Métodos Numéricos

Paul Bosch, Sergio Plaza

Segundo Semestre de 2008. Versión Preliminar en Revisión

Contenidos

Capítulo 1

Sistemas de Ecuaciones Lineales

Sistemas de ecuaciones lineales se utilizan en muchos problemas de ingeniería y de las ciencias, así como en aplicaciones de las matemáticas a las ciencias sociales y al estudio cuantitativo de problemas de administración y economía.

En este capítulo se examinan métodos iterativos y métodos directos para resolver sistemas de ecuaciones lineales

$$\left\{\begin{array}{rcl} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &=& b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &=& b_2 \\ & \vdots & \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n &=& b_n\,. \end{array}\right. \tag{1.1}$$

Este sistema puede ser expresado en la forma matricial como

$$\boxed{A\mathbf{x} = \mathbf{b}} \tag{1.2}$$

donde $A = (a_{ij})_{n\times n} \in M\,(n \times n, \mathbb{R})\,,\ \mathbf{b} = (b_1\,,\, b_2\,,\, \ldots\,,\, b_n)^T \in \mathbb{R}^n$ y $\mathbf{x} = (x_1\,,\, x_2\,,\, \ldots\,,\, x_n)^T \in \mathbb{R}^n$ es la incógnita.

En la siguiente sección estudiaremos algunos métodos directos para resolver sistemas de ecuaciones lineales, posteriormente abordaremos algunos métodos iterativos .

1.1 Normas matriciales

Para estudiar algunos métodos iterativos que nos permitan resolver sistemas de ecuaciones lineales, necesitamos de algunos conceptos básicos de Álgebra Lineal.

1.1.1 Normas vectoriales

Sea V un espacio vectorial de dimensión finita.

Definición 1.1 *Una norma en V es una función $||\ ||: V \longrightarrow \mathbb{R}$ que satisface*

N1) $\|x\| \geqslant 0$ *para todo* $x \in V$,

N2*)* $\|ax\| = |a|\ \|x\|$ *para todo* $x \in V$ *y todo* $a \in \mathbb{R}$,

N3*)* $\|x+y\| \leqslant \|x\| + \|y\|$ *para todo* $x, y \in V$ *(desigualdad triangular).*

Ejemplo 1 En $\mathbb{R}^n$ podemos definir las siguientes normas

1. $\| (x_1, x_2, \ldots, x_n) \|_2 = (\sum_{i=1}^n x_i^2)^{1/2}$ (*norma euclideana*)
2. $\| (x_1, x_2, \ldots, x_n) \|_\infty = \max\{\ |x_i|\ :\ 1 \leqslant i \leqslant n\ \}$
3. $\| (x_1, x_2, \ldots, x_n) \|_1 = \sum_{i=1}^n |x_i|$
4. $||(x_1, x_2, \ldots, x_n)||_p = (\sum_{i=1}^n |x_i|^p)^{1/p}$ (*norma* p).

Ejemplo 2 Denotemos por $V = M(n \times n, \mathbb{R})$ el espacio vectorial real de las matrices de orden $n \times n$ con entradas reales. Definimos las siguientes normas en V

1. $||A||_2 = \sqrt{\text{traza}(AA^T)}$
2. $||A||_F = \left(\sum_{i,j=1}^n a_{ij}^2\right)^{1/2} = \left(\sum_{i=1}^n \sum_{j=1}^n a_{ij}^2\right)^{1/2}$ (norma de Fröbenius)
3. $||A||_{\alpha,\beta} = \sup\{||A\mathbf{x}||_\beta\ :\ \mathbf{x} \in \mathbb{R}^n,\ \text{ con }\ \|\mathbf{x}\|_\alpha = 1\}$, donde $||\ ||_\beta$ y $||\ ||_\beta$, denotan normas cualesquiera norma sobre $\mathbb{R}^n$. Esta norma es llamada *norma subordinada* a las norma en $\mathbb{R}^n$.

   Notemos que también se tiene

$$\|A\|_{\alpha,\beta} = \sup\left\{\ \frac{\|A\mathbf{x}\|_\beta}{\|\mathbf{x}\|_\alpha}\ :\ \mathbf{x} \in \mathbb{R}^n,\ \ \|\mathbf{x}\|_\alpha \neq 0\ \right\}.$$

   La verificación que lo anterior define una norma sobre el espacio vectorial $M(n \times n, \mathbb{R})$ no es difícil, por ejemplo N1 es inmediata, para verificar N2, usamos la propiedad siguiente del supremo de conjuntos en $\mathbb{R}$, $\sup(\lambda A) = \lambda \sup(A)$, para todo $\lambda \geqslant 0$, donde $\lambda A = \{\lambda x : x \in A\}$. Finalmente, para verificar N3, usamos la propiedad $\sup(A+B) \leqslant \sup(A) + \sup(B)$, donde $A + B = \{\ x + y : x \in A,\ y \in B\ \}$.

Observación.

1. Si $A \in M(n \times n, \mathbb{R})$ es no singular, esto es, $\det(A) \neq 0$, entonces

$$\inf\{||A\mathbf{x}||\ :\ ||\mathbf{x}|| = 1\} = \frac{1}{||A^{-1}||}.$$

2. $||A||_2 = \sqrt{\lambda_{\max}}$, donde $\lambda_{\max}$ es el mayor valor propio de AA^T.
3. Si $A \in M(n \times n, \mathbb{R})$ es no singular, entonces

$$||A^{-1}||_2 = \frac{1}{\inf\{||A\mathbf{x}||_2\ :\ ||\mathbf{x}||_2 = 1\}} = \frac{1}{\sqrt{\lambda_{\min}}},$$

   donde $\lambda_{\min}$ es el menor valor propio de AA^T.

4. $||A||_2 = ||A^T||_2$.

5. $$\left|\left|\begin{pmatrix} A & 0 \\ 0 & B \end{pmatrix}\right|\right|_2 = \max\{||A||_2, ||B||_2\}\,.$$

Teorema 1.1 *Para* $A \in M(n\times n, \mathbb{R})$ *y* $\mathbf{x} \in \mathbb{R}^n$ *, se tiene que* $\|A\mathbf{x}\| \leqslant \|A\|\ \|\mathbf{x}\|$.

Demostración. Desde la definición de norma matricial se tiene que $||A|| \geqslant \frac{||A\mathbf{x}||}{||\mathbf{x}||}$, para todo $\mathbf{x} \in \mathbb{R}^n$, con $\mathbf{x} \neq 0$, de donde $||A\mathbf{x}|| \leqslant ||A||\ ||\mathbf{x}||$, y como esta última desigualdad vale trivialmente para $\mathbf{x} = 0$, el resultado se sigue.

Ejemplo 3 Si en $\mathbb{R}^n$ elegimos la $||\ ||_\infty$, entonces la norma matricial subordinada viene dada por

$$\begin{aligned} ||A||_\infty &= \sup\{\|A\mathbf{x}\|_\infty \ : \ \|\mathbf{x}\|_\infty = 1\} \\ &= \max\left\{\sum_{j=1}^n |a_{1j}|, \sum_{j=1}^n |a_{2j}|, \ldots, \sum_{j=1}^n |a_{ij}|, \ldots, \sum_{j=1}^n |a_{nj}|\right\}\,. \end{aligned}$$

Ejemplo 4 Si en $\mathbb{R}^n$ elegimos la $||\ ||_1$, entonces la norma matricial subordinada viene dada por

$$\begin{aligned} ||A||_1 &= \sup\{\|A\mathbf{x}\|_1 \ : \ \|\mathbf{x}\|_1 = 1\} \\ &= \max\left\{\sum_{i=1}^n |a_{i1}|, \sum_{i=1}^n |a_{i2}|, \ldots, \sum_{i=1}^n |a_{ij}|, \ldots, \sum_{i=1}^n |a_{in}|\right\}\,. \end{aligned}$$

Una de las normas más usadas en Algebra Lineal es la *norma de Fröbenius* la cual es dada por

$$||A||_F = \left(\sum_{i=1}^n \sum_{j=1}^m a_{ij}^2\right)^{1/2}, \quad A \in M(m\times n, \mathbb{R})$$

y también las p–normas ($p \geqslant 1$), las cuales son dadas por

$$||A||_p = \sup\left\{\frac{||A\mathbf{x}||_p}{||\mathbf{x}||_p} \ : \ \mathbf{x} \neq 0\right\}\,.$$

Teorema 1.2 *Para cualquier norma matricial subordinada se tiene*

1. $\|\ I\ \| = 1$.
2. $\|AB\| \leqslant \|A\| \cdot \|B\|$.

Demostración. Del teorema (1.1) para todo $\mathbf{x} \in \mathbb{R}^n$ se tiene que

$$||AB\mathbf{x}|| \leqslant ||A|| \cdot ||B\mathbf{x}|| \leqslant ||A|| \cdot ||B||\,||\mathbf{x}||\,,$$

de donde el resultado se sigue.

Ejemplo 5 No toda norma matricial es subordinada a alguna norma en $\mathbb{R}^n$. Para verlo, usamos la parte 2 del teorema anterior.

Definamos la norma matricial

$$||A||_\Delta = \max_{i,j=1,\ldots,n} |a_{i,j}|\,.$$

Afirmamos que esta es una norma matricial (se deja al lector la verificación) y no es subordinada a ninguna norma en $M(n\times n,\mathbb{R})$.

En efecto, tomemos las matrices

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

Tenemos $||A||_\Delta = ||B||_\Delta = 1$ y como

$$AB = \begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix}$$

se tiene que $||AB||_\Delta = 2 > ||A||_\Delta ||B||_\Delta = 1$.

Teorema 1.3 *Para las normas matriciales definidas anteriormente valen las siguientes relaciones. Sea $A \in M(n\times n,\mathbb{R})$, entonces*

1. $||A||_2 \leqslant ||A||_F \leqslant \sqrt{n}\,||A||_2$
2. $||A||_\Delta \leqslant ||A||_2 \leqslant n\,||A||_\Delta$
3. $\frac{1}{\sqrt{n}}\,||A||_\infty \leqslant ||A||_2 \leqslant \sqrt{n}\,||A||_\infty$
4. $\frac{1}{\sqrt{n}}\,||A||_1 \leqslant ||A||_2 \leqslant \sqrt{n}\,||A||_1$.

Teorema 1.4 *Sea $A \in M(n\times n,\mathbb{R})$, con $\|A\| < 1$, entonces $I-A$ es una matriz que tiene inversa, la cual viene dada por $(I-A)^{-1} = \sum_{k=0}^{\infty} A^k$, donde $A^0 = I$. Además, se tiene que $\left\|(I-A)^{-1}\right\| \leqslant \dfrac{1}{1-\|A\|}$.*

Demostración. Es sólo cálculo y se deja a cargo del lector.

Teorema 1.5 *Sean $A,\ B \in M(n\times n,\mathbb{R})$ tales que $\|I-AB\| < 1$ entonces A y B tienen inversas, las cuales son dadas por $A^{-1} = B\left(\sum_{k=0}^{\infty}(I-AB)^k\right)$ y $B^{-1} = \left(\sum_{k=0}^{\infty}(I-AB)^k\right)A$, respectivamente,.*

Demostración. Directa desde el teorema anterior.

Definición 1.2 *Sea $A \in M(n\times n,\mathbb{R})$, decimos que $\lambda \in \mathbb{C}$ es un valor propio de A si la matriz $A-\lambda I$ no es invertible, en otras palabras, λ es solución de la ecuación polinomial*

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

donde $p(\lambda)$ *es el polinomio característico de* A. *Se define el espectro de* A *como el conjunto de sus valores propios, es decir,*

$$\boxed{\sigma(A) = \{\lambda \in \mathbb{C} \ : \ \lambda \text{ valor propio de } A\}} \tag{1.3}$$

Observación. El polinomio característico de A, $p(\lambda)$, tiene grado menor o igual a n.

Definición 1.3 *Sea* $A \in M(n \times n, \mathbb{R})$, *definimos el radio espectral de* A *por*

$$\rho(A) = \max\{ \ |\lambda| \ : \ \lambda \in \sigma(A)\} \tag{1.4}$$

Observación. Geométricamente $\rho(A)$ es el menor radio tal que el círculo centrado en el origen en el plano complejo con radio $\rho(A)$ contiene todos los valores propios de A.

Observación. Si $\lambda = a + ib \in \mathbb{C}$, su *valor absoluto* o *norma* es dado por $|\lambda| = \sqrt{a^2 + b^2}$.

Teorema 1.6 *Se tiene* $\rho(A) = \inf \|A\|$, *donde el ínfimo es tomado sobre todas las normas matriciales definidas sobre el espacio vectorial* $M(n \times n, \mathbb{R})$.

Observación. Desde la definición de las norma matriciales $||\ ||_\infty$ y $||\ ||_1$ vemos que calcularlas para una matriz dada es fácil. Para lo norma $||\ ||_2$ el cálculo no es tan sencillo, pero tenemos el siguiente resultado.

Teorema 1.7 *Sea* $A \in M(n \times n, \mathbb{R})$. *Entonces*

$$||A||_2 = \left(\max\left\{ |\lambda| \ : \ \lambda \in \sigma(A^T A)\right\}\right)^{1/2}, \tag{1.5}$$

en otras palabras, $||A||_2$ *es la raíz cuadrada del mayor valor propio de la matriz* $A^T A$.

Corolario 1.1 *Si* $A \in M(n \times n, \mathbb{R})$ *es simétrica, entonces*

$$||A||_2 = \max\{|\lambda| \ : \ \lambda \in \sigma(A)\}.$$

1.2 Número de condición

Consideremos el sistema $A\mathbf{x} = \mathbf{b}$, donde A es una matriz invertible, por lo tanto tenemos que $\mathbf{x}_T = A^{-1}\mathbf{b}$ es la solución exacta.

Caso 1: Se perturba A^{-1} para obtener una nueva matriz B, la solución $\mathbf{x} = A^{-1}\mathbf{b}$ resulta perturbada y se obtiene un vector $\mathbf{x}_A = B\mathbf{b}$ que debería ser una solución aproximada al sistema original. Una pregunta natural ¿es que ocurre con $E(\mathbf{x}_A) = \|\mathbf{x}_T - \mathbf{x}_A\|$? Tenemos

$$\|\mathbf{x}_T - \mathbf{x}_A\| = \|\mathbf{x}_T - B\mathbf{b}\| = \|\mathbf{x}_T - BA\mathbf{x}_T\| = \|(I - BA)\,\mathbf{x}_T\| \leqslant \|I - BA\| \ \|\mathbf{x}_T\|$$

de donde obtenemos que

$$E(\mathbf{x}_A) = \|\mathbf{x}_T - \mathbf{x}_A\| \leqslant \|I - BA\| \ \|\mathbf{x}_T\|$$

y para el error relativo $E_R(\mathbf{x}_A)$, tenemos

$$\boxed{E_R(\mathbf{x}_A) = \frac{\|\mathbf{x}_T - \mathbf{x}_A\|}{\|\mathbf{x}_T\|} \leqslant \|I - BA\|} \tag{1.6}$$

Caso 2: Si perturbamos $\mathbf{b}$ para obtener un nuevo vector $\mathbf{b}_A$. Si $\mathbf{x}_T$ satisface $A\mathbf{x} = \mathbf{b}$ y $\mathbf{x}_A$ satisface $A\mathbf{x} = \mathbf{b}_A$. Una pregunta natural es determinar cotas para $E(\mathbf{x}_A) = \|\mathbf{x}_T - \mathbf{x}_A\|$ y $E_R(\mathbf{x}_A) = \frac{\|\mathbf{x}_T - \mathbf{x}_A\|}{\|\mathbf{x}_T\|}$. Tenemos que

$$\|\mathbf{x}_T - \mathbf{x}_A\| = \left\|A^{-1}\mathbf{b} - A^{-1}\mathbf{b}_A\right\| = \left\|A^{-1}(\mathbf{b} - \mathbf{b}_A)\right\| \leqslant \left\|A^{-1}\right\| \|\mathbf{b} - \mathbf{b}_A\|,$$

por lo tanto

$$E(\mathbf{x}_A) \leqslant \|A^{-1}\| \|\mathbf{b} - \mathbf{b}_A\| = \|A^{-1}\| E(\mathbf{b}_A),$$

esto es,

$$\boxed{E(\mathbf{x}_A) \leqslant \left\|A^{-1}\right\| E(\mathbf{b}_A)} \tag{1.7}$$

y por otro lado se tiene

$$\|\mathbf{x}_T - \mathbf{x}_A\| \leqslant \left\|A^{-1}\right\| \|\mathbf{b} - \mathbf{b}_A\| = \left\|A^{-1}\right\| \|A\mathbf{x}_T\| \frac{\|\mathbf{b} - \mathbf{b}_A\|}{\|\mathbf{b}\|} \leqslant \left\|A^{-1}\right\| \|A\| \|\mathbf{x}_T\| \frac{\|\mathbf{b} - \mathbf{b}_A\|}{\|\mathbf{b}\|}$$

luego

$$E_R(\mathbf{x}_T) = \frac{\|\mathbf{x}_T - \mathbf{x}_A\|}{\|\mathbf{x}_T\|} \leqslant \left\|A^{-1}\right\| \|A\| \frac{\|\mathbf{b} - \mathbf{b}_A\|}{\|\mathbf{b}\|} = \left\|A^{-1}\right\| \|A\| \, \mathrm{E}_R(\mathbf{b}_A).$$

esto es,

$$\boxed{E_R(\mathbf{x}_A) \leqslant ||A^{-1}|| \cdot ||A|| \cdot E_R(\mathbf{b}_A} \tag{1.8}$$

Definición 1.4 *Sea $A \in M(n \times n, \mathbb{R})$ una matriz invertible, el número condición de A es dado por*

$$\boxed{\kappa(A) = \|A\| \left\|A^{-1}\right\|} \tag{1.9}$$

Observación. Note que $\kappa(A) \geqslant 1$, pues es evidente que $\kappa(I) = 1$ y que

$$1 = ||I|| \leqslant ||A|| \, ||A^{-1}|| = \kappa(A).$$

Ejemplo 6 Considere

$$A = \begin{pmatrix} 1 & 1+\varepsilon \\ 1-\varepsilon & 1 \end{pmatrix},$$

con $\varepsilon > 0$, suficientemente pequeño. Tenemos que $\det(A) = \varepsilon^2 \approx 0$ es muy pequeño, lo cual significa que A es "casi singular". Como

$$A^{-1} = \varepsilon^{-2} \begin{pmatrix} 1 & -1-\varepsilon \\ -1+\varepsilon & 1 \end{pmatrix},$$

obtenemos que $\|A\|_\infty = 2+\varepsilon$, $\left\|A^{-1}\right\|_\infty = \varepsilon^{-2}(2+\varepsilon)$, por lo tanto

$$\kappa(A) = \frac{(2+\varepsilon)^2}{\varepsilon^2} \geqslant \frac{4}{\varepsilon}.$$

Observe que si $\varepsilon < 0.0001$, entonces $k(A) \geqslant 40.000$.

Ejemplo 7 Consideremos la matriz

$$A = \begin{pmatrix} 1 & 1 \\ 1 & 1.00000001 \end{pmatrix}$$

Tenemos

$$A^{-1} = \begin{pmatrix} 1.0 \times 10^8 & -1.0 \times 10^8 \\ -1.0 \times 10^8 & 1.0 \times 10^8 \end{pmatrix}$$

Luego, $||A||_1 = ||A||_2 \approx ||A||_\infty = 2$, $||A^{-1}||_1 = ||A^{-1}||_\infty \approx ||A^{-1}||_2 \approx 2 \times 10^8$, luego $\kappa_1(A) = \kappa_\infty(A) \approx 4 \times 10^8$.

Observación. Si $\kappa(A)$ es demasiado grande diremos que A está *mal condicionada*.

Ejemplo 8 Consideremos la matriz

$$A = \begin{pmatrix} 1 & -1 & -1 & \cdots & -1 \\ 0 & 1 & -1 & \cdots & -1 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & -1 \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix}.$$

La matriz A tiene sólo unos en su diagonal, y sobre ella tiene sólo menos unos. Tenemos que $\det(A) = 1$, pero $\kappa_\infty(A) = n2^{n-1}$, esta matriz está mal condicionada.

Si resolvemos $A\mathbf{x} = b$ numéricamente no obtenemos una solución exacta, $\mathbf{x}_T$, si no una solución aproximada $\mathbf{x}_A$. Una pregunta natural al resolver numéricamente el sistema es ¿qué tan cerca esta $A\mathbf{x}_A$ de $\mathbf{b}$?

Para responder dicha interrogante definamos $\mathbf{r} = \mathbf{b} - A\mathbf{x}_A$ como el *vector residual* y $\mathbf{e} = \mathbf{x}_T - \mathbf{x}_A$ como el *vector de error*.

Observación. Tenemos que

1. $A\mathbf{e} = \mathbf{r}$.
2. $\mathbf{x}_A$ es solución exacta de $A\mathbf{x}_A = \mathbf{b}_A = \mathbf{b} - \mathbf{r}$.

¿Qué relación existe entre $\frac{\|\mathbf{x}_T-\mathbf{x}_A\|}{\|\mathbf{x}_T\|}$ y $\frac{\|\mathbf{b}-\mathbf{b}_A\|}{\|\mathbf{b}\|} = \frac{\|\mathbf{r}\|}{\|\mathbf{b}\|}$?

Teorema 1.8 *Para toda matriz invertible* $A \in M(n\times n,\mathbb{R})$, *se tiene que*

$$\boxed{\frac{1}{\kappa(A)}\frac{\|\mathbf{b}-\mathbf{b}_A\|}{\|\mathbf{b}\|} \leqslant \frac{\|\mathbf{x}_T-\mathbf{x}_A\|}{\|\mathbf{x}_T\|} \leqslant \kappa(A)\frac{\|\mathbf{b}-\mathbf{b}_A\|}{\|\mathbf{b}\|}}, \tag{1.10}$$

es decir,

$$\boxed{\frac{1}{\kappa(A)}\frac{\|\mathbf{r}\|}{\|\mathbf{b}\|} \leqslant \frac{\|\mathbf{e}\|}{\|\mathbf{x}_T\|} \leqslant \kappa(A)\frac{\|\mathbf{r}\|}{\|\mathbf{b}\|}} \tag{1.11}$$

Observación. Si tenemos una matriz B escrita en la forma

$$B = A(I+E)$$

donde I denota la matriz identidad y E es una matriz de error, entonces se tienen las siguientes cotas para el error relativo

$$\boxed{\frac{||\mathbf{x}_T-\mathbf{x}_A||}{||\mathbf{x}_T||} \leqslant \frac{\kappa(A)}{1-\kappa(A)\frac{||AE||}{||A}}\frac{||AE||}{||A||}} \tag{1.12}$$

si $||AE|| < 1$ y

$$\boxed{\frac{||\mathbf{x}_T-\mathbf{x}_A||}{||\mathbf{x}_T||} \leqslant \frac{||E||}{1-||E||}} \tag{1.13}$$

si $||E|| < 1$.

Observación. Sea $A \in M(n\times n,\mathbb{R})$. Denotemos por $\lambda_{\max}$ y $\lambda_{\min}$, respectivamente, el máximo y el mínimo de los valores propios de AA^T. Entonces $\kappa_2(A) = \sqrt{\frac{\lambda_{\max}}{\lambda_{\min}}}$.

Ejemplo 9 Determine $||A||_2$, $||A^{-1}||_2$ y $\kappa_2(A)$, donde

$$A = \begin{pmatrix} \frac{3}{\sqrt{3}} & -\frac{1}{\sqrt{3}} \\ 0 & \frac{\sqrt{8}}{\sqrt{3}} \end{pmatrix}$$

Tenemos que

$$A^T = \begin{pmatrix} \frac{3}{\sqrt{3}} & 0 \\ -\frac{1}{\sqrt{3}} & \frac{\sqrt{8}}{\sqrt{3}} \end{pmatrix}$$

Luego

$$AA^T - \lambda \mathrm{I} = \begin{pmatrix} \frac{10}{3}-\lambda & -\frac{\sqrt{8}}{3} \\ -\frac{\sqrt{8}}{3} & \frac{8}{3}-\lambda \end{pmatrix}.$$

Por lo tanto, $\det(AA^T - \lambda \mathrm{I}) = \lambda^2 - 6\lambda + 8$, de donde los valores propios de AA^T son $\lambda = 2$ y $\lambda = 4$, es decir, $\lambda_{\min} = 2$ y $\lambda_{\max} = 4$. Luego

$$||A||_2 = \sqrt{\lambda_{\max}} = 2 \quad \text{y} \quad ||A^{-1}||_2 = \frac{1}{\sqrt{\lambda_{\min}}} = \frac{1}{\sqrt{2}},$$

y se tiene que $\kappa_2(A) = \frac{2}{\sqrt{2}} = \sqrt{2}$.

1.3 Solución de sistemas de ecuaciones lineales: métodos directos

En esta parte estudiaremos métodos directos, es decir, algebraicos, para resolver sistemas de ecuaciones lineales. Herramienta fundamental aquí es el Álgebra Lineal.

1.3.1 Conceptos básicos

La idea es reducir un sistema de ecuaciones lineales a otro que sea más sencillo de resolver.

Definición 1.5 *Sean $A, B \in M(n \times n, \mathbb{R})$. Decimos que dos sistemas de ecuaciones lineales $A\mathbf{x} = \mathbf{b}$ y $B\mathbf{x} = \mathbf{d}$ son equivalentes si poseen exactamente las mismas soluciones.*

Por lo tanto si podemos reducir un sistema de ecuaciones lineales $A\mathbf{x} = \mathbf{b}$ a un sistema equivalente y más simple $B\mathbf{x} = \mathbf{d}$, podemos resolver este último, y obtenemos de ese modo las soluciones del sistema original.

Definición 1.6 *Sea $A \in M(n \times n, \mathbb{R})$. Llamaremos operaciones elementales por filas sobre A a cada una de las siguientes operaciones*

1. *Intercambio de la fila i con la fila j, denotamos esta operación por $F_i \leftrightarrow F_j$.*
2. *Reemplazar la fila i, F_i, por un múltiplo no nulo λF_i de la fila i, denotamos esta operación por $F_i \mapsto \lambda F_i$*
3. *Reemplazar la fila i, F_i, por la suma de la fila i más un múltiplo no nulo λF_j de la fila j, $i \neq j$, denotamos esta operación por $F_i \mapsto F_i + \lambda F_j$*

Definición 1.7 *Decimos que una matriz A es una matriz elemental si ella se obtiene a partir de la matriz identidad a través de una única operación elemental. Una matriz elemental será denotada por E.*

Ejemplo 10 1. $\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} F_3 \leftrightarrow F_2 \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}$.

2. $\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} F_2 \mapsto rF_2 \begin{pmatrix} 1 & 0 & 0 \\ 0 & r & 0 \\ 0 & 0 & 1 \end{pmatrix}$

3. $\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \; F_3 \mapsto F_3 + rF_2 \; \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & r & 1 \end{pmatrix}$

Observación. Si a una matriz A se le aplican una serie de operaciones elementales $E_1, E_2, \ldots, E_k$ denotaremos el resultado por $E_k E_{k-1} \cdots E_2 E_1 A$.

Además, si $E_k E_{k-1} \cdots E_2 E_1 A = I$, se tiene que A posee inversa y $E_k E_{k-1} \cdots E_2 E_1 = A^{-1}$.

El teorema siguiente resume las condiciones para que una matriz tenga inversa.

Teorema 1.9 *Sea $A \in M(n \times n, \mathbb{R})$ entonces son equivalentes*

1. *A tiene inversa.*
2. $\det(A) \neq 0$.
3. *Las filas de A forman una base de $\mathbb{R}^n$.*
4. *Las columnas de A forman una base de $\mathbb{R}^n$.*
5. *La transformación lineal $T : \mathbb{R}^n \longrightarrow \mathbb{R}^n$ dada por $T(\mathbf{x}) = A\mathbf{x}$, asociada a la matriz A, es inyectiva,*
6. *La transformación $T : \mathbb{R}^n \longrightarrow \mathbb{R}^n$ dada por $T(\mathbf{x}) = A\mathbf{x}$, asociada a la matriz A, es sobreyectiva*
7. *El sistema $A\mathbf{x} = 0$, donde $\mathbf{x} \in \mathbb{R}^n$ posee solución única $\mathbf{x} = 0$.*
8. *Para cada $\mathbf{b} \in \mathbb{R}^n$ existe un único vector $\mathbf{x} \in \mathbb{R}^n$ tal que $A\mathbf{x} = \mathbf{b}$.*
9. *A es producto de matrices elementales.*
10. *Todos los valores propios de A son distinto de cero.*

1.4 Factorización de matrices

En esta sección estudiaremos las condiciones necesarias para que una matriz $A \in M(n \times n, \mathbb{R})$ tenga una factorización de la forma LU donde $L, U \in M(n \times n, \mathbb{R})$, con L una matriz triangular inferior y U una matriz triangular superior.

Observe que si $A \in M(n \times n, \mathbb{R})$ tiene una factorización de la forma LU como arriba, entonces el sistema de ecuaciones lineales $A\mathbf{x} = \mathbf{b}$, con $\mathbf{x}, \mathbf{b} \in \mathbb{R}^n$, puede resolverse de una manera más sencilla, para ello consideramos el cambio de variable $U\mathbf{x} = \mathbf{z}$, y resolvemos primero el sistema $L\mathbf{z} = \mathbf{b}$ primero y enseguida resolvemos el sistema $U\mathbf{x} = \mathbf{z}$, obteniendo así la solución del sistema original, es decir,

$$\boxed{A\mathbf{x} = \mathbf{b} \Longleftrightarrow L\underbrace{U\mathbf{x}}_{\mathbf{z}} = \mathbf{b} \Longleftrightarrow \begin{cases} L\mathbf{z} = \mathbf{b} \\ U\mathbf{x} = \mathbf{z}. \end{cases}} \tag{1.14}$$

Supongamos que $A \in M(n \times n, \mathbb{R})$ tiene una descomposición de la forma LU como arriba, con

$$L=\begin{pmatrix} l_{11} & 0 & 0 & \cdots & 0 \\ l_{21} & l_{22} & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ l_{n1} & l_{n2} & \cdots & l_{nn-1} & l_{nn} \end{pmatrix} \text{ y } U=\begin{pmatrix} u_{11} & u_{12} & \cdots & u_{1n-1} & u_{1n} \\ 0 & u_{22} & \cdots & u_{2n-1} & u_{2n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & u_{nn} \end{pmatrix} \tag{1.15}$$

De esta descomposición tenemos

$$\boxed{\det(A)=\det(L)\det(U)=\left(\prod_{i=1}^{n} l_{ii}\right)\left(\prod_{i=1}^{n} u_{ii}\right)} \tag{1.16}$$

así, $\det(A)\neq 0$ si y sólo si $l_{ii}\neq 0$ y $u_{ii}\neq 0$ para todo $i=1,\ldots,n$.

Para simplificar la notación, escribamos

$$L=(F_1\ F_2\ \ldots\ F_n)^T \quad \text{y} \quad U=(C_1C_2\ \cdots\ C_n)\ ,$$

donde $F_i=(l_{i1}\ l_{i2}\ \cdots\ l_{ii}\ 0\ \cdots\ 0)$ y $C_i=(u_{1i}\ u_{2i}\ \cdots\ u_{ii}\ 0\ \cdots\ 0)^T$, son las filas de L y las columnas de U, respectivamente.

Ahora, al desarrollar la multiplicación $A=LU$,

$$\begin{pmatrix} a_{11} & a_{21} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix}=\begin{pmatrix} F_1 \\ F_2 \\ \vdots \\ F_n \end{pmatrix}\ (C_1C_2\ \cdots\ C_n)=\begin{pmatrix} F_1C_1 & F_1C_2 & \cdots & F_1C_n \\ F_2C_1 & F_2C_2 & \cdots & F_2C_n \\ \vdots & \vdots & \ddots & \vdots \\ F_nC_1 & F_nC_2 & \cdots & F_nC_n \end{pmatrix}$$

donde F_iC_j es considerado como el producto de las matrices $F_i=(l_{i1}\ l_{i2}\ \cdots l_{ii}\ 0\ \cdots 0)$ y $C_i=(u_{1i}\ u_{2i}\ \cdots\ u_{ii}\ 0\ \cdots\ 0)^T$. Tenemos entonces

$$\begin{cases} a_{11} &=& F_1C_1=l_{11}u_{11} \\ a_{12} &=& F_1C_2=l_{11}u_{12} \\ \vdots & \vdots & \vdots \\ a_{1n} &=& F_1C_n=l_{11}u_{1n} \end{cases} \tag{1.17}$$

de aquí, vemos que si fijamos $l_{11}\neq 0$, podemos resolver las ecuaciones anteriores para u_{1i}, $i=1,\ldots,n$. Enseguida, para la segunda fila de A tenemos

$$\begin{cases} a_{21} &=& F_2C_1=l_{21}u_{11} \\ a_{22} &=& F_2C_2=l_{21}u_{12}+l_{22}u_{22} \\ \vdots & \vdots & \vdots \\ a_{2n} &=& F_2C_n=l_{21}u_{1n}+l_{22}u_{2n} \end{cases} \tag{1.18}$$

como u_{11} ya lo conocemos desde las ecuaciones (1.17), podemos encontrar el valor l_{21}, conocido este valor, fijando un valor no cero para l_{22}, y dado que conocemos los valores u_{1i} para $i=2,\ldots,n$, podemos encontrar los valores de u_{2i} para $i=2,\ldots,n$. Siguiendo este método

vemos que es posible obtener la descomposición de A en la forma LU. Note que también podemos comparar las columnas de A con aquellas obtenidas desde el producto LU y proceder a resolver las ecuaciones resultantes. El problema es ahora saber bajo que condiciones dada una matriz A ella puede descomponerse en la forma LU, es decir, siempre y cuando podamos hacer las elecciones anteriores.

Ejemplo 11 Sea $A = \begin{pmatrix} 10 & -3 & 6 \\ 1 & 8 & -2 \\ -2 & 4 & -9 \end{pmatrix}$. Encontremos una descomposición LU para A. Escribamos

$$\begin{pmatrix} 10 & -3 & 6 \\ 1 & 8 & -2 \\ -2 & 4 & -9 \end{pmatrix} = \begin{pmatrix} l_{11} & 0 & 0 \\ l_{21} & l_{22} & 0 \\ l_{31} & l_{32} & l_{33} \end{pmatrix} \begin{pmatrix} u_{11} & u_{12} & u_{13} \\ 0 & u_{22} & u_{23} \\ 0 & 0 & u_{33} \end{pmatrix}$$

desarrollando, obtenemos que

$$\begin{cases} l_{11}u_{11} &= 10 \\ l_{11}u_{12} &= -3 \\ l_{11}u_{13} &= 6 \end{cases}$$

de aquí fijando un valor no cero para l_{11}, digamos $l_{11} = 2$, obtenemos que $u_{11} = 5$, $u_{12} = -\frac{3}{2}$, y $u_{13} = 3$. Enseguida tenemos las ecuaciones

$$\begin{cases} l_{21}u_{11} &= 1 \\ l_{21}u_{12} + l_{22}u_{22} &= 8 \\ l_{21}u_{13} + l_{22}u_{23} &= -2 \end{cases}$$

como $u_{11} = 5$ se tiene que $l_{21} = \frac{1}{5}$. Ahora fijamos el valor de $l_{22} = 1$ y obtenemos los valores $u_{22} = \frac{83}{10}$, y $u_{23} = -\frac{13}{5}$. Finalmente, para la última fila de A, nos queda

$$\begin{cases} l_{31}u_{11} &= -2 \\ l_{31}u_{12} + l_{32}u_{22} &= 4 \\ l_{31}u_{13} + l_{32}u_{23} + l_{33}u_{33} &= -2 \end{cases}$$

como $u_{11} = 5$ de la primera ecuación $l_{31} = -\frac{2}{5}$, reemplazando los valores de $l_{31} = -\frac{2}{5}$, $u_{12} = -\frac{3}{2}$ y $u_{22} = \frac{83}{10}$ en la segunda ecuación encontramos que $l_{32} = \frac{34}{83}$. Ahora, reemplazando los valores ya obtenidos en la tercera ecuación obtenemos que $l_{33}u_{33} = -\frac{1694}{415}$, para encontrar el valor de u_{33} podemos fijar el valor de l_{33}, digamos $l_{33} = -\frac{1}{415}$, y obtenemos que $u_{33} = 1604$. Tenenmos así una descomposición LU para la matriz A, es decir,

$$\begin{pmatrix} 10 & -3 & 6 \\ 1 & 8 & -2 \\ -2 & 4 & -9 \end{pmatrix} = \begin{pmatrix} 2 & 0 & 0 \\ \frac{1}{5} & 1 & 0 \\ -\frac{3}{5} & \frac{34}{83} & -\frac{1}{415} \end{pmatrix} \begin{pmatrix} 5 & -\frac{3}{2} & 3 \\ 0 & \frac{83}{10} & -\frac{13}{5} \\ 0 & 0 & 1604 \end{pmatrix}.$$

No siempre es posible encontrar una descomposición de la forma LU para una matriz A dada, esto se muestra en el siguiente ejemplo.

Ejemplo 12 La matriz $A = \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix}$ no tiene descomposición de la forma LU. Notemos que $\det(A) \neq 0$.

Para verlo supongamos que podemos escribir $A = LU$, donde

$$L = \begin{pmatrix} l_{11} & 0 \\ l_{21} & l_{22} \end{pmatrix} \quad \text{y} \quad U = \begin{pmatrix} u_{11} & u_{12} \\ 0 & u_{22} \end{pmatrix}.$$

Tenemos entonces que

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

de donde $l_{11}u_{11} = 0$, por lo tanto

$$l_{11} = 0 \ \ (*) \quad \text{o} \quad u_{11} = 0 \ \ (**).$$

Como $l_{11}u_{12} = 1$, no es posible que (*) sea verdadero, es decir, debemos tener que $l_{11} \neq 0$, en consecuencia se debe tener que $u_{11} = 0$, pero además se tiene que $u_{12} \neq 0$. Ahoram, como $l_{21}u_{11} = 1$, llegamos a una contradicción. Luego tal descomposición para A no puede existir.

En el algoritmo que vimos para buscar la descomposición de una matriz A en la forma LU, en los sucesivos paso fue necesario fijar valores no cero para los coeficientes l_{ii} de L, de modo a poder resolver las ecuaciones resultantes. También, si procedemos con el algoritmo a comparar las columnas de la matriz producto LU con las columnas de A, vemos que será necesario fijar en paso sucesivos valores no cero para los coeficientes u_{ii} de U, de modo a poder resolver las ecuaciones resultantes. Como es evidente existe una manera simple de hacer esto, por ejemplo si en el primer algoritmo fijamos los valores $l_{ii} = 1$ para $i = 1, \ldots, n$, decimos que tenemos una *descomposición LU de Doolitle* para A, y si en el segundo algoritmo fijamos los valores $u_{ii} = 1$ para $i = 1, \ldots, n$, decimos que tenemos una *descomposición LU de Crout* para A, finalmente si fijamos $U = L^T$ (transpuesta de L) se tiene que $l_{ii} = u_{ii}$, y decimos que tenemos una *descomposición LU de Cholesky* para A. Note que para la existencia de descomposición de Cholesky es necesario que A sea simétrica, pues si $A = LL^T$ entonces $A^T = A$, esto significa que es una condición necesario, pero una no es una condición suficiente.

Sobre la existencia de descomposición LU para una matriz A veremos a continuación algunos teoremas. Primero recordemos que si

$$A = \begin{pmatrix} a_{11} & a_{21} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix}_{n \times n}$$

entonces los menores principales de A son las submatrices A_k, $k = 1, \ldots, n$, definidas por

$$A_1 = (a_{11})_{1\times 1}\,,\ A_2 = \begin{pmatrix} a_{11} & a_{21} \\ a_{21} & a_{22} \end{pmatrix}_{2\times 2}\,,\ \ldots\,,\ A_k = \begin{pmatrix} a_{11} & a_{21} & \cdots & a_{1k} \\ a_{21} & a_{22} & \cdots & a_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ a_{k1} & a_{k2} & \cdots & a_{kk} \end{pmatrix}_{k\times k},$$

$$\ldots\,,\ A_n = A\,.$$

Teorema 1.10 *Si los n menores principales de una matriz A tienen determinante distinto de cero, entonces A tiene una descomposición LU.*

Para tener descomposición de Cholesky, recordemos algunos conceptos y resultados.

Definición 1.8 *Sea $A \in M(n\times n, \mathbb{R})$. Decimos que A es definida positiva si $\langle A\mathbf{x}, \mathbf{x}\rangle > 0$ para todo $\mathbf{x} \in \mathbb{R}^n$, con $\mathbf{x} \neq 0$.*

Ejemplo 13 La matriz

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

es definida positiva. En efecto considere $\mathbf{x} = (x\ y\ z)^T$, tenemos

$$\langle \mathbf{x}, A\mathbf{x}\rangle = (x, y, z) \begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = 2x^2 - 2xy + 2y^2 - 2yz + 2z^2.$$

Luego,

$$\begin{aligned} \langle \mathbf{x}, A\mathbf{x}\rangle &= 2x^2 - 2xy + 2y^2 - 2yz + 2z^2 \\ &= x^2 + (x-y)^2 + (y-z)^2 + z^2 > 0, \end{aligned}$$

para todo $(x, y, z) \neq (0, 0, 0)$.

Teorema 1.11 *Si $A \in M(n\times n, \mathbb{R})$ es definida positiva entonces todos sus valores propios son reales y positivos.*

Observación. Si $A \in M(n\times n, \mathbb{R})$ entonces podemos descomponer A de manera única como la suma de una matriz simétrica, $A_0 = \frac{1}{2}\left(A + A^T\right)$, y una matriz antisimétrica, $A_1 = \frac{1}{2}\left(A - A^T\right)$. Desde la definición de matriz positiva definida se tiene que $\langle A\mathbf{x}, \mathbf{x}\rangle = \langle A_0\mathbf{x}, \mathbf{x}\rangle$, por lo cual sólo necesitamos considerar matrices simétricas.

Teorema 1.12 *Sea $A \in M(n\times n, \mathbb{R})$, entonces A es positiva definida si y sólo si todos los menores principales de A tiene determinante positivo.*

Teorema 1.13 *Sea $A \in M(n\times n, \mathbb{R})$, entonces A simétrica y positiva definida si y sólo si A tiene una descomposición de Cholesky.*

Ejemplo 14 La matriz

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

es definida positiva, pues se tiene que $\det(A_1) = 2 > 0$, $\det(A_2) = \det\begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix} = 3 > 0$ y $\det(A_3) = \det(A) = 4 > 0$. Luego tiene descomposición de Cholesky.

Ahora, para encontrar la descomposición de Cholesky para esta matriz procedemos como sigue.

$$\begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{pmatrix} = \begin{pmatrix} l_{11} & 0 & 0 \\ l_{21} & l_{22} & 0 \\ l_{31} & l_{32} & l_{33} \end{pmatrix} \begin{pmatrix} l_{11} & l_{21} & l_{31} \\ 0 & l_{22} & l_{32} \\ 0 & 0 & l_{33} \end{pmatrix}$$

Luego, $l_{11}^2 = 2$, es decir, $l_{11} = \sqrt{2}$; $-1 = l_{11}l_{21}$, de donde $l_{21} = -\frac{\sqrt{2}}{2}$; $0 = l_{11}l_{31}$, de donde $l_{31} = 0$; $l_{21}^2 + l_{22}^2 = 2$, de donde $l_{22} = \frac{\sqrt{5}}{\sqrt{2}}$; $l_{21}l_{31} + l_{22}l_{32} = -1$, de donde $l_{32} = -\frac{\sqrt{2}}{\sqrt{5}}$; $l_{31}^2 + l_{32}^2 + l_{33}^2 = 2$, de donde $l_{33} = \frac{2\sqrt{2}}{\sqrt{5}}$.

Por lo tanto

$$\begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{pmatrix} = \begin{pmatrix} \sqrt{2} & 0 & 0 \\ -\frac{\sqrt{2}}{2} & \frac{\sqrt{5}}{\sqrt{2}} & 0 \\ 0 & -\frac{\sqrt{2}}{\sqrt{5}} & \frac{2\sqrt{2}}{\sqrt{5}} \end{pmatrix} \begin{pmatrix} \sqrt{2} & -\frac{\sqrt{2}}{2} & 0 \\ 0 & \frac{\sqrt{5}}{\sqrt{2}} & -\frac{\sqrt{2}}{\sqrt{5}} \\ 0 & 0 & \frac{2\sqrt{2}}{\sqrt{5}} \end{pmatrix}.$$

1.5 Método de eliminación gaussiana

En esta sección estudiaremos el método de eliminación gaussiana para resolver sistemas de ecuaciones lineales

$$\begin{pmatrix} a_{11} & a_{21} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix}.$$

Colocamos primero esta información en la forma

$$\left(\begin{array}{c|c} A & b \end{array} \right) = \left(\begin{array}{cccc|c} a_{11} & a_{12} & \cdots & a_{1n} & b_1 \\ a_{21} & a_{22} & \cdots & a_{2n} & b_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} & b_n \end{array} \right) \begin{matrix} F_1 \\ F_2 \\ \vdots \\ F_n \end{matrix}$$

esto es, consideramos la matriz aumentada del sistema. En la notación anterior, los símbolos F_i denotan las filas de la nueva matriz.

Si $a_{11} \neq 0$, el primer paso de la eliminación gaussiana consiste en realizar las operaciones elementales

$$(F_j - m_{j1}F_1) \to F_j, \text{ donde } m_{j1} = \frac{a_{j1}}{a_{11}}, \quad j = 2, 3, ..., n. \tag{1.19}$$

Estas operaciones transforman el sistema en otro en el cual todos los componentes de la primera columna situada bajo la diagonal son cero.

Podemos ver desde otro punto de vista el sistema de operaciones. Esto lo logramos simultáneamente al multiplicar por la izquierda la matriz original A por la matriz

$$M^{(1)} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ -m_{21} & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ -m_{n1} & 0 & \cdots & 1 \end{pmatrix} \tag{1.20}$$

Esta matriz recibe el nombre de primera matriz gaussiana de transformación. El producto de $M^{(1)}$ con la matriz $A^{(1)} = A$ lo denotamos por $A^{(2)}$ y el producto de $M^{(1)}$ con la matriz $b^{(1)} = b$ lo denotamos por $b^{(2)}$, con lo cual obtenemos

$$A^{(2)}\mathbf{x} = M^{(1)}A\mathbf{x} = M^{(1)}b = b^{(2)} \tag{1.21}$$

De manera análoga podemos construir la matriz $M^{(2)}$, en la cual si $a_{22}^{(2)} \neq 0$ los elementos situados bajo de la diagonal en la segunda columna con los elementos negativos de los multiplicadores

$$m_{j2} = \frac{a_{j2}^{(2)}}{a_{22}^{(2)}}, \quad j = 3, 4, \ldots, n \tag{1.22}$$

Así obtenemos

$$A^{(3)}\mathbf{x} = M^{(2)}A^{(2)}\mathbf{x} = M^{(2)}M^{(1)}A\mathbf{x} = M^{(2)}M^{(1)}b = b^{(3)} \tag{1.23}$$

En general, con $A^{(k)}\mathbf{x} = b^{(k)}$ ya formada, multiplicamos por la k–ésima matriz de transformación gaussiana

$$M^{(k)} = \begin{pmatrix} 1 & 0 & & \cdots & & \cdots & 0 \\ 0 & 1 & & & & & \\ & \ddots & 1 & \ddots & & & \vdots \\ & & 0 & & & & \\ \vdots & & & -m_{k+1\,k} & \ddots & & \vdots \\ & & \vdots & \vdots & 0 & & 0 \\ 0 & \cdots & 0 & -m_{n\,k} & 0 & 0 & 1 \end{pmatrix}$$

para obtener

$$A^{(k+1)}\mathbf{x} = M^{(k)}A^{(k)}\mathbf{x} = M^{(k)}\cdots M^{(1)}A\mathbf{x} = M^{(k)}b^{(k)} = b^{(k+1)} = M^{(k)}\cdots M^{(1)}b \qquad (1.24)$$

Este proceso termina con la formación de un sistema $A^{(n)}\mathbf{x} = b^{(n)}$, donde $A^{(n)}$ es una matriz triangular superior

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

dada por

$$A^{(n)} = M^{(n-1)}M^{(n-2)}M^{(n-3)}\cdots M^{(1)}A. \qquad (1.25)$$

El proceso que hemos realizado sólo forma la mitad de la factorización matricial $A = LU$, donde con U denotamos la matriz triangular superior $A^{(n)}$. Si queremos determinar la matriz triangular inferior L, primero debemos recordar la multiplicación de $A^{(k)}\mathbf{x} = b^{(k)}$ mediante la transformación gaussiana $M^{(k)}$ con que obtuvimos

$$A^{(k+1)}\mathbf{x} = M^{(k)}A^{(k)}\mathbf{x} = M^{(k)}b^{(k)} = b^{(k+1)}$$

para poder revertir los efectos de esta transformación debemos considerar primero la matriz $L^{(k)} = \left(M^{(k)}\right)^{-1}$, la cual esta dada por

$$L^{(k)} = \left(M^{(k)}\right)^{-1} = \begin{pmatrix} 1 & 0 & \cdots & \cdots & & \cdots & 0 \\ 0 & 1 & & & & & \vdots \\ \vdots & & 1 & & \vdots & & \\ & \vdots & 0 & \ddots & & & \vdots \\ \vdots & & & m_{k+1\ k} & \ddots & & \\ & \vdots & \vdots & \vdots & 0 & \ddots & 0 \\ 0 & \cdots & 0 & m_{n\ k} & 0 & 0 & 1 \end{pmatrix}.$$

Por lo tanto si denotamos

$$L = L^{(1)}L^{(2)}\cdots L^{(n-1)}$$

obtenemos

$$LU = L^{(1)}L^{(2)}\cdots L^{(n-1)}M^{(n-1)}M^{(n-2)}M^{(n-3)}\cdots M^{(1)} = A.$$

Con lo cual obtenemos el siguiente resultado.

Teorema 1.14 *Si podemos efectuar la eliminación gaussiana en el sistema lineal* $A\mathbf{x} = b$ *sin intercambio de filas, entonces podemos factorizar la matriz* A *como el producto de una matriz triangular inferior* L *y una matriz triangular superior* U*, donde*

$$U = \begin{pmatrix} a_{11}^{(1)} & a_{12}^{(1)} & \cdots & a_{1n}^{(1)} \\ 0 & a_{22}^{(2)} & \ddots & \vdots \\ \vdots & \vdots & \ddots & a_{n-1\,n}^{(n-1)} \\ 0 & \cdots & 0 & a_{nn}^{(n)} \end{pmatrix} \quad y \quad L = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ m_{21} & 1 & \vdots & \vdots \\ \vdots & \vdots & \ddots & 0 \\ m_{n1} & \cdots & m_{n\,n-1} & 1 \end{pmatrix}.$$

Ejemplo 15 Consideremos el siguiente sistema lineal

$$\begin{aligned} x + y + 3w &= 4 \\ 2x + y - z + w &= 1 \\ 3x - y - z + w &= -3 \\ -x + 2y + 3z - w &= 4 \end{aligned}$$

Si realizamos las siguientes operaciones elementales $F_2 \mapsto (FE_2 - 2F_1)$, $F_3 \mapsto (F_3 - 3F_1)$, $F_4 \mapsto (F_4 - (-1)\,F_1)$, $F_3 \mapsto (F_3 - 4F_2)$, $F_4 \mapsto (F_4 - (-3)\,F_2)$ con lo que obtenemos el sistema triangular

$$\begin{aligned} x + y + 3w &= 4 \\ -y - z - 5w &= -7 \\ 3z + 13w &= 13 \\ -13w &= -13 \end{aligned}$$

Por lo tanto la factorización LU está dada por

$$A = \begin{pmatrix} 1 & 1 & 0 & 3 \\ 2 & 1 & -1 & 1 \\ 3 & -1 & -1 & 2 \\ -1 & 2 & 3 & -1 \end{pmatrix} = \underbrace{\begin{pmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ 3 & 4 & 1 & 0 \\ -1 & -3 & 0 & 1 \end{pmatrix}}_{L} \underbrace{\begin{pmatrix} 1 & 1 & 0 & 3 \\ 0 & -1 & -1 & -5 \\ 0 & 0 & 3 & 13 \\ 0 & 0 & 0 & -13 \end{pmatrix}}_{U} = LU.$$

Esta factorización nos permite resolver fácilmente todo sistema que posee como matriz asociada la matriz A. Así, por ejemplo, para resolver el sistema

$$A\mathbf{x} = LU\mathbf{x} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ 3 & 4 & 1 & 0 \\ -1 & -3 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 1 & 0 & 3 \\ 0 & -1 & -1 & -5 \\ 0 & 0 & 3 & 13 \\ 0 & 0 & 0 & -13 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \\ w \end{pmatrix} = \begin{pmatrix} 8 \\ 7 \\ 14 \\ -7 \end{pmatrix}$$

primero realizaremos la sustitución $\mathbf{y} = U\mathbf{x}$. Luego resolvemos $L\mathbf{y} = \mathbf{b}$, es decir,

$$LU\mathbf{x} = L\mathbf{y} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ 3 & 4 & 1 & 0 \\ -1 & -3 & 0 & 1 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix} = \begin{pmatrix} 8 \\ 7 \\ 14 \\ -7 \end{pmatrix}.$$

Este sistema se resuelve para $\mathbf{y}$ mediante un simple proceso de susbtitución hacia adelante, con lo cual obtenemos

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix} = \begin{pmatrix} 8 \\ -9 \\ 26 \\ -26 \end{pmatrix}.$$

Por ultimo resolvemos $U\mathbf{x} = \mathbf{y}$, es decir, resolvemos el sistema por un proceso de susbtitución hacia atrás

$$U\mathbf{x} = \begin{pmatrix} 1 & 1 & 0 & 3 \\ 0 & -1 & -1 & -5 \\ 0 & 0 & 3 & 13 \\ 0 & 0 & 0 & -13 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \\ w \end{pmatrix} = \begin{pmatrix} 8 \\ -9 \\ 26 \\ -26 \end{pmatrix}$$

el cual tiene por solución

$$\mathbf{x} = \begin{pmatrix} x \\ y \\ z \\ w \end{pmatrix} = \begin{pmatrix} 3 \\ -1 \\ 0 \\ 2 \end{pmatrix}.$$

que es la solución buscada.

1.6 Eliminación gaussiana con pivoteo

Al resolver un sistema lineal $A\mathbf{x} = \mathbf{b}$, a veces es necesario intercambiar las ecuaciones, ya sea porque la elimninación gaussiana no puede efectuarse o porque las soluciones que se obtienen no corresponden a as soluciones exactas del sistema. Este proceso es llamado *pivoteo* y nos lleva a la descomposición $PA = LU$ del sistema. Así el sistema se transforma en $LU\mathbf{x} = P\mathbf{b}$ y resolvemos entonces los sistemas

$$\begin{cases} U\mathbf{x} &= \mathbf{z} \\ L\mathbf{z} &= P\mathbf{b} \end{cases}$$

donde P es una *matriz de permutación*, es decir, es una matriz obtenida a partir de la matriz identidad intercambiando algunas de sus filas. Más precisamente.

Definición 1.9 *Una matriz de permutación $P \in M(n \times n, \mathbb{R})$ es una matriz obtenida desde la matriz identidad por permutación de sus filas.*

Observación. Las matrices de permutación poseen dos propiedades de gran utilidad que se relacionan con la eliminación gaussiana. La primera de ellas es que al multiplicar por la izquierda una A por una matriz de permutación P, es decir, al realizar el producto PA se permutan las filas de A. La segunda propiedad establece que si P es una matriz de permutación, entonces $P^{-1} = P^T$.

Si $A \in M(n \times n, \mathbb{R})$ es una matriz invertible entonces podemos resolver el sistema lineal $A\mathbf{x} = \mathbf{b}$ vía eliminación gaussiana, sin excluir el intercambio de filas. Si conociéramos los intercambios que se requieren para resolver el sistema mediante eliminación gaussiana, podríamos arreglar las ecuaciones originales de manera que nos garantice que no se requieren intercambios de filas. Por lo tanto, existe un arreglo de ecuaciones en el sistema que nos permite resolver el sistema con eliminación gaussiana sin intercambio de filas. Con lo que podemos concluir que si $A \in M(n \times n, \mathbb{R})$ es una matriz invertible entonces existe una matriz de permutación P para la cual podemos resolver el sistema $PA\mathbf{x} = P\mathbf{b}$, sin hacer intercambios de filas. Pero podemos factorizar la matriz PA en $PA = LU$, donde L es una triangular inferior y U triangular superior, y P es una matriz de permutación. Dado que $P^{-1} = P^T$ obtenemos que $A = \left(P^T L\right) U$. Sin embargo, la matriz $P^T L$ no es triangular inferior, salvo que $P = I$.

Supongamos que tenemos el sistema

$$\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix} \tag{1.26}$$

Se define la *escala* de cada fila como

$$s_i = \max\{|a_{i1}|, |a_{i2}|, \ldots, |a_{in}|\}, \qquad 1 \leqslant i \leqslant n. \tag{1.27}$$

Elección de la fila pivote. Elegimos como *fila povote* la fila para la cual se tiene que

$$\frac{|a_{i_0 1}|}{s_{i_0}} \tag{1.28}$$

es el mayor.

Intercambiamos en A la fila 1 con la fila i_0, esto es, realizamos la operación elemental $F_1 \leftrightarrow F_{i_0}$. Esto nos produce la primera permutación. Procedemos ahora a producir ceros bajo la primera columna de la matriz permutada, tal como se hizo en la eliminación gaussiana. Denotamos el índice que hemos elegido por p_1, este se convierte en la primera componente de la permutación.

Más especificamente, comenzamos por fijar

$$p = (p_1\, p_2\, \ldots\, p_n) = (1\, 2\, \ldots\, n) \tag{1.29}$$

como la permutación antes de comenzar el pivoteo. Así la primera permutación que hemos hecho es

$$\begin{pmatrix} 1 & 2 & \ldots & i_0 & \ldots & n \\ i_0 & 2 & \ldots & 1 & \ldots & n \end{pmatrix}.$$

Para fijar las ideas, comenzamos con la permutación $(p_1\, p_2\, \ldots\, p_n) = (1\ 2\ \ldots\ n)$. Primero seleccionamos un índice j para el cual se tiene

$$\frac{|a_{p_j j}|}{s_{p_j}}, \quad 2 \leqslant j \leqslant n \tag{1.30}$$

es el mayor, y se intercambia p_1 con p_j en la permutación p. Procedemos a la elmincación gaussiana, restando

$$\frac{a_{p_i 1}}{a_{p_1 1}} \times F_{p_1}$$

a las filas p_i, $2 \leqslant i \leqslant n$.

En general, supongamos que tenemos que producir ceros en la columna k. Impeccionamos los números

$$\frac{|a_{p_i k}|}{s_{p_i}}, \quad k \leqslant i \leqslant n \tag{1.31}$$

y buscamos el mayor de ellos. Si j es el índice del primer cuociente más grande, intercambiamos p_k con p_j en la permutación p y las filas F_{p_k} con la fila F_{p_j} en la matriz que tenemos en esta etapa de la eliminación gaussiana–pivoteo. Enseguida producimos ceros en la columna p_k bajo el elemento $a_{p_k k}$, para ellos restamos

$$\frac{a_{p_i k}}{a_{p_k k}} \times F_{p_k}$$

de la fila F_{p_i}, $k+1 \leqslant i \leqslant n$, y continuamos el proceso hasta obtener una matriz triangular superior.

Veamos con un ejemplo específico como podemos ir guardando toda la información del proceso que vamos realizando en cada etapa del proceso de eliminación gaussiana–pivoteo

Ejemplo 16 Resolver el sistema de ecuaciones lineales

$$\begin{pmatrix} 2 & 3 & -6 \\ 1 & -6 & 8 \\ 3 & -2 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

Solución. Primero calculemos la escala de las filas. Tenemos

$$\begin{aligned} s_1 &= \max\{|2|, |3|, |-6|\} = 6 \\ s_2 &= \max\{|1|, |-6|, |8|\} = 8 \\ s_3 &= \max\{|3|, |-2|, |1|\} = 3 \end{aligned}$$

permutación inicial $p = (1\ 2\ 3)$. Denotemos por s el vector de la escala de las filas, es decir, $s = (6\ 8\ 3)$.

Elección de la primera fila pivote:

$$\begin{aligned}\frac{|a_{11}|}{s_1} &= \frac{2}{6} = \frac{1}{3}\\ \frac{|a_{21}|}{s_2} &= \frac{1}{8}\\ \frac{|a_{31}|}{s_3} &= 1\,,\end{aligned}$$

luego $j = 3$. Antes de proceder a intercambiar la fila F_1 con la fila F_3 guardamos la información en la forma siguiente

$$\left(\begin{array}{ccc|c||c|c} 2 & 3 & -6 & 1 & 6 & 1\\ 1 & -6 & 8 & 1 & 8 & 2\\ 3 & -2 & 1 & 1 & 3 & 3 \end{array}\right)$$

donde la primera parte de esta matriz representa a la matriz A, la segunda al vector b, la tercera al vector s y la última a la permutación inicial $p = (1\,2\,3)$. Como $j = 3$ intercambianos las filas $F_1 \leftrightarrow F_3$ y obtenemos

$$\left(\begin{array}{ccc|c||c|c} 3 & -2 & 1 & 1 & 3 & 3\\ 1 & -6 & 8 & 1 & 8 & 2\\ 2 & 3 & -6 & 1 & 6 & 1 \end{array}\right)$$

en la parte $(A|\mathbf{b})$ de la matriz arriba procedemos a hacer eliminación gaussiana. Los multiplicadores son $m_{21} = \frac{1}{3}$ y $m_{31} = \frac{2}{3}$, obtenemos así

$$\left(\begin{array}{ccc|c||c|c} 3 & -2 & 1 & 1 & 3 & 3\\ 0 & -\frac{16}{3} & \frac{23}{3} & \frac{2}{3} & 8 & 2\\ 0 & \frac{13}{3} & -\frac{20}{3} & \frac{1}{3} & 6 & 1 \end{array}\right)$$

Agregamos la información de los multiplicadores a esta estructura como sigue

$$\left(\begin{array}{ccc|c||c|c|c} 3 & -2 & 1 & 1 & 3 & 3 & \\ 0 & -\frac{16}{3} & \frac{23}{3} & \frac{2}{3} & 8 & 2 & \frac{1}{3}\\ 0 & \frac{13}{3} & -\frac{20}{3} & \frac{1}{3} & 6 & 1 & \frac{2}{3} \end{array}\right)$$

Ahora determinamos la nueva fila pivote. Recuerde que la primera fila de esta nueva matriz permanece inalterada, y sólo debemos trabajar con las filas restantes. Para determinar la nueva fila pivote, calculamos

$$\begin{aligned}\frac{|a_{p_2 2}|}{s_{p_2}} &= \frac{\frac{16}{3}}{8} = \frac{16}{24} = \frac{2}{3}\\ \frac{|a_{p_3 2}|}{s_{p_3}} &= \frac{\frac{13}{3}}{6} = \frac{13}{18}\end{aligned}$$

como $\frac{13}{18} > \frac{2}{3}$ la nueva fila pivote es la fila 3. Intercambiando $F_2 \leftrightarrow F_3$ nos queda

$$\left(\begin{array}{ccc|c||ccc|c} 3 & -2 & 1 & 1 & 3 & 3 & & \\ 0 & \frac{13}{3} & -\frac{20}{3} & \frac{1}{3} & 6 & 1 & \frac{2}{3} & \\ 0 & -\frac{16}{3} & \frac{23}{3} & \frac{2}{3} & 8 & 2 & \frac{1}{3} & \end{array}\right)$$

El multiplicador es $m_{32} = \frac{-\frac{16}{3}}{\frac{13}{3}} = -\frac{16}{13}$. Agregando la información del nuevo multiplicador y aplicando eliminación gaussiana, nos queda

$$\left(\begin{array}{ccc|c||ccc|c} 3 & -2 & 1 & 1 & 3 & 3 & & \\ 0 & \frac{13}{3} & -\frac{20}{3} & \frac{1}{3} & 6 & 1 & \frac{2}{3} & \\ 0 & 0 & -\frac{7}{13} & \frac{94}{117} & 8 & 2 & \frac{1}{3} & -\frac{16}{13} \end{array}\right)$$

De esto, el sistema a resolver es

$$\underbrace{\begin{pmatrix} 3 & -2 & 1 \\ 0 & \frac{13}{3} & -\frac{20}{3} \\ 0 & 0 & -\frac{7}{13} \end{pmatrix}}_{U} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1 \\ \frac{1}{3} \\ \frac{94}{117} \end{pmatrix}$$

La matriz L es

$$L = \begin{pmatrix} 1 & 0 & 0 \\ \frac{2}{3} & 1 & 0 \\ \frac{1}{3} & -\frac{16}{13} & 1 \end{pmatrix}$$

y la matriz de permutación P es

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

Ahora es fácil verificar que $PA = LU$.

Ejemplo 17 Consideremos la matriz

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

puesto que $a_{11} = 0$ la matriz no posee una factorización LU. Pero si realizamos el intercambio de filas $F_1 \leftrightarrow F_2$, seguido de $F_3 \mapsto (F_3 - F_1)$ y de $F_4 \mapsto (F_4 - F_1)$ obtenemos

$$\begin{pmatrix} 1 & 1 & -1 & 2 \\ 0 & 0 & -1 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \end{pmatrix}.$$

Realizando las operaciones elementales $F_2 \leftrightarrow F_4$ y $F_4 \mapsto (F_4 + F_3)$, obtenemos

$$U = \begin{pmatrix} 1 & 1 & -1 & 2 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 2 \end{pmatrix}.$$

La matriz de permutación asociada al cambio de filas $F_1 \leftrightarrow F_2$ seguida del intercambio de filas $F_2 \leftrightarrow F_4$ es

$$P = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}.$$

La eliminación gaussiana en PA se puede realizar sin intercambio de filas usando las operaciones $F_2 \mapsto (F_2 - F_1)$, $F_3 \mapsto (F_3 - F_1)$ y $(F_4 + F_3) \mapsto F_4$, esto produce la factorización LU de PA

$$PA = \begin{pmatrix} 1 & 1 & -1 & 2 \\ 1 & 2 & -1 & 3 \\ 1 & 1 & 0 & 3 \\ 0 & 0 & -1 & 2 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 0 & -1 & 1 \end{pmatrix} \begin{pmatrix} 1 & 1 & -1 & 2 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 2 \end{pmatrix} = LU.$$

Al multiplicar por $P^{-1} = P^T$ obtenemos la factorización

$$A = \left(P^T L\right) U = \begin{pmatrix} 0 & 0 & -1 & 1 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} 1 & 1 & -1 & 2 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 2 \end{pmatrix}.$$

1.7 Matrices especiales

En esta sección nos ocuparemos de clases de matrices en las cuales podemos realizar la eliminación gaussiana sin intercambio de filas.

Decimos que una matriz $A \in M(n \times n, \mathbb{R})$ es *diagonal dominante* si

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

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

Ejemplo 18 Consideremos las matrices

$$A = \begin{pmatrix} 7 & 2 & 0 \\ 3 & 5 & -1 \\ 0 & 5 & -6 \end{pmatrix} \quad \text{y} \quad B = \begin{pmatrix} 6 & 4 & -3 \\ 4 & -2 & 0 \\ -3 & 0 & 1 \end{pmatrix}.$$

La matriz A es diagonal dominante y la matriz B no es diagonal dominante, pues por ejemplo $|6| < |4| + |-3| = 7$.

Teorema 1.15 *Toda matriz $A \in M(n \times n, \mathbb{R})$ diagonal dominante tiene inversa, más aun, en este caso podemos realizar la eliminación gaussiana de cualquier sistema lineal de la forma $A\mathbf{x} = \mathbf{b}$ para obtener su solución única sin intercambio de filas, y los cálculos son estables respecto al crecimiento de los errores de redondeo.*

Teorema 1.16 *Si $A \in M(n \times n, \mathbb{R})$ es una matriz definida positiva entonces*

1. *A tiene inversa.*
2. *$a_{ii} > 0$, para todo $i = 1, 2, \ldots, n$.*
3. *$\max_{1 \leqslant k,\, j \leqslant n} |a_{kj}| \leqslant \max_{1 \leqslant i \leqslant n} |a_{ii}|$.*
4. *$(a_{ij})^2 < a_{ii}\, a_{jj}$, para todo $i \neq j$*

Teorema 1.17 *Una matriz simétrica A es definida positiva si y sólo si la eliminación gaussiana sin intercambio de filas puede efectuarse en el sistema $A\mathbf{x} = \mathbf{b}$ con todos los elementos pivotes positivos. Además, en este caso los cálculos son estables respecto al crecimiento de los errores de redondeo.*

Corolario 1.2 *Una matriz simétrica A es definida positiva si y sólo si A puede factorizarse en la forma LDL^T, donde L es una matriz triangular inferior con unos en su diagonal y D es una matriz diagonal con elementos positivos a lo largo de la diagonal.*

Corolario 1.3 *Una matriz simétrica A es definida positiva si y sólo si A puede factorizarse de la forma LL^T, donde L es una matriz triangular inferior con elementos distintos de cero en su diagonal.*

Corolario 1.4 *Sea $A \in M(n \times n, \mathbb{R})$ una matriz simétrica a la cual puede aplicarse la eliminación gaussiana sin intercambio de filas. Entonces, A puede factorizarse en LDL^T, donde L es una matriz triangular inferior con unos en su diagonal y donde D es una matriz diagonal con $a_{11}^{(1)}, \ldots, a_{nn}^{(n)}$ en su diagonal.*

Ejemplo 19 La matriz

$$A = \begin{pmatrix} 4 & -1 & 1 \\ -1 & 4.25 & 2.75 \\ 1 & 2.75 & 3.5 \end{pmatrix}$$

es definida positiva. La factorización LDL^T de la matriz A es

$$A = \begin{pmatrix} 1 & 0 & 0 \\ -0.25 & 1 & 0 \\ 0.25 & 0.75 & 1 \end{pmatrix} \begin{pmatrix} 4 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & -0.25 & 0.25 \\ 0 & 1 & 0.75 \\ 0 & 0 & 1 \end{pmatrix}$$

mientras que su descomposición de Cholesky es

$$A = \begin{pmatrix} 2 & 0 & 0 \\ -0.5 & 2 & 0 \\ 0.5 & 1.5 & 1 \end{pmatrix} \begin{pmatrix} 2 & -0.5 & 0.5 \\ 0 & 2 & 1.5 \\ 0 & 0 & 1 \end{pmatrix}.$$

Definición 1.10 *Una matriz $A \in M(n \times n, \mathbb{R})$ es llamada matriz banda si existen enteros p y q con $1 < p, q < n$, tales que $a_{ij} = 0$ siempre que $i + p \leqslant j$ o $j + q \leqslant i$. El ancho de la banda de este tipo de matrices esta dado por $w = p + q - 1$.*

Ejemplo 20 La matriz $A = \begin{pmatrix} 1 & 2 & 0 \\ 2 & 1 & 2 \\ 0 & 1 & 1 \end{pmatrix}$ es una matriz de banda con $p = q = 2$ y con ancho de banda 3.

Definición 1.11 *Una matriz $A \in M(n \times n, \mathbb{R})$ se denomina tridiagonal si es una matriz de banda con $p = q = 2$.*

Observación. Los algoritmos de factorización pueden simplificarse considerablemente en el caso de las matrices de banda.

Para ilustrar lo anterior, supongamos que podemos factorizar una matriz tridiagonal A en las matrices triangulares L y U. Ya que A tiene sólo $3n-2$ elementos distintos de cero, habrá apenas $3n-2$ condiciones aplicables para determinar los elementos de L y U, naturalmente a condición de que también se obtengan los elementos cero de A. Supongamos que podemos encontrar las matrices de la forma

$$L = \begin{pmatrix} l_{11} & 0 & \cdots & \cdots & 0 \\ l_{21} & l_{22} & \ddots & \vdots & \vdots \\ 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & l_{n\,n-1} & l_{nn} \end{pmatrix} \text{ y } U = \begin{pmatrix} 1 & u_{12} & 0 & \cdots & 0 \\ 0 & 1 & \ddots & \vdots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ \vdots & \vdots & \ddots & \ddots & u_{n-1\,n} \\ 0 & \cdots & \cdots & 0 & 1 \end{pmatrix}$$

De la multiplicación que incluye $A = LU$, obtenemos

$$\begin{aligned}
a_{11} &= l_{11};\\
a_{i\,i-1} &= l_{i\,i-1} \text{ para cada } i = 2,3,\ldots,n;\\
a_{ii} &= l_{i\,i-1}u_{i-1\,i} + l_{ii} \text{ para cada } i = 2,3,\ldots,n;\\
a_{i\,i+1} &= l_{ii}u_{i\,i+1} \text{ para cada } i = 1,2,3,\ldots,n-1;
\end{aligned}$$

1.8 Solución de sistemas de ecuaciones lineales: métodos iterativos

En general, puede resultar muy engorroso encontrar la solución exacta a un sistema de ecuaciones lineales, y por otra parte, a veces nos basta con tener buenas aproximaciones a dicha solución. Para obtener estas últimas usamos métodos iterativos de punto fijo, que en este caso particular resultan ser más analizar su convergencia.

Teorema 1.18 *Consideremos un método iterativo*

$$\boxed{\mathbf{x}^{(k+1)} = G\mathbf{x}^{(k)} + \mathbf{c}} \tag{1.32}$$

donde $G \in M(n\times n,\mathbb{R})$, $\mathbf{x},\mathbf{c} \in \mathbb{R}^n$. *Entonces el método iterativo es convergente, para cualquier condición inicial* $\mathbf{x}^{(0)}$ *elegida arbitrariamente si y sólo si* $\rho(G) < 1$. *Además, si existe una norma* $|||\ ||$ *en* $M(n\times n,\mathbb{R})$, *en la cual se tiene* $||G|| < 1$, *entonces el método iterativo converge cualesquiera que sea la condición inicial* $\mathbf{x}^{(0)} \in \mathbb{R}^n$ *dada. Si tomamos* $\mathbf{x}^{(k)}$ *como una aproximación al punto fijo* $\mathbf{x}_T$ *del método iterativo (1.32), entonces valen las desigualdades siguientes,*

$$\boxed{\left\|\mathbf{x}^{(k)} - \mathbf{x}\right\| \leqslant \frac{\rho(G)^k}{1-\rho(G)}\left\|\mathbf{x}^{(1)} - \mathbf{x}^{(0)}\right\|} \tag{1.33}$$

y

$$\boxed{\left\|\mathbf{x}^{(k)} - \mathbf{x}\right\| \leqslant \frac{\lambda^k}{1-\lambda}\left\|\mathbf{x}^{(1)} - \mathbf{x}^{(0)}\right\|} \tag{1.34}$$

donde $\lambda = ||G||$.

Observación. Si el método iterativo $\mathbf{x}^{(k+1)} = G\mathbf{x}^{(k)} + \mathbf{c}$ es convergente y tiene por límite a $\mathbf{x} \in \mathbb{R}^n$, entonces

$$\mathbf{x} = \lim_{k\to\infty}\mathbf{x}^{(k+1)} = G\left(\lim_{k\to\infty}\mathbf{x}^{(k)}\right) + \mathbf{c} = G\mathbf{x} + \mathbf{c},$$

de donde $\mathbf{x} = (I-G)^{-1}\mathbf{c}$.

Regresemos al problema inicial de resolver el sistema de ecuaciones lineales

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

donde $A = (a_{ij}) \in M(n\times n,\mathbb{R})$, $\mathbf{b} = (b_1, b_2, \ldots, b_n)^T \in \mathbb{R}^n$ y $\mathbf{x} = (x_1, x_2, \ldots, x_n)^T \in \mathbb{R}^n$ es la incógnita.

Consideremos una matriz $Q \in M(n \times n, \mathbb{R})$, la cual suponemos tiene inversa. Tenemos que $A\mathbf{x} = \mathbf{b}$ es equivalente a escribir $0 = -A\mathbf{x} + \mathbf{b}$, sumando a ambos miembros de esta última igualdad el término $Q\mathbf{x}$, no queda la ecuación $Q\mathbf{x} = Q\mathbf{x} - A\mathbf{x} + \mathbf{b}$, multiplicando esta ecuación por la izquierda por Q^{-1} obtenemos $\mathbf{x} = (I - Q^{-1}A)\mathbf{x} + Q^{-1}\mathbf{b}$, lo que nos sugiere proponer el siguiente método iterativo para resolver la ecuación original $A\mathbf{x} = \mathbf{b}$,

$$\boxed{\mathbf{x}^{(k+1)} = (I - Q^{-1}A)\mathbf{x}^{(k)} + Q^{-1}\mathbf{b}} \tag{1.35}$$

donde $\mathbf{x}^{(0)} \in \mathbb{R}^n$ es arbitrario. Sobre la convergencia del método propuesto, tenemos el siguiente corolario.

Corolario 1.5 *La fórmula de iteración* $\mathbf{x}^{(k+1)} = \left(I - Q^{-1}A\right)\mathbf{x}^{(k)} + Q^{-1}\mathbf{b}$ *define una sucesión que converge a la solución de* $A\mathbf{x} = \mathbf{b}$, *para cualquier condición inicial* $\mathbf{x}^{(0)} \in \mathbb{R}^n$ *si y sólo si* $\rho\left(I - Q^{-1}A\right) < 1$. *Además, si existe una norma* $|||\ ||$ *en* $M(n \times n, \mathbb{R})$, *en la cual se tiene* $||I - Q^{-1}A|| < 1$, *entonces el método iterativo converge cualesquiera que sea la condición inicial* $\mathbf{x}^{(0)} \in \mathbb{R}^n$ *dada.*

Observación La fórmula de iteración

$$\mathbf{x}^{(k+1)} = \left(I - Q^{-1}A\right)\mathbf{x}^{(k)} + Q^{-1}\mathbf{b}$$

define un método iterativo de punto fijo. En efecto, consideremos la función $F : \mathbb{R}^n \longrightarrow \mathbb{R}^n$ definida por $F(\mathbf{x}) = \left(I - Q^{-1}A\right)\mathbf{x} + Q^{-1}\mathbf{b}$. Observemos que si $\mathbf{x}$ es un punto fijo de F entonces $\mathbf{x}$ es una solución de la ecuación $A\mathbf{x} = \mathbf{b}$. Además, si $\mathbf{x} = \left(I - Q^{-1}A\right)\mathbf{x} + Q^{-1}\mathbf{b}$ entonces se tiene que

$$\mathbf{x}^{(k)} - \mathbf{x} = \left(I - Q^{-1}A\right)\left(\mathbf{x}^{(k-1)} - \mathbf{x}\right),$$

por lo tanto

$$\left\|\mathbf{x}^{(k)} - \mathbf{x}\right\| \leqslant \left\|I - Q^{-1}A\right\| \left\|\mathbf{x}^{(k-1)} - \mathbf{x}\right\|,$$

de donde, iterando esta desigualdad obtenemos

$$\left\|\mathbf{x}^{(k)} - \mathbf{x}\right\| \leqslant \left\|I - Q^{-1}A\right\|^k \left\|\mathbf{x}^{(0)} - \mathbf{x}\right\|,$$

luego si $\left\|I - Q^{-1}A\right\| < 1$, se tiene de inmediato que $\lim_{k\to\infty} \left\|\mathbf{x}^{(k)} - \mathbf{x}\right\| = 0$ para cualquier $\mathbf{x}^{(0)} \in \mathbb{R}^n$ dado.

Observación. La condición $\left\|I - Q^{-1}A\right\| < 1$ implica que las matrices $Q^{-1}A$ y A tienen inversas.

Teorema 1.19 *Si* $\left\|I - Q^{-1}A\right\| < 1$ *para alguna norma matricial subordinada en* $M(n \times n, \mathbb{R})$, *entonces el método iterativo* $\mathbf{x}^{(k+1)} = \left(I - Q^{-1}A\right)\mathbf{x}^{(k)} + Q^{-1}\mathbf{b}$ *es convergente a una solución de* $A\mathbf{x} = \mathbf{b}$, *para cualquier vector inicial* $\mathbf{x}^{(0)} \in \mathbb{R}^n$. *Además, si* $\lambda = \left\|I - Q^{-1}A\right\| < 1$, *podemos usar el criterio de parada* $\left\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\right\| < \varepsilon$ *y se tiene que*

$$\left\|\mathbf{x}^{(k)} - \mathbf{x}\right\| \leqslant \frac{\lambda}{1-\lambda}\left\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\right\| \leqslant \cdots \leqslant \frac{\lambda^k}{1-\lambda}\left\|\mathbf{x}^{(1)} - \mathbf{x}^{(0)}\right\|,$$

es decir,

$$\boxed{\left\|\mathbf{x}^{(k)} - \mathbf{x}\right\| \leqslant \frac{\lambda^k}{1-\lambda}\left\|\mathbf{x}^{(1)} - \mathbf{x}^{(0)}\right\|} \tag{1.36}$$

También vale la desigualdad siguiente

$$\boxed{\left\|\mathbf{x}^{(k)}-\mathbf{x}\right\| \leqslant \frac{\rho(I-Q^{-1}A)^k}{1-\rho(I-Q^{-1}A)}\left\|\mathbf{x}^{(1)}-\mathbf{x}^{(0)}\right\|} \tag{1.37}$$

Observación. Desde la definición de radio espectral de una matriz, se tiene que $\rho(A)<1$ implica que existe una norma matricial $||\ ||$ en $M(n\times n,\mathbb{R})$, tal que $||A||<1$. Por otra parte, si $||A||<1$ para alguna norma matricial en $M(n\times n,\mathbb{R})$ entonces se tiene que $\rho(A)<1$. Notemos que si existe una norma matricial $||\ ||$ en $M(n\times n,\mathbb{R})$, tal que $||A||\geqslant 1$ no necesariamente se tiene que $\rho(A)\geqslant 1$.

Observación. Si M_1 y M_2 son dos métodos iterativos convergentes para resolver un sistema de ecuaciones lineales $A\mathbf{x}=\mathbf{b}$, digamos $M_1:\ \mathbf{x}^{(k+1)}=G_1\mathbf{x}^{(k)}+\mathbf{c}_1$ y $M_2:\ \mathbf{x}^{(k+1)}=G_2\mathbf{x}^{(k)}+\mathbf{c}_2$. Si $\rho(G_1))\leqslant\rho(G_2)$, entonces M_1 converge más rápido que M_2 a la solución de la ecuación. Además, si existe una norma $||\cdot||$ en $M(n\times n,\mathbb{R})$ tal que $||G_1||\leqslant||G_2||$, entonces M_1 converge más rápido que M_2 a la solución de la ecuación.

1.9 Método de Richardson

Para este método tomamos $Q_R=I$, luego el método iterativo $\mathbf{x}^{(k+1)}=\left(I-Q^{-1}A\right)\mathbf{x}^{(k)}+Q^{-1}\mathbf{b}$ se transforma en

$$\boxed{\mathbf{x}^{(k+1)}=\underbrace{(I-A)}_{T_R}\mathbf{x}^{(k)}+\mathbf{b}}, \tag{1.38}$$

el cual converge si y sólo si $\rho(T_R)=\rho(I-A)<1$.

Si existe una norma matricial $\|\cdot\|$ para la cual se tiene que $\|T_R\|=||I-A||<1$, entonces el método iterativo de Richardson es convergente.

Ejemplo 21 Resolver el sistema de ecuaciones lineales usando el método de Richardson.

$$\begin{pmatrix} 1 & \frac{1}{2} & \frac{1}{3} \\ \frac{1}{3} & 1 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{3} & 1 \end{pmatrix}\begin{pmatrix} x \\ y \\ z \end{pmatrix}=\begin{pmatrix} \frac{11}{18} \\ \frac{11}{18} \\ \frac{11}{18} \end{pmatrix}$$

Solución. Tenemos

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

Llamando $\mathbf{x}^{(j)} = (x_j \ y_j \ z_j)^T$, nos queda

$$\mathbf{x}^{(k+1)} = T_R\mathbf{x}^{(k)} + \mathbf{b} = \begin{pmatrix} 0 & -\frac{1}{2} & -\frac{1}{3} \\ -\frac{1}{3} & 0 & -\frac{1}{2} \\ -\frac{1}{2} & -\frac{1}{3} & 0 \end{pmatrix} \begin{pmatrix} x_k \\ y_k \\ z_k \end{pmatrix} + \begin{pmatrix} \frac{11}{18} \\ \frac{11}{18} \\ \frac{11}{18} \end{pmatrix}$$

se tiene $||T_R||_\infty = \frac{5}{6} < 1$, luego el método iterativo de Richardson para este sistema converge.

1.10 Método de Jacobi

En este caso tomamos

$$\boxed{Q_J = \operatorname{diag}(A) = \begin{pmatrix} a_{11} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & a_{nn} \end{pmatrix}} \tag{1.39}$$

donde los elementos fuera de la diagonal en Q_J son todos iguales a 0. Notemos que $\det(Q_J) = a_{11}a_{22}\cdots a_{nn}$, luego $\det(Q_J) \neq 0$ si y sólo si $a_{ii} \neq 0$ para todo $i = 1, \ldots, n$, es decir, Q_J tiene inversa si y sólo si $a_{ii} \neq 0$ para todo $i = 1, \ldots, n$.

Como antes, colocando esta matriz $Q_J = \operatorname{diag}(A)$ en la fórmula iterativa definida por

$$\boxed{\mathbf{x}^{(k+1)} = \underbrace{\left(I - \operatorname{diag}(A)^{-1}A\right)}_{T_J}\mathbf{x}^{(k)} + \operatorname{diag}(A)^{-1}\mathbf{b}} \tag{1.40}$$

obtenemos un método iterativo convergente si y sólo si $\rho(T_J) = \rho(I - \operatorname{diag}(A)^{-1}A) < 1$. Si existe una norma matricial $\|\cdot\|$ para la cual se tiene que $\|T_J\| = ||I - \operatorname{diag}(A)^{-1}A|| < 1$, entonces el método iterativo de Jacobi es convergente.

Examinemos un poco más la fórmula iterativa del método de Jacobi. Tenemos, en este caso, que un elemento genérico de $Q^{-1}A$ es de la forma $\frac{a_{ij}}{a_{ii}}$ y todos los elementos de la diagonal de $Q^{-1}A$ son iguales a 1. Luego, tomando la norma $\|\ \|_\infty$ en $M(n \times n, \mathbb{R})$ tenemos que

$$\left\|I - \operatorname{diag}(A)^{-1}A\right\|_\infty = \max_{1 \leqslant i \leqslant n} \left\{ \sum_{j=1\,,j\neq i}^{n} \frac{|a_{ij}|}{|a_{ii}|} \right\}$$

Recordemos que una matriz A es *diagonal dominante* si

$$|a_{ii}| > \sum_{j=1\,,\,j\neq i}^{n} |a_{ij}|\,, \quad 1 \leqslant i \leqslant n,$$

es decir, $\sum\limits_{j=1\,,\,j\neq i}^{n} \left|\frac{a_{ij}}{a_{ii}}\right| < 1$, por lo tanto tenemos el siguiente teorema.

Teorema 1.20 *Si A es una matriz diagonal dominante, entonces el método iterativo de Jacobi es convergente para cualquiera que sea la condición inicial $\mathbf{x}^{(0)}$ elegida.*

1.11 Método de Gauss–Seidel

En esta caso, consideremos la matriz Q como la matriz triangular obtenida considerando la parte triangular inferior de la matriz A incluyendo su diagonal, es decir,

$$\boxed{Q_{\text{G-S}} = \begin{pmatrix} a_{11} & 0 & \cdots & 0 \\ a_{21} & a_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix}} \tag{1.41}$$

Se tiene que $\det(Q_{\text{G-S}}) = a_{11}a_{22}\cdots a_{nn}$. Luego, $\det(Q_{\text{G-S}}) \neq 0$ si y sólo si $a_{ii} \neq 0$ $(i = 1, \ldots, n)$. Usando la matriz $Q_{\text{G-S}}$, obtenemos el método iterativo

$$\boxed{\mathbf{x}^{(k+1)} = \underbrace{\left(I - Q_{\text{G-S}}^{-1}A\right)}_{T_{\text{G-S}}}\mathbf{x}^{(k)} + Q_{\text{G-S}}^{-1}\mathbf{b}} \tag{1.42}$$

Teorema 1.21 *Sea $A \in M(n \times n, \mathbb{R})$. Entonces el método iterativo de Gauss–Seidel es convergente si y sólo si $\rho(T_{G\text{-}S}) = \rho(I - Q_{G\text{-}S}^{-1}A) < 1$, donde $Q_{G\text{-}S}$ es la matriz triangular inferior definida arriba a partir de A.*

Como en el caso del método iterativo de Jacobi, tenemos el siguiente teorema.

Teorema 1.22 *Si $A \in M(n \times n, \mathbb{R})$ es una matriz es diagonal dominante entonces el método de Gauss–Seidel converge para cualquier elección inicial $\mathbf{x}^{(0)}$.*

Ejemplo 22 Considere el sistema de ecuaciones lineales

$$\begin{pmatrix} 4 & 2 & 1 \\ 2 & 5 & 2 \\ 1 & 2 & 6 \end{pmatrix}\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 5 \\ 4 \\ 7 \end{pmatrix}$$

Explicitaremos los métodos iterativos de Jacobi y Gauss–Seidel para este sistema. Estudiaremos la convergencia de ellos sin iterar. Finalmente, sabiendo que la solución exacta del sistema es $(x_T, y_T, z_T) = (1, 0, 1)$ y usando el punto de partida $(1, 1, 1)$ nos podemos preguntar ¿Cuál de ellos converge más rápido?, para las comparaciones usaremos la norma $\| \quad \|_\infty$ en $M(n \times n, \mathbb{R})$.

Primero que nada tenemos que los métodos iterativos de Jacobi y Gauss–Seidel convergen, pues la matriz asociada al sistema es diagonal dominante.

Explicitaremos el método de Jacobi. En este caso tenemos

$$Q_J = \begin{pmatrix} 4 & 0 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & 6 \end{pmatrix}$$

por lo tanto

$$Q_J^{-1} = \begin{pmatrix} \frac{1}{4} & 0 & 0 \\ 0 & \frac{1}{5} & 0 \\ 0 & 0 & \frac{1}{6} \end{pmatrix}$$

luego

$$Q_J^{-1}A = \begin{pmatrix} \frac{1}{4} & 0 & 0 \\ 0 & \frac{1}{5} & 0 \\ 0 & 0 & \frac{1}{6} \end{pmatrix} \begin{pmatrix} 4 & 2 & 1 \\ 2 & 5 & 2 \\ 1 & 2 & 6 \end{pmatrix} = \begin{pmatrix} 1 & \frac{1}{2} & \frac{1}{4} \\ \frac{2}{5} & 1 & \frac{2}{5} \\ \frac{1}{6} & \frac{1}{3} & 1 \end{pmatrix}$$

de donde

$$T_J I - Q_J^{-1}A = \begin{pmatrix} 0 & -\frac{1}{2} & -\frac{1}{4} \\ -\frac{2}{5} & 0 & -\frac{2}{5} \\ -\frac{1}{6} & -\frac{1}{3} & 0 \end{pmatrix} \text{ y } Q_J^{-1}\mathbf{b} = \begin{pmatrix} \frac{1}{4} & 0 & 0 \\ 0 & \frac{1}{5} & 0 \\ 0 & 0 & \frac{1}{6} \end{pmatrix} \begin{pmatrix} 5 \\ 4 \\ 7 \end{pmatrix} = \begin{pmatrix} \frac{5}{4} \\ \frac{4}{5} \\ \frac{7}{6} \end{pmatrix},$$

por lo tanto

$$\begin{cases} x_{k+1} = \dfrac{-2y_k - z_k + 5}{4} \\ y_{k+1} = \dfrac{-2x_k - 2z_k + 4}{5} \\ z_{k+1} = \dfrac{-x_k - 2y_k + 7}{6} \end{cases}$$

Como $||T_J||_\infty = \max\{|\frac{1}{2}|+|\frac{-1}{4}|, |\frac{-2}{5}|+|\frac{-2}{5}|, |\frac{-1}{6}|+|\frac{-1}{3}|\} = \max\{\frac{3}{4}, \frac{4}{5}, \frac{1}{2}\} = \{0.75, 0.8, 0.5\} = 0.8 < 1$, el método de Jacobi converge para cualquier condición inicial $\mathbf{x}^{(0)} \in \mathbb{R}^n$ dada. Ahora explicitemos el método de Gauss–Seidel. En este caso,

$$Q_{\text{G-S}} = \begin{pmatrix} 4 & 0 & 0 \\ 2 & 5 & 0 \\ 1 & 2 & 6 \end{pmatrix} \qquad \text{por lo tanto} \qquad Q_{\text{G-S}}^{-1} = \begin{pmatrix} \frac{1}{4} & 0 & 0 \\ -\frac{1}{10} & \frac{2}{10} & 0 \\ -\frac{1}{120} & -\frac{8}{120} & \frac{20}{120} \end{pmatrix}$$

así obtenemos que

$$Q_{\text{G-S}}^{-1}A = \begin{pmatrix} \frac{1}{4} & 0 & 0 \\ -\frac{1}{10} & \frac{2}{10} & 0 \\ -\frac{1}{120} & -\frac{8}{120} & \frac{20}{120} \end{pmatrix} \begin{pmatrix} 4 & 2 & 1 \\ 2 & 5 & 2 \\ 1 & 2 & 6 \end{pmatrix} = \begin{pmatrix} 1 & \frac{1}{2} & \frac{1}{4} \\ 0 & \frac{8}{10} & \frac{3}{10} \\ 0 & -\frac{2}{120} & \frac{103}{120} \end{pmatrix}$$

luego $I - Q_{\text{G-S}}^{-1}A = \begin{pmatrix} 0 & -\frac{1}{2} & -\frac{1}{4} \\ 0 & \frac{2}{10} & -\frac{3}{10} \\ 0 & \frac{2}{120} & \frac{17}{120} \end{pmatrix}$ y $Q_{\text{G-S}}^{-1}\mathbf{b} = \begin{pmatrix} \frac{5}{4} \\ \frac{3}{10} \\ \frac{103}{120} \end{pmatrix}$, por lo tanto

$$\begin{cases} x_{k+1} = \dfrac{-2y_k - z_k + 5}{4} \\ y_{k+1} = \dfrac{2y_k - 3z_k + 3}{10} \\ z_{k+1} = \dfrac{2y_k - 17z_k + 103}{120} \end{cases}$$

Como $||T_{\text{G-S}}||_\infty = \max\{|\frac{-1}{4}| + |\frac{-1}{4}|, |\frac{2}{10}| + |\frac{-3}{10}|, |\frac{2}{120}| + |\frac{17}{120}\} = \max\{\frac{3}{4}, \frac{5}{10}, \frac{19}{120}\} = 0.75 < 1$. Luego, el método de Gauss–Seidel converge para cualquier condición inicial $\mathbf{x}^{(0)} \in \mathbb{R}^n$ dada.

Notemos que cómo $||T_{\text{G-S}}||_\infty < ||T_J||_\infty$, vemos que el método de Gauss–Seidel converge más rápido que el método de Jacobi.

Ejercicio. Realizar las iteraciones en ambos caso, Jacobi y Gauss-Seidel para este ejemplo.

1.12 Método SOR (successive overrelaxation)

Este método, también llamado sobre-relajación sucesiva.

Supongamos que escogemos la matriz Q como $Q_{\text{SOR}} = \frac{1}{\omega} D - C$, donde $\omega \in \mathbb{R}$, $\omega \neq 0$, es un parámetro, D es una matriz simétrica, positiva definida y C satisface $C + C^T = D - A$, tenemos entonces el método iterativo

$$\boxed{\mathbf{x}^{(k+1)} = \underbrace{\left(I - Q_{\text{SOR}}^{-1} A\right)}_{T_{\text{SOR}}} \mathbf{x}^{(k)} + Q_{\text{SOR}}^{-1} \mathbf{b}} \tag{1.43}$$

Teorema 1.23 *Si A es simétrica y positiva definida, Q_{SOR} es no singular y $0 < \omega < 2$, entonces el método iterativo SOR converge para todo valor inicial dado.*

Recordemos que una matriz $A \in M(n \times n, \mathbb{R})$ de la forma

$$A = \begin{pmatrix} a_1 & b_1 & 0 & 0 & \cdots & 0 \\ c_2 & a_2 & b_2 & 0 & \cdots & 0 \\ 0 & c_3 & a_3 & b_3 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & c_{n-1} & a_{n-1} & b_{n-1} \\ 0 & \cdots & 0 & 0 & c_n & a_n \end{pmatrix}$$

es llamada *tridiagonal*.

Teorema 1.24 *Si A es simétirca, positiva definida y tridiagonal, entonces, en caso se tiene $\rho(T_{G\text{-}S}) = \rho(T_J)^2$, donde $T_J = D^{-1}(C_L + C_U)$ es la matriz de iteración en el método de Jacobi y $T_{G\text{-}S} = (D - C_L)^{-1} C_U$ es la matriz de iteración en el método de Gauss–Seidel. Además, la elección óptima de ω en el método SOR es*

$$\boxed{\omega_{opt} = \frac{2}{1 + \sqrt{1 - \rho(T_J)^2}} = \frac{2}{1 + \sqrt{1 - \rho(T_{G\text{-}S})}}} \tag{1.44}$$

Por otra parte, también se tiene

$$\boxed{\rho(T_{SOR,\omega_{opt}}) = \omega_{opt} - 1} \tag{1.45}$$

1.13 Otra forma de expresar los métodos iterativos para sistemas lineales

Sea

$$A = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{pmatrix}$$

podemos escribir A en la forma

$$\begin{aligned} A &= \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{pmatrix} \\ &= \underbrace{\begin{pmatrix} a_{11} & 0 & \cdots & 0 \\ 0 & a_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & a_{nn} \end{pmatrix}}_{D} - \underbrace{\begin{pmatrix} 0 & 0 & \cdots & 0 \\ -a_{21} & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots \\ -a_{n1} & \cdots & -a_{nn-1} & 0 \end{pmatrix}}_{C_L} - \underbrace{\begin{pmatrix} 0 & -a_{12} & \cdots & -a_{1n} \\ 0 & 0 & \ddots & \vdots \\ 0 & 0 & \cdots & -a_{n-1n} \\ 0 & 0 & \cdots & 0 \end{pmatrix}}_{C_U} \end{aligned}$$

Notemos primero que $A\mathbf{x} = \mathbf{b}$ si y sólo si $(D - C_L - C_U)\mathbf{x} = b$ si y sólo si $D\mathbf{x} = (C_L + C_U)\mathbf{x} + \mathbf{b}$.

Veamos como quedan los métodos anteriores con esta notación.

Método de Jacobi. Si $\det(D) \neq 0$, entonces como $(D - C_L - C_U)\mathbf{x} = b$ se sigue que

$$\mathbf{x} = \underbrace{D^{-1}(C_L + C_U)}_{T_j}\mathbf{x} + \underbrace{D^{-1}b}_{C_J} \tag{1.46}$$

luego el método de Jacobi se puede escribir como

$$\mathbf{x}^{(k+1)} = \underbrace{D^{-1}(C_L + C_U)}_{T_j}\mathbf{x}^{(k)} + \underbrace{D^{-1}b}_{C_J}$$

Método de Gauss–Seidel. Este se puede escribir como

$$(D - C_L)\mathbf{x}^{(k+1)} = C_U\mathbf{x}^{(k)} + \mathbf{b} \tag{1.47}$$

de donde, si $\det(D - C_L) \neq 0$ nos queda

$$\mathbf{x}^{(k+1)} = \underbrace{(D - C_L)^{-1}C_U}_{T_{\text{G-S}}}\mathbf{x}^{(k)} + \underbrace{(D - C_L)^{-1}}_{C_{\text{G-S}}}b$$

Resumen. Supongamos que la matriz A se escribe en la forma $A = D - C_L - C_U$, donde $D = diag(A)$, C_L es el negativo de la parte triangular inferior estricta de A y C_U es el negativo de la parte triangular superior estricta de A. Entonces podemos rescribir los métodos iterativos vistos antes como sigue:

Richardson

$$\begin{cases} Q = I \\ G = I - A \end{cases}$$

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

Jacobi

$$\begin{cases} Q = D \\ G = D^{-1}(C_L + C_U) \end{cases}$$

$$D\mathbf{x}^{(k+1)} = (C_L + C_U)\mathbf{x}^{(k)} + \mathbf{b}$$

Gauss-Seidel

$$\begin{cases} Q = D - C_L \\ G = (D - C_L)^{-1}C_U \end{cases}$$

$$(D - C_L)\mathbf{x}^{(k+1)} = C_U\mathbf{x}^{(k)} + \mathbf{b}$$

SOR

$$\begin{cases} Q = \omega^{-1}(D - \omega C_L) \\ G = (D - \omega C_L)^{-1}(\omega C_U + (1 - \omega)D) \end{cases}$$

$$(D - \omega C_L)\mathbf{x}^{(k+1)} = \omega(C_U\mathbf{x}^{(k)} + \mathbf{b}) + (1 - \omega)D\mathbf{x}^{(k)}$$

$$\omega = \tfrac{1}{\alpha},\ 0 < \omega < 2.$$

1.14 Ejemplos resueltos

Problema 1.1 Considere el sistema de ecuaciones lineales $A\mathbf{x} = \mathbf{b}$, donde

$$A = \begin{pmatrix} 2 & 3 & -1 \\ 4 & 4 & -1 \\ -1 & -3 & 4 \end{pmatrix},\ \mathbf{b} = \begin{pmatrix} 2 \\ 5 \\ 7 \end{pmatrix} \text{ y } \mathbf{x} = \begin{pmatrix} x \\ y \\ z \end{pmatrix}$$

1. Obtenga la descomposición $A = LU$, donde la diagonal de L es dada por $l_{11} = 1$, $l_{22} = 3$ y $l_{33} = 4$.

2. Usando la descomposición $A = LU$ que obtuvo en a) encuentre la solución del sistema $A\mathbf{x} = \mathbf{b}$.

3. Usando el método de eliminación gaussiana sin pivoteo encuentre la solución del sistema $A\mathbf{x} = \mathbf{b}$.

4. Explicite la descomposición de Doolittle de A.

5. ¿A tiene descomposición de Cholesky?

6. Examine la convergencia de los métodos de Richardson, Jacobi y Gauss-Seidel para resolver el sistema $A\mathbf{x} = \mathbf{b}$.

7. Considere la matriz
$$Q = \begin{pmatrix} 2 & 4 & 6 \\ 4 & 4 & 10 \\ -1 & -3 & 4 \end{pmatrix}$$
y proponga un método iterativo convergente para encontrar una solución aproximada al sistema $A\mathbf{x} = \mathbf{b}$. La demostración de la convergencia debe hacerla sin iterar

Solución. a) Tenemos que $l_{11} = 1$, $l_{22} = 3$, $l_{33} = 4$. Escribamos

$$\begin{pmatrix} 1 & 0 & 0 \\ l_{21} & 3 & 0 \\ l_{31} & l_{32} & 4 \end{pmatrix} \begin{pmatrix} u_{11} & u_{12} & u_{13} \\ 0 & u_{22} & u_{23} \\ 0 & 0 & u_{33} \end{pmatrix} = \begin{pmatrix} 2 & 3 & -1 \\ 4 & 4 & -1 \\ -1 & -3 & 4 \end{pmatrix}$$

De aquí,
$$\begin{aligned} u_{11} &= 2 \\ u_{12} &= 3 \\ u_{13} &= -1\,. \end{aligned}$$

Ahora, $l_{21}u_{11} = 4$, de donde $l_{21} = 2$ y

$$\begin{aligned} l_{21}u_{12} + l_{22}u_{22} &= 4 \\ l_{21}u_{13} + l_{22}u_{23} &= -1 \end{aligned}$$

de aquí reemplazando nos queda $6 + 3u_{22} = 4$ luego $u_{22} = -\frac{2}{3}$ y $-2 + 3u_{23} = -1$, de donde $u_{23} = \frac{1}{3}$

Finalmente, $l_{31}u_{11} = -1$ de donde $l_{31} = -\frac{1}{2}$ y tenemos

$$\begin{aligned} l_{31}u_{12} + l_{32}u_{22} &= -3 \\ l_{31}u_{13} + l_{32}u_{23} + l_{33}u_{33} &= -1 \end{aligned}$$

reemplazando obtenemos $-\frac{3}{2} - \frac{2}{3}l_{32} = -3$ de donde $l_{32} = \frac{9}{4}$. Por lo tanto, reemplazando en la última ecuación nos queda $\frac{1}{2} + \frac{9}{12} + 4u_{33} = 4$ de donde $u_{33} = \frac{33}{48}$. Por lo tanto,

$$\begin{pmatrix} 1 & 0 & 0 \\ 2 & 3 & 0 \\ -\frac{1}{2} & \frac{9}{4} & 4 \end{pmatrix} \begin{pmatrix} 2 & 3 & -1 \\ 0 & -\frac{2}{3} & \frac{1}{3} \\ 0 & 0 & \frac{33}{48} \end{pmatrix} = \begin{pmatrix} 2 & 3 & -1 \\ 4 & 4 & -1 \\ -1 & -3 & 4 \end{pmatrix}$$

b) Tenemos que $A = LU$, así $A\mathbf{x} = \mathbf{b}$ se rescribe como $LU\mathbf{x} = \mathbf{b}$, llamando $Ux = z$ nos queda

$$\begin{cases} L\mathbf{z} = \mathbf{b} \\ U\mathbf{x} = \mathbf{z} \end{cases}$$

Ahora, $L = \begin{pmatrix} 1 & 0 & 0 \\ 2 & 3 & 0 \\ -\frac{1}{2} & \frac{9}{4} & 4 \end{pmatrix}$, $\mathbf{z} = \begin{pmatrix} z_1 \\ z_2 \\ z_3 \end{pmatrix}$ y $\mathbf{b} = \begin{pmatrix} 2 \\ 5 \\ 7 \end{pmatrix}$. Luego,

$$\begin{pmatrix} 1 & 0 & 0 \\ 2 & 3 & 0 \\ -\frac{1}{2} & \frac{9}{4} & 4 \end{pmatrix} \begin{pmatrix} z_1 \\ z_2 \\ z_3 \end{pmatrix} = \begin{pmatrix} 2 \\ 5 \\ 7 \end{pmatrix}$$

de donde,

$$\begin{aligned} z_1 &= 2 \\ 2z_1 + 3z_2 &= 5 \\ -\tfrac{1}{2}z_1 + \tfrac{9}{4}z_2 + 4z_3 &= 7 \end{aligned}$$

reemplazando nos queda, $4 + 3z_2 = 5$ luego $z_2 = \frac{1}{3}$, y en la tercera ecuación obtenemos, $-\frac{1}{2}2 + \frac{9}{4} \cdot \frac{1}{3} + 4z_3 = 7$ de donde, $z_3 = \frac{29}{16}$.

Ahora, de $Ux = z$ obtenemos

$$\begin{pmatrix} 2 & 3 & -1 \\ 0 & -\frac{2}{3} & \frac{1}{3} \\ 0 & 0 & \frac{33}{48} \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 2 \\ \frac{1}{3} \\ \frac{29}{16} \end{pmatrix}$$

luego, $\frac{33}{48}z = \frac{29}{16}$ y de aquí $z = \frac{29}{11}$. Además, $-\frac{2}{3}y + \frac{1}{3}z = \frac{1}{3}$, reemplazando nos queda $-2y + \frac{29}{11} = 1$ de donde $y = \frac{9}{11}$. Finalmente, $2x + 3y - z = 2$, remplazando, $2x + 3\frac{9}{11} - \frac{29}{11} = 2$ de donde $x = \frac{12}{11}$. La solución del sistema es $\left(\frac{12}{11}, \frac{9}{11}, \frac{29}{11}\right)$.

c) realizando la eliminación gaussiana, tenemos

$$\left(\begin{array}{ccc|c} 2 & 3 & -1 & 2 \\ 4 & 4 & -1 & 5 \\ -1 & -3 & 4 & 7 \end{array}\right) \begin{array}{c} F_2 - 2F_1 \\ F_3 - (-\frac{1}{2})F_1 \end{array} \left(\begin{array}{ccc|c} 2 & 3 & -1 & 2 \\ 0 & -2 & 1 & 1 \\ 0 & -\frac{3}{2} & \frac{7}{2} & 8 \end{array}\right) \quad F_3 - \tfrac{3}{4}F_2 \quad \left(\begin{array}{ccc|c} 2 & 3 & -1 & 2 \\ 0 & -2 & 1 & 1 \\ 0 & 0 & \frac{11}{4} & \frac{29}{4} \end{array}\right)$$

Luego $\frac{11}{4}z = \frac{29}{4}$, de donde $z = \frac{29}{11}$, reemplazando en $-2y + z = 1$ nos da que $y = \frac{9}{11}$. Finalmente de $2x + 3y - z = 2$ tenemos que $x = \frac{12}{11}$.

d) De lo anterior tenemos la descomposición de Doolittle de A es

$$\begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ -\frac{1}{2} & \frac{3}{4} & 1 \end{pmatrix} \begin{pmatrix} 2 & 3 & -1 \\ 0 & -2 & 1 \\ 0 & 0 & \frac{11}{4} \end{pmatrix}$$

e) No existe descomposición de Cholesky pues la matriz A no es simétrica.

f) Para Jacobi, tenemos

$$Q = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 4 \end{pmatrix}, \qquad Q^{-1} = \begin{pmatrix} \frac{1}{2} & 0 & 0 \\ 0 & \frac{1}{4} & 0 \\ 0 & 0 & \frac{1}{4} \end{pmatrix}, \qquad Q^{-1}A = \begin{pmatrix} 1 & \frac{3}{2} & -\frac{1}{2} \\ 1 & 1 & -\frac{1}{4} \\ -\frac{1}{4} & -\frac{3}{4} & 1 \end{pmatrix},$$

luego

$$J = I - Q^{-1}A = \begin{pmatrix} 0 & -\frac{3}{2} & \frac{1}{2} \\ -1 & 0 & \frac{1}{4} \\ \frac{1}{4} & \frac{3}{4} & 0 \end{pmatrix}$$

$J-\lambda I=\begin{pmatrix} -\lambda & -\frac{3}{2} & \frac{1}{2} \\ -1 & -\lambda & \frac{1}{4} \\ \frac{1}{4} & \frac{3}{4} & -\lambda \end{pmatrix}$, calculando nos queda $\det(J-\lambda I)=-\lambda^3+\frac{29}{16}\lambda-\frac{15}{32}=0$, obtenemos los autovalores $\lambda_1=-1.460624640$, $\lambda_2=1.191215504$, $\lambda_3=0.2694091369$, por lo tanto $\rho(J)>1$, y el método de Jacobi no converge.

Para Gauss-Seidel

$$Q=\begin{pmatrix} 2 & 0 & 0 \\ 4 & 4 & 0 \\ -1 & -3 & 4 \end{pmatrix}, \qquad Q^{-1}=\begin{pmatrix} \frac{1}{2} & 0 & 0 \\ -\frac{1}{2} & \frac{1}{4} & 0 \\ -\frac{1}{4} & \frac{3}{16} & \frac{1}{4} \end{pmatrix}, \qquad Q^{-1}A=\begin{pmatrix} 1 & \frac{3}{2} & -\frac{1}{2} \\ 0 & -\frac{1}{2} & \frac{1}{4} \\ 0 & -\frac{3}{4} & \frac{17}{16} \end{pmatrix},$$

$$GS=I-Q^{-1}A=\begin{pmatrix} 0 & -\frac{3}{2} & \frac{1}{2} \\ 0 & \frac{3}{2} & -\frac{1}{4} \\ 0 & \frac{3}{4} & -\frac{1}{16} \end{pmatrix} \text{ y } (GS-\lambda I)=\begin{pmatrix} -\lambda & -\frac{3}{2} & \frac{1}{2} \\ 0 & \frac{3}{2}-\lambda & -\frac{1}{4} \\ 0 & \frac{3}{4} & -\frac{1}{16}-\lambda \end{pmatrix}$$

solucionando $\det(GS-\lambda I)=-\lambda^3+\frac{23}{16}\lambda^2-\frac{3}{32}\lambda=0$ obtenemos los valores propios

$\lambda_1=0$, $\lambda_2=1.369020376$, $\lambda_3=0.06847962360$, luego $\rho(GS)>1$ y Gauss-Seidel no converge.

Para Richardson nos queda

$R=I-A=\begin{pmatrix} -1 & -3 & 1 \\ -4 & -3 & 1 \\ 1 & 3 & -3 \end{pmatrix}$ y $R-\lambda I=\begin{pmatrix} -1-\lambda & -3 & 1 \\ -4 & -3-\lambda & 1 \\ 1 & 3 & -3-\lambda \end{pmatrix}$, solucionando $\det(R-\lambda I)=-\lambda^3-7\lambda^2+\lambda+18=0$ obtenemos los valores propios $\lambda_1=1.513935115$, $\lambda_2=-6.753410526$ y $\lambda_3=-1.760524588$, luego $\rho(R)>1$ y el método de Richardson no converge.

Ahora, usando la matriz

$$Q=\begin{pmatrix} 2 & 4 & 6 \\ 4 & 4 & 10 \\ -1 & -3 & 4 \end{pmatrix}$$

tenemos

$$Q^{-1}=\begin{pmatrix} -\frac{23}{30} & \frac{17}{30} & -\frac{4}{15} \\ \frac{13}{30} & -\frac{7}{30} & -\frac{1}{15} \\ \frac{2}{15} & -\frac{1}{30} & \frac{2}{15} \end{pmatrix}, Q^{-1}A=\begin{pmatrix} 1 & \frac{23}{30} & -\frac{13}{15} \\ 0 & \frac{17}{30} & -\frac{7}{15} \\ 0 & -\frac{2}{15} & \frac{13}{30} \end{pmatrix}, P=I-Q^{-1}A=\begin{pmatrix} 0 & -\frac{23}{30} & \frac{13}{15} \\ 0 & \frac{13}{30} & \frac{7}{15} \\ 0 & \frac{2}{15} & \frac{17}{30} \end{pmatrix}$$

y $P-\lambda I=\begin{pmatrix} -\lambda & -\frac{23}{30} & \frac{13}{15} \\ 0 & \frac{13}{30}-\lambda & \frac{7}{15} \\ 0 & \frac{2}{15} & \frac{17}{30}-\lambda \end{pmatrix}$,

solucionando $\det(P-\lambda I)=-\lambda^3+\lambda^2-\frac{11}{60}\lambda=0$ obtenemos los valores propios $\lambda_1=0$, $\lambda_2=0.2418011101$ y $\lambda_3=0.7581988896$, de donde $\rho(P)=\lambda_3<1$ y el método propuesto converge.

Problema 1.2 Considere el sistema de ecuaciones lineales

$$\begin{pmatrix} \frac{1}{3} & \frac{1}{4} \\ \\ \frac{1}{4} & \frac{1}{5} \end{pmatrix} \begin{pmatrix} x_1 \\ \\ x_2 \end{pmatrix} = \begin{pmatrix} \frac{7}{12} \\ \\ 0.45 \end{pmatrix}.$$

(a) Para realizar los cálculos con decimales, se considera el vector $\hat{\mathbf{b}} = \begin{pmatrix} 0.58 \\ 0.45 \end{pmatrix}$. Obtenga una cota para el error relativo de la solución del sistema con respecto a la solución del sistema $A\mathbf{x} = \hat{\mathbf{b}}$.

(b) Ahora considere la matriz perturbada

$$\hat{A} = \begin{pmatrix} 0.33 & 0.25 \\ 0.25 & 0.20 \end{pmatrix}.$$

Obtenga una cota para el error relativo de la solución del sistema original con respecto a la solución del sistema $\hat{A}\mathbf{x} = \mathbf{b}$.

Solución. Tenemos el sistema de ecuaciones lineales

$$\begin{pmatrix} \frac{1}{3} & \frac{1}{4} \\ \\ \frac{1}{4} & \frac{1}{5} \end{pmatrix} \begin{pmatrix} x_1 \\ \\ x_2 \end{pmatrix} = \begin{pmatrix} \frac{7}{12} \\ \\ 0.45 \end{pmatrix}.$$

a) El vector $\hat{\mathbf{b}} = \begin{pmatrix} 0.58 \\ 0.45 \end{pmatrix}$ es una perturbación del vector $\mathbf{b} = \begin{pmatrix} \frac{7}{12} \\ \\ 0.45 \end{pmatrix}$. Sean $\mathbf{x}_T$ la solución exacta del sistema $A\mathbf{x} = \mathbf{b}$ y $\mathbf{x}_A$ la solución exacta del sistema $A\mathbf{x} = \hat{\mathbf{b}}$. Se tiene entonces que $\mathbf{x}_A$ es una aproximación a $\mathbf{x}_T$. Para simplificar los cálculos trabajaremos con la norma subordinada $||\cdot||_\infty$. Como $\hat{\mathbf{b}}$ es una perturbación de $\mathbf{b}$, tenemos la fórmula

$$E_R(\mathbf{x}_A) = \frac{||\mathbf{x}_T - \mathbf{x}_A||_\infty}{||\mathbf{x}_T||_\infty} \leqslant ||A||_\infty ||A^{-1}||_\infty \frac{||\mathbf{b} - \hat{\mathbf{b}}||_\infty}{||\mathbf{b}||_\infty}.$$

Ahora,

$$\mathbf{b} - \hat{\mathbf{b}} = \begin{pmatrix} 3.33333333 \times 10^{-3} \\ 0 \end{pmatrix}.$$

Recordemos que si $A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$, entonces

$$A^{-1} = \frac{1}{\det(A)} \begin{pmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{pmatrix}$$

En nuestro caso, $\det \begin{pmatrix} \frac{1}{3} & \frac{1}{4} \\ \\ \frac{1}{4} & \frac{1}{5} \end{pmatrix} = \frac{1}{240}$. Luego,

$$A^{-1} = \frac{1}{\frac{1}{240}} \begin{pmatrix} \frac{1}{5} & -\frac{1}{4} \\ -\frac{1}{4} & \frac{1}{3} \end{pmatrix} = 240 \begin{pmatrix} \frac{1}{5} & -\frac{1}{4} \\ -\frac{1}{4} & \frac{1}{3} \end{pmatrix} = \begin{pmatrix} 48 & -60 \\ -60 & 80 \end{pmatrix}.$$

De esto, tenemos

$$\begin{aligned} ||A||_\infty &= \max\left\{\frac{1}{3}+\frac{1}{4}, \frac{1}{4}+\frac{1}{5}\right\} = \max\left\{\frac{7}{12}, \frac{9}{20}\right\} = \frac{7}{12} \\ ||A^{-1}||_\infty &= \max\{|48|+|-60|, |-60|+|80|\} = 140 \\ ||\mathbf{b}||_\infty &= \max\left\{\frac{7}{12}, 0.45\right\} = \frac{7}{12} \\ ||\hat{\mathbf{b}}||_\infty &= \max\{0.58, 0.45\} = 0.58 \\ ||\mathbf{b}-\hat{\mathbf{b}}||_\infty &= \max\{3.33333333\times 10^{-3}, 0\} = 3.33333333\times 10^{-3}. \end{aligned}$$

Reemplazando nos queda

$$E_R(\mathbf{x}_A) \leqslant \frac{7}{12}\, 140\, \frac{3.3333333\times 10^{-3}}{\frac{7}{12}} = 0.4666666662,$$

esto es, $E_R(\mathbf{x}_A) \leqslant 0.4666666662$.

b) Consideremos ahora la matriz perturbada

$$\hat{A} = \begin{pmatrix} 0.33 & 0.25 \\ 0.25 & 0.20 \end{pmatrix}.$$

Podemos escribir $\hat{A}$ en la forma $\hat{A} = A(I+E)$, donde I es la matriz identidad 2×2 y E es la matriz de error, esto es, $.E = \hat{A} - A$. Como $\hat{A} = A(I+E)$, se tiene que $\hat{A} - A = AE$. Luego,

$$AE = \hat{A} - A = \begin{pmatrix} 0.33 & 0.25 \\ 0.25 & 0.20 \end{pmatrix} - \begin{pmatrix} \frac{1}{3} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{5} \end{pmatrix} = \begin{pmatrix} -3.33333333\times 10^{-3} & 0 \\ 0 & 0 \end{pmatrix},$$

de donde $||AE||_\infty = 3.33333333\times 10^{-3}$.

Veamos si podemos usar la fórmula

$$\frac{||\mathbf{x}_T - \mathbf{x}_A||}{||\mathbf{x}_T||} \leqslant \frac{k(A)}{1 - k(A)\,\frac{||AE||}{||A||}} \cdot \frac{||AE||}{||A||},$$

para ello debemos verificar que

$$||AE|| < \frac{1}{||A^{-1}||}.$$

Como $||A||_\infty = 140$ y $||AE||_\infty = 3.33333333 \times 10^{-3}$, se cumple que

$$||AE||_\infty < \frac{1}{140} = 7.14285714 \times 10^{-3},$$

y podemos aplicar la fórmula anterior. Reemplazando nos queda

$$\begin{aligned}
\frac{||\mathbf{x}_T - \mathbf{x}_A||_\infty}{||\mathbf{x}_T||_\infty} &\leqslant \frac{81.66666667}{1 - 81.66666667 \times \frac{3.33333333 \times 10^{-3}}{0.5833333333}} \times \frac{3.3333333 \times 10^{-3}}{0.5833333333} \\
&= \frac{81.66666667}{1 - 81.66666667 \times 5.7142857 \times 10^{-3}} \times 5.7142857 \times 10^{-3} \\
&= \frac{0.466666667}{0.5333333349} = 0.8749999945
\end{aligned}$$

Luego $E_R(\mathbf{x}_A) \leqslant 0.8749999945$.

Problema 1.3 Dada la matriz

$$A = \begin{bmatrix} 2 & -1 & -1 \\ -1 & 1 & 1 \\ -1 & 1 & 3 \end{bmatrix} \text{ se tiene que } A^{-1} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 5/2 & -1/2 \\ 0 & -1/2 & 1/2 \end{bmatrix}$$

(a) Obtenga la descomposición de Cholesky de A. Utilice lo anterior para resolver el sistema $A\mathbf{x} = \mathbf{b}$ para el vector $b = (1, 1, 2)^T$.

(b) Estime a priori la magnitud del error relativo si la matriz A se perturba quedando finalmente

$$\tilde{A} = \begin{bmatrix} 2 & -0.99 & -0.98 \\ -1 & 1 & 0.99 \\ -1 & 1.02 & 2.99 \end{bmatrix}$$

Solución.

(a) Tenemos que

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

es simétrica. Ahora, $A_1 = [2]$, luego $\det(A_1) = 2 > 0$, $A_2 = \begin{bmatrix} 2 & -1 \\ -1 & 1 \end{bmatrix}$ y $\det(A_2) = 1 > 0$, $A_3 = A = \begin{bmatrix} 2 & -1 & -1 \\ -1 & 1 & 1 \\ -1 & 1 & 3 \end{bmatrix}$ y $\det(A_3) = 2 > 0$. Por lo tanto A es positiva definida, y consecuentemente tiene descomposición de Cholesky.

Para obtener la descompsición de Cholesky de A escribamos

$$\begin{bmatrix} 2 & -1 & -1 \\ -1 & 1 & 1 \\ -1 & 1 & 3 \end{bmatrix} = \begin{bmatrix} \ell_{11} & 0 & 0 \\ \ell_{21} & \ell_{22} & 0 \\ \ell_{31} & \ell_{32} & \ell_{33} \end{bmatrix} \begin{bmatrix} \ell_{11} & \ell_{21} & \ell_{31} \\ 0 & \ell_{22} & \ell_{32} \\ 0 & 0 & \ell_{33} \end{bmatrix}$$

De aquí, $\ell_{11}^2 = 2$, de donde $\ell_{11} = \sqrt{2}$; $\ell_{11}\ell_{21} = -1$, de donde $\ell_{21} = -\frac{\sqrt{2}}{2}$, $\ell_{11}\ell_{31} = -1$; $\ell_{31} = -\frac{\sqrt{2}}{2}$; $\ell_{21}^2 + \ell_{22}^2 = 1$, de donde $\ell_{22} = \frac{\sqrt{2}}{2}$; $\ell_{21}\ell_{31} + \ell_{22}\ell_{32} = 1$, de donde $\ell_{32} = \frac{\sqrt{2}}{2}$; finalmente $\ell_{31}^2 + \ell_{32}^2 + \ell_{33}^2 = 3$, de donde $\ell_{33} = \sqrt{2}$. Por lo tanto,

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

Ahora, resolver el sistema $A\mathbf{x} = \mathbf{b}$, con $\mathbf{b}^T = (1, 1, 2)$ es equivalente a resolver los sistemas

$$\begin{cases} L\mathbf{y} = \mathbf{b} \\ L^T\mathbf{x} = \mathbf{y} \end{cases}$$

El sistema $L\mathbf{y} = \mathbf{b}$

$$\begin{bmatrix} \sqrt{2} & 0 & 0 \\ -\frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} & 0 \\ -\frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} & \sqrt{2} \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \\ \gamma \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 2 \end{bmatrix}$$

De aquí $\sqrt{2}\,\alpha = 1$, de donde $\alpha = \frac{\sqrt{2}}{2}$; $-\frac{\sqrt{2}}{2}\alpha + \frac{\sqrt{2}}{2}\beta = 1$, reemplazando y despejando nos da que $\beta = \frac{3}{2}\sqrt{2}$; finalmente $-\frac{\sqrt{2}}{2}\alpha + \frac{\sqrt{2}}{2}\beta + \sqrt{2}\gamma = 2$, reemplazando y despejando nos da que $\gamma = \frac{\sqrt{2}}{2}$. Debemos resolver ahora el sistema $L^T\mathbf{x} = \mathbf{y}$, es decir,

$$\begin{bmatrix} \sqrt{2} & -\frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \\ 0 & \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \\ 0 & 0 & \sqrt{2} \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} \frac{\sqrt{2}}{2} \\ \frac{3\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} \end{bmatrix}$$

de donde $z = \frac{1}{2}$, $y = \frac{5}{2}$, $x = 2$.

(b) Sean $\mathbf{x}_T$ y $\mathbf{x}_A$ las soluciones exactas de los sistemas $A\mathbf{x} = \mathbf{b}$ y $\tilde{A}\mathbf{x} = \mathbf{b}$, respectivamente, es decir, $\mathbf{x}_T = A^{-1}\mathbf{b}$. Entonces

$$\begin{aligned} E(\mathbf{x}_A = ||\mathbf{x}_A - \mathbf{x}_T|| &= ||\mathbf{x}_A - A^{-1}\mathbf{b}|| \\ &= ||\mathbf{x}_A - A^{-1}\tilde{A}\mathbf{x}_A|| = ||(I - A^{-1}\tilde{A})\mathbf{x}_A|| \\ &\leqslant ||I - A^{-1}\tilde{A}||\;||\mathbf{x}_A|| \end{aligned}$$

de donde

$$E_R(\mathbf{x}_A) = \frac{||\mathbf{x}_A - \mathbf{x}_T||}{||\mathbf{x}_A||} \leqslant ||I - A^{-1}\tilde{A}||\,.$$

Para simplificar los cálculos usamos la norma subordinada $||\;||_\infty$. Ahora, tenemos

$$A^{-1} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & \frac{5}{2} & \frac{-1}{2} \\ 0 & \frac{-1}{2} & \frac{1}{2} \end{bmatrix}$$

$$\tilde{A} = \begin{bmatrix} 2 & -0.99 & -0.98 \\ -1 & 1 & 0.99 \\ -1 & 1.02 & 2.99 \end{bmatrix}$$

luego,

$$A^{-1}\tilde{A} = \begin{bmatrix} 1 & 0.01 & 0.01 \\ 0 & 1.0 & 0 \\ 0 & 0.01 & 1.0 \end{bmatrix}$$

Además, como

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

se tiene que

$$I - A^{-1}\tilde{A} = \begin{bmatrix} 0 & -0.01 & -0.01 \\ 0 & 0 & 0 \\ 0 & -0.01 & 0 \end{bmatrix}$$

luego, $||I-A^{-1}\tilde{A}||_\infty = \max\{0.02, 0, 0.01\} = 0.02$, de donde $\frac{||\mathbf{x}_A-\mathbf{x}_T||_\infty}{||\mathbf{x}_A||_\infty} \leqslant ||I-A^{-1}\tilde{A}||_\infty = 0.02$.

Otra solución. Escribimos $\tilde{A}$ de la forma $\tilde{A} = A(I+E)$. Entonces se tiene las fórmulas

1.
$$E_R(\mathbf{x}_A) = \frac{||\mathbf{x}_T - \mathbf{x}_A||}{||\mathbf{x}_T||} \leqslant \frac{||E||}{1-||E||}$$
si $||E|| < 1$.

2.
$$E_R(\mathbf{x}_A) = \frac{||\mathbf{x}_T - \mathbf{x}_A||}{||\mathbf{x}_T||} \leqslant \frac{k(A)}{1-k(A)\frac{||AE||}{||A||}} \frac{||AE||}{||A||}$$
si $||AE|| < \frac{1}{||A^{-1}||}$.

Usemos por ejemplo la primera de ellas. Debemos calcular la matriz E . De la ecuación $\tilde{A} = A(I+E)$ se tiene que $A^{-1}\tilde{A} = I + E$, de donde, $A^{-1}\tilde{A} - I = E$. Realizando los cálculos, obtenemos que

$$E = \begin{bmatrix} 0 & 0.01 & 0.01 \\ 0 & 0 & 0 \\ 0 & 0.01 & 0 \end{bmatrix}$$

Por simplicidad de los cálculos, usaremos la norma subordinada $||\ ||_\infty$. Luego, $||E||_\infty = 0.02 < 1$, por lo tanto, reemplazando nos queda,

$$\frac{||\mathbf{x}_T - \mathbf{x}_A||}{||\mathbf{x}_T||} \leqslant \frac{||E||}{1-||E||} = \frac{0.02}{1-0.02} = \frac{0.02}{0.98} = 0.0204016...$$

Ahora usamos la segunda de las fórmulas. Tenemos que calcular $k(A) = ||A||\,||A^{-1}||$. Sabemos que

$$A = \begin{bmatrix} 2 & -1 & -1 \\ -1 & 1 & 1 \\ -1 & 1 & 3 \end{bmatrix} \quad \text{luego} \quad ||A||_\infty = 5$$

y

$$A^{-1} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & \frac{5}{2} & \frac{-1}{2} \\ 0 & \frac{-1}{2} & \frac{1}{2} \end{bmatrix} \quad \text{luego} \quad ||A^{-1}||_\infty = 4\,,$$

de donde $k_\infty(A) = 5 \times 4 = 20$. Para aplicar la fórmula debemos verificar que $||AE||_\infty < \frac{1}{||A^{-1}||_\infty}$. Tenemos que

$$AE = \begin{bmatrix} 0 & 0.01 & 0.01 \\ 0 & 0.005 & 0.01 \\ 0 & 0.005 & 0 \end{bmatrix}$$

luego, $||AE||_\infty = 0.02$ y $||A^{-1}||_\infty = 4$, de aquí se tiene que $\frac{1}{||A^{-1}||_\infty} = \frac{1}{4} = 0.25$ y es claro entonces que se satisface la condición $||AE||_\infty < \frac{1}{||A^{-1}||_\infty}$. Reemplazando los valores obtenidos, nos queda

$$\begin{aligned} \frac{||\mathbf{x}_T - \mathbf{x}_A||_\infty}{||\mathbf{x}_T||_\infty} &\leqslant \frac{20}{1 - 20 \cdot \frac{0.02}{5}} \cdot \frac{0.02}{5} = \frac{20 \times 0.004}{1 - 20 \times 0.004} \\ &= \frac{0.08}{1-0.08} = \frac{0.08}{0.92} = 0.08695652... \end{aligned}$$

esto es,

$$\frac{||\mathbf{x}_T - \mathbf{x}_A||_\infty}{||\mathbf{x}_T||_\infty} \leqslant 0.08695652...$$

Problema 1.4 Sean

$$A = \begin{pmatrix} 1 & -1 & -1 & -1 \\ 0 & 1 & -1 & -1 \\ 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 1 \end{pmatrix} \quad \text{y} \quad b = \begin{pmatrix} 5.00 \\ 1.02 \\ 1.04 \\ 1.10 \end{pmatrix}$$

(a) Calcule una solución aproximada $\mathbf{x}_A$ del sistema $A\mathbf{x} = \mathbf{b}$, primero aproximando cada entrada del vector $\mathbf{b}$ al entero más próximo, obteniendo un vector $\tilde{\mathbf{b}}$ y luego resolviendo el sistema $A\mathbf{x} = \tilde{\mathbf{b}}$.

(b) Calcule $||\mathbf{r}||_\infty$ y $k_\infty(A)$, donde $\mathbf{r}$ es el vector residual, es decir $\mathbf{r} = \mathbf{b} - A\mathbf{x}_A$ y $k_\infty(A)$ el número de condicionamiento de la matriz A.

(c) Estime una cota para el error relativo de la solución aproximada, respecto a la solución exacta (no calcule esta última explícitamente).

Solución.

(a) El vector perturbado es $(5\ 1\ 1\ 1)^T$. Luego la solución del sistema perturbado es

$$\begin{pmatrix} 1 & -1 & -1 & -1 \\ 0 & 1 & -1 & -1 \\ 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \\ w \end{pmatrix} = \begin{pmatrix} 5 \\ 1 \\ 1 \\ 1 \end{pmatrix}$$

de donde $(x, y, z, w) = (12, 4, 2, 1) = \mathbf{x}_A$

(b) Tenemos $||b||_\infty = ||\tilde{b}||_\infty = 5$. Por otra parte,

$$\begin{aligned} r = b - A\mathbf{x}_A &= \begin{pmatrix} 5.00 \\ 1.02 \\ 1.04 \\ 1.10 \end{pmatrix} - \begin{pmatrix} 1 & -1 & -1 & -1 \\ 0 & 1 & -1 & -1 \\ 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 12 \\ 4 \\ 2 \\ 1 \end{pmatrix} \\ &= \begin{pmatrix} 5.00 \\ 1.02 \\ 1.04 \\ 1.10 \end{pmatrix} - \begin{pmatrix} 5 \\ 1 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0.02 \\ 0.04 \\ 0.10 \end{pmatrix} \end{aligned}$$

Luego, $||r||_\infty = 0.10$.

Para encontrar la inversa de la matriz A basta resolver el sistema

$$\begin{pmatrix} 1 & -1 & -1 & -1 \\ 0 & 1 & -1 & -1 \\ 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

de donde

$$A^{-1} = \begin{pmatrix} 1 & 1 & 2 & 4 \\ 0 & 1 & 1 & 2 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

Luego, $||A^{-1}||_\infty = 8$. Tenemos que $||A||_\infty = 4$, luego $k_\infty(A) = ||A||_\infty ||A^{-1}||_\infty = 4 \times 8 = 32$.

(c) Usamos la fórmula

$$\frac{1}{k(A)}\,\frac{||r||}{||b||} \leqslant \frac{||\mathbf{x}_T - \mathbf{x}_A||}{||\mathbf{x}_T||} \leqslant k(A)\,\frac{||r||}{||b||}$$

Reemplazando los datos nos queda

$$\frac{1}{32} \times \frac{0.10}{5} \leqslant \frac{||\mathbf{x}_T - \mathbf{x}_A||}{||\mathbf{x}_T||} \leqslant 32 \times \frac{0.10}{5}$$

es decir,

$$0.625 \times 10^{-3} \leqslant \frac{||\mathbf{x}_T - \mathbf{x}_A||}{||\mathbf{x}_T||} \leqslant 0.64\,.$$

Problema 1.5 Encuentre una descomposición del tipo LU para la matriz A siguiente

$$A = \begin{pmatrix} 4 & 3 & 2 & 1 \\ 3 & 4 & 3 & 2 \\ 2 & 3 & 4 & 3 \\ 1 & 2 & 3 & 4 \end{pmatrix}$$

Use la descomposión anterior para solucionar el sistema $A\mathbf{x} = (1\ \ 1\ \ -1\ \ -1)^T$.

Solución. La descomposición de Doolittle de A es dada por

$$A = L \cdot U = \begin{pmatrix} 1 & 0 & 0 & 0 \\ \frac{3}{4} & 1 & 0 & 0 \\ \frac{1}{2} & \frac{6}{7} & 1 & 0 \\ \frac{1}{4} & \frac{5}{7} & \frac{5}{6} & 1 \end{pmatrix} \cdot \begin{pmatrix} 4 & 3 & 2 & 1 \\ 0 & \frac{7}{4} & \frac{3}{2} & \frac{5}{4} \\ 0 & 0 & \frac{12}{7} & \frac{10}{7} \\ 0 & 0 & 0 & \frac{5}{3} \end{pmatrix}$$

Ahora resolver el sistemas $A\mathbf{x} = b$ es equivalente a resolver los sistemas $L\mathbf{y} = b$ y $U\mathbf{x} = y$. Tenemos que $b = (1\ 1\ \ -1\ \ -1)^T$. Llamando $\mathbf{y} = (y_1\ y_2\ y_3\ y_4)^T$, de la forma del sistema $L\mathbf{y} = b$, obtenemos que $y_1 = 1$, $y_2 = \frac{1}{4}$, $y_3 = -\frac{12}{7}$, y $y_4 = 0$. Ahora al resolver el sistema $U\mathbf{x} = b$ obtenemos $x_4 = 0$, $x_3 = -1$, $x_2 = 1$ y $x_1 = 0$.

Problema 1.6 Considere el sistema de ecuaciones lineales $A\mathbf{x} = b$, donde

$$A = \begin{pmatrix} 2 & -3 & 8 & 1 \\ 4 & 0 & 1 & -10 \\ 16 & 4 & -2 & 1 \\ 0 & 7 & -1 & 5 \end{pmatrix} \quad \text{y} \quad \mathbf{b} = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}$$

a) Usando aritmética exacta y el método de eliminación gaussiana con pivoteo resuelva el sistema de ecuaciones lineales $A\mathbf{x} = b$.

b) Denote por B a la matriz PA obtenida en el item anterior Demuestre que el método iterativo de Jacobi aplicado a la matriz B es convergente. Usando como punto inicial (0.0377,0.21819,0.20545,-0.06439) y la máxima capacidad de su calculadora, realice iteraciones con el método de Jacobi para obtener una solución aproximada del sistema $B\mathbf{x} = b$. Use como criterio de parada $\|(x_{n+1}, y_{n+1}, z_{n+1}, w_{n+1}) - (x_n, y_n, z_n, w_n)\| \leqslant 10^{-5}$.

Solución. a) Tenemos que

$$A = \begin{pmatrix} 2 & -3 & 8 & 1 \\ 4 & 0 & 1 & -10 \\ 16 & 4 & -2 & 1 \\ 0 & 7 & -1 & 5 \end{pmatrix} \quad \text{y} \quad b = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}.$$

Escribiendo la matrix ampliada del sistema, nos queda

$$A = \left(\begin{array}{cccc|c} 2 & -3 & 8 & 1 & 1 \\ 4 & 0 & 1 & -10 & 1 \\ 16 & 4 & -2 & 1 & 1 \\ 0 & 7 & -1 & 5 & 1 \end{array}\right),$$

de donde

$$s = \begin{pmatrix} 8 \\ 10 \\ 16 \\ 7 \end{pmatrix} \quad \text{y} \quad P = \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \end{pmatrix}.$$

Luego

$$m_1 = \max\left\{\frac{2}{8}, \frac{4}{10}, \frac{16}{16}, 0\right\} = 1 = m_{31}$$

por lo tanto debemos intercambiar la fila 1 con la fila 3. Realizando este intercambio de filas y la eliminación gaussiana, nos queda

$$\left(\begin{array}{cccc|c} 16 & 4 & -2 & 1 & 1 \\ 4 & 0 & 1 & -10 & 1 \\ 2 & -3 & 8 & 1 & 1 \\ 0 & 7 & -1 & 5 & 1 \end{array}\right) \begin{array}{c} F_2 - \frac{1}{4}F_1 \\ \rightarrow \\ \\ F_3 - \frac{1}{8}F_1 \end{array} \left(\begin{array}{cccc|c} 16 & 4 & -2 & 1 & 1 \\ 0 & -1 & 3/2 & -41/4 & 3/4 \\ 0 & -7/2 & 33/4 & 7/8 & 7/8 \\ 0 & 7 & -1 & 5 & 1 \end{array}\right)$$

de donde

$$L_1 = \begin{pmatrix} 1 \\ 1/4 \\ 1/8 \\ 0 \end{pmatrix}, \quad s_1 = \begin{pmatrix} 16 \\ 10 \\ 8 \\ 7 \end{pmatrix}, \; P_1 = \begin{pmatrix} 3 \\ 2 \\ 1 \\ 4 \end{pmatrix}$$

Para la elección del segundo pivote tenemos

$$m_2 = \max\left\{\frac{1}{10}, \frac{7}{16}, \frac{7}{7}\right\} = 1 = m_{42}$$

por lo tanto debemos intercambiar la fila 2 con la fila 4

$$\left(\begin{array}{ccccc} 16 & 4 & -2 & 1 & 1 \\ 0 & 7 & -1 & 5 & 1 \\ 0 & -7/2 & 33/4 & 7/8 & 7/8 \\ 0 & -1 & 3/2 & -41/4 & 3/4 \end{array}\right) \begin{array}{c} F_3 - \frac{-1}{2}F_2 \\ \rightarrow \\ \\ F_4 - \frac{-1}{7}F_2 \end{array} \left(\begin{array}{cccc|c} 16 & 4 & -2 & 1 & 1 \\ 0 & 7 & -1 & 5 & 1 \\ 0 & 0 & 31/4 & 27/8 & 11/8 \\ 0 & 0 & 19/14 & -227/28 & 25/28 \end{array}\right)$$

luego,

$$L_1^1 = \begin{pmatrix} 1 \\ 0 \\ 1/8 \\ 1/4 \end{pmatrix},\ L_2 = \begin{pmatrix} 0 \\ 1 \\ -1/2 \\ -1/7 \end{pmatrix},\ P_2 = \begin{pmatrix} 3 \\ 4 \\ 1 \\ 2 \end{pmatrix},\ S_2 = \begin{pmatrix} 16 \\ 7 \\ 8 \\ 10 \end{pmatrix}$$

Para la elección del tercer pivote tenemos que

$$m_3 = \max\left\{\frac{31}{32}, \frac{19}{140}\right\} = \frac{31}{32} = m_{33}$$

y obtenemos

$$\left(\begin{array}{cccc|c} 16 & 4 & -2 & 1 & \\ 0 & 7 & -1 & 5 & 1 \\ 0 & 0 & 31/4 & 27/8 & 11/8 \\ 0 & 0 & 19/14 & -227/28 & 25/8 \end{array}\right) \quad \begin{array}{c} \\ \\ \\ F_4 - \frac{38}{217}F_3 \end{array} \quad \left(\begin{array}{cccc|c} 16 & 4 & -2 & 1 & 1 \\ 0 & 7 & -1 & 5 & 1 \\ 0 & 0 & 31/4 & 27/8 & 11/8 \\ 0 & 0 & 0 & -4395/434 & 283/434 \end{array}\right)$$

luego,

$$L_3 = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 38/217 \end{pmatrix}$$

Por lo tanto

$$PA = LU = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 1/8 & -1/2 & 1 & 0 \\ 1/4 & -1/7 & 38/217 & 1 \end{pmatrix} \begin{pmatrix} 16 & 4 & -2 & 1 \\ 0 & 7 & -1 & 5 \\ 0 & 0 & 31/4 & 27/8 \\ 0 & 0 & 0 & -4395/434 \end{pmatrix} \tag{1.48}$$

donde

$$P = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}.$$

Por lo tanto

$$PA = \begin{pmatrix} 16 & 4 & -2 & 1 \\ 0 & 7 & -1 & 5 \\ 2 & -3 & 8 & 1 \\ 4 & 0 & 1 & -10 \end{pmatrix}$$

Resolviendo la ecuación se obtiene que

$$w = -\frac{283}{4395},\ z = \frac{301}{1465},\ y = \frac{959}{4395},\ \ x = 331/8790 \tag{1.49}$$

Ahora, como

$$B = PA = \begin{pmatrix} 16 & 4 & -2 & 1 \\ 0 & 7 & -1 & 5 \\ 2 & -3 & 8 & 1 \\ 4 & 0 & 1 & -10 \end{pmatrix}$$

tenemos que la matriz B es diagonal dominante, por lo tanto el método de Jacobi converge.

Por otra parte, la matriz de Jacobi es dada por

$$\begin{aligned} J &= I - Q^{-1}B \\ &= I_4 - \begin{pmatrix} 1/16 & 0 & 0 & 0 \\ 0 & 1/7 & 0 & 0 \\ 0 & 0 & 1/8 & 0 \\ 0 & 0 & 0 & -1/10 \end{pmatrix} \begin{pmatrix} 16 & 4 & -2 & 1 \\ 0 & 7 & -1 & 5 \\ 2 & -3 & 8 & 1 \\ 4 & 0 & 1 & -10 \end{pmatrix} \\ &= \begin{pmatrix} 0 & -1/4 & 1/8 & -1/16 \\ 0 & 0 & 1/7 & -5/7 \\ -1/4 & 3/8 & 0 & -1/8 \\ 2/5 & 0 & 1/10 & 0 \end{pmatrix} \end{aligned}$$

Para determinar si el método de Jacobi converge determinaremos los valores propios de J. Tenemos

$$\det(J - \lambda I_4) = \det \begin{pmatrix} -\lambda & -1/4 & 1/8 & -1/16 \\ 0 & -\lambda & 1/7 & -5/7 \\ -1/4 & 3/8 & -\lambda & -1/8 \\ 2/5 & 0 & 1/10 & -\lambda \end{pmatrix} = \lambda^4 + \frac{17}{1120}\lambda^2 - \frac{219}{4480}\lambda + \frac{33}{2240} \tag{1.50}$$

de donde se tiene que los valores propios son

$$\begin{aligned} \lambda_{1,2} &= -0.2483754458 \pm 0.3442140503i \\ \lambda_{3,4} &= -0.2483754458 \pm 0.1416897424i \end{aligned} \tag{1.51}$$

Por lo tanto, se tiene que el radio espectral es $\rho = 0.4244686967 < 1$ y el método de Jacobi es convergente.

Realizando las iteraciones, obtenemos la siguinete tabla

Problema 1.7 Para cada $n \in \mathbb{N}$, se definen

$$A_n = \begin{pmatrix} 1 & 2 \\ 2 & 4 + \frac{1}{n^2} \end{pmatrix} \ \text{ y } \ b_n = \begin{pmatrix} 1 \\ 2 - \frac{1}{n^2} \end{pmatrix}.$$

Se desea resolver el sistema $A_n x = b_n$. Un estudiante obtiene como resultado $\tilde{x} = (1, 0)^{\mathrm{T}}$.

n	x_n	y_n	z_n	w_n	E
0					0.000041875
1	0.037658125	0.2182	0.205445	-0.064375	0.00001725
2	0.0376540625	0.218188571428	0.20545734375	-0.064392640625	0.000014084822
3	0.0376595407368	0.21820265625	0.20545622991	-0.064392640625	0.3961078E-5
4	0.0376559047154	0.218202776148	0.205460190988	-0.0643905607143	

1. Se define el vector residual asociado a esta solución como

$$r_n = A_n\tilde{x} - b_n$$

Calcule r_n ¿ Podemos decir que para n grande, la solución es razonablemente confiable? Justifique

2. Resolver $A_n x = b_n$ en forma exacta. Denote por $\bar{x}_n$ la solución exacta y cálcule el error relativo de la aproximación $\tilde{x} = (1,0)^{\mathrm{T}}$.

3. Lo razonablemente esperado es que $\bar{x}_n$ converja a $\tilde{x}$ cuando n tiende a infinito. Para este ejemplo, ¿ se tiene dicha afirmación?. Calcule $K_\infty(A_n)$ y explique el resultado obtenido.

Solución. Para cada $n \in \mathbb{N}$, un resultado para una solución de $A_n x = b_n$ es $\tilde{x} = (1,0)$.

1. El vector residual asociado a esta solución es

$$r_n = A_n\tilde{x} - b_n\,.$$

Tenemos

$$\begin{aligned} r_n &= \begin{pmatrix} 1 & 2 \\ 2 & 4+\frac{1}{n^2} \end{pmatrix}\begin{pmatrix} 1 \\ 0 \end{pmatrix} - \begin{pmatrix} 1 \\ 2-\frac{1}{n^2} \end{pmatrix} \\ &= \begin{pmatrix} 1 \\ 2 \end{pmatrix} - \begin{pmatrix} 1 \\ 2-\frac{1}{n^2} \end{pmatrix} = \begin{pmatrix} 0 \\ \frac{1}{n^2} \end{pmatrix} \end{aligned}$$

Ahora, la solución exacta del sistema es

$$\begin{pmatrix} 1 & 2 \\ 2 & 4+\frac{1}{n^2} \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 1 \\ 2-\frac{1}{n^2} \end{pmatrix}$$

esto del par de ecuaciones lineales

$$\begin{aligned} x + 2y &= 1 \\ 2x + \left(4+\frac{1}{n^2}\right)y &= 2 - \frac{1}{n^2}\,. \end{aligned}$$

De la primera ecuación obtenemos $x = 1-2y$, reemplazando en la segunda ecuación nos queda

$$2 - 4y + 4y + \frac{1}{n^2}y = 2 - \frac{1}{n^2}\,,$$

es decir, $\frac{1}{n^2}y = -\frac{1}{n^2}$, de donde $y = -1$, por lo tanto $x = 3$. Note que la solución es exactamente la misma para cada $n \in \mathbb{N}$, es decir, la solución no depende de $n \in \mathbb{N}$.

Por lo tanto la solución dada no es razonablemente confiable, pues no se aproxima en nada a la solución exacta $x_T = (3, -1)$ del sistema $A_n x = b_n$.

2. Ya calculamos la solución exacta $\overline{x}_n$ de $A_n x = b_n$, la cual es $\overline{x}_n = (3, -1) = x_T$ para todo $n \in \mathbb{N}$.

   Usando la norma $||\cdot||_\infty$ para simplificar los cálculos.

$$\begin{aligned} E_R(\tilde{x}) &= \frac{||x_T - \tilde{x}||_\infty}{||x_T||_\infty} \\ &= \frac{||(3,-1) - (1,0)||_\infty}{||(3,-1)||_\infty} \\ &= \frac{||(2,-1)||_\infty}{||(3,-1)||_\infty} \\ &= \frac{\max\{|2|, |-1|\}}{\max\{|3|, |-1|\}} \\ &= \frac{2}{3}\,. \end{aligned}$$

3. No, pues $x_n = (3, -1) = x_T$ es un vector constante y no se aproxima (obviamente) a $\tilde{x} = (1, 0)$.

   Para calcular $k_\infty(A_n)$, tenemos

$$||A_n||_\infty = \max\{|1| + |2|, |2| + \left|4 + \frac{1}{n^2}\right|\} = 6 + \frac{1}{n^2}$$

   Ahora, $\det(A_n) = 4 + \frac{1}{n^2} - 4 = \frac{1}{n^2}$. Luego

$$\begin{aligned} A_n^{-1} &= \frac{1}{\frac{1}{n^2}} \begin{pmatrix} 4 + \frac{1}{n^2} & -2 \\ -2 & 1 \end{pmatrix} \\ &= n^2 \begin{pmatrix} 4 + \frac{1}{n^2} & -2 \\ -2 & 1 \end{pmatrix} \\ &= \begin{pmatrix} 4n^2 + 1 & -2n^2 \\ -2n^2 & n^2 \end{pmatrix} \end{aligned}$$

y $||A_n^{-1}||_\infty = \max\{|14n^2+1| + |-2n^2| + |n^2|\} = \max\{6n^2+1, 3n^2\} = 6n^2+1$. Por lo tanto

$$\begin{aligned} k_\infty(A_n) &= \left(6 + \frac{1}{n^2}\right)(6n^2+1) \\ &= 36n^2 + 6 + 6 + \frac{1}{n^2} \\ &= 36n^2 + 12 + \frac{1}{n^2} \xrightarrow[n\to\infty]{} \infty, \end{aligned}$$

lo cual significa que la matriz A_n está muy mal condicionada para valores grandes de n (es numéricamente no invertible para n grande), lo cual explica la mala aproximación $\tilde{x}$ de x_n para n grande.

Problema 1.8 Dados

$$A = \begin{pmatrix} 2 & a & 0 \\ a & 2 & a \\ 0 & a & 2 \end{pmatrix} \qquad \text{y} \qquad b = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}$$

(a) ¿Bajo que condiciones del parámetro a, se puede asegurar que la matriz A tiene factorización de Cholesky?

(b) Obtener la factorización de Cholesky , es decir, encuentrarla matriz triangular superior U tal que $A = U^{\mathrm{T}}U$.

(c) Resuelver el sistema $Ax = b$ utilizando la factorización anterior.

Solución

(a) Para que A tenga factorización de Cholesky es necesario y suficiente que A sea simétrica y positiva definida. Ahora bien, A es claramente simétrica. Para encontrar las condiciones sobre el parámetro a, vamos a imponer que A sea positiva definida. Hay al menos 3 formas de hacer esto. Nosotros mostraremos aquí las dos formas más simples.

Primera forma: Calcular los valores propios de A e imponer que sean positivos.

$$\begin{aligned} \det(A - \lambda I) &= \det \begin{pmatrix} 2-\lambda & a & 0 \\ a & 2-\lambda & a \\ 0 & a & 2-\lambda \end{pmatrix} \\ &= (2-\lambda)\left[(2-\lambda)^2 - a^2\right] - a\left[a(2-\lambda) - 0\right] \\ &= (2-\lambda)\left[(2-\lambda)^2 - 2a^2\right], \end{aligned}$$

luego los valores propios de A son $\lambda_1 = 2$ y las soluciones de $|2-\lambda_{2,3}| = \sqrt{2}|a|$, esto es, $2 - \lambda_{2,3} = \pm\sqrt{2}|a|$, lo que nos da $\lambda_{2,3} = 2 \pm \sqrt{2}|a|$.

Nota: $\sqrt{a^2} = |a|$, ya que no sabemos si a es positivo, negativo o cero.

En consecuencia, A es positiva definida si y solamente si $\lambda_1 > 0$, $\lambda_{2,3} > 0$. Como $\lambda_1 = 2 > 0$, basta imponer que $\lambda_{2,3} > 0$, es decir, $2 - \sqrt{2}|a| > 0$ y $2 + \sqrt{2}|a| > 0$.

Ahora, como $2+\sqrt{2}|a|$ es siempre positivo, lo anterior será cierto si y sólo si $2-\sqrt{2}|a|>0$, es decir, $|a|<\sqrt{2}$.

Segunda forma: Calcular los subdeterminantes principales de A y verificar (o imponer) que sean positivos. Tenemos, $\det(A_1)=\det 2=2>0$, $\det(A_2)=\det\begin{pmatrix}2&a\\a&2\end{pmatrix}=4-a^2>0$ si y sólo si $|a|<2$, finalmente, $\det(A_3)=\det(A)=2\cdot\left[4-a^2\right]-a\left[2a-0\right]=8-2a^2-2a^2=8-4a^2>0$ si y sólo si $|a|<\sqrt{2}$, en otras palabras, para satisfacer ambas condiciones al mismo tiempo, debemos pedir que $|a|<\min(2,\sqrt{2})=\sqrt{2}$, es decir, $|a|<\sqrt{2}$.

(b) Escribiendo la igualdad $U^TU=A$, con L una matriz triangular superior de 3×3, obtenemos que

$$s_{11} = \sqrt{a_{11}}=\sqrt{2},$$

$$(j=1)\qquad\begin{cases}(k=2) & s_{12}=\frac{1}{s_{11}}a_{12}=\frac{\sqrt{2}}{2}a,\\ (k=3) & s_{13}=\frac{1}{s_{11}}a_{13}=0,\\ s_{22}=\sqrt{a_{22}-s_{12}^2}=\frac{\sqrt{2}}{2}\sqrt{4-a^2}; & \end{cases}$$

$$(j=2)\qquad\begin{cases}(k=3) & s_{23}=\frac{1}{s_{22}}(a_{23}-s_{12}\cdot s_{13})=\frac{\sqrt{2}a}{\sqrt{4-a^2}}\\ s_{33}=\sqrt{a_{33}-s_{13}^2-s_{23}^2}=\frac{2\sqrt{2-a^2}}{\sqrt{4-a^2}}; & \end{cases}$$

de donde

$$S=\begin{pmatrix}\sqrt{2} & \frac{\sqrt{2}}{2}a & 0\\ 0 & \frac{\sqrt{2}}{2}\sqrt{4-a^2} & \frac{\sqrt{2}a}{\sqrt{4-a^2}}\\ 0 & 0 & \frac{2\sqrt{2-a^2}}{\sqrt{4-a^2}}\end{pmatrix}.$$

(c) $Ax=b$ si y sólo si $U^TUx=b$. Haciendo $Ux=y$ tenemos $U^Ty=b$. Primero resolveremos este último sistema y enseguida resolvemos el sistema $Ux=y$. Tenemos entonces

$$\begin{pmatrix}\sqrt{2} & 0 & 0\\ \frac{\sqrt{2}}{2}a & \frac{\sqrt{2}}{2}\sqrt{4-a^2} & 0\\ 0 & \frac{\sqrt{2}a}{\sqrt{4-a^2}} & \frac{2\sqrt{2-a^2}}{\sqrt{4-a^2}}\end{pmatrix}\begin{pmatrix}y_1\\y_2\\y_3\end{pmatrix}=\begin{pmatrix}1\\2\\3\end{pmatrix}.$$

Resolviendo de arriba hacia abajo (sustitución hacia adelante) obtenemos:

$$y_1 = \frac{\sqrt{2}}{2}, \qquad \frac{\sqrt{2}}{2}ay_1 + \frac{\sqrt{2}}{2}\sqrt{4-a^2}y_2 = 2,$$

$$\Rightarrow \quad \frac{\sqrt{2}}{2}\sqrt{4-a^2}y_2 = 2 - \frac{a}{2} = \frac{4-a}{2},$$

$$\Rightarrow \quad y_2 = \frac{4-a}{\sqrt{2}\sqrt{4-a^2}} = \frac{\sqrt{2}(4-a)}{2\sqrt{4-a^2}},$$

$$\frac{\sqrt{2}a}{\sqrt{4-a^2}}y_2 + \frac{2\sqrt{2-a^2}}{\sqrt{4-a^2}}y_3 = 3,$$

$$\Rightarrow \quad \frac{2\sqrt{2-a^2}}{\sqrt{4-a^2}}y_3 = 3 - \frac{\sqrt{2}a}{\sqrt{4-a^2}} \cdot \frac{\sqrt{2}(4-a)}{2\sqrt{4-a^2}} = 3 - \frac{a(4-a)}{4-a^2}$$

$$\Rightarrow \quad y_3 = \frac{12-4a-2a^2}{2\sqrt{2-a^2}\sqrt{4-a^2}} = \frac{6-2a-a^2}{\sqrt{2-a^2}\sqrt{4-a^2}}.$$

Resolvamos ahora el sistema $Ux = y$

$$\begin{pmatrix} \sqrt{2} & \frac{\sqrt{2}}{2}a & 0 \\ 0 & \frac{\sqrt{2}}{2}\sqrt{4-a^2} & \frac{\sqrt{2}a}{\sqrt{4-a^2}} \\ 0 & 0 & \frac{2\sqrt{2-a^2}}{\sqrt{4-a^2}} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} \frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}(4-a)}{2\sqrt{4-a^2}} \\ \frac{6-2a-a^2}{\sqrt{2-a^2}\sqrt{4-a^2}} \end{pmatrix}.$$

Resolviendo de abajo hacia arriba (sustitución en reversa) obtenemos

$$\frac{2\sqrt{2-a^2}}{\sqrt{4-a^2}}x_3 = \frac{6-2a-a^2}{\sqrt{2-a^2}\sqrt{4-a^2}} \Rightarrow \boxed{x_3 = \frac{6-2a-a^2}{2(2-a^2)}}$$

$$\frac{\sqrt{2}}{2}\sqrt{4-a^2}x_2 + \frac{\sqrt{2}a}{\sqrt{4-a^2}}x_3 = \frac{\sqrt{2}(4-a)}{2\sqrt{4-a^2}}$$

$$\Rightarrow (4-a^2)x_2 + 2ax_3 = (4-a) \Rightarrow (4-a^2)x_2 = 4 - a - \frac{a(6-2a-a^2)}{2-a^2}$$

$$\Rightarrow \boxed{x_2 = \frac{2a^3-2a^2-8a+8}{(2-a^2)(4-a^2)}}$$

$$\sqrt{2}x_1 + \frac{\sqrt{2}}{2}ax_2 = \frac{\sqrt{2}}{2} \Rightarrow 2x_1 = 1 - a\frac{2a^3-2a^2-8a+8}{(2-a^2)(4-a^2)} \Rightarrow \boxed{x_1 = \frac{-a^4+2a^3+2a^2-8a+8}{2(2-a^2)(4-a^2)}}$$

Problema 1.9 Para cada $h > 0$, se definen

$$A_h = \begin{pmatrix} 1 & 2 \\ 2 & 4+h \end{pmatrix} \text{ y } b_h = \begin{pmatrix} 1 \\ 2-h \end{pmatrix}.$$

Se desea resolver el sistema $A_h x = b_h$. Un estudiante obtiene como resultado $\tilde{x} = (1,0)^{\mathrm{T}}$.

(a) Se define el vector residual asociado a esta solución como

$$r_h = A_h\tilde{x} - b_h$$

Calcule r_h . ¿ Podemos decir que para h suficientemente pequeño, ls solución es razonablemente confiable? Justifique.

(b) Resolver $A_h x = b_h$ en forma exacta. Denote por $\bar{x}_h$ la solución exacta y calcule el error relativo de la aproximación $\tilde{x} = (1,0)^{\mathrm{T}}$.

(c) Lo razonablemente esperado es que $\bar{x}_h$ converja a $\tilde{x}$ cuando h tiende a cero. Para este ejemplo, ¿ se tiene dicha afirmación?. Calcule $k_\infty(A_h)$ y explique el resultado obtenido.

Solución.

(a) El vector residual asociado a la solución aproximada $\tilde{x}$ es:

$$r_h = A_h\tilde{x} - b_h\,.$$

Tenemos

$$\begin{aligned} r_h &= \begin{pmatrix} 1 & 2 \\ 2 & 4+h \end{pmatrix}\begin{pmatrix} 1 \\ 0 \end{pmatrix} - \begin{pmatrix} 1 \\ 2-h \end{pmatrix} \\ &= \begin{pmatrix} 1 \\ 2 \end{pmatrix} - \begin{pmatrix} 1 \\ 2-h \end{pmatrix} = \begin{pmatrix} 0 \\ h \end{pmatrix} \end{aligned}$$

De modo que $r_h \to (0,0)$ cuando $h \to 0$. Por lo tanto, la solución es razonablemente confiable para valores de h pequeos, ya que en ese caso $A_h\tilde{x} \approx b_h$.

(b) Ahora, la solución exacta del sistema satisface:

$$\begin{pmatrix} 1 & 2 \\ 2 & 4+h \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 1 \\ 2-h \end{pmatrix},$$

lo cual dá el par de ecuaciones lineales

$$\begin{aligned} x + 2y &= 1 \\ 2x + (4+h)\,y &= 2-h\,. \end{aligned}$$

De la primera ecuación obtenemos $x = 1 - 2y$. Reemplazando en la segunda ecuación nos queda

$$2 - 4y + 4y + hy = 2 - h\,,$$

es decir, $hy = -h$, de donde $y = -1$, y por lo tanto $x = 3$. Note que la solución es exactamente la misma para cada $h > 0$, es decir, la solución no depende de $h > 0$.

Con esto, la solución exacta $\overline{x}_h$ del sistema $A_h x = b_h$ es $\overline{x}_h = (3,-1)$ para todo $h > 0$.

Usando la norma $||\cdot||_\infty$ para simplificar los cálculos tenemos:

$$\begin{aligned} E_R(\tilde{x}) &= \frac{||\bar{x}_h - \tilde{x}||_\infty}{||\bar{x}_h||_\infty} \\ &= \frac{||(3,-1)-(1,0)||_\infty}{||(3,-1)||_\infty} \\ &= \frac{||(2,-1)||_\infty}{||(3,-1)||_\infty} \\ &= \frac{\max\{|2|,|-1|\}}{\max\{|3|,|-1|\}} \\ &= \frac{2}{3}. \end{aligned}$$

(c) No, pues $\bar{x}_h = (3,-1)$ es un vector constante y no se aproxima (obviamente) a $\tilde{x} = (1,0)$.

Para calcular $k_\infty(A_h)$, calculamos primero la norma de A_h :

$$||A_h||_\infty = \max\{|1|+|2|, |2|+\left|4+h\right|\} = 6+h.$$

Ahora, $\det(A_h) = 4+h-4 = h$. Luego

$$\begin{aligned} A_h^{-1} &= \frac{1}{h}\begin{pmatrix} 4+h & -2 \\ -2 & 1 \end{pmatrix} \\ &= \begin{pmatrix} \frac{4}{h}+1 & \frac{-2}{h} \\ \frac{-2}{h} & \frac{1}{h} \end{pmatrix} \end{aligned}$$

y $||A_h^{-1}||_\infty = \max\{\left|\frac{4}{h}+1\right| + \left|\frac{-2}{h}\right|, \left|\frac{-2}{h}\right| + \left|\frac{1}{h}\right|\} = \max\{\frac{6}{h}+1, \frac{3}{h}\} = \frac{6}{h}+1$. Por lo tanto

$$\begin{aligned} k_\infty(A_h) &= (6+h)\left(\tfrac{6}{h}+1\right) \\ &= \tfrac{36}{h}+6+6+h \\ &= \tfrac{36}{h}+12+h \longrightarrow \infty, \qquad \text{cuando } h \to 0 \end{aligned}$$

lo cual significa que la matriz A_h está muy mal condicionada para valores pequeños de h , en otras palabras A_h es numéricamente no invertible para h pequeño, aunque evidentemente si lo es en aritmética exacta, lo cual explica la mala aproximación $\tilde{x}$ de $\bar{x}_h$ para h pequeño.

Problema 1.10 Consideremos el sistema de ecuaciones lineales $Ax = b$, donde

$$A = \begin{pmatrix} 4 & 1 & 0 & 0 & 0 \\ 1 & \frac{5}{4} & 1 & 0 & 0 \\ 0 & 1 & 5 & 1 & 0 \\ 0 & 0 & 1 & \frac{5}{4} & 1 \\ 0 & 0 & 0 & 1 & 5 \end{pmatrix} \qquad \text{y} \qquad b = \begin{pmatrix} 4 \\ 2 \\ 5 \\ 2 \\ 5 \end{pmatrix}$$

(a) Calcule el número de condicionamiento (en norma infinito) de la matriz A y diga si el sistema es estable o no. Justifique.

(b) Un estudiante propone como solución de este sistema , el vector

$$\hat{x} = (1, 0.01, 0.99, 0.01, 1)^{\mathrm{T}}$$

Determine una estimación del error relativo de la aproximación dada por el estudiante (haga esto **SIN** resolver el sistema.)

(c) Encuentre la Factorización de Cholesky de la matriz A

(d) Resuelva el sistema lineal usando la factorización encontrada en la parte (c) y compruebe que la estimación dada por usted en la parte (b) es correcta.

(e) Dé una estimación de en cuanto variaría la solución del sistema $Ax = b$, si la matriz A es perturbada, de manera que la nueva matriz es

$$\hat{A} = \begin{pmatrix} 4 & 1 & 0 & 0 & 0.04 \\ 1 & \frac{5}{4} & 1 & 0 & 0.01 \\ 0 & 1 & 5 & 1 & 0 \\ -0.01 & 0 & 1 & \frac{5}{4} & 1 \\ -0.05 & 0 & 0 & 1 & 5 \end{pmatrix}$$

para lo cual, suponga que $\hat{A} = A(I + E)$, compruebe que se cumple la hipó tesis

$$\|AE\| < \frac{1}{\|A^{-1}\|}$$

y luego concluya.

Solución.

(a) El número de condicionamiento se calcula mediante la fórmula

$$k_\infty(A) = \|A\|_\infty \left\|A^{-1}\right\|_\infty$$

donde

$$\|A\|_\infty = \max_{1 \leqslant i \leqslant n} \left\{ \sum_{j=1}^{n} |a_{ij}| \right\}$$

y por tanto

$$\begin{aligned} \|A\|_\infty &= \max\left\{5, \frac{13}{4}, 7, \frac{13}{4}, 6\right\} = 7 \\ \left\|A^{-1}\right\|_\infty &= \max\left\{ \begin{array}{c} \frac{341+340+84+80+16}{1024}, \frac{85+340+84+80+16}{256}, \\ \frac{21+84+84+80+16}{256}, \frac{5+20+20+80+16}{64}, \frac{1+4+4+16+16}{64} \end{array} \right\} \\ &= \max\left\{\frac{861}{1024}, \frac{605}{256}, \frac{285}{256}, \frac{141}{64}, \frac{41}{64}\right\} \\ &= \frac{605}{256} \end{aligned}$$

Finalmente se tiene

$$k_\infty(A) = \|A\|_\infty \left\|A^{-1}\right\|_\infty \approx 16.54$$

y como el número de condicionamiento es pequeño, podemos asegurar que el sistema es estable.

(b) Calculemos primero, de que vector $\hat{b}$, es $\hat{x}$ solución, para esto basta multiplicar $A\hat{x}$, de donde se obtiene

$$\begin{pmatrix} 4 & 1 & 0 & 0 & 0 \\ 1 & \frac{5}{4} & 1 & 0 & 0 \\ 0 & 1 & 5 & 1 & 0 \\ 0 & 0 & 1 & \frac{5}{4} & 1 \\ 0 & 0 & 0 & 1 & 5 \end{pmatrix} \begin{pmatrix} 1 \\ 0.01 \\ 0.99 \\ 0.01 \\ 1 \end{pmatrix} = \begin{pmatrix} 4.01 \\ 2.0025 \\ 4.97 \\ 2.0025 \\ 5.01 \end{pmatrix}$$

es decir, $\hat{b} = (4.01, 2.0025, 4.97, 2.0025, 5.01)^{\mathrm{T}}$ y por tanto,aplicando la fórmula

$$E_R(\hat{x}) = \frac{\|\bar{x} - \hat{x}\|_\infty}{\|\bar{x}\|_\infty} \leqslant k_\infty(A) \frac{\left\|b - \hat{b}\right\|_\infty}{\|b\|_\infty}$$

pero $\left\|b - \hat{b}\right\|_\infty = \left\|(-0.01, -0.0025, 0.03, -0.0025, -0.01)^{\mathrm{T}}\right\|_\infty = 0.03$ y $\|b\|_\infty = \left\|(4, 2, 5, 2, 5)^{\mathrm{T}}\right\|_\infty = 5$. Sustituyendo en la expresión de arriba tenemos

$$\begin{aligned} E_R(\hat{x}) &\leqslant k_\infty(A) \frac{\left\|b - \hat{b}\right\|_\infty}{\|b\|_\infty} \\ &\leqslant 16.54 \cdot \frac{0.03}{5} \approx 0.09924 \end{aligned}$$

(c) Dado que la matriz es tridiagonal, se comprueba (visto en clases) que el algoritmo se reduce a

$$\begin{aligned} &s_{11} = (a_{11})^{1/2} \\ &\text{Para } j = 1, 2, \cdots, n-1 \\ &\qquad s_{j,j+1} = \frac{1}{s_{jj}} \\ &s_{j+1,j+1} = \left(a_{j+1,j+1} - s_{i,j+1}^2\right)^{1/2} \\ &\text{end} \end{aligned}$$

y por tanto,

$$U = \begin{pmatrix} 2 & \frac{1}{2} & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 2 & \frac{1}{2} & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 2 \end{pmatrix}$$

(d) Sustituyamos en el sistema, la matriz A por su factorización

$$\begin{aligned} Ax &= b \\ U^{\mathrm{T}}Ux &= b \end{aligned}$$

y resolviendo primeramente $U^{\mathrm{T}}y = b$ y después $Ux = y$ se obtiene la solución

$$\begin{pmatrix} 4 & 1 & 0 & 0 & 0 \\ 1 & \frac{5}{4} & 1 & 0 & 0 \\ 0 & 1 & 5 & 1 & 0 \\ 0 & 0 & 1 & \frac{5}{4} & 1 \\ 0 & 0 & 0 & 1 & 5 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{pmatrix} = \begin{pmatrix} 4 \\ 2 \\ 5 \\ 2 \\ 5 \end{pmatrix}$$

de donde

$$\begin{pmatrix} 2 & 0 & 0 & 0 & 0 \\ \frac{1}{2} & 1 & 0 & 0 & 0 \\ 0 & 1 & 2 & 0 & 0 \\ 0 & 0 & \frac{1}{2} & 1 & 0 \\ 0 & 0 & 0 & 1 & 2 \end{pmatrix} \begin{pmatrix} 2 & \frac{1}{2} & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 2 & \frac{1}{2} & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{pmatrix} = \begin{pmatrix} 4 \\ 2 \\ 5 \\ 2 \\ 5 \end{pmatrix}$$

y por tanto

$$\begin{pmatrix} 2 & 0 & 0 & 0 & 0 \\ \frac{1}{2} & 1 & 0 & 0 & 0 \\ 0 & 1 & 2 & 0 & 0 \\ 0 & 0 & \frac{1}{2} & 1 & 0 \\ 0 & 0 & 0 & 1 & 2 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \end{pmatrix} = \begin{pmatrix} 4 \\ 2 \\ 5 \\ 2 \\ 5 \end{pmatrix} \quad \Rightarrow \quad \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \end{pmatrix} = \begin{pmatrix} 2 \\ 1 \\ 2 \\ 1 \\ 2 \end{pmatrix}$$

y

$$\begin{pmatrix} 2 & \frac{1}{2} & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 2 & \frac{1}{2} & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{pmatrix} = \begin{pmatrix} 2 \\ 1 \\ 2 \\ 1 \\ 2 \end{pmatrix} \quad \Rightarrow \quad \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 1 \\ 0 \\ 1 \end{pmatrix}$$

es decir, la solución del sistema es $\bar{x} = (1, 0, 1, 0, 1)^{\mathrm{T}}$. Calculemos ahora el error relativo que cometió el estudiante

$$\begin{aligned} E_R(\hat{x}) &= \frac{\|\bar{x} - \hat{x}\|_\infty}{\|\bar{x}\|_\infty} \\ &= \frac{\left\|(0, -0.01, 0.01, -0.01, 0)^{\mathrm{T}}\right\|_\infty}{\left\|(1, 0, 1, 0, 1)^{\mathrm{T}}\right\|_\infty} \\ &= 0.01 \end{aligned}$$

y evidentemente, $0.01 < 0.09924$ por lo que la estimación hecha era correcta.

(e) Como $\hat{A} = A(I + E)$ entonces $\hat{A} - A = AE$ y por tanto,

$$\begin{aligned} \|AE\|_\infty &= \left\|\hat{A} - A\right\|_\infty \\ &= \left\| \begin{pmatrix} 0 & 0 & 0 & 0 & 0.04 \\ 0 & 0 & 0 & 0 & 0.01 \\ 0 & 0 & 0 & 0 & 0 \\ -0.01 & 0 & 0 & 0 & 0 \\ -0.05 & 0 & 0 & 0 & 0 \end{pmatrix} \right\|_\infty \\ &= 0.05 \end{aligned}$$

Pro otro lado, $\left\|A^{-1}\right\|_\infty = \frac{605}{256}$ (ver parte (b)), y por tanto, la hipótesis

$$\|AE\|_\infty = 0.05 < \frac{1}{\|A^{-1}\|_\infty} \approx 0.423$$

se cumple. Entonces podemos aplicar la fórmula

$$\begin{aligned} E_R(\hat{x}) &\leqslant \frac{k_\infty(A)}{1-k_\infty(A)\frac{\left\|\hat{A}-A\right\|_\infty}{\|A\|_\infty}} \frac{\left\|\hat{A}-A\right\|_\infty}{\|A\|_\infty} \\ &\leqslant \frac{7\frac{605}{256}}{1-7\frac{605}{256}\frac{0.05}{7}} \frac{0.05}{7} \\ &\leqslant \frac{30.25}{225.75} \approx 0.1339 \end{aligned}$$

Algoritmo General de la Factorización de Cholesky:

Sea $A=(a_{ij}) \in M_n$ y denotemos por $S=(s_{ij})$ la matriz de la factorización, entonces el algoritmo es

$$\begin{array}{l} s_{11}=(a_{11})^{1/2} \\ \text{Para } j=1,2,\cdots,n-1 \\ \quad \text{Para } k=j+1,\cdots,n \\ \quad\quad s_{jk}=\dfrac{1}{s_{jj}}\left(a_{jk}-\displaystyle\sum_{i=1}^{j-1} s_{ij}s_{ik}\right) \\ \quad \text{end} \\ s_{j+1,j+1}=\left(a_{j+1,j+1}-\displaystyle\sum_{i=1}^{j} s_{i,j+1}^2\right)^{1/2} \\ \text{end} \end{array}$$

Tenga presente que la inversa de la matriz A es

$$A^{-1}=\begin{pmatrix} \frac{341}{1024} & -\frac{85}{256} & \frac{21}{256} & -\frac{5}{64} & \frac{1}{64} \\ -\frac{85}{256} & \frac{85}{64} & -\frac{21}{64} & \frac{5}{16} & -\frac{1}{16} \\ \frac{21}{256} & -\frac{21}{64} & \frac{21}{64} & -\frac{5}{16} & \frac{1}{16} \\ -\frac{5}{64} & \frac{5}{16} & -\frac{5}{16} & \frac{5}{4} & -\frac{1}{4} \\ \frac{1}{64} & -\frac{1}{16} & \frac{1}{16} & -\frac{1}{4} & \frac{1}{4} \end{pmatrix}$$

Problema 1.11 Considere el siguiente sistema de ecuaciones lineales

$$\begin{pmatrix} 1 & \alpha-1 & 0 & 0 \\ -\alpha & 1 & \alpha-1 & 0 \\ 0 & -\alpha & 1 & \alpha-1 \\ 0 & 0 & -\alpha & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} \alpha \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

con $\alpha \in [0,1]$.

(a) Escriba el método de Jacobi para resolver el sistema anterior y diga para que valores del parámetro α el método iterativo es convergente?

(b) Tomando como punto inicial, el vector $\mathbf{x}^0 = (x_1^0, x_2^0, x_3^0, x_4^0)^{\mathrm{T}} = (0,0,0,0)^{\mathrm{T}}$. Encuentre los valores de las iteraciones $\mathbf{x}^1$ y $\mathbf{x}^2$. Determine la cantidad de iteraciones (para $\alpha = 1/2$) que necesitaría el método de Jacobi para obtener una precisión de 10^{-6}.
Sugerencia: Recuerde que el radio espectral de una matriz simétrica es igual a la norma matricial subordinada 2.

(c) Encuentre el valor del parámetro α que hace que el método de Jacobi converja lo más rápido posible. Denotemos éste valor por α^* y diga para que valores del parámetro α, el método de Gauss-Seidel converge **MAS LENTO** que el método de Jacobi con el parámetro α^*. Justifique su respuesta.

Solución

(a) Dada la matriz del sistema lineal

$$A = \begin{pmatrix} 1 & \alpha-1 & 0 & 0 \\ -\alpha & 1 & \alpha-1 & 0 \\ 0 & -\alpha & 1 & \alpha-1 \\ 0 & 0 & -\alpha & 1 \end{pmatrix},$$

entonces, la matriz de Jacobi viene dada por

$$M_J = -D^{-1}(E+F) = \begin{pmatrix} 0 & 1-\alpha & 0 & 0 \\ \alpha & 0 & 1-\alpha & 0 \\ 0 & \alpha & 0 & 1-\alpha \\ 0 & 0 & \alpha & 0 \end{pmatrix}$$

y por tanto, debemos buscar los valores del parámetro α que garanticen que el radio espectral de esta matriz M_J sea menor que 1. Para esto se calcula el polinomio característico de la matriz como

$$\begin{aligned} p(\lambda) &= \det(M_J - \lambda I) = \det \begin{pmatrix} -\lambda & 1-\alpha & 0 & 0 \\ \alpha & -\lambda & 1-\alpha & 0 \\ 0 & \alpha & -\lambda & 1-\alpha \\ 0 & 0 & \alpha & -\lambda \end{pmatrix} \\ &= -\lambda \det \begin{pmatrix} -\lambda & 1-\alpha & 0 \\ \alpha & -\lambda & 1-\alpha \\ 0 & \alpha & -\lambda \end{pmatrix} - \alpha \begin{pmatrix} 1-\alpha & 0 & 0 \\ \alpha & -\lambda & 1-\alpha \\ 0 & \alpha & -\lambda \end{pmatrix} \\ &= -\lambda\left[-\lambda^3 + 2\alpha(1-\alpha)\lambda\right] - \alpha\left[(1-\alpha)\lambda^2 - \alpha(1-\alpha)^2\right] \\ &= \lambda^4 - 3\alpha(1-\alpha)\lambda^2 + \alpha^2(1-\alpha)^2 \end{aligned}$$

y haciendo el cambio de variable $\beta = \lambda^2$ y $a = \alpha(1-\alpha)$, tenemos que las raices del polinomio $\beta^2 - 3a\beta + a^2$ vienen dadas por la expresión

$$\beta_{12} = \frac{3 \pm \sqrt{5}}{2} a$$

y por tanto, el conjunto de valores propios de la matriz de Jacobi es

$$\sigma(M_J) = \left\{ \pm\sqrt{\frac{3+\sqrt{5}}{2}\alpha(1-\alpha)}, \pm\sqrt{\frac{3-\sqrt{5}}{2}\alpha(1-\alpha)} \right\}$$

y finalmente, el radio espectral es igual a

$$\begin{aligned}\rho(M_J) &= \max_{\lambda\in\sigma(M_J)}|\lambda| \\ &= \sqrt{\frac{3+\sqrt{5}}{2}\alpha(1-\alpha)}\end{aligned}$$

Para garantizar convergencia, se necesita que el radio espectral sea menor que uno, por tanto

$$\begin{aligned}\sqrt{\frac{3+\sqrt{5}}{2}\alpha(1-\alpha)} &< 1 \\ \frac{3+\sqrt{5}}{2}\alpha(1-\alpha) &< 1 \\ \alpha^2-\alpha+\frac{2}{3+\sqrt{5}} &> 0\end{aligned}$$

pero, este último polinomio es positivo para todo valor real del parámetro α (basta comprobar que ambas raices son complejas y además la parábola abre para arriba), por tanto

$$\rho(M_J)<1 \qquad \forall\alpha\in[0,1]$$

es decir, el método de Jacobi converge para todo $\alpha\in[0,1]$.

(b) El método de Jacobi se escribe como

$$\mathbf{x}^{k+1}=M_J\mathbf{x}^k+v_J$$

donde M_J representa la matriz de Jacobi encontrada en el punto anterior y $v_J=D^{-1}b=b$ (note que la matriz diagonal D de la matriz A es la identidad), es decir, $v_J=(\alpha,0,0,0)^{\mathrm{T}}$. Si $\mathbf{x}^0=(0,0,0,0)^{\mathrm{T}}$, entonces evidentemente se tiene que $\mathbf{x}^1=v_J=(\alpha,0,0,0)^{\mathrm{T}}$ y

$$\begin{aligned}\mathbf{x}^2 &= M_J\mathbf{x}^1+v_J \\ &= \begin{pmatrix}0 & 1-\alpha & 0 & 0\\ \alpha & 0 & 1-\alpha & 0\\ 0 & \alpha & 0 & 1-\alpha\\ 0 & 0 & \alpha & 0\end{pmatrix}\begin{pmatrix}\alpha\\0\\0\\0\end{pmatrix}+\begin{pmatrix}\alpha\\0\\0\\0\end{pmatrix} \\ &= \begin{pmatrix}\alpha^2\\ \alpha\\ 0\\ 0\end{pmatrix}\end{aligned}$$

Finalmente, considerando $\alpha=1/2$, se ve claramente que la matriz de Jacobi resultante, es simétrica

$$M_J=\begin{pmatrix}0 & \frac{1}{2} & 0 & 0\\ \frac{1}{2} & 0 & \frac{1}{2} & 0\\ 0 & \frac{1}{2} & 0 & \frac{1}{2}\\ 0 & 0 & \frac{1}{2} & 0\end{pmatrix}$$

y dada la sugerencia, se tiene que

$$\begin{aligned}\|M_J\|_2 &= \rho(M_J)\\ &= \sqrt{\frac{3+\sqrt{5}}{2}\frac{1}{2}(1-\frac{1}{2})}\\ &= \sqrt{\frac{3+\sqrt{5}}{8}}\end{aligned}$$

Por otro lado, se tiene

$$\left\|\mathbf{x}^k-\mathbf{x}^*\right\|_2 \leqslant \frac{\|M_J\|_2^k}{1-\|M_J\|_2}\left\|\mathbf{x}^1-\mathbf{x}^0\right\| \leqslant 10^{-6}$$

y por tanto

$$\begin{aligned}\frac{\left(\frac{3+\sqrt{5}}{8}\right)^{k/2}}{1-\sqrt{\frac{3+\sqrt{5}}{8}}}\left\|(\frac{1}{2},0,0,0)^{\mathrm{T}}\right\|_2 &\leqslant 10^{-6}\\ \frac{1}{0.191}\left(\frac{3+\sqrt{5}}{8}\right)^{k/2}\frac{1}{2} &\leqslant 10^{-6}\\ \left(\frac{3+\sqrt{5}}{8}\right)^{k/2} &\leqslant 0.382\times 10^{-6}\\ \frac{k}{2}\ln\left(\frac{3+\sqrt{5}}{8}\right) &\leqslant \ln(0.382\times 10^{-6})\\ \frac{k}{2} &\geqslant 34.8640\\ k &\geqslant 69.7281\end{aligned}$$

es decir, se necesitarían al menos, 70 iteraciones.

(c) Primero, encontremos el valor de $\alpha \in [0,1]$ para el cual, el radio espectral de la matriz de Jacobi se hace lo más pequeño posible, pero como

$$\rho(M_J) = \sqrt{\frac{3+\sqrt{5}}{2}\alpha(1-\alpha)}$$

entonces, de manera evidente se ve que $\alpha^* = 0$ y $\alpha^* = 1$. Para ambos valores, el radio espectral se hace cero. Por otro lado, para calcular el radio espectral del método de Gauss-Seidel, basta ver que la matriz A es tridiagonal y además, esta se escribe como $A = I - M_J$ y por tanto, si λ es un valor propio de M_J entonces $1-\lambda$ es valor propio de A, y como todos los valores propios de M_J son menores, en valor absoluto, que uno, entonces todos los valores propios de A son positivos y por tanto, la matriz A es definida positiva. Entonces podemos aplicar el teorema visto en clases, de donde

$$\rho(M_{GS}) = \rho(M_J)^2 = \frac{3+\sqrt{5}}{2}\alpha(1-\alpha)$$

lo cual, evidentemente es mayor estricto que cero, y por tanto podemos concluir que para todo $\alpha \in (0,1)$, el radio espectral de Gauss-Seidel es mayor que el radio espectral de Jacobi, para el caso de que $\alpha^* = 0$ o $\alpha^* = 1$, y por tanto, la convergencia es más lenta.

Problema 1.12 Dada la siguiente matriz tridiagonal parámetrica

$$A = \begin{pmatrix} \alpha & \beta & 0 \\ \beta & \alpha & \beta \\ 0 & \beta & \alpha \end{pmatrix}$$

con $\alpha > 0$ y $\beta > 0$.

(a) Demuestre que si $\alpha > \sqrt{2}\beta$ entonces la matriz A es definida positiva.

(b) Encuentre una relación necesaria y suficiente que deban cumplir los parámetros α y β de modo que

(i) El método de Jacobi sea convergente.

(ii) El método de Gauss-Seidel sea convergente.

(c) Considere los valores de los parámetros $\alpha = 4$ y $\beta = 1$ y compare los métodos de Jacobi, Gauss-Seidel y SOR (con el parámetro ω optimo) en cuanto a velocidad de convergencia. Justifique su respuesta (debe calcular los radios espectrales de cada uno de los tres métodos).

(d) Considere el vector $b = (1, 1, 1)^{\mathrm{T}}$ y los valores de los parámetros $\alpha = 4$ y $\beta = 1$ y determine cuantas iteraciones necesitaría el método de Gauss-Seidel para obtener una precisión de $\varepsilon = 10^{-4}$ si partimos desde el punto inicial $\mathbf{x}^0 = (0, 0, 0)^{\mathrm{T}}$.

Solución

(a) Cálculemos los determinantes de los menores principales (denotémos estos por $|A_i|$ con $i = 1, 2, 3$) de la matriz A

$$\begin{aligned} |A_1| &= \alpha > 0 \\ |A_2| &= \det \begin{pmatrix} \alpha & \beta \\ \beta & \alpha \end{pmatrix} = \alpha^2 - \beta^2 > 0 \Rightarrow \alpha > \beta \\ |A_3| &= \det \begin{pmatrix} \alpha & \beta & 0 \\ \beta & \alpha & \beta \\ 0 & \beta & \alpha \end{pmatrix} = \alpha^3 - 2\alpha\beta^2 > 0 \Rightarrow \alpha(\alpha^2 - 2\beta^2) > 0 \\ &\Rightarrow \alpha > \sqrt{2}\beta \end{aligned}$$

y como $\alpha > \sqrt{2}\beta > \beta > 0$, entonces basta que se cumpla la relación $\alpha > \sqrt{2}\beta$ para que se tenga que los determinantes de los tres menores principales son estrictamente mayores que cero y por tanto, la matriz A es definida positiva.

(b) Cualquier método iterativo será convergente si y sólo si, el radio espectral de la matriz del método es menor estricto que uno. Por otro lado, notemos que la matriz A es simétrica, tridiagonal y si $\alpha > \sqrt{2}\beta$ entonces es además definida positiva, en este caso se tiene (aplicando el Teorema visto en clases) que

$$\rho(M_{GS}) = \rho(M_J)^2 < 1$$

y por tanto, ambos métodos serían convergentes. Veamos entonces, los casos (i) y (ii) en general.

(i) La matriz de Jacobi queda

$$M_J = -D^{-1}(E+F) = \begin{pmatrix} 0 & -\frac{\beta}{\alpha} & 0 \\ -\frac{\beta}{\alpha} & 0 & -\frac{\beta}{\alpha} \\ 0 & -\frac{\beta}{\alpha} & 0 \end{pmatrix}$$

y por tanto,

$$p(\lambda) = \det(M_J - \lambda I) = \det \begin{pmatrix} -\lambda & -\frac{\beta}{\alpha} & 0 \\ -\frac{\beta}{\alpha} & -\lambda & -\frac{\beta}{\alpha} \\ 0 & -\frac{\beta}{\alpha} & -\lambda \end{pmatrix} = -\lambda^3 + \lambda\frac{\beta^2}{\alpha^2}$$

de donde,

$$\begin{aligned} p(\lambda) &= 0 \\ -\lambda^3 + 2\lambda\frac{\beta^2}{\alpha^2} &= 0 \\ \lambda(2\frac{\beta^2}{\alpha^2} - \lambda^2) &= 0 \end{aligned}$$

y concluimos que el radio espectral de la matriz de Jacobi es igual a $\sqrt{2}\frac{\beta}{\alpha}$,

$$\rho(M_J) = \sqrt{2}\frac{\beta}{\alpha} < 1 \Longrightarrow \alpha > \sqrt{2}\beta$$

es decir, si $\alpha > \sqrt{2}\beta$ entonces el método de Jacobi converge.

(ii) La matriz de Gauss-Seidel queda

$$\begin{aligned} M_{GS} &= -(D+E)^{-1}F \\ &= -\begin{pmatrix} \alpha & 0 & 0 \\ \beta & \alpha & 0 \\ 0 & \beta & \alpha \end{pmatrix}^{-1} \begin{pmatrix} 0 & \beta & 0 \\ 0 & 0 & \beta \\ 0 & 0 & 0 \end{pmatrix} \\ &= \begin{pmatrix} -\frac{1}{\alpha} & 0 & 0 \\ \frac{1}{\alpha^2}\beta & -\frac{1}{\alpha} & 0 \\ -\frac{1}{\alpha^3}\beta^2 & \frac{1}{\alpha^2}\beta & -\frac{1}{\alpha} \end{pmatrix} \begin{pmatrix} 0 & \beta & 0 \\ 0 & 0 & \beta \\ 0 & 0 & 0 \end{pmatrix} \\ &= \begin{pmatrix} 0 & -\frac{\beta}{\alpha} & 0 \\ 0 & \frac{\beta^2}{\alpha^2} & -\frac{\beta}{\alpha} \\ 0 & -\frac{\beta^3}{\alpha^3} & \frac{\beta^2}{\alpha^2} \end{pmatrix} \end{aligned}$$

y por tanto,

$$\begin{aligned} p(\lambda) &= \det(M_{GS} - \lambda I) = \det \begin{pmatrix} -\lambda & -\frac{\beta}{\alpha} & 0 \\ 0 & \frac{\beta^2}{\alpha^2} - \lambda & -\frac{\beta}{\alpha} \\ 0 & -\frac{\beta^3}{\alpha^3} & \frac{\beta^2}{\alpha^2} - \lambda \end{pmatrix} \\ &= -\lambda\left(\frac{\beta^2}{\alpha^2} - \lambda\right)^2 + \lambda\frac{\beta^4}{\alpha^4} \end{aligned}$$

de donde,

$$\begin{aligned} p(\lambda) &= 0 \\ -\lambda\left(\frac{\beta^2}{\alpha^2}-\lambda\right)^2+\lambda\frac{\beta^4}{\alpha^4} &= 0 \\ -\lambda\left(\left(\frac{\beta^2}{\alpha^2}-\lambda\right)^2-\frac{\beta^4}{\alpha^4}\right) &= 0 \\ \lambda &= 0 \text{ y } \left(\frac{\beta^2}{\alpha^2}-\lambda\right)^2=\frac{\beta^4}{\alpha^4} \\ \frac{\beta^2}{\alpha^2}-\lambda &= -\frac{\beta^2}{\alpha^2} \quad \text{ó} \quad \frac{\beta^2}{\alpha^2}-\lambda=\frac{\beta^2}{\alpha^2} \end{aligned}$$

y concluimos que el radio espectral de la matriz de Gauss Seidel (el máximo valor propio) es igual a $2\frac{\beta^2}{\alpha^2}$,

$$\rho(M_{GS})=2\frac{\beta^2}{\alpha^2}<1 \Longrightarrow \alpha>\sqrt{2}\beta$$

es decir, si $\alpha>\sqrt{2}\beta$ entonces el método de Gauss Seidel converge.

Notemos que la relación en ambos acasos, es la misma que se obtuvo en el item (a), y que permitía asegurar que A es definida positiva.

(c) Si los valores de los parámetros son $\alpha=4$ y $\beta=1$ entonces se cumple la relación $\alpha>\sqrt{2}\beta$, y por tanto se tiene que

$$\rho(M_{GS})=\rho(M_J)^2<1$$

y además que

$$\begin{aligned} \omega^* &= \frac{2}{1+\sqrt{1-\rho(M_J)^2}}=\frac{2}{1+\sqrt{1-2\frac{\beta^2}{\alpha^2}}} \\ &= \frac{2}{1+\sqrt{1-\frac{1}{8}}}=1.0334 \end{aligned}$$

y por tanto, el radio espectral de la matriz de SOR con el parámetro ω^* es igual a

$$\rho(M_{SOR}(\omega^*))=\omega^*-1=0.0334$$

Por otro lado,

$$\begin{aligned} \rho(M_J) &= \sqrt{2}\frac{\beta}{\alpha}=\frac{\sqrt{2}}{4}=0.3536 \\ \rho(M_{GS}) &= 2\frac{\beta^2}{\alpha^2}=\frac{1}{8}=0.125 \end{aligned}$$

y comparando los tres radios espectrales, podemos concluir que el método SOR es el más rápido, luego viene el método de Gauss Seidel y finalmente, el método de Jacobi.

(d) Dado un método iterativo cualquiera, se tiene que

$$\left\|\mathbf{x}^k-\bar{\mathbf{x}}\right\| \leqslant \frac{\|M\|^k}{1-\|M\|}\left\|\mathbf{x}^1-\mathbf{x}^0\right\| \tag{1.52}$$

y, en nuestro caso

$$\begin{aligned}
\mathbf{x}^1 &= M_{GS}\mathbf{x}^0 + v_{GS} \\
&= -(D+E)^{-1}F\mathbf{x}^0 + (D+E)^{-1}b \\
&= \begin{pmatrix} 0 & -\frac{\beta}{\alpha} & 0 \\ 0 & \frac{\beta^2}{\alpha^2} & -\frac{\beta}{\alpha} \\ 0 & -\frac{\beta^3}{\alpha^3} & \frac{\beta^2}{\alpha^2} \end{pmatrix}\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} + \begin{pmatrix} \frac{1}{\alpha} & 0 & 0 \\ -\frac{1}{\alpha^2}\beta & \frac{1}{\alpha} & 0 \\ \frac{1}{\alpha^3}\beta^2 & -\frac{1}{\alpha^2}\beta & \frac{1}{\alpha} \end{pmatrix}\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \\
&= \begin{pmatrix} \frac{1}{\alpha} \\ \frac{1}{\alpha} - \frac{1}{\alpha^2}\beta \\ \frac{1}{\alpha} - \frac{1}{\alpha^2}\beta + \frac{1}{\alpha^3}\beta^2 \end{pmatrix}
\end{aligned}$$

y evaluando en los parámetros $\alpha = 4$ y $\beta = 1$obtenemos

$$\mathbf{x}^1 = \begin{pmatrix} \frac{1}{4} & \frac{3}{16} & \frac{13}{64} \end{pmatrix}^{\mathrm{T}}$$

y por tanto,

$$\begin{aligned}
\left\|\mathbf{x}^1 - \mathbf{x}^0\right\|_1 &= \left\|\mathbf{x}^1\right\|_1 \\
&= \left\|\begin{pmatrix} \frac{1}{4} & \frac{3}{16} & \frac{13}{64} \end{pmatrix}^{\mathrm{T}}\right\|_1 \\
&= \frac{41}{64}
\end{aligned}$$

Por otro lado,

$$M_{GS} = \begin{pmatrix} 0 & -\frac{\beta}{\alpha} & 0 \\ 0 & \frac{\beta^2}{\alpha^2} & -\frac{\beta}{\alpha} \\ 0 & -\frac{\beta^3}{\alpha^3} & \frac{\beta^2}{\alpha^2} \end{pmatrix} = \begin{pmatrix} 0 & -\frac{1}{4} & 0 \\ 0 & \frac{1}{16} & -\frac{1}{4} \\ 0 & -\frac{1}{64} & \frac{1}{16} \end{pmatrix}$$

y por tanto,

$$\begin{aligned}
\left\|M_{GS}\right\|_1 &= \left\|\begin{pmatrix} 0 & -\frac{1}{4} & 0 \\ 0 & \frac{1}{16} & -\frac{1}{4} \\ 0 & -\frac{1}{64} & \frac{1}{16} \end{pmatrix}\right\|_1 \\
&= \max\left\{0, \frac{21}{64}, \frac{5}{16}\right\} = \frac{21}{64}
\end{aligned}$$

Finalmente, aplicando la fórmula (1.52), se tiene que

$$\left\|\mathbf{x}^k - \bar{\mathbf{x}}\right\|_1 \leqslant \frac{\left(\frac{21}{64}\right)^k}{1 - \frac{21}{64}}\frac{41}{64} < 10^{-4}$$

de donde $\left(\frac{21}{64}\right)^k < \frac{43}{41}10^{-4}$, y de aquí $4 + k\left(\log_{10} 21 - \log_{10} 64\right) < \log_{10} 43 - \log_{10} 41$, de donde $k > 8.2224$, y por tanto, concluimos que se necesitarán, al menos, 9 iteraciones.

Problema 1.13 Para resolver el sistema de ecuaciones lineales

$$\begin{pmatrix} 1 & a & b \\ 1 & 1 & c \\ 1 & 1 & 1 \end{pmatrix}\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix}$$

se propone el siguiente método iterativo

$$\begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} x_{n+1} \\ y_{n+1} \\ z_{n+1} \end{pmatrix} + \begin{pmatrix} 0 & a & b \\ 0 & 0 & c \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} x_n \\ y_n \\ z_n \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix}$$

(a) Encuentre los valores de las constantes a, b y c de modo que el método propuesto sea convergente cualesquiera que sean los vectores $\mathbf{x}^{(0)} = (x_0, y_0, z_0)^T$ y $\mathbf{b} = (b_1, b_2, b_3)^T$.

(b) De un ejemplo de valores de las constantes a, b y c de modo que el método propuesto arriba sea convergente y compare dicho método con el método de Jacobi (con los mismos valores para las constantes) en cuanto a velocidad de convergencia.

Solución.

(a) El método iterativo propuesto tiene, primero que nada ser puesto en la forma

$$\mathbf{x}^{(n+1)} = M\mathbf{x}^{(n)} + R$$

Como tenemos

$$\begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} x_{n+1} \\ y_{n+1} \\ z_{n+1} \end{pmatrix} + \begin{pmatrix} 0 & a & b \\ 0 & 0 & c \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} x_n \\ y_n \\ z_n \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix}$$

despejando nos queda

$$\begin{pmatrix} x_{n+1} \\ y_{n+1} \\ z_{n+1} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 0 & -1 & 1 \end{pmatrix} \begin{pmatrix} 0 & a & b \\ 0 & 0 & c \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} x_n \\ y_n \\ z_n \end{pmatrix} + \begin{pmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 0 & -1 & 1 \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix}$$

De donde

$$\begin{pmatrix} x_{n+1} \\ y_{n+1} \\ z_{n+1} \end{pmatrix} = \begin{pmatrix} 0 & -a & -b \\ 0 & a & b-c \\ 0 & 0 & c \end{pmatrix} \begin{pmatrix} x_n \\ y_n \\ z_n \end{pmatrix} + \begin{pmatrix} b_1 \\ b_2 - b_1 \\ b_3 - b_2 \end{pmatrix}.$$

Ahora, como un método iterativo para matrices es convergente cuando $||M||_\infty < 1$. Tenemos

$$||M||_\infty = \max\{|a| + |b|, |a| + |b - c|, |c|\} < 1$$

cuando

$$\begin{cases} |a| + |b| & < & 1 \\ |a| + |b - c| & < & 1 \\ |c| & < & 1 \end{cases}$$

Otra forma, y la más segura es analizar el radio espectral de M. En este caso, tenemos que el método iterativo converge si y sólo si $\rho(M) < 1$.

En nuestro caso, tenemos

$$\det(M - \lambda I) = \det \begin{pmatrix} -\lambda & -a & -b \\ 0 & a - \lambda & b - c \\ 0 & 0 & c - \lambda \end{pmatrix} = \lambda(a - \lambda)(c - \lambda) = 0$$

si y sólo si $\lambda_1 = 0$, $\lambda_2 = a$ y $\lambda_3 = c$. Por lo tanto $\rho(M) = \max\{0, |a|, |c|\} < 1$ si y sólo si $|a| < 1$ y $|c| < 1$.

b) El sistema de ecuaciones lineales es

$$\begin{pmatrix} 1 & a & b \\ 1 & 1 & c \\ 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix}$$

La matriz del método de Jacobi es dada por

$$M_J = \begin{pmatrix} 0 & -a & -b \\ -1 & 0 & -c \\ -1 & -1 & 0 \end{pmatrix}.$$

Su radio espectral es dado por

$$\det \begin{pmatrix} -\lambda & -a & -b \\ -1 & -\lambda & -c \\ -1 & -1 & -\lambda \end{pmatrix} = -\lambda^3 - b - ac + \lambda(a+b+c) = 0$$

Si tomamos, por ejemplo, $a = c = 1/2$ y $b = -1/4$, obtenemos

$$\rho(M_J) = \max\left\{0, \frac{\sqrt{3}}{2}, \left|\frac{-\sqrt{3}}{2}\right|\right\} = \frac{\sqrt{3}}{2}$$

y

$$\rho(M) = \max\{0, |a|, |c|\} = \frac{1}{2}$$

Por lo tanto, para esos valores de a, b y c el método propuesto converge más rápido que el método de Jacobi.

Problema 1.14 Considere el sistema de ecuaciones lienales

$$\begin{cases} ax + by &= p \\ cx + dy &= q \end{cases}$$

con $a, d \neq 0$.

(a) Escribir los métodos de Jacobi y de Gauss–Seidel para este sistema. Probar que ambos convergen o ambos divergen si y sólo si $|bc| < |ad|$.

(b) Si además se tiene que $\frac{cb}{ad} > 0$; ¿Qué puede decir acerca de la convergencia del método de SOR?

(c) Suponga que $a = d = 2$, $b = c = 1$ y $p = q = 3$. para el método de Gauss–Seidel, obtener una expresión explícita para el error en función del número de iteraciones, tomando como punto inicial $\mathbf{x}^{(0)} = (0, 0)$. Use la norma $||\cdot||_\infty$.

Solución. Tenemos el sistema de ecuaciones lineales

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} p \\ q \end{pmatrix}$$

con $a, d \neq 0$.

(a) Para el método de Jacobi, tenemos que

$$Q_J = \begin{pmatrix} a & 0 \\ 0 & d \end{pmatrix}.$$

Luego

$$Q_J^{-1} = \begin{pmatrix} \frac{1}{a} & 0 \\ 0 & \frac{1}{d} \end{pmatrix}$$

así

$$Q_J^{-1}A = \begin{pmatrix} \frac{1}{a} & 0 \\ 0 & \frac{1}{d} \end{pmatrix}\begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} 1 & \frac{b}{a} \\ \frac{c}{d} & 1 \end{pmatrix}$$

y

$$T_J = I - Q_J^{-1}A = \begin{pmatrix} 0 & -\frac{b}{a} \\ -\frac{c}{d} & 0 \end{pmatrix}.$$

Ahora, $\det(T_J - \lambda I) = \lambda^2 - \frac{bc}{ad} = 0$ si y sólo si $\lambda = \pm\sqrt{\left|\frac{bc}{ad}\right|}$. Luego, $\rho(T_J) = \sqrt{\left|\frac{bc}{ad}\right|} < 1$ si y sólo si $\left|\frac{bc}{ad}\right| < 1$.

Para Gauss–Seidel, tenemos

$$Q_{GS} = \begin{pmatrix} a & 0 \\ c & d \end{pmatrix}$$

y

$$Q_{GS}^{-1} = \begin{pmatrix} \frac{1}{a} & 0 \\ -\frac{c}{ad} & \frac{1}{d} \end{pmatrix}.$$

Luego,

$$T_{GS} = I - Q_{GS}^{-1}A = \begin{pmatrix} 0 & -\frac{b}{a} \\ 0 & \frac{bc}{ad} \end{pmatrix}$$

y $\det(T_{GS} - \lambda I) = -\lambda(-\lambda + \frac{bc}{ad}) = 0$ si y sólo si $\lambda = 0$ o $\lambda = \frac{bc}{ad}$. Por lo tanto $\rho(T_{GS}) = \left|\frac{bc}{ad}\right|$. Consecuentemente, el método iterativo de Gauss–Seidel converbe si y sólo si $\frac{|cb|}{|ad|} < 1$.

(b) Tenemos que $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ es claramente tridiagonal. Ahora, A es positiva definida si $\det(A_1) = \det(a) > 0$ y $\det(A) = ad - bc > 0$, esto si $ad > bc$, de donde $\frac{bc}{ad} < 1$ que es la condición dada en la parte (a), esto sólo si $d > 0$. Luego, si estas condiciones son válidas, la matriz A es tridiagonal y positiva definida, luego el método iterativo SOR converge para todo $0 < \omega < 2$. Además, como A es tridiagonal, se tiene que el ω óptimo es dado por

$$\omega = \frac{2}{1 + \sqrt{1 - \rho(T_{GS})}} = \frac{2}{1 + \sqrt{1 - \frac{|bc|}{|ad|}}}$$

Nota: En este caso, se tiene que $\rho(T_{GS}) = (\rho(T_J))^2$.

(c) Reemplazando los valores dados, nos queda

$$A = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} \qquad \text{y} \qquad \begin{pmatrix} p \\ q \end{pmatrix} = \begin{pmatrix} 3 \\ 3 \end{pmatrix},$$

Sea x_T la solución exacta del sistema. Tomamos $x^{(0)} = (0,0)^T$ y usamos la fórmula

$$||x^{(k)} - x_T|| \leqslant \frac{||T_{GS}||^k}{1 - ||T_{GS}||} \, ||x^{(1)} - x^{(0)}|| \, .$$

Ahora, en este caso,

$$T_{GS} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} - \begin{pmatrix} 2 & 0 \\ 1 & 2 \end{pmatrix}^{-1} \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} = \begin{pmatrix} 0 & -\frac{1}{2} \\ 0 & \frac{1}{4} \end{pmatrix}$$

y

$$b_{GS} = Q_{GS}^{-1} b = \begin{pmatrix} \frac{1}{2} & 0 \\ -\frac{1}{2} & \frac{1}{2} \end{pmatrix} \begin{pmatrix} 3 \\ 3 \end{pmatrix} = \begin{pmatrix} \frac{3}{2} \\ 0 \end{pmatrix}$$

luego, $x^{(1)} = T_{GS} x^{(0)} + b_{GS} = \begin{pmatrix} \frac{3}{2} \\ 0 \end{pmatrix}$.

Ahora, para simplificar los cálculos usamos la norma $|| \, ||_\infty$ en el espacio de matrices. Luego, $||T_{GS}||_\infty = \max\{\frac{1}{2}, \frac{1}{4}\} = \frac{1}{2}$ y $||x^{(1)} - x^{(0)}||_\infty = \frac{3}{2}$. Por lo tanto,

$$||x^{(k)} - x_T||_\infty \leqslant \frac{(1/2)^k}{1 - 1/2} \cdot \frac{3}{2} = 3 \left(\frac{1}{2}\right)^k .$$

Problema 1.15 Sea el sistema de ecuaciones lineales paramétrico:

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

donde $a > 0$.

(a) ¿Para qué valores del parámetro a se tiene que la matriz del sistema es definida positiva?

(b) ¿Para que valores de a, los métodos iterativos de Jacobi y Gauss Seidel aplicados para resolver el sistema convergen? ¿Qué puede decir acerca del método de SOR?

Solución. Tenemos

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

la cual es claramente es simétrica.

(a) Para que A sea positiva definida debemos tener que $\det(A_1) = \det(a)_{1\times 1} = a > 0$, que es dada como una condición del problema, $\det(A_2) = \det \begin{pmatrix} a & -1 \\ -1 & 2 \end{pmatrix} = 2a - 1 > 0$, de donde $A > 1/2$, finalmente debe tenerse que $\det(A_3) = \det(A) = (2-a)\det(A_2) = (2-a)(2a-1) > 0$, siendo que $a > 1/2$, de debe tner que $2 - a > 0$, de donde $a > 2$. Por lo tanto, A es positiva definida si y sólo si $1/2 < a < 2$.

Nota. Sin usar propiedades de la función determinamte, se tiene que $\det(A) = -2a^2 + 5a - 2$, y debemos resolver la inecuación $-2a^2 + 5a - 2 > 0$. La función $g(a) = -2a^2 + 5a - 2$ tiene por gráfico una parábola, que corta al eje de las abscisas en los puntos $a_1 = 1/2$ y $a_2 = 2$,

y es concava, por lo tanto la parte del gráfico de esta parábola que queda sobre el eje de las abscisas corresponde a los valores de a, tales que $1/2 < a < 2$. En otras palabras, A es positiva definida si y sólo si $1/2 < a < 2$.

(b). Veamos primero el método de Jacobi. La matriz de iteración en este método es dada por

$$T_J = I - \text{diag}(A)^{-1}A$$

donde diag es la metriz formada por los elemento de la diagonal de A. En nuestro caso tenemos

$$T_J = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} - \begin{pmatrix} \frac{1}{a} & 0 & 0 \\ 0 & \frac{1}{2} & 0 \\ 0 & 0 & \frac{1}{2-a} \end{pmatrix} \begin{pmatrix} a & -1 & 0 \\ -1 & 2 & 0 \\ 0 & 0 & 2-a \end{pmatrix} = \begin{pmatrix} 0 & \frac{1}{a} & 0 \\ \frac{1}{2} & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

es decir

$$T_J = \begin{pmatrix} 0 & \frac{1}{a} & 0 \\ \frac{1}{2} & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

Calculemos el radio espectral de T_J. Tenemos

$$\det(T_J - \lambda I) = \det \begin{pmatrix} -\lambda & \frac{1}{a} & 0 \\ \frac{1}{2} & -\lambda & 0 \\ 0 & 0 & -\lambda \end{pmatrix} = -\lambda \left(\lambda^2 - \frac{1}{2a} \right)$$

de donde los valores propios de T_J son $\lambda_1 = 0$ y $\lambda_{2,3} = \pm\sqrt{\frac{1}{2a}}$. Ahora queremos $\rho(T_J) = \max\{0, \sqrt{\frac{1}{2}}\} < 1$ se debe tener que $\sqrt{\frac{1}{2a}} < 1$, de donde $a > 1/2$. Por lo tanto el método iterativo de Jacobi converge para todos los valores de a, tales que $a > 1/2$.

Si usamos los valores de a para los cuales A es positiva definida, entonces como claramente A es simétrica y tridiagonal, se tiene que

$$\rho(T_{GS}) = \rho(T_J)^2 = \frac{1}{2a},$$

y para SOR, tenemos

$$\omega_{opt} = \frac{2}{1 + \sqrt{1 - \rho(T_J)^2}} = \frac{2}{1 + \sqrt{1 - \rho(T_{GS})}}$$

y

$$\rho(T_{SOR,\omega_{opt}}) = \omega_{opt} - 1.$$

reemplazando los datos nos que

$$\omega_{opt} = \frac{2}{1 + \sqrt{1 - \frac{1}{2a}}} = \frac{2\sqrt{2a}}{\sqrt{2a} + \sqrt{2a-1}}$$

de donde

$$\rho(T_{SOR,\omega_{opt}}) = \frac{2\sqrt{2a}}{\sqrt{2a} + \sqrt{2a-1}} - 1$$

Nota. También se pueden realizar los cálculo directamente. Tenemos

$$\begin{aligned} T_{GS} &= \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} - \begin{pmatrix} a & 0 & 0 \\ -1 & 2 & 0 \\ 0 & 0 & 2-a \end{pmatrix}^{-1} \begin{pmatrix} a & -1 & 0 \\ -1 & 2 & 0 \\ 0 & 0 & 2-a \end{pmatrix} \\ &= \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} - \begin{pmatrix} \frac{1}{a} & 0 & 0 \\ \frac{1}{2a} & \frac{1}{2} & 0 \\ 0 & 0 & -\frac{1}{2-a} \end{pmatrix} \begin{pmatrix} a & -1 & 0 \\ -1 & 2 & 0 \\ 0 & 0 & 2-a \end{pmatrix} \\ &= \begin{pmatrix} 0 & \frac{1}{a} & 0 \\ 0 & \frac{1}{2a} & 0 \\ 0 & 0 & 0 \end{pmatrix} \end{aligned}$$

Ahora $\det(T_{GS} - \lambda I) = \lambda^2\left(\frac{1}{2a} - \lambda\right) = 0$, nos da $\lambda_{1,2} = 0$ y $\lambda_3 = \frac{1}{2a}$. Por lo tanto $\rho(T_{GS}) = \frac{1}{2a} < 1$ si y sólo si $a > 1/2$.

Problema 1.16 Para resolver el sistema de ecuaciones lineales $A\mathbf{x} = \mathbf{b}$, con $A \in M(n \times n\mathbb{R})$, $\mathbf{x}, \mathbf{b} \in \mathbb{R}^n$, con $\det(A) \neq 0$, se propone el método iterativo de punto fijo

$$\mathbf{x}^{(k+1)} = B\mathbf{x}^{(k)} + \mathbf{c}.$$

Si el método es convergente. Determinar la matriz de iteración B y muestre una forma de estimar el error.

Solución. Sea $\mathbf{x}_T = A^{-1}\mathbf{b}$ la solución exacta del sistema. Tenemos entonces que $\mathbf{x}_T = B\mathbf{x}_T + \mathbf{c}$, de donde $A^{-1}\mathbf{b} = BA^{-1}\mathbf{b} + \mathbf{c}$. De esto obtenemos que $\mathbf{c} = A^{-1}\mathbf{b} - BA^{-1}\mathbf{b} = (I - B)A^{-1}\mathbf{b}$. Llamando $M = (I - B)A^{-1}$, nos queda $\mathbf{c} = M\mathbf{b}$. De la definición de M, obtenemos $B = I - AM$. Por lo tanto el método iterativo tiene la forma

$$\mathbf{x}^{(k+1)} = (I - AM)\mathbf{x}^{(k)} + M\mathbf{b}\,.$$

Ahora, para estimar el error $\mathbf{e}^{(k)} = \mathbf{x}^{(k)} - \mathbf{x}_T$, tenemos

$$\begin{aligned} \mathbf{e}^{(k+1)} &= \mathbf{x}^{(k+1)} - \mathbf{x}_T \\ &= [(I - AM)\mathbf{x}^{(k)} + M\mathbf{b}] - [(I - AM)\mathbf{x}_T + M\mathbf{b}] \\ &= (I - AM)(\mathbf{x}^{(k)} - \mathbf{x}_T) \\ &= (I - AM)\mathbf{e}^{(k)} \\ &\vdots \\ &= (I - AM)^{k+1}\mathbf{e}^{(0)} \end{aligned}$$

de donde

$$||\mathbf{e}^{(k)}|| \leqslant ||I - AM||^k \, ||\mathbf{e}^{(0)}|| \overset{k\to\infty}{\longrightarrow} 0$$

si $||I - AM|| < 1$.

Problema 1.17 Considere el siguiente sistema de ecuaciones lineales

$$\begin{aligned} x + ay &= 1 \\ ax + y + bz &= 1 \\ by + z &= 1 \end{aligned}$$

(a) Determine para qué valores de a y b el método iterativo de Jacobi es convergente y a su vez, la matriz del sistema original es definida positiva.

(b) Tome un par de valores cualesquiera de a y b ambos no cero, con $a \neq b$, de la parte (**a**) y determine cuantas iteraciones necesitaría realizar con el método de Gauss Seidel para obtener una precisión de la solución de 10^{-4}, tomando como punto de partida el $(0, 0, 0)$.

(c) Calcule el radio espectral del método iterativo de SOR tomando ω de manera óptima. ¿Se necesitarán más o menos iteraciones para obtener la misma precisión de la parte (**b**) usando este método?. Justifique su respuesta.

Solución. Tenemos el sistema de ecuaciones lineales

$$\begin{pmatrix} 1 & a & 0 \\ a & 1 & b \\ 0 & b & 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \tag{1.53}$$

(a) Para que la martriz A del sistema sea positiva definida, debemos tener que $\det(A_1) = \det(1) = 1 > 0$ (obvio), $\det(A_2) = \det \begin{pmatrix} 1 & a \\ a & 1 \end{pmatrix} = 1 - a^2 > 0$, esto es, $-1 < a < 1$ (cond. 1) y $\det(A_3) = \det(A) = 1 - a^2 - b^2 > 0$, en otras palabras, $a^2 + b^2 < 1$ (cond.2). Claramente, la cond.2 contiene como caso particular la cond.1, por lo tanto, la condición que deben satisfacer los coeficientes de la matriz para ser positiva definida es

$$a^2 + b^2 < 1\,, \tag{1.54}$$

es decir, los pares ordenado $(a, b) \in \mathbb{R}^2$ que satisfacen que la matriz del sistema es positiva definida son aquellos dados por (1.54) y corresponden a los puntos en el interior del círculo unitario del plano.

Alternativamente, podemos imponer la condición de que todos los valores propios de A sean positivos.

Calculando el polinomio caraterísticos de A, obtenemos

$$p(\lambda) = \det(A - \lambda I) = \lambda^3 - 3\lambda^2 + (3 - a^2 - b^2)\lambda - 1 + a^2 + b^2 \tag{1.55}$$

por simple observación vemos que $\lambda_1 = 1$ es una raíz del la ecuación $p(\lambda) = 0$; dividiendo el polinomio característico por $q(\lambda) = \lambda - 1$, nos queda una ecuación de segundo grado, cuyas raíces son $\lambda_2 = 1 + \sqrt{a^2 + b^2}$ y $\lambda_3 = 1 - \sqrt{a^2 + b^2}$. es claro que $\lambda_2 > 0$, sólo nos resta imponer que $\lambda_3 > 0$, esto es, $\sqrt{a^2 + b^2} < 1$, o lo que es lo mismo $a^2 + b^2 < 1$, que es exactamente la misma condición anterior.

Para el método de Jacobi, tenemos

$$Q_J = I \qquad \text{o} \qquad D = I \tag{1.56}$$

matriz identidad 3×3, y la matriz de iteración para el método de Jacobi es entonces

$$T_J = I - Q_J^{-1}A = -D^{-1}(E+F) = \begin{pmatrix} 0 & -a & 0 \\ -a & 0 & -b \\ 0 & -b & 0 \end{pmatrix} \tag{1.57}$$

El polinomio caraterístico de T_J es dado por

$$p_J(\lambda) = \det(\lambda I - T_J) = \lambda^3 - \lambda(a^2+b^2)$$

cuyas raíces son $\lambda_1 = 0$, $\lambda_2 = -\lambda_3 = \sqrt{a^2+b^2}$. Luego el radio espectral de la matriz de iteración del método de Jacobi es $\rho(T_J) = \sqrt{a^2+b^2}$. Por lo tanto, el método de Jacobi converge si y sólo si $a^2+b^2 < 1$.

Usando la condición más débil $||T_J|| < 1$, y considerando la norma $||\ ||_\infty$ nos queda

$$||T_J||_\infty = \max\{|a|, |a|+|b|, |b|\} = |a|+|b|$$

así, $||T_J||_\infty < 1$ si y sólo si $|a|+|b| < 1$. Por lo tanto, si $|a|+|b| < 1$ el método de Jacobi converge.

Nota. La región del plano determinada por la condición $|a|+|b| < 1$ está enteramente contenida dentro del disco unitario abierto $a^2+b^2 < 1$, lo que muestra, que la condición con el radio espectral es más general que la condición con normas.

(b) Tomamos $a = b = \frac{1}{2}$ se tiene que

$$\begin{aligned} T_{GS} &= -(D+E)^{-1}F = -\begin{pmatrix} 1 & 0 & 0 \\ \frac{1}{2} & 1 & 0 \\ 0 & \frac{1}{2} & 1 \end{pmatrix}^{-1} \begin{pmatrix} 0 & \frac{1}{2} & 0 \\ 0 & 0 & \frac{1}{2} \\ 0 & 0 & 0 \end{pmatrix} \\ &= -\begin{pmatrix} 1 & 0 & 0 \\ -\frac{1}{2} & 1 & 0 \\ \frac{1}{4} & -\frac{1}{2} & 1 \end{pmatrix} \begin{pmatrix} 0 & \frac{1}{2} & 0 \\ 0 & 0 & \frac{1}{2} \\ 0 & 0 & 0 \end{pmatrix} \\ &= \begin{pmatrix} 0 & -\frac{1}{2} & 0 \\ 0 & \frac{1}{4} & -\frac{1}{2} \\ 0 & -\frac{1}{8} & \frac{1}{4} \end{pmatrix} \end{aligned}$$

y

$$v_{GS} = (D+E)^{-1}\mathbf{b} = \begin{pmatrix} 1 & 0 & 0 \\ -\frac{1}{2} & 1 & 0 \\ \frac{1}{4} & -\frac{1}{2} & 1 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 1 \\ \frac{1}{2} \\ \frac{3}{4} \end{pmatrix}$$

por tanto, si tomamos como punto inicial $\mathbf{x}^{(0)} = (0,0,0)$ entonces $\mathbf{x}^{(1)} = v_{GS} = (1, \frac{1}{2}, \frac{3}{4})$ y finalmente, si sustituimos en la fórmula

$$\left\|\mathbf{x}^{(n)} - \bar{\mathbf{x}}\right\|_\infty \leqslant \frac{\|T_{GS}\|_\infty^n}{1 - \|T_{GS}\|_\infty} \left\|\mathbf{x}^{(1)} - \mathbf{x}^{(0)}\right\|_\infty \leqslant 10^{-4}$$

los valores

$$\|T_{GS}\|_\infty = \max\left\{\frac{1}{2}, \frac{3}{4}, \frac{3}{8}\right\} = \frac{3}{4}$$

y

$$\left\|\mathbf{x}^{(1)} - \mathbf{x}^{(0)}\right\|_\infty = \max_{1 \leqslant i \leqslant 3}\left\{\left|\mathbf{x}_i^{(1)} - \mathbf{x}_i^{(0)}\right|\right\} = \max\left\{1, \frac{1}{2}, \frac{3}{4}\right\} = 1$$

obtenemos

$$\begin{aligned} \frac{(\frac{3}{4})^n}{1-\frac{3}{4}}1 &\leqslant 10^{-4} \\ 4\left(\frac{3}{4}\right)^n &\leqslant 10^{-4} \\ 4\times 10^4 &\leqslant \left(\frac{4}{3}\right)^n \\ \frac{4+\log 4}{\log 4-\log 3} &\leqslant n \\ 36.8345 &\leqslant n \end{aligned}$$

Es decir, se necesitan 37 iteraciones.

(c) Es claro que la matriz del sistema es simétrica, y siendo positiva definida, el valor de ω en el método de SOR, es dado por la fórmula ($\rho(T_J)$ representa el radio espectral de la matriz de Jacobi el cual fue calculado en el items (a))

$$\omega^* = \frac{2}{1+\sqrt{1-\rho(T_J)^2}} = \frac{2}{1+\sqrt{1-a^2-b^2}} \tag{1.58}$$

y para este valor de ω^*, se tiene que

$$\rho(T_\omega) = \omega^* - 1$$

donde T_ω es la matriz de iteración del método de SOR para el valor óptimo de ω. Tomando ahora $a = b = \frac{1}{2}$ y sustituyendo en (1.58) se tiene que

$$\omega^* = \frac{2}{1+\sqrt{1-\frac{1}{4}-\frac{1}{4}}} = \frac{2}{1+\sqrt{1-\frac{1}{2}}} = 1.1715$$

y por tanto $\rho(T_\omega) = \omega^* - 1 = 0.1715$. Por otro lado, el radio espectral de la metriz de Gauss-Seidel es el cuadrado del radio espectral de la matriz de Jacobi dado que la matriz del sistema original es definida positiva, simétrica y tridiagonal, por tanto $\rho(T_{GS}) = \rho(T_J)^2 = a^2 + b^2 = \frac{1}{2}$. Finalmente, comparando los valores de los radios espectrales, vemos que $\rho(T_\omega) < \rho(T_{GS})$ y por ende, el método SOR necesitará menos iteraciones para converger.

Problema 1.18 Sea $A \in M_{n\times n}(\mathbb{R})$ una matriz invertible y supongamos que $C \in M_{n\times n}(\mathbb{R})$ es una aproximación *conocida* de la inversa A^{-1} de A. Utilizaremos la matriz C para aproximar iterativamente la solución del sistema lineal $Ax = b$, por medio del, así llamado, *método de corrección residual*.

Sea $x^{(0)}$ un punto inicial, y definamos $r^{(0)} = b - Ax^{(0)}$ el *residuo* asociado a $x^{(0)}$. Definamos además $x^{(1)} = x^{(0)} + Cr^{(0)}$. En general, se define

$$r^{(m)} = b - Ax^{(m)} \qquad x^{(m+1)} = x^{(m)} + Cr^{(m)}, \quad \text{para } m \geqslant 0.$$

(a) Escriba este método en la forma estándar

$$x^{(m+1)} = Mx^{(m)} + v,$$

explicitando la matriz de iteración M y el vector de iteración v.

(b) Escogiendo $C = D^{-1}$, con D la parte diagonal de A, verifique que este método es convergente si A es diagonal dominante.

(c) Dado un número real ϵ, defina A como

$$A = \begin{pmatrix} 2 & 1+\epsilon & \epsilon \\ 1-\epsilon & 2 & 1+\epsilon \\ -\epsilon & 1-\epsilon & 2 \end{pmatrix}.$$

Como aproximación de la inversa de A, utilice

$$C = \begin{pmatrix} \frac{3}{4} & -\frac{1}{2} & \frac{1}{4} \\ -\frac{1}{2} & 1 & -\frac{1}{2} \\ \frac{1}{4} & -\frac{1}{2} & \frac{3}{4} \end{pmatrix}.$$

(i) Muestre que $\|I - CA\|_\infty = |\epsilon|$.

(ii) Qué condición debe satisfacer ϵ para asegurar la convergencia del método en este caso?

Solución .

(a) Reeplazando la fórmula para el residuo $r^{(m)}$ en la fórmula para $x^{(m+1)}$, tenemos que

$$x^{(m+1)} = x^{(m} + C(b - Ax^{(m)}) = (I - CA)x^{(m)} + Cb,$$

es decir, $M = I - CA$ y $v = Cb$.

(b) Si $C = D^{-1}$ entonces $M = I - D^{-1}A = I - D^{-1}(D + E + F) = -D^{-1}(E + F)$, y $v = D^{-1}b$, que corresponden a la matriz y al vector de iteración del método de Jacobi. Por lo tanto, para $C = D^{-1}$ este método es convergente, dado que la matriz A sea diagonal dominante.

(c) Haremos los ítemes (i) y (ii) juntos. Multiplicando A por C, tenemos que

$$CA = \begin{pmatrix} 1+\frac{\epsilon}{4} & \frac{\epsilon}{2} & \frac{\epsilon}{4} \\ -\frac{\epsilon}{2} & 1 & \frac{\epsilon}{2} \\ -\frac{\epsilon}{4} & -\frac{\epsilon}{2} & 1-\frac{\epsilon}{4} \end{pmatrix}.$$

Luego,

$$I - CA = \begin{pmatrix} -\frac{\epsilon}{4} & -\frac{\epsilon}{2} & -\frac{\epsilon}{4} \\ \frac{\epsilon}{2} & 0 & -\frac{\epsilon}{2} \\ \frac{\epsilon}{4} & \frac{\epsilon}{2} & \frac{\epsilon}{4} \end{pmatrix}.$$

Por lo tanto,

$$\|I - CA\|_\infty = \max\{\left|-\frac{\epsilon}{4}\right| + \left|-\frac{\epsilon}{2}\right| + \left|-\frac{\epsilon}{4}\right|, \left|\frac{\epsilon}{2}\right| + \left|-\frac{\epsilon}{2}\right|, \left|\frac{\epsilon}{4}\right| + \left|\frac{\epsilon}{2}\right| + \left|\frac{\epsilon}{4}\right|\} = |\epsilon|.$$

Luego, la condición que debe satisfacer ϵ para que el método converja es

$$|\epsilon| < 1 \iff -1 < \epsilon < 1,$$

ya que así $\|I - CA\|_\infty = |\epsilon| < 1$ y esta condición implica la convergencia del método.

Problema 1.19 Sea

$$A = \begin{pmatrix} \alpha & 1 & 0 \\ 1 & \alpha & \beta \\ 0 & \beta & \alpha \end{pmatrix}, \qquad \alpha, \beta \in \mathbb{R}.$$

(a) Encuentre que relación deben cumplir α y β de modo que la matriz A resulte ser definida positiva.

(b) Encuentre que relación deben cumplir α y β de modo que se asegure la convergencia de los métodos iterativos de Jacobi y Gauss-Seidel aplicados para resolver el sistema lineal $A\mathbf{x} = \mathbf{b}$ Qué puede decir acerca de la convergencia del método SOR en este caso?

(c) Fije unos valores para α y β de modo que se cumplan ambos items de arriba, calcule los radios espectrales de los tres métodos (en el caso de SOR diga cual es el parámetro ω óptimo) y compare estos métodos en cuanto a velocidad de convergencia.

Solución. Tenemos

$$A = \begin{pmatrix} \alpha & 1 & 0 \\ 1 & \alpha & \beta \\ 0 & \beta & \alpha \end{pmatrix}$$

Notemos que A es simétrica y tridiagonal.

a) Para que A sea positiva definida se debe tener que $\det(A_1) = \det(\alpha) = \alpha > 0$, $\det(A_2) = \det\begin{pmatrix} \alpha & 1 \\ 1 & \alpha \end{pmatrix} = \alpha^2 - 1 > 0$, es decir, $\alpha > 1$ o $\alpha < -1$, pero como $\alpha > 0$ se debe tener que $\alpha > 1$, esta es una primera condición. Finalmente, debemos tener que $\det(A_3) = \det(A) = \alpha^3 - \alpha\beta^2 - \alpha > 0$, es decir, $\alpha(\alpha^2 - 1 - \beta^2) > 0$, como $\alpha > 1$ se debe tener entonces que $\alpha^2 - (1 + \beta^2) > 0$, esto es,

$$\alpha > \sqrt{1 + \beta^2}\,. \tag{1.59}$$

Esta condición implica las otras dos condiciones anteriores ($\alpha > 0$ y $\alpha > 1$). Por lo tanto para que A sea positiva definida basta imponer la condición (1.59) anterior.

Nota. Esta es la misma condición que se obtiene al usar el resultado que nos dice que *"una matriz simétrica es positiva definida si y sólo si sus valores propios son positivos"*

b) Para tener asegurada la convergencia de los métodos iterativos de Jacobi y de Gauss–Seidel, basta imponer la condición de diagonal dominante, es decir,

(i) $|\alpha| > 0$

(ii) $|\alpha| > 1 + |\beta|$

(iii) $|\alpha| > |\beta|$

Note que las condiciones (i) y (iii) se deducen de (ii). Luego la condición a imponer es (ii), es decir, si $|\alpha| > 1 + |\beta|$ entonces los método iterativos de Jacobi y de Gauss–Seidel convergen. Para la convergencia del método SOR basta imponer la condición $0 < \omega < 2$.

Nota. Como la matriz A es definida positiva, simétrica y tridiagonal, entonces los tres métodos iterativos son convergentes (SOR converge siempre que $0 < \omega < 2$). Por lo tanto, la condición (1.59) es suficiente para la convergencia de los métodos. Note además que la condición (1.59) es precisamente la condición que establece que el radio espectral de la matriz de iteración del método de Jacobi es menor que 1.

c) Tenemos $Q_J = \begin{pmatrix} \alpha & 0 & 0 \\ 0 & \alpha & 0 \\ 0 & 0 & \alpha \end{pmatrix}$ y $T_J = Id - Q_J^{-1}A = \begin{pmatrix} 0 & -\frac{1}{\alpha} & 0 \\ -\frac{1}{\alpha} & 0 & -\frac{\beta}{\alpha} \\ 0 & -\frac{\beta}{\alpha} & 0 \end{pmatrix}$ es la matriz de iteración del método de Jacobi. Ahora, los valores propios de T_J son obtenidos como las soluciones de la ecuación

$$p(\lambda) = \det(\lambda Id - T_J) = \lambda \, \frac{(\lambda^2\alpha^2 - \beta^2 - 1)}{\alpha^2} = 0$$

esto es, $\lambda_1 = 0$ o $\lambda_{2,3} = \pm\frac{\sqrt{1+\beta^2}}{\alpha}$. Luego, el radio espectral es $\rho(T_J) = \frac{\sqrt{1+\beta^2}}{\alpha}$.

Además, como A es simétrica, positiva definida y tridiagonal, se tiene que $\rho(T_{GS}) = \rho(T_J)^2 = \frac{1+\beta^2}{\alpha^2}$ y el ω óptimo está dado en este caso por

$$\omega = \frac{2}{1 + \sqrt{1 - \rho(T_{GS})}} = \frac{2\alpha}{\alpha + \sqrt{\alpha^2 - (\beta^2 + 1)}}.$$

Con este valor de ω el radio espectral del método SOR está dado por $\rho_{\text{SOR}} = \omega - 1$. Se verifica fácilmente que

$$\rho_{\text{SOR}} < \rho(T_{GS}) < \rho(T_J)$$

y en consecuencia el método SOR es el más rápido, en cuanto a velocidad de convergencia (para ω igual al ω óptimo), de los tres métodos. Después le sigue el método de Gauss-Seidel y finalmente el más lento en converger es el método de Jacobi.

Problema 1.20 Considere el sistema de ecuaciones lineales $A\mathbf{x} = \mathbf{b}$, donde

$$A = \begin{pmatrix} 1.2969 & 0.8648 \\ 0.2161 & 0.1441 \end{pmatrix}, \qquad \mathbf{b} = \begin{pmatrix} 0.8642 \\ 0.1440 \end{pmatrix}, \qquad \widetilde{\mathbf{b}} = \begin{pmatrix} 0.86419999 \\ 0.14400001 \end{pmatrix}$$

(a) Calcule el número de condicionamiento de A usando la norma $||\cdot||_\infty$. Es estable el sistema lineal asociado a la matriz A? Justifique su respuesta.

(b) De una cota del error relativo de aproximar la solución del sistema original por la solución del sistema perturbado usando la norma $\|\cdot\|_\infty$. Cómo es este error relativo comparado con el error relativo de la perturbación del lado derecho del sistema lineal?. Explique este resultado.

Sugerencia. Recuerde que si

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \qquad \text{entonces} \qquad A^{-1} = \frac{1}{\det(A)} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$$

Solución.a) Tenemos

$$||A||_\infty = \max\{1.2969 + 0.8648, 0.2161 + 0.1441\} = 2.1617$$

Ahora, $\det(A) = 0.1 \times 10^{-7}$ y

$$A^{-1} = \begin{pmatrix} 0.1441 \times 10^8 & -0.8648 \times 10^8 \\ -0.2161 \times 10^8 & 0.12969 \times 10^9 \end{pmatrix}$$

luego,

$$||A^{-1}||_\infty = \max\{0.10089 \times 10^9, 0.1513 \times 10^9\} = 0.1513 \times 10^9 = 151300000\,.$$

Por lo tanto,

$$k_\infty(A) = ||A||_\infty \, ||A^{-1}||_\infty = 0.1513 \times 10^9 \times 2.1617 = 0.32706521 \times 10^9 = 327065210\,.$$

En consecuencia el sistema lineal dado no es estable, pues está mal condicionado.

b) Tenemos los sistemas de ecuaciones lineales

$$A\mathbf{x} = \mathbf{b} \qquad \text{y} \qquad A\mathbf{x} = \tilde{\mathbf{b}}\,.$$

Sean x_T y x_A sus respectivas soluciones. Debemos estimar

$$E_R(x_A) = \frac{||x_T - x_A||}{||x_T||} \leqslant k(A)\, \frac{||b - \tilde{b}||}{||b||} = k(A) E_R(\tilde{b})$$

Para esto usamos la norma $||\cdot||_\infty$. Tenemos

$$||b - \tilde{b}||_\infty = \max\{10^{-8}, 10^{-8}\} = 10^{-8}\,.$$

Como $||b||_\infty = 0.8642$, obtenemos

$$E_R(\tilde{b}) = \frac{||b - \tilde{b}||_\infty}{||b||_\infty} = \frac{10^{-8}}{0.8642} = 0.115713955 \times 10^{-7}$$

así $k_\infty(A) \cdot E_R(\tilde{b}) = 327065210 \times 0.1157139551 \times 10^{-7} = 3.784600902$. Luego

$$E_R(x_A) \leqslant 3.784600902.$$

Por otro lado,

$$x_T = \begin{pmatrix} 2 \\ -2 \end{pmatrix}, \quad x_A = \begin{pmatrix} 0.9911 \\ -0.4870 \end{pmatrix}.$$

Por lo tanto $E_R(x_A) = 0.7565$, es decir, x_A es una aproximación de x_T con un error relativo del 75.65%. Se observa que el error relativo en la perturbación del lado derecho es mucho menor que 75.65%, lo cual se explica por el hecho de que el sistema lineal es inestable (debido a que la matriz A está muy mal condicionada).

Problema 1.21 Sean

$$A = \begin{pmatrix} 1 & -1 & -1 & -1 \\ 0 & 1 & -1 & -1 \\ 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 1 \end{pmatrix} \quad \text{y} \quad \mathbf{b} = \begin{pmatrix} 5.00 \\ 1.02 \\ 1.04 \\ 1.10 \end{pmatrix}$$

(a) Calcule una solución aproximada $\mathbf{x}_a$ del sistema $A\mathbf{x} = \mathbf{b}$, primero aproximando cada entrada del vector $\mathbf{b}$ al entero más próximo, obteniendo un vector $\tilde{\mathbf{b}}$ y luego resolviendo el sistema $A\mathbf{x} = \tilde{\mathbf{b}}$.

(b) Calcule $||r||_\infty$ y $k_\infty(A)$, donde $\mathbf{r}$ es el vector residual, es decir $\mathbf{r} = \mathbf{b} - A\mathbf{x}_a$ y $_\infty(A)$ el número de condicionamiento de la matriz A.

(c) Estime una cota para el error relativo de la solución aproximada, respecto a la solución exacta (no calcule esta última explícitamente).

Solución.

(a) El vector perturbado es $(5\ 1\ 1\ 1)^T$. Luego la solución del sistema perturbado es

$$\begin{pmatrix} 1 & -1 & -1 & -1 \\ 0 & 1 & -1 & -1 \\ 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \\ w \end{pmatrix} = \begin{pmatrix} 5 \\ 1 \\ 1 \\ 1 \end{pmatrix}$$

de donde $(x, y, z, w) = (12, 4, 2, 1) = \mathbf{x}_a$

(b) Tenemos $||\mathbf{b}||_\infty = ||\tilde{\mathbf{b}}||_\infty = 5$. Por otra parte,

$$\begin{aligned} \mathbf{r} = \mathbf{b} - A\mathbf{x}_a &= \begin{pmatrix} 5.00 \\ 1.02 \\ 1.04 \\ 1.10 \end{pmatrix} - \begin{pmatrix} 1 & -1 & -1 & -1 \\ 0 & 1 & -1 & -1 \\ 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 12 \\ 4 \\ 2 \\ 1 \end{pmatrix} \\ &= \begin{pmatrix} 5.00 \\ 1.02 \\ 1.04 \\ 1.10 \end{pmatrix} - \begin{pmatrix} 5 \\ 1 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0.02 \\ 0.04 \\ 0.10 \end{pmatrix} \end{aligned}$$

Luego, $||\mathbf{r}||_\infty = 0.10$.

Para encontrar la inversa de la matriz A basta resolver el sistema

$$\begin{pmatrix} 1 & -1 & -1 & -1 \\ 0 & 1 & -1 & -1 \\ 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

de donde

$$A^{-1} = \begin{pmatrix} 1 & 1 & 2 & 4 \\ 0 & 1 & 1 & 2 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

Luego, $||A^{-1}||_\infty = 8$. Tenemos que $||A||_\infty = 4$, luego $k_\infty(A) = ||A||_\infty ||A^{-1}||_\infty = 4 \times 8 = 32$.

(c) Usamos la fórmula

$$\frac{1}{k(A)}\,\frac{||r||}{||b||} \leqslant \frac{||X_T - X_A||}{||X_T||} \leqslant k(A)\,\frac{||r||}{||b||}$$

Reemplazando los datos nos queda

$$\frac{1}{32} \times \frac{0.10}{5} \leqslant \frac{||\mathbf{x}_T - \mathbf{x}_a||}{||\mathbf{x}_T||} \leqslant 32 \times \frac{0.10}{5}$$

es decir,

$$0.625 \times 10^{-3} \leqslant \frac{||\mathbf{x}_T - \mathbf{x}_a||}{||\mathbf{x}_T||} \leqslant 0.64\,.$$

Problema 1.22 Poner el enunciado Correcto

Tenemos que

$$A = \begin{pmatrix} 1 & 1-\varepsilon & 3 \\ 2 & 2 & 2 \\ 3 & 6 & 4 \end{pmatrix}$$

donde $\varepsilon > 0$ es suficientemente pequeño.

(a) Suponemos que $0 < \varepsilon < 1$, pues se asume pequeño. Luego, $||A||_\infty = \max\{5-\varepsilon, 6, 13\} = 13$.

Dado que

$$A^{-1} = \frac{1}{6+\varepsilon} \begin{pmatrix} -2 & 2\varepsilon+7 & -(\varepsilon+2) \\ -1 & -\frac{5}{2} & 2 \\ 3 & -\frac{3(\varepsilon+1)}{2} & \varepsilon \end{pmatrix}$$

se tiene $||A^{-1}||_\infty = \frac{1}{6+\varepsilon}\,\max\{11+3\varepsilon, \frac{11}{2}, \frac{9}{2}+\frac{5\varepsilon}{2}\} = \frac{11+3\varepsilon}{6+\varepsilon}$.

Por lo tanto,

$$k_\infty(A) = ||A||_\infty ||A^{-1}||_\infty = 13 \times \frac{11+3\varepsilon}{6+\varepsilon}\,.$$

(b) Dado $b = (5-\varepsilon, 6, 13)^T$ y la solución $\tilde{x} = \frac{1}{6+\varepsilon}(6-\varepsilon, 6, 6+4\varepsilon)^T$, propuesta por el alumno hábil, llamando x_T la solución exacta del sistema $Ax = b$, tenemos

$$E_R(\tilde{x}) = \frac{||x_T - \tilde{x}||}{||x_T||} \leqslant k(A) E_R(\tilde{b})$$

donde $\tilde{b}$ es tal que $\tilde{x}$ es la solución del sistema $Ax = \tilde{b}$. Ahora,

$$\begin{aligned} \tilde{b} &= A\tilde{x} \\ &= \frac{1}{6+\varepsilon}(30+5\varepsilon, 36+6\varepsilon, 78+13\varepsilon) \\ &= (5, 6, 13) \end{aligned}$$

luego, $b - \tilde{b} = (5-\varepsilon, 6, 13) - (5, 6, 13) = (-\varepsilon, 0, 0)$, así $||b-\tilde{b}||_\infty = \varepsilon$ y como $||b||_\infty = 13$, se tiene que $E_R(\tilde{b}) = \frac{\varepsilon}{13}$. Por lo tanto,

$$E_{R,\infty}(\tilde{x}) = \frac{||x_T - \tilde{x}||_\infty}{||x_T||_\infty} \leqslant k_\infty(A) E_{R,\infty}(\tilde{b}) = 13 \times \frac{11+3\varepsilon}{6+\varepsilon} \times \frac{\varepsilon}{13}$$

esto es, nos piden

$$E_{R,\infty}(\tilde{x}) = \frac{11+3\varepsilon}{6+\varepsilon} \cdot \varepsilon \leqslant 10^{-2}\,.$$

Desarrollando, obtenemos la inecuación $3\varepsilon^2 + 11\varepsilon \leqslant 6 \cdot 10^{-2} + \varepsilon 10^{-2}$, de donde obtenemos $3\varepsilon^2 + \frac{1099}{100}\varepsilon - \frac{3}{50} \leqslant 0$. Consideremos la función cuadrática $g(\varepsilon) = 3\varepsilon^2 + \frac{1099}{100}\varepsilon - \frac{3}{50}$ tiene dos raíces reales, que son $\varepsilon_1 \approx 0.005451396436$ y $\varepsilon_2 \approx -3.668784730$, y siendo el coeficiente que acompaña a ε^2 positivo, la parábola está vuelta hacia arriba, luego, $g(\varepsilon) \leqslant 0$ si y sólo si $\varepsilon_2 \leqslant \varepsilon \leqslant \varepsilon_1$, y como sólo nos interesa $\varepsilon > 0$, los valores para los cuales tenemos asegurado que $E_{R,\infty}(\tilde{x}) \leqslant 10^{-2}$ es dado por $0 \leqslant \varepsilon \leqslant \varepsilon_1$.

Problema 1.23

Problema 1.24

Problema 1.25

Problema 1.26

Problema 1.27

Problema 1.28

1.15 Ejercicios Propuestos

Problema 1.1 Para resolver un sistema de ecuaciones lineales $Ax = b$, donde $x, b \in \mathbb{R}^2$, se propone el siguiente método iterativo

$$x^{(k+1)} = Bx^{(k)} + b\,, \qquad k \geqslant 0$$

donde

$$B = \begin{pmatrix} \lambda & c \\ 0 & -\lambda \end{pmatrix}, \qquad \lambda, c \in \mathbb{R}$$

1. ¿Para qué valores de λ y c el método iterativo propuesto es convergente?
2. Sea $\tilde{x}$ el punto fijo de la iteración. Calcule $||\tilde{x} - x^{(k)}||_\infty$ y $||x^{(k+1)} - x^{(k)}||_\infty$ cuando $k \longrightarrow \infty$ ¿Es la convergencia la punto fijo independiente de c ? Justificque.

Problema 1.2 Dado un método iterativo para resolver un sistema de ecuaciones lineales

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

Si $\det(B) = 0$ ¿Puede el método propuesto ser convergente?

Problema 1.3 Si un método iterativo de punto fijo $x_{n+1} = Ax_n$, donde $A \in \mathbb{M}(n \times n, \mathbb{R})$ tiene un punto fijo $\bar{x} \neq 0$. Pruebe que $||A|| \geqslant 1$ para cualquier norma ssubordinada en $\mathbb{M}(n \times n, \mathbb{R})$.

Problema 1.4 Dada una matriz

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}$$

y un vector $\mathbf{b} \in \mathbb{R}^3$, se quiere resolver el sistema $A\mathbf{x} = \mathbf{b}$. Para ello, se propone el método iterativo siguiente:

$$\mathbf{x}^{(k+1)} = -\begin{pmatrix} a_{11} & a_{12} & 0 \\ a_{21} & a_{22} & 0 \\ 0 & 0 & a_{33} \end{pmatrix}^{-1} \cdot \begin{pmatrix} 0 & 0 & a_{13} \\ 0 & 0 & a_{23} \\ a_{31} & a_{32} & 0 \end{pmatrix} \mathbf{x}^{(k)} + \begin{pmatrix} a_{11} & a_{12} & 0 \\ a_{21} & a_{22} & 0 \\ 0 & 0 & a_{33} \end{pmatrix}^{-1} \mathbf{b}$$

(a) Pruebe que el método propuesto resulta convergente cuando se aplica a las matriz A y al vector $\mathbf{b}$ siguientes

$$A = \begin{pmatrix} 8 & 2 & -3 \\ -3 & 9 & 4 \\ 3 & -1 & 7 \end{pmatrix} \qquad \text{y} \qquad \mathbf{b} = \begin{pmatrix} -20 \\ 62 \\ 0 \end{pmatrix}.$$

Indicación: Use que $\begin{pmatrix} 8 & 2 & 0 \\ -3 & 9 & 0 \\ 0 & 0 & 7 \end{pmatrix}^{-1} = \begin{pmatrix} 3/26 & -1/39 & 0 \\ 1/26 & 4/39 & 0 \\ 0 & 0 & 1/7 \end{pmatrix}$.

(b) Considere el vector $\mathbf{x}^{(0)} = (0\,0\,0)^T$. Encuentre el número mínimo de iteraciones necesarias $k \in \mathbb{N}$, de modo de tener una precisión $||\mathbf{x}^{(k)} - \mathbf{x}||_\infty \leqslant 10^{-4}$.

(c) Compare el método anterior con el método de Jacobi en cuanto a velocidad de convergencia. Justifique.

Problema 1.5 Encuentre la inversa A^{-1} de la matriz A, las normas $||A||_1$, $||A^{-1}||_1$, $||A||_2$, $||A^{-1}||_2$, $||A||_\infty$, $||A^{-1}||_\infty$ y los números de condición $\kappa_1(A)$, $\kappa_2(A)$ y $\kappa_\infty(A)$ si

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

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

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

Problema 1.6 Sean $A, B \in \mathbb{M}(n\times n, \mathbb{R})$, matrices invertibles. Pruebe que $\kappa(AB) \leqslant \kappa(A)\kappa(B)$, donde el número de condición es calculado usando cualquier norma subordinada en $\mathbb{M}(n\times n, \mathbb{R})$.

Problema 1.7 Sean $A, B \in \mathbb{M}(3\times 3, \mathbb{R})$ las matrices

$$A = \begin{pmatrix} a & c & 0 \\ c & a & c \\ 0 & c & a \end{pmatrix}, \qquad B = \begin{pmatrix} 0 & b & 0 \\ b & 0 & b \\ 0 & b & 0 \end{pmatrix}$$

1. Probar que $\lim_{n\to\infty} B^n = 0$ si y sólo si $|b| < \sqrt{2}/2$.
2. Dar condiciones necesarias y suficientes sobre $a, c \in \mathbb{R}$ para la convergencia de los métodos de Jacobi y de Gauss–Seidel aplicados a resolver el sistema de ecuacion es lineales $A\mathbf{x} = \mathbf{b}$.

Problema 1.8 Considere el sistema de ecuaciones lineales siguiente

$$\begin{cases} 3x + y + z &= 5 \\ x + 3y - z &= 3 \\ 3x + y - 5z &= -1 \end{cases}$$

1. Explicite el método iterativo de Jacobi para encontrar solución a la ecuación. Justifique porqué este método converge (sin usar calculadora). Comenzando con $(0, 0, 0)$, realice 10 iteraciones ¿Cuál es el error absoluto respecto de la solución exacta?
2. Explicite el método iterativo de Gauss–Seidel y justifique la convergencia o no de este método para la ecuación dada.

Problema 1.9 Sean

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

Proponga un método iterativo convergente para resolver $Ax = b$. La demostración de convergencia debe hacerse sin iterar. Resuelva la ecuación utilizando el método iterativo propuesto. Use como critero de parada $||(x_{n+1}, y_{n+1}, z_{n+1}) - (x_n, y_n, z_n)||_2 \leqslant 10^{-3}$.

Problema 1.10 Considere el sistema de ecuaciones lineales

$$\begin{cases} 3x - 2y + z &= 1 \\ x - \frac{2}{3}\, y + 2z &= 2 \\ -x + 2y - z &= 0 \end{cases}$$

1. Resolver el sistema usando descomposición LU y aritmética exacta.
2. Resolver el sistema usando descomposición LU y aritmética decimal con 4 dígitos y redondeo.
3. Resolver el sistema usando aritmética exacta sin pivoteo.
4. Resolver el sistema usando aritmética decimal con 4 dígitos y redondeo, y sin pivoteo.
5. Resolver el sistema usando aritmética exacta con pivoteo.
6. Resolver el sistema usando aritmética decimal con 4 dígitos y redondeo, y con pivoteo.

7. ¿cuál es su conclusión?

Problema 1.11 1. Obtener la descomposición LU para matrices tridiagonales.

2. Describir el proceso de eliminación gaussiana sin pivoteo para resolver un sistema con matriz tridiagonal.

3. Suponiendo que tenemos una matriz tridiagonal no singular. Explicitar los algoritmos iterativos de Richardson, Jacobi, y Gauss–Seidel.

4. Ilustre todo lo anterior con el sistema

$$\begin{pmatrix} 3 & 1 & 0 \\ 1 & 3 & 1 \\ 0 & 1 & 3 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} -1 \\ 4 \\ 10 \end{pmatrix}$$

Problema 1.12 Considere los siguientes sistemas de ecuaciones lineales

$$\begin{pmatrix} 3 & -2 & 1 \\ 1 & -\frac{2}{3} & 2 \\ -1 & 2 & -1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \\ 0 \end{pmatrix} \tag{1.60}$$

$$\begin{pmatrix} 0 & 1 & 0 \\ 0 & 2 & 1 \\ 2 & 3 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 8 \\ 7 \\ -5 \end{pmatrix} \tag{1.61}$$

$$\begin{pmatrix} 3 & 1 & 1 \\ 1 & 3 & -1 \\ 3 & 1 & -5 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 5 \\ 3 \\ -1 \end{pmatrix} \tag{1.62}$$

1. Estudie la convergencia de los métodos iterativos de Richardson, Jacobi, Gauss–Seidel y de SOR para cada uno de ellos.

2. Caso algunos de los métodos anteriores sea convergente, encontrar la solución del sistema, con precisión $\varepsilon = 10^{-3}$ y usando como criterio de parada $||(x_{n+1}, y_{n+1}, z_{n+1}) - (x_n, y_n, z_n)||_\infty \leqslant \varepsilon$, comenzando con $(x_0, y_0, z_0) = (0, 0, 0)$.

Problema 1.13 Considere el sistema de ecuaciones lineales

$$\begin{pmatrix} -10^{-5} & 1 \\ 2 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

1. Resuelva el sistema usando eliminación gaussiana sin pivoteo y con aritmética de redondeo a 4 dígitos.

2. Resuelva el sistema usando eliminación gaussiana con pivoteo y con aritmética de redondeo a 4 dígitos.

3. ¿Cuál de las soluciones obtenidas en a) y b) es la más aceptable?

4. Calcule el número de condición para la matriz en el sistema inicial y para la matriz del sistema despues de hacer pivoteo ¿Cuál es su conclusión?

Problema 1.14 Dado el sistema lineal de ecuaciones estructurado

$$\begin{pmatrix} a_{11} & a_{12} & & \cdots & a_{1\,n-1} & a_{1\,n} \\ 0 & a_{22} & & \cdots & a_{2\,n-1} & a_{2\,n} \\ 0 & 0 & \ddots & \vdots & \vdots & \vdots \\ \vdots & \vdots & & & & \\ 0 & \cdots & 0 & a_{n-2\,n-2} & a_{n-2\,n-1} & a_{n-2\,n} \\ a_{n-1\,1} & a_{n-1\,2} & \cdots & a_{n-1\,n-2} & a_{n-1\,n-1} & a_{n-1\,n} \\ a_{n\,1} & a_{n\,2} & \cdots & a_{n\,n-2} & a_{n\,n-1} & a_{n\,n} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix}$$

a) Diseñe un algoritmo basado en el método de eliminación gaussiana que permita resolver dicho sistema adaptado a la estructura particular de este, sin hacer ceros donde ya existen.

b) Realice un conteo de las operaciones (multiplicación, división, suma y resta) involucrados sólo en la triangulación del sistema en el algoritmo de la parte a).

Problema 1.15 Dado el sistema lineal $Ax = h$ con

$$A = \begin{pmatrix} a_1 & c_1 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 \\ b_2 & a_2 & c_2 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & b_3 & a_3 & c_3 & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 0 & b_4 & a_4 & c_4 & 0 & \cdots & 0 & 0 & 0 \\ & & & & & & & & & \\ \vdots & \vdots & \vdots & \ddots & \ddots & \ddots & \vdots & \vdots & \vdots & \vdots \\ & & & & & & & & & \\ 0 & 0 & 0 & 0 & & \cdots & & b_{n-1} & a_{n-1} & c_{n-1} \\ 0 & 0 & f_3 & f_4 & & \cdots & & f_{n-2} & b_n & a_n \end{pmatrix},$$

y $h \in \mathbb{R}^n$ un vector columna.

Diseñe un algoritmo en pseudo lenguaje basado en el método de eliminación gaussiana, sin elección de pivote, que utilice el mínimo número de localizaciones de memoria, realice el mínimo número de operaciones (es decir, que obvíe los elementos nulos de la matriz A (sin hacer ceros donde ya existen)) para resolver el sistema $Ax = h$.

Problema 1.16 Resuelva mediante el método de eliminación gaussiana con aritmética exacta el sistema

$$\begin{array}{rcrcrcr} 5x_1 & - & x_2 & & & = & 4 \\ -x_1 & + & 5x_2 & - & x_3 & = & -2 \\ & - & x_2 & + & 5x_3 & = & 1 \end{array}$$

Problema 1.17 Dado el sistema lineal

$$\begin{cases} x &= 0.9x + y + 17 \\ y &= 0.9y - 13 \end{cases}$$

1. Proponga un método iterativo convergente (en $\mathbb{R}^2$). La demostración de convergencia debe hacerse sin iterar.

2. Resuelva el sistema por su método propuesto, con aritmética decimal de redondeo a 3 dígitos y con una precisión de 10^{-3} en la norma $||\ ||_\infty$, eligiendo como punto de partida $(0,0)$. (el redondeo debe hacerse en cada operación).

3. Analice la convergencia del método de Jacobi y del método de Gauss–Seidel. Sin iterar.

Problema 1.18 Dado el sistema lineal

$$\begin{cases} 3x+y+z &= 5 \\ 3x+y-5z &= -1 \\ x+3y-z &= 1 \end{cases}$$

1. Resuelva el sistema usando eliminación gaussiana, con o sin pivoteo, con aritmética decimal de redondeo a 3 dígitos. (el redondeo debe hacerse en cada operación).

2. Analice la convergencia del método de Jacobi, sin iterar.

3. Analice la convergencia del método de Gauss–Seidel, sin iterar.

Problema 1.19 Considere el sistema de ecuaciones lineal

$$\begin{pmatrix} 1 & \frac{1}{2} & \frac{1}{3} \\ \frac{1}{2} & 1 & \frac{1}{6} \\ \frac{1}{3} & \frac{1}{6} & 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}.$$

1. Resuelva el sistema por el método de eliminación de Gauss (sin elección de pivote) redondeando cada operación a 4 dígitos en la mantisa.

2. Resuelva el sistema por el método de Jacobi redondeando cada operación a 4 dígitos en la mantisa, y con una presición de 10^{-3}, comenzando con $x_0 = y_0 = z_0 = 0$.

3. En este caso particular ¿Cuál de los dos métodos es más conveniente? Justifique.

Problema 1.20 Resuelva el sistema $Ax = b$, donde

$$A = \begin{pmatrix} -3 & 2 & 3 & -1 \\ 6 & -2 & -6 & 0 \\ -9 & 4 & 10 & 3 \\ 12 & -4 & -13 & -5 \end{pmatrix} \quad \text{y} \quad b = \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \end{pmatrix}$$

usando descomposición LU con pivoteo.

Problema 1.21 Considere la matriz

$$A = \begin{pmatrix} 0.0001 & 1 \\ 1 & 1 \end{pmatrix}$$

1. Trabajando con aritmética decimal de tres dígitos y redondeo ¿Es posible encontrar una descomposición LU para A?

2. Considere ahora aritmética decimal de cinco dígitos y redondeo. Encuentre una descomposición LU para A, y compare los números de condición de A y de LU ¿Cuál es su conclusión?

3. Ahora considere la matriz
$$A = \begin{pmatrix} 1 & 1 \\ 0.0001 & 1 \end{pmatrix}$$
¿Es posible ahora encontrar una descomposición LU, con aritmética decimal de redondeo a tres dígitos para A?

Problema 1.22 Considere el siguiente sistema de ecuaciones lineales
$$\begin{pmatrix} 10 & -3 & 6 \\ 1 & -8 & -2 \\ -2 & 4 & 9 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 25 \\ -9 \\ -50 \end{pmatrix}$$

1. Resuelva el sistema usando descomposición LU de Doolittle.

2. ¿Cuáles de los métodos iterativos: Richardson, Jacobi, Gauss–Seidel convergen?

3. Use un método iterativo convergente para encontrar la solución del sistema con una precisión de $\varepsilon = 10^{-3}$, con criterio de parada $||(x_n, y_n, z_n) - (x_T, y_T, z_T)||_\infty \leqslant \varepsilon$, donde (x_T, y_T, z_T) es la solución exacta del sistema.

Problema 1.23 Considere la matriz
$$A = \begin{pmatrix} 4 & -1 & -1 \\ -1 & 4 & 1 \\ -1 & 1 & 4 \end{pmatrix}$$

a) Demuestre que existe la descomposión de Cholesky (sin calcularla) de A.

b) Determine la descompisición de Cholesky de A.

Problema 1.24 a) Demuestre que la matriz
$$A = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 2 & 1 \\ 2 & 3 & 1 \end{pmatrix}$$
no tiene descomposición LU.

b) Realice pivoteo en la matriz A y demuestre que la matriz resultante tiene descomposición LU.

Problema 1.25 Considere la matriz
$$A = \begin{pmatrix} 2 & 1 & 0 \\ 0 & 3 & 2 \\ 1 & 2 & 4 \end{pmatrix}$$

1. Usando aritmética exacta, encuentre una descomposición $A = LU$, donde L es triangular inferior con "unos" en la diagonal.

2. Usando la descomposición LU obtenida en a), encuentre la solución del sistema (usando aritmética exacta) $Ax = b$, con $b^T = (5, -3, 8)$.

3. Proponga un método de punto fijo y demuestre la convergencia (sin iterar) para resolver el sistema dado en b), y encuentre una solución aproximada con una presición de 10^{-3}.

Problema 1.26 Consideremos el problema de calcular la temperatura de una barra metálica de largo L, cuyos extremos se mantienen a temperaturas constantes T_0 y T_L conocidas. La ecuación diferencial que gobierna este fenómeno es conocida como *ecuación de conducción del calor*, la cual, en régimen estacionario está dada por

$$-kT''(x) = f(x) \qquad 0 < x < L, \tag{1.63}$$

donde $f(x)$ es una función conocida, que representa una fuente de calor externa y $k > 0$ es una constante que se denomina *coeficiente de difusión o conductividad térmica.* A esta ecuación se le agregan las condiciones de borde:

$$T(0) = T_0, \qquad T(L) = T_L. \tag{1.64}$$

Una de las técnicas más usadas para resolver el problema 1.63-1.64 es el *método de diferencias finitas.* En este método, el intervalo $[0, L]$ se divide en $N + 1$ intervalos de largo $h = \frac{L}{N+1}$ y la solución es buscada en los puntos *internos* definidos por esta división, es decir, en los puntos $x_n = nh$, con $n = 1, \ldots, N$ (los valores de la temperatura en los nodos del borde $x_0 = 0$ y $x_{N+1} = L$ son conocidos). En este método, las derivadas de primer orden se aproximan por el *cuociente de diferencias finitas*

$$f'(x) \approx \frac{f(x+h) - f(x)}{h},$$

mientras que las derivadas de segundo orden, se aproximan por

$$f''(x) \approx \frac{f(x-h) - 2f(x) + f(x+h)}{h^2}. \tag{1.65}$$

Denotemos por T_n los valores aproximados de la evaluaciones de la temperatura exacta en los puntos de la malla, $T(x_n)$. Reemplazando $-T''(x)$ en la ecuación 1.63 por la expresión 1.65, y evaluando la ecuación en los *nodos internos* de la malla $\{x_1, \ldots, x_N\}$, se obtiene el siguiente sistema de ecuaciones lineales:

$$\begin{aligned} 2T_1 - T_2 &= h^2 f(x_1) + T_0, && (1.66)\\ -T_{j-1} + 2T_j - T_{j+1} &= h^2 f(x_j), \quad j = 2, \ldots, N-1; && (1.67)\\ 2T_{N-1} + 2T_N &= h^2 f(x_N) + T_L, && (1.68) \end{aligned}$$

que matricialmente puede ser escrito como

$$\begin{pmatrix} 2 & -1 & 0 & \cdots & 0 \\ -1 & 2 & -1 & \cdots & 0 \\ 0 & -1 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 2 & -1 \\ 0 & \cdots & 0 & -1 & 2 \end{pmatrix} \begin{pmatrix} T_1 \\ T_2 \\ \vdots \\ T_{N-1} \\ T_N \end{pmatrix} = h^2 \begin{pmatrix} f(x_1) \\ f(x_2) \\ \vdots \\ f(x_{N-1}) \\ f(x_N) \end{pmatrix} + \begin{pmatrix} T_0 \\ 0 \\ \vdots \\ 0 \\ T_L \end{pmatrix} \tag{1.69}$$

Denotemos por A la matriz y por F_h el lado derecho de este sistema.

Tomando como parámetros del problema los valores

$$\begin{array}{lll} k=1, & L=1, & f(x)=37\left(\frac{\pi}{2}\right)^2\sin(\frac{\pi}{2}x), \\ T_0=0, & T_L=37, & \end{array} \tag{1.70}$$

se pide resolver numéricamente el sistema 1.69, usando los métodos iterativos de Jacobi, Gauss-Seidel y SOR. Para el método SOR tome distintos valores del parámetro de aceleración $\omega \in (1,2)$. Calcule además el parámetro de aceleración óptimo ω^*.
Para ello, siga los siguientes pasos:

1. Compruebe que la matriz A es definida positiva.
2. Implemente cada uno de los 3 métodos antes mencionados y haga un estudio comparativo de estos para distintos valores de N. Por ejemplo, $N = 50, 100, 200, 1000$. Especifique el criterio de parada utilizado y el número de iteraciones para cada uno de los métodos y para los distintos valores de N. Presente sus resultados en una tabla.
3. Finalmente, determine la solución analítica del problema 1.63-1.64, con los parámetros dados por 1.70 y grafique el error absoluto en la solución, para cada uno de los métodos y para los distintos valores de N . A partir de estos gráficos, comente cual de los tres métodos aproxima mejor a la solución .

Problema 1.27 Demuestre que la matriz no singular

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

no tiene descomposición LU, pero la matriz $A-I$ si tiene descomposición LU. Dar una matriz de permutación P de modo que PA tiene descomposición LU.

Problema 1.28 Sea $a>0$ una constante. Se tiene la siguiente matriz

$$A=\begin{pmatrix} 2\,a & 0 & \frac{2}{3}\,a^3 \\ 0 & \frac{2}{3}\,a^3 & 0 \\ \frac{2}{3}\,a^3 & 0 & \frac{2}{5}\,a^5 \end{pmatrix}$$

(a) Demuestre que A tiene descomposición de Cholesky y usela para calcular A^{-1}.

(b) calcule el número de condición de la matriz A, en términos de A y determine cuando ella está bien condicionada.

(c) Si $\mathbf{b}=(0\ -1\ 1)^T$. Usando la descomposición anterior resuelva el sistema de ecuaciones lineales $A\mathbf{x}=\mathbf{b}$.

Problema 1.29

Problema 1.30

Problema 1.31