición en valores singulares con las matrices unitarias $\tilde{\mathbf{U}} := \mathbf{V}$, $\tilde{\mathbf{V}} := \mathbf{U}$:

$$\mathbf{A}^{-1} = \tilde{\mathbf{U}}\mathbf{\Sigma}^{-1}\tilde{\mathbf{V}}^*,$$

y análogamente a la derivación de (3.32) tenemos

$$\|\mathbf{A}^{-1}\| = \|\mathbf{\Sigma}^{-1}\| = \max_{1\leqslant i\leqslant n}\frac{1}{\sigma_i} = \frac{1}{\sigma_n}.$$

Combinando este resultado con (3.32), tenemos

$$\mathrm{cond}_{\|\cdot\|}(\mathbf{A}) = \frac{\sigma_1}{\sigma_n}. \tag{3.33}$$

Ahora consideramos el sistema lineal $\mathbf{Ax} = \mathbf{b}$ con $\mathbf{b} = \mathbf{u}_1$. Entonces

$$\mathbf{\Sigma V}^*\mathbf{x} = \mathbf{U}^*\mathbf{u}_1 = \mathbf{e}_1 \quad \Longrightarrow \quad \mathbf{V}^*\mathbf{x} = \mathbf{\Sigma}^{-1}\mathbf{e}_1 = \frac{1}{\sigma_1}\mathbf{e}_1 \quad \Longrightarrow \quad \mathbf{x} = \frac{1}{\sigma_1}\mathbf{V}\mathbf{e}_1 = \frac{1}{\sigma_1}\mathbf{v}_1.$$

Por otro lado,

$$\tilde{\mathbf{A}}\mathbf{v}_i = \mathbf{A}\mathbf{v}_i - \varepsilon\mathbf{u}_n\mathbf{v}_1^*\mathbf{v}_i = \mathbf{U\Sigma V}^*\mathbf{v}_i - \varepsilon\mathbf{u}_n\mathbf{v}_1^*\mathbf{v}_i = \mathbf{U}(\sigma_i\mathbf{e}_i) - \varepsilon\mathbf{u}_n\delta_{1i}$$
$$= \begin{cases} \sigma_1\mathbf{u}_1 - \varepsilon\mathbf{u}_n & \text{si } i = 1, \\ \sigma_i\mathbf{u}_i & \text{sino.} \end{cases}$$

Para determinar $\tilde{\mathbf{x}}$, podemos usar la fórmula de Sherman-Morrison. Alternativamente, dado que $\tilde{\mathbf{b}} = \mathbf{u}_1 + \varepsilon\mathbf{u}_n$, podemos tratar un planteo de la forma $\tilde{\mathbf{x}} = \alpha\mathbf{v}_1 + \beta\mathbf{v}_n$. En este caso, el sistema $\tilde{\mathbf{A}}\tilde{\mathbf{x}} = \tilde{\mathbf{b}}$ se transforma a

$$\alpha(\sigma_1\mathbf{u}_1 - \varepsilon\mathbf{u}_n) + \beta\sigma_n\mathbf{u}_n = \mathbf{u}_1 + \varepsilon\mathbf{u}_n,$$

entonces

$$\alpha\sigma_1 = 1 \Longrightarrow \alpha = \frac{1}{\sigma_1}, \quad \beta\sigma_n - \alpha\varepsilon = \varepsilon \Longrightarrow \beta = \frac{\varepsilon}{\sigma_n}\left(\frac{1}{\sigma_1} + 1\right).$$

Ambos métodos entregan

$$\tilde{\mathbf{x}} = \frac{1}{\sigma_1}\mathbf{v}_1 + \frac{\varepsilon}{\sigma_n}\left(\frac{1}{\sigma_1} + 1\right)\mathbf{v}_n,$$

lo que implica

$$\frac{\|\mathbf{x} - \tilde{\mathbf{x}}\|}{\|\mathbf{x}\|} = \frac{\dfrac{\varepsilon}{\sigma_n}\left(\dfrac{1}{\sigma_1} + 1\right)\|\mathbf{v}_n\|}{\dfrac{1}{\sigma_1}\|\mathbf{v}_1\|} = \varepsilon\frac{\sigma_1}{\sigma_n}\left(\frac{1}{\sigma_1} + 1\right).$$

Ahora, evaluando la parte derecha de la desigualdad (3.22) del Teorema 3.8, tenemos

$$\frac{\|\mathbf{x} - \tilde{\mathbf{x}}\|}{\|\mathbf{x}\|} \leqslant \frac{\sigma_1}{\sigma_n}\left(\frac{\|\varepsilon\mathbf{u}_n\|}{\|\mathbf{u}_1\|} + \frac{\|-\varepsilon\mathbf{u}_n\mathbf{v}_1^*\|}{\sigma_1}\right)\frac{1}{1 - \dfrac{\sigma_1}{\sigma_n}\dfrac{\|-\varepsilon\mathbf{u}_n\mathbf{v}_1^*\|}{\sigma_1}}$$
$$= \frac{\sigma_1}{\sigma_n}\varepsilon\left(1 + \frac{\|\mathbf{u}_n\mathbf{v}_1^*\|}{\sigma_1}\right)\frac{1}{1 - \dfrac{\varepsilon}{\sigma_n}\|\mathbf{u}_n\mathbf{v}_1^*\|}.$$

Queda para demostrar que

$$\|\mathbf{u}_n\mathbf{v}_1^*\| = 1. \tag{3.34}$$

Sea $\mathbf{x} \in \mathbb{C}^n$. Entonces podemos escribir

$$\mathbf{x} = \sum_{i=1}^{n} \xi_i\mathbf{v}_i,$$

y entonces para $\mathbf{x} \neq 0$

$$\|\mathbf{u}_n \mathbf{v}_1^* \mathbf{x}\| = \left\| \mathbf{u}_n \left(\mathbf{v}_1^* \sum_{i=1}^{n} \xi_i \mathbf{v}_i \right) \right\| = \|\xi_1 \mathbf{u}_n\| = |\xi_1|$$

$$\Longrightarrow \quad \frac{\|\mathbf{u}_n \mathbf{v}_1^* \mathbf{x}\|}{\|\mathbf{x}\|} = \frac{|\xi_1|}{(|\xi_1|^2 + \cdots + |\xi_n|^2)^{1/2}} \leqslant 1.$$

Por otro lado, para $\mathbf{x} = \mathbf{v}_1$ tenemos

$$\frac{\|\mathbf{u}_n \mathbf{v}_1^* \mathbf{x}\|}{\|\mathbf{x}\|} = 1,$$

asi que (3.34) es válido. Finalmente, definimos

$$f(\varepsilon) := \frac{1}{1 - \dfrac{\varepsilon}{\sigma_n}} \Longrightarrow f'(\varepsilon) = \frac{1}{\sigma_n \left(1 - \dfrac{\varepsilon}{\sigma_n}\right)^2},$$

es decir, para un $\tilde{\varepsilon} \in (0, \varepsilon)$ sigue (desarrollando $f(\varepsilon)$ por $\varepsilon = 0$)

$$f(\varepsilon) = 1 + \frac{\varepsilon}{\sigma_n \left(1 - \dfrac{\tilde{\varepsilon}}{\sigma_n}\right)^2},$$

entonces

$$\begin{aligned} \frac{\|\mathbf{x} - \tilde{\mathbf{x}}\|}{\|\mathbf{x}\|} &\leqslant \frac{\sigma_1}{\sigma_n} \varepsilon \left(1 + \frac{1}{\sigma_1}\right) \left(1 + \frac{\varepsilon}{\sigma_n \left(1 - \tilde{\varepsilon}/\sigma_n\right)^2}\right) \\ &= \frac{\sigma_1}{\sigma_n} \varepsilon \left(1 + \frac{1}{\sigma_1}\right) + \underbrace{\varepsilon^2 \frac{\sigma_1}{\sigma_n^2} \left(1 + \frac{1}{\sigma_1}\right) \frac{1}{\left(1 - \tilde{\varepsilon}/\sigma_n\right)^2}}_{= \mathcal{O}(\varepsilon^2)}. \end{aligned}$$

■

Entonces, si $\mathbf{A} \in \mathbb{C}^{n \times n}$ y $\mathbf{A} = \mathbf{U} \boldsymbol{\Sigma} \mathbf{V}^*$ es la descomposición en valores singulares, sabemos que $\mathbf{A}$ es regular si y sólo si $\boldsymbol{\Sigma}$ es regular, y

$$\begin{aligned} \|\mathbf{A}\|_2 &= \text{máx}\{\sigma_1, \ldots, \sigma_n\}, \\ \|\mathbf{A}^{-1}\|_2 &= \text{máx}\{\sigma_1^{-1}, \ldots, \sigma_n^{-1}\}, \\ \text{cond}_{\|\cdot\|_2}(\mathbf{A}) &= \sigma_1/\sigma_n \quad \text{si } \sigma_1 \geqslant \ldots \geqslant \sigma_n > 0. \end{aligned}$$

En muchos casos, los coeficientes $\mathbf{A}$ y $\mathbf{b}$ de un sistema lineal se conocen solamente aproximadamente, a lo más se puede estimar el orden de magnitud de los errores. Sean $\mathbf{A}_0$ y $\mathbf{b}_0$ dados con

$$|\mathbf{A}_0 - \mathbf{A}| \leqslant \mathbf{E}, \quad \mathbf{E} \in \mathbb{R}_+^{n \times n}, \quad |\mathbf{b}_0 - \mathbf{b}| \leqslant \mathbf{d}, \quad \mathbf{d} \in \mathbb{R}_+^n.$$

En esta situación no hay mucho sentido en resolver el sistema lineal $\mathbf{A}_0 \mathbf{x}_0 = \mathbf{b}_0$ con gran exactitud, sino que se busca una solución "razonable" $\tilde{\mathbf{x}}$, cuya exactitud "corresponde" a la de $\mathbf{A}_0$ y $\mathbf{b}_0$. Sorpresivamente, podemos decidir sin conocer el número de condición de $\mathbf{A}_0$ si $\tilde{\mathbf{x}}$ es una aproximación razonable de $\mathbf{x}_0$.

Teorema 3.10 (Criterio de Prager & Oettli)**.** *Sean*

$$\mathcal{A} := \big\{\mathbf{A} \mid |\mathbf{A} - \mathbf{A}_0| \leqslant \mathbf{E}\big\}, \quad \mathcal{B} := \big\{\mathbf{b} \mid |\mathbf{b} - \mathbf{b}_0| \leqslant \mathbf{d}\big\},$$

donde $\mathbf{E} \in \mathbb{R}^{n\times n}_+$, $\mathbf{d} \in \mathbb{R}^n_+$ *son dados, y* $\tilde{\mathbf{x}} \in \mathbb{R}^n$ *es dado. Entonces existen una matriz* $\mathbf{A} \in \mathcal{A}$ *y un vector* $\mathbf{b} \in \mathcal{B}$ *tales que* $\mathbf{A}\tilde{\mathbf{x}} = \mathbf{b}$ *si y sólo si*

$$|\mathbf{b}_0 - \mathbf{A}_0\tilde{\mathbf{x}}| \leqslant \mathbf{E}|\tilde{\mathbf{x}}| + \mathbf{d}. \tag{3.35}$$

En este caso, el vector $\tilde{\mathbf{x}}$ *se llama* solución aproximada compatible con $\mathbf{A}_0\mathbf{x}_0 = \mathbf{b}_0$.

Demostración. Hay que demostrar dos implicaciones.

"⇒": Supongamos que $\mathbf{A}\tilde{\mathbf{x}} = \mathbf{b}$. Entonces, dado que $\mathbf{b} - \mathbf{A}\tilde{\mathbf{x}} = 0$, podemos escribir:

$$\begin{aligned}
|\mathbf{b}_0 - \mathbf{A}_0\tilde{\mathbf{x}}| &= \big|\mathbf{b}_0 - \mathbf{b} + \mathbf{b} - \mathbf{A}\tilde{\mathbf{x}} + (\mathbf{A} - \mathbf{A}_0)\tilde{\mathbf{x}}\big| \\
&\leqslant |\mathbf{b}_0 - \mathbf{b}| + \big|(\mathbf{A} - \mathbf{A}_0)\tilde{\mathbf{x}}\big| \\
&\leqslant \mathbf{d} + |\mathbf{A}_0 - \mathbf{A}||\tilde{\mathbf{x}}| \\
&\leqslant \mathbf{d} + \mathbf{E}|\tilde{\mathbf{x}}|.
\end{aligned}$$

"⇐": Supongamos que

$$|\mathbf{b}_0 - \mathbf{A}_0\tilde{\mathbf{x}}| \leqslant \mathbf{E}|\tilde{\mathbf{x}}| + \mathbf{d}, \quad \text{con } \mathbf{E} = (\varepsilon_{ij}),\ \mathbf{d} = (\delta_1, \ldots, \delta_n)^{\mathrm{T}}.$$

Entonces definimos

$$\varrho_i := \mathbf{e}_i^{\mathrm{T}}(\mathbf{b}_0 - \mathbf{A}_0\tilde{\mathbf{x}}), \quad \sigma_i := \mathbf{e}_i^{\mathrm{T}}\big(\mathbf{E}|\tilde{\mathbf{x}}| + \mathbf{d}\big), \quad i = 1, \ldots, n.$$

Sabemos que $|\varrho_i| \leqslant \sigma_i$, es decir

$$\left|\frac{\varrho_i}{\sigma_i}\right| \leqslant 1 \quad \text{para } \sigma_i \neq 0.$$

Además, supongamos que $\mathbf{A}_0 = (\alpha_{ij}^{(0)})$, $\mathbf{b}_0 = (\beta_1^{(0)}, \ldots, \beta_n^{(0)})^{\mathrm{T}}$. Ahora vamos a construir explícitamente una matriz $\mathbf{A}$ y un vector $\mathbf{b}$ tales que $\mathbf{A}\tilde{\mathbf{x}} = \mathbf{b}$. Sean $\mathbf{A} = (\alpha_{ij})$, $\mathbf{b} = (\beta_1, \ldots, \beta_n)^{\mathrm{T}}$ y $\tilde{\mathbf{x}} = (\tilde{\xi}_1, \ldots, \tilde{\xi}_n)$ tales que

$$\alpha_{ij} = \begin{cases} \alpha_{ij}^{(0)} + \dfrac{\varrho_i \varepsilon_{ij}\operatorname{sgn}(\tilde{\xi}_j)}{\sigma_i} & \text{si } \sigma_i \neq 0, \\ \alpha_{ij}^{(0)} & \text{sino,} \end{cases} \quad \beta_i = \begin{cases} \beta_i^{(0)} - \dfrac{\varrho_i \delta_i}{\sigma_i} & \text{si } \sigma_i \neq 0, \\ \beta_i^{(0)} & \text{sino.} \end{cases}$$

Entonces tenemos que $|\mathbf{A} - \mathbf{A}_0| \leqslant \mathbf{E}$, $|\mathbf{b} - \mathbf{b}_0| \leqslant \mathbf{d}$, y

$$\begin{aligned}
\mathbf{e}_i^{\mathrm{T}}(\mathbf{A}\tilde{\mathbf{x}} - \mathbf{b}) &= \sum_{j=1}^n \alpha_{ij}\tilde{\xi}_i - \beta_i \\
&= \begin{cases} \displaystyle\sum_{j=1}^n \alpha_{ij}^{(0)}\tilde{\xi}_j - \beta_i^{(0)} = 0 & \text{si } \sigma_i = 0, \\ \underbrace{\displaystyle\sum_{j=1}^n \alpha_{ij}^{(0)}\tilde{\xi}_j - \beta_i^{(0)}}_{=-\varrho_i} + \dfrac{\varrho_i}{\sigma_i}\underbrace{\displaystyle\sum_{j=1}^n \varepsilon_{ij}\operatorname{sgn}(\tilde{\xi}_j)\tilde{\xi}_j + \delta_i}_{=\sigma_i} = 0 & \text{sino.} \end{cases}
\end{aligned}$$

■

El sistema de desigualdades

$$|\mathbf{b}_0 - \mathbf{A}_0\tilde{\mathbf{x}}| \leqslant \mathbf{E}|\tilde{\mathbf{x}}| + \mathbf{d} \Longleftrightarrow -\mathbf{E}|\tilde{\mathbf{x}}| - \mathbf{d} \leqslant \mathbf{b}_0 - \mathbf{A}_0\tilde{\mathbf{x}} \leqslant \mathbf{E}|\tilde{\mathbf{x}}| + \mathbf{d}$$

representa un sistema de desigualdades lineales por trozos, dado que hay desigualdades diferentes para las 2^n distribuciones posibles de los signos de las componentes de $\tilde{\mathbf{x}}$. El conjunto de sus soluciones, es decir, el conjunto de las soluciones aproximadas compatibles con $(\mathbf{E}, \mathbf{d})$ se reduce para $\mathbf{E} = 0$, $\mathbf{d} = 0$ al punto $\mathbf{x}_0$, la solución de $\mathbf{A}_0\mathbf{x}_0 = \mathbf{b}_0$. Si para $(\mathbf{E}, \mathbf{d})$ "pequeño" el conjunto es "grande", eso significa que la matriz $\mathbf{A}$ es mal acondicionada.

Las consideraciones de esta sección son de gran interés para el análisis del efecto de errores de redondeo durante la solución de sistemas lineales. Los computadores y las calculadoras representan un número real con un número fijo de dígitos en la forma

$$\xi = \pm\beta^k \sum_{i=1}^{t} \zeta_i \beta^{-i}, \quad \beta \in \{2, 10, 16\},$$

donde β es la "base" de la representación, k es un número entero (en un cierto rango) y $\zeta_i \in \{0, 1, \ldots, \beta - 1\}$, con $\zeta_1 \neq 0$ si $\xi \neq 0$.

Ahora, las operaciones aritméticas con tales números no entregan precisamente números del mismo tipo. El redondeo tiene como consecuencia que las operaciones aritméticas no pueden ser ejecutadas de forma exacta, sino que sólo de forma "aproximadamente exacta" como "aritmética de máquina" o "pseudo-aritmética". Ahora, si

$$m(\alpha \# \beta), \quad \# \in \{+, -, \cdot, /\}$$

denota esta aritmética, se puede demostrar que casí siempre tenemos

$$m(\alpha \# \beta) = (\alpha \# \beta)(1 + \eta), \quad \# \in \{+, -, \cdot, /\}, \tag{3.36}$$

donde $|\eta| \leqslant \varepsilon$, y $\varepsilon := \beta^{-t+1}$ es la "exactitud de máquina".

Ahora, si usamos el algoritmo de Gauss con esa aritmética de máquina y $\mathbf{x}_\varepsilon$ denota el resultado, se puede demostrar (usando (3.36)) que $(\mathbf{A} + \mathbf{E})\mathbf{x}_\varepsilon = \mathbf{b}$, donde $\mathbf{A}$ y $\mathbf{b}$ son los coeficientes de entrada realmente usados y la matriz $\mathbf{E}$ satisface la siguiente desigualdad:

$$\begin{aligned} \|\mathbf{E}\|_\infty &\leqslant 1{,}2(n^3 + n^2)\varepsilon\gamma \quad \text{si } n\varepsilon \leqslant 0{,}09, \\ \gamma &:= \max\big\{ |\alpha_{ij}^{(k)}| \;\big|\; k \leqslant i, j \leqslant n,\ 1 \leqslant k \leqslant n \big\}. \end{aligned} \tag{3.37}$$

El valor de γ depende *decisivamente* de la estrategia del pivoteo. Junto con la tarea de hacer ejecutable el algoritmo (evitando divisiones por cero), la estrategia de pivoteo sirve para garantizar que γ no crece demasiado. Se puede demostrar que

$$\gamma \leqslant 2^{n-1} \max_{1 \leqslant i,j \leqslant n} |\alpha_{ij}|$$

con búsqueda del pivote en la columna, y

$$\gamma \leqslant \sqrt{n}\sqrt{2 \cdot 3^{1/2} \cdot n^{1/(n-1)}} \max_{1 \leqslant i,j \leqslant n} |\alpha_{ij}|$$

con búsqueda del pivote en la matriz restante.

Los valores de $\gamma/\max_{1\leqslant i,j\leqslant n}|\alpha_{ij}|$ observados en la práctica son de la magnitud entre 1 y 10. Eso significa que $\mathbf{x}_\varepsilon$ es la solución exacta de un sistema lineal con datos de entrada ligeramente modificados, por ejemplo datos compatibles en el senito del criterio de Prager & Oettli con $\mathbf{d} = 0$ y

$$\mathbf{E} = 1{,}2(n^3+n^2)\varepsilon\gamma \begin{bmatrix} 1 & \cdots & 1 \\ \vdots & & \vdots \\ 1 & \cdots & 1 \end{bmatrix}.$$

En virtud de esta discusión, el algoritmo de Gauss con estrategía de pivoteo se llama *benigno* o *estable*.

Si se calcula la descomposición triangular de una matriz singular con estrategía de pivoteo en aritmética de máquina, al lugar de una columna con $\alpha_{ki}^{(i)} = 0$ para $k \geqslant i$ aparece una columna con

$$\alpha_{ki}^{(i)} \leqslant c\varepsilon, \quad k \geqslant i, \tag{3.38}$$

donde c es una constante similar a la de (3.37). Eso significa que bajo el efecto de errores de redondeo, ya no podemos seguramente detecter la singularidad de una matriz. Se terminará la computación al detectar (3.38) con $c = n\gamma$.

Entonces, usando aritmética de máquina es posible que no se puede decidir cual es el rango de una matriz. La descomposición en valores singulares es una buena herramienta para la definción de un "rango numérico" de una matrix $\mathbf{A}$. Por supuesto, esta definición considera la incerteza en los elemntos de $\mathbf{A}$ y la exactitud de la aritmética usada. Por ejemplo, sea $\tilde{\mathbf{A}}$ una aproximación de $\mathbf{A}$ (por ejemplo, por redondeo) con $\|\mathbf{A}-\tilde{\mathbf{A}}\| \leqslant \alpha$, α conocida, y

$$\tilde{\mathbf{A}} = \mathbf{U}\begin{bmatrix}\mathbf{\Sigma} \\ 0\end{bmatrix}\mathbf{V}^*$$

con los valores singulares $\sigma_1 \geqslant \ldots \geqslant \sigma_n > 0$ y $\mathbf{U}$, $\mathbf{V}$ unitarias. Entonces se aceptarán solo aquellos valores singulares como "intrinsicamente diferentes de cero" para los cuales $\sigma_i > \alpha$. Si en tal caso tenemos $\sigma_1 \geqslant \ldots \geqslant \sigma_r > \alpha \geqslant \sigma_{r+1} \geqslant \ldots \geqslant \sigma_1$, el número r se llama "rango numérico" o "pseudo-rango" de $\mathbf{A}$.

Ejemplo 3.5 (Certamen 1, Curso 2010)**.**

a) *Calcular una descomposición triangular* $\mathbf{PAQ} = \mathbf{LR}$*, con búsqueda de pivote en la matriz restante, de la matriz*

$$\mathbf{A} = \begin{bmatrix} 1 & -1 & 2 & 4 \\ 1 & 1 & 3 & -4 \\ -2 & 1 & 3 & 6 \\ 1 & 2 & -1 & 8 \end{bmatrix}. \tag{3.39}$$

Indicar explícitamente las matrices $\mathbf{P}$*,* $\mathbf{Q}$*,* $\mathbf{L}$ *y* $\mathbf{R}$*.*

b) *Resolver el sistema* $\mathbf{Ax} = \mathbf{b}$*, donde* $\mathbf{b} = (15, 7, 24, 17)^{\mathrm{T}}$*.*

c) *Sean* $\mathbf{e} := (1,1,1,1)^{\mathrm{T}}$*,* $\mathbf{E} := \alpha\mathbf{e}\mathbf{e}^{\mathrm{T}}$*, y* $\mathbf{d} := \alpha\mathbf{e}$*. Decidir si los siguientes vectores son una solución aproximada (en el sentido del criterio de Prager & Oettli) del sistema*

$\mathbf{Ax} = \mathbf{b}$ *(i) para* $\alpha = 0{,}1$*, (ii) para* $\alpha = 0{,}5$*:*

$$\mathbf{x}_1 := \begin{pmatrix} 1{,}01 \\ 1{,}98 \\ 3{,}99 \\ 2{,}01 \end{pmatrix}, \quad \mathbf{x}_2 := \begin{pmatrix} 1{,}05 \\ 1{,}95 \\ 3{,}95 \\ 2{,}05 \end{pmatrix}.$$

Solución sugerida.

a) *Se obtiene la siguiente sucesión de esquemas, donde los elementos marcados con estrella corresponden a multipolicadors y los elementos con marco son el pivote actual:*

	1	2	3	4	
1	1	−1	2	4	15
2	1	1	3	−4	7
3	−2	1	3	6	24
4	1	2	−1	$\boxed{8}$	17

$\rightarrow$

	4	2	3	1	
4	8	2	−1	1	17
2	−4	1	3	1	7
3	6	1	3	−2	24
1	4	−1	2	1	15

$\rightarrow$

	4	2	3	1	
4	8	2	−1	1	17
2	$-\frac{1}{2}^*$	2	$\frac{5}{2}$	$\frac{3}{2}$	$\frac{31}{2}$
3	$\frac{3}{4}^*$	$-\frac{1}{2}$	$\boxed{\frac{15}{4}}$	$-\frac{11}{4}$	$\frac{45}{4}$
1	$\frac{1}{2}^*$	−2	$\frac{5}{2}$	$\frac{1}{2}$	$\frac{13}{2}$

$\rightarrow$

	4	3	2	1	
4	8	−1	2	1	17
3	$\frac{3}{4}^*$	$\frac{15}{4}$	$-\frac{1}{2}$	$-\frac{11}{4}$	$\frac{45}{4}$
2	$-\frac{1}{2}^*$	$\frac{5}{2}$	2	$\frac{3}{2}$	$\frac{31}{2}$
1	$\frac{1}{2}^*$	$\frac{5}{2}$	−2	$\frac{1}{2}$	$\frac{13}{2}$

$\rightarrow$

	4	3	2	1	
4	8	−1	2	1	17
3	$\frac{3}{4}^*$	$\frac{15}{4}$	$-\frac{1}{2}$	$-\frac{11}{4}$	$\frac{45}{4}$
2	$-\frac{1}{2}^*$	$\frac{2}{3}^*$	$\frac{7}{3}$	$\boxed{\frac{10}{3}}$	8
1	$\frac{1}{2}^*$	$\frac{2}{3}^*$	$-\frac{5}{3}$	$\frac{7}{3}$	−1

$\rightarrow$

	4	3	1	2	
4	8	−1	1	2	17
3	$\frac{3}{4}^*$	$\frac{15}{4}$	$-\frac{11}{4}$	$-\frac{1}{2}$	$\frac{45}{4}$
2	$-\frac{1}{2}^*$	$\frac{2}{3}^*$	$\frac{10}{3}$	$\frac{7}{3}$	8
1	$\frac{1}{2}^*$	$\frac{2}{3}^*$	$\frac{7}{3}$	$-\frac{5}{3}$	−1

$$
\begin{array}{c|cccc|c}
 & 4 & 3 & 1 & 2 & \\ \hline
4 & 8 & -1 & 1 & 2 & 17 \\
3 & \frac{3}{4}^{*} & \frac{15}{4} & -\frac{11}{4} & -\frac{1}{2} & \frac{45}{4} \\
2 & -\frac{1}{2}^{*} & \frac{2}{3}^{*} & \frac{10}{3} & \frac{7}{3} & 8 \\
1 & \frac{1}{2}^{*} & \frac{2}{3}^{*} & \frac{7}{10}^{*} & -\frac{33}{10} & -\frac{33}{5}
\end{array}
\quad \rightarrow \quad ,
$$

es decir

$$
\mathbf{P} = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix}; \quad \mathbf{Q} = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix},
$$

$$
\mathbf{L} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ \frac{3}{4} & 1 & 0 & 0 \\ -\frac{1}{2} & \frac{2}{3} & 1 & 0 \\ \frac{1}{2} & \frac{2}{3} & \frac{7}{10} & 1 \end{bmatrix}, \quad \mathbf{R} = \begin{bmatrix} 8 & -1 & 1 & 2 \\ 0 & \frac{15}{4} & -\frac{11}{4} & -\frac{1}{2} \\ 0 & 0 & \frac{10}{3} & \frac{7}{3} \\ 0 & 0 & 0 & -\frac{33}{10} \end{bmatrix}.
$$

b) *Utilizando el último esquema obtenemos*

$$
x_2 = \frac{-\frac{33}{5}}{-\frac{33}{10}} = 2, \quad x_1 = \frac{3}{10}\left(8 - \frac{14}{3}\right) = 1,
$$

$$
x_3 = \frac{4}{15}\left(\frac{45}{4} + \frac{11}{4} + 1\right) = 4, \quad x_4 = \frac{1}{8}(17 + 4 - 1 - 4) = 2.
$$

c) *Aquí se calcula para* $\mathbf{b}_0 = \mathbf{b}$, $\mathbf{A}_0 = \mathbf{A}$*:*

$$
|\mathbf{b}_0 - \mathbf{A}_0\mathbf{x}_1| = \begin{pmatrix} 0{,}05 \\ 0{,}08 \\ 0{,}01 \\ 0{,}06 \end{pmatrix}, \quad |\mathbf{b}_0 - \mathbf{A}_0\mathbf{x}_2| = \begin{pmatrix} 0{,}2 \\ 0{,}35 \\ 0 \\ 0{,}4 \end{pmatrix},
$$

$$
\mathbf{e}\mathbf{e}^{\mathrm{T}}|\mathbf{x}_1| + \mathbf{e} = \begin{pmatrix} 9{,}99 \\ 9{,}99 \\ 9{,}99 \\ 9{,}99 \end{pmatrix}, \quad \mathbf{e}\mathbf{e}^{\mathrm{T}}|\mathbf{x}_2| + \mathbf{e} = \begin{pmatrix} 10 \\ 10 \\ 10 \\ 10 \end{pmatrix}.
$$

Puesto que

$$
|\mathbf{b}_0 - \mathbf{A}_0\mathbf{x}_i| < 0{,}1\big(\mathbf{e}\mathbf{e}^{\mathrm{T}}|\mathbf{x}_i| + \mathbf{e}\big) < 0{,}5\big(\mathbf{e}\mathbf{e}^{\mathrm{T}}|\mathbf{x}_i| + \mathbf{e}\big),
$$

para $i = 1, 2$, ambos vectores $\mathbf{x}_1$ y $\mathbf{x}_2$ son una solución aproximada (en el sentido del criterio de Prager & Oettli) para $\alpha = 0{,}1$ y $\alpha = 0{,}5$.

3.3. El método de cuadrados mínimos y la transformación de una matriz $n \times n$ a una matriz triangular superior

En muchas aplicaciones se presenta el siguiente problema. Para una matriz $\mathbf{A} \in \mathbb{R}^{m\times n}$ (en general, $m \gg n$) y un vector $\mathbf{b} \in \mathbb{R}^m$ se busca un vector $\mathbf{x}^* \in \mathbb{R}^n$ tal que

$$\forall \mathbf{x} \in \mathbb{R}^n : \quad \|\mathbf{A}\mathbf{x}^* - \mathbf{b}\|_2^2 \leqslant \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2. \tag{3.40}$$

Por ejemplo, cuando tenemos puntos de mediciones (t_i, y_i), $i = 1, \ldots, m$, $m \gg 3$, buscamos una función

$$t \mapsto \alpha_0^* + \alpha_1^* t + \alpha_2^* t^2$$

que aproxima nuestros datos. Para tal efecto, hay que determinar los coeficientes α_0^*, α_1^* y α_2^* óptimos, en el sentido de

$$\sum_{i=1}^{m} \big(y_i - (\alpha_0^* + \alpha_1^* t_i + \alpha_2^* t_i^2)\big)^2 = \min_{\alpha_0,\alpha_1,\alpha_2\in\mathbb{R}} \sum_{i=1}^{m} \big(y_i - (\alpha_0 + \alpha_1 t_i + \alpha_2 t_i^2)\big)^2.$$

Definiendo

$$\mathbf{b} = \begin{pmatrix} y_1 \\ \vdots \\ y_m \end{pmatrix}, \quad \mathbf{x}^* = \begin{pmatrix} \alpha_0^* \\ \alpha_1^* \\ \alpha_2^* \end{pmatrix}, \quad \mathbf{A} = \begin{bmatrix} 1 & t_1 & t_1^2 \\ \vdots & \vdots & \vdots \\ 1 & t_m & t_m^2 \end{bmatrix}$$

obtenemos el problema planteado (3.40).

El problema (3.40) significa que entre todas las combinaciones lineales de las columnas de $\mathbf{A}$, buscamos la combinación que minimiza la distancia (euclidiana) del vector fijo $\mathbf{b}$. El vector $\mathbf{b}$ puede ser interpretado como una "función" sobre $\{1, \ldots, m\}$. Eso motiva la denominación *aproximación lineal discreta en L^2* para el problema (3.40). Este tipo de aproximación fue por primera vez definido por Gauss ("método de cuadrados mínimos"). Tiene una motivación que proviene de la estadística: si los y_i son hasta ciertos errores ε_i iguales a $\bar\alpha_0 + \bar\alpha_1 t_i + \bar\alpha_2 t_i^2$, entonces resulta que α_0^*, α_1^* y α_2^*, en un sentido, son las mejores aproximaciones a los "verdaderos" valores $\bar\alpha_0$, $\bar\alpha_1$ y $\bar\alpha_2$; de tal forma, el problema (3.40) minimiza la influencia de los errores ε_i para la determinación de los α's ("compensación" de la influencia de los errores).

El problema (3.40) admite una solución elemental si usamos el Teorema 3.6 y las matrices de Householder. Primero, recordamos que

$$\big\|\mathbf{Q}(\mathbf{A}\mathbf{x} - \mathbf{b})\big\|_2^2 = \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2$$

para cada matriz ortonormal $\mathbf{Q} \in \mathbb{R}^{m\times m}$. Supongamos ahora que se conoce una matriz ortonormal $\mathbf{Q}$ tal que

$$\mathbf{Q}\mathbf{A} = \begin{bmatrix} \mathbf{R} \\ 0 \end{bmatrix}, \quad \mathbf{R} \in \mathbb{R}^{n\times n} \text{ triangular superior.}$$

Ahora, definimos

$$\mathbf{Qb} =: \mathbf{c} = \begin{pmatrix} \mathbf{c}_1 \\ \mathbf{c}_2 \end{pmatrix}, \quad \mathbf{c}_1 \in \mathbb{R}^n, \quad \mathbf{c}_2 \in \mathbb{R}^{m-n}.$$

Entonces,

$$\big\|\mathbf{Q}(\mathbf{Ax}-\mathbf{b})\big\|_2^2 = \left\| \begin{bmatrix} \mathbf{R} \\ 0 \end{bmatrix} \mathbf{x} - \begin{pmatrix} \mathbf{c}_1 \\ \mathbf{c}_2 \end{pmatrix} \right\|_2^2 = \|\mathbf{Rx}-\mathbf{c}_1\|_2^2 + \|\mathbf{c}_2\|_2^2.$$

Supongamos que rango$(\mathbf{A}) = n$. En este caso, $\mathbf{R}$ es regular y

$$\forall \mathbf{x} \in \mathbb{R}^n: \quad \|\mathbf{Rx}-\mathbf{c}_1\|_2^2 + \|\mathbf{c}_2\|_2^2 \geqslant \|\mathbf{c}_2\|_2^2 + \| \underbrace{\mathbf{Rx}^* - \mathbf{c}_1}_{=0} \|_2^2,$$

es decir, $\mathbf{x}^*$ es la solución del sistema escalonado $\mathbf{Rx}^* = \mathbf{c}_1$, o sea

$$\mathbf{x}^* = \mathbf{R}^{-1}\mathbf{c}_1.$$

Entonces, cuando se ha encontrado la matriz de transformación $\mathbf{Q}$, solamente hay que aplicar $\mathbf{Q}$ a $\mathbf{b}$ y resolver el sistema escalonado $\mathbf{Rx}^* = \mathbf{c}_1$.

La determinación de $\mathbf{Q}$ (y entonces la de $\mathbf{R}$) se ejecuta en n pasos (o $n-1$ pasos si $n = m$). Sea

$$\mathbf{A} = \begin{bmatrix} \mathbf{a}_1^{(1)} & \cdots & \mathbf{a}_n^{(1)} \end{bmatrix}, \quad \mathbf{b}^{(1)} := \mathbf{b}, \quad \mathbf{a}_j^{(1)} := \begin{pmatrix} \alpha_{1j}^{(1)} \\ \vdots \\ \alpha_{nj}^{(1)} \end{pmatrix}.$$

Definiendo

$$\operatorname{sgn}_0(x) := \begin{cases} 1 & \text{si } x = 0, \\ \operatorname{sgn}(x) & \text{sino,} \end{cases}$$

formamos una matriz $\mathbf{U}_1$ ortonormal y simétrica $\mathbf{U}_1 := \mathbf{I} - \beta_1 \hat{\mathbf{w}}_1 \hat{\mathbf{w}}_1^{\mathrm{T}}$ definida por

$$\beta_1 := \frac{1}{\big\|\mathbf{a}_1^{(1)}\big\|_2 \big(\big|\alpha_{11}^{(1)}\big| + \big\|\mathbf{a}_1^{(1)}\big\|_2\big)}, \quad \hat{\mathbf{w}}_1 = \begin{pmatrix} \operatorname{sgn}_0\big(\alpha_{11}^{(1)}\big)\big(\big|\alpha_{11}^{(1)}\big| + \big\|\mathbf{a}_1^{(1)}\big\|_2\big) \\ \alpha_{21}^{(1)} \\ \vdots \\ \alpha_{m1}^{(1)} \end{pmatrix}. \tag{3.41}$$

Entonces tenemos

$$\mathbf{U}_1 \mathbf{a}_1^{(1)} = -\operatorname{sgn}_0\big(\alpha_{11}^{(1)}\big)\|\mathbf{a}_1\|\mathbf{e}_1.$$

Ahora definimos

$$\begin{aligned} \mathbf{a}_i^{(2)} &:= \mathbf{U}_1 \mathbf{a}_i^{(1)} = \mathbf{a}_i^{(1)} - \beta_1 \big(\hat{\mathbf{w}}_1^{\mathrm{T}} \mathbf{a}_i^{(1)}\big) \hat{\mathbf{w}}_1, \quad i = 2, \ldots, n, \\ \mathbf{b}^{(2)} &:= \mathbf{U}_1 \mathbf{b}^{(1)} = \mathbf{b}^{(1)} - \beta_1 \big(\hat{\mathbf{w}}_1^{\mathrm{T}} \mathbf{b}^{(1)}\big) \hat{\mathbf{w}}_1. \end{aligned}$$

Queremos aplicar la misma técnica que para $\mathbf{a}_1^{(1)}$ a las últimas $m-1$ componentes de $\mathbf{a}_2^{(2)}$, mientras que la nueva transformación debe dejar sin cambio a la fila 1 y la columna 1 de la matriz transformada. Eso se logra definiendo

$$\mathbf{U}_2 = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & & & \\ \vdots & & \mathbf{I} - \beta_2 \hat{\mathbf{w}}_2 \hat{\mathbf{w}}_2^{\mathrm{T}} & \\ 0 & & & \end{bmatrix},$$

donde

$$\beta_2 := \frac{1}{\|\tilde{\mathbf{a}}_2^{(2)}\|_2 \big(|\alpha_{22}^{(2)}| + \|\tilde{\mathbf{a}}_2^{(2)}\|_2\big)}, \quad \hat{\mathbf{w}}_2 = \begin{pmatrix} \operatorname{sgn}_0\big(\alpha_{22}^{(2)}\big)\big(|\alpha_{22}^{(2)}| + \|\tilde{\mathbf{a}}_2^{(2)}\|_2\big) \\ \alpha_{32}^{(2)} \\ \vdots \\ \alpha_{m2}^{(2)} \end{pmatrix}, \quad \tilde{\mathbf{a}}_2^{(2)} := \begin{pmatrix} \alpha_{22}^{(2)} \\ \vdots \\ \alpha_{m2}^{(2)} \end{pmatrix}.$$

Así se continua la construcción. En general, obtenemos el siguiente algoritmo.

Algoritmo 3.1.

1. *Definición de n':*

$$n' \leftarrow \begin{cases} n-1 & \text{si } m = n, \\ n & \text{sino.} \end{cases}$$

2. **do** $i = 1, \ldots, n'$

$$\mathbf{a}_i^{(i)} = \begin{pmatrix} \tilde{\tilde{\mathbf{a}}}_i^{(i)} \\ \tilde{\mathbf{a}}_i^{(i)} \end{pmatrix}, \quad \mathbf{b}^{(i)} = \begin{pmatrix} \tilde{\tilde{\mathbf{b}}}^{(i)} \\ \tilde{\mathbf{b}}^{(i)} \end{pmatrix}, \quad \tilde{\mathbf{a}}_i^{(i)}, \tilde{\mathbf{b}}^{(i)} \in \mathbb{R}^{m-i+1},$$

if $\tilde{\mathbf{a}}_i^{(i)} \neq 0$ **then**

$$\beta_i \leftarrow \frac{1}{\|\tilde{\mathbf{a}}_i^{(i)}\|_2 \big(|\alpha_{ii}^{(i)}| + \|\tilde{\mathbf{a}}_i^{(i)}\|_2\big)}, \quad \hat{\mathbf{w}}_i \leftarrow \begin{pmatrix} \operatorname{sgn}_0\big(\alpha_{ii}^{(i)}\big)\big(|\alpha_{ii}^{(i)}| + \|\tilde{\mathbf{a}}_i^{(i)}\|_2\big) \\ \alpha_{i+1,i}^{(i)} \\ \vdots \\ \alpha_{mi}^{(i)} \end{pmatrix}$$

else

$$\beta_i \leftarrow 0, \quad \hat{\mathbf{w}}_i \leftarrow 0 \quad (\text{o sea, } \mathbf{U}_i = \mathbf{I})$$

endif
do $j = i, \ldots, n$

$$\begin{aligned} \tilde{\tilde{\mathbf{a}}}_j^{(i+1)} &\leftarrow \tilde{\tilde{\mathbf{a}}}_j^{(i)}, \\ \tilde{\mathbf{a}}_j^{(i+1)} &\leftarrow \tilde{\mathbf{a}}_j^{(i)} - \beta_i \big(\hat{\mathbf{w}}_i^{\mathrm{T}} \tilde{\mathbf{a}}_j^{(i)}\big) \hat{\mathbf{w}}_i, \\ \tilde{\tilde{\mathbf{b}}}^{(i+1)} &\leftarrow \tilde{\tilde{\mathbf{b}}}^{(i)}, \end{aligned}$$

$$\tilde{\mathbf{b}}^{(i+1)} \leftarrow \tilde{\mathbf{b}}^{(i)} - \beta_i \big(\hat{\mathbf{w}}_i^{\mathrm{T}} \tilde{\mathbf{b}}^{(i)}\big) \hat{\mathbf{w}}_i$$

enddo

enddo

3. *El resultado de los pasos anteriores es*

$$\underbrace{\mathbf{U}_{n'} \cdot \ldots \cdot \mathbf{U}_1}_{=:\mathbf{Q}} \mathbf{A} = \begin{bmatrix} \mathbf{a}_1^{(2)} & \cdots & \mathbf{a}_n^{(n'+1)} \end{bmatrix} = \begin{bmatrix} \alpha_{11}^{(2)} & \alpha_{12}^{(3)} & \cdots & \alpha_{1n}^{(n'+1)} \\ 0 & \alpha_{22}^{(3)} & & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & \alpha_{nn}^{(n'+1)} \\ 0 & \cdots & \cdots & 0 \\ \vdots & & & \vdots \\ 0 & \cdots & \cdots & 0 \end{bmatrix} = \begin{bmatrix} \mathbf{R} \\ 0 \end{bmatrix},$$

$$\mathbf{U}_{n'} \cdot \ldots \cdot \mathbf{U}_1 \mathbf{b} = \mathbf{b}^{(n'+1)} = \begin{pmatrix} \mathbf{c}_1 \\ \mathbf{c}_2 \end{pmatrix}.$$

Como producto "secundario" de este algoritmo obtenemos una demostración del siguiente teorema.

Teorema 3.11. *Sea* $\mathbf{A} \in \mathbb{R}^{m \times n}$ *con* $m \geqslant n$. *Entonces existe una matriz ortonormal* $\mathbf{Q} \in \mathbb{R}^{m \times m}$ *tal que*

$$\mathbf{QA} = \begin{bmatrix} \mathbf{R} \\ 0 \end{bmatrix},$$

donde $\mathbf{R}$ *es una matriz triangular superior. Si* $\operatorname{rango}(\mathbf{A}) = n$, *entonces el problema* (3.40) *tiene una única solución* $\mathbf{x}$, *la cual se calcula de* $\mathbf{R}\mathbf{x}^* = \mathbf{c}_1$, *donde*

$$\mathbf{Qb} = \begin{pmatrix} \mathbf{c}_1 \\ \mathbf{c}_2 \end{pmatrix}, \quad \mathbf{c}_1 \in \mathbb{R}^n.$$

En este caso, $\mathbf{R}$ *es regular.*

El método descrito aquí es conocido como *transformación de Householder* u *ortogonalización de Householder*. Esta nomenclatura se explica de la siguiente forma:

$$\mathbf{QA} = \begin{bmatrix} \mathbf{R} \\ 0 \end{bmatrix} \Longrightarrow \mathbf{A} = \mathbf{Q}^{\mathrm{T}} \begin{bmatrix} \mathbf{R} \\ 0 \end{bmatrix} = \begin{bmatrix} \mathbf{Q}_1^{\mathrm{T}} & \mathbf{Q}_2^{\mathrm{T}} \end{bmatrix} \begin{bmatrix} \mathbf{R} \\ 0 \end{bmatrix} = \mathbf{Q}_1^{\mathrm{T}} \mathbf{R},$$

es decir, las n columnas de $\mathbf{Q}_1^{\mathrm{T}}$ (o sea las primeras n filas de $\mathbf{Q}$) forman una base ortonormal del subespacio de $\mathbb{R}^m$ generado por las columnas de $\mathbf{A}$, y las $m - n$ últimas filas de $\mathbf{Q}$ son una base ortonormal del complemento ortogonal, es decir, del espacio nulo de $\mathbf{A}^{\mathrm{T}}$. Hay que tomar en cuenta, sin embargo, que la matriz $\mathbf{Q}$ solo aparece en forma factorizada.

Para el cálculo de $\mathbf{A} = \mathbf{UR}$, $\mathbf{U}^{\mathrm{T}}\mathbf{U} = \mathbf{I}$, $\mathbf{U} \in \mathbb{R}^{m \times n}$ (donde $\mathbf{U}$ corresponde a $\mathbf{Q}_1^{\mathrm{T}}$) podríamos también usar el método de ortogonalización de Gram-Schmidt. Pero este método es inestable numéricamente, así que se prefiere la transformación de Householder.

Por otro lado, uno podría aplicar el método en el caso $m = n$, es decir, para la solución de sistemas. El esfuerzo es el doble del algoritmo de Gauss, asi que efectivamente se prefiere el

algoritmo de Gauss, sobre todo en virtud de la equivalencia de las propiedades de estabilidad de ambos algoritmos.

Ejemplo 3.6. *Para la transformación de Householder de la matriz*

$$\mathbf{A} = \begin{bmatrix} 4 & 3 \\ -4 & -1 \\ 4 & 3 \\ -4 & -1 \end{bmatrix}$$

calculamos sucesivamente las siguientes cantidades:

$$\tilde{\mathbf{a}}_1^{(1)} = \begin{pmatrix} 4 \\ -4 \\ 4 \\ -4 \end{pmatrix}, \quad \beta_1 = \frac{1}{8(4+8)} = \frac{1}{96}, \quad \hat{\mathbf{w}}_1 = \begin{pmatrix} 12 \\ -4 \\ 4 \\ -4 \end{pmatrix},$$

luego

$$\mathbf{a}_1^{(2)} = \mathbf{a}_1^{(1)} - \frac{1}{96}(48+16+16+16)\hat{\mathbf{w}}_1 = \begin{pmatrix} -8 \\ 0 \\ 0 \\ 0 \end{pmatrix},$$

$$\mathbf{a}_2^{(2)} = \mathbf{a}_2^{(1)} - \frac{1}{96}(12 \cdot 3 + 4 \cdot 1 + 4 \cdot 3 + 4 \cdot 1) = \begin{pmatrix} -4 \\ \frac{4}{3} \\ \frac{2}{3} \\ \frac{4}{3} \end{pmatrix}.$$

Después, calculamos

$$\left\|\tilde{\mathbf{a}}_2^{(2)}\right\|_2 = \frac{1}{3}(16+16+4)^2 = 2, \beta_2 = \frac{1}{2\left(\frac{4}{3}+2\right)}, \quad \hat{\mathbf{w}}_2 = \begin{pmatrix} \frac{10}{3} \\ \frac{2}{3} \\ \frac{4}{3} \end{pmatrix},$$

$$\hat{\mathbf{w}}_2^{\mathrm{T}}\tilde{\mathbf{a}}_2^{(2)} = \frac{1}{9}(10 \cdot 4 + 2 \cdot 2 + 4 \cdot 4) = \frac{20}{3}.$$

Finalmente resulta

$$\mathbf{a}_2^{(3)} = \begin{pmatrix} -4 \\ \frac{4}{3} - \frac{10}{3} \\ \frac{2}{3} - \frac{2}{3} \\ \frac{4}{3} - \frac{4}{3} \end{pmatrix} = \begin{pmatrix} -4 \\ -2 \\ 0 \\ 0 \end{pmatrix}, \quad \textit{o sea} \quad \underbrace{\mathbf{U}_2\mathbf{U}_1}_{=\mathbf{Q}}\mathbf{A} = \begin{bmatrix} -8 & -4 \\ 0 & -2 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}.$$

Ejemplo 3.7 (Tarea 13, Curso 2006)**.** *Sea* $\mathbf{A}$ *la matriz indicada, y buscamos una descomposicion* $\mathbf{QR}$ *de* $\mathbf{A}$ *con* $\mathbf{R} = \mathbf{Q}^{\mathrm{T}}\mathbf{A}$*. ¿La matriz* $\mathbf{R}$ *dada abajo puede ser correcta?*

$$\mathbf{A} = \begin{bmatrix} 2 & 3 & 2 \\ 2 & 1 & 0 \\ 1 & 2 & 3 \\ -1 & 4 & 2 \end{bmatrix}, \quad \mathbf{R} = \begin{bmatrix} -\sqrt{10} & 43 & 4 \\ 0 & 7 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}.$$

Solución sugerida. *La matriz* $\mathbf{R}$ *debe ser incorrecta, dado que*

$$43^2 + 7^2 \neq 3^2 + 1^2 + 2^2 + 4^2,$$

y porque una aplicación ortogonal no cambia la norma $\|\cdot\|_2$ *de un vector. Sea* $\mathbf{A} = \begin{bmatrix} \mathbf{a}_1 & \cdots & \mathbf{a}_n \end{bmatrix}$ *y* $\mathbf{R} = \begin{bmatrix} \mathbf{r}_1 & \cdots & \mathbf{r}_n \end{bmatrix}$ *con vectores de columnas* $\mathbf{a}_i$ *y* $\mathbf{r}_i$*. Entonces* $\mathbf{Q}\mathbf{a}_i = \mathbf{r}_i$ *implica que* $\|\mathbf{Q}\mathbf{a}_i\|_2 = \|\mathbf{r}_i\|_2$*, y como* $\mathbf{Q}$ *es ortogonal,* $\|\mathbf{a}_i\|_2 = \|\mathbf{r}_i\|_2$*.*

Ejemplo 3.8 (Tarea 15, Curso 2006)**.** *Sean*

$$\mathbf{A} = \begin{bmatrix} -1 & 1 \\ 1 & -1 \\ -1 & -1 \\ 1 & 1 \end{bmatrix}, \quad \mathbf{b}_1 = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}, \quad \mathbf{b}_2 = \begin{pmatrix} -1 \\ -1 \\ 1 \\ 1 \end{pmatrix}.$$

a) *Determinar la descomposición* $\mathbf{QR}$ *de* $\mathbf{A}$*. Para* $\mathbf{Q}$*, indicar sólo los vectores* $\hat{\mathbf{u}}_1$ *y* $\hat{\mathbf{u}}_2$ *de la representación*

$$\mathbf{Q} = \left(\mathbf{I} - \frac{2}{\hat{\mathbf{u}}_2^{\mathrm{T}}\hat{\mathbf{u}}_2}\hat{\mathbf{u}}_2\hat{\mathbf{u}}_2^{\mathrm{T}}\right)\left(\mathbf{I} - \frac{2}{\hat{\mathbf{u}}_1^{\mathrm{T}}\hat{\mathbf{u}}_1}\hat{\mathbf{u}}_1\hat{\mathbf{u}}_1^{\mathrm{T}}\right).$$

b) *Calcular* $\mathbf{c}_1 := \mathbf{Q}\mathbf{b}_1$ *y* $\mathbf{c}_2 := \mathbf{Q}\mathbf{b}_2$*.*

c) *Usando los resultados de (a) y (b), determinar la solución del problema de compensación*

$$\|\mathbf{A}\mathbf{x} - \mathbf{b}_1 - t\mathbf{b}_2\|_2 \stackrel{!}{=} \min_{\mathbf{x}\in\mathbb{R}^2}$$

para $t \in \mathbb{R}$ *arbitrario.*

Solucion sugerida.

a) *La matriz dada es de la forma*

$$\mathbf{A} = \begin{bmatrix} \mathbf{a}_1^{(1)} & \mathbf{a}_2^{(1)} \end{bmatrix}$$

con

$$\mathbf{a}_1^{(1)} = \begin{pmatrix} -1 \\ 1 \\ -1 \\ 1 \end{pmatrix}, \quad \gamma_1 = \left\|\mathbf{a}_1^{(1)}\right\|_2 = 2, \quad \hat{\mathbf{u}}_1 = \begin{pmatrix} -2 \\ 1 \\ -1 \\ 1 \end{pmatrix},$$

$$\hat{\mathbf{u}}_1^{\mathrm{T}}\hat{\mathbf{u}}_1 = 12, \quad \beta_1 = \frac{1}{6}, \quad \hat{\mathbf{u}}_1^{\mathrm{T}}\mathbf{a}_1^{(1)} = 6, \quad \hat{\mathbf{u}}_1^{\mathrm{T}}\mathbf{a}_2^{(1)} = -2.$$

Entonces obtenemos los vectores

$$\mathbf{a}_1^{(2)} = \begin{pmatrix} 2 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \quad \mathbf{a}_2^{(2)} = \frac{1}{3}\begin{pmatrix} 0 \\ -2 \\ -4 \\ 4 \end{pmatrix}.$$

Luego calculamos que

$$\gamma_2 = 2, \quad \hat{\mathbf{u}}_2 = \frac{1}{3}\begin{pmatrix} 0 \\ -8 \\ -4 \\ 4 \end{pmatrix}, \quad \hat{\mathbf{u}}_2^{\mathrm{T}}\mathbf{a}_2^{(2)} = \frac{48}{9}, \quad \beta_2 = \frac{9}{48}, \quad \mathbf{a}_2^{(3)} = \begin{pmatrix} 0 \\ 2 \\ 0 \\ 0 \end{pmatrix},$$

o sea

$$\mathbf{R} = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}.$$

b) *Calculamos sucesivamente*

$$\mathbf{b}_1^{(2)} = \mathbf{b}_1^{(1)} - \beta_1\big(\hat{\mathbf{u}}_1^{\mathrm{T}}\mathbf{b}_1^{(1)}\big)\hat{\mathbf{u}}_1 = \frac{1}{3}\begin{pmatrix} 0 \\ 4 \\ 2 \\ 4 \end{pmatrix}, \quad \mathbf{b}_2^{(2)} = \mathbf{b}_2^{(1)} - \beta_1\big(\hat{\mathbf{u}}_1^{\mathrm{T}}\mathbf{b}_2^{(1)}\big)\hat{\mathbf{u}}_1 = \frac{1}{3}\begin{pmatrix} 0 \\ -4 \\ 4 \\ 2 \end{pmatrix},$$

$$\mathbf{b}_1^{(3)} = \mathbf{b}_1^{(2)} - \beta_2\big(\hat{\mathbf{u}}_2^{\mathrm{T}}\mathbf{b}_1^{(2)}\big)\hat{\mathbf{u}}_2 = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 2 \end{pmatrix}, \quad \mathbf{b}_2^{(3)} = \mathbf{b}_2^{(3)} - \beta_3\big(\hat{\mathbf{u}}_3^{\mathrm{T}}\mathbf{b}_2^{(3)}\big)\hat{\mathbf{u}}_3 = \begin{pmatrix} 0 \\ 0 \\ 2 \\ 0 \end{pmatrix}.$$

c) *Aquí obtenemos*

$$\begin{aligned} \|\mathbf{A}\mathbf{x} - \mathbf{b}_1 - t\mathbf{b}_2\|_2 &= \big\|\mathbf{Q}(\mathbf{A}\mathbf{x} - \mathbf{b}_1 - t\mathbf{b}_2)\big\|_2 \\ &= \left\|\begin{bmatrix}\mathbf{R}\\ 0\end{bmatrix}\mathbf{x} - \mathbf{c}_1 - t\mathbf{c}_2\right\|_2 = \left\|\begin{pmatrix} 2x_1 \\ 2x_2 \\ -2t \\ -2 \end{pmatrix}\right\|_2 \geqslant \left\|\begin{pmatrix} 0 \\ 0 \\ 2t \\ 2 \end{pmatrix}\right\|_2, \end{aligned}$$

es decir, para todo $t \in \mathbb{R}$, el mínimo se asume para $\mathbf{x} = 0$.

Ejemplo 3.9 (Tarea 16, Curso 2006)**.** *Sean*

$$\mathbf{A} = \begin{bmatrix} -3 & 1 & 1 \\ 6 & 1 & 6 \\ -6 & -4 & 1 \\ 0 & 0 & \sqrt{11} \end{bmatrix}, \quad \mathbf{b} = \begin{pmatrix} 3 \\ 2 \\ 2 \\ 2\sqrt{11} \end{pmatrix}.$$

Usar el método de Householder para determinar una descomposición $\mathbf{QR}$ *de* $\mathbf{A}$*. Luego determinar un vector* $\mathbf{x}$ *que minimiza* $\|\mathbf{Ax} - \mathbf{b}\|_2$*. ¿Cuál es entonces el error en la ecuación* $\mathbf{Ax} - \mathbf{b}$*?*

Solución sugerida.

1. *Consideramos la primera columna* $\mathbf{a}_1$ *de* $\mathbf{A}$*. Usando* $\|\mathbf{a}_1\|_2 = 9$*, obtenemos* $\mathbf{u} = (-12, 6, -6, 0)^{\mathrm{T}}$ *y* $\beta = 1/108$*. Con* $\mathbf{P}_1 = \mathbf{I} - \beta\mathbf{u}\mathbf{u}^{\mathrm{T}}$ *determinamos* $\mathbf{P}_1\mathbf{A}$ *calculando para cada columna* $\mathbf{a}_i$ *de* $\mathbf{A}$ *el vector* $\mathbf{P}_1\mathbf{a}_i$*. El resultado es*

$$\mathbf{P}_1\mathbf{A} = \begin{bmatrix} 9 & 3 & 3 \\ 0 & 0 & 5 \\ 0 & -3 & 2 \\ 0 & 0 & \sqrt{11} \end{bmatrix}, \quad \mathbf{P}_1\mathbf{b} = \begin{pmatrix} -1 \\ 4 \\ 0 \\ 2\sqrt{11} \end{pmatrix}.$$

2. *Luego, para la primera columna*

$$\tilde{\mathbf{a}}_1 = \begin{pmatrix} 0 \\ -3 \\ 0 \end{pmatrix}$$

de la matriz 3×2 *restante obtenemos* $\|\tilde{\mathbf{a}}_1\|_2 = 3$*, y entonces* $\tilde{\mathbf{u}} = (3, -3, 0)^{\mathrm{T}}$ *y* $\beta = 1/9$*. Obtenemos*

$$\tilde{\mathbf{P}}_2\tilde{\mathbf{A}} = \begin{bmatrix} -3 & 2 \\ 0 & 5 \\ 0 & \sqrt{11} \end{bmatrix}, \quad \tilde{\mathbf{P}}_2\tilde{\mathbf{b}} = \begin{pmatrix} 0 \\ 4 \\ 2\sqrt{11} \end{pmatrix}.$$

3. *Luego, para la primera columna*

$$\tilde{\tilde{\mathbf{a}}}_1 = \begin{pmatrix} 5 \\ \sqrt{11} \end{pmatrix}$$

de la matriz 2×1 *restante obtenemos* $\|\tilde{\tilde{\mathbf{a}}}_1\|_2 = 6$*, y entonces* $\tilde{\tilde{\mathbf{u}}} = (11, \sqrt{11})^{\mathrm{T}}$ *y* $\beta = 1/66$*. Obtenemos*

$$\tilde{\tilde{\mathbf{P}}}_3\tilde{\tilde{\mathbf{A}}} = \begin{bmatrix} -6 \\ 0 \end{bmatrix}, \quad \tilde{\tilde{\mathbf{P}}}_3\tilde{\tilde{\mathbf{b}}} = \begin{pmatrix} -7 \\ \sqrt{11} \end{pmatrix}.$$

Después de los 3 pasos anteriores, obtenemos la siguiente descomposición $\mathbf{QR}$*, donde* $\mathbf{Q}$ *es el producto de las matrices de Householder ampliadas a la dimensión 4:*

$$\mathbf{QA} = \begin{bmatrix} 9 & 3 & 3 \\ 0 & -3 & 2 \\ 0 & 0 & -6 \\ 0 & 0 & 0 \end{bmatrix}, \quad \mathbf{Qb} = \begin{pmatrix} -1 \\ 0 \\ -7 \\ \sqrt{11} \end{pmatrix}.$$

Particionando la matriz del tamaño 4×3 *en una matriz del tamaño* 3×3 *y otra del tamaño* 1×3*, obtenemos*

$$\min_{\mathbf{x}\in\mathbb{R}^4} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2 = \min_{\mathbf{x}\in\mathbb{R}^4} \|\mathbf{Q}\mathbf{A}\mathbf{x} - \mathbf{Q}\mathbf{b}\|_2 = \min_{\mathbf{x}\in\mathbb{R}^4} \left\| \begin{bmatrix} \mathbf{R} \\ 0 \end{bmatrix} \mathbf{x} - \begin{pmatrix} \mathbf{c}_1 \\ \mathbf{c}_2 \end{pmatrix} \right\|_2 .$$

Obviamente, el sistema $\mathbf{R}\mathbf{x} = \mathbf{c}_1$ *posee una solución única, entonces*

$$\min_{\mathbf{x}\in\mathbb{R}^4} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2 = \|\mathbf{c}_2\|_2 = \sqrt{11}.$$

Ejemplo 3.10 (Certamen 1, Curso 2010)**.** *Resolver el problema de aproximación*

$$\sum_{i=1}^{m} \big(y_i - (\alpha_0^* \varphi_0(t_i) + \alpha_1^* \varphi_1(t_i) + \alpha_2^* \varphi_2 t_i\big)^2 = \min_{\alpha_0,\alpha_1,\alpha_2} \sum_{i=1}^{m} \big(y_i - (\alpha_0 \varphi_0(t_i) + \alpha_1 \varphi_1(t_i) + \alpha_2 \varphi_2 t_i\big)^2$$

para los datos

i	*1*	*2*	*3*	*4*
t_i	*0*	*1*	*2*	*3*
y_i	*1*	*-1*	*1*	*3*

para $\varphi_i(t) = t^i$*,* $i = 0, 1, 2$*, transformando la matriz* $\mathbf{A} \in \mathbb{R}^{4\times 3}$ *a forma triangular superior mediante la transformación de Householder.*
Solución sugerida. *Sea*

$$\mathbf{A} = \begin{bmatrix} 1 & t_1 & t_1^2 \\ 1 & t_2 & t_2^2 \\ 1 & t_3 & t_3^2 \\ 1 & t_4 & t_4^2 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 2 & 4 \\ 1 & 3 & 9 \end{bmatrix}, \quad \mathbf{b} = \begin{pmatrix} 1 \\ -1 \\ 1 \\ 3 \end{pmatrix}.$$

La transformación de Householder se ejecuta en tres pasos.

1. *Desde la matriz* $\mathbf{A}$ *identificamos*

$$\tilde{\mathbf{a}}_1^{(1)} = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}, \quad \beta_1 = \frac{1}{2(1+2)} = \frac{1}{6}, \quad \hat{\mathbf{w}}_1 = \begin{pmatrix} 3 \\ 1 \\ 1 \\ 1 \end{pmatrix}.$$

Si

$$\mathbf{U}_1 = \mathbf{I} - \frac{1}{6}\hat{\mathbf{w}}_1 \hat{\mathbf{w}}_1^{\mathrm{T}} \in \mathbb{R}^{4\times 4},$$

obtenemos

$$\mathbf{a}_1^{(2)} = \mathbf{U}_1 \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} -2 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \quad \mathbf{a}_2^{(2)} = \mathbf{U}_1 \begin{pmatrix} 0 \\ 1 \\ 2 \\ 3 \end{pmatrix} = \begin{pmatrix} -3 \\ 0 \\ 1 \\ 2 \end{pmatrix},$$

$$\mathbf{a}_3^{(2)} = \mathbf{U}_1 \begin{pmatrix} 0 \\ 1 \\ 4 \\ 9 \end{pmatrix} = \begin{pmatrix} -7 \\ -\frac{4}{3} \\ \frac{5}{3} \\ \frac{20}{3} \end{pmatrix}, \quad \mathbf{b}^{(2)} = \mathbf{U}_1 \mathbf{b} = \begin{pmatrix} -2 \\ -2 \\ 0 \\ 2 \end{pmatrix}.$$

2. *Considerando los 3 últimos componentes de los vectores anteriores obtenemos*

$$\tilde{\mathbf{a}}_2^{(2)} = \begin{pmatrix} 0 \\ 1 \\ 2 \end{pmatrix}, \quad \beta_2 = \frac{1}{\sqrt{5}(0+\sqrt{5})} = \frac{1}{5}, \quad \hat{\mathbf{w}}_2 = \begin{pmatrix} \sqrt{5} \\ 1 \\ 2 \end{pmatrix}.$$

Definiendo

$$\mathbf{U}_2 = \mathbf{I} - \frac{1}{5}\hat{\mathbf{w}}_2 \hat{\mathbf{w}}_2{}^{\mathrm{T}} \in \mathbb{R}^{3\times 3}$$

obtenemos

$$\mathbf{a}_2^{(3)} = \mathbf{U}_2 \begin{pmatrix} 0 \\ 1 \\ 2 \end{pmatrix} = \begin{pmatrix} -\sqrt{5} \\ 0 \\ 0 \end{pmatrix},$$

$$\mathbf{a}_3^{(3)} = \mathbf{U}_2 \begin{pmatrix} -\frac{4}{3} \\ \frac{5}{3} \\ \frac{20}{3} \end{pmatrix} = \begin{pmatrix} -3\sqrt{5} \\ -\frac{4}{3} + \frac{4\sqrt{5}}{15} \\ \frac{2}{3} + \frac{8\sqrt{5}}{15} \end{pmatrix} \approx \begin{pmatrix} -6{,}7082 \\ -0{,}7370 \\ 1{,}8592 \end{pmatrix},$$

$$\mathbf{b}^{(3)} = \mathbf{U}_2 \begin{pmatrix} -2 \\ 0 \\ 2 \end{pmatrix} = \begin{pmatrix} -\frac{4}{5}\sqrt{5} \\ -\frac{4}{5} + \frac{2\sqrt{5}}{5} \\ \frac{2}{5} + \frac{4\sqrt{5}}{5} \end{pmatrix} \approx \begin{pmatrix} -1{,}7889 \\ 0{,}0944 \\ 2{,}1889 \end{pmatrix}.$$

3. *Considerando los 2 últimos componentes de los vectores anteriores obtenemos*

$$\tilde{\mathbf{a}}_3^{(3)} = \begin{pmatrix} -\frac{4}{3} + \frac{4\sqrt{5}}{15} \\ \frac{2}{3} + \frac{8\sqrt{5}}{15} \end{pmatrix} \approx \begin{pmatrix} -0{,}7370 \\ 1{,}8592 \end{pmatrix},$$

$$\beta_3 = \frac{1}{2\left(\frac{4}{3}\left(1-\frac{\sqrt{5}}{5}\right)+2\right)} = \frac{1}{\frac{20}{3}-\frac{8\sqrt{5}}{15}} \approx 0{,}1827,$$

$$\hat{\mathbf{w}}_3 = \begin{pmatrix} -\frac{10}{3}+\frac{4\sqrt{5}}{15} \\ \frac{2}{3}+\frac{8\sqrt{5}}{15} \end{pmatrix} \approx \begin{pmatrix} -2{,}7370 \\ 1{,}8592 \end{pmatrix}.$$

Definiendo

$$\mathbf{U}_3 = \mathbf{I} - \frac{1}{\frac{20}{3}-\frac{8\sqrt{5}}{15}} \hat{\mathbf{w}}_3 \hat{\mathbf{w}}_3{}^{\mathrm{T}} \in \mathbb{R}^{2\times 2}$$

obtenemos

$$\mathbf{a}_3^4 = \mathbf{U}_3 \begin{pmatrix} -\frac{4}{3}+\frac{4\sqrt{5}}{15} \\ \frac{2}{3}+\frac{8\sqrt{5}}{15} \end{pmatrix} = \begin{pmatrix} 2 \\ 0 \end{pmatrix},$$

$$\mathbf{b}^{(4)} = \mathbf{U}_3 \begin{pmatrix} -\frac{4}{5}+\frac{2\sqrt{5}}{5} \\ \frac{2}{5}+\frac{4\sqrt{5}}{5} \end{pmatrix} = \begin{pmatrix} 2 \\ \frac{2}{5}\sqrt{5} \end{pmatrix} \approx \begin{pmatrix} 2 \\ 0{,}8944 \end{pmatrix}.$$

Concluimos que la solución del problema está dada por el sistema lineal escaloneado

$$\begin{bmatrix} -2 & -3 & -7 \\ 0 & -\sqrt{5} & -3\sqrt{5} \\ 0 & 0 & -2 \end{bmatrix} \begin{pmatrix} \alpha_0^* \\ \alpha_1^* \\ \alpha_2^* \end{pmatrix} = \begin{pmatrix} -2 \\ -\frac{4}{5}\sqrt{5} \\ 2 \end{pmatrix}$$

con la solución

$$\alpha_0^* = \frac{4}{5} = 0{,}8, \quad \alpha_1^* = -\frac{11}{5} = -2{,}2, \quad \alpha_2^* = 1.$$

La tarea de minimizar $\|\mathbf{Ax}-\mathbf{b}\|_2^2$ también puede ser tratada por métodos del cálculo diferencial de varias variables. Para tal efecto, definimos la función

$$\begin{aligned} \Phi(\mathbf{x}) &:= (\mathbf{Ax}-\mathbf{b})^{\mathrm{T}}(\mathbf{Ax}-\mathbf{b}) \\ &= \mathbf{x}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{Ax} - 2\mathbf{x}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{b} + \mathbf{b}^{\mathrm{T}}\mathbf{b}. \end{aligned}$$

La condición necesaria para un mínimo en $\mathbf{x}^*$ es

$$\frac{\partial}{\partial \xi_i}\Phi(\mathbf{x})\bigg|_{\mathbf{x}=\mathbf{x}^*} = 0, \quad i = 1,\ldots,n,$$

lo cual entrega el sistema lineal (las *ecucaciones normales*)

$$\mathbf{A}^{\mathrm{T}}\mathbf{Ax}^* = \mathbf{A}^{\mathrm{T}}\mathbf{b}, \tag{3.42}$$

con la solución única (si rango$(\mathbf{A}) = n$)

$$\mathbf{x}^* = (\mathbf{A}^{\mathrm{T}}\mathbf{A})^{-1}\mathbf{A}^{\mathrm{T}}\mathbf{b}.$$

El hecho de que $\mathbf{x}^*$ realmente es el mínimo se refleja en la satisfacción de la condición adicional suficiente que

$$\left(\frac{\partial^2}{\partial\xi_i\partial\xi_j}\Phi(\mathbf{x})\bigg|_{\mathbf{x}=\mathbf{x}^*}\right)_{ij}$$

es definida positiva.

Uno podria resolver (3.42) por la descomposición de Cholesky. Pero este camino no es recomendado, con la excepción del caso que las columnas de $\mathbf{A}$ son "casi ortogonales". En efecto, el método de Cholesky es mucho más sensible para errores de redondeo que la transformación de Householder.

Frecuentamente los problemas de compensación son extremadamente mal acondicionados, sobre todo cuando las funciones de planteo no son las apropiadas. Por ejemplo, para un planteo polinomial siempre se recomienda transformar la variable independiente al intervalo $[-1, 1]$ y usar los polinomios de Chebyshev

$$T_0(x) = 1, \quad T_1(x) = x, \quad T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x), \quad n \geqslant 2.$$

la última medida para la solución del problema es la descomposición en valores singulares. Si

$$\mathbf{A} = \mathbf{U}\begin{bmatrix}\mathbf{\Sigma}\\ 0\end{bmatrix}\mathbf{V}^*$$

y la matriz $\mathbf{\Sigma}$ es invertible, entonces la solución del problema (3.40) es

$$\mathbf{x}^* = \mathbf{V}\begin{bmatrix}\mathbf{\Sigma}^{-1} & 0\end{bmatrix}\mathbf{U}^*\mathbf{b}.$$

Si $\mathbf{\Sigma}$ no es invertible, la solución del problema de compensación no es única. En este caso, se usa la solución óptima

$$\mathbf{x}^* = \mathbf{A}^+\mathbf{b} = \mathbf{V}\begin{bmatrix}\mathbf{\Sigma}^+ & 0\end{bmatrix}\mathbf{U}^*\mathbf{b},$$

con la longitud euclidiana mínima. Por supuesto, en la práctica se remplaza $\mathbf{\Sigma}^+$ por $\mathbf{\Sigma}^+(\alpha) = \mathrm{diag}(\sigma_i^+(\alpha))$,

$$\sigma_i^+(\alpha) = \begin{cases}1/\sigma_i & \text{si } \sigma_i \geqslant \alpha,\\ 0 & \text{sino,}\end{cases}$$

donde el parámetro $\alpha > 0$ representa la inexactitud en $\mathbf{A}$ y en la aritmética.

Capítulo 4

Métodos iterativos para la solución de sistemas de ecuaciones lineales

4.1. Un ejemplo

Muchas aplicaciones requieren la solución de sistemas de ecuaciones lineales de extremadamente gran tamaño ($n \geqslant 10^4$), pero donde la matriz de coeficientes tiene una estructura muy especial: cada fila contiene sólo muy pocos, por ejemplo cinco, elementos diferentes de cero, en una configuración muy particular. En esta situación no se recomienda el uso de los métodos discutidos hasta ahora por razones de espacio de almacenaje y tiempo computacional. Vamos a ilustrar el problema en el siguiente ejemplo.

Ejemplo 4.1. *Se busca una función $u : [0,1]^2 \to \mathbb{R}$ que es solución del siguiente problema de valores de frontera (problema de Dirichlet) para la ecuación de Poisson:*

$$\begin{aligned} -\Delta u \equiv -u_{\xi\xi} - u_{\eta\eta} = f(\xi,\eta), \quad (\xi,\eta) \in (0,1)^2, \\ u(\xi,\eta) = 0, \quad \xi \in \{0,1\} \text{ o } \eta \in \{0,1\}. \end{aligned} \tag{4.1}$$

Aquí f es una función real y continua dada sobre $[0,1]^2$. La tarea de determinar la solución u numéricamente se reduce a un sistema de ecuaciones lineales para los valores aproximados

$$\tilde{u}_{ij} \approx u(\xi_i, \eta_j). \tag{4.2}$$

Para una función $z = z(\xi) \in C^4$ tenemos según la fórmula de Taylor

$$\begin{aligned} z(\xi + h) &= z(\xi) + z'(\xi)h + \frac{1}{2}z''(\xi)h^2 + \frac{1}{6}z'''(\xi)h^3 + \frac{1}{24}z^{(4)}(\tilde{\xi})h^4, \\ z(\xi - h) &= z(\xi) - z'(\xi)h + \frac{1}{2}z''(\xi)h^2 - \frac{1}{6}z'''(\xi)h^3 + \frac{1}{24}z^{(4)}(\tilde{\tilde{\xi}})h^4, \end{aligned}$$

donde $\tilde{\xi} = \xi + \delta_1 h$ y $\tilde{\tilde{\xi}} = \xi - \delta_2 h$. Combinando estas dos ecuaciones obtenemos

$$z''(\xi) = \frac{z(\xi + h) - 2z(\xi) + z(\xi - h)}{h^2} - \frac{1}{12}z^{(4)}(\hat{\xi})h^2. \tag{4.3}$$

Entonces, para h "pequeño", el primer término en el lado derecho de (4.3) sirve como buena aproximación de $z''(\xi)$. Ahora $[0,1]^2$ se cubre con una malla cuadrática de puntos (ξ_i, η_j), $0 \leqslant i, j \leqslant N+1$, donde

$$\xi_i = ih, \quad \eta_j = jh, \quad 0 \leqslant i, j \leqslant N+1, \quad h = \frac{1}{N+1}, \quad N \in \mathbb{N}.$$

Ahora aproximamos las segundas derivadas parciales $u_{\xi\xi}$ y $u_{\eta\eta}$ usando (4.3). Usando la ecuación de derivadas parciales, obtenemos la siguiente ecuación para (4.2):

$$4\tilde{u}_{ij} - \tilde{u}_{i-1,j} - \tilde{u}_{i+1,j} - \tilde{u}_{i,j-1} - \tilde{u}_{i,j+1} = h^2 f(\xi_i, \eta_j), \quad 1 \leqslant i, j \leqslant N. \tag{4.4}$$

Entonces, la aproximación genera un sistema de N^2 ecuaciones para N^2 desconocidas. Enumerando los pares (i,j) en el orden

$$(1,1),(2,1),\ldots,(N,1),(1,2),(2,2),\ldots,(N,2),\ldots,(N,N),$$

resulta un sistema lineal $\mathbf{Ax}=\mathbf{b}$ *con*

$$\mathbf{x}=\begin{pmatrix}\tilde{u}_{11}\\ \tilde{u}_{21}\\ \vdots\\ \tilde{u}_{NN}\end{pmatrix},\quad \mathbf{b}=h^2\begin{pmatrix}f(\xi_1,\eta_1)\\ f(\xi_2,\eta_2)\\ \vdots\\ f(\xi_N,\eta_N)\end{pmatrix},\quad \mathbf{A}=\begin{bmatrix}\mathbf{B} & -\mathbf{I} & & & \\ -\mathbf{I} & \ddots & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & -\mathbf{I}\\ & & & -\mathbf{I} & \mathbf{B}\end{bmatrix}\in\mathbb{R}^{N^2\times N^2},$$

donde

$$\mathbf{B}=\begin{bmatrix}4 & -1 & & & \\ -1 & 4 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & -1\\ & & & -1 & 4\end{bmatrix}\in\mathbb{R}^{N\times N}.$$

Si se trata aplicar la descomposición de Cholesky a $\mathbf{A}$*, aprovechando la estructura de bandas, resultan* $N^4/2$ *operaciones aritméticas y* N^3 *elementos de almacenaje; estos números ya son grandes para* $50\leqslant N\leqslant 200$*. Además, no es muy razonable tratar de resolver el sistema exactamente, dado que la solución* $\tilde{u}_{ij}$ *misma sólo representa una aproximación (con un error* $\mathcal{O}(h^2)$ *si* $|f|\leqslant 1$*) de los valores* u_{ij}*. Por supuesto, el caso análogo tri-dimensional es aún mucho mas complicado.*

Comentamos que existe un algoritmo especial precisamente para el sistema discutido en este ejemplo, el algoritmo de Buneman*, que para* $N=2^{m+1}-1$ *necesita sólo* $3N^2(m+1)$ *operaciones aritméticas y approx.* $6N^2$ *elementos de almacenaje. Sin embargo, este este algoritmo ya fracasa si los elementos varian con* (i,j)*, con una matriz que sino tiene la misma estructura que* $\mathbf{A}$*.*

Resumiendo, constatamos que se recomienda buscar métodos simplemente estructurados que aproximan la solución $\mathbf{x}$ de un sistema lineal por una sucesión $\{\mathbf{x}_k\}_{k\in\mathbb{N}}$ infinita, pero donde cada paso $\mathbf{x}_k\to\mathbf{x}_{k+1}$ requiere sólo muy poco esfuerzo computacional.

4.2. Metodología general del desarrollo de métodos iterativos

La idea básica es la siguiente. Queremos resolver el problema

$$\mathbf{Ax}^*=\mathbf{b}. \tag{4.5}$$

La matriz $\mathbf{A}$ se descompone de la siguiente forma:

$$\mathbf{A}=\mathbf{M}+\mathbf{N},$$

entonces (4.5) es equivalente a

$$\mathbf{Mx}^*=\mathbf{b}-\mathbf{Nx}^*. \tag{4.6}$$

Entonces, introduciendo un factor ω y una matriz $\mathbf{C}$ arbitraria, (4.6) es equivalente a

$$(\omega\mathbf{M}+\mathbf{C})\mathbf{x}^* = (\mathbf{C}-\omega\mathbf{N})\mathbf{x}^* + \omega\mathbf{b}. \tag{4.7}$$

Las matrices $\mathbf{C}$ y $\mathbf{M}$ y el factor ω se eligen de tal forma que $\omega\mathbf{M}+\mathbf{C}$ es regular y tiene una estructura simple, de manera que un sistema lineal con esta matriz puede ser resuelto más facilmente que el sistema original (4.5). Ahora, si remplazamos en la última ecuación $\mathbf{x}^*$ en el lado derecho por $\mathbf{x}_k$ y en el lado izquierdo por $\mathbf{x}_{k+1}$, obtenemos el método de iteración

$$(\omega\mathbf{M}+\mathbf{C})\mathbf{x}_{k+1} = (\mathbf{C}-\omega\mathbf{N})\mathbf{x}_k + \omega\mathbf{b}, \tag{4.8}$$

el cual podemos escribir como

$$\mathbf{x}_{k+1} = (\omega\mathbf{M}+\mathbf{C})^{-1}\big((\mathbf{C}-\omega\mathbf{N})\mathbf{x}_k + \omega\mathbf{b}\big) =: \mathbf{\Phi}(\mathbf{x}_k).$$

Según nuestra construcción,

$$\mathbf{x}^* = \mathbf{\Phi}(\mathbf{x}^*), \tag{4.9}$$

es decir, $\mathbf{x}^*$ es el punto fijo de la aplicación $\mathbf{x}\mapsto\mathbf{\Phi}(\mathbf{x})$; por lo tanto,

$$\mathbf{x}_{k+1}-\mathbf{x}^* = \underbrace{(\omega\mathbf{M}+\mathbf{C})^{-1}(\mathbf{C}-\omega\mathbf{N})}_{=:\mathbf{B}(\omega)}(\mathbf{x}_k-\mathbf{x}^*). \tag{4.10}$$

En virtud de (4.10), para cualquier vector inicial $\mathbf{x}_0$, se cumple

$$\mathbf{x}_k-\mathbf{x}^* = \big(\mathbf{B}(\omega)\big)^k(\mathbf{x}_0-\mathbf{x}^*).$$

Teorema 4.1. *El método* (4.8) *converge para todo* $\mathbf{x}_0\in\mathbb{R}^n$ *a* $\mathbf{x}^*$ *si y sólo si*

$$r_\sigma\big(\mathbf{B}(\omega)\big) < 1. \tag{4.11}$$

Demostración. Sea $r_\sigma(\mathbf{B}(\omega)) < 1$. Entonces $\mathbf{I}-\mathbf{B}(\omega)$ es regular, lo que significa que la ecuación

$$\mathbf{x}^* = \mathbf{B}(\omega)\mathbf{x}^* + \omega(\omega\mathbf{M}+\mathbf{C})^{-1}\mathbf{b}$$

tiene una única solución $\mathbf{x}^*$. Según el Teorema 3.5, existe una norma vectorial $\|\cdot\|$ (con una norma matricial inducida) tal que para un $\varepsilon>0$ pequeño dado,

$$\|\mathbf{B}(\omega)\| \leqslant r_\sigma\big(\mathbf{B}(\omega)\big)+\varepsilon < 1.$$

Entonces

$$\|\mathbf{x}_k-\mathbf{x}^*\| \leqslant \|\mathbf{B}(\omega)\|^k\|\mathbf{x}_0-\mathbf{x}^*\|,$$

o sea,

$$\lim_{k\to\infty}\mathbf{x}_{k+1} = \mathbf{x}^*.$$

Por otro lado, en el caso cntrario $r_\sigma(\mathbf{B}(\omega)) \geqslant 1$, existe un valor propio λ de $\mathbf{B}(\omega)$ con $|\lambda|\geqslant 1$. Sea $\mathbf{v}\neq 0$ el vector propio asociado. Entonces

$$\big(\mathbf{B}(\omega)\big)^k\mathbf{v} = \lambda^k\mathbf{v}.$$

Sea $\mathbf{x}_0 = \mathbf{x}^*+\mathbf{v}$. En este caso,

$$\mathbf{x}_k-\mathbf{x}^* = \lambda^k\mathbf{v},$$

es decir,

$$\forall k \in \mathbb{N}: \quad \|\mathbf{x}_k - \mathbf{x}^*\| = |\lambda^k| \|\mathbf{v}\| \geqslant \|\mathbf{v}\| > 0.$$

■

Para cualquier $\varepsilon > 0$ suficientemente pequeño existe una norma tal que en esta norma, la distancia $\|\mathbf{x} - \mathbf{x}^*\|$ se reduce por un factor $r_\sigma(\mathbf{B}(\omega)) + \varepsilon$ en cada paso. Obviamente, hay que elegir $\mathbf{M}$, $\mathbf{N}$, $\mathbf{C}$ y ω de la forma que $r_\sigma(\mathbf{B}(\omega))$ sea lo más pequeño posible. Antes de seguir estudiando la teoría de los métodos (específicamente, el comportamento de $\mathbf{B}(\omega)$ en un caso especial importante), vamos a mencionar los métodos mas importantes que se usan en la práctica, definiendo las matrices $\mathbf{L}$, $\mathbf{D}$ y $\mathbf{U}$ por medio de

$$\mathbf{A} = -\mathbf{L} + \mathbf{D} - \mathbf{U}, \tag{4.12}$$

donde

$$-\mathbf{L} = \begin{bmatrix} 0 & 0 & \cdots & 0 \\ \alpha_{21} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ \alpha_{n1} & \cdots & \alpha_{n,n-1} & 0 \end{bmatrix}, \ \mathbf{D} = \begin{bmatrix} \alpha_{11} & 0 & \cdots & 0 \\ 0 & \alpha_{22} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \alpha_{nn} \end{bmatrix}, \ -\mathbf{U} = \begin{bmatrix} 0 & \alpha_{12} & \cdots & \alpha_{1n} \\ 0 & 0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \alpha_{n-1,n} \\ 0 & \cdots & 0 & 0 \end{bmatrix}. \tag{4.13}$$

1. El *método de Jacobi* es definido por $\mathbf{M} = \mathbf{D}$, $\mathbf{N} = -\mathbf{L} - \mathbf{U}$, $\mathbf{C} = 0$ y $\omega = 1$. La fórmula de iteración vectorial correspondiente es

$$\mathbf{x}_{k+1} = \mathbf{D}^{-1}\big((\mathbf{L} + \mathbf{U})\mathbf{x}_k + \mathbf{b}\big), \quad k \in \mathbb{N}_0; \tag{4.14}$$

para las componentes obtenemos

$$\xi_{i,k+1} = \xi_{i,k} + \frac{1}{\alpha_{ii}} \left(-\sum_{j=1}^{n} \alpha_{ij} \xi_{j,k} + \beta_i \right), \quad i = 1, \ldots, n, \quad k \in \mathbb{N}_0. \tag{4.15}$$

2. El *método de Gauss-Seidel* es definido por $\mathbf{M} = -\mathbf{L} + \mathbf{D}$, $\mathbf{N} = -\mathbf{U}$, $\mathbf{C} = 0$ y $\omega = 1$. Las fórmulas de iteración son

$$(-\mathbf{L} + \mathbf{D})\mathbf{x}_{k+1} = \mathbf{U}\mathbf{x}_k + \mathbf{b}, \quad k \in \mathbb{N}_0; \tag{4.16}$$

$$\begin{aligned} \xi_{i,k+1} =& \xi_{i,k} + \frac{1}{\alpha_{ii}} \left(-\sum_{j=1}^{i-1} \alpha_{ij} \xi_{j,k+1} - \sum_{j=i}^{n} \alpha_{ij} \xi_{j,k} + \beta_i \right), \\ & i = 1, \ldots, n, \quad k \in \mathbb{N}_0. \end{aligned} \tag{4.17}$$

3. El *método SOR* (*successive overrelaxation*) con el *parámetro de relajación* ω corresponde a $\mathbf{M} = -\mathbf{L} + \mathbf{D}$, $\mathbf{N} = -\mathbf{U}$, $\mathbf{C} = (1 - \omega)\mathbf{D}$ y $\omega \neq 0$. Las fórmulas de iteración son

$$(-\omega\mathbf{L} + \mathbf{D})\mathbf{x}_{k+1} = \big((1 - \omega)\mathbf{D} + \omega\mathbf{U}\big)\mathbf{x}_k + \omega\mathbf{b}, \quad k \in \mathbb{N}_0; \tag{4.18}$$

$$\begin{aligned} \xi_{i,k+1} =& \xi_{i,k} + \frac{\omega}{\alpha_{ii}} \left(-\sum_{j=1}^{i-1} \alpha_{ij} \xi_{j,k+1} - \sum_{j=i}^{n} \alpha_{ij} \xi_{j,k} + \beta_i \right), \\ & i = 1, \ldots, n, \quad k \in \mathbb{N}_0. \end{aligned} \tag{4.19}$$

La identidad (4.18) ilustra la idea del método SOR: tenemos como "última" ("mejor") aproximación de $\mathbf{x}^*$ el vector $(\xi_{1,k+1},\dots,\xi_{i-1,k+1},\xi_{i,k},\dots,\xi_{n,k})^{\mathrm{T}}$. Después se calcula primero $\xi_{i,k+1}^{\mathrm{GS}}$, aplicando el método de Gauss-Seidel, luego se determina $\xi_{i,k+1}$ agrandando o achicando la corrección de Gauss-Seidel por el factor ω.

Ejemplo 4.2 (Tarea 18, Curso 2006)**.** *Se considera el sistema lineal* $\mathbf{Ax}=\mathbf{b}$ *con*

$$\mathbf{A}=\begin{bmatrix}5 & -4\\ 1 & -3\end{bmatrix},\quad \mathbf{b}=\begin{pmatrix}12\\ -2\end{pmatrix}.$$

a) *Preparar un dibujo que interpreta ambas ecuaciones del sistema lineal como lineas rectas en el plano* x_1-x_2*. La solución exacta del problema es* $x_1^*=4, x_2^*=2$*.*

b) *Ejecutar desde* $\mathbf{x}_0=(-8,-8)^{\mathrm{T}}$ *tres pasos de cada uno de los métodos de Jacobi, de Gauss-Seidel, y del método SOR con* $\omega=1{,}5$ *aus. Agregar al dibujo las sucesiones*

$$\begin{pmatrix}\xi_{1,0}\\ \xi_{2,0}\end{pmatrix},\ \begin{pmatrix}\xi_{1,1}\\ \xi_{2,0}\end{pmatrix},\ \begin{pmatrix}\xi_{1,1}\\ \xi_{2,1}\end{pmatrix},\ \begin{pmatrix}\xi_{1,2}\\ \xi_{2,1}\end{pmatrix}\dots$$

en los casos de los métodos de Gauss-Seidel y SOR y la sucesión

$$\begin{pmatrix}\xi_{1,0}\\ \xi_{2,0}\end{pmatrix},\ \begin{pmatrix}\xi_{1,1}\\ \xi_{2,1}\end{pmatrix},\ \begin{pmatrix}\xi_{1,2}\\ \xi_{2,2}\end{pmatrix},\ \begin{pmatrix}\xi_{1,3}\\ \xi_{2,3}\end{pmatrix}\dots$$

en el caso del método de Jacobi.

Solucion sugerida.

a) *La Figura 4.1 muestra las dos rectas y las iteradas.*

b) *Sean* $(\xi_{1,k},\xi_{2,k})$ *las iteradas del método de Jacobi,* $(\tilde{\xi}_{1,k},\tilde{\xi}_{2,k})$ *las del método de Gauß-Seidel y* $(\bar{\xi}_{1,k},\bar{\xi}_{2,k})$ *las del método SOR, entonces obtenemos las siguientes sucesiones de iteración:*

$$\begin{aligned}
(\xi_{1,0},\xi_{2,0}) &= (-8,-8),\\
(\xi_{1,1},\xi_{2,1}) &= (-4,-2),\\
(\xi_{1,2},\xi_{2,2}) &= \Big(\frac{4}{5},-\frac{2}{3}\Big),\\
(\xi_{1,3},\xi_{2,3}) &= (1{,}8\bar{6},0{,}9\bar{3});\\
(\tilde{\xi}_{1,0},\tilde{\xi}_{2,0}) &= (-8,-8),\\
(\tilde{\xi}_{1,1},\tilde{\xi}_{2,0}) &= (-4,-8),\\
(\tilde{\xi}_{1,1},\tilde{\xi}_{2,1}) &= \Big(-4,-\frac{2}{3}\Big),\\
(\tilde{\xi}_{1,2},\tilde{\xi}_{2,1}) &= (1{,}8\bar{6},-0.\bar{6}),\\
(\tilde{\xi}_{1,2},\tilde{\xi}_{2,2}) &= (1{,}8\bar{6},1{,}2\bar{8}),\\
(\tilde{\xi}_{1,3},\tilde{\xi}_{2,2}) &= (3{,}43\bar{1},1{,}2\bar{8}),\\
(\tilde{\xi}_{1,3},\tilde{\xi}_{2,3}) &= (3{,}43\bar{1},1{,}81037);\\
(\bar{\xi}_{1,0},\bar{\xi}_{2,0}) &= (-8,-8),
\end{aligned}$$

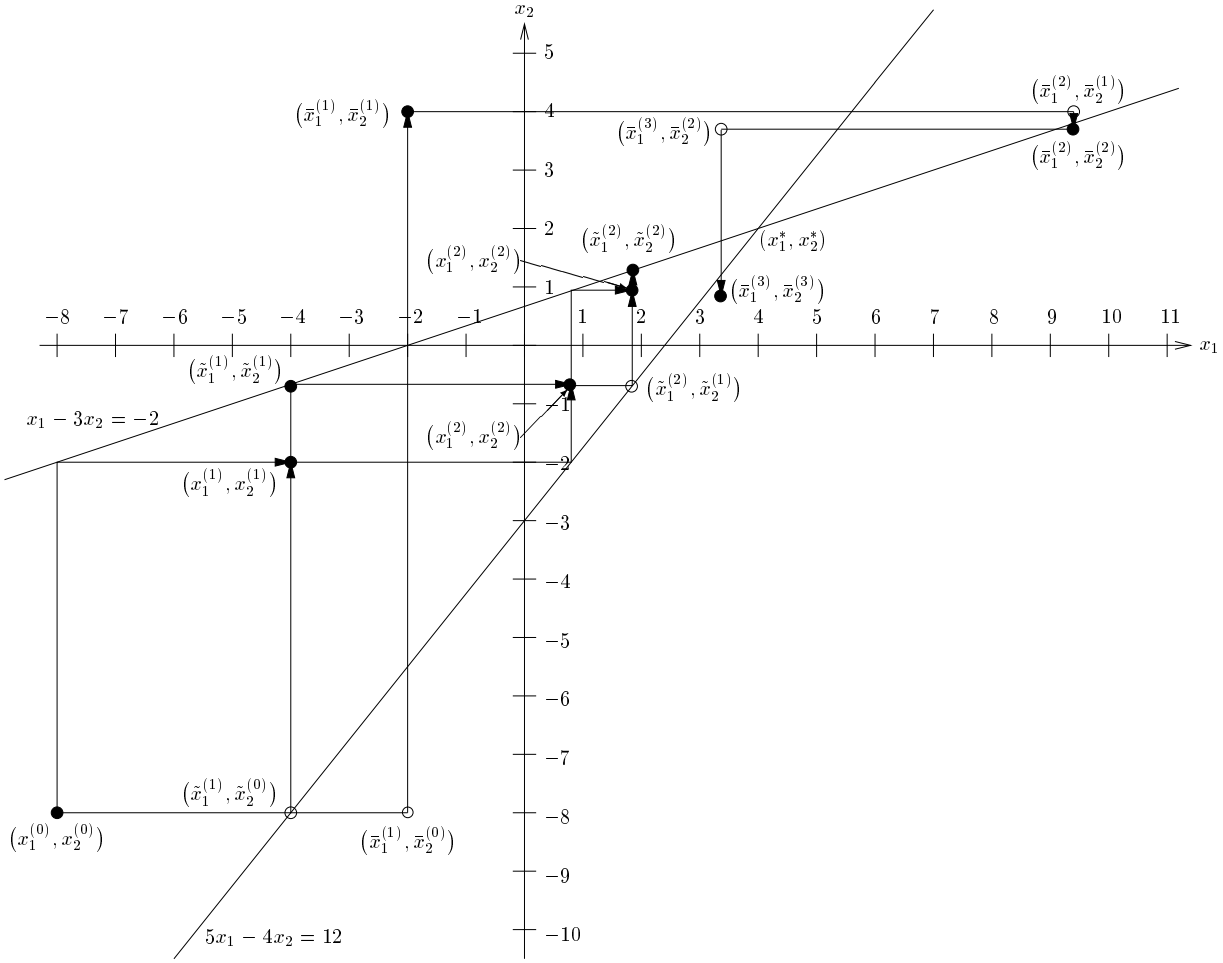


FIGURA 4.1. Interpretación de las ecuaciones de un sistema lineal 2×2 como lineas rectas en el plano x_1-x_2 y sucesiones de soluciones aproximadas

$$
\begin{aligned}
(\bar{\xi}_{1,1}, \bar{\xi}_{2,0}) &= (-2, -8),\\
(\bar{\xi}_{1,1}, \bar{\xi}_{2,1}) &= (-2, 4),\\
(\bar{\xi}_{1,2}, \bar{\xi}_{2,1}) &= (9{,}4, 4),\\
(\bar{\xi}_{1,2}, \bar{\xi}_{2,2}) &= (9{,}4, 3{,}7),\\
(\bar{\xi}_{1,3}, \bar{\xi}_{2,2}) &= (3{,}34, 3{,}7),\\
(\bar{\xi}_{1,3}, \bar{\xi}_{2,3}) &= (3{,}34, 0{,}82).
\end{aligned}
$$

4.3. Teoremas de convergencia para métodos iterativos

Teorema 4.2. *El método de Jacobi* (4.14) *converge si* $\mathbf{A} = \mathbf{D} - \mathbf{L} - \mathbf{U}$ *satisface uno de los siguientes criterios:* $\mathbf{A}$ *es* estrictamente diagonal dominante por filas*:*

$$\forall i = 1, \ldots, n: \quad |\alpha_{ii}| > \sum_{\substack{j=1 \\ j \neq i}}^{n} |\alpha_{ij}|, \tag{4.20}$$

o $\mathbf{A}$ *es* estrictamente diagonal dominante por columnas*:*

$$\forall i = 1, \ldots, n: \quad |\alpha_{ii}| > \sum_{\substack{j=1 \\ j \neq i}}^{n} |\alpha_{ji}|. \tag{4.21}$$

Demostración. Tarea. ■

Los requerimientos (4.20) y (4.21) son bastante restrictivos y no se cumplen para la mayoría de las matrices. En particular, muchas veces la desigualdad estricta en (4.20) o (4.21) no se cumple en todas las filas o columnas, sino que sólo en algunas de ellas, mientras que en las otras $|\alpha_{ii}|$ es igual al lado derecho de (4.20) o (4.21). A continuación veremos que también en este caso podemos asegurar la convergencia del método (4.14), siempre cuando la matriz $\mathbf{A}$ cumpla la propiedad de *irreducibilidad*.

Definición 4.1. *Una matriz* $\mathbf{A} \in \mathbb{C}^{n\times n}$ *se llama* irreducible *si no existe ninguna matriz de permutación* $\mathbf{P}$ *tal que*

$$\mathbf{P}^{\mathrm{T}}\mathbf{A}\mathbf{P} = \begin{bmatrix} \tilde{\mathbf{A}}_{11} & \tilde{\mathbf{A}}_{12} \\ 0 & \tilde{\mathbf{A}}_{22} \end{bmatrix}, \quad \tilde{\mathbf{A}}_{22} \in \mathbb{C}^{k\times k}, \quad k < n.$$

Para decidir sobre la irreducibilidad de una matriz $\mathbf{A}$, necesitamos el concepto del *grafo dirigido* de la matriz.

Definición 4.2. *Sea* $\mathbf{A} = (\alpha_{ij}) \in \mathbb{C}^{n\times n}$*. A la siguiente construcción se refiere como* grafo dirigido $\mathcal{G}(\mathbf{A})$ *de* $\mathbf{A}$*:*

1. $\mathcal{G}(\mathbf{A})$ *incluye* n *vértices* $P_1, \ldots, P_n$*.*
2. *Un arco dirigido* $P_i \mapsto P_j$ *junta* P_i *con* P_j*,* $i \neq j$*, si y sólo si* $\alpha_{ij} \neq 0$*.*
3. *Los* caminos dirigidos *en* $\mathcal{G}(\mathbf{A})$ *son compuestos por arcos dirigidos.*
4. *El grafo dirigido* $\mathcal{G}(\mathbf{A})$ *se llama* conexo *si para cada par* (P_i, P_j) *de vértices,* $1 \leqslant i, j \leqslant n$*,* $i \neq j$*, existe un camino dirigido de* P_i *a* P_j*.*

Ejemplo 4.3. *La Figura 4.2 muestra algunas matrices y sus grafos dirigidos. Nos damos cuenta que los grafos* $\mathcal{G}(\mathbf{A})$*,* $\mathcal{G}(\mathbf{B})$ *y* $\mathcal{G}(\mathbf{C})$ *son conexos; obviamente,* $\mathcal{G}(\mathbf{D})$ *no es conexo.*

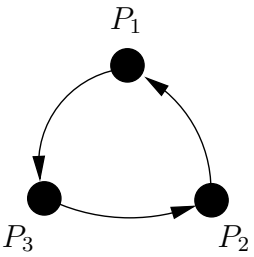
Teorema 4.3. *Una matriz* $\mathbf{A} \in \mathbb{K}^{n\times n}$ *es irreducible si y sólo si su grafo dirigido* $\mathcal{G}(\mathbf{A})$ *es conexo.*

Demostración. Hay que demostrar dos implicaciones.

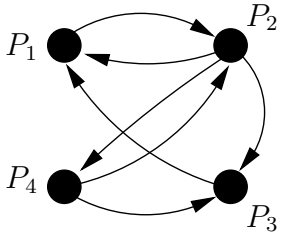
"⇒": Supongamos primero que $\mathcal{G}(\mathbf{A})$ es conexo, y sean los índices i y j, $i \neq j$, $1 \leqslant i, j \leqslant n$ arbitrarios. Entonces existen números $i_1, \ldots, i_m$ tales que

$$\alpha_{i,i_1} \cdot \alpha_{i_1,i_2} \cdot \alpha_{i_2,i_3} \cdot \ldots \cdot \alpha_{i_m,j} \neq 0. \tag{4.22}$$

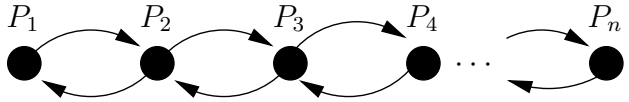
$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 3 \\ 1 & 2 & 0 \\ 0 & 4 & 5 \end{bmatrix} \qquad \mathcal{G}(\mathbf{A}):$$



$$\mathbf{B} = \begin{bmatrix} 2 & 1 & 0 & 0 \\ 1 & 3 & 1 & 1 \\ 1 & 0 & -1 & 0 \\ 0 & 1 & 1 & 2 \end{bmatrix} \qquad \mathcal{G}(\mathbf{B}):$$



$$\mathbf{C} = \begin{bmatrix} 2 & 1 & 0 & \cdots & 0 \\ 1 & 2 & 1 & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & 1 & 2 & 1 \\ 0 & \cdots & 0 & 1 & 2 \end{bmatrix} \qquad \mathcal{G}(\mathbf{C}):$$



$$\mathbf{D} = \begin{bmatrix} 4 & 1 & 0 & 1 & 0 \\ 1 & 2 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ 1 & 2 & 0 & 3 & 0 \\ 0 & 0 & 1 & 0 & 1 \end{bmatrix} \qquad \mathcal{G}(\mathbf{D}):$$

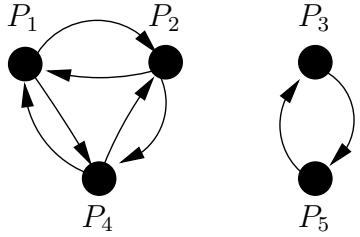


FIGURA 4.2. Algunas matrices y sus grafos dirigidos.

Ahora supongamos que $\mathbf{A}$ es reducible, es decir

$$\alpha_{s,k} = 0, \quad s \in S, \quad k \in K, \quad S \cap K = \varnothing, \quad S \cup K = \{1, \ldots, n\}. \tag{4.23}$$

Entonces sean $i \in S$ y $j \in K$. Dado que $\alpha_{i,i_1} \neq 0$, tenemos que $i_1 \notin K$, o sea, $i_1 \in S$, lo que implica $i_2 \in S$, etc., entonces sería imposible construir la cadena (4.22), una contradicción.

"$\Leftarrow$": Ahora sea $\mathbf{A}$ irreducible e $i \in \{1, \ldots, n\}$ arbitrario. Definimos

$$\mathcal{I} := \big\{k \mid \exists\{i_1, \ldots, i_m\} : \ \alpha_{i,i_1} \cdot \ldots \cdot \alpha_{i_m,k} \neq 0\big\},$$

notando que $\mathcal{I} \neq \varnothing$, puesto que sino $\alpha_{i1} = \ldots = \alpha_{in} = 0$, una contradicción. Supongamos que $\mathcal{I} \neq \{1, \ldots n\}$, y sea $l \in \{1, \ldots n\} \backslash \mathcal{I}$. Demostramos que en este caso $\alpha_{jl} = 0$ para $j \in \mathcal{I}$. (Esto sería una contradicción a la irreducibilidad de $\mathbf{A}$.) Si existiera

$\alpha_{j_0,l} \neq 0$ para un índice $j_0 \in \mathcal{I}$, existirían también índices $\{i_1, \ldots, i_m\}$ tales que

$$\alpha_{i,i_1} \cdot \alpha_{i_1,i_2} \cdot \ldots \cdot \alpha_{i_m,j_0} \neq 0, \quad \alpha_{j_o,i} \neq 0,$$

o sea $l \in \mathcal{I}$, una contradicción. Esto implica que $\mathcal{I} = \{1, \ldots, n\}$, y dado que i es arbitrario, concluimos que $\mathcal{G}(\mathbf{A})$ es conexo.

■

Definición 4.3. *Una matriz* $\mathbf{A} \in \mathbb{C}^{n\times n}$ *se llama* irreduciblemente diagonal dominante *si* $\mathbf{A}$ *es irreducible y*

$$\forall i \in \{1, \ldots, n\} : \ |\alpha_{ii}| \geqslant \sum_{\substack{j=1 \\ j\neq i}}^{n} |\alpha_{ij}| \wedge \exists i_0 \in \{1, \ldots, n\} : \ |\alpha_{i_0 i_0}| > \sum_{\substack{j=1 \\ j\neq i_0}}^{n} |\alpha_{i_0 j}|. \tag{4.24}$$

Teorema 4.4. *Sea* $\mathbf{A}$ *estrictamente o irreduciblemente diagonal dominante. Entonces* $\mathbf{A}$ *es regular y el método de Jacobi converge.*

Demostración. La demostración procede en tres pasos:

1. Demostramos que el método está bien definido.
2. Demostramos que
$$r_\sigma\big(\mathbf{D}^{-1}(\mathbf{L}+\mathbf{U})\big) < 1. \tag{4.25}$$
3. El resultado de 2.) implica que el método de Jacobi converge para $\mathbf{b}$ arbitrario a una solución de $\mathbf{Ax} = \mathbf{b}$, lo que implica la regularidad de $\mathbf{A}$.

Enseguida procedemos a la demostración de 1.) y 2.):

1. Si $\mathbf{A}$ es estrictamente diagonal dominante, esta propiedad es obvia. Si $\mathbf{A}$ es irreduciblemente diagonal dominante, entonces $\alpha_{ii} \neq 0$ para todo i (si existiría un índice i_0 tal que $\alpha_{i_0 i_0} = 0$, tendríamos que $\alpha_{i_0 1} = \ldots = \alpha_{i_0 n} = 0$, una contradicción a la irreducibilidad).
2. Si $\mathbf{A}$ es estrictamente diagonal dominante, (4.25) es una consecuencia del Teorema 4.1. En el otro caso, podemos aplicar el Teorema 3.4, aplicado a
$$\mathbf{B} := \mathbf{D}^{-1}(\mathbf{L}+\mathbf{U}) =: (\beta_{ij}),$$
para concluir que
$$r_\sigma(\mathbf{B}) \leqslant \|\mathbf{B}\|_\infty \leqslant 1.$$
Supongamos que $r_\sigma(\mathbf{B}) = 1$. En este caso existe $\lambda \in \mathbb{C}$, $|\lambda| = 1$, tal que $(\mathbf{B} - \lambda\mathbf{I})\mathbf{x} = 0$ con $\mathbf{x} \neq 0$. Por otro lado, $\mathbf{B} - \lambda\mathbf{I}$ es una matriz irreduciblemente diagonal dominante, dado que difiere de $\mathbf{D}^{-1}\mathbf{A}$ sólo en sus elementos diagonales, donde "1" es remplazado por $-\lambda$ con $|-\lambda| = 1$. Supongamos ahora que $\mathbf{x} = (\xi_1, \ldots, \xi_n)^{\mathrm{T}}$ con $|\xi_1| = \ldots |\xi_n| = \xi$. Esto significa que
$$\forall i \in \{1, \ldots, n\} : \quad \left|\sum_{j=1}^{n} \beta_{ij}\xi_j\right| = \xi \leqslant \xi \sum_{j=1}^{n} |\beta_{ij}|,$$

o sea,

$$\forall i \in \{1, \ldots, n\} : \quad \sum_{j=1}^{n} |\beta_{ij}| \geqslant 1,$$

en contradicción a la supuesta diagonaldominancia irreducible de $\mathbf{A}$. Entonces

$$\mathcal{I} := \left\{ j \mid \forall i : \ |\xi_j| \geqslant |\xi_i|, \ \exists i_0 : \ |\xi_j| > |\xi_{i_0}| \right\} \neq \varnothing.$$

Sea $j \in \mathcal{I}$. Observando que $\beta_{ii} \neq 0$, tenemos que

$$0 = \sum_{i=1}^{n} \beta_{ji}\xi_i - \lambda\xi_j \Longleftrightarrow \lambda\xi_j = \sum_{i=1}^{n} \beta_{ji}\xi_i$$

$$\Longrightarrow |\xi_j| \leqslant \sum_{i=1}^{n} |\beta_{ji}||\xi_i| \Longrightarrow \sum_{i=1}^{n} |\beta_{ji}| \frac{|\xi_i|}{|\xi_j|} \geqslant 1.$$

Notamos que $|\xi_i|/|\xi_j| \leqslant 1$ y $|\xi_i|/|\xi_j| = 1$ para $i \in \mathcal{I}$. Entonces

$$1 \leqslant \sum_{i \in \mathcal{I}} |\beta_{ji}| + \sum_{i \notin \mathcal{I}} |\beta_{ji}| \frac{|\xi_i|}{|\xi_j|} < \sum_{i=1}^{n} |\beta_{ji}| \leqslant 1.$$

Esto es una contradicción en el caso que

$$\sum_{i \notin \mathcal{I}} |\beta_{ji}| \neq 0.$$

Concluimos que para $j \in \mathcal{I}$, tenemos que

$$\sum_{i \notin \mathcal{I}} |\beta_{ji}| = 0,$$

o sea, usando que $\beta_{ji} = \alpha_{ji}/\alpha_{jj}$,

$$\sum_{i \notin \mathcal{I}} |\alpha_{ji}| = 0 \quad \text{para } j \in \mathcal{I},$$

lo que implica que $\mathbf{A}$ es reducible, una contradicción.

■

Uno podría pensar que según nuestra construcción, el método de Gauss-Seidel siempre converge "mejor" que el de Jacobi. Pero eso no es válido en general.

Teorema 4.5 (Stein-Rosenberg)**.** *Sea* $\mathbf{A} \in \mathbb{R}^{n \times n}$, $\alpha_{ij} \leqslant 0$ *para* $i \neq j$ *y* $\alpha_{ii} \neq 0$ *para* $i = 1, \ldots, n$, $\mathbf{J} := \mathbf{D}^{-1}(\mathbf{L} + \mathbf{U})$ *y* $\mathbf{H} := (-\mathbf{L} + \mathbf{D})^{-1}\mathbf{U}$. *En este caso, exactamente una de las siguientes alternativas es válida:*

1. $r_\sigma(\mathbf{H}) = r_\sigma(\mathbf{J}) = 0$,
2. $r_\sigma(\mathbf{H}) = r_\sigma(\mathbf{J}) = 1$,
3. $0 < r_\sigma(\mathbf{H}) < r_\sigma(\mathbf{J}) < 1$,
4. $r_\sigma(\mathbf{H}) > r_\sigma(\mathbf{J}) > 1$.

Demostración. Ver Varga, *Matrix Iterative Analysis.* ■

El Teorema 4.5 dice que para matrices del tipo indicado, el método de Gauss-Seidel converge mejor si alguno de los métodos converge. Eso implica en particular que ambos métodos convergen para el sistema del Ejemplo 4.1.

En la práctica, las siguientes matrices son importantes: matrices simétricas definidas positivas y M-matrices.

Definición 4.4. *Una matriz* $\mathbf{A} \in \mathbb{R}^{n\times n}$ *se llama* M-matriz *si* $\alpha_{ij} \leqslant 0$ *para* $i \neq j$ *y* $\mathbf{A}^{-1}$ *existe y* $\mathbf{A}^{-1} \geqslant 0$.

Teorema 4.6. *Una matriz* $\mathbf{A} \in \mathbb{R}^{n\times n}$ *es una M-matriz si y sólo si* $\alpha_{ii} > 0$ *para* $i = 1, \ldots, n$, $\alpha_{ij} \leqslant 0$ *para* $i \neq j$ *y* $r_\sigma(\mathbf{D}^{-1}(\mathbf{L}+\mathbf{U})) < 1$, *donde* $\mathbf{A} = -\mathbf{L} + \mathbf{D} - \mathbf{U}$ *es la descomposición* (4.12), (4.13) *de* $\mathbf{A}$.

Demostración. Hay que demostrar dos implicaciones.

"⇐": Sea $\alpha_{ii} > 0$ para $i = 1, \ldots, n$, $\alpha_{ij} \leqslant 0$ para $i \neq j$ y $r_\sigma(\mathbf{D}^{-1}(\mathbf{L}+\mathbf{U})) < 1$. Definimos $\mathbf{J} := \mathbf{D}^{-1}(\mathbf{L}+\mathbf{U})$. Entonces $r_\sigma(-\mathbf{J}) < 1$, y la matriz

$$(\mathbf{I}-\mathbf{J})^{-1} = \sum_{k=0}^{\infty} \mathbf{J}^k$$

existe, y $(\mathbf{I}-\mathbf{J})^{-1} \geqslant 0$ dado que $\mathbf{J} \geqslant 0$. Pero por otro lado, sabemos que

$$\mathbf{I}-\mathbf{J} = \mathbf{D}^{-1}\mathbf{A},$$

es decir,

$$(\mathbf{I}-\mathbf{J})^{-1} = \mathbf{A}^{-1}\mathbf{D},$$

lo que implica que $\mathbf{A}^{-1}$ existe y $\mathbf{A}^{-1} \geqslant 0$.

"⇒": Supongamos que $\mathbf{A}$ es una M-matriz. Sea $\alpha_{ii} \leqslant 0$ para algún i. En este caso, usando que $\alpha_{ij} \leqslant 0$ para $i \neq j$, tenemos $\mathbf{A}\mathbf{e}_i \leqslant 0$, es decir, multiplicando con la matriz $\mathbf{A}^{-1} \geqslant 0$, $\mathbf{A}^{-1}\mathbf{A}\mathbf{e}_i \leqslant 0$, o sea $\mathbf{e}_i \leqslant 0$, una contradicción. Entonces, $\alpha_{ii} > 0$ para todo i, lo que implica que $\mathbf{J} := \mathbf{D}^{-1}(\mathbf{L}+\mathbf{U})$ es bien definida, $\mathbf{J} \geqslant 0$ y $\mathbf{A}^{-1}\mathbf{D} = (\mathbf{I}-\mathbf{J})^{-1}$ existe. Sea λ un valor propio de $\mathbf{J}$ con vector propio $\mathbf{x} \neq 0$. En este caso,

$$|\lambda||\mathbf{x}| \leqslant \mathbf{J}|\mathbf{x}| \Longrightarrow (\mathbf{I}-\mathbf{J})|\mathbf{x}| \leqslant \big(1-|\lambda|\big)|\mathbf{x}|,$$

y dado que $(\mathbf{I}-\mathbf{J})^{-1} \geqslant 0$,

$$|\mathbf{x}| \leqslant \big(1-|\lambda|\big)(\mathbf{I}-\mathbf{J})^{-1}|\mathbf{x}|.$$

Dado que $|\mathbf{x}| \neq 0$, $(\mathbf{I}-\mathbf{J})^{-1}|\mathbf{x}| \neq 0$ y $(\mathbf{I}-\mathbf{J})^{-1}|\mathbf{x}| \geqslant 0$. Entonces, podemos concluir que $|\lambda| < 1$, o sea $r_\sigma(\mathbf{J}) < 1$.

■

Entonces el método de Jacobi converge para cada M-matriz, y usando el Teorema 4.5, podemos concluir que converge también el método de Gauss-Seidel.

Definición 4.5. *Sea* $\mathbf{A} \in \mathbb{R}^{n\times n}$. *Llamamos a* $\mathbf{A} = \mathbf{N} - \mathbf{P}$ *una* partición regular *de* $\mathbf{A}$ *si* $\mathbf{N} \in \mathbb{R}^{n\times n}$ *es regular y* $\mathbf{N}^{-1} \geqslant 0$, $\mathbf{P} \geqslant 0$.

Teorema 4.7. *Sea* $\mathbf{A} \in \mathbb{R}^{n\times n}$*, con la partición regular* $\mathbf{A} = \mathbf{N}-\mathbf{P}$*. En este caso,* $r_\sigma(\mathbf{N}^{-1}\mathbf{P}) < 1$ *si y sólo si* $\mathbf{A}^{-1} \geqslant 0$*.*

Demostración.

"⇒": Trivialmente tenemos que $\mathbf{H} := \mathbf{N}^{-1}\mathbf{P} \geqslant 0$. Ahora, si $r_\sigma(\mathbf{H}) < 1$, entonces $r_\sigma(-\mathbf{H}) < 1$, o sea $(\mathbf{I}-\mathbf{H})^{-1}$ existe y $(\mathbf{I}-\mathbf{H})^{-1} \geqslant 0$, por lo tanto

$$\mathbf{A}^{-1} = (\mathbf{I}-\mathbf{H})^{-1}\mathbf{N}^{-1} \geqslant 0.$$

"⇐": Sea $\mathbf{A}^{-1} \geqslant 0$. Sabemos que

$$\mathbf{A}^{-1} = (\mathbf{N}-\mathbf{P})^{-1} = (\mathbf{I}-\mathbf{N}^{-1}\mathbf{P})^{-1}\mathbf{N}^{-1}.$$

Ahora, con $\mathbf{H} := \mathbf{N}^{-1}\mathbf{P} \geqslant 0$,

$$\begin{aligned} 0 &\leqslant (\mathbf{I}+\mathbf{H}+\mathbf{H}^2+\ldots+\mathbf{H}^m)\mathbf{N}^{-1} \\ &= (\mathbf{I}-\mathbf{H}^{m+1})(\mathbf{I}-\mathbf{H})^{-1}\mathbf{N}^{-1} \\ &= (\mathbf{I}-\mathbf{H}^{m+1})\mathbf{A}^{-1} \leqslant \mathbf{A}^{-1}, \end{aligned}$$

lo que implica que $\mathbf{I}+\mathbf{H}+\mathbf{H}^2+\ldots+\mathbf{H}^m$ converge cuando $m \to \infty$, por lo tanto, $r_\sigma(\mathbf{H}) < 1$.

■

Teorema 4.8. *Sea* $\mathbf{A}$ *estrictamente o irreduciblemente diagonal dominante con* $\alpha_{ii} > 0$ *para* $i = 1,\ldots,n$ *y* $\alpha_{ij} \leqslant 0$ *para* $i \neq j$ *(es decir,* $\mathbf{A}$ *es una* L-matriz*). En este caso,* $\mathbf{A}$ *es una M-matriz.*

Demostración. En este caso, $\mathbf{N} = \mathbf{D}$ y $\mathbf{P} = \mathbf{L}+\mathbf{U}$ es una partición regular. Segun el Teorema 4.4, $r_\sigma(\mathbf{D}^{-1}(\mathbf{L}+\mathbf{U})) < 1$. Luego aplicamos el Teorema 4.7. ■

Ejemplo 4.4 (Certamen 1, Curso 2010)**.** *Se consideran las matrices*

$$\mathbf{A}_1 = \begin{bmatrix} 10 & 2 & -6 & 0 \\ 3 & 5 & 1 & 1 \\ 2 & -2 & 12 & 4 \\ 1 & 0 & -2 & 3 \end{bmatrix}, \quad \mathbf{A}_2 = \begin{bmatrix} 4 & 1 & 1 & 0 \\ 1 & 3 & -1 & 1 \\ 0 & 2 & 4 & -2 \\ 1 & 1 & -2 & 4 \end{bmatrix}, \quad \mathbf{A}_3 = \begin{bmatrix} 2 & 0 & -1 & 0 \\ 0 & 2 & 0 & 1 \\ -3 & 0 & 5 & 0 \\ 0 & 2 & 0 & -4 \end{bmatrix}.$$

a) *Demostrar que para cada una de ellas que el método de Jacobi converge a la solución de* $\mathbf{A}_i\mathbf{x}_i = \mathbf{b}_i$ *para* $\mathbf{b}_i \in \mathbb{R}^4$ *y vectores iniciales* $\mathbf{x}_{i,0} \in \mathbb{R}^4$ *arbitrarios.*

b) *Utilizando el vector inicial* $\mathbf{x}_{i,0} = (1,1,1,1)^{\mathrm{T}}$*, calcular para* $i = 2$ *e* $i = 3$ *una nueva aproximación de la solución de* $\mathbf{A}_i\mathbf{x}_i = \mathbf{b}_i$ *para*

$$\mathbf{b}_2 = \begin{pmatrix} 9 \\ 8 \\ 8 \\ 13 \end{pmatrix}, \quad \mathbf{b}_3 = \begin{pmatrix} -2 \\ -1 \\ 24 \\ -6 \end{pmatrix},$$

utilizando los métodos de Jacobi y de Gauss-Seidel.

Solución sugerida.

a) *Se demuestra facilmente que* $\mathbf{A}_1$ *y* $\mathbf{A}_2$ *son irreduciblemente diagonal dominantes. Por otro lado, la matriz* $\mathbf{A}_3$ *corresponde a dos sistemas lineales desacoplados para* (x_1, x_3) *y* (x_2, x_4)*, respectivamente, con las respectivas matrices*

$$\mathbf{A}_{3,1} = \begin{bmatrix} 2 & -1 \\ -3 & 5 \end{bmatrix} \quad y \quad \mathbf{A}_{3,2} = \begin{bmatrix} 2 & 1 \\ 3 & -4 \end{bmatrix},$$

ambas de las cuales son estrictamente diagonaldominantes; por lo tanto se puede concluir ambos métodos convergen también en el caso de $\mathbf{A}_3$*.*

b) *Utilizando el método de Jacobi para* $\mathbf{A}\mathbf{x}_2 = \mathbf{b}_2$ *se genera la sucesión de vectores*

$$\mathbf{x}_{2,1} = \begin{pmatrix} 1{,}7500 \\ 2{,}3333 \\ 2{,}0000 \\ 3{,}2500 \end{pmatrix}, \ \mathbf{x}_{2,2} = \begin{pmatrix} 1{,}1667 \\ 1{,}6667 \\ 2{,}4583 \\ 3{,}2292 \end{pmatrix}, \ \mathbf{x}_{2,3} = \begin{pmatrix} 1{,}2188 \\ 2{,}0208 \\ 2{,}7812 \\ 3{,}7708 \end{pmatrix}, \ \mathbf{x}_{2,4} = \begin{pmatrix} 1{,}0495 \\ 1{,}9306 \\ 2{,}8750 \\ 3{,}8307 \end{pmatrix}$$

etc., mientras que el método de Gauss-Seidel entrega

$$\mathbf{x}_{2,1} = \begin{pmatrix} 1{,}7500 \\ 2{,}0833 \\ 1{,}4583 \\ 3{,}0208 \end{pmatrix}, \ \mathbf{x}_{2,2} = \begin{pmatrix} 1{,}3646 \\ 1{,}6910 \\ 2{,}6649 \\ 3{,}8186 \end{pmatrix}, \ \mathbf{x}_{2,3} = \begin{pmatrix} 1{,}1610 \\ 1{,}8951 \\ 2{,}9617 \\ 3{,}9668 \end{pmatrix}, \ \mathbf{x}_{2,4} = \begin{pmatrix} 1{,}0358 \\ 1{,}9864 \\ 2{,}9902 \\ 3{,}9896 \end{pmatrix}$$

etc. Para el sistema $\mathbf{A}\mathbf{x}_3 = \mathbf{b}_3$ *obtenemos las respectivas sucesiones*

$$\mathbf{x}_{3,1} = \begin{pmatrix} -0{,}5 \\ -1 \\ 5{,}4 \\ -1 \end{pmatrix}, \ \mathbf{x}_{3,2} = \begin{pmatrix} 1{,}7 \\ 0 \\ 4{,}5 \\ -2 \end{pmatrix}, \ \mathbf{x}_{3,3} = \begin{pmatrix} 1{,}25 \\ 0{,}5 \\ 5{,}82 \\ -1{,}5 \end{pmatrix}, \ \mathbf{x}_{3,4} = \begin{pmatrix} 1{,}91 \\ 0{,}25 \\ 5{,}55 \\ -1{,}25 \end{pmatrix}$$

etc. para el método de Jacobi y

$$\mathbf{x}_{3,1} = \begin{pmatrix} -0{,}5 \\ -1 \\ 4{,}5 \\ -2 \end{pmatrix}, \ \mathbf{x}_{3,2} = \begin{pmatrix} 1{,}25 \\ 0{,}5 \\ 5{,}55 \\ -1{,}25 \end{pmatrix}, \ \mathbf{x}_{3,3} = \begin{pmatrix} 1{,}775 \\ 0{,}125 \\ 5{,}865 \\ -1{,}4375 \end{pmatrix}, \ \mathbf{x}_{3,4} = \begin{pmatrix} 1{,}9325 \\ 0{,}2188 \\ 5{,}9595 \\ -1{,}3906 \end{pmatrix}$$

etc. para el método de Gauss-Seidel.

Ejemplo 4.5 (Tarea 19, Curso 2006)**.** *Analizar si las siguientes matrices poseen algunas de las siguientes propiedades: irreducible, irreduciblemente diagonal dominante, estrictamente diagonal dominante, L-matriz, M-matriz:*

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} 2 & -1 & 0 & 0 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 2 & -1 \\ 0 & 0 & -1 & 2 \end{bmatrix}, \quad \mathbf{C} = \begin{bmatrix} 2 & -2 & 0 & 0 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 2 & -1 \\ 0 & 0 & -2 & 2 \end{bmatrix}.$$

Solución sugerida.

a) *La matriz* $\mathbf{A}$ *es irreducible, pero no posee ninguna de las otras propiedades.*

b) *La matriz* $\mathbf{B}$ *igualmente es irreducible. Dado que es diagonal dominante y estrictamente diagonal dominante en por lo menos una fila, la estructura de signos implica que* $\mathbf{B}$ *es una L-matriz irreduciblemente diagonal dominante. Esto es una condición suficiente para asegurar que es una M-matriz.*

c) *La matriz* $\mathbf{C}$ *es una L-matriz irreducible, pero no es una L-matriz irreduciblemente diagonal dominante. Como* $\mathbf{C}$ *es singular, no puede ser M-matriz.*

Ya sabemos que el método de Jacobi converge si $\mathbf{A}$ es una M-matriz. Incluso, tenemos el siguiente teorema.

Teorema 4.9. *Si* $\mathbf{A}$ *es una M-matriz, entonces el método SOR converge para* $0 < \omega \leqslant 1$.

Demostración. La matriz $\mathbf{A}$ puede ser escrita como

$$\begin{aligned}\mathbf{A} =& \frac{1}{\omega}\big(\mathbf{D} - \omega\mathbf{L} - (1-\omega)\mathbf{D} - \omega\mathbf{U}\big)\\ =& \mathbf{N} - \mathbf{P}, \quad \mathbf{N} := \frac{1}{\omega}(\mathbf{D} - \omega\mathbf{L}), \quad \mathbf{P} := \frac{1}{\omega}\big((1-\omega)\mathbf{D} + \omega\mathbf{U}\big).\end{aligned} \tag{4.26}$$

Demostramos ahora que (4.26) es una partición regular. Para tal efecto, demostramos que $(\mathbf{D}-\omega\mathbf{L})^{-1} \geqslant 0$; el resto es una consecuencia del Teorema 4.8. Pero, en virtud de $\omega\mathbf{D}^{-1}\mathbf{L} \geqslant 0$, podemos escribir:

$$(\mathbf{D} - \omega\mathbf{L})^{-1} = (\mathbf{I} - \omega\mathbf{D}^{-1}\mathbf{L})^{-1}\mathbf{D}^{-1} = \lim_{n\to\infty}\left(\sum_{k=0}^{n}(\omega\mathbf{D}^{-1}\mathbf{L})^k\right)\mathbf{D}^{-1} \geqslant 0.$$

■

El Teorema 4.9 no es muy interesante para las aplicaciones por que se puede mostrar que para una M-matriz, la función

$$\omega \mapsto r_\sigma\big(\mathbf{B}(\omega)\big) = r_\sigma\Big((\mathbf{D} - \omega\mathbf{L})^{-1}\big((1-\omega)\mathbf{D} + \omega\mathbf{U}\big)\Big)$$

es estrictamente decreciente para $0 \leqslant \omega < \tilde{\omega}$ con $\tilde{\omega} > 1$. Lo que es interesante es el problema de la convergencia del método SOR para $\omega > 1$, y el problema de existencia de un posible parámetro óptimo, ω_{opt}. En lo siguiente, siempre usamos

$$\mathbf{A} = -\mathbf{L} + \mathbf{D} - \mathbf{U}, \quad \mathbf{B}(\omega) := (\mathbf{D} - \omega\mathbf{L})^{-1}\big((1-\omega)\mathbf{D} + \omega\mathbf{U}\big).$$

Teorema 4.10. *La matriz* $\mathbf{B}(\omega)$ *satisface*

$$r_\sigma\big(\mathbf{B}(\omega)\big) \geqslant |\omega - 1|. \tag{4.27}$$

Demostración. Usando que

$$(\mathbf{D} - \omega\mathbf{L})^{-1} = (\mathbf{I} - \omega\mathbf{D}^{-1}\mathbf{L})^{-1}\mathbf{D}^{-1},$$

podemos escribir

$$\begin{aligned}\det\big(\mathbf{B}(\omega) - \lambda\mathbf{I}\big) &= \det\Big((\mathbf{D} - \omega\mathbf{L})^{-1}\big((1-\omega)\mathbf{D} + \omega\mathbf{U} - \lambda(\mathbf{D} - \omega\mathbf{L})\big)\Big)\\ &= \det\big((1-\omega-\lambda)\mathbf{I} + \omega\mathbf{D}^{-1}\mathbf{U} + \lambda\omega\mathbf{D}^{-1}\mathbf{L}\big).\end{aligned}$$

Evaluando esta fórmula para $\lambda = 0$, obtenemos

$$\det\big(\mathbf{B}(\omega)\big) = \prod_{i=1}^{n} \lambda_i\big(\mathbf{B}(\omega)\big) = (1-\omega)^n.$$

Esto implica que

$$r_\sigma\big(\mathbf{B}(\omega)\big) = \max_{1\leqslant i\leqslant n} \big|\lambda_i\big(\mathbf{B}(\omega)\big)\big| \geqslant |\omega - 1|.$$

■

En consecuencia, para $\omega \in \mathbb{R}$ nos interesa solamente el intervalo $0 < \omega < 2$. Ya sabemos que para M-matrices, $r_\sigma(\mathbf{B}(\omega))$ es una función decreciente en el intervalo $0 < \omega \leqslant \tilde{\omega}$ con $\tilde{\omega} \geqslant 1$. Para matrices definidas positivas tenemos el siguiente teorema.

Teorema 4.11. *Sea $\mathbf{A} \in \mathbb{R}^{n\times n}$ simétrica y definida positiva. Entonces $r_\sigma(\mathbf{B}(\omega)) < 1$ para $0 < \omega < 2$.*

Demostración. Definimos la función

$$f(\mathbf{x}) := \frac{1}{2}\mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x} - \mathbf{b}^{\mathrm{T}}\mathbf{x},$$

notando que

$$f(\mathbf{x}) = -\frac{1}{2}\mathbf{b}^{\mathrm{T}}\mathbf{A}^{-1}\mathbf{b} + \frac{1}{2}(\mathbf{x} - \mathbf{A}^{-1}\mathbf{b})^{\mathrm{T}}(\mathbf{x} - \mathbf{A}^{-1}\mathbf{b}) \geqslant -\frac{1}{2}\mathbf{b}^{\mathrm{T}}\mathbf{A}^{-1}\mathbf{b}.$$

La definción de f implica $\nabla f(\mathbf{x}) = \mathbf{A}\mathbf{x} - \mathbf{b}$ y $\nabla^2 f(\mathbf{x}) = \mathbf{A}$, entonces $\mathbf{x}^* := \mathbf{A}^{-1}\mathbf{b}$ es el mínimo unicamente definido de f. Además, usando la notación de la descripición del método (4.18), (4.19), es decir, escribiendo

$$\mathbf{y}_0 := \mathbf{x}_0, \quad \mathbf{y}_{kn+i} := \begin{pmatrix} \xi_{1,k+1} \\ \vdots \\ \xi_{i,k+1} \\ \xi_{i+1,k} \\ \vdots \\ \xi_{n,k} \end{pmatrix}, \quad 1 \leqslant i \leqslant n, \quad k \in \mathbb{N}_0,$$

podemos definir

$$\mathbf{r}_j := \mathbf{A}\mathbf{y}_j - \mathbf{b}, \quad j \in \mathbb{N}_0; \quad \varrho_{j,i} := \mathbf{e}_i^{\mathrm{T}}\mathbf{r}_j.$$

Entonces, usando (4.19), podemos escribir

$$\mathbf{y}_{kn+i} - \mathbf{y}_{kn+i-1} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \xi_{i,k+1} - \xi_{i,k} \\ 0 \\ \vdots \\ 0 \end{pmatrix} = -\frac{\omega}{\alpha_{ii}}\big(\mathbf{e}_i^{\mathrm{T}}\mathbf{r}_{kn+i-1}\big)\mathbf{e}_i = -\frac{\omega\varrho_{kn+i-1,i}}{\alpha_{ii}}\mathbf{e}_i$$

y $\mathbf{x}_k = \mathbf{y}_{kn}$ para $k \in \mathbb{N}_0$. Una computación directa entrega

$$f(\mathbf{y}_{kn+i}) - f(\mathbf{y}_{kn+i-1}) = -\frac{\omega(2-\omega)}{2\alpha_{ii}}\varrho^2_{kn+i-1,i}. \tag{4.28}$$

Entonces, para $0 < \omega < 2$ la sucesión $\{f(\mathbf{y}_j)\}_{j\in\mathbb{N}}$ es monotonamente no creciente y acotada hacia abajo. Por lo tanto, existe

$$\lim_{k\to\infty} \varrho_{kn+i-1,i} = 0 \quad \text{para } i = 1, \ldots, n, \tag{4.29}$$

y entonces también

$$\lim_{k\to\infty} (\mathbf{y}_{kn+i} - \mathbf{y}_{kn+i-1}) = 0,$$

y dado que $\mathbf{r}_{j+1} - \mathbf{r}_j = \mathbf{A}(\mathbf{y}_{j+1} - \mathbf{y}_j)$, existe también el límite

$$\lim_{k\to\infty} (\mathbf{r}_{kn+i} - \mathbf{r}_{kn+i-1}) = 0, \quad i = 1, \ldots, n. \tag{4.30}$$

Todavía hay que demostrar que

$$\lim_{k\to\infty} \mathbf{r}_{kn} = 0. \tag{4.31}$$

Ahora, debido a (4.29) y (4.30),

$$\begin{aligned} |\varrho_{j+1,i} - \varrho_{j,i}| &\leqslant \varepsilon \quad \text{para } i = 1, \ldots, n \text{ y } j \geqslant j_0 = nk_0, \\ |\varrho_{kn+i-1,i}| &\leqslant \varepsilon \quad \text{para } k \geqslant k_0(\varepsilon) \text{ e } i = 1, \ldots, n. \end{aligned}$$

Entonces $|\varrho_{kn,1}| \leqslant \varepsilon$. Pero

$$\begin{aligned} |\varrho_{kn+1,2} - \varrho_{kn,2}| \leqslant \varepsilon \wedge |\varrho_{kn+1,2}| \leqslant \varepsilon &\Longrightarrow |\varrho_{kn,2}| \leqslant 2\varepsilon, \\ |\varrho_{kn+2,3} - \varrho_{kn+1,3}| \leqslant \varepsilon \wedge |\varrho_{kn+2,3}| \leqslant \varepsilon &\Longrightarrow |\varrho_{kn+1,2}| \leqslant 2\varepsilon, \\ |\varrho_{kn+1,3} - \varrho_{kn,3}| \leqslant \varepsilon \wedge |\varrho_{kn+1,3}| \leqslant 2\varepsilon &\Longrightarrow |\varrho_{kn,3}| \leqslant 3\varepsilon, \end{aligned}$$

y finalmente $|\varrho_{kn,n}| \leqslant n\varepsilon$, es decir, $\|\mathbf{r}_{kn}\|_\infty \leqslant n\varepsilon$, lo que implica (4.31), o sea

$$\mathbf{x}_k \overset{k\to\infty}{\longrightarrow} \mathbf{x}^* = \mathbf{A}^{-1}\mathbf{b}$$

El resultado sigue con el Teorema 4.1. ■

El Teorema 4.11 entrega una interpretación interesante del método SOR como método de minimización de la función f, cuyo gradiente es el "residuo" $\mathbf{Ax} = \mathbf{b}$. Las superficies $f(\mathbf{x}) = c$ en $\mathbb{R}^n$ son elipsoides concéntricos, y mediante el método SOR se alcanza el centro común $\mathbf{x}^*$ (el único mínimo de f en $\mathbb{R}^n$) a lo largo de las direcciones de coordenadas con descenso monótono de f.

Para una clase especial de matrices existe un resultado cuantitativo acerca de la dependencia de $r_\sigma(\mathbf{B}(\omega))$ de ω.

Definición 4.6. *Sea*

$$\mathbf{A} = \mathbf{D}(\mathbf{I} - \hat{\mathbf{L}} - \hat{\mathbf{U}}), \quad \hat{\mathbf{L}} := \mathbf{D}^{-1}\mathbf{L}, \quad \hat{\mathbf{U}} := \mathbf{D}^{-1}\mathbf{U},$$

partiendo de la partición usual de $\mathbf{A}$ *en una matriz diagonal* $\mathbf{D}$ *y en matrices* $\mathbf{L}$ *y* $\mathbf{U}$ *estrictamente triangulares. La matriz* $\mathbf{A}$ *se llama* ordenada consistentemente *si para cada* $\alpha, \beta \neq 0$,

$$\lambda \in \sigma\left(\alpha\hat{\mathbf{L}} + \frac{1}{\alpha}\hat{\mathbf{U}}\right) \Longleftrightarrow \lambda \in \sigma\left(\beta\hat{\mathbf{L}} + \frac{1}{\beta}\hat{\mathbf{U}}\right).$$

Teorema 4.12. *Supongamos que la matriz* $\mathbf{A} \in \mathbb{C}^{N\times N}$ *puede ser particionada como*

$$\mathbf{A} = \begin{bmatrix} \mathbf{D}_1 & \mathbf{A}_{12} & 0 & \cdots & \cdots & \cdots & 0 \\ \mathbf{A}_{21} & \mathbf{D}_2 & \mathbf{A}_{23} & 0 & & & \vdots \\ 0 & \mathbf{A}_{32} & \mathbf{D}_3 & \mathbf{A}_{34} & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & & & \ddots & \mathbf{A}_{n-1,n-2} & \mathbf{D}_{n-1} & \mathbf{A}_{n-1,n} \\ 0 & \cdots & \cdots & \cdots & 0 & \mathbf{A}_{n,n-1} & \mathbf{D}_n \end{bmatrix},$$

donde $\mathbf{D}_1, \ldots, \mathbf{D}_n$ *son matrices diagonales regulares. Entonces* $\mathbf{A}$ *es ordenada consistentemente.*

Demostración. Tarea. ■

La matriz del Ejemplo 4.1 no posee la forma requerida en el Teorema 4.12. Pero si cambiamos la enumeración a $(N,1),(N,2),(N-1,1),(N,3),(N-1,2),\ldots$, la matriz de coeficientes si asume esta forma. En este caso, se puede demostrar que la matriz es ordenada consistentemente (Tarea).

Teorema 4.13. *Sea* $\mathbf{A}$ *ordenada consistentemente. Entonces sabemos que para* $\mathbf{J} := \mathbf{D}^{-1}(\mathbf{L}+\mathbf{U})$,

a) $\lambda \in \sigma(\mathbf{J}) \Longleftrightarrow -\lambda \in \sigma(\mathbf{J})$,
b) $\mu \in \sigma(\mathbf{B}(\omega)) \Longleftrightarrow \exists \lambda \in \sigma(\mathbf{J}) : \mu\lambda^2\omega^2 = (\mu+\omega-1)^2$.

Demostración.

a) Sea

$$\mathbf{J}(\alpha) := \alpha\mathbf{D}^{-1}\mathbf{L} + \frac{1}{\alpha}\mathbf{D}^{-1}\mathbf{U}.$$

Entonces $\mathbf{J}(-1) = -\mathbf{J}(1)$, mientras que $\mathbf{J}(1)$ y $\mathbf{J}(-1)$ tienen los mismos valores propios según hipótesis.

b)"⇒": Para $\mu \neq 0$, sabemos que

$$\begin{aligned} \det\big(\mathbf{B}(\omega) - \mu\mathbf{I}\big) &= \det\big(\omega\mathbf{D}^{-1}\mathbf{U} + \mu\omega\mathbf{D}^{-1}\mathbf{L} + (1+\omega-\mu)\mathbf{I}\big) \\ &= \det\left(\omega\sqrt{\mu}\left[\frac{1}{\sqrt{\mu}}\mathbf{D}^{-1}\mathbf{U} + \sqrt{\mu}\mathbf{D}^{-1}\mathbf{L}\right] + (1+\omega-\mu)\mathbf{I}\right) \\ &= \det\big(\omega\sqrt{\mu}\mathbf{J}(\sqrt{\mu}) + (1+\omega-\mu)\mathbf{I}\big) \\ &= (\omega\sqrt{\mu})^n \det\left(\mathbf{J}(\sqrt{\mu}) - \frac{\mu+\omega-1}{\omega\sqrt{\mu}}\mathbf{I}\right). \end{aligned}$$

Entonces,

$$0 \neq \mu \in \sigma\big(\mathbf{B}(\omega)\big) \iff \frac{\mu+\omega-1}{\omega\sqrt{\mu}} \in \sigma\big(\mathbf{J}(\sqrt{\mu})\big) \iff \frac{\mu+\omega-1}{\omega\sqrt{\mu}} \in \sigma\big(\mathbf{J}(1)\big).$$

Pero si 0 es un valor propio de $\mathbf{B}(\omega)$, $(0 \in \sigma(\mathbf{B}(\omega)))$, entonces

$$\det\big(\mathbf{B}(\omega)\big) = (1-\omega)^n = 0 \Longrightarrow \omega = 1;$$

para este caso, b) es trivial.

"$\Leftarrow$": Sea λ un valor propio de $\mathbf{J}$ y $\mu\lambda^2\omega^2 = (\mu+\omega-1)^2$. En el caso $\mu \neq 0$, eso significa

$$\lambda = \pm\frac{\mu+\omega-1}{\omega\sqrt{\mu}};$$

usando la discusión de (a), podemos elegir

$$\lambda = \frac{\mu+\omega-1}{\omega\sqrt{\mu}},$$

es decir, $\lambda \in \sigma(\mathbf{J}(1))$ y también $\lambda \in \sigma(\mathbf{J}(\sqrt{\mu}))$, por lo tanto, $\mu \in \sigma(\mathbf{B}(\omega))$. Si $\mu = 0$, tenemos $\omega = 1$, pero $\det(\mathbf{B}(\omega)) = 0$, o sea, $0 \in \sigma(\mathbf{B}(1))$.

■

Podemos concluir que si $\mathbf{A}$ es ordenada consistentemente, entonces

$$r_\sigma\big((\mathbf{D}-\mathbf{L})^{-1}\mathbf{U}\big) = r_\sigma^2\big(\mathbf{D}^{-1}(\mathbf{L}+\mathbf{U})\big),$$

es decir, esencialmente el método de Gauss-Seidel converge dos veces más rápido que el método de Jacobi: notamos que para $\omega = 1$ en b), $\mu = \lambda^2$.

Teorema 4.14. *Sea* $\mathbf{A}$ *ordenada consistentemente. Supongamos que los valores propios* λ_i *de* $\mathbf{J} := \mathbf{D}^{-1}(\mathbf{L}+\mathbf{U})$ *satisfacen* $\lambda_i \in (-1,1)$ *para* $i = 1,\ldots,n$. *Entonces para*

$$\omega_{\text{opt}} := \frac{2}{1+\sqrt{1-\hat{\varrho}^2}}, \quad \hat{\varrho} := r_\sigma(\mathbf{J}), \tag{4.32}$$

el radio espectral de $\mathbf{B}(\omega)$ *es dado por*

$$r_\sigma\big(\mathbf{B}(\omega)\big) = \begin{cases} \left(\dfrac{\hat{\varrho}\omega}{2} + \dfrac{1}{2}\sqrt{\hat{\varrho}^2\omega^2 - 4(\omega-1)}\right)^2 & \text{para } 0 \leqslant \omega \leqslant \omega_{\text{opt}}, \\ \omega - 1 & \text{para } \omega_{\text{opt}} \leqslant \omega \leqslant 2. \end{cases} \tag{4.33}$$

La función $\omega \mapsto r_\sigma(\mathbf{B}(\omega))$ *es estrictamente decreciente sobre el intervalo* $[0, \omega_{\text{opt}}]$.

Demostración. La solución de la ecuación cuadrática $\mu\lambda^2\omega^2 = (\mu+\omega-1)^2$ del Teorema 4.13 entrega para $i = 1, 2$

$$\begin{aligned} \mu_i &:= \frac{1}{2}\left(\lambda_i^2\omega^2 - 2(\omega-1) \pm \sqrt{\big(\lambda_i^2\omega^2 - 2(\omega-1)\big)^2 - 4(\omega-1)^2}\right) \\ &= \frac{1}{2}\left(\lambda_i^2\omega^2 - 2(\omega-1) \pm \sqrt{\lambda_i^4\omega^4 - 4(\omega-1)\lambda_i^2\omega^2}\right) \\ &= \frac{1}{4}\lambda_i^2\omega^2 + \frac{1}{4}\lambda_i^2\omega^2 + 1 - \omega \pm \sqrt{\frac{\lambda_i^2\omega^2}{4} + (1-\omega)\cdot 2\frac{\lambda_i\omega}{2}} \end{aligned}$$

$$= \frac{1}{4}\left(\lambda_i\omega \pm \sqrt{\lambda_i^2\omega^2 + 4(1-\omega)}\right)^2 .$$

Aqui el radicando es no negativo siempre que

$$0 \leqslant \omega \leqslant \frac{2}{1+\sqrt{1-\lambda_i^2}} =: \omega_i. \quad i = 1, 2.$$

Para $\omega \geqslant \omega_i$ tenemos $|\mu_i| = \omega - 1$, lo cual se deriva usando que para $z \in \mathbb{C}$, $|z|^2 = (\text{Re } z)^2 + (\text{Im } z)^2$. Dado que $0 \leqslant \lambda_i^2 < 1$, tenemos que $\omega_i \geqslant 1$ y ω es monotonamente creciente con λ_i, lo que entrega la segunda parte de (4.33). Para $0 \leqslant \omega \leqslant \omega_i$ el valor mayor de valor absoluto de μ_i resulta de

$$|\mu_i| = \frac{1}{4}\left(|\lambda_i|\omega + \sqrt{\lambda_i^2\omega^2 + 4(1-\omega)}\right)^2 \geqslant \omega - 1,$$

y este valor crece monótonamente con λ_i, lo que demuestra la primera parte de (4.33). Diferenciando con respecto a ω obtenemos el último enunciado. ■

El Teorema 4.14 parte de la hipótesis que los valores propios de $\mathbf{J}$ son, en particular, reales. El siguiente lema informa que esto efectivamente es válido para una matriz $\mathbf{A}$ simétrica y definida positiva.

Lema 4.1. *Para una matriz* $\mathbf{A}$ *simétrica y definida positiva, los valores propios de* $\mathbf{J} = \mathbf{I} - \mathbf{D}^{-1}\mathbf{A}$ *son reales.*

Demostración. Puesto que la diagonal $\mathbf{D}$ de $\mathbf{A}$ es positiva, existe una matriz diagonal $\mathbf{F}$ con $\mathbf{D} = \mathbf{F}^2$, e

$$\begin{aligned}\mathbf{I} - \mathbf{D}^{-1}\mathbf{A} &= \mathbf{F}^{-1}(\mathbf{I} - \mathbf{F}^{-1}\mathbf{A}\mathbf{F}^{-1})\mathbf{F} \\ &= \mathbf{F}^{-1}\mathbf{M}\mathbf{F}.\end{aligned}$$

Dado que la matriz

$$\mathbf{M} = \mathbf{I} - \mathbf{F}^{-1}\mathbf{A}\mathbf{F}^{-1}$$

es simétrica y por lo tanto solo posee valores propios reales, tambien la transformada $\mathbf{I}-\mathbf{D}^{-1}\mathbf{A}$ solo posee valores propios reales. ■

Ejemplo 4.6. *Supongamos que*

$$\hat{\varrho} = r_\sigma(\mathbf{J}) = r_\sigma\big(\mathbf{D}^{-1}(\mathbf{L}+\mathbf{U})\big) = 0{,}9.$$

En este caso, de (4.32) *obtenemos*

$$\omega_{\text{opt}} = \frac{2}{1+\sqrt{1-0{,}9^2}} \approx 1{,}39286;$$

la Figura 4.3 muestra la función $\omega \mapsto r_\sigma(\mathbf{B}(\omega))$*, dada por* (4.33)*, que resulta en este caso.*

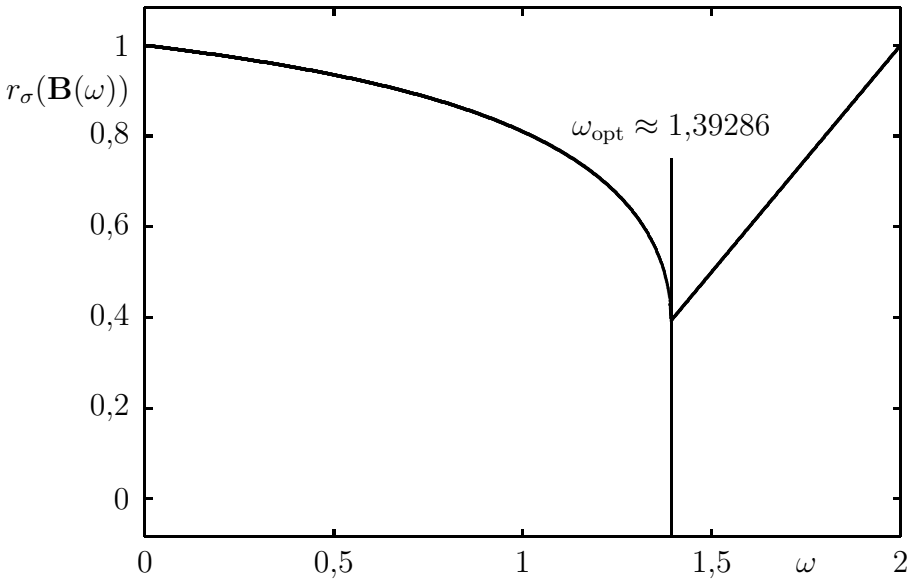


FIGURA 4.3. El radio espectral $r_\sigma(\mathbf{B}(\omega))$ dado por (4.33) para $\hat{\varrho} = 0{,}9$ (Ejemplo 4.6).

Ejemplo 4.7 (Tarea 20 b), Curso 2006)**.** *Demostrar que la siguiente matriz es ordenada consistentemente, y determinar (si posible) el parametro ω óptimo para el método SOR.*

$$\mathbf{A} = \begin{bmatrix} 2 & -1 & 1 & 0 \\ -1 & 4 & 0 & 1 \\ 1 & 0 & 4 & -1 \\ 0 & 1 & -1 & 2 \end{bmatrix}$$

Solución sugerida. *Descomponiendo* $\mathbf{A}$ *en bloques de los tamaños 1, 2 y 1,*

$$\mathbf{A} = \left[\begin{array}{c|cc|c} 2 & -1 & 1 & 0 \\ \hline -1 & 4 & 0 & 1 \\ 1 & 0 & 4 & -1 \\ \hline 0 & 1 & -1 & 2 \end{array}\right],$$

nos fijamos que $\mathbf{A}$ *es una matriz tridiagonal por bloques con bloques diagonales diagonales y regulares. La simetría y el Lema 4.1 implican que todos los valores propios de la matriz* $\mathbf{J} = \mathbf{I} - \mathbf{D}^{-1}\mathbf{A}$ *son reales. Dado que* $\mathbf{A}$ *es irreduciblemente diagonal dominante, el método de Jacobi converge, tal que los valores propios pertenecen a* $[-1, 1]$*. Entonces existe* ω_{opt}*. Finalmente, tenemos que*

$$\mathbf{J} = \begin{bmatrix} 0 & 1/2 & -1/2 & 0 \\ 1/4 & 0 & 0 & -1/4 \\ -1/4 & 0 & 0 & 1/4 \\ 0 & -1/2 & 1/2 & 0 \end{bmatrix} \Longrightarrow p(\lambda) = \det(\mathbf{J} - \lambda\mathbf{I}) = \lambda^2(\lambda^2 - 1/2),$$

entonces

$$\omega_{\text{opt}} = \frac{2}{1+\sqrt{1-\frac{1}{2}}} \approx 1{,}171573.$$

Ejemplo 4.8 (Tarea 21, Curso 2006)**.** *Sea*

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & -c \\ 0 & 1 & -d \\ -a & -b & 1 \end{bmatrix}, \quad a, b, c, d \in \mathbb{R}.$$

a) *¿Para qué valores de $z = ac + bd$ el método de Jacobi converge para la solución del sistema $\mathbf{Ax} = \mathbf{r}$?*
b) *¿Para qué valores de a, b, c, d la matriz $\mathbf{A}$ es irreducible?*
c) *Indicar la fórmula de iteración del método SOR para $\mathbf{Ax} = \mathbf{r}$ en forma explícita (o sea, sin matrices inversas) en la forma $\mathbf{x}_{k+1} = \mathbf{B}(\omega)\mathbf{x}_k + \mathbf{v}(\omega)$.*
d) *Sean $a = 0{,}5$, $b = 0{,}4$, $c = 0{,}7$, $d = -0{,}4$ y $\mathbf{r} = (-5, 9, 7)^{\mathrm{T}}$. (La solución exacta es $\tilde{\mathbf{x}} = (2, 5, 10)^{\mathrm{T}}$.) Partiendo de $\mathbf{x}_0 = (1, 1, 1)^{\mathrm{T}}$, calcular $\mathbf{x}_2$ con el método de Gauss-Seidel.*
e) *Sean $\mathbf{H}_1 := \mathbf{B}(1)$ y $\mathbf{H}$ la matriz de iteración del método de Jacobi. Demostrar que $\mathbf{A}$ es ordenada consistentemente y que $(r_\sigma(\mathbf{H}))^2 = r_\sigma(\mathbf{H}_1)$. Determinar $r_\sigma(\mathbf{H}_1)$ para los valores numéricos de (d).*
f) *Sea $0 < z < 1$. Demostrar que el método SOR aplicado a $\mathbf{A}$ posee un parámetro óptimo $\omega = \omega_{\text{opt}}$, y calcular el valor de ω_{opt} para los valores numéricos de (d). ¿Cuál es el valor del radio espectral correspondiente?*
g) *Partiendo de $\mathbf{x}_0$ especificado en (d), determinar $\tilde{\mathbf{x}}_2$ usando el método SOR y el parámetro ω óptimo.*

Solución sugerida.

a) *En este caso,*

$$\mathbf{J} = \begin{bmatrix} 0 & 0 & c \\ 0 & 0 & d \\ a & b & 0 \end{bmatrix} \Longrightarrow \det(\mathbf{J} - \lambda\mathbf{I}) = -\lambda(\lambda^2 + z),$$

entonces $r_\sigma(\mathbf{J}) < 1$ si y sólo si $|z| < 1$.
b) *P_1 puede ser conectado con P_3 si y sólo si $c \neq 0$, y P_2 con P_3 si y sólo si $d \neq 0$. Lo mismo es válido para los arcos dirigidos $P_3 \to P_1$ y $p_3 \to P_2$, o sea A es irreducible si y sólo si $a \neq 0$, $b \neq 0$, $c \neq 0$, $d \neq 0$.*
c)

$$\begin{aligned} \mathbf{B}(\omega) &= (\mathbf{D} - \omega\mathbf{L})^{-1}((1-\omega)\mathbf{D} + \omega\mathbf{U}) \\ &= \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -\omega a & -\omega b & 1 \end{bmatrix}^{-1} \begin{bmatrix} 1-\omega & 0 & \omega c \\ 0 & 1-\omega & \omega d \\ 0 & 0 & 1-\omega \end{bmatrix} \end{aligned}$$

$$= \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ \omega a & \omega b & 1 \end{bmatrix} \begin{bmatrix} 1-\omega & 0 & \omega c \\ 0 & 1-\omega & \omega d \\ 0 & 0 & 1-\omega \end{bmatrix}$$

$$= \begin{bmatrix} 1-\omega & 0 & \omega c \\ 0 & 1-\omega & \omega d \\ \omega(1-\omega)a & \omega(1-\omega)b & 1-\omega+\omega^2 z \end{bmatrix},$$

$$\mathbf{v}(\omega) = \omega \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ \omega a & \omega b & 1 \end{bmatrix} \mathbf{r}.$$

d)

$$\mathbf{x}_{k+1} = \begin{bmatrix} 0 & 0 & 0{,}7 \\ 0 & 0 & -0{,}4 \\ 0 & 0 & 0{,}19 \end{bmatrix} \mathbf{x}_k + \begin{pmatrix} -5 \\ 9 \\ 8{,}1 \end{pmatrix} \Longrightarrow \mathbf{x}_1 = \begin{pmatrix} -4{,}3 \\ 8{,}6 \\ 8{,}29 \end{pmatrix}, \quad \mathbf{x}_2 = \begin{pmatrix} 0{,}803 \\ 5{,}684 \\ 9{,}6751 \end{pmatrix}$$

e) *Puesto que*

$$\det\left(\alpha\mathbf{L} + \frac{1}{\alpha}\mathbf{U} - \lambda\mathbf{I}\right) = -\lambda(\lambda^2 + z)$$

independentemente de α*, la matriz* $\mathbf{A}$ *es ordenada consistentemente. Usando (a), vemos que* $r_\sigma(\mathbf{H}) = \sqrt{0{,}19}$*; entonces* $(r_\sigma(\mathbf{H}))^2 = r_\sigma(\mathbf{H}_1)$*, es decir* $r_\sigma(\mathbf{H}_1) = 0{,}19$.

f) *Según (a),* $r_\sigma(\mathbf{H}) < 1$ *y todos los valores propios de* $\mathbf{H}$ *son reales. Según (e),* $\mathbf{A}$ *es ordenada consistentemente, entonces existe* ω_{opt}*. Aquí tenemos*

$$\omega_{\text{opt}} = \frac{2}{1 + \sqrt{1 - (r_\sigma(\mathbf{H}))^2}} = \frac{20}{19} = 1{,}0526, \quad r_\sigma(\mathbf{H}_{\omega_{\text{opt}}}) = 0{,}0526.$$

g)

$$\mathbf{x}^1 = \begin{pmatrix} -4{,}5789 \\ 9 \\ 8{,}795 \end{pmatrix}, \quad \mathbf{x}^2 = \begin{pmatrix} 1{,}4583 \\ 5{,}2968 \\ 10{,}0485 \end{pmatrix}.$$

Se reconoce la gran ganancia en velocidad de convergencia usando $\omega = \omega_{\text{opt}}$. Pero en la práctica, se recomienda *sobreestimar* ω, dado que en el lado izquierdo de ω_{opt} la tangente es vertical (como ilustra la Figura 4.3). Entonces, necesitamos una cota superior de $r_\sigma(\mathbf{D}^{-1}(\mathbf{L}+\mathbf{U}))$. Veremos en el Capítulo 5 como nos podemos conseguir una tal cota.

Sin demostración comunicamos el siguiente teorema.

Teorema 4.15. *Sea* $\mathbf{A}$ *una matriz tridiagonal por bloques con bloques diagonales* $\mathbf{I}$*, simétrica, y definida positiva. Entonces para*

$$\kappa := \big(\text{cond}_{\|\cdot\|_2}(\mathbf{A})\big)^{1/2}, \quad \beta := \left(2 + \frac{4}{\kappa}\right)^2$$

tenemos la inclusión

$$\frac{\kappa - 1}{\kappa + 1} \geqslant r_\sigma\big(\mathbf{B}(\omega_{\text{opt}})\big) \geqslant \frac{\kappa - \beta}{\kappa + \beta}. \tag{4.34}$$

Este resultado significa que para una matriz mal condicionada, el método SOR con ω_{opt} aún es muy lento, pero mucho más rápido que los métodos de Gauss-Seidel o Jacobi, dado que bajo las hipótesis del Teorema 4.15 tenemos que

$$r_\sigma^2\big(\mathbf{D}^{-1}(\mathbf{L}+\mathbf{U})\big) = r_\sigma^2(\mathbf{L}+\mathbf{U}) = r_\sigma\big((\mathbf{D}-\mathbf{L})^{-1}\mathbf{U}\big),$$

donde

$$r_\sigma\big(\mathbf{D}^{-1}(\mathbf{L}+\mathbf{U})\big) = 1-\lambda_{\text{mín}}(\mathbf{A}) = \lambda_{\text{máx}}(\mathbf{A})-1 = 1-\frac{\lambda_{\text{máx}}(\mathbf{A})}{\text{cond}_{\|\cdot\|_2}(\mathbf{A})}$$

con $1 \leqslant \lambda_{\text{máx}}(\mathbf{A}) \leqslant 2$.

Ejemplo 4.9. *Si* $\text{cond}_{\|\cdot\|_2}(\mathbf{A}) = 10000$, *tenemos* $\kappa = 100$ *y* $r_\sigma \geqslant 0{,}9998$ *para el método de Jacobi y* $r_\sigma \geqslant 0{,}9996$ *para el método de Gauss-Seidel, pero*

$$r_\sigma\big(\mathbf{B}(\omega_{\text{opt}})\big) \leqslant 1-\frac{2}{101} = 0{,}980198$$

para el método SOR con ω_{opt}. *Notar que depués de 1000 pasos,*

$$0{,}9998^{1000} = 0{,}8187, \quad 0{,}9996^{1000} = 0{,}67026, \quad 0{,}980198^{1000} = 2{,}059\times 10^{-9}.$$

Una pregunta obvia es cómo se puede estimar el radio espectral de la matriz de iteración con poco esfuerzo computacional. A parte de considerar vectores iniciales especiales, podemos considerar la expresión $\|\mathbf{A}\mathbf{x}_k - \mathbf{b}\|^{1/k}$.

Teorema 4.16. *Sea* $\mathbf{A}$ *regular y el sistema* $\mathbf{A}\mathbf{x}^* = \mathbf{b}$ *equivalente a* $\mathbf{x}^* = \mathbf{G}\mathbf{x}^* + \mathbf{g}$, *y la sucesión* $\{\mathbf{x}_k\}_{k\in\mathbb{N}_0}$ *definida por*

$$\mathbf{x}_{k+1} = \mathbf{G}\mathbf{x}_k + \mathbf{g}, \quad k \in \mathbb{N}_0. \tag{4.35}$$

Entonces, para cualquier norma $\|\cdot\|$, *tenemos que*

$$\limsup_{k\to\infty} \|\mathbf{A}\mathbf{x}_k - \mathbf{b}\|^{1/k} \leqslant r_\sigma(\mathbf{G}), \tag{4.36}$$

y existe un vector $\mathbf{x}_0$ *para el cual* (4.36) *vale con "=".*

Demostración. Tarea. ■

4.4. Métodos de iteración por bloque

Se ofrece la siguiente generalización de los métodos de iteración discutidos hasta ahora. Se particiona la matriz $\mathbf{A}$, el vector de solución $\mathbf{x}$ y la parte derecha $\mathbf{b}$ en bloques y subvectores, respectivamente:

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} & \cdots & \mathbf{A}_{1n} \\ \mathbf{A}_{21} & \mathbf{A}_{22} & \cdots & \mathbf{A}_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{A}_{n1} & \cdots & \cdots & \mathbf{A}_{nn} \end{bmatrix}, \quad \mathbf{x} = \begin{pmatrix} \mathbf{x}_1 \\ \vdots \\ \mathbf{x}_n \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} \mathbf{b}_1 \\ \vdots \\ \mathbf{b}_n \end{pmatrix},$$

y en la derivación de los métodos ponemos

$$\mathbf{D} = \begin{bmatrix} \mathbf{A}_{11} & & \\ & \ddots & \\ & & \mathbf{A}_{nn} \end{bmatrix}, \; -\mathbf{L} = \begin{bmatrix} 0 & \cdots & \cdots & 0 \\ \mathbf{A}_{21} & 0 & & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ \mathbf{A}_{n1} & \cdots & \mathbf{A}_{n,n-1} & 0 \end{bmatrix}, \; -\mathbf{U} = \begin{bmatrix} 0 & \mathbf{A}_{12} & \cdots & \mathbf{A}_{1n} \\ \vdots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \mathbf{A}_{n-1,n} \\ 0 & \cdots & \cdots & 0 \end{bmatrix}. \tag{4.37}$$

Por ejemplo, la iteración del método *Bloque-SOR* es definida por

$$\begin{aligned} &\mathbf{A}_{ii}\mathbf{x}_{i,k+1} = \omega \left(\mathbf{b}_i - \sum_{j=1}^{i-1} \mathbf{A}_{ij}\mathbf{x}_{j,k+1} - \sum_{j=i}^{n} \mathbf{A}_{ij}\mathbf{x}_{j,k} \right) + \mathbf{A}_{ii}\mathbf{x}_{i,k}, \\ &i = 1, \ldots, n, \quad k \in \mathbb{N}_0. \end{aligned} \tag{4.38}$$

Este procedimiento requiere en cada paso la solución de n sistemas lineales "pequeños". Sin embargo, tal procedimiento puede ser ventajoso cuando, por ejemplo, las matrices $\mathbf{A}_{ii}$ son simplemente estructuradas (por ejemplo, tridiagonales). Los Teoremas 4.10–4.14 pueden ser generalizados facilmente. A modo de ejemplo, tenemos el siguiente teorema.

Teorema 4.17. *Sea* $\mathbf{A}$ *una matriz tridiagonal por bloques y definida positiva. Entonces, el método* (4.38) *converge para* $0 < \omega < 2$. *El parámetro óptimo* ω_{opt} *es dado por*

$$\omega_{\text{opt}} = \frac{2}{1 + \sqrt{1 - \mu^2}}, \quad \mu = r_\sigma\big(\mathbf{D}^{-1}(\mathbf{L} + \mathbf{U})\big),$$

donde $\mathbf{D}$, $\mathbf{L}$ *y* $\mathbf{U}$ *estan dadas por* (4.37), *y la función*

$$\omega \mapsto r_\sigma\Big((\mathbf{D} - \omega\mathbf{L})^{-1}\big((1-\omega)\mathbf{D} + \omega\mathbf{U}\big)\Big)$$

tiene las mismas propiedades que las especificadas en el Teorema 4.14.

4.5. El método de gradientes conjugados (cg) de Hestenes y Stiefel

Ya mencionamos en la demostración del Teorema 4.11 que el método SOR para la solución de $\mathbf{Ax} = \mathbf{b}$, donde $\mathbf{A}$ es simétrica y definida positiva, puede ser interpretado como un método de minimización para la función

$$f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x} - \mathbf{b}^{\mathrm{T}}\mathbf{x} \tag{4.39}$$

con el gradiente

$$\nabla f(\mathbf{x}) = \mathbf{A}\mathbf{x} - \mathbf{b}, \tag{4.40}$$

donde se procede en orden cíclico a lo largo de las direcciones de coordenadas. El caso $n = 2$ ya ilustra que los ejes de las coordenadas no necesariamente deben ser las más ventajosas.

En el caso $n = 2$, las curvas $f(\mathbf{x}) = c$ son elipses concéntricas. Una minimización de f a lo largo de los ejes principales de la elipse entregaría el mínimo de f, o sea, la solución de $\mathbf{Ax}^* = \mathbf{b}$, en dos pasos. El resultado análogo también es válido para $n \geqslant 3$. Los ejes principales forman un caso especial la las llamadas *direcciones* $\mathbf{A}$*-ortogonales*, que en este caso también son ortogonales.

Definición 4.7. *Sea* $\mathbf{A} \in \mathbb{R}^{n\times n}$ *simétrica y definida positiva. Un sistema* $\{\mathbf{p}_0, \mathbf{p}_1, \ldots, \mathbf{p}_{n-1}\}$ *de vectores se llama* $\mathbf{A}$-ortogonal *si para* $0 \leqslant j, k \leqslant n-1$,

$$\mathbf{p}_j^{\mathrm{T}} \mathbf{A} \mathbf{p}_k = \kappa_k \delta_{jk}, \quad \kappa_j > 0, \quad \delta_{jk} = \begin{cases} 0 & \text{si } j \neq k, \\ 1 & \text{si } j = k. \end{cases}$$

Pero, para la solución del sistema $\mathbf{A}\mathbf{x} = \mathbf{b}$, la determinación de los ejes principales de $\mathbf{A}$, es decir, de sus vectores propios, significaria un esfuerzo computacional altamente exagerado. Demostraremos ahora que la minimización a lo largo de direcciones $\mathbf{A}$-ortogonales, tambien llamadas $\mathbf{A}$-*conjugadas*, también entrega un método finito.

Ejemplo 4.10. *Consideramos el sistema* $\mathbf{A}\mathbf{x} = \mathbf{b}$ *con*

$$\mathbf{A} = \begin{bmatrix} 2 & -1 \\ -1 & 3 \end{bmatrix}, \quad \mathbf{b} = \begin{pmatrix} 3 \\ 1 \end{pmatrix}, \tag{4.41}$$

con la solución exacta $\mathbf{x}^* = (2,1)^{\mathrm{T}}$*; la matriz* $\mathbf{A}$ *es simétrica y definida positiva. Un ejemplo de direcciones* $\mathbf{A}$ *ortogonales son*

$$\mathbf{p}_0 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad \mathbf{p}_1 = \frac{1}{\sqrt{5}} \begin{pmatrix} 1 \\ 2 \end{pmatrix}.$$

La Figura 4.4 muestra las elipses concéntricas $f(\mathbf{x}) = c$*, donde* f *es definida por* (4.39) *y* $c = -3, -2, -1, 0, \ldots$*; el mínimo es* $f(\mathbf{x}^*) = -3{,}5$.

Teorema 4.18. *Sea la función* f *definida por* (4.39)*, con* $\mathbf{A} \in \mathbb{R}^{n\times n}$ *simétrica y definida positiva. Sea* $\{\mathbf{p}_0, \mathbf{p}_1, \ldots, \mathbf{p}_{n-1}\}$ *un sistema de direcciones* $\mathbf{A}$*-ortogonales. Sea* $\mathbf{x}_0$ *arbitrario, y*

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \sigma_k \mathbf{p}_k, \quad \sigma_k := \frac{\mathbf{p}_k^{\mathrm{T}}(\mathbf{A}\mathbf{x}_k - \mathbf{b})}{\mathbf{p}_k^{\mathrm{T}} \mathbf{A} \mathbf{p}_k}, \quad k = 0, \ldots, n-1. \tag{4.42}$$

Entonces

(i) $\mathbf{x}_{k+1}$ *minimiza* $f(\mathbf{x}_k - \sigma \mathbf{p}_k)$ *con respecto a* σ*,*

(ii) $(\mathbf{A}\mathbf{x}_k - \mathbf{b})^{\mathrm{T}} \mathbf{p}_j = 0$ *para* $j = 0, \ldots, k-1$*, es decir,*

$$\mathbf{x}_k = \operatorname{argmin}\big\{ f(\mathbf{x}) \mid \mathbf{x} = \mathbf{x}_0 + \operatorname{span}\{\mathbf{p}_0, \ldots, \mathbf{p}_{k-1}\} \big\}, \tag{4.43}$$

(iii) $\mathbf{x}_n = \mathbf{x}^* = \mathbf{A}^{-1}\mathbf{b}$.

Demostración.

(i) Utilizando el cálculo diferencial, obtenemos

$$\begin{aligned} & \frac{\mathrm{d}}{\mathrm{d}\sigma} f(\mathbf{x}_k - \sigma \mathbf{p}_k) = 0 \\ \iff \quad & -\big(\nabla f(\mathbf{x}_k - \sigma \mathbf{p}_k)\big)^{\mathrm{T}} \mathbf{p}_k = 0 \\ \iff \quad & \mathbf{p}_k^{\mathrm{T}} \big(\mathbf{A}(\mathbf{x}_k - \sigma \mathbf{p}_k) - \mathbf{b}\big) = 0 \\ \iff \quad & \sigma = \frac{\mathbf{p}_k^{\mathrm{T}}(\mathbf{A}\mathbf{x}_k - \mathbf{b})}{\mathbf{p}_k^{\mathrm{T}} \mathbf{A} \mathbf{p}_k}. \end{aligned}$$

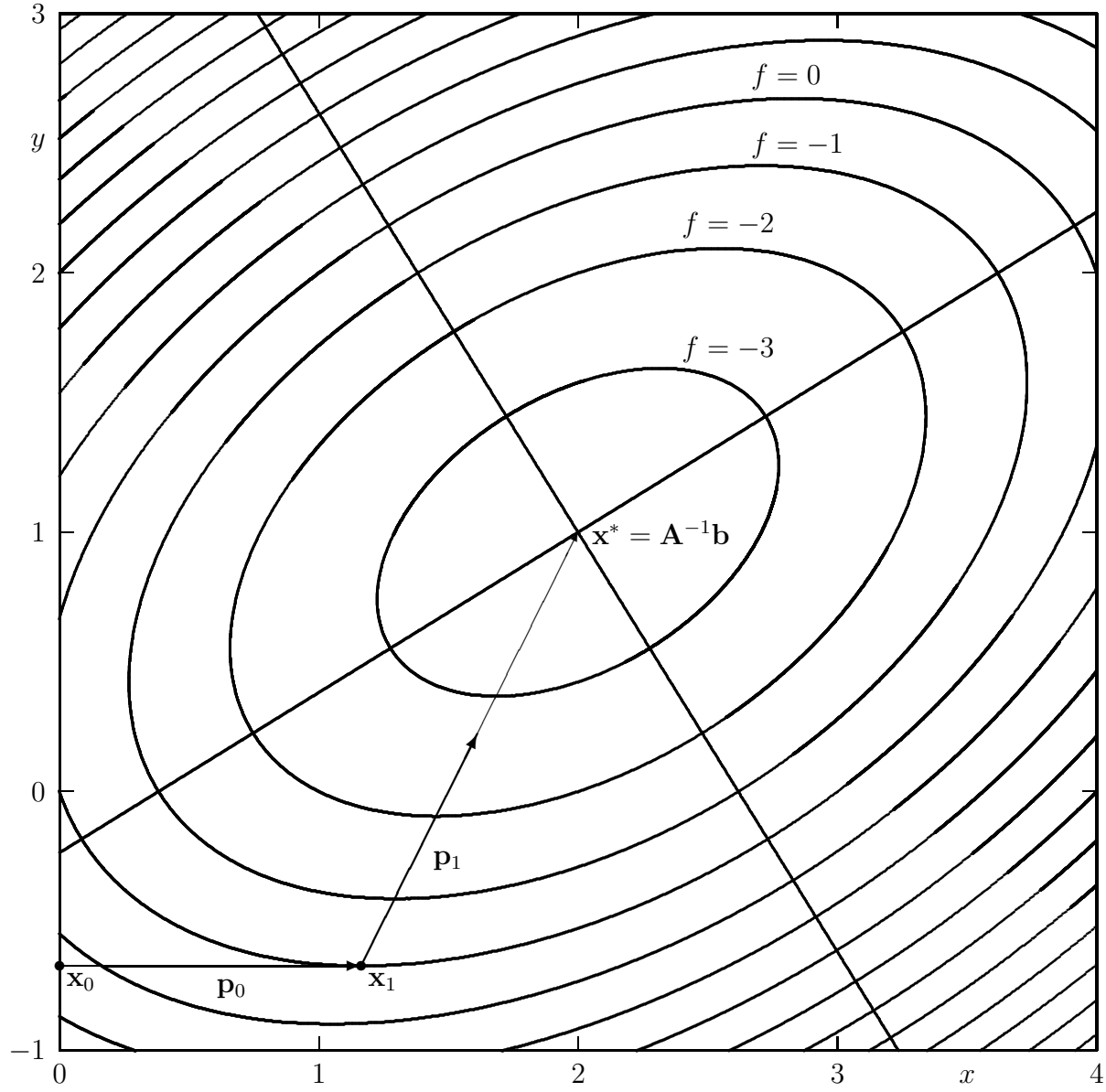


FIGURA 4.4. Ejemplos 4.10 y 4.11: Las curvas $f(\mathbf{x}) = c$, $c = -3, -2, -1, 0, \ldots$, direcciones $\mathbf{A}$-ortogonales $\mathbf{p}_0$ y $\mathbf{p}_1$, la solución exacta $\mathbf{x}^* = \mathbf{A}^{-1}\mathbf{b}$, y $\mathbf{x}_0$ y $\mathbf{x}_1$.

(ii) Procedemos por inducción. Para $k = 1$, (ii) es la consecuencia de (i). Supongamos ahora que

$$(\mathbf{A}\mathbf{x}_k - \mathbf{b})^{\mathrm{T}}\mathbf{p}_j = 0, \quad j = 0, \ldots, k-1.$$

Hay que demostrar que

$$(\mathbf{A}\mathbf{x}_{k+1} - \mathbf{b})^{\mathrm{T}}\mathbf{p}_j = 0, \quad j = 0, \ldots, k. \tag{4.44}$$

Para $j = k$, (4.44) es una consecuencia obvia de (i). Pero para $j < k$, calculamos que

$$\begin{aligned}(\mathbf{A}\mathbf{x}_{k+1} - \mathbf{b})^{\mathrm{T}}\mathbf{p}_j &= \big(\mathbf{A}(\mathbf{x}_k - \sigma_k\mathbf{p}_k) - \mathbf{b}\big)^{\mathrm{T}}\mathbf{p}_j \\ &= (\mathbf{A}\mathbf{x}_k - \mathbf{b} - \sigma_k\mathbf{A}\mathbf{p}_k)^{\mathrm{T}}\mathbf{p}_j \\ &= (\mathbf{A}\mathbf{x}_k - \mathbf{b})^{\mathrm{T}}\mathbf{p}_j - \sigma_k\mathbf{p}_k^{\mathrm{T}}\mathbf{A}\mathbf{p}_j = 0.\end{aligned}$$

Esta última expresión es cero debido a la hipótesis de inducción y la definición de los vectores $\mathbf{p}_j$.

(iii) Para $k = n$, sabemos que

$$(\mathbf{A}\mathbf{x}_n - \mathbf{b})^{\mathrm{T}}\begin{bmatrix}\mathbf{p}_0 & \cdots & \mathbf{p}_{n-1}\end{bmatrix} = 0.$$

Dado que $\begin{bmatrix}\mathbf{p}_0 & \cdots & \mathbf{p}_{n-1}\end{bmatrix}$ es una matriz regular, se tiene que $\mathbf{A}\mathbf{x}_n = \mathbf{b}$.

■

Ejemplo 4.11. *Continuamos considerando el sistema del Ejemplo 4.10, partiendo de*

$$\mathbf{x}_0 = \begin{pmatrix}0\\ 1-2\sqrt{0{,}7}\end{pmatrix} = \begin{pmatrix}0\\ -0{,}67332005\end{pmatrix}.$$

En este caso, obtenemos de (4.42) *con* $k = 0$

$$\begin{aligned}\sigma_0 &= \frac{(1,0)\left(\begin{bmatrix}2 & -1\\ -1 & 3\end{bmatrix}\begin{pmatrix}0\\ 1-2\sqrt{0{,}7}\end{pmatrix} - \begin{pmatrix}3\\ 1\end{pmatrix}\right)}{(1,0)\begin{bmatrix}2 & -1\\ -1 & 3\end{bmatrix}\begin{pmatrix}1\\ 0\end{pmatrix}} \\ &= \frac{(1,0)\begin{pmatrix}2\sqrt{0{,}7} - 1 - 3\\ -6\sqrt{0{,}7} + 3 - 1\end{pmatrix}}{(1,0)\begin{pmatrix}2\\ -1\end{pmatrix}} = \frac{2\sqrt{0{,}7} - 4}{2} = \sqrt{0{,}7} - 2 = -1{,}16333997347,\end{aligned}$$

entonces

$$\mathbf{x}_1 = \begin{pmatrix}0\\ 1-2\sqrt{0{,}7}\end{pmatrix} - (\sqrt{0{,}7} - 2)\begin{pmatrix}1\\ 0\end{pmatrix} = \begin{pmatrix}2-\sqrt{0{,}7}\\ 1-2\sqrt{0{,}7}\end{pmatrix} = \begin{pmatrix}1{,}16333997347\\ -0{,}673320053068\end{pmatrix}.$$

Luego calculamos

$$\sigma_1 = \sqrt{5}\frac{(1,2)\left(\begin{bmatrix}2 & -1\\ -1 & 3\end{bmatrix}\begin{pmatrix}2-\sqrt{0{,}7}\\ 1-2\sqrt{0{,}7}\end{pmatrix} - \begin{pmatrix}3\\ 1\end{pmatrix}\right)}{(1,2)\begin{bmatrix}2 & -1\\ -1 & 3\end{bmatrix}\begin{pmatrix}1\\ 2\end{pmatrix}} = -\sqrt{\frac{7}{2}},$$

$$\mathbf{x}_2 = \begin{pmatrix}2-\sqrt{0{,}7}\\ 1-2\sqrt{0{,}7}\end{pmatrix} - \left(-\sqrt{\frac{7}{2}}\right)\cdot\frac{1}{\sqrt{5}}\begin{pmatrix}1\\ 2\end{pmatrix} = \begin{pmatrix}2\\ 1\end{pmatrix} = \mathbf{x}^*.$$

La Figura 4.4 muestra $\mathbf{x}_0$ *y* $\mathbf{x}_1$.

La construcción del Teorema 4.18 implica que la dirección $\mathbf{p}_k$ sólo se necesita cuando $\mathbf{x}_k$ ya ha sido determinado. Podemos aprovechar esa observación para generar las direcciones $\mathbf{A}$-ortogonales $\mathbf{p}_j$ mediante un método de ortogonalización sucesiva (con respecto al producto escalar $(\mathbf{x}, \mathbf{y}) = \mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{y}$) durante la computación, partiendo de

$$\mathbf{p}_0 := \mathbf{A}\mathbf{x}_0 - \mathbf{b} = \nabla f(\mathbf{x}_0),$$

según

$$\mathbf{p}_k = \nabla f(\mathbf{x}_k) + \sum_{j=0}^{k-1} \beta_{kj}\mathbf{p}_j, \tag{4.45}$$

donde $\nabla f(\mathbf{x}_k) \neq 0$ (sino ya se ha encontrado el mínimo deseado), donde β_{kj} se determina de tal forma que

$$\mathbf{p}_j^{\mathrm{T}}\mathbf{A}\mathbf{p}_k = 0, \quad j = 0, \ldots, k-1. \tag{4.46}$$

La observación importante para poder ejecutar el método es que resultará que

$$\beta_{k,0} = \ldots = \beta_{k,k-2} = 0, \quad \beta_{k,k-1} = \frac{(\nabla f(\mathbf{x}_k))^{\mathrm{T}}\nabla f(\mathbf{x}_k)}{(\nabla f(\mathbf{x}_{k-1}))^{\mathrm{T}}\nabla f(\mathbf{x}_{k-1})}. \tag{4.47}$$

Eso significa que cada paso requiere sólo muy poco esfuerzo computacional y espacio de almacenaje.

Teorema 4.19. *Sea la función f definida por* (4.39)*, con $\mathbf{A} \in \mathbb{R}^{n\times n}$ simétrica y definida positiva. Sea $\mathbf{x}_0 \in \mathbb{R}^n$ arbitrario y*

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \sigma_k\mathbf{p}_k,$$

donde

$$\mathbf{p}_k = \begin{cases} \nabla f(\mathbf{x}_k) & \text{para } k = 0, \\ \nabla f(\mathbf{x}_k) + \dfrac{\|\nabla f(\mathbf{x}_k)\|_2^2}{\|\nabla f(\mathbf{x}_{k-1})\|_2^2}\mathbf{p}_{k-1} & \text{para } k > 0, \end{cases}$$

$$\sigma_k = \frac{(\nabla f(\mathbf{x}_k))^{\mathrm{T}}\mathbf{p}_k}{\mathbf{p}_k^{\mathrm{T}}\mathbf{A}\mathbf{p}_k}$$

Entonces, $\nabla f(\mathbf{x}_j) \neq 0$ para $j = 0, \ldots, k$ implica que $\{\mathbf{p}_0, \ldots, \mathbf{p}_k\}$ son direcciones $\mathbf{A}$-ortogonales, es decir existe un índice $N \leqslant n$ tal que $\mathbf{x}_N = \mathbf{A}^{-1}\mathbf{b}$.

Demostración. En lo siguiente, sea

$$\mathbf{r}_j := \nabla f(\mathbf{x}_j) = \mathbf{A}\mathbf{x}_j - \mathbf{b}, \quad \beta_j := \frac{\|\mathbf{r}_j\|_2^2}{\|\mathbf{r}_{j-1}\|_2^2}.$$

Para $k = 0$, notamos que si $\mathbf{r}_0 \neq 0$, entonces $\mathbf{p}_0 = \mathbf{r}_0 \neq 0$. Para $k = 1$, hay que demostrar que $\mathbf{r}_1 \neq 0$ implica que $\mathbf{p}_1^{\mathrm{T}}\mathbf{A}\mathbf{p}_0 = 0$, pero $\mathbf{p}_1 \neq 0$. Pero sabemos que $\mathbf{p}_1 = \mathbf{r}_1 + \beta_1\mathbf{r}_0$ implica que

$$-\mathbf{p}_1^{\mathrm{T}}\mathbf{A}\mathbf{p}_0 = \mathbf{p}_1^{\mathrm{T}}\left(\frac{1}{\sigma_0}(\mathbf{r}_1 - \mathbf{r}_0)\right)$$

$$= \frac{1}{\sigma_0}(\mathbf{r}_1^{\mathrm{T}} + \beta_1 \mathbf{r}_0^{\mathrm{T}})(\mathbf{r}_1 - \mathbf{r}_0)$$
$$= \frac{1}{\sigma_0}\left(\mathbf{r}_1^{\mathrm{T}}\mathbf{r}_1 - \mathbf{r}_1^{\mathrm{T}}\mathbf{r}_0 + \frac{\mathbf{r}_1^{\mathrm{T}}\mathbf{r}_1}{\mathbf{r}_0^{\mathrm{T}}\mathbf{r}_0}\mathbf{r}_0^{\mathrm{T}}\mathbf{r}_1 - \mathbf{r}_1^{\mathrm{T}}\mathbf{r}_1\right).$$

Ahora, en virtud de

$$\begin{aligned}
\mathbf{x}_1 &= \mathbf{x}_0 - \sigma_0 \mathbf{r}_0, \\
\sigma_0 &= \frac{\mathbf{r}_0^{\mathrm{T}}\mathbf{r}_0}{\mathbf{r}_0^{\mathrm{T}}\mathbf{A}\mathbf{r}_0}, \\
\mathbf{r}_1 &= \mathbf{A}\mathbf{x}_1 - \mathbf{b} \\
&= \mathbf{A}\mathbf{x}_0 - \sigma_0 \mathbf{A}\mathbf{r}_0 - \mathbf{b} \\
&= \mathbf{r}_0 - \frac{\mathbf{r}_0^{\mathrm{T}}\mathbf{r}_0}{\mathbf{r}_0^{\mathrm{T}}\mathbf{A}\mathbf{r}_0}\mathbf{A}\mathbf{r}_0,
\end{aligned}$$

sabemos que

$$\mathbf{r}_0^{\mathrm{T}}\mathbf{r}_1 = \mathbf{r}_0^{\mathrm{T}}\mathbf{r}_0 - \frac{\mathbf{r}_0^{\mathrm{T}}\mathbf{r}_0}{\mathbf{r}_0^{\mathrm{T}}\mathbf{A}\mathbf{r}_0}\mathbf{r}_0^{\mathrm{T}}\mathbf{A}\mathbf{r}_0 = 0.$$

Etonces, resulta $\mathbf{p}_1^{\mathrm{T}}\mathbf{A}\mathbf{p}_0 = 0$. Puesto que $\mathbf{r}_1^{\mathrm{T}}\mathbf{p}_1 = \mathbf{r}_1^{\mathrm{T}}\mathbf{r}_1 \neq 0$, se tiene que $\mathbf{p}_1 \neq 0$.

Finalmente, consideremos el paso $k \to k+1$. Hay que demostrar ahora que si $\mathbf{r}_{k+1} \neq 0$ y $\{\mathbf{p}_0, \ldots, \mathbf{p}_k\}$ son $\mathbf{A}$-ortogonales, entonces $\{\mathbf{p}_0, \ldots, \mathbf{p}_{k+1}\}$ son $\mathbf{A}$-ortogonales, es decir, $\mathbf{p}_{k+1}^{\mathrm{T}}\mathbf{A}\mathbf{p}_j = 0$ para $j = 0, \ldots, k$, y $\mathbf{p}_{k+1} \neq 0$. Para tal efecto, notamos que $\mathbf{r}_{k+1}^{\mathrm{T}}\mathbf{p}_k = 0$ implica

$$\mathbf{r}_{k+1}^{\mathrm{T}}\mathbf{p}_{k+1} = \mathbf{r}_{k+1}^{\mathrm{T}}\mathbf{r}_{k+1} \neq 0,$$

es decir, $\mathbf{p}_{k+1} \neq 0$. Luego consideramos que

$$\begin{aligned}
-\mathbf{A}\mathbf{p}_j &= \frac{1}{\sigma_j}(\mathbf{r}_{j+1} - \mathbf{r}_j) \\
&= \frac{1}{\sigma_j}(\mathbf{p}_{j+1} - \beta_{j+1}\mathbf{p}_j - \mathbf{p}_j + \beta_j \mathbf{p}_{j-1}),
\end{aligned}$$

entonces, según (ii) del Teorema 4.18,

$$\begin{aligned}
-\mathbf{p}_{k+1}^{\mathrm{T}}\mathbf{A}\mathbf{p}_j &= -\big(\mathbf{r}_{k+1}^{\mathrm{T}} + \beta_{k+1}\mathbf{p}_k^{\mathrm{T}}\big)\mathbf{A}\mathbf{p}_j \\
&= -\mathbf{r}_{k+1}^{\mathrm{T}}\mathbf{A}\mathbf{p}_j \\
&= \frac{1}{\sigma_j}\big(\mathbf{r}_{k+1}^{\mathrm{T}}\mathbf{p}_{j+1} - \beta_{j+1}\mathbf{r}_{k+1}^{\mathrm{T}}\mathbf{p}_j - \mathbf{r}_{k+1}^{\mathrm{T}}\mathbf{p}_j + \beta_j \mathbf{r}_{k+1}^{\mathrm{T}}\mathbf{p}_{j-1}\big) = 0.
\end{aligned}$$

Para $j = k$, podemos escribir

$$\begin{aligned}
-\mathbf{p}_{k+1}^{\mathrm{T}}\mathbf{A}\mathbf{p}_k &= \big(\mathbf{r}_{k+1}^{\mathrm{T}} + \beta_{k+1}\mathbf{p}_k^{\mathrm{T}}\big) \cdot \left(\frac{1}{\sigma_k}(\mathbf{r}_{k+1} - \mathbf{r}_k)\right) \\
&= \frac{1}{\sigma_k}\big(\mathbf{r}_{k+1}^{\mathrm{T}}\mathbf{r}_{k+1} + \beta_{k+1}\mathbf{p}_k^{\mathrm{T}}\mathbf{r}_{k+1} - \mathbf{r}_{k+1}^{\mathrm{T}}\mathbf{r}_k - \beta_{k+1}\mathbf{r}_k^{\mathrm{T}}\mathbf{p}_k\big).
\end{aligned}$$

Tomando en cuenta que $\mathbf{p}_k^{\mathrm{T}}\mathbf{r}_{k+1} = 0$, obtenemos

$$-\mathbf{p}_{k+1}^{\mathrm{T}}\mathbf{A}\mathbf{p}_k = \frac{1}{\sigma_k}\left(\mathbf{r}_{k+1}^{\mathrm{T}}\mathbf{r}_{k+1}\left[1 - \frac{\mathbf{r}_k^{\mathrm{T}}\mathbf{p}_k}{\mathbf{r}_k^{\mathrm{T}}\mathbf{r}_k}\right] - \mathbf{r}_{k+1}^{\mathrm{T}}\mathbf{r}_k\right).$$

Usando que $[\ldots] = 0$, podemos seguir escribiendo

$$-\mathbf{p}_{k+1}^{\mathrm{T}}\mathbf{A}\mathbf{p}_k = -\frac{1}{\sigma_k}\mathbf{r}_{k+1}^{\mathrm{T}}(\mathbf{p}_k - \beta_k\mathbf{p}_{k-1}).$$

Finalmente, sabemos que $\mathbf{r}_{k+1}^{\mathrm{T}}\mathbf{p}_k = 0$, por lo tanto

$$\begin{aligned}
-\mathbf{p}_{k+1}^{\mathrm{T}}\mathbf{A}\mathbf{p}_k &= \frac{\beta_k}{\sigma_k}\mathbf{r}_{k+1}^{\mathrm{T}}\mathbf{p}_{k-1} \\
&= \frac{\beta_k}{\sigma_k}\big(\mathbf{A}(\mathbf{x}_k - \sigma_k\mathbf{p}_k) - \mathbf{b}\big)^{\mathrm{T}}\mathbf{p}_{k-1} \\
&= \frac{\beta_k}{\sigma_k}\big(\mathbf{r}_k^{\mathrm{T}}\mathbf{p}_{k-1} - \sigma_k\mathbf{p}_k^{\mathrm{T}}\mathbf{A}\mathbf{p}_{k-1}\big) = 0,
\end{aligned}$$

lo que concluye la demostración. ■

Ejemplo 4.12 (Tarea 22, Curso 2006)**.** *Resolver el sistema* $\mathbf{A}\mathbf{x} = \mathbf{b}$ *con*

$$\mathbf{A} = \begin{bmatrix} 5 & 1 & 1 \\ 1 & 5 & -1 \\ 1 & -1 & 5 \end{bmatrix}, \quad \mathbf{b} = \begin{pmatrix} 4 \\ 2 \\ -4 \end{pmatrix}$$

usando el método cg de Hestenes y Stiefel, $\mathbf{x}_0 = 0$*, y calculando exactamente con fracciones.*

Solución sugerida. *Con* $\mathbf{D}(\mathbf{x}) = \mathbf{A}\mathbf{x} - \mathbf{b}$ *obtenemos sucesivamente*

$$\mathbf{D}(\mathbf{x}_0) = \begin{pmatrix} -4 \\ -2 \\ 4 \end{pmatrix}, \quad \mathbf{p}_0 = \begin{pmatrix} -4 \\ -2 \\ 4 \end{pmatrix}, \quad (\mathbf{D}(\mathbf{x}_0))^{\mathrm{T}}\mathbf{p}_0 = 36, \quad \mathbf{A}\mathbf{p}_0 = 18\begin{pmatrix} -1 \\ -1 \\ 1 \end{pmatrix},$$

$$\mathbf{p}_0^{\mathrm{T}}\mathbf{A}\mathbf{p}_0 = 180, \quad \sigma_0 = \frac{1}{5}, \quad \mathbf{x}_1 = \frac{2}{5}\begin{pmatrix} 2 \\ 1 \\ -2 \end{pmatrix};$$

$$\mathbf{D}(\mathbf{x}_1) = \frac{2}{5}\begin{pmatrix} -1 \\ 4 \\ 1 \end{pmatrix}, \quad \mathbf{p}_1 = \frac{18}{25}\begin{pmatrix} -1 \\ 2 \\ 1 \end{pmatrix}, \quad (\mathbf{D}(\mathbf{x}_1))^{\mathrm{T}}\mathbf{p}_1 = 4\frac{18}{25}, \quad \mathbf{A}\mathbf{p}_1 = \frac{18}{25}\begin{pmatrix} -2 \\ 8 \\ 2 \end{pmatrix},$$

$$\mathbf{p}_1^{\mathrm{T}}\mathbf{A}\mathbf{p}_1 = 20\left(\frac{18}{25}\right)^2, \quad \sigma_1 = \frac{5}{18}, \quad \mathbf{x}_2 = \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix}.$$

Entonces, a pesar de su estructura iterativa, el método entrega (si se usa aritmética exacta) la solución de un sistema lineal con una matriz $\mathbf{A}$ simétrica y definida positiva después de a lo más n pasos. Si $\mathbf{A}$ es dispersa, es decir, posee sólo pocos elementos diferentes de cero, cada paso cuesta poco esfuerzo computacional. Sin embargo, el método es muy sensitivo con respecto a errores de redondeo; por lo tanto, después de n pasos, obtenemos

solamente una solución aproximada (falsificada) y no exacta. Se puede empezar el método de nuevo con $\mathbf{x}_n$ como vector inicial, o simplemente se puede continuar.

Puede parecer sorpresivo que un método iterativo, como SOR, aún puede competir con el método cg. Eso tiene que ver con que en la práctica, para sistemas lineales de gran tamaño no se necesita la solución exacta, y frecuentamente se desea terminar el método después de pocos pasos de iteración. Ahora, mientras que un método tal como el método SOR garantiza una reducción del error más o menos igual en cada paso, el método cg es un poco irregular en este aspecto, como ilustra el siguiente teorema.

Teorema 4.20. *Sea $\mathbf{A} \in \mathbb{R}^{n\times n}$ simétrica y definida positiva con los valores propios $\lambda_1 \geqslant \lambda_2 \geqslant \ldots \geqslant \lambda_n > 0$, $\mathbf{b} \in \mathbb{R}^n$, $\mathbf{x}^* = \mathbf{A}^{-1}\mathbf{b}$ y*

$$E(\mathbf{x}) := \frac{1}{2}(\mathbf{x}-\mathbf{x}^*)^{\mathrm{T}}\mathbf{A}(\mathbf{x}-\mathbf{x}^*).$$

Entonces la sucesión $\{\mathbf{x}_k\}$ generada por el método cg satisface

$$\begin{aligned} E(\mathbf{x}_{k+1}) &= \frac{1}{2}\min_{P_k\in\Pi_k}(\mathbf{x}_0-\mathbf{x}^*)^{\mathrm{T}}\mathbf{A}\big(\mathbf{I}+\mathbf{A}P_k(\mathbf{A})\big)^2(\mathbf{x}_0-\mathbf{x}^*) \\ &\leqslant \min_{P_k\in\Pi_k}\max_{1\leqslant i\leqslant n}\big|1+\lambda_i P_k(\lambda_i)\big|^2 E(\mathbf{x}_0) \\ &\leqslant \left(\frac{\lambda_{k+1}-\lambda_n}{\lambda_{k+1}+\lambda_n}\right)^2 E(\mathbf{x}_0), \quad k=0,\ldots,n-1. \end{aligned} \tag{4.48}$$

Recordamos que Π_k es el espacio de los polinomios con coeficientes reales del grado máximo k.

Demostración. Primero demostramos que

$$\mathbf{p}_j = P_j(\mathbf{A})\mathbf{r}_0, \quad P_j \in \Pi_j, \quad \mathbf{r}_0 = \mathbf{A}\mathbf{x}_0 - \mathbf{b}. \tag{4.49}$$

Para tal efecto, notamos primero que

$$\mathbf{p}_0 = \mathbf{r}_0 = P_0(\mathbf{A})\mathbf{r}_0, \quad P_0 \equiv 1 \in \Pi_0.$$

Luego supongamos que se ha demostrado (4.49) hasta el índice $j-1$. Entonces tenemos que

$$\begin{aligned} \mathbf{p}_j &= \mathbf{A}\mathbf{x}_j - \mathbf{b} + \beta_j\mathbf{p}_{j-1} = \mathbf{A}\left(\mathbf{x}_0 - \sum_{i=0}^{j-1}\sigma_i\mathbf{p}_i\right) - \mathbf{b} + \beta_j\mathbf{p}_{j-1} \\ &= \mathbf{A}\mathbf{x}_0 - \mathbf{b} - \sum_{i=0}^{j-1}\sigma_i\mathbf{A}\mathbf{p}_i + \beta_j\mathbf{p}_{j-1} = \mathbf{r}_0 - \sum_{i=0}^{j-1}\sigma_i\mathbf{A}\mathbf{p}_i + \beta_j\mathbf{p}_{j-1} \\ &= P_j(\mathbf{A})\mathbf{r}_0, \end{aligned}$$

donde

$$P_j(\tau) = 1 + \beta_j P_{j-1}(\tau) - \sum_{i=0}^{j-1}\sigma_i\tau P_i(\tau).$$

Notando que $P_i \in \Pi_{j-1}$ para $i = 0,\ldots,j-1$ (según hipótesis), concluimos que $P_j \in \Pi_j$.

En virtud de lo anterior, podemos escribir

$$\begin{aligned}\mathbf{x}_j - \mathbf{x}^* &= \mathbf{x}_0 - \mathbf{x}^* - \sum_{i=0}^{j-1} \sigma_i \mathbf{p}_i \\ &= \mathbf{x}_0 - \mathbf{x}^* - \sum_{i=0}^{j-1} \sigma_i P_i(\mathbf{A})\mathbf{r}_0 \\ &= \left(\mathbf{I} - \sum_{i=0}^{j-1} \sigma_i P_i(\mathbf{A})\mathbf{A}\right)(\mathbf{x}_0 - \mathbf{x}^*) \\ &= \big(\mathbf{I} + \mathbf{A}Q_{j-1}(\mathbf{A})\big)(\mathbf{x}_0 - \mathbf{x}^*),\end{aligned}$$

donde definimos

$$Q_{j-1}(\tau) := -\sum_{i=0}^{j-1} \sigma_i P_i(\tau); \qquad Q_{j-1} \in \Pi_{j-1}.$$

Sea ahora

$$\mathbf{A} = \mathbf{V}^{\mathrm{T}} \operatorname{diag}(\lambda_1, \ldots, \lambda_n)\mathbf{V}, \quad \mathbf{V}^{\mathrm{T}}\mathbf{V} = \mathbf{V}\mathbf{V}^{\mathrm{T}} = \mathbf{I}.$$

Entonces, usando $\mathbf{\Lambda} := \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$, podemos escribir

$$\mathbf{V}(\mathbf{x}_j - \mathbf{x}^*) = \big(\mathbf{I} + \mathbf{\Lambda}Q_{j-1}(\mathbf{\Lambda})\big)\mathbf{V}(\mathbf{x}_0 - \mathbf{x}^*).$$

Definimos $\mathbf{y} := (\eta_1, \ldots, \eta_n)^{\mathrm{T}} = \mathbf{V}(\mathbf{x}_0 - \mathbf{x}^*)$, y notamos que $\mathbf{I} + \mathbf{\Lambda}Q_{j-1}(\mathbf{\Lambda})$ es una matriz diagonal. Entonces, podemos escribir

$$\begin{aligned}E(\mathbf{x}_{k+1}) &= \frac{1}{2}(\mathbf{x}_{k+1} - \mathbf{x}^*)^{\mathrm{T}}\mathbf{V}^{\mathrm{T}}\mathbf{\Lambda}\mathbf{V}(\mathbf{x}_{k+1} - \mathbf{x}^*) \\ &= \frac{1}{2}\mathbf{y}^{\mathrm{T}}\big(\mathbf{I} + \mathbf{\Lambda}Q_k(\mathbf{\Lambda})\big)^2\mathbf{\Lambda}\mathbf{y} \\ &= \frac{1}{2}\sum_{i=1}^{n} \eta_i^2 \lambda_i \big(1 + \lambda_i Q_k(\lambda_i)\big)^2.\end{aligned}$$

En virtud de (4.43), éste es el valor más pequeño que puede ser alcanzado a través de la construcción

$$\mathbf{p}_j = F_j(\mathbf{A})\mathbf{r}_0, \quad \mathbf{x}_{j+1} = \mathbf{x}_j - \tau_j \mathbf{p}_j, \quad j = 0, \ldots, k, \quad F_j \in \Pi_j.$$

Tomando en cuenta que la dependencia de los coeficientes β_j y σ_i es presente sólo en el polinomio Q_k, podemos escribir

$$\begin{aligned}E(\mathbf{x}_{k+1}) &= \min_{F_k \in \Pi_k} \frac{1}{2}\sum_{i=1}^{n} \eta_i^2 \lambda_i \big(1 + \lambda_i F_k(\lambda_i)\big)^2 \\ &\leqslant \min_{F_k \in \Pi_k} \max_{1 \leqslant i \leqslant n} \big(1 + \lambda_i F_k(\lambda_i)\big)^2 \cdot \frac{1}{2}\sum_{i=1}^{n} \eta_i^2 \lambda_i \qquad (4.50) \\ &= \min_{F_k \in \Pi_k} \max_{1 \leqslant i \leqslant n} \big(1 + \lambda_i F_k(\lambda_i)\big)^2 E(\mathbf{x}_0).\end{aligned}$$

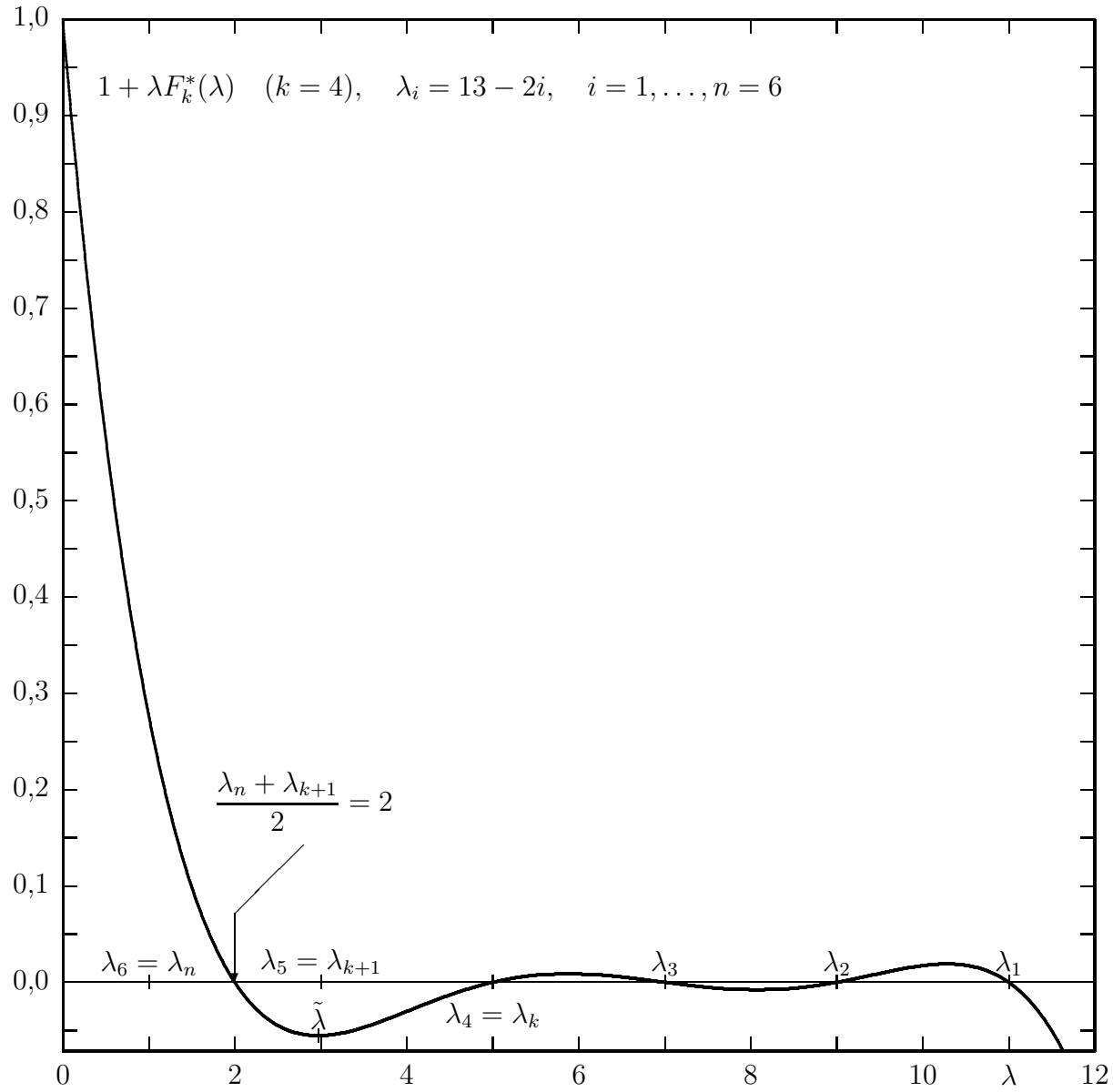


FIGURA 4.5. La función $1 + \lambda F_k^*(\lambda)$ para $k = 4$ y $\lambda_i = 13 - 2i$, $i = 1, \ldots, n = 6$.

Para acotar el lado derecho de (4.50), escogimos la siguiente función $F_k^* \in \Pi_k$:

$$F_k^*(\lambda) = \frac{1}{\lambda}\left(\frac{2 \cdot (-1)^{k+1}}{\lambda_1 \cdot \ldots \cdot \lambda_k(\lambda_{k+1} + \lambda_n)}\left(\lambda - \frac{\lambda_{k+1} + \lambda_n}{2}\right)\prod_{m=1}^{k}(\lambda - \lambda_m) - 1\right). \tag{4.51}$$

Es decir,

$$1 + \lambda F_k^*(\lambda) = 0 \quad \text{para } \lambda \in \{\lambda_1, \ldots, \lambda_k\} \cup \left\{\frac{\lambda_{k+1} + \lambda_n}{2}\right\}.$$

Dado que $1+\lambda F_k^*(\lambda) \in \Pi_{k+1}$, este polinomio es similar al polinomio graficado en la Figura 4.5. Hasta un valor

$$\tilde{\lambda} \in \left(\frac{\lambda_{k+1} + \lambda_n}{2}, \lambda_k \right),$$

el polinomio $1+\lambda F_k^*(\lambda)$ es monótonamente decreciente y convexo, y monótonamente creciente (pero decreciente en valor absoluto) en $[\tilde{\lambda}, \lambda_k]$. El valor $\tilde{\lambda}$ es definido por

$$\tilde{\lambda} \in \left[\frac{\lambda_{k+1} + \lambda_n}{2}, \lambda_k \right], \quad (F_k^*)'(\tilde{\lambda})\tilde{\lambda} + F_k^*(\tilde{\lambda}) = 0.$$

Concluimos que

$$\lambda \in [\lambda_n, \lambda_{k+1}] \Longrightarrow \left|1 + \lambda F_k^*(\lambda)\right| \leqslant \left| 1 - \frac{2\lambda}{\lambda_{k+1} + \lambda_n} \right| \leqslant \frac{\lambda_{k+1} - \lambda_n}{\lambda_{k+1} + \lambda_n}.$$

■

Entonces, si por ejemplo,

$$\lambda_1 \gg \lambda_n, \quad \lambda_i = \lambda_{i-1} - \varepsilon, \quad i = 2, \ldots, n-1, \quad \varepsilon \ll \lambda_1 - \lambda_n,$$

los $n-1$ primeros pasos generan solamente una reducción muy pequeña del error.

Teorema 4.21. *Sea $\mathbf{A} \in \mathbb{R}^{n\times n}$ simétrica y definida positiva con $k < n$ valores propios reales y distintos $\tilde{\lambda}_1, \ldots, \tilde{\lambda}_k$. En este caso, el método cg ya converge después de k iteraciones para $\mathbf{A}\mathbf{x}^* = \mathbf{b}$, o sea, $\mathbf{x}_k = \mathbf{x}^*$.*

Demostración. Usamos el Teorema 4.20, que indica que la cantidad

$$E(\mathbf{x}) := \frac{1}{2}(\mathbf{x} - \mathbf{x}^*)^{\mathrm{T}} \mathbf{A} (\mathbf{x} - \mathbf{x}^*)$$

satisface la desigualdad

$$E(\mathbf{x}_{k+1}) \leqslant \min_{P_k \in \Pi_k} \max_{1 \leqslant i \leqslant n} \left|1 + \lambda_i P_k(\lambda_i)\right|^2 E(\mathbf{x}_0),$$

donde Π_k es el espacio de todos los polinomios del grado maximo k. Ahora hay que demostrar que bajo los hipótesis de la tarea, $E(\mathbf{x}_k) = 0$, es decir que existe un polinomio $P_{k-1} \in \Pi_{k-1}$ tal que

$$1 + \lambda_i P_{k-1}(\lambda_i) = 0, \quad i = 1, \ldots, n. \tag{4.52}$$

Ahora sabemos que el polinomio

$$\tilde{P}_k(\lambda) = (\lambda - \tilde{\lambda}_1) \cdot \cdots \cdot (\lambda - \tilde{\lambda}_k), \quad \tilde{P}_k \in \Pi_k$$

satisface $\tilde{P}_k(\lambda_i) = 0$, $i = 1, \ldots, n$, con $\tilde{P}(0) = (-1)^k \tilde{\lambda}_1 \cdot \cdots \cdot \tilde{\lambda}_k$. Entonces

$$P_{k-1}(\lambda) := \frac{1}{\lambda} \left(\frac{(-1)^k}{\tilde{\lambda}_1 \cdot \cdots \cdot \tilde{\lambda}_k} \tilde{P}_k(\lambda) - 1 \right)$$

es un polinomio en Π_{k-1} tal que se satisface (4.52). ■

Capítulo 5

El problema de valores propios de una matriz

Discutiremos ahora el problema de la localización y la determinación numérica de los valores propios reales y complejos de una matriz $\mathbf{A} \in \mathbb{C}^{n\times n}$ (o $\mathbf{A} \in \mathbb{R}^{n\times n}$) y los vectores propios asociados, o sea el problema de determinar los ceros del polinomio

$$p_n(\lambda; \mathbf{A}) := \det(\mathbf{A} - \lambda\mathbf{I})$$

y la solución de los sistemas homogéneos $(\mathbf{A} - \lambda_i\mathbf{I})\mathbf{x}_i = 0$. Formalmente, la solución del problema es dada por (a) la determinación de los coeficientes de $p_n(\lambda; \mathbf{A})$, (b) la computación (exacta o aproximada) de sus ceros y (c) la solución de los sistemas lineales homogéneos. Sin embargo, en la práctica, este camino es absolutamente inútil, bajo el aspecto del esfuerzo computacional tanto que bajo el de la estabilidad numérica. Para ilustrar el último punto, consideremos un pequeño ejemplo.

Ejemplo 5.1. *La matriz*

$$\mathbf{A} = \begin{bmatrix} 1000 & 1 \\ 1 & 1000 \end{bmatrix}$$

tiene los valores propios $\lambda_1 = 1001$ *y* $\lambda_2 = 999$. *Ahora, si modificamos* $\mathbf{A}$ *a*

$$\tilde{\mathbf{A}} = \begin{bmatrix} 1000{,}001 & 1 \\ 1 & 1000 \end{bmatrix},$$

obtenemos $\lambda_1 = 1001{,}00050\ldots$ *y* $\lambda_2 = 999{,}00050\ldots$. *Sabemos que*

$$p_2(\lambda, \mathbf{A}) = \lambda^2 - 2000\lambda + 999999,$$
$$p_2(\lambda, \tilde{\mathbf{A}}) = \lambda^2 - 2000{,}001\lambda + 1000000.$$

Ahora, si el coeficiente 10^6 *en* $p_2(\lambda, \tilde{\mathbf{A}})$ *se cambia a* 1000002 *(correspondiente a la magnitud de errores de redondeo en una aritmética con 6 dígitos), el polinomio modificado tiene los ceros*

$$\lambda^2 - 2000{,}001\lambda + 1000002 = 0 \Longleftrightarrow \lambda = 1000{,}0005 \pm 0{,}99999927\mathrm{i},$$

es decir que la influencia del error en los coeficientes de $p_2(\lambda : \mathbf{A})$ *es casi 2000 veces mayor que la de los errores en la matriz original.*

5.1. La localización de valores propios y la sensitividad del problema

Para la siguiente discusión es útil recordar el siguiente teorema.

Teorema 5.1. *Sea* $\mathbf{A} \in \mathbb{C}^{n\times n}$ *y* $\|\cdot\|$ *una norma matricial inducida por una norma vectorial. Entonces cada valor propio* $\lambda(\mathbf{A})$ *satisface* $|\lambda(\mathbf{A})| \leqslant r_\sigma(\mathbf{A}) \leqslant \|\mathbf{A}\|$.

Teorema 5.2 (Los círculos de Gershgorin)**.** *Para una matriz* $\mathbf{A} \in \mathbb{C}^{n\times n}$ *definimos los círculos*

$$\mathcal{K}_i := \left\{ \lambda \in \mathbb{C} \;\middle|\; |\lambda - \alpha_{ii}| \leqslant \sum_{j=1, j\neq i}^{n} |\alpha_{ij}| \right\}, \quad \bar{\mathcal{K}}_i := \left\{ \lambda \in \mathbb{C} \;\middle|\; |\lambda - \alpha_{ii}| \leqslant \sum_{j=1, j\neq i}^{n} |\alpha_{ji}| \right\},$$

para $i = 1, \ldots, n$*. Sea* $\lambda(\mathbf{A})$ *un valor propio de* $\mathbf{A}$*, entonces*

$$\lambda(\mathbf{A}) \in \left(\bigcup_{i=1}^{n} \mathcal{K}_i \right) \cap \left(\bigcup_{i=1}^{n} \bar{\mathcal{K}}_i \right).$$

Demostración. Sean $\lambda \in \sigma(\mathbf{A})$ y $\mathbf{x}$ el vector propio asociado, es decir $\mathbf{Ax} = \lambda\mathbf{x}$, e $i \in \{1, \ldots, n\}$ tal que $|x_i| = \|\mathbf{x}\|_\infty$. Entonces, la componente i de la ecuación vectorial $\mathbf{Ax} = \lambda\mathbf{x}$ es

$$\sum_{j=1}^{n} \alpha_{ij} x_j = \lambda x_i \Longleftrightarrow \sum_{\substack{j=1 \\ j\neq i}}^{n} \alpha_{ij} x_j + \alpha_{ii} x_i = \lambda x_i,$$

lo cual entrega que

$$\lambda - \alpha_{ii} = \sum_{\substack{j=1 \\ j\neq i}}^{n} \alpha_{ij} \frac{x_j}{x_i}.$$

Dado que $x_i \neq 0$, podemos concluir que

$$|\lambda - \alpha_{ii}| \leqslant \sum_{\substack{j=1 \\ j\neq i}}^{n} |\alpha_{ij}| \frac{|x_j|}{|x_i|} \leqslant \sum_{\substack{j=1 \\ j\neq i}}^{n} |\alpha_{ij}|,$$

y luego se toma en cuenta que $\mathbf{A}$ y $\mathbf{A}^*$ tienen los mismos valores propios. ■

Dado que las matrices $\mathbf{A}$ y $\mathbf{D}^{-1}\mathbf{AD}$ poseen los mismos valores propios, a veces se puede precisar el resultado del Teorema 5.2 significativamente.

Ejemplo 5.2. *La matriz*

$$\mathbf{A} = \begin{bmatrix} 1 & 10^{-3} & 10^{-4} \\ 10^{-3} & 2 & 10^{-3} \\ 10^{-4} & 10^{-3} & 3 \end{bmatrix}$$

es simétrica, entonces sus valores propios son reales, y según el Teorema 5.2, cada valor propio λ *de* $\mathbf{A}$ *satisface*

$$\lambda \in [1 - 0{,}0011, 1 + 0{,}0011] \cup [2 - 0{,}002, 2 + 0{,}002] \cup [3 - 0{,}0011, 3 + 0{,}0011].$$

Ahora, usando las matrices

$$\mathbf{D}_1 := \operatorname{diag}(1, 100, 10), \quad \mathbf{D}_2 := \operatorname{diag}(100, 1, 100), \quad \mathbf{D}_3 := \operatorname{diag}(10, 100, 1),$$

obtenemos las siguientes inclusiones respectivas:

$$\lambda \in U_1 := [1 - 2 \times 10^{-5}, 1 + -2 \times 10^{-5}] \cup [2 - 0{,}11, 2 + 0{,}11] \cup [3 - 0{,}0011, 3 + 0{,}0011],$$
$$\lambda \in U_2 := [1 - 0{,}1001, 1 + 0{,}1001] \cup [2 - 2 \times 10^{-5}, 2 + 2 \times 10^{-5}] \cup [3 - 0{,}1001, 3 + 0{,}1001],$$

$$\lambda \in U_3 := [1-0{,}0011, 1+0{,}0011] \cup [2-0{,}02, 2+0{,}02] \cup [3-2\times 10^{-5}, 3+2\times 10^{-5}],$$

lo que implica

$$\lambda \in U_1 \cap U_2 \cap U_3 = \bigcup_{i=1}^{3} [i-2\times 10^{-5}, i+2\times 10^{-5}].$$

Teorema 5.3. *Consideremos las hipótesis del Teorema 5.2. Sea* $\{i_1, \ldots, i_n\}$ *una permutación de* $\{1, \ldots, n\}$ *y*

$$(\mathcal{K}_{i_1} \cup \ldots \cup \mathcal{K}_{i_m}) \cap \mathcal{K}_{i_s} = \varnothing \quad \text{para } s = m+1, \ldots, n.$$

Entonces $\mathcal{K}_{i_1} \cup \ldots \cup \mathcal{K}_{i_m}$ *contiene exactamente* m *valores propios de* $\mathbf{A}$*, contados con su multiplicidad, es decir, cada componente de conectividad por camino de* $\mathcal{K}_{i_1} \cup \ldots \cup \mathcal{K}_{i_m}$ *contiene tantos valores propios de* $\mathbf{A}$ *que círculos.*

Demostración. Sean $\mathbf{D} = \operatorname{diag}(\alpha_{11}, \ldots, \alpha_{nn})$ y $\mathbf{B}(\tau) := \mathbf{D} + \tau(\mathbf{A} - \mathbf{D})$, $0 \leqslant \tau \leqslant 1$, es decir, $\mathbf{B}(0) = \mathbf{D}$ y $\mathbf{B}(1) = \mathbf{A}$. Todos los valores propios de $\mathbf{B}(\tau)$ estan contenidos en $\mathcal{K}_1(\tau) \cup \ldots \cup \mathcal{K}_n(\tau)$, donde definimos

$$\mathcal{K}_i(\tau) = \left\{ z \in \mathbb{C} \;\middle|\; |z - \alpha_{ii}| \leqslant \tau \sum_{j=1, j\neq i}^{n} |\alpha_{ij}| \right\}, \quad i = 1, \ldots, n.$$

Obviamente, el Teorema 5.3 es válido para $\mathbf{B}(0)$, además, los valores propios dependen de forma continua de τ (ver Lema 5.1 abajo). Pero como $\mathcal{K}_{i_1}(0) \cup \cdots \cup \mathcal{K}_{i_m}(0)$ contiene exactamente m valores propios de $\mathbf{B}(0)$, y

$$\forall \tau \in [0,1] : \left(\bigcup_{j=1}^{m} \mathcal{K}_{i_j}(\tau) \right) \cap \mathcal{K}_{i_s}(1) \subset \left(\bigcup_{j=1}^{m} \mathcal{K}_{i_j}(1) \right) \cap \mathcal{K}_{i_s}(1) = \varnothing,$$

entonces $\mathcal{K}_{i,1}(\tau) \cup \cdots \cup \mathcal{K}_{i,m}(\tau)$ contiene exactamente m valores propios para $0 \leqslant \tau \leqslant 1$. ■

Lema 5.1. *Sean* $\mathbf{A}, \mathbf{B} \in \mathbb{C}^{n\times n}$*,* $\lambda_1, \ldots, \lambda_n$ *los valores propios de* $\mathbf{A}$*, contados con su multiplicidad, y* $\lambda'_1, \ldots, \lambda'_n$ *los valores propios de* $\mathbf{B}$ *contados con su multiplicidad. Sean*

$$\varrho := \max\{|\alpha_{ij}|, |\beta_{ij}| : 1 \leqslant i, j \leqslant n\}, \quad \delta := \frac{1}{n\varrho} \sum_{i=1}^{n} \sum_{j=1}^{n} |\alpha_{ij} - \beta_{ij}|.$$

Entonces existe una enumeración de los valores propios λ_i *y* λ'_i *tal que a cada* λ_i *corresponde un valor* λ'_i *con*

$$|\lambda_i - \lambda'_i| \leqslant 2(n+1)^2 \varrho \sqrt[n]{\delta}.$$

Demostración. Ver A.M. Ostrowski, Solution of Equations in Euclidean and Banach Spaces, Academic Press, 3rd ed., 1973, pp. 334–335, 276–279. ■

Ejemplo 5.3. *En virtud del Teorema 5.3, podemos mejorar ahora el resultado del Ejemplo 5.2: los valores propios de la matriz* $\mathbf{A}$ *en este ejemplo pueden ser enumerados de tal forma que*

$$\lambda_i \in [i - 2\times 10^{-5}, i + 2\times 10^{-5}], \quad i = 1, 2, 3.$$

Cuando conocemos un vector propio $\mathbf{x}$ aproximado (de hecho, como tal vector podemos usar cualquier vector $\mathbf{x}$ con $\mathbf{Ax} \neq 0$), podemos usar el cuociente de Rayleigh

$$R(\mathbf{x}; \mathbf{A}) := \frac{\mathbf{x}^*\mathbf{Ax}}{\mathbf{x}^*\mathbf{x}}$$

para definir una aproximación del valor propio correspondiente, para la cual también podemos definir una inclusión.

Teorema 5.4. *Sea $\mathbf{A} \in \mathbb{C}^{n\times n}$ similar a una matriz diagonal con los valores propios $\lambda_1, \ldots, \lambda_n$. Sean $\mathbf{x} \in \mathbb{C}^n$, $\mathbf{x} \neq 0$ y $\mathbf{Ax} \neq 0$. Sea $\lambda := R(\mathbf{x}; \mathbf{A})$. Entonces*

(i) $\forall c \in \mathbb{C} : \|\mathbf{Ax} - \lambda\mathbf{x}\|_2^2 \leqslant \|\mathbf{Ax} - c\mathbf{x}\|_2^2$.

(ii) *Existe un valor $\lambda_j \neq 0$, $\lambda_j \in \sigma(\mathbf{A})$, tal que*

$$\left|\frac{\lambda_j - \lambda}{\lambda_j}\right| \leqslant \frac{\|\mathbf{Ax} - \lambda\mathbf{x}\|_2}{\|\mathbf{Ax}\|_2} \operatorname{cond}_{\|\cdot\|_2}(\mathbf{U}),$$

donde $\mathbf{U} = \begin{bmatrix}\mathbf{u}_1 & \ldots & \mathbf{u}_n\end{bmatrix}$ es un sistema completo de vectores propios de $\mathbf{A}$.

(iii) *Si $\mathbf{A}$ es normal (es decir, $\mathbf{AA}^* = \mathbf{A}^*\mathbf{A}$), entonces existe un valor $0 \leqslant \lambda_j \in \sigma(\mathbf{A})$ tal que*

$$\left|\frac{\lambda_j - \lambda}{\lambda_j}\right| \leqslant \frac{\|\mathbf{Ax} - \lambda\mathbf{x}\|_2}{\|\mathbf{Ax}\|_2}.$$

Demostración.

(i) Supongamos que $\|\mathbf{x}\|_2 = 1$. Entonces

$$\begin{aligned}\|\mathbf{Ax} - c\mathbf{x}\|_2^2 &= (\mathbf{x}^*\mathbf{A}^* - \bar{c}\mathbf{x}^*)(\mathbf{Ax} - c\mathbf{x}) \\ &= \mathbf{x}^*\mathbf{A}^*\mathbf{Ax} - \bar{c}\mathbf{x}^*\mathbf{Ax} - c\mathbf{x}^*\mathbf{A}^*\mathbf{x} + |c|^2 \\ &= \|\mathbf{Ax}\|_2^2 + |c - \mathbf{x}^*\mathbf{Ax}|^2 - |\mathbf{x}^*\mathbf{Ax}|^2 \\ &\geqslant \|\mathbf{Ax}\|_2^2 - |\mathbf{x}^*\mathbf{Ax}|^2\end{aligned}$$

con igualdad para $c = \lambda$.

(ii) Sean $\mathbf{U}^{-1}\mathbf{AU} = \operatorname{diag}(\lambda_1, \ldots, \lambda_n) =: \mathbf{\Lambda}$, $\mathbf{y} := \mathbf{U}^{-1}\mathbf{x}$ y

$$\operatorname{glb}(\mathbf{U}) := \sup\{\alpha \mid \alpha\|\mathbf{x}\| \leqslant \|\mathbf{Ux}\|\} = \frac{1}{\|\mathbf{U}^{-1}\|}.$$

Entonces

$$\begin{aligned}\frac{\|\mathbf{Ax} - \lambda\mathbf{x}\|_2}{\|\mathbf{Ax}\|_2} &= \frac{\|\mathbf{U}(\mathbf{\Lambda} - \lambda\mathbf{I})\mathbf{U}^{-1}\mathbf{x}\|_2}{\|\mathbf{U\Lambda U}^{-1}\mathbf{x}\|_2} \\ &\geqslant \frac{\operatorname{glb}(\mathbf{U})\|(\mathbf{\Lambda} - \lambda\mathbf{I})\mathbf{y}\|_2}{\|\mathbf{U}\|_2\|\mathbf{\Lambda y}\|_2}\end{aligned}$$

$$= \frac{1}{\|\mathbf{U}\|_2\|\mathbf{U}^{-1}\|_2}\left(\frac{\displaystyle\sum_{i=1}^{n}|\lambda_i-\lambda|^2|\eta_i|^2}{\displaystyle\sum_{i=1}^{n}|\lambda_i|^2|\eta_i|^2}\right)^{1/2}$$

$$= \frac{1}{\mathrm{cond}_{\|\cdot\|_2}(\mathbf{U})}\left(\frac{|\lambda|^2\displaystyle\sum_{\substack{i=1\\ \lambda_i=0}}^{n}|\eta_i|^2+\sum_{\substack{i=1\\ \lambda_i\neq 0}}^{n}\left|\frac{\lambda_i-\lambda}{\lambda_i}\right|^2|\eta_i\lambda_i|^2}{\displaystyle\sum_{\substack{i=1\\ \lambda_i\neq 0}}^{n}|\lambda_i\eta_i|^2}\right)^{1/2}$$

$$\geqslant \frac{1}{\mathrm{cond}_{\|\cdot\|_2}(\mathbf{U})}\min_{\substack{1\leqslant i\leqslant n\\ \lambda_i\neq 0}}\left|\frac{\lambda_i-\lambda}{\lambda_i}\right|.$$

(iii) Si $\mathbf{A}$ es normal, existe un sistema de vectores propios unitario, o sea, $\mathrm{cond}_{\|\cdot\|_2}(\mathbf{U})=1$. ■

Para las matrices hermitianas el cuociente de Rayleigh tiene muchas propiedades interesantes.

Teorema 5.5. *Sea $\mathbf{A}\in\mathbb{C}^{n\times n}$ hermitiana con un sistema $\mathbf{X}=\begin{bmatrix}\mathbf{x}_1 & \mathbf{x}_2 & \dots & \mathbf{x}_n\end{bmatrix}$ unitario, donde $\mathbf{A}\mathbf{x}_j=\lambda\mathbf{x}_j$ para $j=1,\dots,n$. Sea $\tilde{\mathbf{x}}\in\mathbb{C}^n$ tal que*

$$\tilde{\mathbf{x}}^*\tilde{\mathbf{x}}=1,\quad \tilde{\mathbf{x}}=\mathbf{x}_j+\sum_{k=1}^{n}\varepsilon_k\mathbf{x}_k;\quad |\varepsilon_k|\leqslant\varepsilon,\quad k=1,\dots,n.$$

Entonces

$$\left|R(\tilde{\mathbf{x}};A)-\lambda_j\right|\leqslant\sum_{\substack{i=1\\ i\neq j}}^{n}|\lambda_i-\lambda_j||\varepsilon_i|^2\leqslant 2\|\mathbf{A}\|(n-1)\varepsilon^2, \tag{5.1}$$

o sea el error en $R(\tilde{\mathbf{x}};A)$ es cuadráticamente pequeño en términos de los errores de la aproximación del vector propio.

Demostración. Utilizando que

$$\mathbf{A}=\sum_{i=1}^{n}\lambda_i\mathbf{x}_i\mathbf{x}_i^*,$$

podemos escribir

$$R(\tilde{\mathbf{x}};\mathbf{A})=\left(\mathbf{x}_j+\sum_{k=1}^{n}\varepsilon_k\mathbf{x}_k\right)^*\left(\sum_{i=1}^{n}\lambda_i\mathbf{x}_i\mathbf{x}_i^*\right)\left(\mathbf{x}_j+\sum_{k=1}^{n}\varepsilon_k\mathbf{x}_k\right)$$

$$=\sum_{i=1}^{n}\lambda_i\left[\left(\mathbf{x}_j+\sum_{k=1}^{n}\varepsilon_k\mathbf{x}_k\right)^*\mathbf{x}_i\right]\left[\mathbf{x}_i^*\left(\mathbf{x}_j+\sum_{k=1}^{n}\varepsilon_k\mathbf{x}_k\right)\right]$$

$$
\begin{aligned}
&= \sum_{i=1}^{n} \lambda_i \left(\delta_{ij} + \sum_{k=1}^{n} \bar{\varepsilon}_k \delta_{ki} \right) \left[\mathbf{x}_i^* \left(\mathbf{x}_j + \sum_{k=1}^{n} \varepsilon_k \mathbf{x}_k \right) \right] \\
&= \sum_{i=1}^{n} \lambda_i \left| \delta_{ij} + \sum_{k=1}^{n} \varepsilon_k \delta_{ki} \right|^2 \\
&= \sum_{\substack{i=1 \\ i \neq j}}^{n} \lambda_i |\varepsilon_i|^2 + \lambda_j |1 + \varepsilon_j|^2 \\
&= \sum_{\substack{i=1 \\ i \neq j}}^{n} (\lambda_i - \lambda_j) |\varepsilon_i|^2 + \lambda_j \left(|1 + \varepsilon_j|^2 + \sum_{\substack{i=1 \\ i \neq j}}^{n} |\varepsilon_i|^2 \right).
\end{aligned}
$$

Usando que

$$
|1 + \varepsilon_j|^2 + \sum_{\substack{i=1 \\ i \neq j}}^{n} |\varepsilon_i|^2 = \tilde{\mathbf{x}}^* \tilde{\mathbf{x}} = 1,
$$

llegamos a

$$
R(\tilde{\mathbf{x}}; \mathbf{A}) - \lambda_j = \sum_{\substack{i=1 \\ i \neq j}}^{n} \lambda_i |\varepsilon_i|^2 + \lambda_j |1 + \varepsilon_j|^2,
$$

lo que implica (5.1) si tomamos valores absolutos, aplicamos la desigualdad del triángulo y la cota trivial

$$
|\lambda_i - \lambda_j| \leqslant 2 r_\sigma(\mathbf{A}) \leqslant 2 \|\mathbf{A}\|.
$$

■

Teorema 5.6. *(El "principio minimax" de Courant) Sea $\mathbf{A} \in \mathbb{C}^{n \times n}$ hermitiana. Los valores propios de $\mathbf{A}$, contados con su multiplicidad, sean $\lambda_1 \geqslant \lambda_2 \geqslant \ldots \geqslant \lambda_n$. Sea $\mathcal{V}_j$ el sistema de todos los subespacios j-dimensionales de $\mathbb{C}^n$, donde definimos $\mathcal{V}_0 := \{0\}$. Entonces,*

$$
\lambda_k = \min_{V \in \mathcal{V}_{k-1}} \max \big\{ R(\mathbf{x}; \mathbf{A}) \mid \mathbf{x} \neq 0,\ \forall \mathbf{v} \in V :\ \mathbf{x}^* \mathbf{v} = 0 \big\}, \tag{5.2}
$$

$$
\lambda_k = \max_{V \in \mathcal{V}_{n-k}} \min \big\{ R(\mathbf{x}; \mathbf{A}) \mid \mathbf{x} \neq 0,\ \forall \mathbf{v} \in V :\ \mathbf{x}^* \mathbf{v} = 0 \big\}. \tag{5.3}
$$

Demostración. Sea $\mathbf{u} = \begin{bmatrix} \mathbf{u}_1 & \cdots & \mathbf{u}_n \end{bmatrix}$ un sistema unitario completo de vectores propios de $\mathbf{A}$, o sea, $\mathbf{A}\mathbf{u}_i = \lambda_i \mathbf{u}_i$, $\mathbf{U}^*\mathbf{U} = \mathbf{I}$. Si $\mathbf{x} \in \mathbb{C}^n$ es arbitrario, podemos escribir

$$
\mathbf{x} = \sum_{i=1}^{n} \gamma_i \mathbf{u}_i; \quad \gamma_i = \mathbf{u}_i^* \mathbf{x}, \quad i = 1, \ldots, n.
$$

Definiendo $\mathbf{\Lambda} = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$, podemos escribir

$$
R(\mathbf{x}; \mathbf{A}) = \frac{\mathbf{x}^* \mathbf{U} \mathbf{\Lambda} \mathbf{U}^* \mathbf{x}}{\mathbf{x}^* \mathbf{U}^* \mathbf{U} \mathbf{x}} = \sum_{i=1}^{n} \frac{|\gamma_i|^2}{|\gamma_1|^2 + \ldots + |\gamma_n|^2} \lambda_i,
$$

lo que implica que

$$R(\mathbf{x};\mathbf{A})\begin{cases}\geqslant \lambda_k & \text{si } \gamma_{k+1}=\ldots=\gamma_n=0,\\ \leqslant \lambda_k & \text{si } \gamma_1=\ldots=\gamma_{k-1}=0.\end{cases}$$

Ahora, demostramos que si $V\in\mathcal{V}_{k-1}$, existe $\tilde{\mathbf{x}}\in\mathbb{C}^n$ tal que

$$\tilde{\mathbf{x}}\neq 0,\quad \tilde{\mathbf{x}}=\sum_{i=1}^{k}\gamma_i\mathbf{u}_i,\quad \forall\mathbf{v}\in V:\ \tilde{\mathbf{x}}^*\mathbf{v}=0. \tag{5.4}$$

Para demostrar (5.4), consideremos una base ortonormalizada $\mathbf{v}_1,\ldots,\mathbf{v}_{k-1}$ de V. Sea $\mathbf{g}=(\gamma_1,\ldots,\gamma_k)^{\mathrm{T}}$, $\mathbf{g}\neq 0$ la solución de

$$\begin{bmatrix}\mathbf{v}_1^*\\ \vdots\\ \mathbf{v}_{k-1}^*\end{bmatrix}\begin{bmatrix}\mathbf{u}_1 & \cdots & \mathbf{u}_k\end{bmatrix}\mathbf{g}=0.$$

En este caso,

$$\text{máx}\big\{R(\mathbf{x};\mathbf{A})\mid \mathbf{x}\neq 0,\ \forall\mathbf{v}\in V:\ \mathbf{x}^*\mathbf{v}=0\big\}\geqslant\lambda_k. \tag{5.5}$$

Tenemos igualdad en (5.5) para $V=\text{span}\{\mathbf{u}_1,\ldots,\mathbf{u}_{k-1}\}$, es decir,

$$\mathbf{g}=\begin{pmatrix}0\\ \vdots\\ 0\\ \gamma_k\end{pmatrix},\quad \gamma_k\neq 0.$$

Si $V\in\mathcal{V}_{n-k}$, existe $\tilde{\mathbf{x}}\in\mathbb{C}^n$ tal que

$$\tilde{\mathbf{x}}\neq 0,\quad \tilde{\mathbf{x}}=\sum_{j=k}^{n}\gamma_j\mathbf{u}_j,\quad \forall\mathbf{v}\in V:\ \tilde{\mathbf{x}}^*\mathbf{v}=0. \tag{5.6}$$

Para demostrar (5.6), consideremos una base ortonormalizada $\mathbf{v}_1,\ldots,\mathbf{v}_{n-k}$ de V. Sea $\mathbf{g}=(\gamma_k,\ldots,\gamma_n)^{\mathrm{T}}$, $\mathbf{g}\neq 0$ la solución de

$$\begin{bmatrix}\mathbf{v}_1^*\\ \vdots\\ \mathbf{v}_{n-k}^*\end{bmatrix}\begin{bmatrix}\mathbf{u}_k & \cdots & \mathbf{u}_n\end{bmatrix}\mathbf{g}=0.$$

En este caso,

$$\text{mín}\big\{R(\mathbf{x};\mathbf{A})\mid \mathbf{x}\neq 0,\ \forall\mathbf{v}\in V:\ \mathbf{x}^*\mathbf{v}=0\big\}\leqslant\lambda_k. \tag{5.7}$$

Tenemos igualdad en (5.7) para $V=\text{span}\{\mathbf{u}_{k+1},\ldots,\mathbf{u}_n\}$, es decir,

$$\mathbf{g}=\begin{pmatrix}\gamma_k\\ 0\\ \vdots\\ 0\end{pmatrix},\quad \gamma_k\neq 0.$$

■

El siguiente teorema formula una consecuencia del Teorema 5.6.

Teorema 5.7. *Sean $\mathbf{A}, \mathbf{B} \in \mathbb{C}^{n\times n}$ ambas hermitianas. Sean $\lambda_i(\mathbf{A})$ y $\lambda_i(\mathbf{B})$ los valores propios de $\mathbf{A}$ y $\mathbf{B}$, enumerados tales que $\lambda_1(\mathbf{A}) \geqslant \ldots \geqslant \lambda_n(\mathbf{A})$, $\lambda_1(\mathbf{B}) \geqslant \ldots \geqslant \lambda_n(\mathbf{B})$. Entonces,*

$$\forall k = 1, \ldots, n: \quad \big|\lambda_k(\mathbf{A}) - \lambda_k(\mathbf{B})\big| \leqslant r_\sigma(\mathbf{B} - \mathbf{A}). \tag{5.8}$$

Demostración. Definir $\mathbf{B} := \mathbf{A} + (\mathbf{B} - \mathbf{A})$, $\mathbf{C} := \mathbf{B} - \mathbf{A}$, aplicar el Teorema 5.6 (Tarea). ■

Por supuesto, el Teorema 5.7 representa un resultado mucho más ventajoso que el Lema 5.1. También es válido para valores propios múltiples. Por supuesto, la restricción que ambas matrices deben ser hermitianas es muy fuerte. Además, tenemos también que los valores propios de una matriz diagonalizable dependen de forma Lipschitz continua de los coeficientes:

Teorema 5.8. *Si $\mathbf{A} \in \mathbb{C}^{n\times n}$ es diagonalizable, es decir, existe un sistema $\mathbf{U} = \begin{bmatrix}\mathbf{u}_1 & \cdots & \mathbf{u}_n\end{bmatrix}$ de vectores propios de $\mathbf{A}$, y $\mathbf{B} \in \mathbb{C}^{n\times n}$ es arbitraria, entonces para cada valor propio $\lambda_j(\mathbf{B})$ existe un valor propio $\lambda_{i(j)}(\mathbf{A})$ tal que*

$$\big|\lambda_{i(j)}(\mathbf{A}) - \lambda_j(\mathbf{B})\big| \leqslant \operatorname{cond}_{\|\cdot\|_\infty}(\mathbf{U})\|\mathbf{B} - \mathbf{A}\|_\infty. \tag{5.9}$$

Demostración. Nos restringimos al caso no trivial. Sea $\lambda(\mathbf{B}) \in \sigma(\mathbf{B})$, $\lambda(\mathbf{B}) \neq \lambda_i(\mathbf{A})$ para $i = 1, \ldots, n$, con el vector propio $\mathbf{x} \neq 0$, es decir, $\mathbf{B}\mathbf{x} = \lambda(\mathbf{B})\mathbf{x}$. Entonces

$$\mathbf{B}\mathbf{x} - \mathbf{A}\mathbf{x} = \lambda(\mathbf{B})\mathbf{x} - \mathbf{A}\mathbf{x} \Longleftrightarrow \mathbf{x} = \big(\lambda(\mathbf{B})\mathbf{I} - \mathbf{A}\big)^{-1}(\mathbf{B} - \mathbf{A})\mathbf{x}$$

implica que

$$\|\mathbf{x}\|_\infty \leqslant \big\|\big(\lambda(\mathbf{B})\mathbf{I} - \mathbf{A}\big)^{-1}\big\|_\infty \|\mathbf{B} - \mathbf{A}\|_\infty \|\mathbf{x}\|_\infty.$$

Ahora, usando $\mathbf{I} = \mathbf{U}\mathbf{U}^{-1}$ y $\mathbf{A} = \mathbf{U}\mathbf{\Lambda}_\mathbf{A}\mathbf{U}^{-1}$, donde $\mathbf{\Lambda}_\mathbf{A}$ es una matriz diagonal de los valores propios de $\mathbf{A}$, tenemos que

$$\|\mathbf{x}\|_\infty \leqslant \frac{1}{\min\limits_{1\leqslant i\leqslant n} |\lambda(\mathbf{B}) - \lambda_i(\mathbf{A})|}\|\mathbf{B} - \mathbf{A}\|_\infty \|\mathbf{x}\|_\infty.$$

■

No existe un resultado análogo para matrices *no* diagonalizables. El siguiente ejemplo ilustra que es muy natural que un tal resultado no existe.

Ejemplo 5.4. *La matriz*

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

posee los valores propios $\lambda_1 = \lambda_2 = 0$, y no es diagonalizable. La matriz

$$\mathbf{B} = \begin{bmatrix} 0 & 1 \\ \varepsilon & 0 \end{bmatrix}$$

posee los valores propios $\pm\sqrt{\varepsilon}$, mientras que en cualquier norma matricial, $\|\mathbf{A} - \mathbf{B}\| \leqslant C\varepsilon$ (con una constante C que depende de la norma).

Finalmente, nos interesan resultados asintóticos sobre los valores y vectores propios. El siguiente teorema, que se comunica sin demostración, representa un resultado típico.

Teorema 5.9. *Sea* $\mathbf{A} \in \mathbb{C}^{n\times n}$ *diagonalizable,* $\mathbf{X} = \begin{bmatrix}\mathbf{x}_1 & \cdots & \mathbf{x}_n\end{bmatrix}$ *un sistema completo de vectores propios de* $\mathbf{A}$*,* $\mathbf{A}\mathbf{x}_i = \lambda_i \mathbf{x}_i$ *y* $\|\mathbf{x}_i\|_2 = 1$ *para* $i = 1, \ldots, n$*. Ademas definimos*

$$\mathbf{X}^{-1} =: \mathbf{Y} =: \begin{bmatrix}\mathbf{y}_1^* \\ \vdots \\ \mathbf{y}_n^*\end{bmatrix},$$

o sea, $\mathbf{y}_1^*, \ldots, \mathbf{y}_n^*$ *son los vectores propios de la izquierda de* $\mathbf{A}$*:* $\mathbf{y}_i^*\mathbf{A} = \mathbf{y}_i^*\lambda_i$ *para* $i = 1, \ldots, n$*. Sea* λ_j *un valor propio simple de* $\mathbf{A}$*. Entonces, para* $\mathbf{F} \in \mathbb{C}^{n\times n}$ *con* $\|\mathbf{F}\|_2$ *suficientemente pequeña existe un valor propio* μ_j *de* $\mathbf{A} + \mathbf{F}$ *con un vector propio* $\mathbf{z}_j$*,* $\|\mathbf{z}_j\|_2 = 1$*, tal que*

$$\mu_j = \lambda_j + \frac{\mathbf{y}_j^*\mathbf{F}\mathbf{x}_j}{\|\mathbf{y}_j\|_2\|\mathbf{x}_j\|_2} \cdot \frac{\|\mathbf{y}_j\|_2\|\mathbf{x}_j\|_2}{\mathbf{y}_j^*\mathbf{x}_j} + \mathcal{O}\big(\|\mathbf{F}\|_2^2\big),$$

$$\mathbf{z}_j = \mathbf{x}_j + \left(\sum_{\substack{i=1\\ i\neq j}}^{n} \frac{\mathbf{y}_j^*\mathbf{F}\mathbf{x}_i}{\|\mathbf{y}_j\|_2\|\mathbf{x}_j\|_2} \frac{1}{\lambda_i - \lambda_j}\mathbf{x}_i\right) \frac{\|\mathbf{y}_j\|_2\|\mathbf{x}_j\|_2}{\mathbf{y}_j^*\mathbf{x}_j} + \mathcal{O}\big(\|\mathbf{F}\|_2^2\big).$$

Demostración. Se usa el Teorema de Funciones Implícitas para el problema $\mathbf{g}(\mathbf{x}, \lambda, \varepsilon) = 0$ con

$$\mathbf{g}(\mathbf{x}_j, \lambda_j, 0) = 0, \quad \mathbf{g}(\mathbf{x}, \lambda, \varepsilon) = \begin{pmatrix}(\mathbf{A} + \varepsilon\mathbf{F}_0)\mathbf{x} - \lambda\mathbf{x} \\ \mathbf{x}^{\mathrm{T}}\mathbf{x} - 1\end{pmatrix}, \quad \mathbf{F} = \varepsilon\mathbf{F}_0,$$

con ε en una vecindad apropiada de cero, y se representa la solución $(\mathbf{x}, \lambda)$ como función de ε. ∎

Obviamente, el factor de amplificación del error decisivo es $\|\mathbf{y}_j\|_2\|\mathbf{x}_j\|_2/|\mathbf{y}_j^*\mathbf{x}_j|$ para un valor propio (este factor puede ser grande para matrices no normales), mientras que para un vector propio, tambien juega un rol importante *la separación* de los valores propios.

5.2. Transformación de similaridad unitaria de una matriz $n \times n$ a una forma de Hessenberg o tridiagonal

La solución del problema de valores propios para una matriz no esparsa siempre empieza con la transformación de la matriz a una forma "condensada". Esa transformación genera nueavos errores de redondeo. Para que la matriz transformada aun sea usable, los valores propios no deben ser más falsificados que como si la matriz fuera modificada dentro de la exactitud aritmética. Por lo tanto, sólo es practicable la transformación a la forma de Hessenberg. La transformación parte de la matriz $\mathbf{A}$, luego determinamos matrices unitarias y hermitianas $\mathbf{U}_1, \ldots, \mathbf{U}_{n-2}$ tales que $\mathbf{U}_{n-2} \cdot \ldots \cdot \mathbf{U}_1\mathbf{A}\mathbf{U}_1 \cdot \ldots \cdot \mathbf{U}_{n-2}$ es una matriz del tipo Hessenberg. Ahora, si la matriz $\mathbf{A}$ es hermitiana, entonces

$$(\mathbf{U}_{n-2} \cdot \ldots \cdot \mathbf{U}_1\mathbf{A}\mathbf{U}_1 \cdot \ldots \cdot \mathbf{U}_{n-2})^* = \mathbf{U}_{n-2} \cdot \ldots \cdot \mathbf{U}_1\mathbf{A}\mathbf{U}_1 \cdot \ldots \cdot \mathbf{U}_{n-2},$$

o sea la matriz del tipo Hessenberg es hermitiana y por lo tanto, tridiagonal.

La transfomacion procede en $n-2$ pasos. Supongamos que después de $j-1$ pasos, tenemos la matriz

$$\mathbf{A}_j = \mathbf{U}_{j-1} \cdot \ldots \cdot \mathbf{U}_1 \mathbf{A} \mathbf{U}_1 \cdot \ldots \cdot \mathbf{U}_{j-1} = \begin{bmatrix} \alpha_{11}^{(j)} & \alpha_{12}^{(j)} & \cdots & \alpha_{1,j-1}^{(j)} & \alpha_{1j}^{(j)} & \cdots & \alpha_{1n}^{(j)} \\ \alpha_{21}^{(j)} & \alpha_{22}^{(j)} & \cdots & \alpha_{2,j-1}^{(j)} & \alpha_{2j}^{(j)} & \cdots & \alpha_{2n}^{(j)} \\ 0 & \alpha_{32}^{(j)} & & \vdots & \vdots & & \vdots \\ \vdots & 0 & \ddots & \alpha_{j,j-1}^{(j)} & \alpha_{jj}^{(j)} & & \alpha_{jn}^{(j)} \\ \vdots & \vdots & \ddots & 0 & \alpha_{j+1,j}^{(j)} & & \vdots \\ \vdots & \vdots & & \vdots & \vdots & & \vdots \\ 0 & \cdots & \cdots & 0 & \alpha_{nj}^{(j)} & & \alpha_{nn}^{(j)} \end{bmatrix}.$$

Si definimos $\mathbf{A}_{j+1} := \mathbf{U}_j \mathbf{A}_j \mathbf{U}_j$, tenemos (ver Teorema 3.6)

$$\mathbf{U}_j = \begin{bmatrix} \mathbf{I} & 0 \\ 0 & \hat{\mathbf{U}}_j \end{bmatrix}, \quad \hat{\mathbf{U}}_j = \mathbf{I} - \beta_k \hat{\mathbf{w}}_j \hat{\mathbf{w}}_j^*,$$

donde se determina $\hat{\mathbf{U}}_j$ de tal forma que

$$\hat{\mathbf{U}}_j \begin{pmatrix} \alpha_{j+1,j}^{(j)} \\ \vdots \\ \alpha_{nj}^{(j)} \end{pmatrix} = -\exp(\mathrm{i}\varphi)\sigma_j \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

El Teorema 3.6 entrega las fórmulas

$$\hat{\mathbf{w}}_j = \begin{pmatrix} \exp(\mathrm{i}\varphi_j)\big(\big|\alpha_{j+1,j}^{(j)}\big| + \sigma_j\big) \\ \alpha_{j+2,j}^{(j)} \\ \vdots \\ \alpha_{nj}^{(j)} \end{pmatrix}, \quad \sigma_j = \left(\sum_{k=j+1}^{n} \big|\alpha_{kj}^{(j)}\big|^2 \right)^{1/2},$$

$$\alpha_{j+1,j}^{(j)} = \exp(\mathrm{i}\varphi_j)\big|\alpha_{j+1,j}^{(j)}\big|, \quad \beta_j = \frac{1}{\sigma_j(\sigma_j + |\alpha_{j+1,j}^{(j)}|)}.$$

Puesto que las primeras j columnas de $\mathbf{U}_j$ son columnas unitarias, la multiplicación de $\mathbf{U}_j \mathbf{A}_j$ desde la derecha con $\mathbf{U}_j$ no cambia los ceros recien generados en la columna j. La multiplicación de $\mathbf{A}_j$ desde la izquierda por $\mathbf{U}_j$ no cambia las primeras j filas. Por lo tanto, hemos desarrollado completamente la transformación.

Para la ejecución práctica de la transformación, aprovechamos la estructura especial de $\mathbf{U}_j$. Sea

$$\mathbf{A}_j = \begin{bmatrix} \mathbf{A}_{11}^{(j)} & \mathbf{A}_{12}^{(j)} \\ \mathbf{A}_{21}^{(j)} & \mathbf{A}_{22}^{(j)} \end{bmatrix}$$

En este caso, obtenemos

$$\mathbf{A}_{j+1} = \begin{bmatrix} \mathbf{A}_{11}^{(j)} & \mathbf{A}_{12}^{(j)}\hat{\mathbf{U}}_j \\ \hat{\mathbf{U}}_j\mathbf{A}_{21}^{(j)} & \hat{\mathbf{U}}_j\mathbf{A}_{22}^{(j)}\hat{\mathbf{U}}_j \end{bmatrix},$$

donde

$$\begin{aligned} \hat{\mathbf{U}}_j\mathbf{A}_{21}^{(j)} &= \mathbf{A}_{21}^{(j)} - \beta_j\mathbf{w}_j\mathbf{w}_j^*\mathbf{A}_{21}^{(j)} \\ &= \mathbf{A}_{21}^{(j)} - \hat{\mathbf{u}}_j\hat{\mathbf{z}}_j^*, \\ \mathbf{A}_{12}^{(j)}\hat{\mathbf{U}}_j &= \mathbf{A}_{12}^{(j)} - \beta_j\mathbf{A}_{12}^{(j)}\mathbf{w}_j\mathbf{w}_j^* \\ &= \mathbf{A}_{12}^{(j)} - \hat{\mathbf{y}}_j\hat{\mathbf{u}}_j^*, \\ \hat{\mathbf{U}}_j\mathbf{A}_{22}^{(j)}\hat{\mathbf{U}}_j &= (\mathbf{I} - \beta_j\hat{\mathbf{w}}_j\hat{\mathbf{w}}_j^*)\big(\mathbf{A}_{22}^{(j)} - \beta_j\mathbf{A}_{22}^{(j)}\hat{\mathbf{w}}\hat{\mathbf{w}}^*\big) \\ &= \mathbf{A}_{22}^{(j)} - \hat{\mathbf{u}}_j\big(\hat{\mathbf{w}}_j^*\mathbf{A}_{22}^{(j)}\big) - \big(\mathbf{A}_{22}^{(j)}\hat{\mathbf{w}}_j\big)\hat{\mathbf{u}}_j^* + \big(\hat{\mathbf{w}}_j^*\mathbf{A}_{22}^{(j)}\hat{\mathbf{w}}_j\big)\hat{\mathbf{u}}_j\hat{\mathbf{u}}_j^* \\ &= \mathbf{A}_{22}^{(j)} - \hat{\mathbf{u}}_j\left(\hat{\mathbf{s}}_j^* - \frac{\gamma_j}{2}\hat{\mathbf{u}}_j^*\right) - \left(\hat{\mathbf{t}}_j - \frac{\gamma_j}{2}\hat{\mathbf{u}}_j\right)\hat{\mathbf{u}}_j^*, \end{aligned}$$

donde definimos

$$\hat{\mathbf{u}}_j := \beta_j\hat{\mathbf{w}}_j, \quad \hat{\mathbf{t}}_j := \mathbf{A}_{22}^{(j)}\hat{\mathbf{w}}_j, \quad \gamma_j := \hat{\mathbf{w}}_j^*\hat{\mathbf{t}}_j, \quad \hat{\mathbf{s}}_j^* := \hat{\mathbf{w}}_j^*\mathbf{A}_{22}^{(j)}, \quad \hat{\mathbf{z}}_j^* := \hat{\mathbf{w}}_j^*\mathbf{A}_{21}^{(j)}, \quad \hat{\mathbf{y}}_j := \mathbf{A}_{12}^{(j)}\hat{\mathbf{w}}_j.$$

Pero, según hipótesis,

$$\mathbf{A}_{21}^{(j)} = \begin{bmatrix} 0 & \cdots & 0 & \alpha_{j+1,j}^{(j)} \\ \vdots & & \vdots & \vdots \\ 0 & \cdots & 0 & \alpha_{nj}^{(j)} \end{bmatrix},$$

de manera que no hay que calcular $\hat{\mathbf{U}}_j\mathbf{A}_{21}^{(j)}$ explícitamente:

$$\hat{\mathbf{U}}_j\mathbf{A}_{21}^{(j)} = \begin{bmatrix} 0 & \cdots & 0 & -\exp(\mathrm{i}\varphi_j)\sigma_j \\ \vdots & & \vdots & 0 \\ \vdots & & \vdots & \vdots \\ 0 & \cdots & 0 & 0 \end{bmatrix}.$$

Además, en el caso $\mathbf{A}$ hermitiana tomamos en cuenta que

$$\mathbf{A}_{12}^{(j)}\hat{\mathbf{U}}_j = \big(\hat{\mathbf{U}}_j\mathbf{A}_{21}^{(j)}\big)^*, \quad \hat{\mathbf{t}}_j = \hat{\mathbf{s}}_j,$$

entonces una computación explícita no es necesaria, y el esfuerzo computacional total se reduce a menos que la mitad. Para $\mathbf{A}$ general, necesitamos $\frac{5}{3}n^3 + \mathcal{O}(n^2)$ operaciones; para $\mathbf{A}$ hermitiana, solo $\frac{2}{3}n^3 + \mathcal{O}(n^2)$ operaciones esenciales.

5.3. Computación de los valores propios de una matriz tridiagonal hermitiana

Consideremos la matriz $\mathbf{T} \in \mathbb{C}^{n\times n}$ dada por

$$\mathbf{T} = \begin{bmatrix} \alpha_1 & \beta_1 & & \\ \gamma_1 & \ddots & \ddots & \\ & \ddots & \ddots & \beta_{n-1} \\ & & \gamma_{n-1} & \alpha_n \end{bmatrix}, \quad \gamma_i = \bar{\beta}_i, \quad i = 1, \ldots, n-1; \quad \alpha_i \in \mathbb{R}, \quad i = 1, \ldots, n. \tag{5.10}$$

Se supone que $\gamma_i \neq 0$ para $i = 1, \ldots, n-1$, sino la matriz tridiagonal puede ser particionada en dos matrices tridiagonales de tamaño menor cuyos problemas de valores propios pueden ser estudiados separadamente.

Teorema 5.10. *Si* $\mathbf{T}$ *es una matriz tridiagonal hermitiana de la forma indicada en* (5.10) *y* $\gamma_i \neq 0$ *para* $i = 1, \ldots, n-1$*, entonces* $\mathbf{T}$ *sólo tiene valores propios reales simples.*

Demostración. Tarea. ■

Comentamos si $\mathbf{T}$ es una matriz real *no* simétrica con $\beta_i\gamma_i > 0$, $i = 1, \ldots, n-1$, entonces mediante una transformación de similaridad con

$$\mathbf{D} = \operatorname{diag}(\delta_1, \ldots, \delta_n), \quad \delta_1 := 1, \quad \delta_{i+1} := \delta_i\sqrt{\beta_i/\gamma_i},$$

$\mathbf{T}$ puede ser transformada a una matriz simétrica y tridiagonal $\hat{\mathbf{T}} := \mathbf{DTD}^{-1}$. Entonces tales matrices $\mathbf{T}$ también poseen solo valores propios reales y simples.

Para la computación de valores propios de $\mathbf{T}$, necesitamos el Teorema de la Ley de Inercia de Sylvester.

Definición 5.1. *Sea* $\mathbf{A} \in \mathbb{C}^{n\times n}$ *hermitiana. Entonces se define como* inercia *de* $\mathbf{A}$ *al triple* (m, z, p)*, donde* m*,* z *y* p *es el número de valores propios negativos, zero, y positivos, respectivamente.*

Teorema 5.11 (Ley de Inercia de Sylvester)**.** *Si* $\mathbf{A} \in \mathbb{C}^{n\times n}$ *es hermitiana y* $\mathbf{X} \in \mathbb{C}^{n\times n}$ *es regular, entonces* $\mathbf{A}$ *y* $\mathbf{X}^*\mathbf{AX}$ *tienen la misma inercia.*

Demostración. Supongamos que

$$\lambda_1(\mathbf{A}) \geqslant \lambda_2(\mathbf{A}) \geqslant \ldots \geqslant \lambda_n(\mathbf{A})$$

son los valores propios de $\mathbf{A}$, contados con su multiplicidad, y que para algún $r \in \{1, \ldots, n\}$, $\lambda_r(\mathbf{A})$ es un valor propio de $\mathbf{A}$ positivo. Definimos el subespacio $S_0 \subseteq \mathbb{R}^n$ a través de

$$S_0 := \operatorname{span}\{\mathbf{X}^{-1}\mathbf{q}_1, \ldots, \mathbf{X}^{-1}\mathbf{q}_r\}, \quad \mathbf{q}_1 \neq 0, \ldots, \mathbf{q}_r \neq 0,$$

donde $\mathbf{Aq}_i = \lambda_i(\mathbf{A})\mathbf{q}_i$ para $i = 1, \ldots, r$. Utilizando la caracterización minimax de $\lambda_r(\mathbf{X}^*\mathbf{AX})$, donde se supone que

$$\lambda_1(\mathbf{X}^*\mathbf{AX}) \geqslant \ldots \geqslant \lambda_n(\mathbf{X}^*\mathbf{AX})$$

son los valores propios de $\mathbf{X}^*\mathbf{AX}$, obtenemos

$$\lambda_r(\mathbf{X}^*\mathbf{AX}) = \max_{V\in\mathcal{V}_{n-r}} \min\big\{R(\mathbf{x}; \mathbf{X}^*\mathbf{AX}) \mid \mathbf{x} \neq 0,\ \forall \mathbf{v} \in V :\ \mathbf{x}^*\mathbf{v} = 0\big\}. \tag{5.11}$$

Ahora, escogiendo

$$\tilde{V} := S_0^{\perp} := \big\{\mathbf{w} \in \mathbb{R}^n \mid \forall \mathbf{v} \in S_0 : \ \mathbf{w}^*\mathbf{v} = 0\big\} \in \mathcal{V}_{n-r},$$

deducimos de (5.11) que

$$\begin{aligned}\lambda_r(\mathbf{X}^*\mathbf{A}\mathbf{X}) &\geqslant \text{mín}\big\{R(\mathbf{x};\mathbf{X}^*\mathbf{A}\mathbf{X}) \mid \mathbf{x} \neq 0,\ \forall \mathbf{v} \in \tilde{V} : \ \mathbf{x}^*\mathbf{v} = 0\big\} \\ &= \text{mín}\big\{R(\mathbf{x};\mathbf{X}^*\mathbf{A}\mathbf{X}) \mid \mathbf{x} \neq 0,\ \mathbf{x} \in S_0\big\} \\ &\geqslant \lambda_r(\mathbf{A}).\end{aligned}$$

Si $\sigma_1(\mathbf{X}) \geqslant \ldots \geqslant \sigma_n(\mathbf{X})$ son los valores singulares de $\mathbf{X}$, podemos demostrar que para cada $\mathbf{y} \in \mathbb{R}^n$,

$$R(\mathbf{y}, \mathbf{X}^*\mathbf{X}) \geqslant \sigma_n(\mathbf{X})^2.$$

Entonces, conluimos que

$$\lambda_r(\mathbf{X}^*\mathbf{A}\mathbf{X}) \geqslant \min_{\mathbf{y}\in S_0} \left\{ \frac{\mathbf{y}^*(\mathbf{X}^*\mathbf{A}\mathbf{X})\mathbf{y}}{\mathbf{y}^*(\mathbf{X}^*\mathbf{X})\mathbf{y}} \frac{\mathbf{y}^*(\mathbf{X}^*\mathbf{X})\mathbf{y}}{\mathbf{y}^*\mathbf{y}} \right\} \geqslant \lambda_r(\mathbf{A})\sigma_n(\mathbf{X})^2. \tag{5.12}$$

Un argumento análogo, con los roles de $\mathbf{A}$ y $\mathbf{X}^*\mathbf{A}\mathbf{X}$ intercambiados muestra que

$$\lambda_r(\mathbf{A}) \geqslant \lambda_r(\mathbf{X}^*\mathbf{A}\mathbf{X})\sigma_n(\mathbf{X}^{-1})^2 = \frac{\lambda_r(\mathbf{X}^*\mathbf{A}\mathbf{X})}{\sigma_1(\mathbf{X})^2}. \tag{5.13}$$

Combinando (5.12) y (5.13), concluimos que $\lambda_r(\mathbf{A})$ y $\lambda_r(\mathbf{X}^*\mathbf{A}\mathbf{X})$ tienen el mismo signo, por lo tanto, $\mathbf{A}$ y $\mathbf{X}^*\mathbf{A}\mathbf{X}$ tienen el mismo número de valores propios positivos. Aplicando este resultado a $-\mathbf{A}$, concluimos que $\mathbf{A}$ y $\mathbf{X}^*\mathbf{A}\mathbf{X}$ tienen el mismo número de valores propios negativos, y obviamente, el mismo número de valores propios zero (debidamente contados con su multiplicidad). ■

El Teorema 5.11 implica que las matrices $\mathbf{A} - \mu\mathbf{I}$ y $\mathbf{X}^*(\mathbf{A} - \mu\mathbf{I})\mathbf{X}$ tienen los mismos números de valores propios positivos, cero, y negativos, es decir, $\mathbf{A}$ tiene los mismos números de valores propios $> \mu$, $= \mu$, y $< \mu$ ($\mu \in \mathbb{R}$). Queremos aplicar este resultado ahora a la matriz $\mathbf{T}$ con

$$\mathbf{X} \in \mathbb{C}^{n\times n}, \quad \text{donde } \mathbf{X}^{-1} = \begin{bmatrix} 1 & \xi_1 & & \\ & \ddots & \ddots & \\ & & \ddots & \xi_{n-1} \\ & & & 1 \end{bmatrix},$$

donde se debe cumplir que

$$\mathbf{X}^*(\mathbf{T} - \mu\mathbf{I})\mathbf{X} = \mathbf{Q} = \text{diag}(q_1, \ldots, q_n), \quad q_i \in \mathbb{R},$$

o sea

$$\mathbf{T} - \mu\mathbf{I} = (\mathbf{X}^*)^{-1}\mathbf{Q}\mathbf{X}^{-1} = \begin{bmatrix} 1 & & & \\ \bar{\xi}_1 & \ddots & & \\ & \ddots & \ddots & \\ & & \bar{\xi}_{n-1} & 1 \end{bmatrix} \begin{bmatrix} q_1 & & & \\ & q_2 & & \\ & & \ddots & \\ & & & q_n \end{bmatrix} \begin{bmatrix} 1 & \xi_1 & & \\ & \ddots & \ddots & \\ & & \ddots & \xi_{n-1} \\ & & & 1 \end{bmatrix}.$$

Entonces, los q_i son cuocientes de subdeterminantes principales sucesivos de $\mathbf{T} - \mu\mathbf{I}$. Obviamente,

$$\begin{aligned} q_1 &= \alpha_1 - \mu, \\ q_1\xi_1 &= \beta_1 \quad (\Longrightarrow q_1\bar{\xi}_1 = \gamma_1 = \bar{\beta}_1), \\ q_2 + q_1|\xi_1|^2 &= \alpha_2 - \mu, \\ q_2\xi_2 &= \beta_2, \quad \text{etc.} \end{aligned}$$

En general, tenemos

$$\begin{gathered} q_k + q_{k-1}|\xi_{k-1}^2| = \alpha_k - \mu, \quad q_k\ \xi_k = \beta_k, \quad k = 1, \ldots, n, \\ \xi_0 := 0, \quad q_0 := 1, \quad \beta_n := 0. \end{gathered}$$

Dado que $\beta_k \neq 0$ para $k = 1, \ldots, n-1$, el valor ξ_k existe para $q_k \neq 0$, $k = 1, \ldots, n-1$, es decir, $\xi_k = \beta_k/q_k$ para $k = 1, \ldots, n-1$. Si sucede que $q_k = 0$, remplazamos este valor por $\varepsilon \ll 1$, es decir remplazamos α_k por $\alpha_k + \varepsilon$. Debido al Teorema 4.7, eso cambia los valores propios sólo en ε. Entonces, siempre se calcula

$$q_k = \alpha_k - \mu - \frac{(\beta_{k-1})^2}{q_{k-1}}, \quad k = 1, \ldots, n, \quad q_0 := 1, \quad \beta_0 := 0. \tag{5.14}$$

Según el Teorema 5.11, sabemos que

$$\#\{k \mid q_k < 0,\ 1 \leqslant k \leqslant n\} = \#\{\lambda \mid \lambda \text{ es valor propio de } \mathbf{T},\ \lambda < \mu\}.$$

Este resultado lo podemos aprovechar directamente para crear un *método de bisección* para calcular valores propios arbitrarios λ_j de $\mathbf{T}$. Partiendo de la enumeración $\lambda_1 \leqslant \lambda_2 \leqslant \ldots \leqslant \lambda_n$ y, por ejemplo, de la inclusión trivial

$$[a_0, b_0] := [-\|\mathbf{T}\|_\infty, \|\mathbf{T}\|_\infty], \tag{5.15}$$

la cual incluye todos los valores propios de $\mathbf{T}$, ponemos para $s \in \mathbb{N}_0$:

$$\mu_s := \frac{a_s + b_s}{2}, \tag{5.16}$$

$$m := \#\{q_k \mid q_k < 0, \text{calculados de (5.14) con } \mu = \mu_s\}, \tag{5.17}$$

$$a_{s+1} := \begin{cases} a_s & \text{si } m \geqslant j, \\ \mu_s & \text{sino,} \end{cases} \qquad b_{s+1} := \begin{cases} \mu_s & \text{si } m \geqslant j, \\ b_s & \text{sino.} \end{cases} \tag{5.18}$$

Para este método sabemos que

$$\lim_{s\to\infty} \mu_s = \lambda_j. \tag{5.19}$$

Este método es muy robusto y extremadamente eficiente.

Ejemplo 5.5. *Queremos determinar el valor propio λ_2 de*

$$\mathbf{T} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 3 & 2 & 0 \\ 0 & 2 & 5 & 3 \\ 0 & 0 & 3 & 7 \end{bmatrix},$$

s	μ	q_1	q_2	q_3	q_4	m	
0	1,5	−0,500000	3,500000	2,357143	1,681818	1	$\Rightarrow \lambda_2 > 1{,}5$
1	1,75	−0,750000	2,583333	1,701613	−0,039100	2	$\Rightarrow \lambda_2 < 1{,}75$
2	1,625	−0,625000	2,975000	2,030462	0,942511	1	$\Rightarrow \lambda_2 > 1{,}625$
3	1,6875	−0,687500	2,767045	1,866915	0,491712	1	$\Rightarrow \lambda_2 > 1{,}6875$
4	1,71875	−0,718750	2,672554	1,784555	0,237975	1	$\Rightarrow \lambda_2 > 1{,}71875$
5	1,734375	−0,734375	2,627327	1,743165	0,102603	1	$\Rightarrow \lambda_2 > 1{,}734375$
6	1,7421875	−0,742187	2,605181	1,722410	0,032577	1	$\Rightarrow \lambda_2 > 1{,}7421875$
7	1,74609375	−0,746094	2,594220	1,712017	−0,003050	2	$\Rightarrow \lambda_2 < 1{,}74609375$
8	1,744140625	−0,744141	2,599691	1,717215	0,014816	1	$\Rightarrow \lambda_2 > 1{,}744140625$

CUADRO 5.1. Ejemplo 5.5 (método de bisección).

empezando con $[a_0, b_0] := [1, 2]$. *El método de bisección entrega la información del Cuadro 5.1.*

Ejemplo 5.6 (Tarea 30, Curso 2006)**.** *Se considera la matriz*

$$\mathbf{A} = \begin{bmatrix} 10 & -6 & 8 \\ -6 & 17 & 2 \\ 8 & 2 & 20 \end{bmatrix}.$$

a) *Transformar* $\mathbf{A}$ *a forma tridiagonal y aplicar el método de bisección para demostrar que* $\mathbf{A}$ *es definida positiva.*

b) *Usando el método de bisección, determinar el valor propio más pequeño hasta un error del valor absoluto* $\leqslant 0{,}5$.

Solución sugerida.

a) *La transformación de la matriz a forma tridiagonal necesita un paso, es decir* $\mathbf{T} = \mathbf{A}_2 = \mathbf{P}_1\mathbf{A}_1\mathbf{P}_1$ *con* $\mathbf{A}_1 = \mathbf{A}$. *Sabemos que*

$$\mathbf{P}_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & & \\ 0 & & \hat{\mathbf{P}}_1 \end{bmatrix}, \quad \hat{\mathbf{P}}_1 = \mathbf{I} - \beta_1 \hat{\mathbf{w}}_1 \hat{\mathbf{w}}_1^*.$$

Aquí

$$\sigma_1 = \sqrt{36 + 64} = 10, \quad \beta_1 = \frac{1}{10(10+6)} = \frac{1}{160}, \quad \hat{\mathbf{w}}_1 = \begin{pmatrix} -16 \\ 8 \end{pmatrix},$$

$$\mathbf{P}_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & -0{,}6 & 0{,}8 \\ 0 & 0{,}8 & 0{,}6 \end{bmatrix}, \quad \mathbf{T} = \mathbf{P}_1\mathbf{A}\mathbf{P}_1 = \begin{bmatrix} 10 & 10 & 0 \\ 10 & 17 & 2 \\ 0 & 2 & 20 \end{bmatrix}.$$

Aplicando el Teorema de Gershgorin, vemos que $\mathbf{T}$ *solo tiene valores propios no negativos; dado que* $\mathbf{T}$ *es regular,* 0 *no es valor propio; entonces los valores propios de* $\mathbf{T}$ *(y los de* $\mathbf{A}$*) son positivos.*

b) *El valor propio mas pequeño es* λ_1*, entonces* $j = 1$. *El valor propio esta contenido en el intervalo* $[a_0 := 0, b_0 := 32]$. *Obtenemos la siguiente tabla.*

k	a_k	b_k	μ_k	q_1	q_2	q_3	m
0	0	32	16	-6	$16.\bar{6}$	3,76	1
1	0	16	8	2	-41	12,097	1
2	0	8	4	6	$-3.\bar{6}$	$17.0\bar{9}$	1
3	0	4	2	8	2,5	16,4	0
4	2	4	3	7	$-2/7$	31	1
5	2	3	2,5	7,5	$1,1\bar{6}$	13,5	0

Entonces sabemos que $\lambda_1 \in [2{,}5, 3]$.

Ejemplo 5.7 (Certamen 2, Curso 2010)**.** *Se considera la matriz*

$$\mathbf{A} = \begin{bmatrix} -10 & 3 & -4 \\ 3 & 2 & 1 \\ -4 & 1 & 16 \end{bmatrix}.$$

a) *Demostrar sin calcular el polinomio característico que* $\mathbf{A}$ *tiene tres valores propios reales distintos.*
b) *Transformar* $\mathbf{A}$ *unitariamente a forma tridiagonal simétrica.*
c) *Determinar números* α_i, β_i, $i = 1, 2, 3$, *tales que* $\alpha_i \leqslant \lambda_i \leqslant \beta_i$, $i = 1, 2, 3$, *donde* $\lambda_1, \lambda_2, \lambda_3$ *son los valores propios de* $\mathbf{A}$, *y* $|\beta_i - \alpha_i| \leqslant 0{,}25$, *mediante el método de bisección.*

Solución sugerida.

a) *Puesto que* $\mathbf{A}$ *es simétrica, sus valores propios son reales. Los círculos de Gershgorin son*

$$\mathcal{K}_1 = [-17, -3], \quad \mathcal{K}_2 = [-2, 6], \quad \mathcal{K}_3 = [11, 21].$$

Dado que $\mathcal{K}_i \cap \mathcal{K}_j = \varnothing$ *para* $i \neq j$, *cada uno de los círculos contiene exactamente un valor propio, es decir* $\lambda_i \in \mathcal{K}_i$ *para* $i = 1, 2, 3$.

b) *Siguiendo el procedimiento canónico, determinamos*

$$\mathbf{U}_1 = \begin{bmatrix} 1 & 0 \\ 0 & \hat{\mathbf{U}}_1 \end{bmatrix} \in \mathbb{R}^{3\times 3}$$

tal que $\hat{\mathbf{U}}_1 = \mathbf{I} - \beta_1 \hat{\mathbf{w}}\hat{\mathbf{w}}^{\mathrm{T}} \in \mathbb{R}^{2\times 2}$*; con*

$$\sigma_1 = \sqrt{s^2 + (-4)^2} = 5$$

se tiene aquí

$$\hat{\mathbf{w}} = \begin{pmatrix} 3 + \sigma_1 \\ 4 \end{pmatrix} = \begin{pmatrix} 8 \\ -4 \end{pmatrix}, \quad \beta_1 = \frac{1}{5 \cdot (5+3)} = \frac{1}{40}; \quad \hat{\mathbf{U}}_1 = \frac{1}{5}\begin{bmatrix} -3 & 4 \\ 4 & 3 \end{bmatrix}$$

y la matriz tridiagonal deseada

$$\mathbf{T} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & -0{,}6 & 0{,}8 \\ 0 & 0{,}8 & 0{,}6 \end{bmatrix} \begin{bmatrix} -10 & 3 & -4 \\ 3 & 2 & 1 \\ -4 & 1 & 16 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & -0{,}6 & 0{,}8 \\ 0 & 0{,}8 & 0{,}6 \end{bmatrix} = \begin{bmatrix} -10 & -5 & 0 \\ -5 & 10 & 7 \\ 0 & 7 & 8 \end{bmatrix}.$$

c) *El método de bisección, aplicado a la matriz* $\mathbf{T}$*, requiere de la computación sucesiva de las cantidades*

$$
\begin{aligned}
&q_0 = 1, \quad \beta_0 = 0;\\
&q_1 = \alpha_1 - \mu - \frac{\beta_0^2}{q_0} = -10 - \mu,\\
&q_2 = \alpha_2 - \mu - \frac{\beta_1^2}{q_1} = 10 - \mu - \frac{25}{q_1},\\
&q_3 = \alpha_3 - \mu - \frac{\beta_2^2}{q_2} = 8 - \mu - \frac{49}{q_2},
\end{aligned}
$$

donde el valor μ *se ajusta según lo especificado en (5.16)–(5.18). Se recomienda empezar la iteración con un intervalo cuya longitud sea una potencia de* 2*. Así obtenemos los resultados*

s	a	b	q_1	q_2	q_3	m
0	-19.0000	-3.0000	1.0000	-4.0000	31.2500	1
1	-19.0000	-11.0000	5.0000	20.0000	20.5500	0
2	-15.0000	-11.0000	3.0000	14.6667	17.6591	0
3	-13.0000	-11.0000	2.0000	9.5000	14.8421	0
4	-12.0000	-11.0000	1.5000	4.8333	9.3621	0
5	-11.5000	-11.0000	1.2500	1.2500	-19.9500	1
6	-11.5000	-11.2500				

para $j = 1$*, por lo tanto,* $\lambda_1 \in [-11{,}5, -11{,}25]$*,*

s	a	b	q_1	q_2	q_3	m
0	-2.0000	6.0000	-12.0000	10.0833	1.1405	1
1	2.0000	6.0000	-14.0000	7.7857	-2.2936	2
2	2.0000	4.0000	-13.0000	8.9231	-0.4914	2
3	2.0000	3.0000	-12.5000	9.5000	0.3421	1
4	2.5000	3.0000	-12.7500	9.2108	-0.0699	2
5	2.5000	2.7500				

para $j = 2$*, por lo tanto,* $\lambda_2 \in [2{,}5, 2{,}75]$*, y*

s	a	b	q_1	q_2	q_3	m
0	8.0000	24.0000	-26.0000	-5.0385	1.7252	2
1	16.0000	24.0000	-30.0000	-9.1667	-6.6545	3
2	16.0000	20.0000	-28.0000	-7.1071	-3.1055	3
3	16.0000	18.0000	-27.0000	-6.0741	-0.9329	3
4	16.0000	17.0000	-26.5000	-5.5566	0.3183	2
5	16.5000	17.0000	-26.7500	-5.8154	-0.3241	3
6	16.5000	16.7500				

para $j = 3$*, por lo tanto,* $\lambda_3 \in [16{,}5, 16{,}75]$*. (Los valores exactos son* $\lambda_1 = -11{,}3301$*,* $\lambda_2 = 2{,}7080$ *y* $\lambda_3 = 16{,}6221$*.)*

5.4. Determinación de los vectores propios de una matriz tridiagonal hermitiana

En lo siguiente, se supone que $\mathbf{T}$ es una matriz tridiagonal hermitiana con $\gamma_i \neq 0$, $i = 1, \ldots, n$, y sea μ una aproximación de un valor propio λ de $\mathbf{T}$ determinada con exactitud de máquina (por ejemplo, usando el método de bisección). Sabemos que para λ arbitrario, $\text{rango}(\mathbf{T} - \lambda\mathbf{I}) \geqslant n-1$, y que ningún de los elementos subdiagonales de $\mathbf{T}$ desaparece, tal que la descomposición triangular de $\mathbf{T} - \mu\mathbf{I}$ puede ser realizada completamente con intercambios de filas. Para la búsqueda del pivote en la columna (medida indispensable aquí) tenemos la descomposición triangular

$$\mathbf{P}(\mathbf{T} - \mu\mathbf{I}) = \mathbf{LR}, \quad |\varrho_{jj}| \geqslant |\beta_j|, \quad j = 1, \ldots, n-1.$$

Si $\mu = \lambda_j$, entonces para una computación sin errores de redondeo tenemos $\varrho_{nn} = 0$ y una solución $\mathbf{x}$ de $\mathbf{Rx} = 0$ con $\xi_n = 1$ sera un vector propio de $\mathbf{T}$. En la práctica, no es asegurado que siempre resulta un valor de ϱ_{nn} pequeño, incluso cuando μ es una muy buena aproximación del valor propio. El siguiente teorema informa como a pesar de lo anterior, podemos determinar una buena aproximación del vector propio, siempre que la aproximación del valor propio es suficientemente buena.

Teorema 5.12. *Sea $\mathbf{T}$ una matriz tridiagonal y hermitiana, $\mathbf{P}(\mathbf{T}-\mu\mathbf{I}) = \mathbf{LR}$ una descomposición triangular (determinada con búsqueda del pivote en la columna), y μ una aproximación del valor propio λ_j de $\mathbf{T}$ con*

$$\mu = \lambda_j + \varepsilon\vartheta f(n)\|\mathbf{T}\|_2, \quad |\vartheta| \leqslant 1, \quad \varepsilon \text{ suficientemente pequeño.}$$

Sean todos los elementos subdiagonales de $\mathbf{T}$ diferentes de cero. Entonces existe (por lo menos) un índice $i \in \{1, \ldots, n\}$ tal que la solución $\mathbf{x}_i$ de

$$\mathbf{Rx}_i = \mathbf{e}_i\varrho_{ii}, \quad \mathbf{x}_i = \begin{pmatrix} \xi_{i1} \\ \vdots \\ \xi_{in} \end{pmatrix}$$

(con $\xi_{nn} := 1$ si $i = n$ y $\varrho_{nn} = 0$) satisface

$$\frac{\mathbf{x}_i}{\|\mathbf{x}_i\|_2} = \alpha\mathbf{u}_j + \mathbf{d}, \quad \|\mathbf{d}\|_2 \leqslant \frac{n^3 f(n)\varepsilon\|\mathbf{T}\|_2}{\min\{|\lambda_i - \lambda_j|, i \neq j\}} + \mathcal{O}(\varepsilon^2), \tag{5.20}$$

$|\alpha| = 1$, *donde* $\mathbf{U} := \begin{bmatrix} \mathbf{u}_1 & \cdots & \mathbf{u}_n \end{bmatrix}$ *es un sistema ortonormalizado de vectores propios de $\mathbf{T}$ y $\mathbf{Tu}_i = \lambda_i\mathbf{u}_i$, $i = 1, \ldots, n$.*

Demostración. Sea $\mathbf{y}_i := \varrho_{ii}\mathbf{PL}^{\mathrm{T}}\mathbf{e}_i$. Entonces, con $\mathbf{T} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^*$, $\mathbf{\Lambda} := \text{diag}(\lambda_1, \ldots, \lambda_n)$, tenemos

$$(\mathbf{T} - \mu\mathbf{I})\mathbf{x}_i = \mathbf{P}^{\mathrm{T}}\mathbf{LR}\mathbf{x}_i = \varrho_{ii}\mathbf{P}^{\mathrm{T}}\mathbf{L}\mathbf{e}_i = \mathbf{y}_i.$$

En el caso $\varrho_{nn} = 0$, ya no hay que demostrar nada, entonces podemos asumir que $\varrho_{nn} \neq 0$, y definimos

$$\mathbf{x}'_i := \frac{1}{\varrho_{ii}}\mathbf{x}_i, \quad \mathbf{y}'_i := \frac{1}{\varrho_{ii}}\mathbf{y}_i, \quad i = 1, \ldots, n,$$

y se supone que $\mathbf{x}_i$ e $\mathbf{y}_i$, $i = 1, \ldots, n$, poseen las representaciones

$$\mathbf{x}_i' := \sum_{k=1}^{n} \tilde{\xi}_{ik}\mathbf{u}_k, \quad \mathbf{y}_i' := \sum_{k=1}^{n} \tilde{\eta}_{ik}\mathbf{u}_k.$$

Entonces sabemos que

$$(\lambda_k - \mu)\tilde{\xi}_{ik} = \tilde{\eta}_{ik}, \quad i, k = 1, \ldots, n.$$

Dado que los elementos de $\mathbf{Le}_i$ son $\leqslant 1$ en valor absoluto, sabemos que

$$\begin{pmatrix} \tilde{\eta}_{i1} \\ \vdots \\ \tilde{\eta}_{in} \end{pmatrix} = \mathbf{U}^*\mathbf{P}^{\mathrm{T}}\mathbf{L}\mathbf{e}_i \Longrightarrow |\tilde{\eta}_{ik}| \leqslant \sqrt{n}, \quad i, k = 1, \ldots, n, \tag{5.21}$$

$$\max_{1\leqslant i\leqslant n} |\tilde{\eta}_{ij}| = \max_{1\leqslant i\leqslant n} \left|\mathbf{e}_j^{\mathrm{T}}\mathbf{U}^*\mathbf{P}^{\mathrm{T}}\mathbf{L}\mathbf{e}_i\right| = \|\mathbf{L}^{\mathrm{T}}\mathbf{P}\mathbf{U}\mathbf{e}_j\|_\infty \geqslant \frac{\|\mathbf{P}\mathbf{U}\mathbf{e}_j\|_\infty}{\|(\mathbf{L}^{-1})^{\mathrm{T}}\|_\infty} \geqslant \frac{1}{n^{3/2}}. \tag{5.22}$$

Pero definiendo

$$\sigma := \min_{i\neq j}\{|\lambda_i - \lambda_j|\} > 0,$$

sabemos que

$$\forall i = 1, \ldots, n, k \neq j: \quad |\tilde{\xi}_{ik}| = \frac{|\tilde{\eta}_{ik}|}{|\lambda_k - \lambda_j - \varepsilon\vartheta f(n)\|\mathbf{T}\|_2} \leqslant \frac{\sqrt{n}}{\sigma - \varepsilon f(n)\|\mathbf{T}\|_2}$$

(para ε suficientemente pequeño, $\sigma - \|\mathbf{T}\|_2\varepsilon f(n) > 0$), mientras que para un índice i apropiado,

$$|\tilde{\xi}_{ij}| = \frac{|\tilde{\eta}_{ij}|}{|\varepsilon\vartheta f(n)\|\mathbf{T}\|_2|} \geqslant \frac{1}{\varepsilon n^{3/2} f(n)\|\mathbf{T}\|_2}. \tag{5.23}$$

Para este índice i, tenemos que

$$\|\mathbf{x}_i'\|_2 = \left(\sum_{k=1}^{n} |\tilde{\xi}_{ik}|^2\right)^{1/2} = |\tilde{\xi}_{ij}| \left(1 + \sum_{\substack{k=1 \\ k\neq j}}^{n} \frac{|\tilde{\xi}_{ik}|^2}{|\tilde{\xi}_{ij}|^2}\right)^{1/2} = |\tilde{\xi}_{ij}|(1 + \vartheta_{ij}),$$

donde, tomando en cuenta que $1 \leqslant \sqrt{1+\xi} \leqslant 1 + \xi$ para $0 \leqslant \xi \leqslant 1$,

$$0 \leqslant \vartheta_{ij} \leqslant \frac{\|\mathbf{T}\|_2^2 n^4 f^2(n)\varepsilon^2}{(\sigma - \varepsilon f(n)\|\mathbf{T}\|_2)^2}.$$

Entonces,

$$\frac{\mathbf{x}_i}{\|\mathbf{x}_i\|_2} = \frac{\mathbf{x}_i'}{\|\mathbf{x}_i'\|_2} = \frac{\tilde{\xi}_{ij}}{|\tilde{\xi}_{ij}|(1+\vartheta_{ij})}\mathbf{u}_j + \sum_{\substack{k=1 \\ k\neq j}}^{n} \frac{\tilde{\xi}_{ik}}{|\tilde{\xi}_{ij}|(1+\vartheta_{ij})}\mathbf{u}_k = \frac{\tilde{\xi}_{ij}}{|\tilde{\xi}_{ij}|}\mathbf{u}_j + \mathbf{d},$$

donde el vector $\mathbf{d}$ satisface

$$\|\mathbf{d}\|_2 \leqslant 1 - \frac{1}{1 + \dfrac{n^4 f^2(n)\|\mathbf{T}\|_2^2\varepsilon^2}{(\sigma - \varepsilon f(n)\|\mathbf{T}\|_2)^2}} + \frac{(n-1)\sqrt{n}n^{3/2}f(n)\|\mathbf{T}\|_2\varepsilon}{\sigma - \varepsilon f(n)\|\mathbf{T}\|_2} \leqslant \frac{n^3 f(n)\varepsilon}{\sigma}\|\mathbf{T}\|_2 + \mathcal{O}(\varepsilon^2).$$

■

Según nuestra derivación, es obvio que el factor n^3 en el enunciado del Teorema (5.12) es muy pesimista. Se puede suponer que en la mayoría de los cases, el factor 1 es más apropiado que n^3. Se puede demostrar que $i = n$ es apropiado si con

$$\mathbf{R} = \begin{bmatrix} \mathbf{R}_{11} & \mathbf{r} \\ 0 & \varrho_{nn} \end{bmatrix},$$

la norm a $\|\mathbf{R}_{11}^{-1}\mathbf{r}\|_2$ es "pequeña" (por ejemplo, $\leqslant n$). En la práctica, se procede de la siguiente forma: se resuelven los sistemas

$$\mathbf{R}\mathbf{x}_i = \varrho_{ii}\mathbf{e}_i, \quad i = n, n-1, \ldots; \tag{5.24}$$

la computación termina cuando por primera vez,

$$\|\mathbf{x}_i\|_\infty \geqslant \frac{|\varrho_{ii}|}{100n\varepsilon}, \tag{5.25}$$

donde el factor $1/(100n)$ remplaza (de forma un poco arbitraria) eñ término $1/(n^{3/2}f(n))$ (no se conoce el verdadero en μ). Si (5.25) no se cumple para ningún i, también la aproximación μ puede ser considerada de mala calidad. "Normalmente", el test ya está satisfecho para $i = n$. Además, se puede domostrar que modificaciones de $\mathbf{T}$ del orden $\varepsilon\|\mathbf{T}\|_2$ causan errores en los vectores propios del orden

$$\frac{n\varepsilon\|\mathbf{T}\|_2}{\min\limits_{i\neq j}|\lambda_i - \lambda_j|}.$$

Entonces, el Teorema 5.12 representa un resultado excelente. Además, podemos demostrar que los errores de redondeo cometidos al calcular la descomposición triangular de $\mathbf{P}(\mathbf{T} - \mu\mathbf{I})$ y durante la solución de (5.24) no afectan seriamente el resultado.

5.5. Valores propios de una matriz de tipo Hessenberg

En los capítulos anteriores vimos que es relativamente fácil determinar un valor propio individual (y el vector propio asociado) de una matriz hermitiana. Por otro lado, vimos que cada matriz de $\mathbb{C}^{n\times n}$ puede ser transformada unitariamente a la forma de Hessenberg superior. Para la determinación de un valor propio de una tal matriz, es importante que el polinomio característico $p_n(\lambda; \mathbf{A})$, y posiblemente su su derivada $p'_n(\lambda; \mathbf{A})$, pueden ser evaluados facilmente.

Sea $\gamma \neq 0$ arbitrario. Para la componente ξ_n del sistema lineal

$$\begin{aligned} (\alpha_{11} - \lambda)\xi_1 + \alpha_{12}\xi_2 + \cdots + \alpha_{1n}\xi_n &= -\gamma, \\ \alpha_{21}\xi_1 + (\alpha_{22} - \lambda)\xi_2 + \cdots + \alpha_{2n}\xi_n &= 0, \\ &\vdots \\ \alpha_{n1}\xi_1 + \alpha_{n2}\xi_2 + \cdots + (\alpha_{nn} - \lambda)\xi_n &= 0 \end{aligned} \tag{5.26}$$

($\gamma \neq 0$ arbitrario) tenemos, según la regla de Cramer,

$$\xi_n = \frac{1}{\det(\mathbf{A}-\lambda\mathbf{I})}\begin{vmatrix} a_{11}-\lambda & a_{12} & \dots & a_{1,n-1} & -\gamma \\ a_{21} & a_{22}-\lambda & & a_{2,n-1} & 0 \\ \vdots & \ddots & \ddots & \vdots & \vdots \\ \vdots & & \ddots & \alpha_{n-1,n-1}-\lambda & 0 \\ \alpha_{n1} & \dots & \dots & \alpha_{n,n-1} & 0 \end{vmatrix} = \frac{(-1)^n\gamma\alpha_2 1\cdot\ldots\cdot\alpha_{n,n-1}}{\det(\mathbf{A}-\lambda\mathbf{I})},$$

es decir, cuando ponemos $\xi_n := 1$ y solvemos el sistema lineal por substitución, obtenemos

$$\gamma = \frac{(-1)^n \det(\mathbf{A}-\lambda\mathbf{I})}{\alpha_{21}\alpha_{32}\cdot\ldots\cdot\alpha_{n,n-1}}. \tag{5.27}$$

Esto significa que hasta un factor común, obtenemos $\det(\mathbf{A}-\lambda\mathbf{I})$ y $(\mathrm{d}/\mathrm{d}\lambda)\det(\mathbf{A}-\lambda\mathbf{I})$ por la siguiente recursión, conocida como *Método de Hyman*;

$$\begin{aligned} x_n(\lambda) &:= 1 \quad (\text{corresponde a } \xi_n), \\ x_n'(\lambda) &:= 0 \quad (\text{corresponde a } \frac{d}{d\lambda}\xi_n,) \\ x_{n-i}(\lambda) &:= \frac{1}{\alpha_{n-i+1,n-i}}\left(\lambda x_{n-i+1}(\lambda) - \sum_{k=n-i+1}^{n}\alpha_{n-i+1,k}x_k(\lambda)\right), \quad i=1,\dots,n-1, \\ x_{n-i}'(\lambda) &:= \frac{1}{\alpha_{n-i+1,n-i}}\left(x_{n-i+1}(\lambda) + \lambda x_{n-i+1}'(\lambda) - \sum_{k=n-i+1}^{n}\alpha_{n-i+1,k}x_k'(\lambda)\right). \end{aligned}$$

Ahora, considerando la primera ecuación de (5.26) obtenemos

$$\det(\mathbf{A}-\lambda\mathbf{I}) = (-1)^{n+1}\alpha_{21}\alpha_{32}\cdot\dots\cdot\alpha_{n,n-1}\left((\alpha_{11}-\lambda)x_1(\lambda) + \sum_{k=2}^{n}\alpha_{1k}x_k(\lambda)\right)$$

y análogamente

$$\frac{\mathrm{d}}{\mathrm{d}\lambda}\det(\mathbf{A}-\lambda\mathbf{I}) = (-1)^{n+1}\alpha_{21}\alpha_{32}\cdot\dots\cdot\alpha_{n,n-1}\left((\alpha_{11}-\lambda)x_1'(\lambda) - x_1(\lambda) + \sum_{k=2}^{n}\alpha_{1k}x_k'(\lambda)\right).$$

Con métodos iterativos para los ceros de un polinomio podemos facilmente calcular los valores propios. Hay que considerar que este método es muy diferente a la computación de los coeficientes del polinomio característico.

5.6. La iteración directa según von Mises y el método de Wielandt

Ahora consideramos métodos que nos entregan un vector propio aproximado. Ya vimos como mediante el cuociente de Rayleigh podemos obtener un valor propio aproximado usando el vector propio aproximado. La versión básica de la iteración de von Mises (pero de poca utilidad práctica) es la siguiente.

1. Sea $0 \neq \mathbf{x}_0 \in \mathbb{C}^n$ apropiado.
2. Para $k = 0, 1, 2, \dots$, sea $\mathbf{x}_{k+1} = \mathbf{A}\mathbf{x}_k$.

Teorema 5.13. *Sea* $\mathbf{A} \in \mathbb{C}^{n\times n}$ *diagonalizable y* $\mathbf{U} = \begin{bmatrix}\mathbf{u}_1 & \cdots & \mathbf{u}_n\end{bmatrix}$ *un sistema de vectores propios de* $\mathbf{A}$*, con* $\mathbf{A}\mathbf{u}_i = \lambda_i \mathbf{u}_i$*,* $i = 1, \ldots, n$*. Ademas sea*

$$\mathbf{x}_0 = \sum_{i=1}^{n} \xi_i \mathbf{u}_i, \quad \xi_1 \neq 0, \quad |\lambda_1| > |\lambda_2| \geqslant \ldots \geqslant |\lambda_n|.$$

Entonces la sucesión de vectores $\{\mathbf{x}_k\}_{k\in\mathbb{N}_0}$ *satisface*

$$\mathbf{x}_k = \xi_1 \lambda_1^k \left(\mathbf{u}_1 + \mathcal{O}\left(\left| \frac{\lambda_2}{\lambda_1} \right|^k \right) \right), \tag{5.28}$$

$$R(\mathbf{x}_k; \mathbf{A}) = \lambda_1 \left[1 + \mathcal{O}\left(\left| \frac{\lambda_2}{\lambda_1} \right|^k \right) \right]. \tag{5.29}$$

Demostración. Calculamos que

$$\begin{aligned}
\mathbf{x}_k &= \mathbf{A}^k \mathbf{x}_0 \\
&= \begin{bmatrix}\mathbf{u}_1 & \cdots & \mathbf{u}_n\end{bmatrix} \operatorname{diag}(\lambda_1^k, \ldots, \lambda_n^k) \begin{bmatrix}\mathbf{u}_1 & \cdots & \mathbf{u}_n\end{bmatrix}^{-1} \begin{bmatrix}\mathbf{u}_1 & \cdots & \mathbf{u}_n\end{bmatrix} \begin{pmatrix}\xi_1 \\ \vdots \\ \xi_n\end{pmatrix} \\
&= \sum_{i=1}^{n} \lambda_i^k \xi_i \mathbf{u}_i \\
&= \xi_1 \lambda_1^k \left(\mathbf{u}_1 + \sum_{i=2}^{n} \frac{\xi_i}{\xi_1} \left(\frac{\lambda_i}{\lambda_1} \right)^k \mathbf{u}_i \right),
\end{aligned}$$

lo cual implica (5.28) si tomamos en cuenta que

$$\left| \frac{\lambda_i}{\lambda_1} \right|^k \leqslant \left| \frac{\lambda_2}{\lambda_1} \right|^k.$$

Por otro lado,

$$\begin{aligned}
R(\mathbf{x}_k; \mathbf{A}) &= \frac{\mathbf{x}_k^* \mathbf{x}_{k+1}}{\mathbf{x}_k^* \mathbf{x}_k} \\
&= \frac{\bar{\xi}_1 \bar{\lambda}_1^k \left[\mathbf{u}_1^* + \mathcal{O}\left(\left| \frac{\lambda_2}{\lambda_1} \right|^k \right) \right] \xi_1 \lambda_1^{k+1} \left[\mathbf{u}_1 + \mathcal{O}\left(\left| \frac{\lambda_2}{\lambda_1} \right|^{k+1} \right) \right]}{\bar{\xi}_1 \bar{\lambda}_1^k \left[\mathbf{u}_1^* + \mathcal{O}\left(\left| \frac{\lambda_2}{\lambda_1} \right|^k \right) \right] \xi_1 \lambda_1^k \left[\mathbf{u}_1 + \mathcal{O}\left(\left| \frac{\lambda_2}{\lambda_1} \right|^k \right) \right]} \\
&= \lambda_1 \left[1 + \mathcal{O}\left(\left| \frac{\lambda_2}{\lambda_1} \right|^k \right) \right].
\end{aligned}$$

■

Comentamos primero que para $|\lambda_1| \neq 1$, la versión presentada aquí rapidamente entrega números extremadamente grandes o pequeños. Por lo tanto, se prefiere calcular $\{\mathbf{x}_k\}$ de la siguiente forma:

$$\tilde{\mathbf{x}}_{k+1} := \mathbf{A}\mathbf{x}_k, \quad \mathbf{x}_{k+1} := \frac{\tilde{\mathbf{x}}_{k+1}}{\|\tilde{\mathbf{x}}_{k+1}\|}.$$

En este caso,

$$\mathbf{x}_k = \vartheta_k \left[\mathbf{u}_1 + \mathcal{O}\left(\left|\frac{\lambda_2}{\lambda_1}\right|^k\right)\right], \quad |\vartheta_k| = 1, \quad \mathbf{x}_k^* \tilde{\mathbf{x}}_{k+1} = \lambda_1 \left[1 + \mathcal{O}\left(\left|\frac{\lambda_2}{\lambda_1}\right|^k\right)\right].$$

Si $\mathbf{x}_k$ se normaliza a $(\mathbf{x}_k)_j = 1$ para una componente j con $(\mathbf{u}_1)_j \neq 0$, entonces $\{\mathbf{x}_k\}_{k\in\mathbb{N}}$ converge para $k \to \infty$ a un múltiple de $\mathbf{u}_1$.

El Teorema 5.13 es análogamente válido para $\lambda_1 = \ldots = \lambda_r$, $|\lambda_1| > |\lambda_{r+1}| \geqslant \ldots \geqslant |\lambda_n|$, cuando

$$\mathbf{x}_0 = \sum_{i=1}^{n} \xi_i \mathbf{u}_i, \quad \sum_{i=1}^{r} |\xi_i| \neq 0.$$

La presuposición $\xi_1 \neq 0$ o $|\xi_1| + \ldots + |\xi_r| \neq 0$ siempre está satisfecha en la práctica debido a errores de redondeo, includo cuando $\mathbf{x}_0$ no es apropiado.

La matriz $\mathbf{A}$ no necesariamente debe ser diagonalizable. Pero si $\mathbf{A}$ no lo es, el término de error $\mathcal{O}(|\lambda_2/\lambda_1|^k)$ es remplazado por $\mathcal{O}(1/k)$.

Cuando $\mathbf{A}$ posee diferentes valores propios dominantes (del mismo valor absoluto), por ejemplo en el caso de un valor propio dominante en valor absoluto complejo de una matriz $\mathbf{A}$ real, no hay convergencia (pero existen generalizaciones de este método para este caso).

Ejemplo 5.8 (Tarea 28, Curso 2006)**.**

$$\mathbf{A} = \begin{bmatrix} 1 & -1 & 1 \\ -1 & 2 & 1 \\ 1 & 1 & 10 \end{bmatrix}$$

a) *Demostrar que* $\mathbf{A}$ *satisface las condiciones para la ejecución de la iteración directa (método de von Mises), y que* $\mathbf{x}_0 := (0, 0, 1)^{\mathrm{T}}$ *es un vector apropiado para iniciar la iteración.*

b) *Determinar* $\mathbf{x}_3$ *(usando el algoritmo básico, sin normalizar), y calcular usando* $\mathbf{x}_3$ *y* $\mathbf{x}_2$ *un valor aproximado del valor propio. Para esta aproximación estimar el error rigurosamente.*

Solución sugerida.

a) *La matriz* $\mathbf{A}$ *es real y simétrica, entonces posee un sistema completo de vectores propios* $\mathbf{Q} := [\mathbf{u}_1\ \mathbf{u}_2\ \mathbf{u}_3]$*. Según el Teorema de Gershgrorin hay dos valores propios,* λ_2 *y* λ_3*, en el intervalo* $[-1, 4]$*; el tercer,* λ_1*, de valor absoluto máximo, pertenece al intervalo* $[8, 12]$*. Ahora sea* $\mathbf{A}\mathbf{u}_i = \lambda_i \mathbf{u}_i$*. Supongamos que* $\xi_1 = 0$ *en la representación* $\mathbf{x}_0 = \xi_1\mathbf{u}_1 + \xi_2\mathbf{u}_2 + \xi_3\mathbf{u}_3$*. En este caso, tendriamos* $\mathbf{A}\mathbf{x}_0 = \xi_2\lambda_2\mathbf{u}_2 + \xi_3\lambda_3\mathbf{u}_3$*, entonces*

$$\|\mathbf{A}\mathbf{x}_0\| \leqslant \max\{|\lambda_2|, |\lambda_3|\}\|\mathbf{x}_0\| \leqslant 4\|\mathbf{x}_0\| = 4,$$

mientras que

$$\|\mathbf{A}\mathbf{x}_0\| = \left\|(1,1,10)^{\mathrm{T}}\right\| \geqslant 10,$$

una contradicción. Entonces $\xi_1 \neq 0$*, y el vector* $\mathbf{x}_0$ *es apropiado.*

b) *Sin normalizar obtenemos*

$$\mathbf{x}_1 = (1,1,10)^{\mathrm{T}}, \quad \mathbf{x}_1 = (10,11,102)^{\mathrm{T}}, \quad \mathbf{x}_1 = (101,114,1041)^{\mathrm{T}}.$$

El valor propio aproximado correspondiente es

$$\tilde{\lambda}_1 = \frac{\mathbf{x}_2^{\mathrm{T}}\mathbf{x}_3}{\mathbf{x}_2^{\mathrm{T}}\mathbf{x}_2} = \frac{108446}{10625} = 10{,}206682.$$

Según el Teorema 4.4, sabemos que para una matriz $\mathbf{A}$ *normal (por ejemplo,* $\mathbf{A}$ *simétrica), entonces existe un valor propio* $\lambda_j \neq 0$ *de* $\mathbf{A}$ *con*

$$\left|\frac{\lambda_j - \lambda}{\lambda_j}\right| \leqslant \frac{\|\mathbf{A}\mathbf{x} - \lambda\mathbf{x}\|}{\|\mathbf{A}\mathbf{x}\|_2},$$

donde $\mathbf{x}$ *y* λ *son aproximaciones del vector y del valor propio, respectivamente. Aqui hay que usar* $\mathbf{x} = \mathbf{x}_2$*,* $\lambda = \tilde{\lambda}_1 = 10{,}206682$*, y* $\mathbf{x}_3 = \mathbf{A}\mathbf{x}$ *para obtener*

$$\left|\frac{\lambda_j - \tilde{\lambda}_1}{\lambda_j}\right| \leqslant \frac{\|\mathbf{x}_3 - \lambda\mathbf{x}_2\|_2}{\|\mathbf{x}_3\|_2} = \frac{2{,}031146}{1052{,}0827} = 0{,}001931.$$

Entonces,

$$|\lambda_j - \tilde{\lambda}_1| \leqslant 0{,}001931|\lambda_j| \leqslant 0{,}001931 \cdot 12 = 0{,}023167.$$

Ejemplo 5.9 (Certamen 2, Curso 2010)**.** *Se considera la matriz*

$$\mathbf{A} = \begin{bmatrix} -10 & -1 & -1 \\ -1 & 2 & 1 \\ -1 & 1 & 10 \end{bmatrix}$$

a) *Demostrar que* $\mathbf{A}$ *posee tres valores propio* $\lambda_1 < \lambda_2 < \lambda_3$*, en particular* $\lambda_1, \lambda_2, \lambda_3 \in \mathbb{R}$*, y que* $\mathbf{x}_0 := (0,0,1)^{\mathrm{T}}$ *es un vector apropiado para iniciar la iteración directa (método de von Mises).*

b) *Determinar* $\mathbf{x}_3$ *(usando el algoritmo básico, sin normalizar), y calcular usando* $\mathbf{x}_3$ *y* $\mathbf{x}_2$ *un valor aproximado del valor propio. Para esta aproximación estimar el error rigurosamente.*

Solución sugerida.

a) *Dado que* $\mathbf{A}$ *es simétrica, sus valores propios son reales, y los círculos de Gershgorin son*

$$\mathcal{K}_1 = [-12,-8], \quad \mathcal{K}_2 = [0,4], \quad \mathcal{K}_3 = [8,12].$$

Dado que $\mathcal{K}_i \cap \mathcal{K}_j = \varnothing$ *para* $i \neq j$*, cada uno de los círculos contiene exactamente un valor propio, es decir* $\lambda_i \in \mathcal{K}_i$ *para* $i = 1,2,3$*. En el presente caso, aun no podemos*

decidir si el valor propio de valor absoluto máximo pertenece a $\mathcal{K}_1$ o a $\mathcal{K}_3$. Para obtener más información, calculamos el polinomio carácteristico:

$$p(\lambda;\mathbf{A}) = \det(\mathbf{A}-\lambda\mathbf{I}) = \begin{vmatrix} -10-\lambda & -1 & -1 \\ -1 & 2-\lambda & 1 \\ -1 & 1 & 10-\lambda \end{vmatrix} = -\lambda^3+2\lambda^2+103\lambda-200.$$

Debido al signo del coeficiente de λ^3 y sabiendo ya que hay tres valores propios distintos reales, concluimos que

$$p(\lambda)\begin{cases} >0 & \text{para } \lambda<\lambda_1, \\ =0 & \text{para } \lambda=\lambda_1, \\ <0 & \text{para } \lambda_1<\lambda<\lambda_2, \\ =0 & \text{para } \lambda=\lambda_2, \\ >0 & \text{para } \lambda_2<\lambda<\lambda_3, \\ =0 & \text{para } \lambda=\lambda_3, \\ <0 & \text{para } \lambda>\lambda_3. \end{cases}$$

Ahora, evaluando $p(2;\mathbf{A}) = 6 > 0$ concluimos que $\lambda_2 < 2$. Por otro lado, la traza de $\mathbf{A}$ es la suma de sus valores propios. En nuestro caso, $\lambda_1+\lambda_2+\lambda_3 = 2$, es decir

$$\lambda_1+\lambda_3 = 2-\lambda_2 > 0,$$

es decir $\lambda_1 > -\lambda_3$. Dado que $\lambda_1 < 0$ y $\lambda_3 > 0$, esta desigualdad implica que

$$|\lambda_1| = -\lambda_1 < \lambda_3 = |\lambda_3|.$$

Por lo tanto, λ_3 es el valor propio de mayor valor absoluto. Como $\mathbf{A}$ es simétrica, $\mathbf{A}$ posee un sistema de vectores propios ortonormales. Sean $\mathbf{u}_1$, $\mathbf{u}_2$, $\mathbf{u}_3$ los vectores propios correspondiente a los valores propios respectivos λ_1, λ_2, λ_3. Sea $\mathbf{x}_0 = \xi_1\mathbf{u}_1+\xi_2\mathbf{u}_2+\xi_3\mathbf{u}_3$. De acuerdo al Teorema 5.13 hay que demostrar que $\xi_3 \neq 0$. Ahora, si fuera $\xi_3 = 0$, se tendria que

$$\begin{aligned} R(\mathbf{x}_0;\mathbf{A}) = R(\xi_1\mathbf{u}_1+\xi_2\mathbf{u}_2;\mathbf{A}) &= \frac{(\xi_1\mathbf{u}_1^{\mathrm{T}}\xi_2\mathbf{u}_2^{\mathrm{T}})(\lambda_1\xi_1\mathbf{u}_1+\lambda_2\xi_2\mathbf{u}_2)}{\xi_1^2+\xi_2^2} \\ &= \frac{\lambda_1\xi_1^2+\lambda_2\xi_2^2}{\xi_1^2+\xi_2^2} = \frac{\xi_1^2}{\xi_1^2+\xi_2^2}\lambda_1 + \frac{\xi_2^2}{\xi_1^2+\xi_2^2}\lambda_2 \in [\lambda_1,\lambda_2] \subset [-12,2]. \end{aligned}$$

Pero, efectivamente,

$$R(\mathbf{x}_0;\mathbf{A}) = (0,0,1)(-1,1,10)^{\mathrm{T}} = 10 \notin [-12,2],$$

es decir $\xi_3 \neq 0$; por lo tanto, $\mathbf{x}_0$ es apropiado.

b) *Iterando obtenemos*

$$\mathbf{x}_1 = \mathbf{A}\mathbf{x}_0 = \begin{pmatrix} -1 \\ 1 \\ 10 \end{pmatrix}, \quad \mathbf{x}_2 = \mathbf{A}\mathbf{x}_1 = \begin{pmatrix} -1 \\ 13 \\ 102 \end{pmatrix}, \quad \mathbf{x}_3 = \mathbf{A}\mathbf{x}_2 = \begin{pmatrix} -105 \\ 129 \\ 1034 \end{pmatrix},$$

y el valor propio aproximado

$$\lambda_3 \approx \tilde{\lambda}_3 = R(\mathbf{x}_2;\mathbf{A}) = \frac{\mathbf{x}_2^{\mathrm{T}}\mathbf{x}_3}{\mathbf{x}_2^{\mathrm{T}}\mathbf{x}_2} = 10{,}1428.$$

Para estimar el error, utilizamos la parte (iii) del Teorema 5.4. Aquí esto significa que existe un valor propio λ_j de $\mathbf{A}$ tal que

$$\left|\frac{\lambda_j - \tilde{\lambda}_3}{\lambda_j}\right| \leqslant \frac{\|\mathbf{A}\mathbf{x}_2 - 10{,}1428\mathbf{x}_2\|_2}{\|\mathbf{A}\mathbf{x}_2\|_2} = \frac{\|\mathbf{x}_3 - 10{,}1428\mathbf{x}_2\|}{\|\mathbf{x}_3\|_2} = \frac{94{,}9019}{1047{,}3} = 0{,}0906.$$

Evidentemente, $j = 3$, y puesto que $\lambda_3 \in [8, 12]$, llegamos a

$$|\lambda_3 - \tilde{\lambda}_3| \leqslant 12 \times 0{,}0906 = 1{,}0874.$$

Mejores cotas son posibles.

Nos damos cuenta que la velocidad de convergencia de la iteración directa depende decisivamente del cuociente $|\lambda_2/\lambda_1| < 1$. Ahora, sean $\lambda_1, \ldots, \lambda_n$ los valores propios de $\mathbf{A}$. Si μ es una aproximación del valor propio λ_i y $0 < |\lambda_i - \mu| < |\lambda_j - \mu|$ para todo $j \in \{1, \ldots, n\}\backslash\{i\}$, entonces $\mathbf{A} - \mu\mathbf{I}$ es regular y los valores propios $\tau_k = 1/(\lambda_k - \mu)$ de $(\mathbf{A} - \mu\mathbf{I})^{-1}$ satisfacen

$$\forall j \in \{1, \ldots, n\}\backslash\{i\} : \quad |\tau_i| > |\tau_j|.$$

Ademas, $\max_{i\neq j} |\tau_j|/|\tau_i|$ es pequeño cuando la aproximación del valor propio era buena. Entonces, la iteración directa aplicada a $(\mathbf{A}-\mu\mathbf{I})^{-1}$ converge rapidamente. Esta es la idea del método de interación "inversa", conocida también como *Método de Wielandt*. La observación decisiva para su ejecución es que no hay que calcular explícitamente la matriz $(\mathbf{A} - \mu\mathbf{I})^{-1}$; al lugar de eso, en cada paso se resuleve el sistema lineal

$$(\mathbf{A} - \mu\mathbf{I})\tilde{\mathbf{x}}_{k+1} = \mathbf{x}_k, \quad \mathbf{x}_{k+1} := \frac{1}{\|\tilde{\mathbf{x}}_{k+1}\|}\tilde{\mathbf{x}}_{k+1},$$

lo que cuesta poco esfuerzo computacional si una vez por siempre se ha calculado una descomposición triangular (por lo menos, con búsqueda del privot en la columna) de $\mathbf{A} - \mu\mathbf{I}$. Entonces, si $\mathbf{P}(\mathbf{A} - \mu\mathbf{I})\mathbf{Q} = \mathbf{L}\mathbf{R}$, calculamos para $k \in \mathbb{N}_0$ sucesivamente los vectores $\mathbf{z}_k$, $\mathbf{v}_k$, $\mathbf{w}_k$ y $\tilde{\mathbf{x}}_{k+1}$:

$$\mathbf{z}_k = \mathbf{P}\mathbf{x}_k, \quad \mathbf{L}\mathbf{v}_k = \mathbf{z}_k, \quad \mathbf{R}\mathbf{w}_k = \mathbf{v}_k, \quad \mathbf{Q}^{\mathrm{T}}\tilde{\mathbf{x}}_{k+1} = \mathbf{w}_k.$$

Ahora, usando la nueva aproximación $\tilde{\mathbf{x}}_{k+1}$, podriamos calcular una nueva aproximación del valor propio, determinar una nueva descomposición triangular, etc. Pero, en general, el esfuerzo computacional para realizar eso es exagerado.

El siguiente teorema provee información acerca de un vector inicial apropiado.

Teorema 5.14. *Sea $\mathbf{A} \in \mathbb{C}^{n\times n}$ diagonalizable, $\mathbf{A}\mathbf{u}_i = \lambda\mathbf{u}_i$, $i = 1, \ldots, n$, $\|\mathbf{u}_i\|_2 = 1$ para todo i, donde $\mathbf{U} = \begin{bmatrix}\mathbf{u}_1 & \ldots & \mathbf{u}_n\end{bmatrix}$ es el sistema completo de vectores propios de $\mathbf{A}$ y $\mathbf{P}(\mathbf{A}-\mu\mathbf{I})\mathbf{Q} = \mathbf{L}\mathbf{R}$ con matrices de permutación $\mathbf{P}$ y $\mathbf{Q}$, $\mathbf{L}$ una matriz triangular inferior con diagonal $(1, \ldots, 1)$ y elementos de valor absoluto $\leqslant 1$, y $\mathbf{R}$ una matriz triangular superior. Además definimos el índice s a través de*

$$|\varrho_{ss}| = \min_{1\leqslant i\leqslant n} |\varrho_{ii}|,$$

y $\hat{\mathbf{R}} := \operatorname{diag}(\varrho_{11}^{-1}, \ldots, \varrho_{nn}^{-1})\mathbf{R}$. *En este caso, si* μ *no es un valor propio, sabemos que*

$$\min_{1\leqslant j\leqslant n} |\lambda_j - \mu| \leqslant \sqrt{n}|\varrho_{ss}| \operatorname{cond}_{\|\cdot\|_2}(\mathbf{U}) \leqslant \min_{1\leqslant j\leqslant n} |\lambda_j - \mu| \|\mathbf{L}^{-1}\|_2 \|\hat{\mathbf{R}}^{-1}\|_2 \operatorname{cond}_{\|\cdot\|_2}(\mathbf{U})\sqrt{n}.$$

Si definimos $\mathbf{x}_1$ *por*

$$\mathbf{R}\mathbf{Q}^{\mathrm{T}}\mathbf{x}_1 = \varrho_{ss}\mathbf{e}_1$$

(lo que corresponde a $\mathbf{x}_0 := \varrho_{ss}\mathbf{P}^{\mathrm{T}}\mathbf{L}\mathbf{e}_s$*), y el índice* k *por*

$$|\xi_k| = \max_{1\leqslant i\leqslant n} |\xi_i|, \quad \textit{donde } \mathbf{x}_1 = \sum_{i=1}^{n} \xi_i \mathbf{u}_i,$$

entonces sabemos que el valor propio corespondiente λ_k *satisface*

$$|\lambda_k - \mu| \leqslant n^{3/2}|\varrho_{ss}| \operatorname{cond}_{\|\cdot\|_2}(\mathbf{U}).$$

Eso significa que si los valores propios de $\mathbf{A}$ *son suficientemente separados (comparado con* $\min_{1\leqslant j\leqslant n} |\lambda_j - \mu|$*), entonces*

$$|\lambda_k - \mu| = \min_{1\leqslant j\leqslant n} |\lambda_j - \mu|$$

y $\mathbf{x}_1$ *es una aproximación apropiada para iniciar la iteración inversa.*

Demostración. Cambiar el valor ϱ_{ss} en la descomposición triangular a cero es equivalente a cambiar $\mathbf{A} - \mu\mathbf{I}$ a

$$\mathbf{B} := \mathbf{A} - \mu\mathbf{I} - \varrho_{ss}\mathbf{P}^{\mathrm{T}}\mathbf{L}\mathbf{e}_s\mathbf{e}_s^{\mathrm{T}}\mathbf{Q}^{\mathrm{T}},$$

y la matriz $\mathbf{B}$ es singular. En este caso, el Teorema 5.7, aplicado al valor propio cero de $\mathbf{B}$ y el valor propio $\lambda_i - \mu$ de $\mathbf{A}$, entrega que

$$|\lambda_i - \mu| = \big|0 - (\lambda_i - \mu)\big| \leqslant \operatorname{cond}_{\|\cdot\|_2}(\mathbf{U})|\varrho_{ss}|\|\mathbf{P}^{\mathrm{T}}\mathbf{L}\mathbf{e}_s\mathbf{e}_s^{\mathrm{T}}\mathbf{Q}\|_2 \leqslant \sqrt{n}|\varrho_{ss}| \operatorname{cond}_{\|\cdot\|_2}(\mathbf{U})$$

para un índice i apropiado. Ahora, sea j_0 definido por

$$|\lambda_{j_0} - \mu| = \min_{1\leqslant i\leqslant n} |\lambda_i - \mu|.$$

Dado que $\|\mathbf{u}_{j_0}\|_2 = 1$, sabemos que

$$(\mathbf{A} - \mu\mathbf{I})^{-1}\mathbf{u}_{j_0} = \frac{1}{\lambda_{j_0} - \mu}\mathbf{u}_{j_0},$$

lo que implica que

$$\frac{1}{|\lambda_{j_0} - \mu|} \leqslant \big\|(\mathbf{A} - \mu\mathbf{I})^{-1}\big\|_2 = \big\|(\mathbf{P}^{\mathrm{T}}\mathbf{L}\mathbf{R}\mathbf{Q}^{\mathrm{T}})^{-1}\big\|_2 \leqslant \|\mathbf{L}^{-1}\|_2\|\hat{\mathbf{R}}^{-1}\|_2\frac{1}{|\varrho_{ss}|},$$

es decir,

$$|\varrho_{ss}| \leqslant \min_{1\leqslant j\leqslant n} |\lambda_j - \mu|\|\mathbf{L}^{-1}\|_2\|\hat{\mathbf{R}}^{-1}\|_2.$$

Según la definición de $\mathbf{x}_1$ y del índice k, tenemos que

$$1 \leqslant \|\mathbf{x}_1\|_2 \leqslant n|\xi_k|.$$

Además,

$$\mathbf{x}_0 = \sum_{i=1}^{n} \xi_i(\lambda_i - \mu)\mathbf{u}_i,$$
$$|\xi_k||\lambda_k - \mu| \leqslant \|\mathbf{U}^{-1}\mathbf{x}_0\|_2 = \|\mathbf{U}^{-1}\varrho_{ss}\mathbf{P}^{\mathrm{T}}\mathbf{L}\mathbf{e}_s\|_2 \leqslant |\varrho_{ss}|\|\mathbf{U}^{-1}\|_2\sqrt{n},$$
$$|\lambda_k - \mu| \leqslant n^{3/2}|\varrho_{ss}|\|\mathbf{U}^{-1}\|_2 \leqslant n^{3/2}|\varrho_{ss}|\,\mathrm{cond}_{\|\cdot\|_2}(\mathbf{U}).$$

■

Si la descomposición triangular se ha ejecutado con búsqueda del pivote en la matriz, entonces los elementos de $\hat{\mathbf{R}}$ satisfacen $\hat{\varrho}_{ii} = 1$ y $|\hat{\varrho}_{ij} \leqslant 1$. En este caso, $\|\mathbf{L}^{-1}\|_2\|\hat{\mathbf{R}}^{-1}\|_2$ puede ser estimado como una función solamente de n. Si $\mu = \lambda_j$ para un índice j, entonces $\varrho_{ss} = 0$ y $\mathbf{x}_1$ mismo es el vector propio asociado.

Vemos que ϱ_{ss} converge linealmente a cero con respecto a $\mín_{1\leqslant j\leqslant n}|\lambda_j - \mu|$. Pero no aparecen problemas numéricos al determinar $\mathbf{x}_1$.

Tambien aquí se puede formular un teorema análogo al Teorema 5.12, es decir, cuando $\mu - \lambda_j \approx f(n)\vartheta\varepsilon$, tenemos $\mathbf{x}_1 \approx \mathbf{u}_j$ hasta un error del tipo $\mathcal{O}(\varepsilon)$, pero donde hay términos posiblemente grandes como amplificadores del error.

Ejemplo 5.10 (Tarea 29, Curso 2006)**.** *Se considera la matriz*

$$\mathbf{A} = \begin{bmatrix} 1000 & 10 & 1 \\ -1000 & -20 & -2 \\ 1000 & 20 & 3 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 1 & -1 & 1 \end{bmatrix}\begin{bmatrix} 1000 & 10 & 1 \\ 0 & -10 & -1 \\ 0 & 0 & 1 \end{bmatrix} = \mathbf{LR}.$$

a) *Ejecutar un paso del método de Wielandt para determinar el valor propio más pequeño de* $\mathbf{A}^{\mathrm{T}}\mathbf{A}$ *usando* $\mu = 0$*. Elegir el vector inicial para la iteración de Wielandt como* $(a, b, c)^{\mathrm{T}}$*,* $a, b, c = \pm 1$ *de tal forma que* $\|(\mathbf{R}^{-1})^{\mathrm{T}}\mathbf{x}_0\|_\infty$ *sea lo más grande posible.*

b) *Determinar una cota inferior realista para* $\|\mathbf{A}^{-1}\|_2$*.*

Solución sugerida.

a) *Usamos que*

$$\mathbf{R}^{-1} = \begin{bmatrix} 0{,}001 & 0{,}001 & 0 \\ 0 & -0{,}1 & -0{,}1 \\ 0 & 0 & 1 \end{bmatrix}, \quad (\mathbf{R}^{-1})^{\mathrm{T}} = \begin{bmatrix} 0{,}001 & 0 & 0 \\ 0{,}001 & -0{,}1 & 0 \\ 0 & -0{,}1 & 1 \end{bmatrix},$$

entonces $\mathbf{x}_0 = (1, -1, 1)^{\mathrm{T}}$*. Aprovechando que* $\mathbf{A}^{\mathrm{T}}\mathbf{A} = \mathbf{R}^{\mathrm{T}}\mathbf{L}^{\mathrm{T}}\mathbf{LR}$*, podemos resolver el sistema* $\mathbf{A}^{\mathrm{T}}\mathbf{A}\mathbf{x}_1 = \mathbf{x}_0$ *con*

$$\mathbf{A}^{\mathrm{T}}\mathbf{A} = \begin{bmatrix} 3000000 & 50000 & 6000 \\ 50000 & 900 & 110 \\ 6000 & 110 & 14 \end{bmatrix}$$

para obtener

$$\mathbf{x}_1 = \begin{pmatrix} 0{,}0014 \\ -0{,}3604 \\ 2{,}3010 \end{pmatrix}.$$

Entonces tenemos que

$$R(\mathbf{x}_1, \mathbf{A}^{\mathrm{T}}\mathbf{A}) = \frac{\mathbf{x}_1^{\mathrm{T}}\mathbf{A}\mathbf{x}_1}{\mathbf{x}_1^{\mathrm{T}}\mathbf{x}_1} = \frac{\mathbf{x}_1^{\mathrm{T}}\mathbf{x}_0}{\mathbf{x}_1^{\mathrm{T}}\mathbf{x}_1} = 0{,}4909,$$

lo que representa un valor aproximado del valor propio menor de $\mathbf{A}^{\mathrm{T}}\mathbf{A}$.

b) *Sabemos que el valor propio mas pequeño de* $\mathbf{A}^{\mathrm{T}}\mathbf{A}$ *satisface* $\lambda_{\mín}(\mathbf{A}^{\mathrm{T}}\mathbf{A}) \leqslant 0{,}4909$, *entonces* $\varrho(\mathbf{A}^{-1}\mathbf{A}^{-\mathrm{T}}) \geqslant 2{,}0371$. *Dado que para cada matriz* $\mathbf{B}$, $\mathbf{B}\mathbf{B}^{\mathrm{T}}$ *y* $\mathbf{B}^{\mathrm{T}}\mathbf{B}$ *tienen los mismos valores propios, sabemos ahora que*

$$\|\mathbf{A}^{-1}\|_2 = \sqrt{\lambda_{\máx}(\mathbf{A}^{-\mathrm{T}}\mathbf{A}^{-1})} = \sqrt{\lambda_{\máx}(\mathbf{A}^{-1}\mathbf{A}^{-\mathrm{T}})} \geqslant \sqrt{2{,}0371} = 1{,}4272.$$

Ejemplo 5.11 (Certamen 2, Curso 2010)**.** *Se considera la matriz*

$$\mathbf{A} = \begin{bmatrix} -10 & 1 & -1 \\ 1 & 2 & 1 \\ 1 & 1 & 13 \end{bmatrix}.$$

a) *Demostrar sin calcular el polinomio característico que* $\mathbf{A}$ *tiene tres valores propios reales distintos,* $\lambda_1 < \lambda_2 < \lambda_3$.

b) *Partiendo de* $\mathbf{x}_0 = (1,1,1)^{\mathrm{T}}$, *y eligiendo un valor* $\mu \in \{-7, 1, 8\}$ *apropiado (con justificación), calcular un paso de iteración inversa (método de Wielandt) para determinar una mejor aproximación del vector propio que corresponde a* λ_2.

c) *Utilizando el resultado de (b), calcular una mejor aproximación de* λ_2.

Solución sugerida.

a) *Los círculos de Gershgorin son*

$$\mathcal{K}_1 = \big\{z \in \mathbb{C} \,\big|\, |z+10| \leqslant 2\big\}, \quad \mathcal{K}_2 = \big\{z \in \mathbb{C} \,\big|\, |z-2| \leqslant 2\big\},$$
$$\mathcal{K}_3 = \big\{z \in \mathbb{C} \,\big|\, |z-13| \leqslant 2\big\}.$$

Dado que $\mathcal{K}_i \cap \mathcal{K}_j = \varnothing$ *para* $i \neq j$, *cada uno de los círculos contiene exactamente un valor propio, es decir* $\lambda_i \in \mathcal{K}_i$, $i = 1,2,3$, *por lo tanto* $\lambda_i \neq \lambda_j$ *para* $i \neq j$. *Además, los valores propios deben tener partes reales diferentes, por lo tanto sabemos que* $\lambda_1 \in [-12, -8]$, $\lambda_2 \in [0, 4]$ *y* $\lambda_3 \in [11, 15]$.

b) *De los tres valores propuestos, se debe escoger aquél que está mas cerca de* λ_2 *que de* λ_1 *o* λ_3. *De acuerdo al resultado anterior, sabemos que*

$$|\lambda_1 + 7| \leqslant 5, \quad |\lambda_2 + 7| \geqslant 7, \quad |\lambda_3 + 7| \geqslant 18;$$
$$|\lambda_1 - 1| \geqslant 9, \quad |\lambda_2 - 1| \leqslant 3, \quad |\lambda_3 - 1| \geqslant 10;$$
$$|\lambda_1 - 8| \geqslant 16, \quad |\lambda_2 - 8| \geqslant 4, \quad |\lambda_3 - 8| \geqslant 3.$$

Solamente el valor $\mu = 1$ *está mas cerca de* λ_2 *que de* λ_1 *o* λ_3. *La iteración inversa consiste en resolver el sistema* $(\mathbf{A} - \mu I)\mathbf{x}_1 = \mathbf{x}_0$, *en este caso*

$$\begin{bmatrix} -11 & 1 & -1 \\ 1 & 1 & 1 \\ 1 & 1 & 12 \end{bmatrix} \mathbf{x}_1 = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}, \quad \textit{con el resultado obvio} \quad \mathbf{x}_1 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}.$$

c) *Una mejor aproximación de* λ_2 *está dada por* $R(\mathbf{x}_1; \mathbf{A}) = 2$. *(Los verdaderos valores propios de* $\mathbf{A}$ *son* $\lambda_1 = -10{,}0398$, $\lambda_2 = 1{,}9925$ *y* $\lambda_3 = 13{,}0473$.*)*