CAPÍTULO 3. SOLUCIÓN NUMÉRICA DE SISTEMAS DE ECUACIONES

INTRODUCCIÓN

Un sistema de **n**-ecuaciones (con coeficientes reales) en las **n**-incógnitas $x_1, x_2, \ldots, x_n$ es un conjunto de **n** ecuaciones de la forma

$$\begin{cases} f_1(x_1, x_2, \ldots, x_n) = 0 \\ f_2(x_1, x_2, \ldots, x_n) = 0 \\ \vdots \\ f_n(x_1, x_2, \ldots, x_n) = 0 \end{cases} \qquad \textbf{(3.1)}$$

donde

$$f_i : \mathbf{D}_i \rightarrow \mathbf{R}, \quad \mathbf{D}_i \subseteq \mathbf{R}^n$$
$$X = (x_1, x_2, \ldots, x_n) \rightarrow f_i(X) = y$$

Si para cada $i = 1,2,\ldots,n$, la función f_i es de la forma

$$f_i(x_1, x_2, \ldots, x_n) = a_{i1}x_1 + a_{i2}x_2 + \ldots + a_{in}x_n - b_i$$

con $a_{i1}, a_{i2}, \ldots, a_{in}$ y b_i constantes reales, el sistema se dice **lineal** (con coeficientes reales); en cualquier otro caso el sistema se dice **no-lineal**.

Si $C = (c_1, c_2, \ldots, c_n) \in \mathbf{R}^n$ es tal que $f_i(c_1, c_2, \ldots, c_n) = 0$ para cada $i = 1,2,\ldots,n$, entonces se dice que C es una **solución real** del sistema (3.1).

El objetivo de este capítulo es estudiar algunos métodos numéricos para encontrar una solución real de un sistema del tipo (3.1).

3.1 SOLUCIÓN NUMÉRICA DE SISTEMAS DE ECUACIONES LINEALES

Un sistema de **n**-ecuaciones lineales (con coeficientes reales) en las **n**-incógnitas $x_1, x_2, \ldots, x_n$ puede escribirse en la forma

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n = b_1 \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n = b_2 \\ \vdots \\ a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{nn}x_n = b_n \end{cases} \quad \text{con } a_{ij},\ b_i \in \mathbf{R},\ i,j = 1,2,\ldots,n \qquad \textbf{(3.2)}$$

El sistema (3.2) puede escribirse en la forma matricial equivalente $AX = b$ con

$$A=\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix}, \quad X=\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \quad \text{y} \quad b=\begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix}$$

La matriz A es llamada matriz de coeficientes del sistema, el vector columna X el vector de incógnitas y el vector b el vector de términos independientes.

Nota: Consideraremos únicamente sistemas de ecuaciones lineales $AX=b$ con $A\in\mathbf{R}_{n\times n}$ que tengan solución única para cada vector $b\in\mathbf{R}^n$, es decir, con A invertible.

Los métodos numéricos que estudiaremos para la solución de un sistema de ecuaciones lineales se clasifican en dos tipos: **directos** e **iterativos**.

Los **métodos directos** nos proporcionan una solución del sistema en un número finito de pasos. Si usamos aritmética finita para los cálculos, obtendremos por lo general una solución aproximada, debido únicamente a los errores de redondeo, puesto que no hay errores de truncamiento o de fórmula. Los métodos directos más usados tienen como base la **eliminación de Gauss**.

En los **métodos iterativos** se parte de una aproximación inicial a la solución del sistema dado y se genera, a partir de dicha aproximación, una sucesión de vectores que si converge lo hace a la solución del sistema. Al igual que en el capítulo 2, tendremos fórmulas para calcular los términos de la sucesión, así que en general no se espera calcular el límite de la sucesión, por lo que debemos tomar algún término de la sucesión como una solución aproximada del sistema. Esta vez, además de los errores de redondeo si se usa aritmética finita, habrá errores de truncamiento o de fórmula. Los métodos iterativos más simples y conocidos están basados en **iteraciones de Punto Fijo**.

3.2 MÉTODOS DIRECTOS

CASO 1: La matriz A (de coeficientes del sistema $AX=b$) es triangular (superior o inferior) con todas sus componentes sobre la diagonal principal no-nulas.

Supongamos que el sistema es de la forma

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \ \ldots\ + a_{1i}x_i + \ \ldots\ + a_{1,n-1}x_{n-1} + a_{1n}x_n = b_1 \\ a_{22}x_2 + \ \ldots\ + a_{2i}x_i + \ \ldots\ + a_{2,n-1}x_{n-1} + a_{2n}x_n = b_2 \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \vdots \\ a_{ii}x_i + \ \ldots\ + a_{i,n-1}x_{n-1} + a_{in}x_n = b_i \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \vdots \\ a_{n-1,n-1}x_{n-1} + a_{n-1,n}x_n = b_{n-1} \\ a_{nn}x_n = b_n \end{cases}$$

Como $a_{nn}\neq 0$, entonces podemos despejar x_n de la última ecuación y obtenemos

$$x_n = \frac{b_n}{a_{n\,n}}$$

Conocido x_n, usamos la penúltima ecuación para obtener

$$x_{n-1} = \frac{b_{n-1} - a_{n-1,\,n} x_n}{a_{n-1,\,n-1}}$$

Conocidos x_n y x_{n-1}, obtenemos (de la antepenúltima ecuación)

$$x_{n-2} = \frac{b_{n-2} - \left(a_{n-2,\,n-1} x_{n-1} + a_{n-2,\,n} x_n\right)}{a_{n-2,\,n-2}}$$

En general, conocidos $x_n, x_{n-1}, \ldots, x_{i+1}$, obtenemos

$$x_i = \frac{b_i - \sum_{k=i+1}^{n} a_{i\,k} x_k}{a_{i\,i}}, \quad i = n-1, n-2, \ldots, 1$$

El método anterior para determinar la solución del sistema se denomina **sustitución reversiva**, **regresiva** o **hacia atrás**.

Si la matriz de coeficientes del sistema es triangular inferior, para resolver el sistema podemos proceder de manera similar al caso anterior, pero empezando por despejar x_1 de la primera ecuación. El procedimiento en este caso se denomina **sustitución progresiva** o **hacia adelante**.

Algoritmo 3.1 (Sustitución regresiva) Para encontrar una solución aproximada $\tilde{X}$ de un sistema triangular superior $AX = b$ con $A = \left(a_{i\,j}\right)_{n \times n}$ invertible.

Entrada: El orden n del sistema; los coeficientes $a_{i\,j}$, $i = 1,2,\ldots,n$, $j = i,\ldots,n$; los términos independientes b_i, $i = 1,2,\ldots,n$.

Salida: Una solución aproximada $\tilde{X} = \left(x_1, x_2, \ldots, x_n\right)$.

Paso 1: Tomar $x_n = \frac{b_n}{a_{n\,n}}$.

Paso 2: Para $i = n-1, n-2, \ldots, 1$, tomar

$$x_i = \frac{b_i - \sum_{k=i+1}^{n} a_{i\,k} x_k}{a_{i\,i}}$$

Paso 3: Salida: "Una solución aproximada del sistema es $\tilde{X} = \left(x_1, x_2, \ldots, x_n\right)$".
Terminar.

CASO 2: La matriz A (de coeficientes del sistema $AX = b$) es tal que no se requieren intercambios de filas para culminar con éxito la eliminación Gaussiana.

Digamos que el sistema $AX = b$ tiene la forma

$$\begin{cases} E_1 : a_{11}x_1 + \cdots + a_{1j}x_j + \cdots + a_{1n}x_n = b_1 \\ E_2 : a_{21}x_1 + \cdots + a_{2j}x_j + \cdots + a_{2n}x_n = b_2 \\ \quad \vdots \qquad\qquad\qquad\qquad\qquad\quad \vdots \\ E_j : a_{j1}x_1 + \cdots + a_{jj}x_j + \cdots + a_{jn}x_n = b_j \\ \quad \vdots \qquad\qquad\qquad\qquad\qquad\quad \vdots \\ E_i : a_{i1}x_1 + \cdots + a_{ij}x_j + \cdots + a_{in}x_n = b_i \\ \quad \vdots \qquad\qquad\qquad\qquad\qquad\quad \vdots \\ E_n : a_{n1}x_1 + \cdots + a_{nj}x_j + \cdots + a_{nn}x_n = b_n \end{cases}$$

El proceso de **eliminación Gaussiana** (simple) consiste en lo siguiente:

i) Eliminamos el coeficiente de x_1 en cada una de las ecuaciones $E_2, E_3, \ldots, E_n$ para obtener un sistema equivalente $A^{(1)}X = b^{(1)}$, realizando las operaciones elementales

$$\left(E_i - \left(\frac{a_{i1}}{a_{11}} \right) E_1 \right) \to E_i^{(1)}, \quad i = 2,3,\ldots,n$$

ii) Eliminamos el coeficiente de x_2 en cada una de las ecuaciones $E_3^{(1)}, E_4^{(1)}, \ldots, E_n^{(1)}$, para obtener un sistema equivalente $A^{(2)}X = b^{(2)}$, realizando las operaciones elementales

$$\left(E_i^{(1)} - \left(\frac{a_{i2}^{(1)}}{a_{22}^{(1)}} \right) E_2^{(1)} \right) \to E_i^{(2)}, \quad i = 3,4,\ldots,n$$

(debe ocurrir que $a_{22}^{(1)} \neq 0$).

iii) En general, eliminados los coeficientes de $x_1, x_2, \ldots, x_{j-1}$, **eliminamos** el coeficiente de x_j en cada una de las ecuaciones $E_{j+1}^{(j-1)}, E_{j+2}^{(j-1)}, \ldots, E_n^{(j-1)}$, para obtener un sistema equivalente $A^{(j)}X = b^{(j)}$, realizando las operaciones elementales

$$\left(E_i^{(j-1)} - \left(\frac{a_{ij}^{(j-1)}}{a_{jj}^{(j-1)}} \right) E_j^{(j-1)} \right) \to E_i^{(j)}, \quad i = j+1,\ldots,n$$

(debe ocurrir que $a_{jj}^{(j-1)} \neq 0$).

Los números

$$m_{ij} = \frac{a_{ij}^{(j-1)}}{a_{jj}^{(j-1)}}, \quad j = 1,\ldots,n-1, \quad i = j+1,\ldots,n$$

se llaman **multiplicadores** (si $j = 1$, $\frac{a_{i1}^{(0)}}{a_{11}^{(0)}} \equiv \frac{a_{i1}}{a_{11}}$) .

El sistema (reducido) resultante tendrá la forma

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1j}x_j + \ldots + a_{1,n-1}x_{n-1} + a_{1n}x_n = b_1 \\ a_{22}^{(1)}x_2 + \ldots + a_{2j}^{(1)}x_j + \ldots + a_{2,n-1}^{(1)}x_{n-1} + a_{2n}^{(1)}x_n = b_2^{(1)} \\ \vdots \\ a_{jj}^{(j-1)}x_j + \ldots + a_{j,n-1}^{(j-1)}x_{n-1} + a_{jn}^{(j-1)}x_n = b_j^{(j-1)} \\ \vdots \\ a_{n-1,n-1}^{(n-2)}x_{n-1} + a_{n-1,n}^{(n-2)}x_n = b_{n-1}^{(n-2)} \\ a_{nn}^{(n-1)}x_n = b_n^{(n-1)} \end{cases}$$

el cual se resuelve por el método de sustitución regresiva, para obtener la solución del sistema original.

CASO 3: La matriz A **(de coeficientes del sistema** $AX = b$**) es tal que se requieren intercambios de filas para culminar con éxito el proceso de eliminación Gaussiana.**

Procedemos exactamente como en el caso 2, solo que cuando encontremos $a_{jj}^{(j-1)} = 0$ para algún $j = 1,2,\ldots,n-1$ (recuerde que si $j = 1$, $a_{jj}^{(j-1)} = a_{11}^{(0)} \equiv a_{11}$), continuamos de la siguiente manera:

Se busca en la j-ésima columna de $A^{(j-1)}$ (si $j = 1$, $A^{(j-1)} = A^{(0)} \equiv A$) desde la fila $(j+1)$-ésima hasta la n-ésima, el **primer** elemento distinto de cero (Por qué debe existir tal elemento?). Si $a_{kj}^{(j-1)} \neq 0$ es tal elemento, entonces se efectúa la operación elemental

$$E_j^{(j-1)} \leftrightarrow E_k^{(j-1)} : \text{intercambio de las ecuaciones j-ésima y k-ésima}$$

y se continua con el proceso de eliminación Gaussiana.

Una vez que se ha hecho la eliminación Gaussiana completa, se realiza la sustitución regresiva para obtener la solución única del sistema dado.

Los procedimientos descritos anteriormente quedan incluidos en el siguiente algoritmo, en el cual incluso la matriz A puede ser no invertible (singular) .

Algoritmo 3.2 (**Eliminación Gaussiana con sustitución regresiva**) Para obtener una solución aproximada $\tilde{X}$ de un sistema de la forma

$$\begin{cases} E_1: a_{11}x_1 + \ \ldots \ + a_{1n}x_n = b_1 \\ E_2: a_{21}x_1 + \ \ldots \ + a_{2n}x_n = b_2 \\ \quad \vdots \qquad\qquad\qquad\quad \vdots \\ E_n: a_{n1}x_1 + \ \ldots \ + a_{nn}x_n = b_n \end{cases}$$

Entrada: El orden n del sistema; las componentes a_{ij}, $i = 1,2,\ldots,n$, $j = 1,2,\ldots,n+1$ de la matriz aumentada $(A : b)$ con $a_{i,n+1} = b_i$, $i = 1,2,\ldots,n$.

Salida: Una solución aproximada $\tilde{X} = (x_1, x_2, \ldots, x_n)$ del sistema dado o un mensaje.

Paso 1: Para $j = 1,2,\ldots,n-1$, seguir los pasos 2-4 **(Proceso de eliminación):**

Paso 2: Hallar el menor entero k tal que $j \le k \le n$ y $a_{kj} \neq 0$ (a_{kj} es el contenido en la posición de memoria (k, j) en ese momento).

Si **no** existe tal k, entonces A **no** es invertible, por tanto, **salida:** "El sistema no tiene solución única". **Terminar**.

Paso 3: Si existe tal k y $k \neq j$, hacer

$E_j \leftrightarrow E_k$ (intercambio de las filas j-ésima y k-ésima)

Paso 4: Para $i = j+1,\ldots,n$, seguir los pasos 5 y 6:

Paso 5: Tomar $m_{ij} = \dfrac{a_{ij}}{a_{jj}}$.

Paso 6: Efectuar $(E_i - m_{ij}E_j) \to E_i$.

(Hasta aquí llega la eliminación Gaussiana)

Paso 7: Si $a_{nn} = 0$, entonces, **salida:** "El sistema no tiene solución única". **Terminar**.

Paso 8: Tomar $x_n = \dfrac{a_{n,n+1}}{a_{nn}}$ **(Aquí empieza la sustitución regresiva)**.

Paso 9: Para $i = n-1,\ldots,1$ tomar

$$x_i = \frac{a_{i,n+1} - \sum_{k=i+1}^{n} a_{ik}x_k}{a_{ii}}$$

Paso 10: Salida: "Una solución aproximada del sistema es $\tilde{X} = (x_1, x_2, \ldots, x_n)$".

Terminar.

Hay sistemas de ecuaciones lineales, como vimos en el capítulo 1, que son sensibles a pequeños cambios en los datos; de tales **sistemas** decimos que están **mal condicionados**.

En la práctica, por lo general, cuando se requiere resolver un sistema $AX = b$, asociado con un problema, los datos (coeficientes y términos independientes) no se conocen de manera exacta, debido por ejemplo a errores de medición, es decir, se dispone realmente de un sistema perturbado. Por otra parte, aunque los datos se conozcan de manera exacta, éstos al ser entrados al computador serán transformados (por el compilador) en números de máquina, lo que sabemos introduce errores de redondeo. En cualquier caso, interesa saber si tales errores pueden afectar de manera significativa la solución del problema. Una manera de estudiar estos comportamientos es a través del **número de condición** de la matriz de coeficientes del sistema.

3.3 SISTEMAS MAL CONDICONADOS Y NÚMERO DE CONDICIÓN DE UNA MATRIZ

Para llegar a la idea del número de condición de una matriz empecemos considerando el siguiente ejemplo que muestra dos sistemas de ecuaciones lineales mal condicionados.

Ejemplo 3.1 Consideremos los siguientes sistemas de ecuaciones lineales

$$\begin{cases} x + y = 2 \\ 10.05x + 10y = 21 \end{cases} \qquad \textbf{(3.3)}$$

y

$$\begin{cases} 4.1x + 2.8y = 4.1 \\ 9.7x + 6.6y = 9.7 \end{cases} \qquad \textbf{(3.4)}$$

En el capítulo 1, vimos que la solución exacta del sistema (3.3) es $X_1 = \begin{pmatrix} 20 \\ -18 \end{pmatrix}$ y si cambiamos el coeficiente **10.05** por **10.1** (un cambio relativo de aproximadamente **.5%**), la solución exacta del sistema perturbado

$$\begin{cases} x + y = 2 \\ 10.1x + 10y = 21 \end{cases} \qquad \textbf{(3.3')}$$

es $\tilde{X}_1 = \begin{pmatrix} 10 \\ -8 \end{pmatrix}$, que muestra un cambio relativo del **50%** en el valor de x y de aproximadamente el **56%** en el valor de y.

Análogamente, el sistema (3.4) tiene solución exacta $X_2 = \begin{pmatrix} 1.0 \\ 0.0 \end{pmatrix}$ y si cambiamos el término independiente **4.1** por **4.11** (un cambio relativo aproximado de **.2%** en el término independiente), la solución exacta del sistema perturbado

$$\begin{cases} 4.1x + 2.8y = 4.11 \\ 9.7x + 6.6y = 9.7 \end{cases} \qquad \textbf{(3.4')}$$

es $\tilde{X}_2 = \begin{pmatrix} .34 \\ .97 \end{pmatrix}$, que muestra un cambio relativo aproximado de **66%** en el valor de x. ♦

Se observa entonces que un cambio "pequeño" en uno de los datos (coeficientes y términos independientes) ha producido un cambio "grande" en la solución, es decir, la solución del sistema perturbado es "muy diferente" de la solución del sistema original.

Los anteriores son ejemplos de problemas **mal condicionados**.

Un problema se dice **bien condicionado** si "pequeños" cambios en los datos introducen, correspondientemente, un cambio "pequeño" en la solución. El buen o mal condicionamiento de un problema es inherente al problema y no depende del algoritmo empleado para resolverlo.

El mal condicionamiento en el sistema (3.3) puede visualizarse gráficamente, al graficar las dos rectas: L_1: $x + y = 2$ y L_2: $10.05x + 10y = 21$. Como las pendientes de estas dos rectas son casi iguales, es difícil ver exactamente dónde se cortan, esta dificultad visual, digamos que se mide cuantitativamente en los resultados numéricos obtenidos.

Observe que si A es la matriz de coeficientes del sistema (3.3), entonces $\det A = -.05$ y se puede pensar que el mal condicionamiento está relacionado con el tamaño del determinante de la matriz de coeficientes, pero recuerde que si una ecuación de un sistema se multiplica por un escalar, el determinante de la matriz de coeficientes queda multiplicado por ese escalar mientras los dos sistemas siguen teniendo exactamente las mismas soluciones, es decir, son equivalentes.

El objetivo siguiente es desarrollar una teoría que permita estudiar el condicionamiento de un sistema lineal $AX = b$.

Empezamos con la siguiente definición:

Definición 3.1 Si X es la solución exacta de un sistema lineal $AX = b$, A invertible, $b \neq 0$, y $\tilde{X}$ es una solución aproximada de dicho sistema, entonces llamamos **vector error** de $\tilde{X}$ con respecto a X al vector E definido por

$$\boxed{E = \tilde{X} - X}$$

y **vector error residual** correspondiente a la solución aproximada $\tilde{X}$, al vector R definido por

$$\boxed{R = A\tilde{X} - b}$$

Observe que E usualmente no se conoce (pues X **no** se conoce), mientras que R siempre puede conocerse.

Como $R = A\tilde{X} - b$, entonces R mide hasta dónde la solución aproximada $\tilde{X}$ satisface el sistema $AX = b$. Observe que $\tilde{X}$ es tal que $A\tilde{X} = R + b$, es decir, $\tilde{X}$ es solución de una perturbación del sistema $AX = b$.

Nótese que $R = 0$ implica $\tilde{X} = X$, es decir, $R = 0$ implica $E = 0$. Será que $\| R \|$ "pequeña" implica $\| E \|$ también "pequeña", donde $\| \cdot \|$ es alguna norma vectorial?

Empecemos recordando qué es una norma vectorial y qué es una norma matricial.

Definición 3.2 Una **norma vectorial** en $\mathbf{R}^n$ es una función

$$\begin{aligned} \|.\|: \quad & \mathbf{R}^n \to \mathbf{R} \\ & X \to \|X\| \end{aligned}$$

tal que para todo $X, Y \in \mathbf{R}^n$ y todo $\alpha \in \mathbf{R}$:

i) $\|X\| \geq 0, \quad \|X\| = 0$ si y sólo si $X = 0$

ii) $\|\alpha X\| = |\alpha| \, \|X\|$

iii) $\|X + Y\| \leq \|X\| + \|Y\|$. Ñ

Ejemplo 3.2 Las siguientes son algunas normas vectoriales en $\mathbf{R}^n$. Si $X = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \in \mathbf{R}^n$, entonces

1) La **norma euclidiana** (o **norma 2**) definida por

$$\boxed{\|X\|_2 = \left(\sum_{i=1}^{n} x_i^2 \right)^{\frac{1}{2}}}$$

2) La **norma suma** (o **norma 1**) definida por

$$\boxed{\|X\|_1 = \sum_{i=1}^{n} |x_i|}$$

3) La **norma del máximo** (o **norma** ∞) definida por

$$\boxed{\|X\|_\infty = \max_{1 \leq i \leq n} |x_i|}$$

Estas normas en $\mathbf{R}^n$, inducen las siguientes nociones de **distancia** entre dos vectores $X, Y \in \mathbf{R}^n$:

1) $d_2(X,Y) = \|X - Y\|_2 = \left(\sum_{i=1}^{n} (x_i - y_i)^2 \right)^{\frac{1}{2}}$ (distancia asociada con la norma euclidiana).

2) $d_1(X,Y) = \|X - Y\|_1 = \sum_{i=1}^{n} |x_i - y_i|$ (distancia asociada con la norma suma)

3) $d_\infty(X,Y) = \|X - Y\|_\infty = \max_{1 \leq i \leq n} |x_i - y_i|$ (distancia asociada con la norma del máximo). ♦

Definición 3.3 Una **norma matricial** en $\mathbf{R}_{n \times n}$ es una función:

$$\begin{aligned} \|.\|: \ \mathbf{R}_{n\times n} &\to \mathbf{R} \\ A &\to \|A\| \end{aligned}$$

tal que para todo $A,B \in \mathbf{R}_{n\times n}$ y todo $\alpha \in \mathbf{R}$:

i) $\|A\| \geq 0$, $\|A\| = 0$ si y sólo si $A = 0$

ii) $\|\alpha A\| = |\alpha| \, \|A\|$

iii) $\|A+B\| \leq \|A\| + \|B\|$

iv) $\|AB\| \leq \|A\| \, \|B\|$. Ñ

Aunque hay diversas formas de construir normas matriciales, aquí solamente consideraremos las normas matriciales que serán obtenidas a partir de las normas vectoriales dadas en el ejemplo 3.2 según se indica en el siguiente teorema, teorema cuya demostración puede ser consultada en Kincaid 1972, páginas 163 y 164.

Teorema 3.1 Sea $\|.\|$ cualquier norma vectorial en $\mathbf{R}^n$. Entonces la función $\|.\|$ de $\mathbf{R}_{n\times n}$ en $\mathbf{R}$, definida por

$$\boxed{\|A\| = \underset{X\neq 0}{\text{Max}} \frac{\|AX\|}{\|X\|}, \ A \in \mathbf{R}_{n\times n}} \qquad \textbf{(3.5)}$$

es una norma matricial en $\mathbf{R}_{n\times n}$. Ñ

La norma matricial dada por (3.5) se dirá la **norma matricial inducida** por la correspondiente norma vectorial $\|.\|$.

Note que (3.5) implica que

$$\boxed{\|AX\| \leq \|A\| \, \|X\|} \qquad \textbf{(3.6)}$$

para cada $X \in \mathbf{R}^n$ y cada $A \in \mathbf{R}_{n\times n}$, pues si $X \in \mathbf{R}^n$, $X \neq 0$, entonces

$$\frac{\|AX\|}{\|X\|} \leq \underset{X\neq 0}{\text{Max}} \frac{\|AX\|}{\|X\|} = \|A\|$$

Para $X = 0$ claramente se satisface (3.6). Ñ

Nótese, además, que

$$\boxed{\|A\| = \underset{X\neq 0}{\text{Max}} \frac{\|AX\|}{\|X\|} = \underset{X\neq 0}{\text{Max}} \left\| A \frac{X}{\|X\|} \right\| = \underset{\|Z\|=1}{\text{Max}} \|AZ\|}$$

Las normas matriciales inducidas por las normas vectoriales $\|\cdot\|_2$, $\|\cdot\|_1$ y $\|\cdot\|_\infty$ son:

1) $\| A \|_2 = \underset{\| X \|_2 = 1}{\text{Max}} \| AX \|_2$, difícil de calcular con la información que se conoce hasta aquí, pues calcular esta norma es resolver un problema de máximo en varias variables.

2) $\| A \|_1 = \underset{\| X \|_1 = 1}{\text{Max}} \| AX \|_1$, fácil de calcular, ya que se puede demostrar que

$$\boxed{\| A \|_1 = \underset{1 \le j \le n}{\text{Max}} \sum_{i=1}^{n} | a_{ij} |}$$

3) $\| A \|_\infty = \underset{\| X \|_\infty = 1}{\text{Max}} \| AX \|_\infty$, fácil de calcular, ya que como en el caso 2), se puede demostrar, véase Burden 1985, páginas 453 y 454, que

$$\boxed{\| A \|_\infty = \underset{1 \le i \le n}{\text{Max}} \sum_{j=1}^{n} | a_{ij} |}$$

Debido a la facildad del cálculo de las normas $\| \cdot \|_1$ y $\| \cdot \|_\infty$, las usaremos en lo que sigue.

Una **distancia** entre las matrices $A, B \in \mathbf{R}_{n \times n}$ se puede definir como $d(A,B) = \| A - B \|$, donde $\| \cdot \|$ es cualquier norma matricial.

Definición 3.4 El **radio espectral** de una matriz $A \in \mathbf{R}_{n \times n}$, $\rho(A)$, se define como

$$\rho(A) = \text{Max}\{ | \lambda | \ / \ \lambda \text{ es valor propio de } A \}$$

Recuerde que si λ es un número complejo, digamos $\lambda = \alpha + i\beta$ con α y β en $\mathbf{R}$, entonces $| \lambda | = | \alpha + i\beta | = \sqrt{\alpha^2 + \beta^2}$.

El siguiente teorema, cuya demostración puede ser consultada en Ortega 1990, páginas 21 y 22, relaciona el radio espectral de una matriz A con $\| A \|_2$.

Teorema 3.2 Si $A \in \mathbf{R}_{n \times n}$, entonces

i) $\sqrt{\rho(A^T A)} = \| A \|_2$, y en consecuencia, si A es simétrica $\rho(A) = \| A \|_2$.

ii) $\rho(A) \le \| A \|$ para cualquier norma matricial inducida. Ñ

Con respecto al ejemplo 3.1, tenemos:

Para el sistema (3.3), $X_1 = \begin{pmatrix} 20 \\ -18 \end{pmatrix}$ es su solución exacta y si consideramos como una solución aproximada a $\tilde{X}_1 = \begin{pmatrix} 10 \\ -8 \end{pmatrix}$, que es la solución exacta del sistema perturbado (3.3'), entonces el vector **error** de $\tilde{X}_1$ con respecto a X_1, es

$$E_1 = \tilde{X}_1 - X_1 = \begin{pmatrix} -10 \\ 10 \end{pmatrix}$$

y el vector **error residual** correspondiente a la solución aproximada $\tilde{X}_1$, es

$$R_1 = A\tilde{X}_1 - b = \begin{pmatrix} 1 & 1 \\ 10.05 & 10 \end{pmatrix}\begin{pmatrix} 10 \\ -8 \end{pmatrix} - \begin{pmatrix} 2 \\ 21 \end{pmatrix} = \begin{pmatrix} 2 \\ 20.5 \end{pmatrix} - \begin{pmatrix} 2 \\ 21 \end{pmatrix} = \begin{pmatrix} 0 \\ -.5 \end{pmatrix}$$

Entonces

$$\|E_1\|_1 = |-10| + |10| = 20, \quad \|R_1\|_1 = |0| + |-.5| = .5$$

$$\|E_1\|_2 = \sqrt{(-10)^2 + 10^2} \approx 14.14, \quad \|R_1\|_2 = \sqrt{(-.5)^2} = .5$$

$$\|E_1\|_\infty = \text{Max}\{|-10|, |10|\} = 10, \quad \|R_1\|_\infty = \text{Max}\{|0|, |-.5|\} = .5$$

así que un vector error residual "pequeño" (relativo al vector de términos independientes $b = \begin{pmatrix} 2 \\ 21 \end{pmatrix}$, $\|b\|_1 = 23$, $\|b\|_2 \approx 21.095$, $\|b\|_\infty = 21$) corresponde a un vector error relativamente "grande".

Para el sistema (3.4), $X_2 = \begin{pmatrix} 1.0 \\ 0.0 \end{pmatrix}$ es la solución exacta y si consideramos como una solución aproximada a $\tilde{X}_2 = \begin{pmatrix} .34 \\ .97 \end{pmatrix}$, que es la solución exacta del sistema perturbado (3.4'), entonces el vector **error** de $\tilde{X}_2$ con respecto a X_2, es

$$E_2 = \begin{pmatrix} .34 \\ .97 \end{pmatrix} - \begin{pmatrix} 1.0 \\ 0 \end{pmatrix} = \begin{pmatrix} -.66 \\ .97 \end{pmatrix}$$

y el vector **error residual** correspondiente a la solución aproximada $\tilde{X}_2$, es

$$R_2 = \begin{pmatrix} 4.11 \\ 9.7 \end{pmatrix} - \begin{pmatrix} 4.1 \\ 9.7 \end{pmatrix} = \begin{pmatrix} .01 \\ 0 \end{pmatrix}$$

Como

$$\|E_2\|_1 = 1.63, \quad \|R_2\|_1 = .01; \quad \|E_2\|_2 \approx 1.17, \quad \|R_2\|_2 = .01; \quad \|E_2\|_\infty = .97, \quad \|R_2\|_\infty = .01$$

entonces, nuevamente, un vector error residual "pequeño" **no** corresponde a un vector error "pequeño". ♦

El ejemplo anterior pone de manifiesto que $\|R\|$ "pequeño", **no** necesariamente implica que $\|E\|$ también sea "pequeño". Sin embargo, a partir del siguiente teorema podremos probar que, satisfecha cierta condición, $\frac{\|R\|}{\|b\|}$ "pequeño" implica $\frac{\|E\|}{\|X\|}$ también "pequeño".

Teorema 3.3 Sea $A \in \mathbf{R}_{n\times n}$ una matriz no-singular y X la solución exacta del sistema $AX = b,\ b \neq 0$. Si $\tilde{X}$ es una solución aproximada del sistema $AX = b$, entonces para cualquier norma matricial inducida se tiene que

$$\boxed{\frac{\|R\|}{\|b\|}\frac{1}{\|A\|\|A^{-1}\|} \leq \frac{\|E\|}{\|X\|} \leq \|A\|\|A^{-1}\|\frac{\|R\|}{\|b\|}} \qquad \textbf{(3.7)}$$

Demostración: Como $R = A\tilde{X} - b = A\tilde{X} - AX = A(\tilde{X} - X) = AE$, y A es invertible, entonces

$E = A^{-1}R,\quad b = AX,\quad X = A^{-1}b$ y aplicando la desigualdad (3.6), se obtiene

$$\|R\| \leq \|A\|\|E\|, \text{ es decir, } \frac{\|R\|}{\|A\|} \leq \|E\|, \text{ y } \|E\| \leq \|A^{-1}\|\|R\|$$

de donde

$$\frac{\|R\|}{\|A\|} \leq \|E\| \leq \|A^{-1}\|\|R\| \qquad \textbf{(3.8)}$$

Aplicando la misma desigualdad (3.6), se tiene que

$$\|b\| \leq \|A\|\|X\|, \text{ es decir, } \frac{\|b\|}{\|A\|} \leq \|X\|, \text{ y } \|X\| \leq \|A^{-1}\|\|b\|$$

de donde

$$\frac{\|b\|}{\|A\|} \leq \|X\| \leq \|A^{-1}\|\|b\|$$

o equivalentemente

$$\frac{1}{\|A^{-1}\|\|b\|} \leq \frac{1}{\|X\|} \leq \frac{\|A\|}{\|b\|} \qquad \textbf{(3.9)}$$

Combinando (3.8) y (3.9), obtenemos las siguientes cotas para el **error relativo**, $\frac{\|E\|}{\|X\|}$, en términos del **error residual relativo**, $\frac{\|R\|}{\|b\|}$:

$$\frac{\|R\|}{\|b\|}\frac{1}{\|A\|\|A^{-1}\|} \leq \frac{\|E\|}{\|X\|} \leq \|A\|\|A^{-1}\|\frac{\|R\|}{\|b\|} \qquad \textbf{(3.10)}$$

que era lo que quería demostrarse. Ñ

De acuerdo con este teorema 3.3, si se satisface la condición $\| A \| \| A^{-1} \| \approx 1$, entonces $\frac{\| R \|}{\| b \|}$ y $\frac{\| E \|}{\| X \|}$ son más o menos del mismo tamaño. Así que si $\frac{\| R \|}{\| b \|}$ es "pequeño", también lo será $\frac{\| E \|}{\| X \|}$, y si $\frac{\| R \|}{\| b \|}$ es "grande", también lo será $\frac{\| E \|}{\| X \|}$; por lo tanto si $\| A \| \| A^{-1} \| \approx 1$, podremos distinguir una solución aproximada, $\tilde{X}$, buena de una mala observando el error residual relativo $\frac{\| R \|}{\| b \|}$.

El número $Cond(A) = \| A \| \| A^{-1} \|$ se llamará **NÚMERO DE CONDICIÓN** o **CONDICIONAL** de la matriz no-singular A, relativo a la norma matricial usada. Aunque el valor de $Cond(A)$ depende de la norma matricial usada; sin embargo $Cond(A) \geq 1$, cualquiera sea la norma matricial inducida, pues

$$I_n = AA^{-1}, \quad \| I_n \| \leq \| A \| \| A^{-1} \| \quad \text{y} \quad \| I_n \| = \underset{X \neq 0}{Max} \left\| I_n \frac{X}{\| X \|} \right\| = \underset{X \neq 0}{Max} \frac{\| X \|}{\| X \|} = 1$$

De acuerdo con la relación (3.7), dada en el teorema 3.3, vemos que si $Cond(A) \approx 1$, entonces el error relativo, $\frac{\| E \|}{\| X \|}$, y el error residual relativo, $\frac{\| R \|}{\| b \|}$, son más o menos del mismo tamaño y podremos distinguir una solución aproximada "buena" de una "mala" observando el error residual relativo; pero entre más grande sea $Cond(A)$, menor es la información que se puede obtener del error relativo, a partir del error residual relativo.

De lo anterior se espera que A tenga un buen comportamiento, en el sentido de que un error residual relativo pequeño implique, correspondientemente, una buena solución aproximada de $AX = b$, si $Cond(A) \approx 1$, caso en el cual diremos que A está **bien condicionada** (el **sistema** $AX = b$ está **bien condicionado**). Si $Cond(A) >> 1$, es posible que A tenga un mal comportamiento, en el sentido que un error residual relativo pequeño puede corresponder a una solución aproximada mala, y diremos que A está **mal condicionada** (el **sistema** $AX = b$ está **mal condicionado)**.

A pesar de las definiciones anteriores, no debemos olvidar que lo que realmente nos interesa es poder determinar cuando una solución aproximada $\tilde{X}$ de un sistema $AX = b$ es "buena", y tratar de distinguir si el sistema $AX = b$ está bien o mal condicionado.

Para la matriz

$$A = \begin{pmatrix} 1 & 1 \\ 10.05 & 10 \end{pmatrix}$$

del ejemplo 3.1, tenemos

$$A^{-1} = \frac{1}{-.05} \begin{pmatrix} 10 & -1 \\ -10.05 & 1 \end{pmatrix}$$

$$\| A \|_\infty = \text{Max}\{\, |1| + |1| \,,\, |10.05| + |10| \,\} = \text{Max}\{\, 2 \,,\, 20.05 \,\} = 20.05$$

$$\| A^{-1} \|_\infty = \frac{1}{.05} \left\| \begin{pmatrix} 10 & -1 \\ -10.05 & 1 \end{pmatrix} \right\|_\infty = \frac{1}{.05} 11.05 = 221$$

luego

$$\text{Cond}_\infty(A) = \| A \|_\infty \| A^{-1} \|_\infty = (20.05)(221) = 4431.05 >> 1$$

Este número de condición nos dice que un error residual relativo $\frac{\| R \|_\infty}{\| b \|_\infty}$ pequeño, puede corresponder a un error relativo $\frac{\| \tilde{X} - X \|_\infty}{\| X \|_\infty}$ muy grande, así que A puede considerarse mal condicionada.

Veamos qué puede decirse, en este caso, de la calidad de la solución aproximada $\tilde{X}_1 = \begin{pmatrix} 10 \\ -8 \end{pmatrix}$ del sistema

$$\begin{cases} x + y = 2 \\ 10.05x + 10y = 21 \end{cases}$$

Para este ejemplo tenemos

$$\frac{\| R_1 \|_\infty}{\| b \|_\infty} = \frac{.5}{21} \quad \text{y} \quad \text{Cond}_\infty(A) = 4431.05$$

así que la desigualdad (3.7) dada en el teorema 3.3, se convierte en

$$\frac{.5}{21} \frac{1}{4431.05} \le \frac{\| \tilde{X}_1 - X_1 \|_\infty}{\| X_1 \|_\infty} \le 4431.05 \frac{.5}{21}$$

esto es,

$$5.37... \times 10^{-6} \le \frac{\| \tilde{X}_1 - X_1 \|_\infty}{\| X_1 \|_\infty} \le 105.5...$$

lo que indica que aunque el error residual relativo es pequeño, $\frac{.5}{21}$, el número de condición es tan grande (4431.05) que hace que la solución calculada pueda tener un error relativo de hasta 105.5..., así que nada puede decirse de la cercanía entre $\tilde{X}_1$ y X_1. ♦

Instrucciones en DERIVE:

NORMA_INF(A): **S**implifica en la norma del máximo de la matriz A, $\| A \|_\infty$.

COND_INF(A): **S**implifica en el número de condición relativo a la norma del máximo de la matriz A, es decir, simplifica en el número $Cond_{\infty}(A) = \| A \|_{\infty} \left\| A^{-1} \right\|_{\infty}$. à

Existe otro número asociado con una matriz, al cual se le denomina también número de condición. A continuación nos referiremos a tal número:

Del teorema 3.2 se sabe que $\rho(A) \leq \| A \|$ para toda norma matricial inducida, así que

$$Cond(A) = \| A \| \left\| A^{-1} \right\| \geq \rho(A)\rho\left(A^{-1}\right)$$

pero como los valores propios de A^{-1} son los recíprocos de los valores propios de A, se tiene que

$$Cond(A) \geq \frac{\underset{\lambda \in \sigma(A)}{Max} |\lambda|}{\underset{\lambda \in \sigma(A)}{Min} |\lambda|} \equiv Cond_{*}(A)$$

con $\sigma(A) = \{\lambda \in \mathbf{C} \ / \ \lambda$ es valor propio de $A\}$: **espectro** de A. (Recuerde que $\rho\left(A^{-1}\right) = \underset{\lambda \in \sigma\left(A^{-1}\right)}{Max} |\lambda| = \frac{1}{\underset{\lambda \in \sigma(A)}{Min} |\lambda|}$)

El número $Cond_{*}(A) = \frac{\underset{\lambda \in \sigma(A)}{Max} |\lambda|}{\underset{\lambda \in \sigma(A)}{Min} |\lambda|}$ se denomina **número de condición espectral de** A. Según se acaba de probar $Cond(A) \geq Cond_{*}(A)$.

Para la matriz $A = \begin{pmatrix} 1 & 1 \\ 10.05 & 10 \end{pmatrix}$, se tiene que

$$\det(A - \lambda I) = \begin{vmatrix} 1-\lambda & 1 \\ 10.05 & 10-\lambda \end{vmatrix} = \lambda^2 - 11\lambda - .05$$

así que los valores propios de A son $\lambda_1 \approx 11.00454358$, $\lambda_2 \approx -4.5435778 \times 10^{-3}$, y por tanto

$$Cond_{*}(A) \approx \frac{11.00454358}{4.5435778 \times 10^{-3}} \approx 2421.999592 >> 1 \quad \blacklozenge$$

Dado un sistema $AX = b$, si δA y δb denotan perturbaciones en A y b, respectivamente, el siguiente teorema, cuya demostración puede ser consultada en Ortega, 1990, páginas 32 y 33, establece una cota para el error relativo $\frac{\| \tilde{X} - X \|}{\| X \|}$, en términos de las perturbaciones relativas $\frac{\| \delta A \|}{\| A \|}$, $\frac{\| \delta b \|}{\| b \|}$ y $Cond(A)$, donde X es la solución exacta de $AX = b$ y $\tilde{X}$ es la solución exacta del sistema perturbado $(A + \delta A)X = b + \delta b$.

Teorema 3.4 Supóngase que A es no-singular y que $\| \delta A \| < \frac{1}{\| A^{-1} \|}$ (esta hipótesis asegura que $A + \delta A$ es invertible y que $1 - \text{Cond}(A)\frac{\| \delta A \|}{\| A \|} > 0$). Si $\tilde{X}$ es la solución exacta del sistema perturbado $(A + \delta A)X = b + \delta b$, entonces $\tilde{X}$ aproxima a la solución exacta X del sistema $AX = b$, $b \neq 0$, con la siguiente estimación de error

$$\frac{\| \tilde{X} - X \|}{\| X \|} \leq \frac{\text{Cond}(A)}{1 - \text{Cond}(A)\frac{\| \delta A \|}{\| A \|}} \left(\frac{\| \delta b \|}{\| b \|} + \frac{\| \delta A \|}{\| A \|} \right) \qquad \textbf{(3.11)}$$

Ñ

La desigualdad (3.11) dice que si la matriz A está bien condicionada, es decir, si $\text{Cond}(A) \approx 1$, entonces cambios "pequeños" en A y b producen, correspondientemente, cambios "pequeños" en la solución del sistema (el sistema $AX = b$ está bien condicionado). Por otro lado, si A está mal condicionada, entonces cambios "pequeños" en A y b **pueden** producir "grandes" cambios en la solución del sistema (el sistema $AX = b$ está mal condicionado).

Ejercicio 3.1 Estime la cota de error dada en el teorema 3.4 para los sistemas (3.4) y (3.4') del ejemplo 3.1. ♦

Ejercicio 3.2 a) Calcule $\text{Cond}(A)$ usando $\| \cdot \|_2$, $\| \cdot \|_1$ y $\| \cdot \|_\infty$ para las siguientes matrices:

$$\begin{pmatrix} 1 & 2 \\ 1.0001 & 2 \end{pmatrix}, \quad \begin{pmatrix} 4.56 & 2.18 \\ 2.79 & 1.38 \end{pmatrix}$$

b) Qué puede decir del condicionamiento de los siguientes sistemas de ecuaciones lineales?

i) $\begin{cases} 3.9x_1 + 1.6x_2 = 5.5 \\ 6.8x_1 + 2.9x_2 = 9.7 \end{cases}$ **ii)** $\begin{cases} 4.56x_1 + 2.18x_2 = 6.74 \\ 2.79x_1 + 1.38x_2 = 4.17 \end{cases}$ ♦

ESTABILIDAD NUMÉRICA EN LA ELIMINACIÓN GAUSSIANA

Volvamos al método de eliminación Gaussiana (simple) y consideremos el siguiente ejemplo:

Ejemplo 3.3 Resuelva el siguiente sistema de ecuaciones lineales usando eliminación Gaussiana con sustitución regresiva y aritmética (decimal) con redondeo a tres dígitos:

$$\begin{cases} E_1: \ .03x_1 + 58.9x_2 = 59.2 \\ E_2: \ 5.31x_1 - 6.10x_2 = 47.0 \end{cases}$$

Usando eliminación Gaussiana, obtenemos

$$(A : b) = \begin{pmatrix} .03 & 58.9 & : & 59.2 \\ 5.31 & -6.10 & : & 47.0 \end{pmatrix} \xrightarrow{E_{21}\left(\frac{5.31}{.03}\right),\,(m_{21}=177)} \begin{pmatrix} .03 & 58.9 & : & 59.2 \\ 0 & -10400 & : & -10500 \end{pmatrix}$$

y por sustitución regresiva

$$\tilde{x}_2 = \frac{-10500}{-10400} = 1.01$$

$$\tilde{x}_1 = \frac{59.2 - 58.9(1.01)}{.03} = \frac{59.2 - 59.5}{.03} = \frac{-.3}{.03} = -10.0$$

luego la solución calculada es $\tilde{X} = \begin{pmatrix} \tilde{x}_1 \\ \tilde{x}_2 \end{pmatrix} = \begin{pmatrix} -10.0 \\ 1.01 \end{pmatrix}$.

Instrucción en DERIVE:

PIVOT(A, i, ,j): Usa operaciones elemetales de fila para **S**implificar (o apro**X**imar) en una matriz, obtenida de la matriz A, que tiene ceros en la columna j y por debajo de la fila i. à

Qué puede decir de la calidad de la solución aproximada $\tilde{X}$?

Para intentar responder esta pregunta encontremos las cotas para el error relativo $\frac{\|\tilde{X} - X\|}{\|X\|}$, dadas por el teorema 3.3.

Como $A = \begin{pmatrix} .03 & 58.9 \\ 5.31 & -6.10 \end{pmatrix}$, entonces usando aritmética con redondeo a tres dígitos para todos los cálculos se obtienen las siguientes aproximaciones:

$$A^{-1} = \frac{1}{-313}\begin{pmatrix} -6.10 & -58.9 \\ -5.31 & .03 \end{pmatrix}$$

$$\| A \|_\infty = \text{Max}\{\, 58.9, 11.4 \,\} = 58.9$$

$$\| A^{-1} \|_\infty = \frac{1}{313}\,\text{Max}\{\, 65.0, 5.34 \,\} = \frac{1}{313}\,65.0 = .208$$

entonces $\text{Cond}_\infty(A) = \| A \|_\infty \| A^{-1} \|_\infty = (58.9)(.208) = 12.3$, que no es muy grande comparado con uno, así que la matriz A puede considerarse bien condicionada.

(Por ciertas consideraciones teóricas sobre el número de condición de una matriz A, las cuales pueden ser consultadas en Burden, 1985, páginas 481 y 482, cuando se trabaja en aritmética finita (decimal) con redondeo a **t**-dígitos y $\text{Cond}(A) \geq 10^t$ se espera un mal comportamiento de A con respecto a la solución de $AX = b$ y A se considera mal condicionada. En este ejemplo $\text{Cond}(A) = 12.3 < 10^3$).

Ahora,

$$A\tilde{X} - b = \underbrace{\begin{pmatrix} .03 & 58.9 \\ 5.31 & -6.10 \end{pmatrix}\begin{pmatrix} -10.0 \\ 1.01 \end{pmatrix} - \begin{pmatrix} 59.2 \\ 47.0 \end{pmatrix}}_{\text{Estas operaciones se realizan en doble precisión (6 dígitos)}}$$

$$= \begin{pmatrix} 59.189 \\ -59.261 \end{pmatrix} - \begin{pmatrix} 59.2 \\ 47.0 \end{pmatrix} = \begin{pmatrix} -.011 \\ -106.261 \end{pmatrix}$$

(Para evitar la pérdida de cifras significativas, se debe **calcular** el vector error residual, $R = A\tilde{X} - b$, **en doble precisión**).

Convirtiendo este último resultado a tres dígitos usando redondeo, se obtiene

$$R = \begin{pmatrix} -.011 \\ -106 \end{pmatrix}$$

Entonces $\| R \|_\infty = 106$ y $\| b \|_\infty = 59.2$ y por tanto

$$\frac{\| R \|_\infty}{\| b \|_\infty} \frac{1}{Cond_\infty(A)} = \frac{106}{59.2}\frac{1}{12.3} = .145... \leq \frac{\| \tilde{X} - X \|}{\| X \|} \leq 22.02... = 12.3\frac{106}{59.2} = Cond_\infty(A)\frac{\| R \|_\infty}{\| b \|_\infty}$$

pero como $Cond_\infty(A) \approx 1$, entonces se espera que $\frac{\| R \|_\infty}{\| b \|_\infty}$ y $\frac{\| \tilde{X} - X \|_\infty}{\| X \|_\infty}$ sean más o menos del mismo tamaño, y ya que $\frac{\| R \|_\infty}{\| b \|_\infty}$ es grande, se espera que $\frac{\| \tilde{X} - X \|_\infty}{\| X \|_\infty}$ sea también grande. ♦

En situaciones como la observada en este ejemplo, se sugiere hacer un **refinamiento iterativo** sobre la solución calculada $\tilde{X}$ o usar esta solución calculada como aproximación inicial en un método iterativo, con el propósito de tratar de mejorar la solución aproximada y lograr que el error residual relativo sea más pequeño. El método de refinamiento iterativo puede ser consultado en Kincaid, 1972, páginas 174-176.

La solución exacta del sistema en consideración es $X = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 10 \\ 1 \end{pmatrix}$. A qué se debe la diferencia entre la solución exacta y la solución calculada?

Observe que el error en el cálculo de $\tilde{x}_2$ con respecto a x_2 fue de sólo .01 (un error relativo de **1%**) y este error fue multiplicado por un factor de aproximadamente -2000 al obtener $\tilde{x}_1$, debido al **orden** en que se realizó la eliminación Gaussiana.

Instrucción en DERIVE:

RESUELVA_1(A,b): **S**implifica en la solución exacta X del sistema $AX = b$. El vector b se entra como un vector fila. à

En el ejemplo anterior la eliminación Gaussiana condujo a una respuesta defectuosa de un sistema de ecuaciones lineales bien condicionado. Esto muestra la **inestabilidad numérica** del algoritmo de eliminación Gaussiana (consecuencia de la división por un número (pivote) pequeño). Hay, sin

embargo, situaciones en las cuales el algoritmo de eliminación Gaussiana es numéricamente estable. El siguiente teorema cuya demostración puede consultarse en Burden, 1985, páginas 366 y 367, se refiere a una de tales situaciones:

Teorema 3.5 Si $A = (a_{ij})_{n\times n}$ es una **matriz estrictamente dominante diagonalmente** (**E.D.D.**) por filas, es decir, si

$$|a_{ii}| > \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}| \quad \text{para cada } i = 1,2,\ldots,n$$

entonces A es invertible (no-singular). Además, se puede realizar eliminación Gaussiana **sin intercambio de filas** en cualquier sistema $AX = b$ para obtener su única solución, y los cálculos son **estables** con respecto al crecimiento de los errores de redondeo. Ñ

Nótese que como consecuencia del teorema anterior se tiene que: Si $A = (a_{ij})_{n\times n}$ es **E.D.D.** por filas, entonces A tiene factorización LU, es decir, $A = LU$, con L triangular inferior con unos en su diagonal principal y U triangular superior (escalonada).

Observe que la matriz de coeficientes del ejemplo 3.3 anterior, **no** es **E.D.D.** por filas.

El teorema 3.5 también es válido para **matrices reales**, **simétricas y definidas positivas** (véase Burden, 1985, página 368). Una matriz $A \in \mathbf{R}_{n\times n}$, simétrica, se dice **definida positiva** si satisface **una** cualquiera de las siguientes condiciones (las cuales son equivalentes):

i) $X^T AX > 0$ para todo $X \in \mathbf{R}^n$, $X \neq 0$.

ii) Todos los valores propios de A son positivos.

iii) Todos los pivotes obtenidos en la eliminación Gaussiana sobre A, sin intercambio de filas, son positivos.

iv) Todas las submatrices principales de A tienen determinante positivo.

(Las submatrices principales de la matriz $A = (a_{ij})_{n\times n}$ son las matrices

$$A_k = \begin{pmatrix} a_{11} & \cdots & a_{1k} \\ \vdots & \vdots & \vdots \\ a_{k1} & \cdots & a_{kk} \end{pmatrix}, \; k = 1,2,\ldots,n \;)$$

Ñ

Nótese nuevamente que: Si $A \in \mathbf{R}_{n\times n}$ es simétrica y definida positiva, entonces A tiene factorización $A = LU$, con L triangular inferior con sus componentes sobre la diagonal principal iguales a uno y U triangular superior (ver más adelante factorización de Choleski).

Observe que la matriz de coeficientes del ejemplo 3.3 anterior, **no** es simétrica.

3.4 ESTRATEGIAS DE PIVOTEO

El ejemplo 3.3 anterior, muestra una de las dificultades que pueden surgir al aplicar el método de eliminación Gaussiana cuando el pivote $a_{jj}^{(j-1)}$ es "pequeño" comparado con algunos elementos $a_{it}^{(j-1)}$ para $j \le i,t \le n$.

Para tratar de evitar tales dificultades, se introduce en el método de eliminación Gaussiana una **estrategia** llamada **de pivoteo**, la cual consiste en seleccionar el pivote de acuerdo con un cierto criterio. Nosotros usaremos dos estrategias: la estrategia de **pivoteo máximo por columna** o **pivoteo parcial** y la estrategia de **pivoteo escalado de fila** o **escalamiento**.

3.4.1 Pivoteo máximo por columna o pivoteo parcial: Esta estrategia difiere de eliminación Gaussiana simple, únicamente en la escogencia del pivote $a_{jj}^{(j-1)}$, la cual se hace ahora, así:

Para $j = 1,2,\ldots,n-1$, se determina el menor entero k, $j \le k \le n$, tal que

$$\boxed{a_{kj}^{(j-1)} \neq 0 \quad \text{y} \quad \left| a_{kj}^{(j-1)} \right| = \max_{j \le i \le n} \left| a_{ij}^{(j-1)} \right|}$$

es decir, seleccionamos el primer elemento diferente de cero sobre la columna **j**-ésima a partir de la **j**-ésima fila y que tenga mayor valor absoluto (para $j = 1$, $a_{kj}^{(j-1)} = a_{k1}^{(0)} \equiv a_{k1}$).

Si tal k no existe, el sistema no tiene solución única y el proceso se puede terminar.

Si tal k existe y $k \neq j$, entonces hacemos intercambio de las ecuaciones **j**-ésima y **k**-ésima:

$$E_j^{(j-1)} \leftrightarrow E_k^{(j-1)}$$

y continuamos con la eliminación Gaussiana. Ñ

Ilustremos esta estrategia para resolver el sistema

$$\begin{cases} E_1: \ .03x_1 + 58.9x_2 = 59.2 \\ E_2: \ 5.31x_1 - 6.10x_2 = 47.0 \end{cases}$$

que es el mismo del ejemplo 3.3, usando aritmética con redondeo a tres dígitos.

Como para $j = 1$, se tiene que

$$\text{Max}\left\{ \left| a_{11} \right|, \ \left| a_{21} \right| \right\} = \text{Max}\left\{ .03, 5.31 \right\} = 5.31 = \left| a_{21} \right| \neq 0$$

entonces $k = 2 \neq 1 = j$, así que intercambiamos E_1 y E_2 y continuamos con la eliminación

$$(A:b) = \begin{pmatrix} .03 & 58.9 & : & 59.2 \\ 5.31 & -6.10 & : & 47.0 \end{pmatrix} \xrightarrow{P_{12}} \begin{pmatrix} 5.31 & -6.10 & : & 47.0 \\ .03 & 58.9 & : & 59.2 \end{pmatrix}$$

$$\xrightarrow{E_{21}\left(\frac{.03}{5.31}\right),\left(m_{21}=5.65\times10^{-3}\right)} \begin{pmatrix} 5.31 & -6.10 & : & 47.0 \\ 0 & 58.9 & : & 58.9 \end{pmatrix}$$

Por sustitución regresiva, obtenemos

$$\tilde{x}_2 = \frac{58.9}{58.9} = 1.00\,, \quad \tilde{x}_1 = \frac{47.0 + 6.10(1.00)}{5.31} = \frac{53.1}{5.31} = 10.0$$

Observe que en este caso $\tilde{X} = \begin{pmatrix} \tilde{x}_1 \\ \tilde{x}_2 \end{pmatrix}$ es la solución exacta del sistema dado. ♦

Instrucción en DERIVE:

SWAP(A, i, j): Intercambia las filas (o elementos) i y j de la matriz A (de un vector). à

Nota: En el procedimiento de pivoteo máximo por columna (pivoteo parcial) cada multiplicador m_{ij} es tal que

$$\left| m_{ij} \right| = \frac{\left| a_{ij}^{(j-1)} \right|}{\left| a_{jj}^{(j-1)} \right|} \le 1$$

y aunque esta estrategia permite resolver satisfactoriamente muchos sistemas de ecuaciones lineales, hay casos donde fracasa, como se ilustra en el siguiente ejemplo.

Ejemplo 3.4 Consideremos el sistema

$$\begin{cases} E_1: \; 30.0x_1 + 58900x_2 = 59200 \\ E_2: \; 5.31x_1 \; - 6.10x_2 = \; 47.0 \end{cases}$$

el cual es un sistema equivalente al del ejemplo 3.3 (los coeficientes de la primera ecuación en el sistema del ejemplo 3.3, han sido multiplicados por 10^3). El pivoteo máximo por columna con aritmética de redondeo a tres dígitos, nos lleva a los siguientes resultados:

$$(A:b) = \begin{pmatrix} 30.0 & 58900 & : & 59200 \\ 5.31 & -6.10 & : & 47.0 \end{pmatrix} \xrightarrow{E_{21}\left(\frac{5.31}{30.0}\right),\,(m_{21}=.177)} \begin{pmatrix} 30.0 & 58900 & : & 59200 \\ 0 & -10400 & : & -10500 \end{pmatrix}$$

y por sustitución regresiva $\tilde{x}_2 = 1.01$ y $\tilde{x}_1 = -10.0$, que es la misma solución que se obtiene si usamos eliminación Gaussiana simple.
En casos como el de este ejemplo, donde un pivote es mucho más "pequeño" que alguno de los coeficientes de la ecuación que él encabeza, se recomienda la técnica conocida como pivoteo escalado de fila o escalamiento, la cual es nuestra segunda estrategia.

3.4.2 Pivoteo escalado de fila: Esta técnica sólo difiere de la eliminación Gaussiana simple, al igual que el pivoteo parcial, en la escogencia del pivote.
Esta vez el pivote $a_{jj}^{(j-1)}$ se escoge como se indica a continuación:

Para $j = 1,2,...,n-1$, hacemos lo siguiente:

a) Para $i = j, j+1, ..., n$, calculamos

$$S_i = \max_{j \le l \le n} \left| a_{il}^{(j-1)} \right| : \text{Factor de escala}$$

Si $S_i = 0$, entonces el sistema no tiene solución única y el proceso se puede terminar.

b) Para $i = j, j+1, \ldots, n$, calculamos

$$\frac{\left| a_{ij}^{(j-1)} \right|}{S_i}$$

c) Encontramos el menor entero k con $j \le k \le n$ tal que

$$a_{kj}^{(j-1)} \neq 0 \quad \text{y} \quad \frac{\left| a_{kj}^{(j-1)} \right|}{S_k} = \max_{j \le i \le n} \frac{\left| a_{ij}^{(j-1)} \right|}{S_i}$$

Si tal k **no** existe, entonces el sistema no tiene solución única y el proceso se puede terminar.

Si tal k existe y $k \neq j$, entonces hacemos intercambio de las ecuaciones **j-**ésima y **k**-ésima:

$$E_j^{(j-1)} \leftrightarrow E_k^{(j-1)}$$

y continuamos con la eliminación Gaussiana. Ñ

Apliquemos esta estrategia para resolver el sistema del ejemplo 3.4, usando aritmética con redondeo a tres dígitos.

Para $j = 1$:

a) $S_1 = \text{Max}\{ |a_{11}|, |a_{12}| \} = \text{Max}\{ 30.0, 58900 \} = 58900 \neq 0$, y

$S_2 = \text{Max}\{ |a_{21}|, |a_{22}| \} = \text{Max}\{ 5.31, 6.10 \} = 6.10 \neq 0$

b) Ahora,

$$\frac{|a_{11}|}{S_1} = \frac{30.0}{58900}, \text{ y}$$

$$\frac{|a_{21}|}{S_2} = \frac{5.31}{6.10}$$

c) $\text{Max}\left\{ \frac{|a_{11}|}{S_1}, \frac{|a_{21}|}{S_2} \right\} = \text{Max}\left\{ \frac{30.0}{58900}, \frac{5.31}{6.0} \right\} = \frac{5.31}{6.0} = \frac{|a_{21}|}{S_2} \neq 0$, así que $k = 2 \neq 1 = j$ y por tanto intercambiamos las ecuaciones E_1 y E_2 y continuamos con la eliminación Gaussiana:

$$(A : b) = \begin{pmatrix} 30.0 & 58900 & : & 59200 \\ 5.31 & -6.10 & : & 47.0 \end{pmatrix} \xrightarrow{P_{12}} \begin{pmatrix} 5.31 & -6.10 & : & 47.0 \\ 30.0 & 58900 & : & 59200 \end{pmatrix}$$

$$\xrightarrow{E_{21}\left(\frac{30.0}{5.31}\right),\,(m_{21}=5.65)} \begin{pmatrix} 5.31 & -6.10 & : & 47.0 \\ 0 & 58900 & : & 58900 \end{pmatrix}$$

Por sustitución regresiva

$$\tilde{x}_2 = 1.00\,, \quad \tilde{x}_1 = \frac{47.0 + 6.10(1.00)}{5.31} = \frac{53.1}{5.31} = 10.0$$

Observe que en este caso $\tilde{X} = \begin{pmatrix} \tilde{x}_1 \\ \tilde{x}_2 \end{pmatrix} = \begin{pmatrix} 10.0 \\ 1.00 \end{pmatrix}$ es la solución exacta del sistema. ♦

3.5 FACTORIZACIÓN TRIANGULAR

Consideremos un sistema $AX = b$, con A no-singular y $b \neq 0$. Con respecto a la matriz A, se sabe que existen matrices P de permutación, L triangular inferior con sus componentes sobre la diagonal principal iguales a uno y U triangular superior (escalonada) tales que $PA = LU$. Una forma eficiente computacionalmente de encontrar P, L y U, usando eliminación Gaussiana, se muestra en el siguiente ejemplo:

Ejemplo 3.5 Resuelva el siguiente sistema de ecuaciones lineales usando eliminación Gaussiana con pivoteo parcial y obtenga una factorización $PA = LU$ para la matriz A de coeficientes, asociada con este método:

$$\begin{cases} -x_1 + 2x_2 - 3x_3 = -2 \\ 3x_1 - 3x_2 - x_3 = -4 \\ x_1 + x_2 = 3 \end{cases}$$

Empezamos introduciendo un vector $p = (p_1, p_2, p_3)^T$ el cual inicializamos con $p_i = i,\ i = 1,2,3$ y donde se almacenarán los intercambios necesarios en el proceso de eliminación Gaussiana con pivoteo parcial (el número de componentes del vector p coincide con el orden n del sistema a resolver):

$$p = (1,2,3)^T,\ (A:b) = \begin{pmatrix} -1 & 2 & -3 & : & -2 \\ 3 & -3 & -1 & : & -4 \\ 1 & 1 & 0 & : & 3 \end{pmatrix} \xrightarrow{\text{Max}\{1,3,1\}=3,\ p=(2,1,3)^T} \begin{pmatrix} 3 & -3 & -1 & : & -4 \\ -1 & 2 & -3 & : & -2 \\ 1 & 1 & 0 & : & 3 \end{pmatrix}$$

$$\xrightarrow[E_{21}\left(-\frac{1}{3}\right),\ E_{31}\left(\frac{1}{3}\right)]{\left(m_{21}=-\frac{1}{3}\right),\left(m_{31}=\frac{1}{3}\right)} \begin{pmatrix} 3 & -3 & -1 & : & -4 \\ \left(-\frac{1}{3}\right) & 1 & -\frac{10}{3} & : & -\frac{10}{3} \\ \left(\frac{1}{3}\right) & 2 & \frac{1}{3} & : & \frac{13}{3} \end{pmatrix}$$

(Observe que cada multiplicador m_{ij} es almacenado en la posición correspondiente (i,j) en la matriz de trabajo)

$$\xrightarrow{\text{Max}\{1,2\}=2,\ p=(2,3,1)^T} \begin{pmatrix} 3 & -3 & -1 & : & -4 \\ \left(\frac{1}{3}\right) & 2 & \frac{1}{3} & : & \frac{13}{3} \\ \left(-\frac{1}{3}\right) & 1 & -\frac{10}{3} & : & -\frac{10}{3} \end{pmatrix}$$

(Observe que la permutación se hace para las filas 2 y 3 completas, es decir, incluyendo los multiplicadores)

$$\xrightarrow[E_{32}\left(\frac{1}{2}\right)]{\left(m_{32}=\frac{1}{2}\right)} \begin{pmatrix} 3 & -3 & -1 & : & -4 \\ \left(\frac{1}{3}\right) & 2 & \frac{1}{3} & : & \frac{13}{3} \\ \left(-\frac{1}{3}\right) & \left(\frac{1}{2}\right) & -\frac{7}{2} & : & -\frac{11}{2} \end{pmatrix}$$

La eficiencia en el método indicado se debe a que en la misma matriz de trabajo se almacenan los multiplicadores que van a conformar la matriz L (en el ejemplo son los números que se encuentran dentro de paréntesis), lo que significa un ahorro de memoria, y como los intercambios necesarios afectan simultáneamente a las matrices L y U, se evita tener que volver a la matriz original a realizar los intercambios ya observados y repetir la eliminación Gaussiana con pivoteo parcial. De esta manera, al terminar el proceso de eliminación podemos leer en la matriz final la parte estrictamente triangular inferior de L (son los números entre paréntesis) y la matriz triangular superior U (que es la parte triangular superior de la matriz final), y en el vector p final quedan almacenados los intercambios realizados que se usan para producir la matriz de permutación P.

Para el ejemplo 3.5,

$$L = \begin{pmatrix} 1 & 0 & 0 \\ \frac{1}{3} & 1 & 0 \\ -\frac{1}{3} & \frac{1}{2} & 1 \end{pmatrix}, \quad U = \begin{pmatrix} 3 & -3 & -1 \\ 0 & 2 & \frac{1}{3} \\ 0 & 0 & -\frac{7}{2} \end{pmatrix}, \text{ y como } p = (2,3,1)^T, \text{ entonces } \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix} = P$$

(Verifique que $PA = LU$) .

Para obtener la solución del sistema original, usamos sustitución regresiva en el sistema reducido

$$UX = \underbrace{\begin{pmatrix} -4 \\ \frac{13}{3} \\ -\frac{11}{2} \end{pmatrix}}_{c}$$

y obtenemos

$$x_3 = \frac{11}{7}, \quad x_2 = \frac{40}{21} \text{ y } x_1 = \frac{23}{21}$$

así que la solución (exacta) del sistema dado es $X = \left(\frac{23}{21}, \frac{40}{21}, \frac{11}{7}\right)^T$.

Cómo se resuelve el sistema $AX = b$ a partir de la factorización $PA = LU$ obtenida? ♦

3.5.1 Algunas aplicaciones de la factorización $PA = LU$**:** La factorización $PA = LU$ es utilizada eficientemente en aquellos casos donde se trabaja repetídamente con la misma matriz A. Dos de esos casos se presentan a continuación.

1) Resolver varios sistemas $AX = b$ con la misma matriz de coeficientes A, ya que en P, L y U está almacenado **todo** el proceso de eliminación Gaussiana. El algoritmo se basa en la siguiente equivalencia

$$\begin{aligned} AX = b &\Leftrightarrow PAX = Pb \\ &\Leftrightarrow LUX = Pb \\ &\Leftrightarrow UX = L^{-1}Pb \\ &\Leftrightarrow UX = c \text{ y } c = L^{-1}Pb \\ &\Leftrightarrow UX = c \text{ y } Lc = Pb \\ &\Leftrightarrow Lc = Pb \text{ y } UX = c \end{aligned}$$

Los pasos a seguir son:

Paso 1. Calcular Pb.
Paso 2. Resolver, para c, $Lc = Pb$ por sustitución progresiva.
Paso 3. Resolver, para X, $UX = c$ por sustitución regresiva.

Como **ejercicio**, resuelva el sistema del ejemplo anterior, usando este algoritmo.

2) Encontrar la matriz **inversa** A^{-1} de una matriz invertible A, resolviendo los **n**-sistemas

$$AX = e^{(j)}, \quad j = 1,2,...,n$$

donde $e^{(j)} = (0,...,0,1,0,...,0)^T \in \mathbf{R}^n$.

↑
posición j

La solución X del sistema $AX = e^{(j)}$, $j = 1,2,...,n$, produce la correspondiente columna j-ésima de la matriz A^{-1}.

Como **ejercicio**, compare el número de operaciones para encontrar A^{-1} usando el método de Gauss-Jordan, con el número de operaciones resolviendo los **n**-sistemas indicados antes. Ñ

Ejercicio 3.3 Calcule la inversa de la matriz A de coeficientes del sistema del ejemplo 3.5, usando el método de Gauss-Jordan y también usando la factorización $PA = LU$. ♦

3.6 SISTEMAS TRIDIAGONALES

Un caso muy importante de sistemas de ecuaciones lineales, que requiere un tratamiento especial, es el de los **sistemas tridiagonales**. Tales sistemas aparecen en diversas aplicaciones, como por ejemplo al utilizar métodos de diferencias finitas en la solución de problemas con valores en la frontera para ecuaciones diferenciales ordinarias y, como veremos más adelante, en el problema de la interpolación segmentaria cúbica.

Un sistema tridiagonal es de la forma

$$\begin{cases} d_1x_1 + c_1x_2 = b_1 \\ a_2x_1 + d_2x_2 + c_2x_3 = b_2 \\ \quad a_3x_2 + d_3x_3 + c_3x_4 = b_3 \\ \qquad\qquad \vdots \\ a_{n-1}x_{n-2} + d_{n-1}x_{n-1} + c_{n-1}x_n = b_{n-1} \\ \quad a_nx_{n-1} + d_nx_n = b_n \end{cases}$$

La matriz de coeficientes del sistema es

$$A = \begin{pmatrix} d_1 & c_1 & 0 & 0 & \cdots & 0 \\ a_2 & d_2 & c_2 & 0 & & \\ 0 & a_3 & d_3 & c_3 & & \\ \vdots & & & & & \vdots \\ & & & a_{n-1} & d_{n-1} & c_{n-1} \\ 0 & & \cdots & 0 & a_n & d_n \end{pmatrix}$$

la cual se dice una matriz **TRIDIAGONAL** (caso especial de las llamadas matrices banda).

En general, una matriz $A = (a_{ij})_{n\times n}$ se dice **TRIDIAGONAL** si $a_{ij} = 0$ para $|i-j| > 1$, $i,j = 1,2,...,n$.

Suponiendo que la matriz A de coeficientes del sistema tridiagonal es **E.D.D.** por filas, podemos usar eliminación Gaussiana simple para resolver el sistema (recuerde que la eliminación Gaussiana simple es adecuada para resolver sistemas con matriz de coeficientes **E.D.D.** por filas). Otra forma de resolver un sistema tridiagonal con matriz de coeficientes A **E.D.D.** por filas es a partir de la factorización directa $A = LU$ (encontrando directamente las componentes de las matrices L y U, es decir, sin usar eliminación Gaussiana), donde L es triangular inferior con todos sus elementos diagonales iguales a uno y U es triangular superior. Para encontrar tales matrices L y U, partimos de que ellas deben ser de la forma

$$L = \begin{pmatrix} 1 & 0 & 0 & 0 & \cdots & 0 \\ \gamma_2 & 1 & 0 & 0 & \cdots & 0 \\ 0 & \gamma_3 & 1 & 0 & \cdots & 0 \\ \vdots & & & \ddots & & \\ 0 & & & \gamma_{n-1} & 1 & 0 \\ 0 & \cdots & & 0 & \gamma_n & 1 \end{pmatrix}, \quad U = \begin{pmatrix} \alpha_1 & c_1 & 0 & 0 & \cdots & 0 \\ 0 & \alpha_2 & c_2 & 0 & \cdots & 0 \\ 0 & 0 & \alpha_3 & c_3 & \cdots & 0 \\ \vdots & & & \ddots & & \\ 0 & & & 0 & \alpha_{n-1} & c_{n-1} \\ 0 & \cdots & & 0 & 0 & \alpha_n \end{pmatrix}$$

Como

$$LU = \begin{pmatrix} \alpha_1 & c_1 & 0 & 0 & \cdots & 0 \\ \gamma_2\alpha_1 & \gamma_2 c_1 + \alpha_2 & c_2 & 0 & & 0 \\ 0 & \gamma_3\alpha_2 & \gamma_3 c_2 + \alpha_3 & c_3 & & 0 \\ \vdots & & & & & \vdots \\ 0 & & & & \gamma_{n-1}\alpha_{n-2} & \gamma_{n-1}c_{n-2} + \alpha_{n-1} & c_{n-1} \\ 0 & & \cdots & & 0 & \gamma_n\alpha_{n-1} & \gamma_n c_{n-1} + \alpha_n \end{pmatrix}$$

$$= \begin{pmatrix} d_1 & c_1 & & & & \\ a_2 & d_2 & c_2 & & & \\ & a_3 & d_3 & c_3 & & \\ & & \ddots & \ddots & \ddots & \\ & & & a_{n-1} & d_{n-1} & c_{n-1} \\ & & & & a_n & d_n \end{pmatrix} = A$$

igualando componente a componente las matrices LU y A, se tiene que

$$\alpha_1 = d_1$$

$$\gamma_i\alpha_{i-1} = a_i, \quad i = 2,3,...,n$$

$$\gamma_i c_{i-1} + \alpha_i = d_i, \quad i = 2,3,...,n$$

y resolviendo para γ_i y α_i, obtenemos

$$\left.\begin{aligned} \alpha_1 &= d_1 \\ \gamma_i &= \frac{a_i}{\alpha_{i-1}} \\ \alpha_i &= d_i - \gamma_i c_{i-1} \end{aligned}\right\} \text{ para } i = 2,3,...,n$$

($\alpha_i \neq 0$, porque en U todos los elementos diagonales son distintos de cero).

Así que un **algoritmo** para determinar L y U, en el caso que nos ocupa es el siguiente:

Paso 1: Haga $\alpha_1 = d_1$.

Paso 2: Para $i = 2,3,...,n$, haga

$$\gamma_i = \frac{a_i}{\alpha_{i-1}}$$
$$\alpha_i = d_i - \gamma_i c_{i-1}$$

Una vez encontradas L y U se resuelve el sistema $AX = b$, resolviendo para c, $Lc = b$ y luego resolviendo para X, $UX = c$. Ñ

Ejemplo 3.6 Resolver el siguiente sistema tridiagonal usando aritmética (decimal) con redondeo a tres dígitos y la factorización $A = LU$.

$$\begin{cases} .5x_1 + .25x_2 = .35 \\ .35x_1 + .8\,x_2 + .4x_3 = .77 \\ .25x_2 + x_3 + .5x_4 = -.5 \\ x_3 - 2x_4 = -2.25 \end{cases}$$

Es claro que el sistema es tridiagonal **E.D.D.** por filas (ya que $.5 > .25,\ .8 > .75,\ 1 > .75,\ 2 > 1$).

A partir del sistema dado se tiene que

$$\begin{aligned} &d_1 = .5,\ d_2 = .8,\ d_3 = 1.0,\ d_4 = -2.0 \\ &c_1 = .25,\ c_2 = .4,\ c_3 = .5 \\ &a_2 = .35,\ a_3 = .25,\ a_4 = 1.0 \end{aligned}$$

y las matrices L y U son de la forma

$$L = \begin{pmatrix} 1 & 0 & 0 & 0 \\ \gamma_2 & 1 & 0 & 0 \\ 0 & \gamma_3 & 1 & 0 \\ 0 & 0 & \gamma_4 & 1 \end{pmatrix}, \qquad U = \begin{pmatrix} \alpha_1 & .25 & 0 & 0 \\ 0 & \alpha_2 & .4 & 0 \\ 0 & 0 & \alpha_3 & .5 \\ 0 & 0 & 0 & \alpha_4 \end{pmatrix}$$

Usando el algoritmo ya descrito, se obtiene

$$\begin{aligned} &\alpha_1 = .5 \\ &\gamma_2 = \frac{.35}{.5} = .7 \\ &\alpha_2 = .8 - (.7)(.25) = .625 \\ &\gamma_3 = \frac{.25}{.625} = .4 \\ &\alpha_3 = 1.0 - (.4)(.4) = .84 \\ &\gamma_4 = \frac{1.0}{.84} = 1.19 \\ &\alpha_4 = -2.0 - (1.19)(.5) = -2.60 \end{aligned}$$

Luego

$$L = \begin{pmatrix} 1 & 0 & 0 & 0 \\ .7 & 1 & 0 & 0 \\ 0 & .4 & 1 & 0 \\ 0 & 0 & 1.19 & 1 \end{pmatrix} \quad \text{y} \quad U = \begin{pmatrix} .5 & .25 & 0 & 0 \\ 0 & .625 & .4 & 0 \\ 0 & 0 & .84 & .5 \\ 0 & 0 & 0 & -2.60 \end{pmatrix}$$

Ahora, resolvemos el sistema

$$Lc = b \Leftrightarrow \begin{pmatrix} 1 & 0 & 0 & 0 \\ .7 & 1 & 0 & 0 \\ 0 & .4 & 1 & 0 \\ 0 & 0 & 1.19 & 1 \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \\ c_3 \\ c_4 \end{pmatrix} = \begin{pmatrix} .35 \\ .77 \\ -.5 \\ -2.25 \end{pmatrix}$$

por sustitución progresiva, y se obtiene

$$c_1 = .35,\ c_2 = .525,\ c_3 = -.71,\ c_4 = -1.41$$

Enseguida resolvemos el sitema

$$UX = c \Leftrightarrow \begin{pmatrix} .5 & .25 & 0 & 0 \\ 0 & .625 & .4 & 0 \\ 0 & 0 & .84 & .5 \\ 0 & 0 & 0 & -2.60 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} .35 \\ .525 \\ -.71 \\ -1.41 \end{pmatrix}$$

por sustitución regresiva, y obtenemos

$$x_4 = .542,\ x_3 = -1.17,\ x_2 = 1.59,\ x_1 = -.096$$

Luego la solución (aproximada) obtenida para el sistema dado, es $\tilde{X} = (-.096, 1.59, -1.17, .542)^T$.

3.7 FACTORIZACÍON DE CHOLESKI

Se puede probar (véase Kincaid 1972, página 133) que si A es una matriz real, simétrica y definida positiva, entonces A tiene una única factorización de la forma $A = LL^T$ en la cual L es una matriz triangular inferior con sus elementos en la diagonal principal todos positivos (no necesariamente con unos en su diagonal principal). Esta factorización se conoce como **factorización de Choleski** (recuerde que si A es definida positiva, entonces se puede realizar eliminación Gaussiana sobre A sin intercambio de filas y todos los pivotes que resultan son positivos).

Se puede demostrar que si $A \in \mathbf{R}_{n\times n}$ y tiene factorización de Choleski, entonces A es definida positiva (**ejercicio!**).

Para ilustrar cómo se obtiene la **factorización directa de Choleski**, es decir, sin usar eliminación Gaussiana, supongamos que la matriz A es de orden 4. Entonces

$$\underbrace{\begin{pmatrix} l_{11} & 0 & 0 & 0 \\ l_{21} & l_{22} & 0 & 0 \\ l_{31} & l_{32} & l_{33} & 0 \\ l_{41} & l_{42} & l_{43} & l_{44} \end{pmatrix}}_{L} \underbrace{\begin{pmatrix} l_{11} & l_{21} & l_{31} & l_{41} \\ 0 & l_{22} & l_{32} & l_{42} \\ 0 & 0 & l_{33} & l_{43} \\ 0 & 0 & 0 & l_{44} \end{pmatrix}}_{L^T} = \underbrace{\begin{pmatrix} a_{11} & a_{21} & a_{31} & a_{41} \\ a_{21} & a_{22} & a_{32} & a_{42} \\ a_{31} & a_{32} & a_{33} & a_{43} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{pmatrix}}_{A \text{ (simétrica)}}$$

Como $l_{11}l_{11} = a_{11}$, entonces escogemos $l_{11} = \sqrt{a_{11}}$, lo que requiere que $a_{11} > 0$.

$l_{21}l_{11} = a_{21}$, entonces $l_{21} = \dfrac{a_{21}}{l_{11}}$, $l_{11} \neq 0$.

$l_{21}^2 + l_{22}^2 = a_{22}$, escogemos $l_{22} = \sqrt{a_{22} - l_{21}^2}$. Observe que a_{22} debe ser mayor o igual que cero.

$l_{31}l_{11} = a_{31}$, luego $l_{31} = \dfrac{a_{31}}{l_{11}}$.

$l_{31}l_{21} + l_{32}l_{22} = a_{32}$, luego $l_{32} = \dfrac{a_{32} - l_{31}l_{21}}{l_{22}}$, $l_{22} \neq 0$, así que $a_{22} > l_{21}^2 \geq 0$, esto es $a_{22} > 0$.

$l_{31}^2 + l_{32}^2 + l_{33}^2 = a_{33}$, escogemos $l_{33} = \sqrt{a_{33} - \left(l_{31}^2 + l_{32}^2\right)}$, $a_{33} \geq 0$.

En general, para encontrar la fila **i** de L, $(i \geq 3)$, asumiendo que ya se conocen las primeras $i-1$ filas de L, procedemos así:

De $l_{i\,1}l_{11} = a_{i\,1}$, obtenemos $l_{i\,1} = \dfrac{a_{i\,1}}{l_{11}}$.

$l_{i\,1}l_{21} + l_{i\,2}l_{22} = a_{i\,2}$, así que $l_{i\,2} = \dfrac{a_{i\,2} - l_{i\,1}l_{21}}{l_{22}}$.

$l_{i\,1}l_{31} + l_{i\,2}l_{32} + l_{i\,3}l_{33} = a_{i\,3}$, así que

$l_{i3} = \dfrac{a_{i3} - (l_{i1}l_{31} + l_{i2}l_{32})}{l_{33}}$, $l_{33} \neq 0$, $a_{33} > (l_{31}^2 + l_{32}^2) \geq 0$, así que $a_{33} > 0$.

(Fila **i-ésima** de L**)** $\times$ **(**Columna **j-ésima** de L^T **)**, **j** < **i**:

$l_{i1}l_{j1} + l_{i2}l_{j2} + \ldots + l_{ij}l_{jj} = a_{ij}$, entonces para cada $i = 3,\ldots,n$:

$$l_{ij} = \frac{a_{ij} - \sum_{k=1}^{j-1} l_{ik}l_{jk}}{l_{jj}}, \quad j = 2,\ldots,i-1$$

(Fila **i-ésima** de L**)** ´ **(**Columna **i-ésima** de L^T **)**:

$l_{i1}^2 + l_{i2}^2 + \ldots + l_{ii}^2 = a_{ii}$, escogemos

$$l_{ii} = \sqrt{a_{ii} - \sum_{k=1}^{i-1} l_{ik}^2}$$

Algoritmo 3.3 (Factorización directa de Choleski) Para factorizar una matriz $A = (a_{ij})_{n\times n}$ real, simétrica y definida positiva en la forma $A = LL^T$, donde L es triangular inferior.

Entrada: La dimensión **n** de la matriz A, los elementos a_{ij}, $1 \leq i,j \leq n$ (basta almacenar la parte triangular inferior de A, por ser A simétrica).

Salida: Los elementos l_{ij}, $1 \leq i \leq n$, $1 \leq j \leq i$ de la matriz L.

Paso 1: Tomar $l_{11} = \sqrt{a_{11}}$.

Paso 2: Para $i = 2,\ldots,n$, tomar $l_{i1} = \dfrac{a_{i1}}{l_{11}}$ (primera columna de L).

Paso 3: Tomar $l_{22} = \sqrt{a_{22} - l_{21}^2}$.

Paso 4: Para $i = 3,\ldots,n$, seguir los pasos 5 y 6:

Paso 5: Para $j = 2,\ldots,i-1$, tomar

$$l_{ij} = \frac{\left(a_{ij} - \sum_{k=1}^{j-1} l_{ik}l_{jk}\right)}{l_{jj}}$$

Paso 6: Tome $l_{ii} = \left(a_{ii} - \sum_{k=1}^{i-1} l_{ik}^2\right)^{\frac{1}{2}}$.

Paso 7: Salida: "Las componentes l_{ij} de la matriz L para $i = 1,2,\ldots,n,\ j = 1,2,\ldots,i$ ".

Terminar.

Ejemplo 3.7 Diga si la siguiente matriz es simétrica y definida positiva, y si lo es, encuentre la factorización directa de Choleski:

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

Es claro que la matriz A es simétrica, ya que $A^T = A$. Para ver si la matriz A es definida positiva, realicemos eliminación Gaussiana sobre la matriz A:

$$A = \begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{pmatrix} \xrightarrow{E_{21}\left(\frac{-1}{2}\right)} \begin{pmatrix} 2 & -1 & 0 \\ 0 & \frac{3}{2} & -1 \\ 0 & -1 & 2 \end{pmatrix} \xrightarrow{E_{32}\left(-\frac{2}{3}\right)} \begin{pmatrix} 2 & -1 & 0 \\ 0 & \frac{3}{2} & -1 \\ 0 & 0 & \frac{4}{3} \end{pmatrix}$$

Como no hubo necesidad de intercambio de filas y **todos los pivotes**, $2, \frac{3}{2}, \frac{4}{3}$, resultaron positivos, la matriz dada es definida positiva y por lo tanto tiene factorización de Choleski.

Sea $L = \begin{pmatrix} l_{11} & 0 & 0 \\ l_{21} & l_{22} & 0 \\ l_{31} & l_{32} & l_{33} \end{pmatrix}$ tal que $LL^T = A$, entonces

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

De acuerdo con el algoritmo 3.3, empezamos calculando l_{11}.

$l_{11}^2 = 2$, entonces $l_{11} = \sqrt{2}$.

$l_{21}l_{11} = -1$, entonces $l_{21} = -\frac{1}{l_{11}} = -\frac{1}{\sqrt{2}} = -\frac{\sqrt{2}}{2}$.

$l_{31}l_{11} = 0$, entonces $l_{31} = 0$.

$l_{21}^2 + l_{22}^2 = 2$, entonces $l_{22} = \sqrt{2 - l_{21}^2} = \sqrt{2 - \frac{1}{2}} = \sqrt{\frac{3}{2}} = \frac{\sqrt{3}}{\sqrt{2}} = \frac{\sqrt{6}}{2}$.

$l_{31}l_{21} + l_{32}l_{22} = -1$, entonces $l_{32} = \frac{-1 - l_{31}l_{21}}{l_{22}} = \frac{-1-(0)(-1)}{\frac{\sqrt{6}}{2}} = -\frac{2}{\sqrt{6}} = -\frac{2\sqrt{6}}{6} = -\frac{\sqrt{6}}{3}$.

$l_{31}^2 + l_{32}^2 + l_{33}^2 = 2$, entonces

$$l_{33} = \sqrt{2 - \left(l_{31}^2 + l_{32}^2\right)} = \sqrt{2 - \frac{6}{9}} = \sqrt{2 - \frac{2}{3}} = \sqrt{\frac{4}{3}} = \frac{2}{\sqrt{3}} = \frac{2\sqrt{3}}{3}$$

Así que

$$L = \begin{pmatrix} \sqrt{2} & 0 & 0 \\ -\frac{\sqrt{2}}{2} & \frac{\sqrt{6}}{2} & 0 \\ 0 & -\frac{\sqrt{6}}{3} & \frac{2\sqrt{3}}{3} \end{pmatrix}$$

es tal que $LL^T = A$. ♦

Ejercicio 3.4 Encuentre, si es posible, la factorización directa de Choleski para la siguiente matriz

$$A = \begin{pmatrix} 4 & 1 & 1 & 1 \\ 1 & 3 & -1 & 1 \\ 1 & -1 & 2 & 0 \\ 1 & 1 & 0 & 2 \end{pmatrix}$$ ♦

3.8 MÉTODOS ITERATIVOS

Como ya se mencionó al comienzo de este capítulo, en los métodos iterativos para resolver sistemas de ecuaciones lineales, se parte de una aproximación inicial a la solución, la cual se va "mejorando" sucesivamente aplicando cierto algoritmo. Un método iterativo puede ser convergente o divergente, y aunque el método iterativo converja, en general, sólo podemos esperar la obtención de una solución aproximada, por efecto de los errores de truncamiento y/o redondeo. Entre las ventajas de los métodos iterativos, comparados con los directos, están la simplicidad y uniformidad de las operaciones que se realizan, ya que se usa repetídamente un proceso sencillo; y su relativa insensibilidad al crecimiento de los errores de redondeo, es decir, usualmente los métodos iterativos son estables.

Las **matrices** asociadas con los sistemas de ecuaciones lineales se clasifican en **densas** y **esparcidas**. Las matrices densas tienen pocos elementos nulos y su orden es relativamente pequeño (≤ 100); para sistemas con matrices densas se recomienda usar métodos directos. Las matrices esparcidas tienen pocos elementos no nulos y surgen, por ejemplo, al resolver ecuaciones diferenciales por métodos de diferencias finitas; su orden puede ser muy grande. Los métodos iterativos son recomendados para resolver sistemas con matrices esparcidas.

Los métodos iterativos que estudiaremos son generalizaciones del método de iteración de punto fijo:

Dado un sistema $AX = b$, o equivalentemente, $\underbrace{AX - b}_{F(X)} = 0$, donde A no-singular y $b \neq 0$, lo transformamos en un sistema equivalente $X = \underbrace{BX + c}_{G(X)}$ para alguna matriz B y algún vector c.

Se construye entonces la sucesión de vectores $\left\{X^{(k)}\right\}_k$ a partir de la fórmula de iteración

$$X^{(k)} = BX^{(k-1)} + c, \quad k = 1,2,...$$

y se espera que $\left\{X^{(k)}\right\}_k$ "converja" a la única solución X del sistema $AX = b$ $(\Leftrightarrow X = BX + c)$, donde la convergencia se entiende en el siguiente sentido.

Definición 3.5 Sea $\|\cdot\|$ una norma vectorial en $\mathbf{R}^n$. Decimos que la sucesión $\left\{X^{(k)}\right\}_k$ de vectores de $\mathbf{R}^n$ **converge** al vector $X \in \mathbf{R}^n$, según la norma vectorial $\|\cdot\|$ dada, si $\lim_{k\to\infty} \left\| X - X^{(k)} \right\| = 0$. Ñ

Recuerde que $\lim_{k\to\infty} \left\| X - X^{(k)} \right\| = 0$ significa que, dado $\varepsilon > 0$, existe $N \in \mathbf{N} = \{0,1,2,...\}$ tal que si $k \geq N$, entonces $\left\| X - X^{(k)} \right\| < \varepsilon$. Puede probarse que si una sucesión $\left\{X^{(k)}\right\}_k$ de vectores de $\mathbf{R}^n$ converge al vector $X \in \mathbf{R}^n$ según una norma vectorial $\|\cdot\|$ dada, entonces también converge al vector X según cualquier norma vectorial en $\mathbf{R}^n$, por tal razón diremos simplemente que $\left\{X^{(k)}\right\}_k$ converge al vector X en lugar de $\left\{X^{(k)}\right\}_k$ converge al vector X según la norma $\|\cdot\|$ dada.

Digamos que $\left\{X^{(k)}\right\}_k$ converge al vector X. Si usamos la norma vectorial $\|\cdot\|_\infty$, entonces $\lim_{k\to\infty} \left\| X - X^{(k)} \right\|_\infty = 0$, y si $X = (x_1,...x_i,...,x_n)^T$ y $X^{(k)} = \left(x_1^{(k)},...x_i^{(k)},...,x_n^{(k)}\right)^T$, esto último significa que: dado $\varepsilon > 0$, existe $N \in \mathbf{N}$ tal que si $k \geq N$, entonces $\left| x_i - x_i^{(k)} \right| < \varepsilon$ para todo $i = 1,2,...,n$ (ya que $\left\| X - X^{(k)} \right\|_\infty = \underset{1 \leq i \leq n}{\text{Max}} \left| x_i - x_i^{(k)} \right|$).

Un primer método construido siguiendo la idea anterior es el siguiente:

3.8.1 Método iterativo de Jacobi o de desplazamientos simultáneos: Dado un sistema lineal de ecuaciones

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + ... + a_{1i}x_i + ... + a_{1n}x_n = b_1 \\ a_{21}x_1 + a_{22}x_2 + ... + a_{2i}x_i + ... + a_{2n}x_n = b_2 \\ \vdots \\ a_{i1}x_1 + a_{i2}x_2 + ... + a_{ii}x_i + ... + a_{in}x_n = b_i \\ \vdots \\ a_{n1}x_1 + a_{n2}x_2 + ... + a_{ni}x_i + ... + a_{nn}x_n = b_n \end{cases}$$

Si $a_{ii} \neq 0$ para todo $i = 1,2,...,n$, entonces despejando x_i de la **i**-ésima ecuación, obtenemos el sistema equivalente

$$x_i = \sum_{\substack{j=1 \\ j\neq i}}^{n} \left(-\frac{a_{ij}}{a_{ii}} \right) x_j + \frac{b_i}{a_{ii}}, \quad i = 1,2,...,n$$

Si $B = \left(b_{ij}\right)_{n\times n}$ con

$$b_{ij} = \begin{cases} -\dfrac{a_{ij}}{a_{ii}}, & i \neq j \\ \\ 0, & i = j \end{cases} \qquad i,j = 1,2,...,n$$

y $c = \left(c_1, c_2, ..., c_n\right)^T$ con $c_i = \dfrac{b_i}{a_{ii}}$, $i = 1,2,...,n$, entonces el sistema $AX = b$ dado es equivalente al sistema $X = BX + c$. Así que la sucesión $\left\{ X^{(k)} \right\}_k$ de vectores, correspondiente a este método, se genera a partir de la fórmula de iteración

$$X^{(k)} = BX^{(k-1)} + c, \quad k = 1,2,...$$

donde, conocido $X^{(k-1)} = \left(x_1^{(k-1)}, ..., x_i^{(k-1)}, ..., x_n^{(k-1)} \right)^T$, se calcula la aproximación siguiente $X^{(k)} = \left(x_1^{(k)}, ..., x_i^{(k)}, ..., x_n^{(k)} \right)^T$, mediante la fórmula

$$x_i^{(k)} = \frac{b_i - \sum_{\substack{j=1 \\ j\neq i}}^{n} a_{ij} x_j^{(k-1)}}{a_{ii}}, \quad i = 1,2,...,n \qquad \textbf{(3.12)}$$

que es llamada **fórmula (escalar) de iteración del método de Jacobi**.

Escogida alguna norma vectorial (por ejemplo $\|.\|_\infty$), alguna tolerancia $\varepsilon > 0$ y un número máximo de iteraciones N, se termina el proceso iterativo, indicado en la fórmula (3.12), cuando se satisfaga alguno de los siguientes criterios de aproximación:

i) $\left\| R^{(k)} \right\| < \varepsilon$, siendo $R^{(k)} = AX^{(k)} - b$,

ii) $\left\| X^{(k)} - X^{(k-1)} \right\| < \varepsilon$,

iii) $\dfrac{\left\| X^{(k)} - X^{(k-1)} \right\|}{\left\| X^{(k)} \right\|} < \varepsilon$

o en su defecto, cuando se alcance el número máximo de iteraciones N.

Algoritmo 3.4 (Método de Jacobi) Para encontrar una solución aproximada $\tilde{X}$ de un sistema $AX = b$, con $A = (a_{ij})_{n\times n} \in \mathbf{R}_{n\times n}$ invertible, $b \neq 0$, y $a_{ii} \neq 0$, $i = 1,2,...,n$.

Entrada: El orden **n** del sistema; las componentes (no nulas) a_{ij}, $i,j = 1,2,...,n$ de la matriz A; las componentes b_i, $i = 1,...,n$ del vector de términos independientes; las componentes $x0_i$, $i = 1,2,...,n$ de una aproximación inicial $X0 = X^{(0)}$; una tolerancia Tol y un número máximo de iteraciones N.

Salida: Una solución aproximada $\tilde{X} = (x_1, x_2, ..., x_n)^T$ o un mensaje.

Paso 1: Tomar $k = 1$.

Paso 2: Mientras que $k \leq N$, seguir los pasos 3-6:

Paso 3: Para $i = 1,...,n$, tomar

$$x_i = \frac{b_i - \sum_{\substack{j=1 \\ j\neq i}}^{n} a_{ij} x0_j}{a_{ii}}$$

Paso 4: Si $\| X - X0 \| < Tol$, entonces: **salida:** "Una solución aproximada es $\tilde{X} = (x_1, x_2, ..., x_n)^T$". **Terminar**.

Paso 5: Tomar $k = k + 1$.

Paso 6: Para $i = 1,...,n$, tomar $x0_i = x_i$.

Paso 7: Salida: "Se alcanzó el número máximo de iteraciones N pero no la tolerancia".
Terminar.

ANALISIS DE CONVERGENCIA

Para entrar en el **análisis de la convergencia** del método de Jacobi, empezamos descomponiendo la matriz A en la forma

$$A = D - L - U$$

donde D es la matriz diagonal cuya diagonal es la diagonal principal de A (es decir, $D = \text{diag}(a_{11}, a_{22}, ..., a_{nn})$), $-L$ es la matriz triangular inferior obtenida de la parte triangular estrictamente inferior de A, y $-U$ es la matriz triangular superior obtenida de la parte triangular estrictamente superior de A. Entonces

$$\begin{aligned} AX = b &\Leftrightarrow (D - L - U)X = b \\ &\Leftrightarrow DX - (L + U)X = b \\ &\Leftrightarrow DX = (L + U)X + b \\ &\Leftrightarrow X = D^{-1}(L + U)X + D^{-1}b \end{aligned}$$

lo que nos conduce a la **fórmula vectorial de iteración del método de Jacobi**

$$\boxed{X^{(k)} = D^{-1}(L + U)X^{(k-1)} + D^{-1}b, \quad k = 1,2,...}$$

que se usa para efectos teóricos, mientras que la fórmula (3.12) se usa para los cálculos numéricos.

La matriz $B_J = D^{-1}(L + U)$ se llama **matriz de iteración del método de Jacobi**.

Veamos un caso en el cual la sucesión $\left\{ X^{(k)} \right\}_k$, generada por el esquema iterativo $X^{(k)} = BX^{(k-1)} + c$, converge.

Teorema 3.6 Si $\| B \| < 1$ para alguna norma matricial inducida, entonces la sucesión $\left\{ X^{(k)} \right\}_k$, generada por la fórmula de iteración

$$X^{(k)} = BX^{(k-1)} + c, \quad k = 1,2,...$$

converge a la única solución X del sistema $X = BX + c$, cualquiera sea la aproximación inicial $X^{(0)}$, y se tienen las siguientes cotas para el error de truncamiento $\left\| X - X^{(k)} \right\|$:

i) $\left\| X - X^{(k)} \right\| \le \| B \|^k \left\| X - X^{(0)} \right\|, \; k \ge 1$

ii) $\left\| X - X^{(k)} \right\| \le \dfrac{\| B \|^k}{1 - \| B \|} \left\| X^{(1)} - X^{(0)} \right\|, \; k \ge 1$

iii) $\left\| X - X^{(k)} \right\| \le \dfrac{\| B \|}{1 - \| B \|} \left\| X^{(k)} - X^{(k-1)} \right\|, \; k \ge 1$

Demostración: Para demostrar que el sistema $X = BX + c$ tiene sólo una solución, basta probar que el sistema homogéneo $X = BX$, es decir, $(I - B)X = 0$, tiene solución única $X = 0$, ya que si $(I - B)X = 0$ tiene solución única, entonces $I - B$ es invertible y entonces $(I - B)X = c$ tiene solución única, y como $(I - B)X = c$ es equivalente a $X = BX + c$, entonces este último sistema tiene solución única.

Veamos entonces que el sistema $X = BX$ tiene solución única $X = 0$:

Si X es solución de $X = BX$, entonces

$$0 \le \| X \| = \| BX \| \le \| B \| \| X \|$$

Si $0 < \| B \| < 1$ y $X \neq 0$, entonces $\| B \| \| X \| < \| X \|$, lo cual implicaría que $\| X \| < \| X \|$, lo cual es un absurdo. Luego $X = 0$.

Si $\| B \| = 0$, entonces $B = 0$ y se tendría $X^{(k)} = c$ para todo $k = 1,2,\ldots$, y $\left\{ X^{(k)} \right\}_k$ converge a c, que es la única solución de $X = c$.

Ahora,

$$X - X^{(k)} = B\left(X - X^{(k-1)} \right), \quad k = 1,2,\ldots$$

y usando inducción sobre k se tiene que

$$0 \leq \left\| X - X^{(k)} \right\| \leq \| B \| \left\| X - X^{(k-1)} \right\| \leq \| B \|^2 \left\| X - X^{(k-2)} \right\| \leq \ldots \leq \| B \|^k \left\| X - X^{(0)} \right\| \qquad \textbf{(3.13)}$$

Como $\| B \|^k \to 0$ cuando $k \to \infty$ (pues $0 \leq \| B \| < 1$), entonces $\left\| X - X^{(k)} \right\| \to 0$ cuando $k \to \infty$, es decir, $\left\{ X^{(k)} \right\}_k$ converge a X.

De (3.13), obtenemos

$$\textbf{i)} \quad \left\| X - X^{(k)} \right\| \leq \| B \|^k \left\| X - X^{(0)} \right\|$$

De otro lado,

$$\begin{aligned} \left\| X - X^{(0)} \right\| = \left\| X - X^{(1)} + X^{(1)} - X^{(0)} \right\| &\leq \left\| X - X^{(1)} \right\| + \left\| X^{(1)} - X^{(0)} \right\| \\ &\leq \| B \| \left\| X - X^{(0)} \right\| + \left\| X^{(1)} - X^{(0)} \right\| \end{aligned}$$

(La última desigualdad se tiene por la relación (3.13))

Así que

$$\left(1 - \| B \| \right) \left\| X - X^{(0)} \right\| \leq \left\| X^{(1)} - X^{(0)} \right\|$$

y como $0 \leq \| B \| < 1$, entonces

$$\left\| X - X^{(0)} \right\| \leq \frac{1}{1 - \| B \|} \left\| X^{(1)} - X^{(0)} \right\| \qquad \textbf{(3.14)}$$

y entonces multiplicando a ambos lados de (3.14) por $\| B \|^k$, obtenemos

$$\left\| X - X^{(k)} \right\| \underset{\text{usando } \textbf{i})}{\leq} \| B \|^k \left\| X - X^{(0)} \right\| \leq \frac{\| B \|^k}{1 - \| B \|} \left\| X^{(1)} - X^{(0)} \right\|$$

de donde se obtiene

$$\textbf{ii)} \quad \left\| X - X^{(k)} \right\| \leq \frac{\| B \|^k}{1 - \| B \|} \left\| X^{(1)} - X^{(0)} \right\|$$

Finalmente,

$$\left\| X - X^{(k-1)} \right\| = \left\| X - X^{(k)} + X^{(k)} - X^{(k-1)} \right\| \leq \left\| X - X^{(k)} \right\| + \left\| X^{(k)} - X^{(k-1)} \right\|$$
$$\leq \| B \| \left\| X - X^{(k-1)} \right\| + \left\| X^{(k)} - X^{(k-1)} \right\|$$

así que

$$(1 - \| B \|) \left\| X - X^{(k-1)} \right\| \leq \left\| X^{(k)} - X^{(k-1)} \right\|$$

y como $0 \leq \| B \| < 1$, entonces

$$\left\| X - X^{(k-1)} \right\| \leq \frac{1}{1 - \| B \|} \left\| X^{(k)} - X^{(k-1)} \right\| \qquad \textbf{(3.15)}$$

y entonces multiplicando a ambos lados de (3.15) por $\| B \|$, obtenemos

$$\left\| X - X^{(k)} \right\| \leq \| B \| \left\| X - X^{(k-1)} \right\| \leq \frac{\| B \|}{1 - \| B \|} \left\| X^{(k)} - X^{(k-1)} \right\|$$

lo que implica

iii) $$\left\| X - X^{(k)} \right\| \leq \frac{\| B \|}{1 - \| B \|} \left\| X^{(k)} - X^{(k-1)} \right\| \qquad \tilde{N}$$

Compare este teorema con el teorema 2.1, de punto fijo.

Nota: La desigualdad **ii)** $\left\| X - X^{(k)} \right\| \leq \frac{\| B \|^k}{1 - \| B \|} \left\| X^{(1)} - X^{(0)} \right\|$, en el teorema 3.6 anterior, válida para $\| B \| < 1$, nos dice que entre más "pequeña" sea $\| B \|$, más "rápida" será la convergencia del método iterativo. Las cotas para el error de truncamiento $\left\| X - X^{(k)} \right\|$, dadas en el teorema 3.6, permiten analizar la calidad de una solución aproximada $X^{(k)}$.

A continuación probaremos que la matriz $B_J = (b_{ij})_{n \times n}$, de iteración del metodo de Jacobi, tiene la propiedad $\| B_J \|_\infty < 1$ si la matriz $A = (a_{ij})_{n \times n}$, de coeficientes del sistema $AX = b$, es **E.D.D.** por filas:

Recordemos que en la matriz $B_J = (b_{ij})_{n \times n}$,

$$b_{ij} = \begin{cases} -\dfrac{a_{ij}}{a_{ii}} & , \ i \neq j \\ \\ 0 & , \ i = j \end{cases} \qquad i, j = 1, 2, .., n$$

Ahora, si A es **E.D.D.** por filas, entonces para todo $i = 1,2,\ldots,n$, se tiene

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

luego

$$\sum_{\substack{j=1 \\ j\neq i}}^{n} \left|\frac{a_{ij}}{a_{ii}}\right| = \sum_{\substack{j=1 \\ j\neq i}}^{n} \left|-\frac{a_{ij}}{a_{ii}}\right| < 1,\ i = 1,2,\ldots,n$$

de donde

$$\underset{1\le i\le n}{\text{Max}} \sum_{\substack{j=1 \\ j\neq i}}^{n} \left|-\frac{a_{ij}}{a_{ii}}\right| < 1$$

Pero

$$\|B_J\|_\infty = \underset{1\le i\le n}{\text{Max}} \sum_{j=1}^{n} |b_{ij}| = \underset{1\le i\le n}{\text{Max}} \sum_{\substack{j=1 \\ j\neq i}}^{n} \left|-\frac{a_{ij}}{a_{ii}}\right| \quad (\text{pues } b_{ii} = 0)$$

así que $\|B_J\|_\infty < 1$. Ñ

Aplicando el teorema 3.6 a la situación anterior, obtenemos el siguiente teorema:

Teorema 3.7 Si la matriz A en un sistema $AX = b$ es **E.D.D.** por filas, entonces el método iterativo de Jacobi converge a la única solución X del sistema $AX = b$, cualquiera sea $X^{(0)}$, y se tienen las cotas para el error $\left\| X - X^{(k)} \right\|_\infty$, dadas en el teorema 3.6. Ñ

Se dispone, además, del siguiente resultado cuya demostración puede consultarse en Burden, 1985, páginas 469 y 470:

Teorema 3.8 Para cualquier $X^{(0)} \in \mathbf{R}^n$, la sucesión $\left\{ X^{(k)} \right\}_k$ definida por la fórmula de iteración $X^{(k)} = BX^{(k-1)} + c,\ \ k = 1,2,\ldots,\ c \neq 0$, converge a la única solución X del sistema $X = BX + c$ si y sólo si $\rho(B) < 1$. Ñ

Como una aplicación particular del teorema 3.8 se tiene que: El método iterativo de Jacobi converge a la única solución X del sistema $X = B_J X + c$ si y sólo si $\rho(B_J) < 1$.

Ejemplo 3.8 Use el método iterativo de Jacobi para resolver el sistema

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

Lo primero que observamos es que el método de Jacobi es aplicable a este sistema, ya que la matriz de coeficientes del sistema dado tiene todas sus componentes diagonales no nulas.

Veamos si el método de Jacobi converge o nó, en este caso.

Empecemos observando que la matriz A de coeficientes del sistema **no** es **E.D.D.** por filas (ya que $|2| \leq |-1| + |1|$), así que para el estudio de la convergencia del método de Jacobi debemos encontrar la matriz de iteración B_J .

Una forma de obtener B_J es despejando x_1, x_2 y x_3 de la primera, segunda y tercera ecuación del sistema, respectivamente, y escribiendo el sistema resultante en forma matricial, lo que nos da

$$\underbrace{\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}}_{X} = \underbrace{\begin{pmatrix} 0 & \frac{1}{2} & -\frac{1}{2} \\ -1 & 0 & -3 \\ -\frac{3}{5} & -\frac{3}{5} & 0 \end{pmatrix}}_{B_J} \underbrace{\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}}_{X} + \underbrace{\begin{pmatrix} -\frac{1}{2} \\ 0 \\ \frac{4}{5} \end{pmatrix}}_{c}$$

Como $\|B_J\|_\infty = 4 > 1$ ($\|B_J\|_1 = \frac{7}{2} > 1$) no podemos concluir sobre la convergencia del método de Jacobi (a partir de las normas $\|\cdot\|_\infty$, $\|\cdot\|_1$), así que debemos estudiar el radio espectral $\rho(B_J)$, de la matriz de iteración B_J .

Como

$$\det(B_J - \lambda I) = \begin{vmatrix} -\lambda & \frac{1}{2} & -\frac{1}{2} \\ -1 & -\lambda & -3 \\ -\frac{3}{5} & -\frac{3}{5} & -\lambda \end{vmatrix} = -\lambda^3 + \frac{8}{5}\lambda + \frac{3}{5}$$

entonces la ecuación característica de la matriz B_J es $-\lambda^3 + \frac{8}{5}\lambda + \frac{3}{5} = 0$, cuyas raíces son

$$\lambda_1 = -1,\ \lambda_2 \approx 1.42195 \text{ y } \lambda_3 \approx -.421954$$

y como

$$\rho(B_J) \approx \text{Max}\left\{ |-1|, |1.42195|, |-.421954| \right\} = 1.42195 > 1$$

entonces el **método iterativo de Jacobi no converge (diverge)**, según el teorema 3.8.

Instrucciones en DERIVE:

BJ(A): **S**implifica en la matriz de iteración, B_J , del método de Jacobi.

CHARPOLY(M, w): **S**implifica en el polinomio característico $p_M(w)$ de la matriz M.

EIGENVALUES(M, w): **S**implifica en los valores propios $w_1, w_2, \ldots, w_n$ de la matriz M. à

En situaciones como la del ejemplo 3.8, se recomienda reordenar las ecuaciones del sistema dado, de modo que la matriz de coeficientes del sistema reordenado sea lo más cercana posible a una matriz **E.D.D.** por filas (colocando primero las ecuaciones que tienen un coeficiente dominante y luego las restantes).

En el caso del ejemplo, la reordenación más conveniente para el sistema es

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

Observe que la matriz de coeficientes de este sistema reordenado tampoco es **E.D.D.** por filas.

Las **fórmulas escalares de iteración** para el método de Jacobi son ahora

$$\begin{cases} x_1^{(k)} = \dfrac{-1 + x_2^{(k-1)} - x_3^{(k-1)}}{2}, \\ x_2^{(k)} = \dfrac{4 - 3x_1^{(k-1)} - 5x_3^{(k-1)}}{3}, \\ x_3^{(k)} = \dfrac{-x_1^{(k-1)} - x_2^{(k-1)}}{3}, \end{cases} \qquad k = 1,2,\ldots$$

cuya forma vectorial es

$$\underbrace{\begin{pmatrix} x_1^{(k)} \\ x_2^{(k)} \\ x_3^{(k)} \end{pmatrix}}_{X^{(k)}} = \underbrace{\begin{pmatrix} 0 & \frac{1}{2} & -\frac{1}{2} \\ -1 & 0 & -\frac{5}{3} \\ -\frac{1}{3} & -\frac{1}{3} & 0 \end{pmatrix}}_{B_J} \underbrace{\begin{pmatrix} x_1^{(k-1)} \\ x_2^{(k-1)} \\ x_3^{(k-1)} \end{pmatrix}}_{X^{(k-1)}} + \underbrace{\begin{pmatrix} -\frac{1}{2} \\ \frac{4}{3} \\ 0 \end{pmatrix}}_{c}, \qquad k = 1,2,\ldots$$

Como $\| B_J \|_\infty = \dfrac{8}{3} > 1$ ($\| B_J \|_1 = \dfrac{13}{6} > 1$), entonces nada se puede afirmar sobre la convergencia del método de Jacobi (a partir de las normas $\| B_J \|_\infty$, $\| B_J \|_1$), por tanto debemos encontrar $\rho(B_J)$. La ecuación característica de B_J es

$$-\lambda^3 + \frac{2}{9}\lambda + \frac{1}{9} = 0$$

cuyas raíces son

$$\lambda_1 \approx .631096 \ , \ \lambda_2 \approx -.315548 + .276567i \ , \ \lambda_3 \approx -.315548 - .276567i$$

y como

$$\begin{aligned} \rho(B_J) &\approx \text{Max}\{\, |\, .631096\,|\,,\ |\, -.315548 + .276567i\,|\,,\ |\, -.315548 - .276567i\,|\,\} \\ &\approx \text{Max}\{\, .631096,\ .419595\,\} = .631096 < 1 \end{aligned}$$

entonces, según el teorema 3.8, el método iterativo de Jacobi converge a la única solución X del sistema reordenado, cualquiera sea la aproximación inicial $X^{(0)}$.

Iterando con el método de Jacobi, tomando como aproximación inicial $X^{(0)} = (0,0,0)^T$, y usando como criterio de aproximación $\left\| X^{(k)} - X^{(k-1)} \right\|_\infty < .001$, obtenemos

$$X^{(1)} = (-.5,\ 1.3333,\ 0)^T\ ,\ X^{(2)} = (.16667,\ 1.8333,\ -.27778)^T\ ,\ldots,$$
$$X^{(15)} = (.99798,\ 1.9990,\ -.99842)^T\ ,\ X^{(16)} = (.99873,\ 1.9994,\ -.99901)^T$$

Como $\left\| X^{(16)} - X^{(15)} \right\|_\infty \approx 7.5E-4 < .001$ y $k = 16$ es el primer entero positivo para el cual $\left\| X^{(k)} - X^{(k-1)} \right\|_\infty < .001$, entonces

$$X^{(16)} = (.99873,\ 1.9994,\ -.99901)^T = \tilde{X} \approx X$$

Instrucción en DERIVE:

JACOBI(A, b, $X^{(0)}$, N): apro**X**ima las primeras N iteraciones en el método de Jacobi aplicado al sistema $AX = b$, con aproximación inicial $X^{(0)}$. Para el ejemplo, apro**X**ime la expresión JACOBI($[[2,-1,1],[3,3,5],[1,1,3]],[-1,4,0],[0,0,0],16$). à

Cuál es la **calidad de la solución aproximada** obtenida $X^{(16)}$?

Como no se pueden aplicar las cotas para el error de truncamiento $\left\| X^{(16)} - X \right\|$, dadas en el teorema 3.6 (ya que no se satisface la condición $\|B_J\| < 1$ para las normas calculadas $\|B_J\|_\infty$, $\|B_J\|_1$), entonces vamos a usar las cotas para el error relativo $\dfrac{\left\| X^{(16)} - X \right\|_\infty}{\|X\|_\infty}$, dadas en el teorema 3.3:

$$\frac{\left\| R^{(16)} \right\|_\infty}{\|b\|_\infty} \frac{1}{\text{Cond}_\infty(A)} \le \frac{\left\| X^{(16)} - X \right\|_\infty}{\|X\|_\infty} \le \text{Cond}_\infty(A) \frac{\left\| R^{(16)} \right\|_\infty}{\|b\|_\infty}$$

En esta desigualdad $R^{(16)} = AX^{(16)} - b$, con

$$A = \begin{pmatrix} 2 & -1 & 1 \\ 3 & 3 & 5 \\ 3 & 3 & 9 \end{pmatrix} \quad y \quad b = \begin{pmatrix} -1 \\ 4 \\ 0 \end{pmatrix}$$

Si hacemos los cálculos indicados, obtenemos

$$5\times10^{-6} < 1.6...\times10^{-5} \le \frac{\left\| X^{(16)} - X \right\|_\infty}{\left\| X \right\|_\infty} \le .011.. < .05 = 5\times10^{-2}$$

Extendiendo de manera natural, los conceptos de cifras significativas y cifras decimales exactas para vectores, dados en el capítulo 1 para escalares, podemos concluir, a partir de la última desigualdad, que $X^{(16)} = (.99873, 1.9994, -.99901)^T$ aproxima a la solución exacta X del sistema dado con una precisión de por lo menos dos cifras significativas (y no más de cinco). Se puede verificar que la solución exacta del sistema dado es $X = (1,\ 2,\ -1)^T$. ♦

Ejemplo 3.9 Use el método iterativo de Jacobi para resolver el sistema

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

Observe que el método de Jacobi es aplicable a este sistema, pero nuevamente como en el ejemplo anterior, la matriz de coeficientes del sistema dado no es **E.D.D.** por filas, así que para estudiar la convergencia del método de Jacobi debemos encontrar la matriz de iteración B_J.

Se ve fácilmente que la matriz de iteración del método de Jacobi es, en este caso, la siguiente matriz

$$B_J = \begin{pmatrix} 0 & \frac{1}{2} & -5 & 0 \\ 0 & 0 & \frac{1}{3} & -\frac{8}{3} \\ -5 & \frac{1}{2} & 0 & 0 \\ \frac{1}{3} & -\frac{11}{3} & \frac{1}{3} & 0 \end{pmatrix}$$

Como $\left\| B_J \right\|_\infty = \frac{11}{2} > 1$ ($\left\| B_J \right\|_1 > 1$), entonces todavía no podemos concluir acerca de la convergencia del método de Jacobi (apartir de las normas $\left\| B_J \right\|_\infty$, $\left\| B_J \right\|_1$). Encontremos entonces el radio espectral de la matriz de iteración B_J.

Se puede ver que la ecuación característica de la matriz B_J es

$$\lambda^4 - \frac{629}{18}\lambda^2 + \frac{31}{18}\lambda + 240 = 0$$

cuyas raíces son

$$\lambda_1 = 5,\ \lambda_2 \approx -3.01292,\ \lambda_3 \approx 3.11967 \text{ y } \lambda_4 \approx -5.10674$$

por tanto

$$\rho(B_J) \approx 5.10674 > 1$$

lo que implica que el **método de Jacobi diverge**, en este caso.

Sin embargo, si reordenamos las ecuaciones del sistema en la forma

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

obtenemos un sistema equivalente $AX = b$ donde A sí es **E.D.D.** por filas, así que, según el teorema 3.7, el método iterativo de Jacobi converge a la única solución del sistema, cualquiera sea la aproximación inicial $X^{(0)}$, y se tienen cotas para el error de truncamiento $\left\| X^{(k)} - X \right\|_\infty$.

La forma matricial del método de Jacobi, para el sistema reordenado, es

$$\underbrace{\begin{pmatrix} x_1^{(k)} \\ x_2^{(k)} \\ x_3^{(k)} \\ x_4^{(k)} \end{pmatrix}}_{X^{(k)}} = \underbrace{\begin{pmatrix} 0 & \frac{1}{10} & -\frac{2}{10} & 0 \\ \frac{1}{11} & 0 & \frac{1}{11} & -\frac{3}{11} \\ -\frac{2}{10} & \frac{1}{10} & 0 & 0 \\ 0 & -\frac{3}{8} & \frac{1}{8} & 0 \end{pmatrix}}_{B_J} \underbrace{\begin{pmatrix} x_1^{(k-1)} \\ x_2^{(k-1)} \\ x_3^{(k-1)} \\ x_4^{(k-1)} \end{pmatrix}}_{X^{(k-1)}} + \underbrace{\begin{pmatrix} \frac{6}{10} \\ \frac{25}{11} \\ -\frac{11}{10} \\ -\frac{11}{8} \end{pmatrix}}_{c}$$

Observe que $\| B_J \|_\infty = \frac{4}{8} = \frac{1}{2} < 1$ ($\|B_J\|_1 = \frac{23}{40} < 1$).

Investigue **cuántas iteraciones k** serán necesarias en el método de Jacobi (usando la norma $\|\cdot\|_\infty$), para que $X^{(k)}$ aproxime a la solución exacta X del sistema dado, con por lo menos tres cifras significativas, tomando como aproximación inicial $X^{(0)} = (0,0,0,0)^T$? (**ejercicio!**)

Los resultados obtenidos en las iteraciones aplicando el método de Jacobi, empezando con $X^{(0)} = (0,0,0,0)^T$ y usando como criterio de aproximación $\left\| X^{(k)} - X^{(k-1)} \right\|_\infty < .001$, son

$$X^{(1)} = (.60000,\ 2.2727,\ -1.1000,\ -1.3750)^T$$
$$X^{(2)} = (1.0473,\ 2.6023,\ -.99273,\ -2.3648)^T$$
$$\vdots$$
$$X^{(8)} = (1.1040,\ 2.9958,\ -1.0210,\ -2.6262)^T$$
$$X^{(9)} = (1.1038,\ 2.9965,\ -1.0212,\ -2.6261)^T = \tilde{X} \approx X$$

Sólo fueron necesarias $k = 9$ iteraciones para alcanzar la tolerancia dada. Con cuántas cifras decimales exactas aproxima $X^{(9)}$ a X? (**ejercicio**) ♦

3.8.2 Método iterativo de Gauss-Seidel o de desplazamientos sucesivos: Una posible mejora en el algoritmo de Jacobi puede ser la siguiente: para calcular $x_i^{(k)}$ se usan las componentes de $X^{(k-1)}$, pero como $x_1^{(k)}, x_2^{(k)}, \ldots, x_{i-1}^{(k)}$ ya han sido calculadas y supuestamente son mejores aproximaciones de las componentes $x_1, x_2, \ldots, x_{i-1}$ de la solución exacta que $x_1^{(k-1)}, \ldots, x_{i-1}^{(k-1)}$ (asumiendo convergencia), parece más recomendable calcular $x_i^{(k)}$ usando los valores $x_1^{(k)}, x_2^{(k)}, \ldots, x_{i-1}^{(k)}$ calculados recientemente. Esta técnica se conoce como **método iterativo de Gauss-Seidel** o **de desplazamientos sucesivos**.

Se inicia el proceso iterativo con una aproximación inicial $X^{(0)} = \left(x_1^{(0)}, x_2^{(0)}, \ldots, x_i^{(0)}, \ldots, x_n^{(0)}\right)^T$. A partir de este vector se obtiene la primera aproximación $X^{(1)} = \left(x_1^{(1)}, x_2^{(1)}, \ldots, x_i^{(1)}, \ldots, x_n^{(1)}\right)^T$, mediante las siguientes fórmulas (suponemos que $a_{ii} \neq 0 \quad para \ todo \ i = 1,2,\ldots,n$):

$$x_1^{(1)} = \frac{b_1 - \sum_{j=2}^{n} a_{1j} x_j^{(0)}}{a_{11}} \quad , \qquad x_2^{(1)} = \frac{b_2 - a_{21} x_1^{(1)} - \sum_{j=3}^{n} a_{2j} x_j^{(0)}}{a_{22}}$$

y en general, para $i = 2, \ldots, n-1$, se calcula

$$x_i^{(1)} = \frac{b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(1)} - \sum_{j=i+1}^{n} a_{ij} x_j^{(0)}}{a_{ii}}$$

y

$$x_n^{(1)} = \frac{b_n - \sum_{j=1}^{n-1} a_{nj} x_j^{(1)}}{a_{nn}}$$

El paso genérico es:

Conocida la aproximación $X^{(k-1)} = \left(x_1^{(k-1)}, \ldots, x_i^{(k-1)}, \ldots, x_n^{(k-1)}\right)^T$, se obtiene la aproximación siguiente $X^{(k)} = \left(x_1^{(k)}, \ldots, x_i^{(k)}, \ldots, x_n^{(k)}\right)^T$, usando las fórmulas

$$\boxed{x_1^{(k)} = \frac{b_1 - \sum_{j=2}^{n} a_{1j} x_j^{(k-1)}}{a_{11}}}$$

para $i = 2,3,\ldots,n-1$,

$$x_i^{(k)} = \frac{b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(k)} - \sum_{j=i+1}^{n} a_{ij} x_j^{(k-1)}}{a_{ii}} \qquad \textbf{(3.16)}$$

y

$$x_n^{(k)} = \frac{b_n - \sum_{j=1}^{n-1} a_{nj} x_j^{(k)}}{a_{nn}}$$

que son llamadas **fórmulas escalares de iteración del método de Gauss-Seidel**.

Se termina el proceso iterativo con alguno de los criterios de aproximación mencionados anteriormente.

Algoritmo 3.5 (Método de Gauss-Seidel) Para encontrar una solución aproximada $\tilde{X}$ de un sistema $AX = b$, $A = (a_{ij})_{n\times n} \in \mathbf{R}_{n\times n}$ invertible, $b \neq 0$ y $a_{ii} \neq 0$ para todo $i = 1,2,...,n$.

Entrada: El orden **n** del sistema; las componentes no nulas a_{ij}, $i,j = 1,2,...,n$ de la matriz A; las componentes b_i, $i = 1,2,...,n$ del vector de términos independientes; las componentes $x0_i$, $i = 1,2,...,n$ de una aproximación inicial $X0 = X^{(0)}$; una tolerancia Tol, y un número máximo de iteraciones N.

Salida: Una solución aproximada $\tilde{X} = (x_1, x_2, ..., x_n)^T$ o un mensaje.

Paso 1: Tomar $k = 1$.

Paso 2: Mientras que $k \leq N$, seguir los pasos 3-8:

Paso 3: Tomar $x_1 = \dfrac{b_1 - \sum_{j=2}^{n} a_{1j} x0_j}{a_{11}}$.

Paso 4: Para $i = 2,..., n-1$, tomar

$$x_i = \frac{b_i - \sum_{j=1}^{i-1} a_{ij} x_j - \sum_{j=i+1}^{n} a_{ij} x0_j}{a_{ii}}$$

Paso 5: Tomar $x_n = \dfrac{b_n - \sum_{j=1}^{n-1} a_{nj} x_j}{a_{nn}}$.

Paso 6: Si $\| X - X0 \| < Tol$, entonces **salida:** "Una solución aproximada del sistema es $\tilde{X} = (x_1, x_2, ..., x_n)^T$". **Terminar**.

Paso 7: Tomar $k = k + 1$.

Paso 8: Para $i = 1,2,...,n$, tomar $x0_i = x_i$.

Paso 9: Salida: "Se alcanzó el número máximo de iteraciones N pero no la tolerancia". **Terminar**.

ANÁLISIS DE CONVERGENCIA

Al igual que en el método de Jacobi, con el propósito de analizar la convergencia del método de Gauss-Seidel, veamos cuál es la fórmula vectorial de iteración del método de Gauss-Seidel

$$X^{(k)} = BX^{(k-1)} + c, \quad k = 1,2,...$$

Multiplicando a ambos lados de la ecuación (3.16) por a_{ii} y asociando los **k**-ésimos términos iterados, obtenemos

$$\sum_{j=1}^{i} a_{ij} x_j^{(k)} = \sum_{j=i+1}^{n} (-a_{ij}) x_j^{(k-1)} + b_i \qquad \textbf{(3.17)}$$

Si D es la matriz diagonal cuya diagonal es la diagonal de A, $-L$ es la matriz triangular inferior formada por la parte estrictamente inferior de A y $-U$ es la matriz triangular superior formada por la parte estrictamente superior de A, como en el método de Jacobi, entonces al poner a variar **i** de **1** a **n** en la ecuación (3.17), obtenemos el sistema

$$(D - L)X^{(k)} = UX^{(k-1)} + b$$

o equivalentemente

$$\boxed{X^{(k)} = \underbrace{(D - L)^{-1}U}_{B_G} X^{(k-1)} + (D - L)^{-1}b, \quad k = 1,2,...}$$

siempre que la matriz $D - L$ (triangular inferior) sea invertible, o sea si $a_{ii} \neq 0$ para cada $i = 1,2,...,n$.

La fórmula anterior se conoce como **fórmula vectorial de iteración del método de Gauss-Seidel**, y la matriz $B_G = (D - L)^{-1}U$ se llama **matriz de iteración del método de Gauss-Seidel**.

Al igual que en el método de Jacobi, se tiene para el método de Gauss-Seidel el siguiente resultado, el cual puede ser consultado en Kincaid, 1972, páginas 189 y 190.

Teorema 3.9 Si la matriz A de coeficientes de un sistema $AX = b$ es **E.D.D.** por filas, entonces el método iterativo de Gauss-Seidel converge a la única solución X del sistema $AX = b$, para cualquier elección de $X^{(0)}$. Ñ

Recuérdese que según el teorema 3.6: Si $\| B_G \| < 1$ para alguna norma matricial inducida, entonces la sucesión $\left\{X^{(k)}\right\}_k$, $k = 0,1,...$, obtenida en el método de Gauss-Seidel, converge a la única solución X del sistema $X = B_G X + c$ para cualquier $X^{(0)} \in \mathbf{R}^n$, y se tienen las cotas para el error de truncamiento $\left\| X - X^{(k)} \right\|$, dadas en el teorema 3.6. Ñ

También se tiene, de acuerdo con el teorema 3.8, que: Para cualquier $X^{(0)} \in \mathbf{R}^n$, la sucesión $\left\{X^{(k)}\right\}_k$, con

$$X^{(k)} = B_G X^{(k-1)} + c, \quad k = 1,2,..., \; c \neq 0$$

converge a la única solución X del sistema $X = B_G X + c$ $(\Leftrightarrow AX = b)$ si y sólo si $\rho(B_G) < 1$. Ñ

Ejemplo 3.10 Apliquemos el método de Gauss-Seidel para resolver el sistema

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

Este sistema es el mismo del ejemplo 3.8.

Como la matriz de coeficientes no es **E.D.D.** por filas, nada podemos decir todavía acerca de la convergencia del método de Gauss-Seidel, así que debemos encontrar la matriz de iteración del método de Gauss-Seidel, B_G. Una forma de encontrarla es la siguiente:

Las fórmulas escalares de iteración del método de Gauss-Seidel para el sistema dado, son

$$\begin{cases} x_1^{(k)} = \dfrac{-1 + x_2^{(k-1)} - x_3^{(k-1)}}{2}, \\ x_2^{(k)} = -x_1^{(k)} - 3x_3^{(k-1)}, \qquad k = 1,2,... \\ x_3^{(k)} = \dfrac{4 - 3x_1^{(k)} - 3x_2^{(k)}}{5}, \end{cases}$$

Reemplazando $x_1^{(k)}$ en la fórmula de iteración para $x_2^{(k)}$, obtenemos la siguiente expresión para $x_2^{(k)}$, en términos de $x_2^{(k-1)}$ y $x_3^{(k-1)}$:

$$x_2^{(k)} = \frac{1 - x_2^{(k-1)} - 5x_3^{(k-1)}}{2}$$

Ahora se reemplazan $x_1^{(k)}$ y la última expresión obtenida para $x_2^{(k)}$, en la fórmula de iteración para $x_3^{(k)}$, con lo cual se obtiene

$$x_3^{(k)} = \frac{4 + 9x_3^{(k-1)}}{5}$$

Así que, finalmente, se tiene el siguiente esquema de iteración

$$\begin{cases} x_1^{(k)} = \dfrac{-1 + x_2^{(k-1)} - x_3^{(k-1)}}{2}, \\ x_2^{(k)} = \dfrac{1 - x_2^{(k-1)} - 5x_3^{(k-1)}}{2}, \\ x_3^{(k)} = \dfrac{4 + 9x_3^{(k-1)}}{5}, \end{cases} \qquad k = 1,2,\ldots$$

el cual escrito en forma vectorial es

$$\underbrace{\begin{pmatrix} x_1^{(k)} \\ x_2^{(k)} \\ x_3^{(k)} \end{pmatrix}}_{X^{(k)}} = \underbrace{\begin{pmatrix} 0 & \frac{1}{2} & -\frac{1}{2} \\ 0 & -\frac{1}{2} & -\frac{5}{2} \\ 0 & 0 & \frac{9}{5} \end{pmatrix}}_{B_G} \underbrace{\begin{pmatrix} x_1^{(k-1)} \\ x_2^{(k-1)} \\ x_3^{(k-1)} \end{pmatrix}}_{X^{(k-1)}} + \underbrace{\begin{pmatrix} -\frac{1}{2} \\ \frac{1}{2} \\ \frac{4}{5} \end{pmatrix}}_{c}, \quad k = 1,2,\ldots$$

que constituye la fórmula vectorial de iteración del método de Gauss-Seidel para el sistema dado.

Otra forma de obtener la matriz de iteración B_G, es a partir de su fórmula $B_G = (D-L)^{-1}U$.

Como $D-L = \begin{pmatrix} 2 & 0 & 0 \\ 1 & 1 & 0 \\ 3 & 3 & 5 \end{pmatrix}$, entonces $(D-L)^{-1} = \begin{pmatrix} \frac{1}{2} & 0 & 0 \\ -\frac{1}{2} & 1 & 0 \\ 0 & -\frac{3}{5} & \frac{1}{5} \end{pmatrix}$, y como $U = \begin{pmatrix} 0 & 1 & -1 \\ 0 & 0 & -3 \\ 0 & 0 & 0 \end{pmatrix}$, entonces

$$B_G = (D-L)^{-1}U = \begin{pmatrix} 0 & \frac{1}{2} & -\frac{1}{2} \\ 0 & -\frac{1}{2} & -\frac{5}{2} \\ 0 & 0 & \frac{9}{5} \end{pmatrix}$$

Como $\| B_G \|_\infty = 3 > 1$ ($\| B_G \|_1 > 1$), todavía no podemos concluir acerca de la convergencia del método de Gauss-Seidel; pero como $\rho(B_G) = \text{Max}\left\{ 0, \left| -\frac{1}{2} \right|, \left| \frac{9}{5} \right| \right\} = \frac{9}{5} > 1$, entonces el método de Gauss-Seidel diverge (recuerde que si una matriz es triangular superior o inferior, entonces los valores propios de tal matriz son los números que aparecen en su diagonal principal).

Observe que **el número cero es siempre un valor propio de la matriz de iteración** B_G, así que, en particular, B_G siempre es singular (no invertible).

Instrucción en DERIVE:

BG(A): Simplifica en la matriz de iteración, B_G , del método de Gauss-Seidel. à

Una reordenación del sistema dado, de modo que la matriz de coeficientes del sistema resultante, sea lo más cercana posible a una matriz **E.D.D.** por filas, es

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

Las fórmulas escalares de iteración del método de Gauss-Seidel para encontrar una aproximación de la solución de este sistema reordenado, son

$$\begin{cases} x_1^{(k)} = \dfrac{-1 + x_2^{(k-1)} - x_3^{(k-1)}}{2}, \\ x_2^{(k)} = \dfrac{4 - 3x_1^{(k)} - 5x_3^{(k-1)}}{3}, \qquad k = 1,2,... \\ x_3^{(k)} = \dfrac{-x_1^{(k)} - x_2^{(k)}}{3}, \end{cases}$$

Dado que

$$D - L = \begin{pmatrix} 2 & 0 & 0 \\ 3 & 3 & 0 \\ 1 & 1 & 3 \end{pmatrix}, \qquad (D-L)^{-1} = \begin{pmatrix} \frac{1}{2} & 0 & 0 \\ -\frac{1}{2} & \frac{1}{3} & 0 \\ 0 & -\frac{1}{9} & \frac{1}{3} \end{pmatrix} \quad \text{y} \quad U = \begin{pmatrix} 0 & 1 & -1 \\ 0 & 0 & -5 \\ 0 & 0 & 0 \end{pmatrix}$$

entonces

$$B_G = (D-L)^{-1}U = \begin{pmatrix} 0 & \frac{1}{2} & -\frac{1}{2} \\ 0 & -\frac{1}{2} & -\frac{7}{6} \\ 0 & 0 & \frac{5}{9} \end{pmatrix}$$

Como $\| B_G \|_\infty > 1$ ($\| B_G \|_1 > 1$), todavía no podemos concluir sobre la convergencia del método de Gauss-Seidel, pero como $\rho(B_G) = \text{Max}\left\{ 0, \left| -\frac{1}{2} \right|, \left| \frac{5}{9} \right| \right\} = \frac{5}{9} < 1$, entonces el método de Gauss-Seidel converge a la única solución del sistema dado, cualquiera sea la aproximación inicial.

Si usamos el método de Gauss-Seidel con aproximación inicial $X^{(0)} = (0,0,0)^T$ y criterio de aproximación $\left\| X^{(k)} - X^{(k-1)} \right\|_\infty < .001$, se obtienen los siguientes resultados:

$$X^{(1)} = (-.50000,\ 1.8333,\ -.44444)^T, \quad X^{(2)} = (.63889,\ 1.4352,\ -.69136)^T, ...$$
$$X^{(12)} = (.99858,\ 1.9988,\ -.99914)^T, \quad X^{(13)} = (.99898,\ 1.9996,\ -.99952)^T = \tilde{X} \approx X$$

Aquí $k = 13$ es el menor número de iteraciones para el cual se satisface $\left\| X^{(k)} - X^{(k-1)} \right\|_\infty < .001$.

Al igual que en el ejemplo 3.8, nos podemos preguntar por la calidad de la solución aproximada obtenida $X^{(13)}$ (**ejercicio!**). ♦

Instrucción en DERIVE:

G_SEIDEL($A, b, X^{(0)}, N$): apro**X**ima las primeras N iteraciones en el método de Gauss-Seidel aplicado al sistema $AX = b$, tomando como aproximación inicial $X^{(0)}$. Para el ejemplo, apro**X**ime la expresión G_SEIDEL($[[2,-1,1],[3,3,5],[1,1,3]], [-1,4,0], [0,0,0], 13$). à

Ejemplo 3.11 Si aplicamos el método de Gauss-Seidel para resolver el sistema

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

(que es el mismo sistema del ejemplo 3.9), encontramos que la matriz de coeficientes de este sistema no es **E.D.D.** por filas, así que para estudiar la convergencia del método de Gauss-Seidel debemos considerar la matriz de iteración B_G.

Se puede ver que

$$B_G = (D-L)^{-1}U = \begin{pmatrix} 0 & \frac{1}{2} & -5 & 0 \\ 0 & 0 & \frac{1}{3} & -\frac{8}{3} \\ 0 & -\frac{5}{2} & \frac{151}{6} & -\frac{4}{3} \\ 0 & -\frac{2}{3} & \frac{11}{2} & \frac{28}{3} \end{pmatrix}$$

y como $\| B_G \|_\infty > 1$ ($\| B_G \|_1 > 1$), entonces por ahora nada podemos afirmar acerca de la convergencia del método de Gauss-Seidel; pero se puede ver que la ecuación característica de la matriz de iteración B_G, es

$$\lambda^4 - \frac{621}{18}\lambda^3 + \frac{4343}{18}\lambda^2 = 0$$

cuyas raíces son $\lambda_{1,2} = 0$ (es decir, $\lambda = 0$ es raíz doble), $\lambda_3 \approx 24.7523$, $\lambda_4 \approx 9.74768$. Por lo tanto $\rho(B_G) > 1$, lo que implica que el método de Gauss-Seidel diverge.

Si reordenamos el sistema de modo que la matriz de coeficientes del nuevo sistema equivalente sea lo más cercana posible a una matriz **E.D.D.** por filas, obtenemos el sistema

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

cuya matriz de coeficientes es **E.D.D.** por filas. Así que el método de Gauss-Seidel converge a la única solución del sistema dado, cualquiera sea la aproximación inicial y se tienen cotas para el error de truncamiento $\left\| X^{(k)} - X \right\|$, según el teorema 3.9. Si iteramos con el método de Gauss-Seidel, empezando con $X^{(0)} = (0,0,0,0)^T$ y usando como criterio de aproximación $\left\| X^{(k)} - X^{(k-1)} \right\|_\infty < .001$, se obtienen los siguientes resultados:

$$X^{(1)} = (.60000,\ 2.3273,\ -.98727,\ -2.3711)^T$$
$$X^{(2)} = (1.0302,\ 2.9233,\ -1.0137,\ -2.5980)^T$$
$$\vdots$$
$$X^{(5)} = (1.1038,\ 2.9964,\ -1.0211,\ -2.6263)^T = \tilde{X} \approx X \quad \blacklozenge$$

Si comparamos los resultados de los ejemplos 3.8 y 3.9, obtenidos por el método de Jacobi, con los obtenidos en los ejemplos 3.10 y 3.11 por el método de Gauss-Seidel, vemos que el método de Gauss-Seidel es de convergencia más rápida; esto es lo que generalmente ocurre cuando ambos métodos convergen. Anotamos que hay sistemas lineales para los cuales un método converge y el otro diverge.

3.8.3 Método SOR (Successive Over-Relaxation): Este método fue ideado para acelerar la convergencia del método de Gauss-Seidel. La idea del método es que para producir un nuevo valor $x_i^{(k)}$ se **ponderan** los valores $x_i^{(k)}$ actual, obtenido por Gauss-Seidel, y $x_i^{(k-1)}$ anterior, como se indica a continuación:

Dada una aproximación inicial $X^{(0)}$ y calculada la aproximación $X^{(k-1)} = \left(x_1^{(k-1)}, x_2^{(k-1)}, \ldots, x_i^{(k-1)}, \ldots, x_n^{(k-1)} \right)^T$, se calcula la aproximación siguiente $X^{(k)} = \left(x_1^{(k)}, x_2^{(k)}, \ldots, x_i^{(k)}, \ldots, x_n^{(k)} \right)^T$, de acuerdo con las fórmulas siguientes:

$$x_1^{(k)} = (1-w)x_1^{(k-1)} + w\left(\frac{b_1 - \sum_{j=2}^{n} a_{1j} x_j^{(k-1)}}{a_{11}} \right)$$

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

$$x_i^{(k)} = (1-w)x_i^{(k-1)} + w\left(\frac{b_i - \sum_{j=1}^{i-1} a_{ij}x_j^{(k)} - \sum_{j=i+1}^{n} a_{ij}x_j^{(k-1)}}{a_{ii}}\right) \qquad \textbf{(3.18)}$$

y

$$x_n^{(k)} = (1-w)x_n^{(k-1)} + w\left(\frac{b_n - \sum_{j=1}^{n-1} a_{nj}x_j^{(k)}}{a_{nn}}\right)$$

donde **w** es un parámetro, llamado de aceleración. Más adelante mostraremos que la variación de **w** debe ser $0 < w < 2$, para que el método pueda converger.

Las fórmulas anteriores son llamadas **fórmulas escalares de iteración del método SOR**.

Para $0 < w < 1$ el método se denomina de **sub-relajación** y se puede usar para obtener convergencia en **algunos** sistemas para los cuales el método de Gauss-Seidel no es convergente.

Para $1 < w < 2$ el método se denomina de **sobre-relajación** y se puede usar para acelerar la convergencia en **algunos** sistemas que son convergentes por el método de Gauss-Seidel.

Observe que si $w = 1$, el método se convierte en el método de Gauss-Seidel.

La escogencia del valor óptimo de w se hace de modo que $\rho(B_w)$ sea mínimo, donde B_w es la **matriz de iteración del método SOR**.

Se puede obtener, siguiendo la misma idea que se usó en el método de Gauss-Seidel, la **fórmula vectorial de iteración del método SOR**:

$$X^{(k)} = \underbrace{(D - wL)^{-1}\left[(1-w)D + wU\right]}_{B_w} X^{(k-1)} + w(D - wL)^{-1}b \qquad \textbf{(3.19)}$$

donde D, $-L$ y $-U$ son matrices tales que $A = D - L - U$, como en los métodos de Jacobi y Gauss-Seidel.

Dada la dificultad de obtener el w óptimo, a menudo se trabaja experimentando con distintos valores de w.

La variación del parámetro w, con $0 < w < 2$, se debe al siguiente resultado:

Si $a_{ii} \neq 0$ para cada $i = 1,2,\ldots,n$, entonces $\rho(B_w) \geq |w - 1|$.

En efecto:

$$\det(B_w) = \det\left\{ (D - wL)^{-1}\left[(1-w)D + wU\right] \right\}$$
$$= \det\left[(D - wL)^{-1}\right]\det\left[(1-w)D + wU\right]$$

Pero

$$\det\left[(D - wL)^{-1}\right] = \frac{1}{\det(D - wL)} = \frac{1}{\det D}$$

y

$$\det\left[(1-w)D + wU\right] = \det\left[(1-w)D\right] = (1-w)^n \det D$$

así que

$$\det(B_w) = \frac{1}{\det D}(1-w)^n \det D = (1-w)^n$$

De otro lado, como $\det(B_w) = \lambda_1\lambda_2...\lambda_n$ donde $\lambda_1, \lambda_2, ..., \lambda_n$ son los valores propios de la matriz B_w, entonces

$$\left[\rho(B_w)\right]^n \geq |\lambda_1||\lambda_2|...|\lambda_n| = |\lambda_1\lambda_2...\lambda_n| = |\det(B_w)| = \left|(1-w)^n\right| = |1-w|^n$$

y por tanto $\rho(B_w) \geq |w-1|$. Ñ

Como consecuencia del resultado anterior, si $|w-1| \geq 1$, es decir, si $w \geq 2$ o $w \leq 0$, entonces $\rho(B_w) \geq 1$ y entonces el método **SOR** diverge, según el teorema 3.8. Luego sólo si $0 < w < 2$ ($|w-1| < 1$), es posible que $\rho(B_w) < 1$ y así el método **SOR posiblemente** sea convergente.

En el siguiente teorema, cuya prueba puede consultarse en Ortega, 1990, página 123, se establecen condiciones suficientes para la convergencia del método SOR:

Teorema 3.10 Si A es una matriz real, simétrica y definida positiva, entonces el método **SOR** converge a la única solución X del sistema $AX = b$ para cualquier elección de la aproximación inicial $X^{(0)} \in \mathbf{R}^n$ y cualquier valor de w con $0 < w < 2$. Ñ

También se tiene, como en los métodos de Jacobi y Gauss-Seidel, que:

Si $\|B_w\| < 1$ para alguna norma matricial inducida, entonces el método **SOR** converge a la única solución X del sistema $X = B_w X + c$, $c \neq 0$, cualquiera sea la aproximación inicial $X^{(0)} \in \mathbf{R}^n$, y se tienen las cotas para el error de truncamiento $\left\|X - X^{(k)}\right\|$, dadas en el teorema 3.6. Ñ

Para cualquier $X^{(0)} \in \mathbf{R}^n$, el método **SOR** converge a la única solución X del sistema $X = B_w X + c$, $c \neq 0$ si y sólo si $\rho(B_w) < 1$. Ñ

Algoritmo 3.6 (Método SOR) Para encontrar una solución aproximada $\tilde{X}$ de un sistema $AX = b$ con A invertible, $b \neq 0$ y $a_{ii} \neq 0$ para todo $i = 1,2,...,n$, dado un valor del parámetro w con $0 < w < 2$:

Entrada: El orden **n** del sistema; las componentes no nulas a_{ij}, $i,j = 1,2,...,n$ de la matriz A; las componentes b_i, $i = 1,2,...,n$ del vector de términos independientes b; las componentes $x0_i$, $i = 1,2,...,n$ de una aproximación inicial $X0 = X^{(0)}$; un valor del parámetro **w**; una tolerancia Tol, y un número máximo de iteraciones N.

Salida: Una solución aproximada $\tilde{X} = (x_1,..., x_n)^T$ o un mensaje.

Paso 1: Tomar $k = 1$.

Paso 2: Mientras $k \le N$, seguir los pasos 3-8:

Paso 3: Tomar $x_1 = (1-w)x0_1 + w\left(\dfrac{b_1 - \sum_{j=2}^{n} a_{1j}x0_j}{a_{11}}\right)$

Paso 4: Para $i = 2,..., n-1$, tomar

$$x_i = (1-w)x0_i + w\left(\frac{b_i - \sum_{j=1}^{i-1} a_{ij}x_j - \sum_{j=i+1}^{n} a_{ij}x0_j}{a_{ii}}\right)$$

Paso 5: Tomar $x_n = (1-w)x0_n + w\left(\dfrac{b_n - \sum_{j=1}^{n-1} a_{nj}x_j}{a_{nn}}\right)$

Paso 6: Si $\| X - X0 \| < Tol$, entonces **salida:** "Una solución aproximada es $\tilde{X} = (x_1, x_2,..., x_n)^T$". **Terminar**.

Paso 7: Tomar $k = k + 1$.

Paso 8: Para $i = 1,2,...,n$, tomar $x0_i = x_i$.

Paso 9: Salida: "Se alcanzó el número máximo de iteraciones N pero no la tolerancia". **Terminar**.

Ejemplo 3.12 Consideremos el siguiente sistema de ecuaciones lineales

$$\begin{cases} 4x_1 + 3x_2 = 1 \\ 3x_1 + 4x_2 - x_3 = 1 \\ -x_2 + 4x_3 = 1 \end{cases}$$

Como la matriz de coeficientes de este sistema es simétrica y definida positiva (**verifíquelo!**), entonces el método **SOR** converge a la única solución del sistema dado, cualquiera sea la aproximación inicial y cualquiera sea w con $0 < w < 2$, según el teorema 3.10.

Si usamos $X^{(0)} = (0,0,0)^T$ y criterio de aproximación $\left\| X^{(k)} - X^{(k-1)} \right\|_\infty < .001$, se obtienen los siguientes resultados para distintos valores de w, siendo las fórmulas escalares de iteración del método SOR, en este caso, las siguientes:

$$\begin{cases} x_1^{(k)} = (1-w)x_1^{(k-1)} + w\left(\dfrac{1 - 3x_2^{(k-1)}}{4}\right) \\ x_2^{(k)} = (1-w)x_2^{(k-1)} + w\left(\dfrac{1 - 3x_1^{(k)} + x_3^{(k-1)}}{4}\right), \quad k = 1,2,\ldots, \quad 0 < w < 2 \\ x_3^{(k)} = (1-w)x_3^{(k-1)} + w\left(\dfrac{1 + x_2^{(k)}}{4}\right) \end{cases}$$

Como el método de Gauss-Seidel converge, en este caso, podemos pensar en utilizar el método SOR para acelerar la convergencia del método de Gauss-Seidel; así que tomaremos valores de w con $1 \le w < 2$.

Para $w = 1.0$ (Gauss-Seidel):

$$X^{(1)} = (.25000,\ 6.2500 \times 10^{-2},\ .26563)^T$$
$$\vdots$$
$$X^{(13)} = (1.1546 \times 10^{-3},\ .33237,\ .33309)^T \approx X$$

Para $w = 1.2$:

$$X^{(1)} = (.30000,\ 3.0000 \times 10^{-2},\ .30900)^T$$
$$\vdots$$
$$X^{(9)} = (3.5372 \times 10^{-4},\ .33310,\ .33329)^T \approx X$$

Para $w = 1.25$:

$$X^{(1)} = (.31250,\ 1.9531 \times 10^{-2},\ .31860)^T$$
$$\vdots$$
$$X^{(8)} = (6.4330 \times 10^{-5},\ .33331,\ .33333)^T \approx X$$

Para $w = 1.3$:

$$X^{(1)} = \left(.32500,\ 8.1250\times 10^{-3},\ .32764\right)^T$$
$$\vdots$$
$$X^{(6)} = \left(-1.2411\times 10^{-3},\ .33427,\ .33335\right)^T \approx X$$

Para $w = 1.4$:

$$X^{(1)} = \left(.35000,\ -1.7500\times 10^{-2},\ .34388\right)^T$$
$$\vdots$$
$$X^{(8)} = \left(9.7704\times 10^{-4},\ .33286,\ .33315\right)^T \approx X$$

Observando los resultados anteriores, se puede concluir que, para este ejemplo, el valor óptimo del parámetro w debe estar cerca de 1.3, y para este valor de w la convergencia del método SOR es más rápida. ♦

Instrucciones en DERIVE:

BW(A,w): **S**implifica en la matriz de iteración, B_w, del método SOR.

SOR(A,b,w, $X^{(0)}$,N): apro**X**ima las primeras N iteraciones en el método SOR aplicado al sistema $AX = b$, para el valor dado de w y tomando aproximación inicial $X^{(0)}$. Para el ejemplo, aproXime la expresión SOR([[4,3,0],[3,4,−1],[0,−1,4]],[1,1,1],1.3,[0,0,0],6). à

Ejemplo 3.13 Si aplicamos el método **SOR** para resolver el sistema de ecuaciones lineales

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

con $X^{(0)} = (0,0,0,0)^T$ y criterio de aproximación $\left\| X^{(k)} - X^{(k-1)} \right\|_\infty < .001$, se obtienen los siguientes resultados:

Para $w = 1.5$:

$$X^{(1)} = \left(.90000,\ 3.5318,\ -1.3902,\ -4.3098\right)^T$$
$$X^{(2)} = \left(1.3968,\ 3.4072,\ -.86286,\ -1.9859\right)^T$$
$$\vdots$$
$$X^{(13)} = \left(1.1041,\ 2.9965,\ -1.0210,\ -2.6260\right)^T \approx X$$

Para $w = 1.8$:

$$X^{(1)} = (1.0800,\ 4.2676,\ -1.6006,\ -5.7158)^T$$
$$X^{(2)} = (1.5604,\ 3.4762,\ -.63554,\ -.39176)^T$$
$$\vdots$$
$$X^{(42)} = (1.1040,\ 2.9967,\ -1.0207,\ -2.6262)^T \approx X$$

Analice la convergencia del método SOR en cada uno de los casos anteriores. ♦

3.9 SOLUCIÓN NUMÉRICA DE SISTEMAS NO-LINEALES

Consideremos un sistema no-lineal

$$\begin{cases} f_1(x_1, x_2, \ldots, x_n) = 0 \\ f_2(x_1, x_2, \ldots, x_n) = 0 \\ \qquad\qquad \vdots \\ f_n(x_1, x_2, \ldots, x_n) = 0 \end{cases}$$

donde para cada $i = 1,2,\ldots,n$,

$$f_i: \quad \mathbf{D}_i \quad \rightarrow \quad \mathbf{R},\ \mathbf{D}_i \subseteq \mathbf{R}^n$$
$$X = (x_1, \ldots, x_n) \ \rightarrow \ f_i(X) = y$$

y alguna f_i es no-lineal.

Si hacemos $F \equiv \begin{pmatrix} f_1 \\ f_2 \\ \vdots \\ f_n \end{pmatrix}$ y $X = (x_1, x_2, \ldots, x_n)$, el sistema anterior puede ser escrito en la forma vectorial $F(X) = 0$ con $F: \mathbf{D} \rightarrow \mathbf{R}^n,\ \mathbf{D} \subseteq \mathbf{R}^n$.

El problema de hallar las soluciones reales de un sistema no-lineal es mucho más difícil que el problema de hallar las raíces reales de una sola ecuación no-lineal (en una variable), y mucho más que el problema de resolver un sistema lineal de ecuaciones.

Aunque estudiaremos ciertos métodos que convergen a una solución, **no** existe un criterio general para saber cuántas soluciones tiene un sistema no-lineal dado, e incluso es posible que el sistema no tenga solución, como ocurre con el siguiente sistema

$$\begin{cases} xy = 1 \\ y = 0 \end{cases}$$

Un posible método (directo) para resolver sistemas no-lineales de ecuaciones puede ser el de reducción de variables.

Por ejemplo, para el sistema

$$\begin{cases} E_1: x^2 + y^2 = 1 \\ E_2: \quad xy = 0 \end{cases}$$

es claro que sus soluciones son $(1,0)$, $(0,1)$, $(-1,0)$ y $(0,-1)$; que son los puntos de intersección, en el plano xy, de la circunferencia unitaria $x^2 + y^2 = 1$ con los ejes coordenados $x = 0$ y $y = 0$. Ver la FIGURA 3.1.

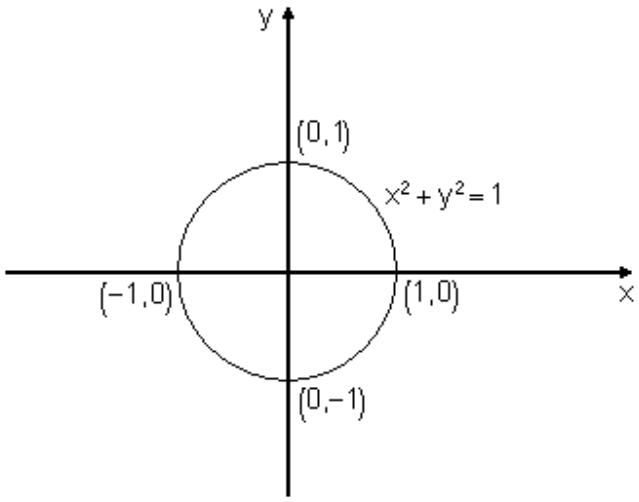

FIGURA 3.1

Una manera de resolver el sistema, en este caso, es la siguiente:

Despejando x de la ecuación E_1, obtenemos E_1': $x = \pm\sqrt{1-y^2}$, y sustituyendo en la ecuación E_2, obtenemos E_2': $\left(\pm\sqrt{1-y^2}\right)y = 0$. Si resolvemos la ecuación E_2' para y, obtenemos $y = 0$, $y = 1$ o $y = -1$. Sustituyendo en la ecuación E_1', los valores de y obtenidos, tenemos que: si $y = 0$, entonces $x = 1$ o $x = -1$, lo que produce las soluciones $(1,0)$ y $(-1,0)$; y si $y = 1$ o $y = -1$, entonces $x = 0$, lo que produce las soluciones $(0,1)$ y $(0,-1)$. ¨

En general, este método **no** es fácil de aplicar, pues la eliminación de variables puede ser muy difícil ó incluso imposible de realizar.

Estudiaremos los métodos iterativos de Punto Fijo y Newton-Raphson para aproximar soluciones reales de sistemas no-lineales.

3.9.1 Iteración de Punto Fijo para sistemas no-lineales: Dado un sistema no-lineal de n-ecuaciones con n-incógnitas, $F(X) = 0$ ($F: \mathbf{D} \to \mathbf{R}^n$, $\mathbf{D} \subseteq \mathbf{R}^n$), el método de Punto Fijo para resolver este sistema consiste en transformar dicho sistema en otro equivalente (por lo menos en forma local) del tipo $X = G(X)$ para alguna función G: $\mathbf{D'} \to \mathbf{R}^n$, $\mathbf{D'} \subseteq \mathbf{D}$.

Por ejemplo, si tenemos un sistema de dos ecuaciones con dos incógnitas

$$\begin{cases} f_1(x,y)=0 \\ f_2(x,y)=0 \end{cases} \qquad \left(F(X)=0 \quad \text{con} \quad F \equiv \begin{pmatrix} f_1 \\ f_2 \end{pmatrix} \right)$$

lo escribimos en la forma equivalente

$$\begin{cases} x=g_1(x,y) \\ y=g_2(x,y) \end{cases} \qquad \left(X=G(X) \quad \text{con} \quad G \equiv \begin{pmatrix} g_1 \\ g_2 \end{pmatrix} \right)$$

(despejando, por ejemplo, si es posible, x de $f_1(x,y)=0$ e y de $f_2(x,y)=0$).

La iteración vectorial de punto fijo correspondiente

$$X^{(k)} = G\left(X^{(k-1)}\right), \quad k=1,2,...$$

se convierte en

$$\begin{cases} x^{(k)} = g_1\left(x^{(k-1)}, y^{(k-1)}\right), \\ \\ y^{(k)} = g_2\left(x^{(k-1)}, y^{(k-1)}\right), \end{cases} \qquad k=1,2,....$$

que no es otra cosa que la **iteración de Jacobi** para sistemas no-lineales.

Otra forma de iteración vectorial de punto fijo es

$$\begin{cases} x^{(k)} = g_1\left(x^{(k-1)}, y^{(k-1)}\right), \\ \\ y^{(k)} = g_2\left(x^{(k)}, y^{(k-1)}\right), \end{cases} \qquad k=1,2,....$$

que usa el valor ya calculado $x^{(k)}$, y que no es otra cosa que el **método de Gauss-Seidel** para sistemas no-lineales.

El siguiente teorema general, cuya demostración puede ser consultada en Ortega, 1990, página 153, da condiciones suficientes (no necesarias) para la convergencia de la sucesión $\left\{X^{(k)}\right\}_k$ generada por la fórmula de iteración de Punto Fijo

$$X^{(k)} = G\left(X^{(k-1)}\right), \quad k=1,2,...$$

Teorema 3.11 Sea $\mathbf{D} = \left\{ (x_1, x_2, ..., x_n) \in \mathbf{R}^n \,/\, a_i \le x_i \le b_i,\ i=1,2,...,n \right\}$ para alguna colección de constantes reales $a_i, b_i,\ i=1,2,...,n$. Supongamos que $G: \mathbf{R}^n \to \mathbf{R}^n$ con $G \equiv \begin{pmatrix} g_1 \\ \vdots \\ g_n \end{pmatrix}$, es tal que $G(\mathbf{D}) \subseteq \mathbf{D}$, es decir, para todo $X \in \mathbf{D}$, $G(X) \in \mathbf{D}$; y que para cada $j=1,2,...,n$, g_j y sus primeras

derivadas parciales $\frac{\partial g_j(X)}{\partial x_i}$, $i = 1,2,...,n$, son continuas en $\mathbf{D}$. Entonces G tiene un punto fijo en $\mathbf{D}$.

Si, además, existe una constante M, con $0 \le M < 1$, tal que para cada $i=1,2,...,n$

$$\left| \frac{\partial g_j(X)}{\partial x_i} \right| \le \frac{M}{n}, \text{ siempre que } X \in \mathbf{D},$$

entonces G tiene un único punto fijo $P \in \mathbf{D}$ y la sucesión $\left\{X^{(k)}\right\}_k$ definida por la iteración

$$X^{(k)} = G\left(X^{(k-1)}\right), \quad k=1,2,...$$

converge al punto fijo P, cualquiera sea la escogencia de $X^{(0)} \in \mathbf{D}$, y se tiene la siguiente cota de error

$$\left\| X^{(k)} - P \right\|_\infty \le \frac{M^k}{1-M} \left\| X^{(1)} - X^{(0)} \right\|_\infty \quad \tilde{N}$$

Ejemplo 3.14 Consideremos el siguiente sistema no-lineal

$$\begin{cases} x^2 - 10x + y^2 + 8 = 0 & \left(f_1(x,y) = x^2 - 10x + y^2 + 8\right) \\ xy^2 + x - 10y + 8 = 0 & \left(f_2(x,y) = xy^2 + x - 10y + 8\right) \end{cases}$$

Las gráficas de $x^2 - 10x + y^2 + 8 = 0$ y $xy^2 + x - 10y + 8 = 0$, en un mismo plano coordenado, se muestran en la FIGURA 3.2.

Observando la FIGURA 3.2 vemos que el sistema dado tiene únicamente dos soluciones reales. Para encontrar estas soluciones por el método de Punto Fijo, empezamos por transformar el sistema dado en otro equivalente de la forma $X = G(X)$. Uno de esos sistemas es

$$\begin{cases} x = \frac{x^2 + y^2 + 8}{10} & \left(g_1(x,y) = \frac{x^2 + y^2 + 8}{10}\right) \\ y = \frac{xy^2 + x + 8}{10} & \left(g_2(x,y) = \frac{xy^2 + x + 8}{10}\right) \end{cases}$$

(Para obtener el sistema equivalente se despejaron las variables x e y de las ecuaciones $f_1(x,y) = 0$ y $f_2(x,y) = 0$, respectivamente, de aquellas posiciones donde éllas eran dominantes, de acuerdo con la solución buscada).

Observe que $G \equiv \begin{pmatrix} g_1 \\ g_2 \end{pmatrix}$ es tal que: g_1 y g_2 son continuas en todo $\mathbf{R}^2$, porque son polinómicas.

$\frac{\partial g_1(X)}{\partial x} = \frac{2x}{10}$, $\frac{\partial g_2(X)}{\partial x} = \frac{y^2 + 1}{10}$ son continuas en todo $\mathbf{R}^2$.

$\frac{\partial g_1(X)}{\partial y} = \frac{2y}{10}$, $\frac{\partial g_2(X)}{\partial y} = \frac{2xy}{10}$ son continuas en todo $\mathbf{R}^2$.

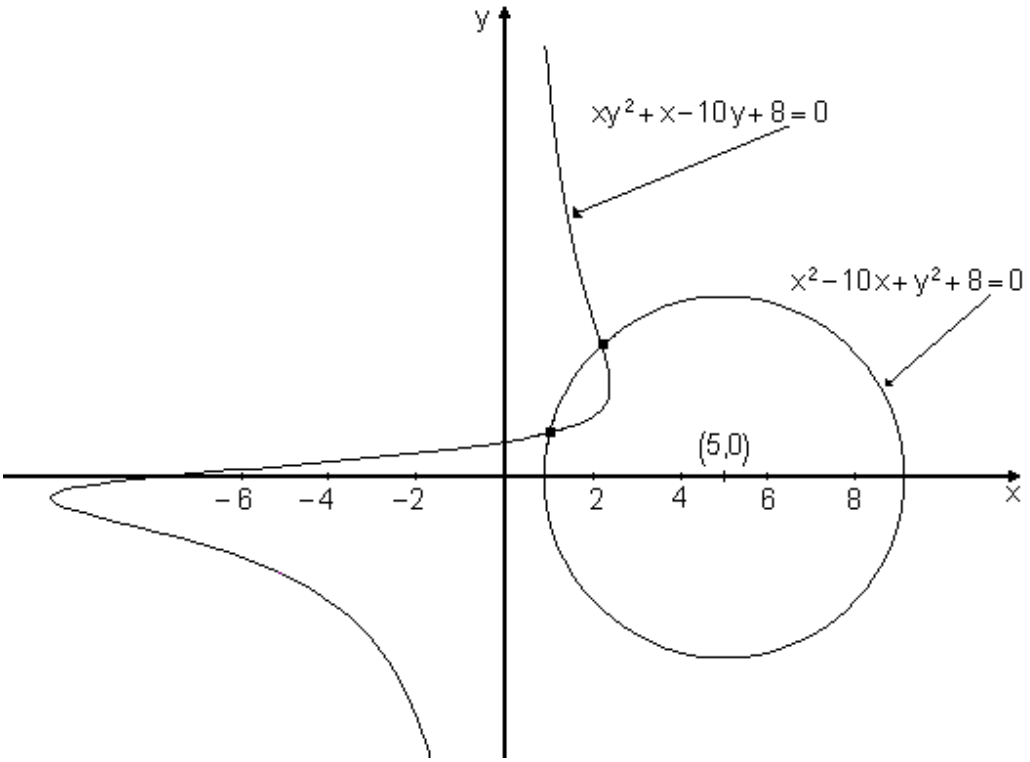

FIGURA 3.2

Ahora,

$$0 \le x,y \le 1.5 \Rightarrow 0 \le x^2, y^2 \le 2.25$$
$$\Rightarrow 0 \le x^2 + y^2 \le 4.5$$
$$\Rightarrow 0 \le \frac{x^2+y^2+8}{10} \le \frac{4.5+8}{10} = \frac{12.5}{10} = 1.25 \le 1.5$$

También,

$$0 \le x,y \le 1.5 \Rightarrow 0 \le xy^2 + x \le (1.5)(2.25) + 1.5 = 4.875$$
$$\Rightarrow 0 \le \frac{xy^2+x+8}{10} \le \frac{4.875+8}{10} = \frac{12.875}{10} = 1.2875 \le 1.5$$

Luego si $\mathbf{D} = \left\{(x,y) \in \mathbf{R}^2 \,/\, 0 \le x,y \le 1.5\right\}$, entonces $G(\mathbf{D}) \subseteq \mathbf{D}$, así que G satisface las hipótesis del teorema 3.11 en $\mathbf{D}$, y por lo tanto G tiene por lo menos un punto fijo en $\mathbf{D}$.

Ahora, si $X \in \mathbf{D}$,

$$\left| \frac{\partial g_1(X)}{\partial x} \right| = \left| \frac{x}{5} \right| \le \frac{1.5}{5} = .3$$
$$\left| \frac{\partial g_1(X)}{\partial y} \right| = \left| \frac{y}{5} \right| \le .3$$

$$\left|\frac{\partial g_2(X)}{\partial x}\right| = \left|\frac{y^2+1}{10}\right| \le \frac{3.25}{10} = .325$$
$$\left|\frac{\partial g_2(X)}{\partial y}\right| = \left|\frac{2xy}{10}\right| \le \frac{4.5}{10} = .45$$

Escogiendo M de modo que $\frac{M}{2} = .45$, es decir, escogiendo $M = .90$ tendremos que $0 \le M < 1$, y para cada $X \in \mathbf{D}$,

$$\left|\frac{\partial g_j(X)}{\partial x}\right| \le \frac{M}{2} \quad y \quad \left|\frac{\partial g_j(X)}{\partial y}\right| \le \frac{M}{2}, \quad j = 1,2$$

En consecuencia, G tiene un único punto fijo $P \in \mathbf{D}$ y cualquiera sea $X^{(0)} \in \mathbf{D}$, la sucesión $\left\{X^{(k)}\right\}_k$, con $X^{(k)} = G\left(X^{(k-1)}\right)$, $k = 1,2,...$ converge a P.

Si al sistema dado le aplicamos la iteración funcional de Punto Fijo

$$x^{(k)} = \frac{\left(x^{(k-1)}\right)^2 + \left(y^{(k-1)}\right)^2 + 8}{10} \quad \left(x^{(k)} = g_1\left(x^{(k-1)}, y^{(k-1)}\right)\right)$$
$$y^{(k)} = \frac{x^{(k-1)}\left(y^{(k-1)}\right)^2 + x^{(k-1)} + 8}{10} \quad \left(y^{(k)} = g_2\left(x^{(k-1)}, y^{(k-1)}\right)\right)$$

empezando con $X^{(0)} = (.5,.5)$ y criterio de aproximación $\left\| X^{(k)} - X^{(k-1)} \right\|_\infty < .001$, obtenemos los resultados que se muestran en la TABLA 3.1 siguiente.

k	$x^{(k)}$	$y^{(k)}$	$\left\| X^{(k)} - X^{(k-1)} \right\|_\infty$
0	.5	.5	
1	.85	.8625	.3625
2	.946641	.948232	.096641
3	.979527	.979781	.032886
4	.991944	.991984	.012417
5	.996799	.996805	4.855×10^{-3}
6	.998723	.998724	1.924×10^{-3}
7	.999490	.999490	7.67×10^{-4}

TABLA 3.1

Luego $P \approx (.999490, .999490)$.

Instrucción en DERIVE:

FIXED_POINT($[g_1(x,y), g_2(x,y)], [x,y], [x_0, y_0], N$): apro**X**ima las primeras N iteraciones en el método de Punto Fijo aplicado al sistema no-lineal $\begin{cases} x = g_1(x,y) \\ y = g_2(x,y) \end{cases}$, tomando como aproximación inicial el punto (x_0, y_0). Para el ejemplo, apro**X**ime la expresión FIXED_POINT($\left[\frac{x^2+y^2+8}{10}, \frac{xy^2+x+8}{10}\right], [x,y], [0.5, 0.5], 7$). à

Usando la cota de error dada por el teorema 3.11 con $M = .90$, obtenemos

$$\left\| X^{(7)} - P \right\|_\infty \le \frac{(.9)^7}{1-.9}(.3625) \approx 1.73$$

la cual no indica la precisión real de $X^{(7)}$ con respecto a P, pues como el lector puede verificar fácilmente, $P = (1,1)$ y realmente $\left\| X^{(7)} - P \right\|_\infty \approx 5.1 \times 10^{-4}$.

Si usamos las fórmulas de iteración del método de Gauss-Seidel

$$x^{(k)} = \frac{\left(x^{(k-1)}\right)^2 + \left(y^{(k-1)}\right)^2 + 8}{10} \quad \left(x^{(k)} = g_1\left(x^{(k-1)}, y^{(k-1)}\right)\right)$$

$$y^{(k)} = \frac{x^{(k)}\left(y^{(k-1)}\right)^2 + x^{(k)} + 8}{10} \quad \left(y^{(k)} = g_2\left(x^{(k)}, y^{(k-1)}\right)\right)$$

con $X^{(0)} = (.5, .5)$ y criterio de aproximación $\left\| X^{(k)} - X^{(k-1)} \right\|_\infty < .001$, se obtienen los resultados de la TABLA 3.2 siguiente.

k	$x^{(k)}$	$y^{(k)}$	$\left\| x^{(k)} - x^{(k-1)} \right\|_\infty$
0	.5	.5	
1	.85	.90625	.40625
2	.954379	.973820	.104379
3	.985916	.992089	.031537
4	.995627	.997556	9.711×10^{-3}
5	.998639	.999240	3.012×10^{-3}
6	.999576	.999763	9.37×10^{-4}

TABLA 3.2

Observe que $\left\| X^{(6)} - P \right\|_\infty \approx 4.24 \times 10^{-4}$, lo que asegura una precisión en la aproximación de P de tres cifras decimales exactas. Como **ejercicio** haga un análisis similar para encontrar una aproximación de la otra solución del sistema dado. ♦

3.9.2 Método de Newton-Raphson para sistemas no-lineales: Consideremos un sistema no-lineal de dos ecuaciones con dos incógnitas

$$\begin{cases} f_1(x,y)=0 \\ f_2(x,y)=0 \end{cases} \qquad \left(F(x,y)=0 \quad \text{con} \quad F \equiv \begin{pmatrix} f_1 \\ f_2 \end{pmatrix} \right)$$

y sea $\alpha=(\alpha_1,\alpha_2)$ una raíz de $F(X)=0$ con $X=(x,y)$.

Supongamos que conocemos una aproximación $X^{(k)}=\left(x^{(k)},y^{(k)}\right)$ de α. Para generar la aproximación siguiente $X^{(k+1)}=\left(x^{(k+1)},y^{(k+1)}\right)$, mediante el método de Newton-Raphson, procedemos como se indica a continuación:

Si las funciones $f_1(x,y)$ y $f_2(x,y)$ y todas sus derivadas parciales de orden menor o igual que dos son continuas en una vecindad de $X^{(k)}$, entonces para $\left(x^{(k+1)},y^{(k+1)}\right)$ en esa vecindad se tiene

$$\begin{aligned} f_1\left(x^{(k+1)},y^{(k+1)}\right) = f_1\left(x^{(k)},y^{(k)}\right) &+ \left(x^{(k+1)}-x^{(k)}\right)\frac{\partial f_1}{\partial x}\left(x^{(k)},y^{(k)}\right) + \left(y^{(k+1)}-y^{(k)}\right)\frac{\partial f_1}{\partial y}\left(x^{(k)},y^{(k)}\right) \\ &+ \frac{1}{2}\left[\left(x^{(k+1)}-x^{(k)}\right)^2\frac{\partial^2 f_1}{\partial x^2}(\xi,\eta) + 2\left(x^{(k+1)}-x^{(k)}\right)\left(y^{(k+1)}-y^{(k)}\right)\frac{\partial^2 f_1}{\partial x\,\partial y}(\xi,\eta)\right. \\ &\left. + \left(y^{(k+1)}-y^{(k)}\right)^2\frac{\partial^2 f_1}{\partial y^2}(\xi,\eta)\right] \end{aligned}$$

con ξ entre $x^{(k)}$ y $x^{(k+1)}$, y η entre $y^{(k)}$ y $y^{(k+1)}$.

Si suponemos que $f_1\left(x^{(k+1)},y^{(k+1)}\right)\approx 0$ y que el residuo

$$\begin{aligned} R_2\left(f_1,x^{(k+1)},y^{(k+1)}\right) = \frac{1}{2}\left[\left(x^{(k+1)}-x^{(k)}\right)\frac{\partial^2 f_1}{\partial x^2}(\xi,\eta) + 2\left(x^{(k+1)}-x^{(k)}\right)\left(y^{(k+1)}-y^{(k)}\right)\frac{\partial^2 f_1}{\partial x\partial y}\right. \\ \left. + \left(y^{(k+1)}-y^{(k)}\right)^2\frac{\partial^2 f_1}{\partial y^2}(\xi,\eta)\right] \approx 0 \end{aligned}$$

obtenemos

$$0 \approx f_1\left(x^{(k)},y^{(k)}\right) + \left(x^{(k+1)}-x^{(k)}\right)\frac{\partial f_1}{\partial x}\left(x^{(k)},y^{(k)}\right) + \left(y^{(k+1)}-y^{(k)}\right)\frac{\partial f_1}{\partial y}\left(x^{(k)},y^{(k)}\right) \qquad \textbf{(3.20)}$$

Al trabajar, de manera similar, con $f_2(x,y)$ obtenemos

$$0 \approx f_2\left(x^{(k)},y^{(k)}\right) + \left(x^{(k+1)}-x^{(k)}\right)\frac{\partial f_2}{\partial x}\left(x^{(k)},y^{(k)}\right) + \left(y^{(k+1)}-y^{(k)}\right)\frac{\partial f_2}{\partial y}\left(x^{(k)},y^{(k)}\right) \qquad \textbf{(3.21)}$$

Si las cuasi-igualdades (3.20) y (3.21) las manejamos como igualdades y las escribimos en forma matricial, obtenemos el siguiente **sistema de dos ecuaciones lineales en las dos incógnitas** $x^{(k+1)}$, $y^{(k+1)}$:

$$\underbrace{\begin{pmatrix} \frac{\partial f_1}{\partial x}\left(x^{(k)},y^{(k)}\right) & \frac{\partial f_1}{\partial y}\left(x^{(k)},y^{(k)}\right) \\ \frac{\partial f_2}{\partial x}\left(x^{(k)},y^{(k)}\right) & \frac{\partial f_2}{\partial y}\left(x^{(k)},y^{(k)}\right) \end{pmatrix}}_{J\left(X^{(k)}\right)} \underbrace{\begin{pmatrix} x^{(k+1)}-x^{(k)} \\ y^{(k+1)}-y^{(k)} \end{pmatrix}}_{X^{(k+1)}-X^{(k)}} = -\underbrace{\begin{pmatrix} f_1\left(x^{(k)},y^{(k)}\right) \\ f_2\left(x^{(k)},y^{(k)}\right) \end{pmatrix}}_{F\left(X^{(k)}\right)}$$

La matriz $J\left(X^{(k)}\right)$, en la ecuación anterior, se denomina **matriz Jacobiana del sistema**.

Despejando $\begin{pmatrix} x^{(k+1)} \\ y^{(k+1)} \end{pmatrix}$, siempre que $\left[\, J\left(X^{(k)}\right)\right]^{-1}$ exista, obtenemos

$$\begin{pmatrix} x^{(k+1)} \\ y^{(k+1)} \end{pmatrix} = \begin{pmatrix} x^{(k)} \\ y^{(k)} \end{pmatrix} - \left[J\left(X^{(k)}\right)\right]^{-1} F\left(X^{(k)}\right)$$

o sea

$$\boxed{X^{(k+1)} = X^{(k)} - \left[J\left(X^{(k)}\right)\right]^{-1} F\left(X^{(k)}\right),\ \ k = 0,1,...}$$

la cual es la **fórmula vectorial de iteración del método de Newton-Raphson** para el sistema no-lineal $F(X) = 0$ (compare esta fórmula con la fórmula de iteracion del método de Newton-Raphson obtenida en el capítulo 2).

Para implementar este método **no** es necesario, ni conveniente calcular $\left[\, J\left(X^{(k)}\right)\right]^{-1}$; se recomienda el siguiente **algoritmo**:

Dada una aproximación inicial $X^{(0)}$, una tolerancia $\varepsilon > 0$, y un número máximo de iteraciones **N**; para $k = 0,1,...,N$, hacemos:

Paso 1: Definimos $Z^{(k+1)} = X^{(k+1)} - X^{(k)}$

Paso 2: Resolvemos, para $Z^{(k+1)}$, el sistema lineal

$$J\left(X^{(k)}\right) Z^{(k+1)} = -F\left(X^{(k)}\right)$$

Paso 3: calculamos $X^{(k+1)} = Z^{(k+1)} + X^{(k)}$.

Paso 4: Si $\left\| Z^{(k+1)} \right\| = \left\| X^{(k+1)} - X^{(k)} \right\| < \varepsilon$ para alguna norma vectorial $\| \,.\, \|$, entonces $X^{(k+1)}$ es una aproximación de una solución del sistema $F(X) = 0$. De lo contrario se vuelve a iterar.

Ejemplo 3.15 Si usamos el método de Newton-Raphson para resolver el siguiente sistema de ecuaciones no-lineales

$$\begin{cases} x^2 - 10x + y^2 + 8 = 0 \\ xy^2 + x - 10y + 8 = 0 \end{cases}$$

que es el mismo del ejemplo 3.14, obtenemos los resultados que aparecen en las TABLAS 3.3 y 3.4, donde se usó como criterio de aproximación $\left\| X^{(k)} - X^{(k-1)} \right\|_\infty < .001$.

k	$x^{(k)}$	$y^{(k)}$	$\left\| X^{(k)} - X^{(k-1)} \right\|_\infty$
0	.5	.5	
1	.937685	.939169	.439169
2	.998694	.998388	.061009
3	.999999	.999999	1.611×10^{-3}
4	1.00000	1.00000	1.0×10^{-6}

TABLA 3.3

Instrucción en DERIVE:

NEWTONS($\left[f_1(x,y), f_2(x,y)\right], [x,y], [x_0, y_0], N$): apro**X**ima las primeras N iteraciones del método de Newton-Raphson aplicado al sistema no-lineal $\begin{cases} f_1(x,y) = 0 \\ f_2(x,y) = 0 \end{cases}$, tomando como aproximación inicial el punto (x_0, y_0). Para el ejemplo, apro**X**ime la expresión NEWTONS($\left[x^2 - 10x + y^2 + 8, xy^2 + x - 10y + 8\right], [x,y], [0.5, 0.5], 4$). à

De acuerdo con los resultados de la TABLA 3.3, se tiene que $X^{(4)} = (1.00000, 1.00000) \approx X = (1,1)$.

k	$x^{(k)}$	$y^{(k)}$	$\left\| X^{(k)} - X^{(k-1)} \right\|_\infty$
0	2.0	3.0	
1	2.19444	3.02778	.19444
2	2.19345	3.02048	7.3×10^{-3}
3	2.19344	3.02047	2.0×10^{-5}

TABLA 3.4

De acuerdo con los resultados que aparecen en la TABLA 3.4, se tiene que $X^{(3)} = (2.19344, 3.02047)$ es una aproximación de la otra solución del sistema dado.

Veamos como se calcula la primera iteración $X^{(1)} = (2.19444, 3.02778)$, que aparece en la TABLA 3.4, usando el método de Newton-Raphson:

Primero que todo, sean $f_1(x,y) = x^2 - 10x + y^2 + 8$, $f_2(x,y) = xy^2 + x - 10y + 8$. Entonces

$$\frac{\partial f_1}{\partial x}(x,y) = 2x - 10, \quad \frac{\partial f_1}{\partial y}(x,y) = 2y$$

$$\frac{\partial f_2}{\partial x}(x,y) = y^2 + 1, \quad \frac{\partial f_2}{\partial y}(x,y) = 2xy - 10$$

Es claro que las funciones $f_1(x,y)$, $f_2(x,y)$ y sus derivadas parciales de orden menor o igual que dos son continuas en todo $\mathbf{R}^2$, por ser polinómicas.

Para calcular la primera iteración $X^{(1)} = \begin{pmatrix} x^{(1)} \\ y^{(1)} \end{pmatrix}$, siguiendo los pasos en el algoritmo anterior, procedemos así:

Definimos $Z^{(1)} = X^{(1)} - X^{(0)}$ ($k = 0$), y resolvemos el sistema lineal

$$J\left(X^{(0)}\right) Z^{(1)} = -F\left(X^{(0)}\right)$$

En este sistema

$$F\left(X^{(0)}\right) = \begin{pmatrix} f_1\left(X^{(0)}\right) \\ f_2\left(X^{(0)}\right) \end{pmatrix}$$

siendo

$$f_1\left(X^{(0)}\right) = f_1(2.0,\ 3.0) = 1.0 \ , \ f_2\left(X^{(0)}\right) = f_2(2.0,\ 3.0) = -2$$

y

$$J\left(X^{(0)}\right) = \begin{pmatrix} \frac{\partial f_1}{\partial x}(2.0,3.0) & \frac{\partial f_1}{\partial y}(2.0,3.0) \\ \frac{\partial f_2}{\partial x}(2.0,3.0) & \frac{\partial f_2}{\partial y}(2.0,3.0) \end{pmatrix} = \begin{pmatrix} -6 & 6 \\ 10 & 2 \end{pmatrix}$$

Luego el sistema a resolver es

$$\begin{pmatrix} -6 & 6 \\ 10 & 2 \end{pmatrix} \begin{pmatrix} z_1^{(1)} \\ z_2^{(1)} \end{pmatrix} = \begin{pmatrix} -1 \\ 2 \end{pmatrix}$$

La solución única de este sistema es

$$\begin{pmatrix} z_1^{(1)} \\ z_2^{(1)} \end{pmatrix} = \frac{1}{-72} \begin{pmatrix} 2 & -6 \\ -10 & -6 \end{pmatrix} \begin{pmatrix} -1 \\ 2 \end{pmatrix} = \frac{1}{36} \begin{pmatrix} 7 \\ 1 \end{pmatrix}$$

Entonces

$$X^{(1)} = Z^{(1)} + X^{(0)} = \frac{1}{36} \begin{pmatrix} 7 \\ 1 \end{pmatrix} + \begin{pmatrix} 2 \\ 3 \end{pmatrix} = \frac{1}{36} \begin{pmatrix} 79 \\ 109 \end{pmatrix} \approx \begin{pmatrix} 2.19444 \\ 3.02778 \end{pmatrix}$$

Procediendo de manera similar se obtienen $X^{(2)}$ y $X^{(3)}$. ♦

Ejemplo 3.15 Consideremos el sistema no-lineal

$$\begin{cases} x^2 + y^2 - x = 0 \\ x^2 - y^2 - y = 0 \end{cases}$$

Las gráficas de las ecuaciones $x^2 + y^2 - x = 0$ y $x^2 - y^2 - y = 0$ ($f_1(x,y) = x^2 + y^2 - x$, $f_2(x,y) = x^2 - y^2 - y$), se muestran en la FIGURA 3.3 siguiente.

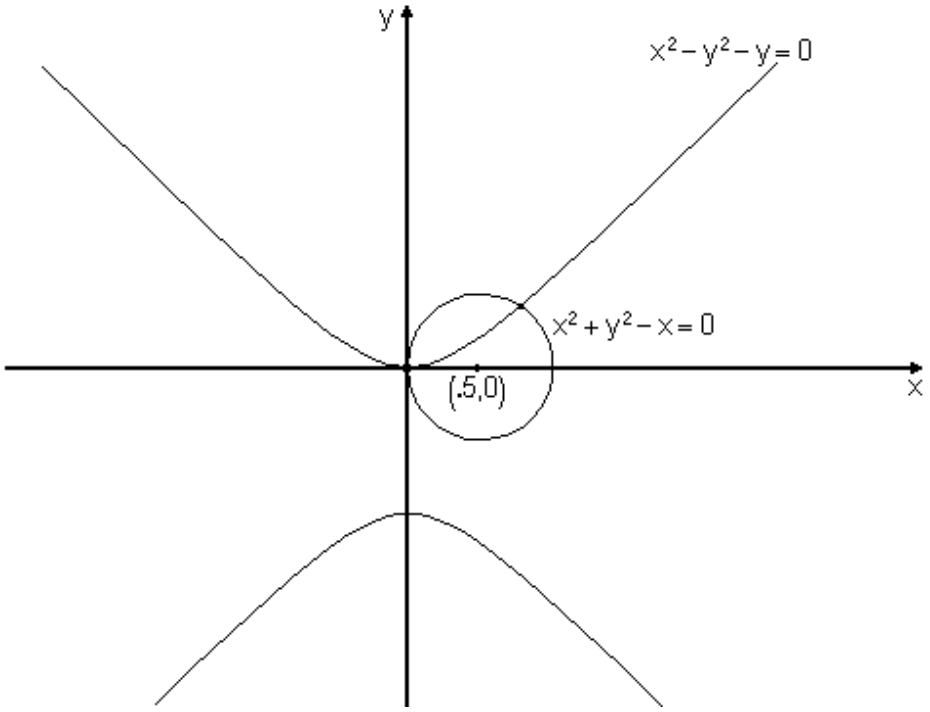

FIGURA 3.3

Por simple inspección sobre el sistema dado, u observando la FIGURA 3.3, se ve que $(0,0)$ es una solución del sistema, y que el sistema dado solamente tiene dos soluciones reales.

Si usamos el método de Newton-Raphson para encontrar la solución no nula del sistema dado, con criterio de aproximación $\left\| X^{(k)} - X^{(k-1)} \right\|_\infty < 10^{-3}$, se obtiene como solución aproximada $X^{(4)} = (.771845, .419643)$ tomando como aproximación inicial $X^{(0)} = (1.0,.5)$; si tomamos aproximación inicial $X^{(0)} = (1.0, 1.0)$, se obtiene la solución aproximada $X^{(4)} = (.771845, .419644)$, y si tomamos aproximación inicial $X^{(0)} = (.5,.5)$, obtenemos la solución aproximada $X^{(5)} = (.771845, .419643)$. ♦

Como una aplicación del método de Newton-Raphson para sistemas no-lineales, tenemos el **método de Bairstow** para encontrar **ceros complejos** de funciones polinómicas, método al cual nos referiremos a continuación.

3.10 CEROS COMPLEJOS DE POLINOMIOS: MÉTODO DE BAIRSTOW

Para hallar una raíz compleja de una ecuación polinómica con coeficientes reales puede utilizarse el método de Newton-Raphson con una aproximación inicial compleja y aritmética compleja. Otra forma de enfocar el problema de las raíces complejas de una ecuación polinómica con coeficientes reales se basa en el hecho de que si $z = a + bi$ es un cero complejo de multiplicidad **m** de un polinomio $p(x)$, entonces su conjugado $\bar{z} = a - bi$ es también un cero de multiplicidad **m** de $p(x)$ y $\left(x^2 - 2ax + a^2 + b^2\right)^m$ es un factor de $p(x)$. Es decir, las raíces complejas de ecuaciones polinómicas con coeficientes **reales** se producen en pares conjugados y por ésto, se debe buscar un factor cuadrático, más que lineal, del polinomio. Esta es la base del **método de Bairstow**, el cual nos permitirá encontrar raíces reales o complejas de una ecuación polinómica con coeficientes reales, realizando únicamente aritmética real. A continuación describimos este método:

Dado un polinomio con coeficientes reales

$$p(x) = a_0 + a_1 x + a_2 x^2 + \ldots + a_n x^n, \quad a_n \neq 0, \quad n \geq 2 \tag{3.22}$$

El algoritmo de la división de Euclides nos permite expresarlo en la forma

$$p(x) = \left(x^2 - ux - v\right)q(x) + r(x - u) + s \tag{3.23}$$

donde **u** y **v** son **constantes reales**, $q(x)$ es un polinomio de grado $n-2$ con **coeficientes reales**,

$$q(x) = b_0 + b_1 x + \ldots + b_{n-2} x^{n-2} \tag{3.24}$$

y $r(x-u)+s$ es un residuo lineal con **coeficientes reales r** y **s**.

Las expresiones $x^2 - ux - v$ y $r(x-u)+s$ han sido escritas así para simplificar los cálculos posteriores.

El objetivo es encontrar **u** y **v** tales que $x^2 - ux - v$ sea un factor cuadrático de $p(x)$.

Al sustituir (3.24) en (3.23), se obtiene

$$\begin{aligned} p(x) &= \left(x^2 - ux - v\right)q(x) + r(x-u) + s \\ &= \left(x^2 - ux - v\right)\left(b_0 + b_1 x + b_2 x^2 + \ldots + b_{n-3} x^{n-3} + b_{n-2} x^{n-2}\right) + r(x-u) + s \\ &= \left(-vb_0 - ru + s\right) + \left(-ub_0 - vb_1 + r\right)x + \left(b_0 - ub_1 - vb_2\right)x^2 + \ldots \\ &\quad + \left(b_{n-4} - ub_{n-3} - vb_{n-2}\right)x^{n-2} + \left(b_{n-3} - ub_{n-2}\right)x^{n-1} + b_{n-2}x^n \end{aligned} \tag{3.25}$$

Igualando (3.22) y (3.25), obtenemos

$$\begin{cases} a_n = b_{n-2} \\ a_{n-1} = b_{n-3} - ub_{n-2} \\ a_{n-2} = b_{n-4} - ub_{n-3} - vb_{n-2} \\ a_{n-3} = b_{n-5} - ub_{n-4} - vb_{n-3} \\ \quad \vdots \\ a_3 = b_1 - ub_2 - vb_3 \\ a_2 = b_0 - ub_1 - vb_2 \\ \\ a_1 = r - ub_0 - vb_1 \\ a_0 = s - ur - vb_0 \end{cases}$$

igualdades que pueden escribirse en la forma

$$\begin{cases} b_{n-2} = a_n \\ b_{n-3} = a_{n-1} + ub_{n-2} \\ b_{n-4} = a_{n-2} + ub_{n-3} + vb_{n-2} \\ b_{n-5} = a_{n-3} + ub_{n-4} + vb_{n-3} \\ \quad \vdots \\ b_1 = a_3 + ub_2 + vb_3 \\ b_0 = a_2 + ub_1 + vb_2 \\ \\ r = a_1 + ub_0 + vb_1 \\ s = a_0 + ur + vb_0 \end{cases} \qquad \textbf{(3.26)}$$

Para que $x^2 - ux - v$ sea un factor de $p(x)$, como queremos, es necesario que $r = 0$ y $s = 0$, pero r y s son funciones no-lineales de u y v (observe que b_0 y b_1 son funciones de u y v).

Debemos entonces resolver el sistema no-lineal

$$\begin{cases} r(u,v) = 0 \qquad (f_1(u,v) = 0) \\ \\ s(u,v) = 0 \qquad (f_2(u,v) = 0) \end{cases} \qquad \textbf{(3.27)}$$

para las incógnitas u y v.

Usando el **método de Newton-Raphson** para sistemas no-lineales, si

$$J = \begin{pmatrix} r_u & r_v \\ s_u & s_v \end{pmatrix}$$

donde r_u, r_v, s_u, s_v son las derivadas parciales de r y s con respecto a u y v, respectivamente, entonces en la **k**-ésima iteración (para determinar u y v), tenemos

$$\underbrace{\begin{pmatrix} u^{(k+1)} \\ v^{(k+1)} \end{pmatrix}}_{X^{(k+1)}} = \underbrace{\begin{pmatrix} u^{(k)} \\ v^{(k)} \end{pmatrix}}_{X^{(k)}} - \left[J\left(X^{(k)}\right)\right]^{-1} \underbrace{\begin{pmatrix} r\left(X^{(k)}\right) \\ s\left(X^{(k)}\right) \end{pmatrix}}_{F\left(X^{(k)}\right)} \tag{3.28}$$

Aparentemente **no** se dispone de las derivadas parciales de r y s, por **no** conocerse una relación explícita para r y s; sin embargo, si usamos la relación de recurrencia para a_j, b_j, u, v, r y s dada en **(3.26)** y **derivamos parcialmente** con respecto a u, obtenemos

$\left(b_{n-2}\right)_u = 0$, pues $b_{n-2} = a_n$ con a_n constante, y

$$\begin{cases} \left(b_{n-3}\right)_u = b_{n-2} \\ \left(b_{n-4}\right)_u = b_{n-3} + u\left(b_{n-3}\right)_u \\ \left(b_{n-5}\right)_u = b_{n-4} + u\left(b_{n-4}\right)_u + v\left(b_{n-3}\right)_u \\ \quad \vdots \\ \left(b_0\right)_u = b_1 + u\left(b_1\right)_u + v\left(b_2\right)_u \\ \\ r_u = b_0 + u\left(b_0\right)_u + v\left(b_1\right)_u \\ s_u = r + ur_u + v\left(b_0\right)_u \end{cases} \tag{3.29}$$

Si definimos $c_k = \left(b_{k-2}\right)_u$, $k = n-1, n-2, \ldots, 2$, $c_1 = r_u$ y $c_0 = s_u$, las relaciones en **(3.29)** pueden escribirse como sigue

$$\begin{cases} c_{n-1} = b_{n-2} \\ c_{n-2} = b_{n-3} + u\,c_{n-1} \\ c_{n-3} = b_{n-4} + u\,c_{n-2} + v\,c_{n-1} \\ \quad \vdots \\ c_2 = b_1 + uc_3 + vc_4 \\ c_1 = b_o + uc_2 + vc_3 \\ c_0 = r + uc_1 + vc_2 \end{cases} \tag{3.30}$$

de modo que

$$J\left(X^{(k)}\right) = \begin{pmatrix} r_u\left(X^{(k)}\right) & r_v\left(X^{(k)}\right) \\ s_u\left(X^{(k)}\right) & s_v\left(X^{(k)}\right) \end{pmatrix}$$

se convierte en

$$J\left(X^{(k)}\right)=\begin{pmatrix} c_1^{(k)} & r_v\left(X^{(k)}\right) \\ c_0^{(k)} & s_v\left(X^{(k)}\right) \end{pmatrix}$$

donde $c_1^{(k)}$ y $c_0^{(k)}$ son los valores de c_1 y c_0, obtenidos en las ecuaciones (3.30) en la **iteración k-**ésima.

Para obtener expresiones para $r_v\left(X^{(k)}\right)$ y $s_v\left(X^{(k)}\right)$, derivamos parcialmente las mismas relaciones en (3.26) pero con respecto a v, con lo cual obtenemos

$$\begin{cases} \left(b_{n-4}\right)_v = b_{n-2} \\ \left(b_{n-5}\right)_v = u\left(b_{n-4}\right)_v + b_{n-3} = b_{n-3} + u\left(b_{n-4}\right)_v \\ \left(b_{n-6}\right)_v = u\left(b_{n-5}\right)_v + b_{n-4} + v\left(b_{n-4}\right)_v = b_{n-4} + u\left(b_{n-5}\right)_v + v\left(b_{n-4}\right)_v \\ \quad\vdots \\ \left(b_0\right)_v = u\left(b_1\right)_v + b_2 + v\left(b_2\right)_v = b_2 + u\left(b_1\right)_v + v\left(b_2\right)_v \\ \\ r_v = u\left(b_0\right)_v + b_1 + v\left(b_1\right)_v = b_1 + u\left(b_0\right)_v + v\left(b_1\right)_v \\ s_v = ur_v + b_0 + v\left(b_0\right)_v = b_0 + ur_v + v\left(b_0\right)_v \end{cases} \qquad \textbf{(3.31)}$$

Si definimos $d_k = \left(b_{k-3}\right)_v$, $k = n-1, n-2, \ldots, 3$, $d_2 = r_v$ y $d_1 = s_v$, entonces las ecuaciones (3.31) se convierten en

$$\begin{cases} d_{n-1} = b_{n-2} \\ d_{n-2} = b_{n-3} + ud_{n-1} \\ d_{n-3} = b_{n-4} + ud_{n-2} + vd_{n-1} \\ \quad\vdots \\ d_3 = b_2 + ud_4 + vd_5 \\ d_2 = b_1 + ud_3 + vd_4 \\ d_1 = b_0 + ud_2 + vd_3 \end{cases} \qquad \textbf{(3.32)}$$

de modo que

$$J\left(X^{(k)}\right)=\begin{pmatrix} c_1^{(k)} & d_2^{(k)} \\ c_0^{(k)} & d_1^{(k)} \end{pmatrix}$$

Si observamos las ecuaciones (3.30) y (3.32), vemos que ellas producen los mismos valores para $k = n-1, n-2, \ldots, 1$, es decir, $d_k = c_k$, $k = n-1, n-2, \ldots, 1$; por tanto (3.32) es redundante, y $J\left(X^{(k)}\right)$ puede calcularse como

$$J\left(X^{(k)}\right)=\begin{pmatrix} c_1^{(k)} & c_2^{(k)} \\ c_0^{(k)} & c_1^{(k)} \end{pmatrix}$$

Resumimos la discusión precedente en el siguiente algoritmo:

Algoritmo 3.7 (Método de Bairstow) Para encontrar un factor cuadrático x^2-ux-v de un polinomio con coeficientes reales

$$p(x)=a_0+a_1x+a_2x^2+...+a_nx^n,\ \ a_n\neq 0,\ \ n\geq 2$$

Entrada: El grado **n** del polinomio $p(x)$; los coeficientes $a_0,a_1,...,a_n$ del polinomio $p(x)$; unas aproximaciones iniciales u_0, v_0 de **u** y **v**, respectivamente; una tolerancia **Tol**, y un número máximo de iteraciones N.

Salida: Un factor cuadrático aproximado x^2-ux-v del polinomio $p(x)$ o un mensaje.

Paso 1: Hacer

$$\begin{aligned} b_n &= a_n \\ c_n &= 0 \\ c_{n-1} &= a_n \end{aligned}$$

(observe que se cambió la notación de subíndices)

Paso 2: Tomar $i=1$.

Paso 3: Mientras que $i\leq N$ seguir los pasos 4-10:

Paso 4: Hacer $b_{n-1}=a_{n-1}+u_0b_n$.

Paso 5: Para $k=n-2,n-3,...,0$, hacer

$$\begin{aligned} b_k &= a_k+u_0b_{k+1}+v_0b_{k+2} \\ c_k &= b_{k+1}+u_0c_{k+1}+v_0c_{k+2} \end{aligned}$$

(con este cambio de subíndices $b_1=r$ y $b_0=s$)

Paso 6: Hacer $J=c_0c_2-c_1^2$ (aquí J es igual a $-\det\begin{pmatrix} c_1 & c_2 \\ c_0 & c_1 \end{pmatrix}$).

Paso 7: Hacer

$$u_1=u_0+\frac{c_1b_1-c_2b_0}{J}$$

$$v_1=v_0+\frac{c_1b_0-c_0b_1}{J}$$

(Se está usando la regla de Cramer para resolver el sistema)

Paso 8: Si $|b_1| = |r| < Tol$ y $|b_0| = |s| < Tol$, entonces : **salida:** "Un factor cuadrático aproximado del polinomio dado, y los correspondientes valores de r y s son: $x^2 - u_1 x - v_1;\ r = b_1,\ s = b_0$ ". **Terminar**.

Paso 9: Tomar $i = i + 1$.

Paso 10: Hacer $u_0 = u_1$
$v_0 = v_1$

Paso 11: Salida: "Se alcanzó el número máximo de iteraciones N pero no la tolerancia". **Terminar**.

Ejemplo 3.16 Encontrar **todas** las raíces de la ecuación $x^4 - 4x^3 + 7x^2 - 5x - 2 = 0$, usando el método de Bairstow.

Solución: Con un programa diseñado siguiendo el algoritmo 3.7 anterior, se obtienen los siguientes resultados:

Para las aproximaciones iniciales $u_0 = 3$ y $v_0 = -4$, y una tolerancia $Tol = 10^{-10}$, se obtiene en la séptima iteración que $u = 2.2756822037$ y $v = -3.6273650847$ y los valores de r y s correspondientes son $r = -5.7853028 \times 10^{-16}$ y $s = -1.1423154 \times 10^{-15}$.

Las raíces del factor cuadrático aproximado $x^2 - ux - v$ obtenido, son complejas conjugadas con parte real 1.1378411018 y parte imaginaria 1.5273122509, y si usamos Deflación se obtiene el polinomio reducido de grado dos

$$q(x) = x^2 - 1.7243177963x - .5513644073$$

cuyas raíces son $x_1 = 2.0000000000$ y $x_2 = -.2756822037$. ♦

Instrucción en DERIVE:

BAIRSTOW($p(x), x, u_0, v_0, N$): apro**X**ima las primeras N iteraciones en el método de Bairstow aplicado al polinomio $p(x)$ para obtener valores aproximados u_N y v_N de los coeficientes u y v de un factor cuadrático $x^2 - ux - v$ del polinomio $p(x)$, tomando como aproximaciones iniciales u_0 y v_0. Para este ejemplo, apro**X**ime la expresión BAIRSTOW($x^4 - 4x^3 + 7x^2 - 5x - 2, x, 3, -4, 7$). à

Ejemplo 3.17 Encontrar **todas** las raíces de la ecuación

$$p(x) = x^7 - 28x^6 + 322x^5 - 1960x^4 + 6769x^3 - 13132x^2 + 13068x - 5040 = 0$$

usando el método de Bairstow.

Solución: Si aplicamos el método de Bairstow con Deflación para hallar todas las raíces de la ecuación polinómica dada, se obtienen los resultados que aparecen a continuación, usando como tolerancia $Tol = 10^{-10}$:

Para los valores iniciales $u_0 = 2$ y $v_0 = 3$, se obtiene en la octava iteración $u = 4.0000000000$, $v = -3.0000000000$ y los valores de r y s correspondientes son $r = -1.8927082 \times 10^{-11}$ y $s = -3.6550318 \times 10^{-11}$. Las dos raíces del factor cuadrático aproximado obtenido son 3.00000000000, 1.0000000000 y el polinomio reducido $q_5(x)$, de grado cinco, es $q_5(x) = x^5 - 24x^4 + 223x^3 - 996x^2 + 2116x - 1680$.

Si aplicamos Deflación, es decir, aplicamos el método de Bairstow al polinomio reducido $q_5(x)$, se obtiene para los valores iniciales $u_0 = 3$ y $v_0 = 5$, en la décima iteración los valores de $u = 8.00000000000$, $v = -12.0000000000$, $r = 1.4388490 \times 10^{-13}$ y $s = 9.1660013 \times 10^{-13}$. Las raíces del factor cuadrático aproximado correspondiente son 6.0000000000 y 2.00000000000, y el polinomio reducido, de grado tres, es $q_3(x) = x^3 - 16x^2 + 83x - 140$.

Al aplicar nuevamente Deflación, para los valores iniciales $u_0 = 8$ y $v_0 = 30$, se obtiene en la séptima iteración $u = 9.0000000000$, $v = -20.0000000000$, $r = 8.6330942 \times 10^{-13}$ y $s = 8.3950624 \times 10^{-12}$. Las raíces del factor cuadrático correspondiente son 5.00000000000 y 4.0000000000, y el polinomio reducido de grado uno es $q_1(x) = x - 7$, que nos lleva a obtener como última raíz aproximada de la ecuación polinómica original el valor 7.0.

Las raíces exactas de la ecuación polinómica dada, son 1, 2, 3, 4, 5, 6 y 7. ♦

TALLER 3.

1. Use el método de eliminación Gaussiana simple (sin pivoteo) con sustitución regresiva y aritmética exacta para resolver, si es posible, los sistemas lineales $AX = b$ siguientes, y encuentre matrices **P** de permutación, **L** triangular inferior con sus elementos diagonales iguales a **1** y **U** escalonada (triangular superior) tales que $PA = LU$.

a)
$$\begin{cases} x_1 - x_2 + 3x_3 = 2 \\ 3x_1 - 3x_2 + x_3 = -1 \\ x_1 + x_2 = 3 \end{cases}$$

b)
$$\begin{cases} x_1 - \frac{1}{2}x_2 + x_3 = 4 \\ 2x_1 - x_2 - x_3 + x_4 = 5 \\ x_1 + x_2 = 2 \\ x_1 - \frac{1}{2}x_2 + x_3 + x_4 = 5 \end{cases}$$

2. Use el algoritmo 3.2 y aritmética de precisión sencilla en un computador para resolver, si es posible, los siguientes sistemas de ecuaciones

a) $$\begin{cases} \frac{1}{4}x_1+\frac{1}{5}x_2+\frac{1}{6}x_3=9 \\ \frac{1}{3}x_1+\frac{1}{4}x_2+\frac{1}{5}x_3=8 \\ \frac{1}{2}x_1+\ x_2+\ 2x_3=8 \end{cases}$$

b) $$\begin{cases} x_1+\frac{1}{2}x_2+\frac{1}{3}x_3+\frac{1}{4}x_4=\frac{1}{6} \\ \frac{1}{2}x_1+\frac{1}{3}x_2+\frac{1}{4}x_3+\frac{1}{5}x_4=\frac{1}{7} \\ \frac{1}{3}x_1+\frac{1}{4}x_2+\frac{1}{5}x_3+\frac{1}{6}x_4=\frac{1}{8} \\ \frac{1}{4}x_1+\frac{1}{5}x_2+\frac{1}{6}x_3\ +\frac{1}{7}x_4=\frac{1}{9} \end{cases}$$

3. Resuelva los siguientes sistemas $AX=b$ por Eliminación Gaussiana sin pivoteo. Chequee si **A** tiene factorización **LU** con **L** triangular inferior con sus elementos diagonales iguales a **1** y **U** escalonada (triangular superior).

a) $A=\begin{pmatrix} 1 & 1 & -1 \\ 1 & 2 & -2 \\ -2 & 1 & 1 \end{pmatrix}, \quad b=\begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}$

b) $A=\begin{pmatrix} 4 & 3 & 2 & 1 \\ 3 & 4 & 3 & 2 \\ 2 & 3 & 4 & 3 \\ 1 & 2 & 3 & 4 \end{pmatrix}, \quad b=\begin{pmatrix} 1 \\ 1 \\ -1 \\ -1 \end{pmatrix}$

4. Dados los cuatro sistemas de ecuaciones lineales $AX=b^{(1)}$, $AX=b^{(2)}$, $AX=b^{(3)}$ y $AX=b^{(4)}$, donde

$$A=\begin{pmatrix} 2 & -3 & 1 \\ 1 & 1 & -1 \\ -1 & 1 & -3 \end{pmatrix}, \quad b^{(1)}=\begin{pmatrix} 2 \\ -1 \\ 0 \end{pmatrix}, \quad b^{(2)}=\begin{pmatrix} 6 \\ 4 \\ 5 \end{pmatrix}, \quad b^{(3)}=\begin{pmatrix} 0 \\ 1 \\ -3 \end{pmatrix}, \quad b^{(4)}=\begin{pmatrix} -1 \\ 0 \\ 0 \end{pmatrix}$$

a) Resuelva los sistemas lineales aplicando eliminación Gaussiana a la matriz aumentada $\left(A \vdots b^{(1)}\ b^{(2)}\ b^{(3)}\ b^{(4)}\right)$, y luego haciendo sustitución regresiva.

b) Resuelva los sistemas lineales usando eliminación Gaussiana para obtener matrices **P**, **L** y **U** tales que $PA=LU$, y luego siguiendo los pasos siguientes:

Paso1: Calcular $Pb^{(i)}$, $i=1,2,3,4$.

Paso 2: Resolver, para $c^{(i)}$, $Lc^{(i)}=Pb^{(i)}$, $i=1,2,3,4$, por sustitución progresiva.

Paso 3: Resolver, para $X^{(i)}$, $UX^{(i)}=c^{(i)}$, $i=1,2,3,4$, por sustitución regresiva.

c) Resuelva los sistemas lineales aplicando el método de Gauss-Jordan a la matriz aumentada de **a)**.

d) Resuelva los sistemas lineales encontrando la inversa de la matriz A y calculando los productos $A^{-1}b^{(i)}$, $i=1,2,3,4$.

e) Cuál método de los anteriores parece ser más fácil? Cuál método de los anteriores requiere más operaciones?

5. Encontrar $\|X\|_\infty$ y $\|X\|_2$ para cada uno de los siguientes vectores:

a) $X = \left(3,-4,0,\frac{3}{2}\right)^T$ **b)** $X = (2,1,-3,4)^T$

c) $X = \left(\text{sen}k, \cos k, 2^k\right)^T$, para un entero positivo fijo k.

6. a) Verificar que la función $\|\cdot\|_1$ definida en $\mathbf{R}^n$ por

$$\|X\|_1 = \sum_{i=1}^{n} |x_i|$$

es una norma vectorial.

b) Encontrar $\|X\|_1$ para cada uno de los vectores dados en el ejercicio 5.

7. Demuestre que para todo $X \in \mathbf{R}^n$,

$$\|X\|_\infty \le \|X\|_2 \le \|X\|_1$$

y que las igualdades pueden ocurrir, aún para vectores no nulos.

8. Demuestre que para todo $X \in \mathbf{R}^n$,

$$\|X\|_1 \le n\|X\|_\infty \text{ y } \|X\|_2 \le \sqrt{n}\|X\|_\infty$$

9. a) Encuentre $\|A\|_2$, $\|A\|_1$ y $\|A\|_\infty$ para cada una de las siguientes matrices

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

b) Calcule el radio espectral $\rho(A)$ para cada una de las matrices dadas en **a)**.

10. Demuestre que si A es simétrica, entonces $\|A\|_2 = \rho(A)$.

11. Calcule $\text{Cond}_\infty(A)$ y $\text{Cond}_*(A)$ para cada una de las matrices dadas en el ejercicio **9.a)**.

12. Sea $A \in \mathbf{R}_{n\times n}$. Demuestre que $\text{Cond}(A) = \text{Cond}(\alpha A)$ para cualquier escalar $\alpha \in \mathbf{R}$, $\alpha \neq 0$.

13. a) Considere el sistema lineal $AX = b$ dado por

$$\begin{pmatrix} 1 & 1 \\ 1 & 1.01 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 2 \\ 2.01 \end{pmatrix}$$

y calcule su solución exacta X.

b) Considere ahora el sistema perturbado $(A + \delta A)X = b$ dado por

$$\begin{pmatrix} 1 & 1 \\ 1 & 1.011 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 2 \\ 2.01 \end{pmatrix}$$

y calcule su solución exacta $\tilde{X}$.

c) Calcule $\dfrac{\|x - \tilde{x}\|_\infty}{\|x\|_\infty}$ y compárela con la cota de error obtenida a partir del teorema 3.4. Es la matriz A mal condicionada?

14. Considere el sistema

$$\begin{cases} .780x_1 + .563x_2 = .217 \\ .913x_1 + .659x_2 = .254 \end{cases}$$

Calcule el vector error residual $R = A\tilde{X} - b$, para las dos soluciones aproximadas $\tilde{X}_1 = (.341, -.087)^T$ y $\tilde{X}_2 = (.999, -1.001)^T$ y concluya, a partir únicamente del tamaño de estos errores residuales, cuál es la mejor aproximación de la solución del sistema. Verifique que la solución exacta del sistema es $X = (1, -1)^T$.

15. El sistema

$$\begin{cases} x + y = 0 \\ x + .999999y = 1 \end{cases}$$

tiene solución exacta $x = 10^6$, $y = -10^6$. Encuentre la solución exacta del sistema

$$\begin{cases} x + y = 0 \\ x + 1.000001y = 1 \end{cases}$$

Comente ampliamente los resultados.

16. Considere las matrices

$$A = \begin{pmatrix} 1 & -1 \\ 1 & -1.00001 \end{pmatrix}, \quad B = \begin{pmatrix} 1 & -1 \\ -1 & 1.00001 \end{pmatrix}$$

Muestre que $\text{Cond}_*(A) \approx 1$ y $\text{Cond}_*(B) \approx 4 \times 10^5$. Muestre, sin embargo, que $\text{Cond}_2(A) = \text{Cond}_2(B)$. Concluya que $\text{Cond}_*(.)$ no es un buen número de condición para matrices no simétricas. Es A mal condicionada o bien condicionada?

17. La matriz de Hilbert $H^{(n)} = (h_{ij})_{n\times n}$ definida por $h_{ij} = \dfrac{1}{i+j-1}$, $1 \le i, j \le n$ es un importante ejemplo en el álgebra lineal numérica.

a) Encuentre la matriz $H^{(4)}$, demuestre que

$$\left[H^{(4)}\right]^{-1} = \begin{pmatrix} 16 & -120 & 240 & -140 \\ -120 & 1200 & -2700 & 1680 \\ 240 & -2700 & 6480 & -4200 \\ -140 & 1680 & -4200 & 2800 \end{pmatrix}$$

y calcule $\text{Cond}_\infty\left(H^{(4)}\right)$.

b) Resuelva el sistema lineal

$$H^{(4)} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 1 \end{pmatrix}$$

usando aritmética con redondeo a tres dígitos y compare el error real en la aproximación calculada con la cota de error dada en el teorema 3.3. Es la matriz $H^{(4)}$ mal condicionada?

18. Considere el sistema lineal

$$\begin{cases} 2x_1 + \dfrac{2}{3}x_2 + \dfrac{1}{3}x_3 = 1 \\ x_1 + 2x_2 - x_3 = 0 \\ 6x_1 + 2x_2 + 2x_3 = -2 \end{cases}$$

y verifique que su solución es $x_1 = 2.6$, $x_2 = -3.8$, $x_3 = -5.0$.

a) Usando aritmética de punto flotante decimal con redondeo a cuatro dígitos, resuelva el sistema anterior por el método de eliminación Gaussiana sin pivoteo.

b) Repita la parte **a)**, usando pivoteo parcial y pivoteo escalado de fila.

c) Cuál de las tres soluciones calculadas en **a)** y **b)** es la mejor? Explique.

19. a) Muestre todos los pasos intermedios, ésto es, los multiplicadores, los factores de escala S_i, y el vector de intercambios **p**, al aplicar el pivoteo escalado de fila sobre la matriz siguiente:

$$A = \begin{pmatrix} 1 & -2 & 3 \\ 3 & -5 & -1 \\ 2 & -4 & 2 \end{pmatrix}$$

b) Use la información de la parte **a)** para encontrar matrices **P**, **L** y **U**, correspondientes al método de pivoteo escalado de fila, tales que $PA = LU$.

c) Use la factorización $PA = LU$, para calcular $\det A$.

20. Resuelva cada uno de los siguientes sistemas de ecuaciones lineales, usando aritmética de computador en precisión simple y eliminación Gaussiana con: **i)** sin pivoteo, **ii)** pivoteo parcial, **iii)** pivoteo escalado de fila.

a)
$$\begin{cases} .2641x_1 + .1735x_2 + .8642x_3 = -.7521 \\ -.8641x_1 - .4243x_2 + .0711x_3 = .2501 \\ .9411x_1 + .0175x_2 + .1463x_3 = .6310 \end{cases}$$

b)
$$\begin{cases} x_1 + \frac{1}{2}x_2 + \frac{1}{3}x_3 = 2 \\ \frac{1}{2}x_1 + \frac{1}{3}x_2 + \frac{1}{4}x_3 = -1 \\ \frac{1}{3}x_1 + \frac{1}{4}x_2 + \frac{1}{5}x_3 = 0 \end{cases}$$

c)
$$\begin{cases} x_1 + \frac{1}{2}x_2 + \frac{1}{3}x_3 + \frac{1}{4}x_4 + \frac{1}{5}x_5 = 1 \\ \frac{1}{2}x_1 + \frac{1}{3}x_2 + \frac{1}{4}x_3 + \frac{1}{5}x_4 + \frac{1}{6}x_5 = 1 \\ \frac{1}{3}x_1 + \frac{1}{4}x_2 + \frac{1}{5}x_3 + \frac{1}{6}x_4 + \frac{1}{7}x_5 = 1 \\ \frac{1}{4}x_1 + \frac{1}{5}x_2 + \frac{1}{6}x_3 + \frac{1}{7}x_4 + \frac{1}{8}x_5 = 1 \\ \frac{1}{4}x_1 + \frac{1}{5}x_2 + \frac{1}{6}x_3 + \frac{1}{7}x_4 + \frac{1}{9}x_5 = 1 \end{cases}$$

En cada caso estime el número de condición, relativo a la norma $\|\cdot\|_\infty$, de la matriz de coeficientes y concluya sobre el bien o mal condicionamiento de esta matriz, y si es posible, sobre la bondad de la solución calculada.

21. Determine cuáles de las siguientes matrices son
i) simétricas,
ii) singulares,
iii) estrictamente dominantes diagonalmente (**E.D.D.**) por filas,

iv) definidas positivas.

a) $\begin{pmatrix} 2 & 1 \\ 1 & 3 \end{pmatrix}$ **b)** $\begin{pmatrix} -2 & 1 \\ 1 & -3 \end{pmatrix}$ **c)** $\begin{pmatrix} 2 & 1 & 0 \\ 0 & 3 & 2 \\ 1 & 2 & 4 \end{pmatrix}$ **d)** $\begin{pmatrix} 2 & -1 & 0 \\ -1 & 4 & 2 \\ 0 & 2 & 2 \end{pmatrix}$

22. Defina la matriz tridiagonal de orden **n**

$$A_n = \begin{pmatrix} 2 & -1 & 0 & & \cdots & 0 \\ -1 & 2 & -1 & 0 & & \\ 0 & -1 & 2 & -1 & & \\ \vdots & & & & & \vdots \\ 0 & & \cdots & & -1 & 2 \end{pmatrix}$$

a) Encuentre una fórmula general para $A_n = LU$, con L triangular inferior con sus elementos diagonales iguales a uno y U triangular superior.

b) Use la factorización obtenida en **a)** para resolver el sistema $A_n X = b_n$ donde $b_n = (1,1,..,1)^T$, para $n = 3,\ 4,\ 5,\ 6$.

c) Con base en la respuesta obtenida en **a)**, muestre que la matriz A_n es invertible.

23. Pruebe que si $A = LL^T$ con $L \in \mathbf{R}_{n\times n}$ no singular, entonces A es simétrica y definida positiva.

24. Usando el método de Choleski, encuentre la factorización $A = LL^T$ para las siguientes matrices:

a) $A = \begin{pmatrix} 2.25 & -3.0 & 4.5 \\ -3.0 & 5.0 & -10.0 \\ 4.5 & -10.0 & 34.0 \end{pmatrix}$ **b)** $A = \begin{pmatrix} 15 & -18 & 15 & -3 \\ -18 & 24 & -18 & 4 \\ 15 & -18 & 18 & -3 \\ -3 & 4 & -3 & 1 \end{pmatrix}$

25. Considere la matriz $A = \begin{pmatrix} 2 & 3 \\ 1 & 4 \end{pmatrix}$. Verifique que la matriz A no es estrictamente dominante diagonalmente (por filas), pero que los métodos iterativos de Jacobi y Gauss-Seidel, para resolver cualquier sistema $AX = b$, convergen.

26. Demuestre que para la matriz

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

las iteraciones de Jacobi convergen y las de Gauss-Seidel divergen.

27. Considere el sistema

$$\begin{cases} 2x + y + z = 4 \\ x + 2y + z = 4 \\ x + y + 2z = 4 \end{cases}$$

a) Muestre que la matriz de coeficientes del sistema no es estrictamente dominante diagonalmente (por filas).

b) Partiendo de $X^{(0)} = (.8,.8,.8)^T$, muestre que las iteraciones de Jacobi oscilan entre los valores $(1.2,1.2,1.2)^T$ y $(.8,.8,.8)^T$.

c) Muestre que las iteraciones de Gauss-Seidel convergen a la solución $X = (1,1,1)^T$, calculando iteraciones hasta que $\left\| X^{(k)} - X^{(k-1)} \right\|_\infty < 10^{-3}$.

28. Para cada uno de los siguientes sistemas de ecuaciones, explique si los métodos iterativos de Jacobi y Gauss-Seidel convergen o no. En los casos donde haya convergencia calcule las iteraciones hasta que $\left\| X^{(k)} - X^{(k-1)} \right\|_\infty < 10^{-3}$.

a) $\begin{cases} 2x_1 + x_2 = 1 \\ x_1 + 6x_2 = 3 \\ 2x_2 + x_3 = 1 \end{cases}$ **b)** $\begin{cases} 2x_1 + x_2 - 3x_3 = -1 \\ -x_1 + 3x_2 + 2x_3 = 12 \\ 3x_1 + x_2 - 3x_3 = 0 \end{cases}$ **c)** $\begin{cases} -x_1 + 2x_2 + 3x_3 = 0 \\ 4x_1 - x_2 + x_3 = 6 \\ 2x_1 + 3x_2 - x_3 = -2 \end{cases}$

d) $\begin{cases} 2x_1 + x_2 - x_3 = 1 \\ x_1 + x_2 = -1 \\ x_1 - x_2 + 2x_3 = 2 \end{cases}$ **e)** $\begin{cases} x_1 + 2x_2 + x_4 = 0 \\ 3x_1 - x_2 + 4x_3 = 2 \\ x_1 - x_3 + 3x_4 = -1 \\ 2x_1 + x_2 - x_3 = 1 \end{cases}$

29. Para cada uno de los sistemas del ejercicio 28, si la matriz de coeficientes no es estrictamente dominante diagonalmente (por filas), reordénelo de modo que el nuevo sistema equivalente tenga matriz de coeficientes lo más cercana posible a ser estrictamente dominante diagonalmente (por filas) y estudie la convergencia o divergencia de los métodos iterativos de Jacobi y Gauss-Seidel para estos sistemas reordenados. En los casos donde haya convergencia calcule las iteraciones hasta que $\left\| X^{(k)} - X^{(k-1)} \right\|_\infty < 10^{-3}$.

30. Para cada uno de los sistemas reordenados del ejercicio 30, use el método **SOR** con $w = 1.2$, $w = .8$.

31. Resuelva los siguientes sistemas de ecuaciones no-lineales usando el método de Punto Fijo y el método de Newton-Raphson. En los casos donde haya convergencia del método, calcule las iteraciones hasta que $\left\| X^{(k)} - X^{(k-1)} \right\|_\infty < 10^{-3}$. En cada caso haga una gráfica que ilustre cuántas soluciones reales tiene el sistema.

a) $\begin{cases} x^2 + y^2 = 4 \\ x^3 - y = 0 \end{cases}$ **b)** $\begin{cases} x^2 - y^2 = 4 \\ e^{-x} + xy = 1 \end{cases}$ **c)** $\begin{cases} 4x^2 - y^2 = 0 \\ 4xy^2 - x = 1 \end{cases}$

32. Use el método de Newton-Raphson para aproximar un punto crítico de la función

$$f(x,y) = x^4 + xy + (1+y)^2$$

33. Considere el polinomio

$$p(x) = 3x^5 - 7x^4 - 5x^3 + x^2 - 8x + 2$$

a) Haga una gráfica que ilustre cuántas raíces reales tiene la ecuación $p(x) = 0$.

b) Aplique el algoritmo 3.7 (método de Bairstow) con punto inicial $(u_0, v_0) = (3,1)$ y $Tol = 10^{-3}$. Una vez que haya encontrado un factor cuadrático $x^2 - ux - v$, use Deflación para encontrar todas las raíces de la ecuación $p(x) = 0$.