TEMA 3: Métodos numéricos para la resolución del sistemas lineales y no lineales

Índice

1. Introducción **3-2**

2. Resolución de sistemas de ecuaciones no lineales **3-3**
2.1. Generalización de la iteración del punto fijo 3-3
2.2. Cotas del error 3-6
2.3. Método de Newton-Raphson en varias variables 3-8
2.4. Métodos Cuasi-Newton 3-10

3. Métodos directos para sistemas lineales **3-10**
3.1. Resolución de sistemas por eliminación gaussiana 3-10
3.1.1. Estrategias de pivotaje 3-16
3.2. Factorización LU 3-18
3.3. Método de Cholesky 3-20
3.4. Análisis del error 3-24

4. Métodos iterativos para sistemas lineales **3-27**
4.1. Iteración del punto fijo. Condiciones de convergencia 3-27
4.1.1. Condición alternativa de convergencia 3-28
4.1.2. Cotas del error 3-31
4.2. Métodos iterativos usuales 3-32
4.2.1. El método de Jacobi 3-33
4.2.2. El método de Gauss-Seidel 3-34
4.2.3. Convergencia de los métodos 3-35

1. Introducción

La primera parte del tema está dedicada al diseño de métodos numéricos que permitan aproximar la raíz de un sistema de ecuaciones no lineales. Dichos métodos son una generalización de los vistos en el Tema 2 para el caso de una variable. En particular, consideraremos el método de Newton-Raphson.

La modelización de problemas de la Física, Mecánica, etc. conducen en muchas ocasiones, directamente o después de un proceso de discretización, a la resolución de un sistema de ecuaciones lineales de dimensión finita. Además, otros problemas del Análisis Numérico como los de aproximación, interpolación, etc. recurren a la resolución de sistemas. Es por ello fundamental desarrollar métodos numéricos eficientes que obtengan la solución de sistemas de ecuaciones lineales, de lo que se ocupará el resto del tema.

Estudiaremos la resolución de sistemas de n ecuaciones lineales con n incógnitas, los cuales representaremos en forma matricial como $Ax = b$, donde $A = (a_{ij})_{1\leqslant i,j\leqslant n}$ es la matriz de coeficientes del sistema, $x = (x_i)_{1\leqslant i\leqslant n}$ el vector de las incógnitas y $b = (b_i)_{1\leqslant i\leqslant n}$ el vector de los términos independientes:

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

Nos centraremos en el caso de coeficientes reales, ya que si fueran complejos el problema se podría trasladar a la resolución de dos sistemas reales de tamaño n. Además consideraremos que $|A| \neq 0$, lo que nos indica que el sistema tiene solución y es única.

Partiendo del sistema $Ax = b$, donde A es una matriz cuadrada de orden n no singular, una primera idea que podemos aplicar para obtener la solución del sistema es calcularla como $x = A^{-1}b$, donde con A^{-1} representamos la inversa de la matriz A. Una posible forma de hallar la inversa, sería calcular los adjuntos de los elementos de A^t y dividir por el determinante de A, con lo que el número de operaciones a realizar sería de orden $n^2n!$.

Otra manera de resolver este problema sería utilizar la regla de Cramer, donde la componente i-ésima del vector solución se calcula como:

$$x_i = \frac{|A_i|}{|A|}, \quad 1 \leqslant i \leqslant n,$$

donde A_i es la matriz que resulta al sustituir en A la columna i-ésima por el vector de términos independientes. Este método requiere efectuar, teniendo en cuenta la definición de determinante, un número de operaciones del orden de $n^2n!$.

Nota. En un determinante hay $n!$ términos que hay que sumar, lo cual supone $n! - 1$ sumas. Cada término es el producto de n elementos, luego requiere $n - 1$ multiplicaciones. Así pues, en cada determinante se realizan $n!(n-1)$ productos y $n! - 1$ sumas. Como en total son $n + 1$ determinantes, el número de operaciones que se realiza es de: $n!(n-1)(n+1)$ productos, $(n! - 1)(n+1)$ sumas, y n cocientes correspondientes a la división por $|A|$.

En la práctica el tiempo de cálculo necesario es exagerado, incluso para un ordenador, y además, los errores de redondeo pueden llegar a ser tan grandes que invaliden el resultado final.

Por todo lo anterior parece necesario construir otro tipo de métodos que sean más eficientes. Los métodos que vamos a construir se agrupan en dos clases: métodos directos y métodos iterativos. Los primeros persiguen la obtención de la solución exacta del sistema al cabo de un número finito de operaciones elementales; en dichos métodos no existe error de truncatura, pero sí son sensibles a los errores de redondeo. Los métodos iterativos tienen como objetivo la construcción de una sucesión de vectores de $\mathbb{R}^n$ cuyo límite sea la solución del sistema, x.

Dentro de los métodos directos destacaremos el método de Gauss, así como aquellos que permiten factorizar la matriz A como producto de una matriz triangular inferior L y una matriz triangular superior U, entre los que destaca el método de Cholesky.

Respecto de los métodos iterativos, centraremos nuestro estudio en aquellos que resultan ser una generalización del proceso de iteración del punto fijo, como son el método de Jacobi y el de Gauss-Seidel.

2. Resolución de sistemas de ecuaciones no lineales

Hemos tratado en el tema anterior las ecuaciones de la forma $f(x) = 0$, donde f es una función real de variable real. Casi todas las cuestiones que se han considerado allí pueden ser generalizadas sin mucha dificultad al caso en que $f(x)$ sea una función *vectorial*

$$\begin{array}{rcl} f : D \subset \mathbb{R}^n & \longrightarrow & \mathbb{R}^n \\ x = (x_1, \ldots, x_n) & \longrightarrow & f(x) = \big(f_1(x), \ldots, f_n(x)\big) \end{array}$$

es decir, al caso en que nos encontramos ante un sistema de n ecuaciones no lineales con n incógnitas.

Así pues, nuestro problema se formula ahora en los siguientes términos: encontrar $c = (c_1, c_2, \ldots, c_n)$ solución del sistema de n ecuaciones con n incógnitas dado por:

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

donde $f_1, f_2, \ldots, f_n$ son las funciones componentes de una cierta función vectorial f.

2.1. Generalización de la iteración del punto fijo

Siguiendo el esquema unidimensional nos plantearemos, en primer lugar, la generalización del proceso de *iteración del punto fijo*. La extensión de los métodos de iteración del punto fijo a los sistemas de ecuaciones no lineales es sencilla desde el punto de

vista teórico, pero sin embargo desde el punto de vista práctico presenta mayores problemas que en el caso unidimensional. Por una parte, existen grandes dificultades para localizar a priori una región en la que se encuentre la raíz, pues no disponemos ahora de herramientas del tipo de los Teoremas de Bolzano y de Rolle y, por otra parte, es mucho más costoso comprobar las condiciones de convergencia global.

Dada una ecuación $f(x) = 0$, $x \in \mathbb{R}^n$, con $f : \mathbb{R}^n \to \mathbb{R}^n$, debemos reemplazar dicha ecuación por otra equivalente de la forma $x = g(x)$, donde $x \in \mathbb{R}^n$ y $g : \mathbb{R}^n \to \mathbb{R}^n$, esto es, en términos de las componentes de $f(x)$ y $g(x)$ la equivalencia se expresa como

$$f_i(x_1, \ldots, x_n) = 0 \quad \longleftrightarrow \quad x_i = g_i(x_1, \ldots, x_n), \quad i = 1, 2, \ldots, n.$$

Por ejemplo, la ecuación

$$f(x_1, x_2, x_3) = (x_1 + e^{x_2} x_3,\ 3x_2 - \operatorname{sen} x_1 + x_3,\ x_3 x_1 + \ln x_2) = (0, 0, 0)$$

puede expresarse en la forma

$$\begin{aligned} x_1 &= -e^{x_2} x_3, \\ x_2 &= \frac{\operatorname{sen} x_1 - x_3}{3}, \\ x_3 &= -\frac{\ln x_2}{x_1}, \end{aligned}$$

es decir que una función de iteración para la ecuación $f(x) = 0$ es

$$g(x_1, x_2, x_3) = \left(-e^{x_2} x_3,\ \frac{\operatorname{sen} x_1 - x_3}{3},\ -\frac{\ln x_2}{x_1} \right).$$

A la hora de plantearnos resultados similares al Teorema de convergencia global visto para el caso de una variable, hemos de reflexionar en primer lugar sobre el hecho de que la sucesión $x^{(m+1)} = g(x^{(m)})$ es ahora una sucesión vectorial, pues cada uno de sus elementos es un vector de $\mathbb{R}^n$. Para analizar la convergencia de la sucesión se considerará a $\mathbb{R}^n$ como espacio vectorial normado, con cualquiera de las posibles normas, pues todas son equivalentes. Recordemos que las normas más habituales en $\mathbb{R}^n$ son las siguientes:

$$\begin{aligned} \|x\|_1 &= \sum_{i=1}^{n} |x_i| && \text{(Norma uno)} \\ \|x\|_2 &= \left(\sum_{i=1}^{n} x_i^2 \right)^{1/2} && \text{(Norma euclídea)} \\ \|x\|_\infty &= \max_{1 \leqslant i \leqslant n} |x_i| && \text{(Norma infinito)} \end{aligned}$$

donde en todos los casos $x = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$. Teniendo en cuenta la definición de sucesión convergente en los espacios normados, es evidente que

$$\lim_{m \to \infty} x^{(m)} = c \quad \Longleftrightarrow \quad \lim_{m \to \infty} \|x^{(m)} - c\| = 0.$$

La equivalencia de las normas hace que sea indiferente la norma elegida a la hora de determinar la convergencia.

Nuestro primer objetivo será demostrar un teorema de convergencia global que generalice el que existía para el caso unidimensional.

Teorema 3.1. *Sea $g : S \subset \mathbb{R}^n \to \mathbb{R}^n$ verificando:*

(i) Existe un conjunto $D = [a_1, b_1] \times \cdots \times [a_n, b_n] \subset S$ tal que $g(x) \in D$, para todo $x \in D$.

(ii) Para alguna norma sobre $\mathbb{R}^n$, existe L $(0 \leqslant L < 1)$ tal que

$$\|g(x) - g(y)\| \leqslant L\|x - y\|, \ \forall x, y \in D.$$

Entonces existe un único punto fijo c de g en D y la sucesión definida por $x^{(m+1)} = g(x^{(m)})$, con $x^{(0)}$ un elemento arbitrario de D, converge a c.

Nota. Al igual que ocurría en el caso de las funciones de una variable, si g es suficientemente regular se puede sustituir la hipótesis (ii) por otra que resulte más fácil de comprobar. En particular, se puede demostrar que si todas las componentes de g, g_i, $i = 1, \ldots, n$, son de clase 1 en D y se verifica que

$$\left|\frac{\partial g_i(x)}{\partial x_j}\right| \leqslant \frac{L}{n}, \ \forall x \in D, \text{ con } 0 \leqslant L < 1, \text{ para } i, j = 1, \ldots, n \tag{1}$$

entonces se cumple la condición (ii) con la misma constante L si se toma la norma 1 o la norma infinito. Detallamos la prueba en el caso de la norma 1. En primer lugar, observemos que de la fórmula de Taylor en varias variables se deduce que para todo $x, y \in D$, podemos escribir

$$g_i(x) - g_i(y) = \sum_{j=1}^{n} \frac{\partial g_i(\xi_i)}{\partial x_j}(x_j - y_j),$$

donde ξ_i es un punto intermedio entre x e y.

Ahora bien, si se verifica (1), se deduce que

$$|g_i(x) - g_i(y)| \leqslant \frac{L}{n} \sum_{j=1}^{n} |x_j - y_j|,$$

y sumando sobre i de 1 a n se obtiene la desigualdad

$$\sum_{i=1}^{n} |g_i(x) - g_i(y)| \leqslant L \sum_{j=1}^{n} |x_j - y_j|,$$

que, de acuerdo con la definición de la norma 1, se transforma en

$$\|g(x) - g(y)\|_1 \leqslant L\|x - y\|_1.$$

Así pues, si $g \in C^1(D)$ y verifica la condición (i) del Teorema y la condición dada en (1), se puede asegurar la convergencia de la iteración del punto fijo a la solución de $x = g(x)$ en D, cualquiera que sea $x^{(0)} \in D$.

2.2. Cotas del error

De manera totalmente análoga a como se hace para el caso de una variable, se puede demostrar que en las hipótesis del Teorema de Convergencia Global, se tienen las siguientes acotaciones del error:

(i) $\|x^{(m)} - c\| \leqslant \dfrac{L^m}{1-L}\|x^{(1)} - x^{(0)}\|$,

(ii) $\|x^{(m)} - c\| \leqslant \dfrac{L}{1-L}\|x^{(m)} - x^{(m-1)}\|$.

Aunque formalmente las cotas son válidas cualquiera que sea la norma elegida, las más ajustadas serán aquellas que corresponden a las normas para las que se verifique la segunda condición del teorema de convergencia.

Ejemplo 3.2. *Consideremos el sistema de ecuaciones no lineales definido por:*

$$\begin{aligned} 8000x_1 + 0.5x_2^2 + 2x_3^3 &= 18008, \\ x_1 + 10500x_2 + x_3^3 &= 43003, \\ x_1^3 + 0.5x_2 + 7850x_3 &= 78510. \end{aligned}$$

Construya un sistema equivalente de la forma $g(x_1, x_2, x_3) = (x_1, x_2, x_3)$ *de manera que la función* g *verifique las hipótesis del teorema de convergencia global en* $D = [1.5, 2.5] \times [3.5, 4.5] \times [9.5, 10.5]$.

Resolución. La posibilidad más simple de obtener una ecuación equivalente del tipo buscado es

$$\begin{aligned} x_1 &= \frac{18008 - 0.5x_2^2 - 2x_3^3}{8000}, \\ x_2 &= \frac{43003 - x_1 - x_3^3}{10500}, \\ x_3 &= \frac{78510 - x_1^3 - 0.5x_2}{7850}, \end{aligned}$$

que corresponde a la siguiente función de iteración:

$$g(x) = \left(\frac{18008 - 0.5x_2^2 - 2x_3^3}{8000}, \frac{43003 - x_1 - x_3^3}{10500}, \frac{78510 - x_1^3 - 0.5x_2}{7850} \right).$$

Puesto que $g \in C^1(D)$, para ver si estamos en las hipótesis de convergencia, vamos a analizar si se verifican las condiciones:

(i) $g(x) \in D$, para todo $x \in D$.

(ii) Existe L $(0 \leqslant L < 1)$ tal que $\left| \dfrac{\partial g_i(x)}{\partial x_j} \right| \leqslant \dfrac{L}{n}$, con $i, j = 1, 2, 3$.

Para probar la condición (i), observemos que, en este caso, por el tipo de función con la que estamos trabajando, es sencillo acotar superior e inferiormente los valores que pueden tomar las funciones componentes g_1, g_2 y g_3 en D. En efecto, es evidente que

$$g_1(x_1, 4.5, 10.5) \leqslant g_1(x_1, x_2, x_3) \leqslant g_1(x_1, 3.5, 9.5),$$

para todo $(x_1, x_2, x_3) \in D$ y, por consiguiente, $1.960328 \leqslant g_1(x_1, x_2, x_3) \leqslant 2.227671$, de donde se sigue que $g_1(x_1, x_2, x_3) \in [1.5, 2.5]$, para todo $(x_1, x_2, x_3) \in D$.

De modo análogo,

$$3.9850 \leqslant g_2(x_1, x_2, x_3) \leqslant 4.0137, \quad \forall (x_1, x_2, x_3) \in D,$$

y

$$9.998996815 \leqslant g_3(x_1, x_2, x_3) \leqslant 10.000621, \quad \forall (x_1, x_2, x_3) \in D.$$

es decir, que se verifica la condición $g(D) \subset D$.

Para probar (ii), observemos que

$$\frac{\partial g_1}{\partial x_1}(x_1, x_2, x_3) = \frac{\partial g_2}{\partial x_2}(x_1, x_2, x_3) = \frac{\partial g_3}{\partial x_3}(x_1, x_2, x_3) = 0,$$

$$\begin{aligned}
\frac{\partial g_1}{\partial x_2}(x_1, x_2, x_3) &= \frac{-x_2}{8000}, \quad \frac{\partial g_1}{\partial x_3}(x_1, x_2, x_3) = \frac{-6x_3^2}{8000}, \\
\frac{\partial g_2}{\partial x_1}(x_1, x_2, x_3) &= \frac{-1}{10500}, \quad \frac{\partial g_2}{\partial x_3}(x_1, x_2, x_3) = \frac{-3x_3^2}{10500}, \\
\frac{\partial g_3}{\partial x_1}(x_1, x_2, x_3) &= \frac{-3x_1^2}{7850}, \quad \frac{\partial g_3}{\partial x_2}(x_1, x_2, x_3) = \frac{-0.5}{7850},
\end{aligned}$$

de donde se deduce que $\max_{1 \leqslant i,j \leqslant 3} |\partial g_i / \partial x_j| = 0.08268$, por lo que tomando $L = 3 \cdot 0.08268 = 0.24804$ se tiene (1).

Así pues, la función g tiene un único punto fijo en D y la sucesión vectorial definida por

$$\begin{aligned}
x_1^{(m+1)} &= \frac{18008 - 0.5(x_2^{(m)})^2 - 2(x_3^{(m)})^3}{8000}, \\
x_2^{(m+1)} &= \frac{43003 - x_1^{(m)} - (x_3^{(m)})^3}{10500}, \\
x_3^{(m+1)} &= \frac{78510 - (x_1^{(m)})^3 - 0.5(x_2^{(m)})}{7850},
\end{aligned}$$

con $(x_1^{(0)}, x_2^{(0)}, x_3^{(0)}) \in D$ converge a dicho punto fijo. ◀

2.3. Método de Newton-Raphson en varias variables

Al igual que en el caso de una variable nos podemos plantear de nuevo las cuestiones relativas al orden y a la velocidad de convergencia. La definición de orden de convergencia para las sucesiones vectoriales es idéntica a las reales sin más que sustituir $\lim_{n\to\infty} |e_{n+1}|/|e_n|^p$ por $\lim_{n\to\infty} \|e_{n+1}\|/\|e_n\|^p$. En estas condiciones, tiene perfecto sentido plantearnos la posibilidad de generalizar el método de Newton-Raphson. Es posible introducir este método imponiendo condiciones sobre la función de iteración que garanticen la convergencia cuadrática de modo similar a como se realizó en el caso de una variable.

El el caso de una variable, para resolver una ecuación $f(x) = 0$, en el método de Newton se construye un proceso iterativo de la forma:

$$x^{(m+1)} = x^{(m)} - \frac{f(x^{(m)})}{f'(x^{(m)})}. \tag{2}$$

En el caso de varias variables, la derivada de la función se sustituye por su matriz jacobiana, de manera que (2) se transforma en:

$$x^{(m+1)} = x^{(m)} - J_f^{-1}(x^{(m)}) f(x^{(m)}). \tag{3}$$

donde J_f es la matriz jacobianas de f:

$$J_f(x) = \left(\frac{\partial f_i(x)}{\partial x_j}\right).$$

No detallaremos en este caso un resultado sobre convergencia global por su complejidad y su escasa utilidad desde un punto de vista práctico, pero sí señalaremos que si f tiene derivadas parciales segundas continuas en un entorno de la raíz y además su Jacobiano no se anula en dicho entorno, se puede garantizar la convergencia del método de Newton-Raphson y además dicha convergencia es cuadrática.

Para hacer los cálculos del método de Newton, no se requiere invertir la matriz jacobiana en cada iteración, sino que suele usarse la relación

$$J_f(x^{(m)})(x^{(m+1)} - x^{(m)}) = -f(x^{(m)}). \tag{4}$$

Entonces, si conocemos $x^{(m)}$, usamos (4) para hallar $\Delta x_{m+1} = x^{(m+1)} - x^{(m)}$, de donde obtenemos

$$x^{(m+1)} = x^{(m)} + \Delta x_{m+1}.$$

A continuación, hallemos las fórmulas explícitas del método en el caso de dos variables. Utilizamos notación de subíndice para las derivadas parciales con el objeto de simplificar las expresiones. En el caso $n = 2$, la inversa de la matriz jacobiana puede

calcularse fácilmente:

$$J_f(x,y) = \begin{pmatrix} \frac{\partial f_1}{\partial x}(x,y) & \frac{\partial f_1}{\partial y}(x,y) \\ \frac{\partial f_2}{\partial x}(x,y) & \frac{\partial f_2}{\partial y}(x,y) \end{pmatrix} = \begin{pmatrix} (f_1)_x & (f_1)_y \\ (f_2)_x & (f_2)_y \end{pmatrix},$$

$$J_f^{-1}(x,y) = \frac{1}{|J_f(x,y)|} \begin{pmatrix} (f_2)_y & -(f_1)_y \\ -(f_2)_x & (f_1)_x \end{pmatrix},$$

donde $|J_f(x,y)| = (f_1)_x(f_2)_y - (f_2)_x(f_1)_y$ es el determinante de la matriz jacobiana. Denotando por x_m e y_m a las componentes de la aproximación m-ésima a la solución, la ecuación (3) en componentes se escribirá como:

$$x_{m+1} = x_m - \left[\frac{f_1(f_2)_y - f_2(f_1)_y}{|J_f|}\right]_{(x_m,y_m)},$$

$$y_{m+1} = y_m - \left[\frac{f_2(f_1)_x - f_1(f_2)_x}{|J_f|}\right]_{(x_m,y_m)},$$

donde con el subíndice del corchete se indica que todas las funciones del interior del mismo están evaluadas en el punto señalado.

Observemos que el esfuerzo de cálculo requerido para aplicar el método de Newton-Raphson es muy grande pues se necesita evaluar en cada iteración dos funciones y cuatro derivadas parciales.

Ejemplo 3.3. *Dado el sistema de ecuaciones no lineales*

$$\frac{x^2 - y^2}{2} - x + \frac{7}{24} = 0,$$

$$xy - y + \frac{1}{9} = 0,$$

halle las primeras iteraciones del método de Newton-Raphson partiendo del vector nulo.

Resolución. Se puede comprobar que el sistema dado tiene una solución única en $[0, 0.4] \times [0, 0.4]$. Vamos a efectuar las primeras iteraciones del método de Newton-Raphson partiendo de $x^{(0)} = 0$ e $y^{(0)} = 0$. Para ello observemos que, en este caso,

$$f_1(x,y) = \frac{x^2 - y^2}{2} - x + \frac{7}{24}, \quad f_2(x,y) = xy - y + \frac{1}{9},$$

y derivando parcialmente respecto de x e y se tiene

$$(f_1)_x(x,y) = x - 1, \quad (f_1)_y(x,y) = -y,$$

$$(f_2)_x(x,y) = y, \qquad (f_2)_y(x,y) = x - 1,$$

de donde el determinante de la matriz jacobiana es

$$J_f(x,y) = (x-1)^2 + y^2.$$

Aplicando la fórmula del método de Newton se obtienen los siguientes resultados:

m	$x^{(m)}$	$y^{(m)}$
0	0	0
1	0.2916	0.1111
2	0.3346	0.1636
3	0.3333	0.1667

donde se pone de manifiesto la convergencia de la sucesión. ◀

2.4. Métodos Cuasi-Newton

Uno de los problemas del método de Newton es que requiere mucho esfuerzo computacional: en cada iteración hay que calcular una matriz jacobiana, lo que supone evaluar n^2 derivadas parciales, y luego resolver un sistema de n ecuaciones y n incógnitas con dicha matriz. El esfuerzo computaciones total para realizar una iteración del método de Newton conlleva, por tanto, al menos $n^2 + n$ evaluaciones de funciones escalares (n^2 para la matriz jacobiana y n para las f_i) junto con un número de operaciones aritméticas $O(n^3)$ para resolver el sistema lineal. Para valores altos de n, este número de operaciones es prohibitivo, por lo que se recurre a otros métodos como el método de Broyden.

3. Métodos directos para sistemas lineales

3.1. Resolución de sistemas por eliminación gaussiana

Una matriz A se dice triangular superior (respectivamente triangular inferior) si todos los elementos situados bajo la diagonal principal son nulos, es decir $a_{ij} = 0$ si $i > j$ (respectivamente $a_{ij} = 0$ si $i < j$). Un sistema de ecuaciones se dice triangular superior (respectivamente triangular inferior) si su matriz de coeficientes lo es.

Consideremos el siguiente sistema triangular superior:

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1n} \\ 0 & a_{22} & a_{23} & \cdots & a_{2n} \\ 0 & 0 & a_{33} & \cdots & a_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & a_{nn} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \\ b_3 \\ \vdots \\ b_n \end{pmatrix}$$

en el cual supondremos que $|A| \neq 0$, es decir $a_{ii} \neq 0$, $\forall i$.

Si utilizamos la regla de Cramer para resolver este sistema no aprovecharemos el gran número de ceros que posee esta matriz, por lo que para resolver un sistema triangular superior se utiliza el procedimiento conocido como *sustitución regresiva*, que podemos expresar en la forma:

$$\begin{aligned} x_n &= \frac{b_n}{a_{nn}}, \\ x_i &= \frac{1}{a_{ii}}\left(b_i - \sum_{j=i+1}^{n} a_{ij}x_j\right), \quad i = n-1, n-2, \ldots, 1. \end{aligned}$$

Observemos que en primer lugar se despeja el valor de x_n, este valor se sustituye en la ecuación $n-1$, para a continuación despejar x_{n-1} y así sucesivamente hasta calcular x_1. Este proceso reduce drásticamente el número de operaciones que supondría el uso de la regla de Cramer, ya que el número de operaciones en este caso es n^2.

Otro tipo de sistemas que conviene destacar son los sistemas diagonales, es decir aquellos en los que la matriz de coeficientes verifica que $a_{ij} = 0$ para todo valor de $i \neq j$,

$$\begin{pmatrix} a_{11} & 0 & 0 & \cdots & 0 \\ 0 & a_{22} & 0 & \cdots & 0 \\ 0 & 0 & a_{33} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & a_{nn} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \\ \vdots \\ b_n \end{pmatrix}$$

y cuya solución se obtiene de forma inmediata como

$$x_i = \frac{b_i}{a_{ii}}, \quad i = 1, 2, \ldots, n, \qquad (a_{ii} \neq 0, \ \forall\, i).$$

Teniendo en cuenta que estos sistemas son de fácil solución y que el número de operaciones realizadas para calcular tales soluciones es sensiblemente menor que el que resulta utilizando otras técnicas, como por ejemplo, la regla de Cramer, vamos a intentar reducir el problema general a la resolución de un sistema triangular (o diagonal), siendo esta la esencia de los métodos directos.

Para ello es útil recordar la solución de un sistema de ecuaciones lineales no cambia si se hacen las siguientes transformaciones:

- ✯ Multiplicar una ecuación por una constante no nula.
- ✯ Intercambiar ecuaciones entre sí.
- ✯ Sumar a una ecuación un múltiplo de otra.
- ✯ Aplicar reiteradamente cualesquiera de las operaciones anteriores (en particular sumar a una ecuación una combinación lineal del resto).

Estudiaremos en primer lugar el método de Gauss, basado en la transformación de la matriz A de partida en una matriz triangular superior.

El método de Gauss consiste en la reducción del sistema a otro de matriz de coeficientes triangular a través de $n-1$ etapas, en cada una de las cuales se va consiguiendo dejar en forma triangular una fila más.

Comenzaremos el estudio de este método presentando un ejemplo:

Ejemplo 3.4.

$$\begin{array}{rcrcrcl} 3x_1 & - & x_2 & + & 2x_3 & = & 12 \\ x_1 & + & 2x_2 & + & 3x_3 & = & 11 \\ 2x_1 & - & 2x_2 & - & x_3 & = & 2 \end{array} \quad \longrightarrow \quad \begin{pmatrix} 3 & -1 & 2 \\ 1 & 2 & 3 \\ 2 & -2 & -1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} \begin{pmatrix} 12 \\ 11 \\ 2 \end{pmatrix}$$

$$\begin{pmatrix} 3 & -1 & 2 \\ 1 & 2 & 3 \\ 2 & -2 & -1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 12 \\ 11 \\ 2 \end{pmatrix} \quad \begin{matrix} F_2 - \frac{1}{3}F_1 \\ \\ F_3 - \frac{2}{3}F_1 \end{matrix}$$

$$\begin{pmatrix} 3 & -1 & 2 \\ 0 & 7/3 & 7/3 \\ 0 & -4/3 & -7/3 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 12 \\ 7 \\ -6 \end{pmatrix} \quad F_3 + \frac{4}{7}F_2$$

$$\begin{pmatrix} 3 & -1 & 2 \\ 0 & 7/3 & 7/3 \\ 0 & 0 & -1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 12 \\ 7 \\ -2 \end{pmatrix} \quad \begin{matrix} x_1 = 3, \\ x_2 = 1, \\ x_3 = 2. \end{matrix}$$

De forma general, si partimos de un sistema de la forma $Ax = b$, el método de Gauss nos permite tranformar la matriz A de partida en una matriz triangular superior U a través de $n-1$ pasos, obteniéndose la siguiente sucesión de matrices intermedias:

$$A = \bar{A}_1 \longrightarrow A_1 \longrightarrow \bar{A}_2 \longrightarrow A_2 \longrightarrow \cdots \longrightarrow \bar{A}_n = A_n = U,$$

con $A_k = (a_{ij}^{(k)})_{1 \leqslant i,j \leqslant n}$ y $\bar{A}_k = (\bar{a}_{ij}^{(k)})_{1 \leqslant i,j \leqslant n}$. En el caso en que $\bar{a}_{kk}^{(k)} = 0$, la matriz A_k se obtiene a partir de $\bar{A}_k$ situando en la posición (k,k) un elemento no nulo de la columna k por debajo de la diagonal principal, para lo cual las filas de $\bar{A}_k$ se intercambian. Esto es siempre posible ya que hemos considerado que $|A| \neq 0$ y por lo tanto la matriz A_k es también no singular, $\forall k = 1, \ldots, n$.

En el caso de que $\bar{a}_{kk}^{(k)} \neq 0$, $\forall k$, no es necesario realizar cambios de filas y $\bar{A}_k = A_k$, con lo que se tiene

$$\begin{aligned} A = A_1 \longrightarrow A_2 \longrightarrow \cdots \longrightarrow A_n = U, \\ b = b^{(1)} \longrightarrow b^{(2)} \longrightarrow \cdots \longrightarrow b^{(n)}, \end{aligned}$$

donde la matriz A_k tiene ceros en los elementos de las $k-1$ primeras columnas bajo la diagonal principal y donde $b^{(k)} = (b_i^{(k)})_{1 \leqslant i \leqslant n}$ es el vector que contiene los términos independientes en el paso k, con $k = 1, 2, \ldots, n$.

En un paso k cualquiera obtendremos la matriz A_{k+1} a partir de la A_k haciendo ceros los elementos de la columna k que se encuentran bajo la diagonal principal.

Sea A_k la matriz

$$A_k = \begin{pmatrix} a_{11}^{(k)} & a_{12}^{(k)} & \cdots & a_{1k}^{(k)} & \cdots & \cdots & a_{1n}^{(k)} \\ 0 & a_{22}^{(k)} & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & a_{kk}^{(k)} & a_{k,k+1}^{(k)} & \cdots & a_{kn}^{(k)} \\ 0 & 0 & \cdots & a_{k+1,k}^{(k)} & a_{k+1,k+1}^{(k)} & \cdots & a_{k+1,n}^{(k)} \\ 0 & 0 & \cdots & a_{k+2,k}^{(k)} & a_{k+2,k+1}^{(k)} & \cdots & a_{k+2,n}^{(k)} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & a_{nk}^{(k)} & a_{n,k+1}^{(k)} & \cdots & a_{nn}^{(k)} \end{pmatrix}$$

con $a_{kk}^{(k)} \neq 0$. Para hacer ceros en las posiciones $(k+1,k)$, $(k+2,k)$, $\ldots$, (n,k), restaremos a cada fila un múltiplo adecuado de la fila k, esto es

$$F_{k+1} - \frac{a_{k+1,k}^{(k)}}{a_{kk}^{(k)}} F_k, \quad F_{k+2} - \frac{a_{k+2,k}^{(k)}}{a_{kk}^{(k)}} F_k, \quad \ldots, \quad F_n - \frac{a_{nk}^{(k)}}{a_{kk}^{(k)}} F_k,$$

con lo que obtendremos la siguiente matriz

$$A_{k+1} = \begin{pmatrix} a_{11}^{(k+1)} & a_{12}^{(k+1)} & \cdots & a_{1k}^{(k+1)} & \cdots & \cdots & a_{1n}^{(k+1)} \\ 0 & a_{22}^{(k+1)} & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & a_{kk}^{(k+1)} & a_{k,k+1}^{(k+1)} & \cdots & a_{kn}^{(k+1)} \\ 0 & 0 & \cdots & 0 & a_{k+1,k+1}^{(k+1)} & \cdots & a_{k+1,n}^{(k+1)} \\ 0 & 0 & \cdots & 0 & a_{k+2,k+1}^{(k+1)} & \cdots & a_{k+2,n}^{(k+1)} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & a_{n,k+1}^{(k+1)} & \cdots & a_{nn}^{(k+1)} \end{pmatrix}.$$

De esta forma, el elemento (i,j) de la matriz A_{k+1} se ha calculado utilizando la expresión

$$a_{ij}^{(k+1)} = \begin{cases} a_{ij}^{(k)}, & 1 \leqslant i \leqslant k, \quad 1 \leqslant j \leqslant n, \\ a_{ij}^{(k)} - m_{ik} a_{kj}^{(k)}, & k+1 \leqslant i \leqslant n, \quad k+1 \leqslant j \leqslant n, \\ 0, & k+1 \leqslant i \leqslant n, \quad 1 \leqslant j \leqslant k. \end{cases}$$

Como se observa las k primeras filas de A_k no se modifican.

Es conveniente destacar en el proceso los siguientes elementos:

✯ La ecuación k-ésima se denomina ecuación pivotal.

✯ El elemento (k,k) lo denotaremos como el elemento pivote del paso k.

✯ Para hacer cero en la posición (i,k) utilizamos el siguiente cociente, $m_{ik} = a_{ik}^{(k)}/a_{kk}^{(k)}$, que llamaremos el multiplicador (i,k).

En lo que refiere a los términos independientes, las componentes del vector $b^{(k+1)}$ se obtendrán a partir de las de $b^{(k)}$ de la siguiente forma:

$$b_i^{(k+1)} = \begin{cases} b_i^{(k)}, & 1 \leqslant i \leqslant k, \\ b_i^{(k)} - m_{ik} b_k^{(k)}, & k+1 \leqslant i \leqslant n. \end{cases}$$

Ejemplo 3.5. *Resuelva utilizando el método de eliminación gaussiana, el siguiente sistema lineal de ecuaciones:*

$$\begin{aligned} 3x_1 - x_2 + 2x_3 &= 12, \\ x_1 + 2x_2 + 3x_3 &= 11, \\ 2x_1 - 2x_2 - x_3 &= 2. \end{aligned}$$

Resolución. En las distintas etapas del proceso de eliminación de Gauss, se obtiene que:

$$A = A_1 = \begin{pmatrix} 3 & -1 & 2 \\ 1 & 2 & 3 \\ 2 & -2 & -1 \end{pmatrix}, \qquad b^{(1)} = \begin{pmatrix} 12 \\ 11 \\ 2 \end{pmatrix} = b,$$

$$\left.\begin{matrix} m_{21} = 1/3 \\ m_{31} = 2/3 \end{matrix}\right\} \Longrightarrow A_2 = \begin{pmatrix} 3 & -1 & 2 \\ 0 & 7/3 & 7/3 \\ 0 & -4/3 & -4/3 \end{pmatrix}, \quad b^{(2)} = \begin{pmatrix} 12 \\ 7 \\ -6 \end{pmatrix},$$

$$m_{32} = -4/7 \Longrightarrow A_3 = \begin{pmatrix} 3 & -1 & 2 \\ 0 & 7/3 & 7/3 \\ 0 & 0 & -1 \end{pmatrix}, \qquad b^{(3)} = \begin{pmatrix} 12 \\ 7 \\ -2 \end{pmatrix}.$$

El sistema final resultante es $A_3 x = b^{(3)}$:

$$\begin{pmatrix} 3 & -1 & 2 \\ 0 & 7/3 & 7/3 \\ 0 & 0 & -1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 12 \\ 7 \\ -2 \end{pmatrix} \tag{5}$$

y sus soluciones son $x_1 = 3,\ x_2 = 1,\ x_3 = 2$. ◀

Teniendo en cuenta la descripción que acabamos de hacer del método de Gauss, se deduce el siguiente algoritmo:

ALGORITMO DE GAUSS:

Para $k = 1$ hasta $n - 1$ hacer

(Estudiar a_{kk})

  Para $i = k + 1$ hasta n hacer

    $m_{ik} = a_{ik}/a_{kk}$

    Para $j = k$ hasta n hacer

      $a_{ij} = a_{ij} - m_{ik}a_{kj}$

    fin j

  $b_i = b_i - m_{ik}b_k$

  fin i

fin k

El estudio del elemento a_{kk} supone comprobar si dicho elemento es nulo, en cuyo caso situaremos en esa posición un elemento no nulo de la columna k por debajo de la diagonal principal, para lo cual será necesario intercambiar filas. También podemos plantearnos la elección de alguna estrategia de pivotaje, cuyo estudio abordaremos más adelante.

A la hora de evaluar el número de operaciones que se realizan en el proceso hay que tener en cuenta los pasos del proceso de eliminación y la etapa de sustitución regresiva. Se puede comprobar que el número total de operaciones realizadas es $(4n^3 + 9n^2 - 7n)/6$.

A continuación se expone una tabla en la que se ve claramente que el método de Gauss reduce notablemente el número de operaciones que se realizarían aplicando la regla de Cramer (expresaremos el número de operaciones a través de la mayor potencia de 10 que en él aparece):

n	GAUSS	CRAMER
10	10^2	10^8
50	10^4	10^{64}
100	10^5	10^{160}

La parte de eliminación del método de Gauss también nos permite comprobar si $|A| = 0$, lo cual ocurrirá cuando en el paso k se verifique

$$a_{kk}^{(k)} = a_{k+1,k}^{(k)} = a_{k+2,k}^{(k)} = \cdots = a_{nk}^{(k)} = 0.$$

Nota. Obsérvese que las operaciones del proceso de eliminación de Gauss no alteran el valor absoluto del determinante de la matriz de partida, por lo que podemos asegurar que todas las matrices A_k tienen en mismo determinante en valor absoluto. En cuanto

a la matriz A_k, desarrollando reiteradamente por los adjuntos de los elementos de la primera columna de las diversas submatrices, su determinante viene dado por:

$$|A_k| = a_{11}^{(k)} a_{22}^{(k)} \cdots a_{k-1,k-1}^{(k)} \begin{vmatrix} a_{kk}^{(k)} & a_{k,k+1}^{(k)} & \dots & a_{kn}^{(k)} \\ \dots & \dots & \dots & \dots \\ a_{nk}^{(k)} & a_{n,k+1}^{(k)} & \dots & a_{nn}^{(k)} \end{vmatrix}.$$

Ejemplo 3.6. *Utilice la eliminación gaussiana para resolver el sistema:*

$$\begin{aligned} 2x_1 + 3x_2 + x_3 &= 6, \\ 4x_1 + 6x_2 + 3x_3 &= 13, \\ 6x_1 + 9x_2 + x_3 &= 16. \end{aligned}$$

Resolución. En este caso se obtiene que $m_{21} = 2$ y $m_{31} = 3$. Esto nos lleva al sistema

$$\begin{aligned} 2x_1 + 3x_2 + x_3 &= 6, \\ x_3 &= 1, \\ -2x_3 &= -2. \end{aligned}$$

En este caso $a_{22}^{(2)} = a_{32}^{(2)} = 0$, de donde se deduce que $|A| = 0$, con lo que el sistema o no posee solución o tiene infinitas soluciones. Este ejemplo corresponde a la segunda situación y las soluciones son:

$$x_3 = 1, \quad x_1 = \frac{5 - 3x_2}{2}.$$ ◀

3.1.1. Estrategias de pivotaje

En la aplicación del método de Gauss, sólo hemos contemplado hasta ahora la posibilidad de elección de la ecuación pivotal cuando, en un paso k genérico, el elemento $a_{kk}^{(k)}$ era nulo. Aún en el caso de que $a_{kk}^{(k)} = 0$ el algoritmo que se utilizó se basaba en una elección al azar, buscando como elemento pivote el primer elemento no nulo de la columna k bajo la diagonal principal. Veamos a continuación un ejemplo que muestra que no es indiferente la elección de la ecuación pivotal.

Ejemplo 3.7 (Forsythe, 1967). *Resuelva utilizando aritmética de punto flotante con tres cifras significativas el sistema:*

$$\begin{aligned} (0.100)10^{-3}x + (0.100)10^{1}y &= (0.100)10^{1}, \\ (0.100)10^{1}x + (0.100)10^{1}y &= (0.200)10^{1}. \end{aligned}$$

Resolución. Puede comprobarse que la solución exacta utilizando seis digitos significativos es: $x = 1.00010$, $y = 0.999900$, lo que nos sirve de referencia para valorar la aproximación que se obtiene al utilizar un número limitado de dígitos significativos.
Si efectuamos el proceso normal de eliminación de Gauss, tenemos que $m_{21} = (0.100)10^{5}$, lo que nos lleva al sistema

$$\begin{aligned} (0.100)10^{-3}x + (0.100)10^{1}y &= (0.100)10^{1}, \\ -(0.100)10^{5}y &= -(0.100)10^{5}, \end{aligned}$$

y así, utilizando sustitución regresiva, se deduce que $x = 0.00$ e $y = 1.00$, solución bastante distinta de la solución auténtica.

Sin embargo, cambiando los papeles entre la primera y la segunda ecuación, el sistema obtenido es

$$\begin{aligned}(0.100)10^1x &+ (0.100)10^1y = (0.200)10^1,\\ (0.100)10^{-3}x &+ (0.100)10^1y = (0.100)10^1,\end{aligned}$$

y las soluciones que a partir de él se obtienen son $x = 1.00$ e $y = 1.00$ con la consiguiente mejora de la exactitud. ◀

Hay que tener en cuenta que en el proceso de eliminación de Gauss se realizan divisiones por los elementos $a_{kk}^{(k)}$ y que si dicho elemento es muy pequeño, entonces un error de redondeo, aunque sea pequeño en valor absoluto, puede producir errores muy grandes en el cociente, con el consiguiente perjuicio para la precisión del resultado. la forma más usual de elegir el elemento pivote corresponde a la estrategia de *pivotaje parcial* que se describe a continuación.

En cada paso k ($1 \leqslant k \leqslant n-1$) se realizarán los cambios necesarios entre filas para situar en la posición (k,k) el mayor elemento en valor absoluto de la columna k entre las filas k y n, ambas inclusive, es decir, situaremos en la posición (k,k) al elemento $a_{lk}^{(k)}$ verificando:

$$|a_{lk}^{(k)}| = \max_{k\leqslant i\leqslant n}\{|a_{ik}^{(k)}|\},$$

pasando a ser la ecuación l-ésima la ecuación pivotal.

Ejemplo 3.8. *Resuelva el sistema*

$$\begin{aligned}x_1 &+ x_2 - x_3 = 0,\\ 2x_1 &+ x_2 + x_3 = 7,\\ 3x_1 &- 2x_2 - x_3 = -4,\end{aligned}$$

utilizando el método de Gauss con estrategia de pivotaje parcial.

Resolución. Para la eliminación de la incógnita x_1 utilizaremos como pivote el elemento en la posición $(3,1)$, siendo este el coeficiente de x_1 en la tercera ecuación, con lo que intercambiando la primera y la tercera ecuación se tiene

$$\begin{aligned}3x_1 &- 2x_2 - x_3 = -4,\\ 2x_1 &+ x_2 + x_3 = 7,\\ x_1 &+ x_2 - x_3 = 0.\end{aligned}$$

Ahora, puesto que $m_{21} = 2/3$ y $m_{31} = 1/3$ se obtiene el sistema

$$\begin{aligned}3x_1 &- 2x_2 - x_3 = -4,\\ \frac{7}{3}x_2 &+ \frac{5}{3}x_3 = \frac{29}{3},\\ \frac{5}{3}x_2 &- \frac{2}{3}x_3 = \frac{4}{3}.\end{aligned}$$

Para la eliminación de x_2 la elección del pivote sería entre $7/3$ y $5/3$ y como $|7/3| > |5/3|$, entonces el pivote sería $7/3$ y no es necesario intercambiar ninguna ecuación. Puesto que $m_{32} = 5/7$, el sistema resultante es

$$\begin{aligned} 3x_1 - 2x_2 - x_3 &= -4, \\ \frac{7}{3}x_2 + \frac{5}{3}x_3 &= \frac{29}{3}, \\ -\frac{39}{21}x_3 &= \frac{173}{21}, \end{aligned} \tag{6}$$

el cual se resuelve por sustitución regresiva, obteniéndose $x_1 = 1$, $x_2 = 2$ y $x_3 = 3$.

Es ilustrativo resolver el ejercicio viendo la evolución de las *matrices ampliadas* de los diferentes sistemas que se van formando. Denotamos por $F_i^{(k)}$ a la fila i-ésima de la matriz ampliada correspondiente al sistema obtenido en el paso k-ésimo.

$$\left(\begin{array}{ccc|c} 1 & 1 & -1 & 0 \\ 2 & 1 & 1 & 7 \\ 3 & -2 & -1 & -4 \end{array}\right) \begin{array}{l} F_1^{(1)} \\ F_2^{(1)} \\ F_3^{(1)} \end{array} \sim \left(\begin{array}{ccc|c} 3 & -2 & -1 & -4 \\ 1 & 1 & -1 & 0 \\ 2 & 1 & 1 & 7 \end{array}\right) \begin{array}{l} F_1^{(2)} = F_3^{(1)} \\ F_2^{(2)} = F_2^{(1)} \\ F_3^{(2)} = F_1^{(1)} \end{array} \sim$$

$$\left(\begin{array}{ccc|c} 3 & -2 & -1 & -4 \\ 0 & \frac{7}{3} & \frac{5}{3} & \frac{29}{3} \\ 0 & \frac{5}{3} & \frac{-2}{3} & \frac{4}{3} \end{array}\right) \begin{array}{l} F_1^{(3)} = F_1^{(2)} \\ F_2^{(3)} = F_2^{(2)} - \frac{2}{3}F_1^{(2)} \\ F_3^{(3)} = F_3^{(2)} - \frac{1}{3}F_1^{(2)} \end{array} \sim \left(\begin{array}{ccc|c} 3 & -2 & -1 & -4 \\ 0 & \frac{7}{3} & \frac{5}{3} & \frac{29}{3} \\ 0 & 0 & \frac{-39}{21} & \frac{173}{21} \end{array}\right) \begin{array}{l} F_1^{(4)} = F_1^{(3)} \\ F_2^{(4)} = F_2^{(3)} \\ F_3^{(4)} = F_3^{(3)} - \frac{5}{7}F_2^{(3)} \end{array}$$

donde la última matriz ampliada corresponde al sistema triangular (6). ◀

3.2. Factorización LU

Consideremos el sistema de ecuaciones lineales $Ax = b$. Si se conoce una descomposición de la matriz A como producto de una matriz triangular inferior $L = (l_{ij})_{1 \leqslant i,j \leqslant n}$ por una matriz triangular superior $U = (u_{ij})_{1 \leqslant i,j \leqslant n}$, $A = LU$, entonces la resolución del sistema anterior queda reducida a la resolución de dos sistemas triangulares

$$Ax = b \quad \longrightarrow \quad (LU)x = b \quad \longrightarrow \quad \begin{cases} Ly = b \\ Ux = y \end{cases}$$

que podemos expresar como

$$\begin{pmatrix} l_{11} & 0 & \cdots & \cdots & \cdots & 0 \\ l_{21} & l_{22} & 0 & \cdots & \cdots & 0 \\ l_{31} & l_{32} & l_{33} & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ l_{n1} & l_{n2} & \cdots & \cdots & \cdots & l_{nn} \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \\ \vdots \\ b_n \end{pmatrix}$$

y

$$\begin{pmatrix} u_{11} & u_{12} & \cdots & \cdots & \cdots & u_{1n} \\ 0 & u_{22} & u_{23} & \cdots & \cdots & u_{2n} \\ 0 & 0 & u_{33} & u_{34} & \cdots & u_{3n} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \cdots & \cdots & u_{nn} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{pmatrix}.$$

Resolviendo el primero por sustitución progresiva calcularemos el vector y. Sustituyendo éste en el segundo sistema y utilizando sustitución regresiva podremos calcular el vector x que es la solución del sistema de partida.

De esta forma nos estamos planteando una nueva estrategia para resolver un sistema de ecuaciones lineales. La pregunta que debemos hacernos ahora es la siguiente: ¿cómo se obtienen las matrices L y U?

En primer lugar es evidente que a partir de la relación $A = LU$ obtenemos ciertas ecuaciones que permitirán el cálculo de los elementos de L y U:

$$l_{11}u_{11} = a_{11},$$

$$u_{1j} = \frac{a_{1j}}{l_{11}}, \quad j \geqslant 2,$$

$$l_{ij} = \frac{a_{i1}}{u_{11}}, \quad i \geqslant 2,$$

$$l_{kk}u_{kk} = a_{kk} - \sum_{r=1}^{k-1} l_{kr}u_{kr}, \quad k \geqslant 2,$$

$$u_{kj} = \frac{1}{l_{kk}}\left(a_{kj} - \sum_{r=1}^{k-1} l_{kr}u_{rj}\right), \quad j > k,\ k \geqslant 2,$$

$$l_{ik} = \frac{1}{u_{kk}}\left(a_{ik} - \sum_{r=1}^{k-1} l_{ir}u_{rk}\right), \quad i > k,\ k \geqslant 2.$$

El número de ecuaciones que se obtienen es inferior al número de incógnitas a determinar ($n^2 < n^2 + n$), por lo que en ocasiones se suelen fijar algunos de los elementos de tales matrices con el fin de obtener una única solución. Por ejemplo, se suele considerar que $l_{ii} = 1$, para todo valor de i.

Observaciones:

(O1) Los elementos de la matriz $A = LU$ pueden calcularse como:

$$a_{ij} = \sum_{k=1}^{\min\{i,j\}} l_{ik}u_{kj}, \quad 1 \leq i, j \leqslant n.$$

(O2) No toda matriz admite una descomposición LU. Por ejemplo, si $a_{11} = 0$, se tiene que:

$$\left.\begin{array}{l} a_{11} = l_{11}u_{11} = 0 \\ a_{1j} = l_{11}u_{1j},\ j \geqslant 2 \\ a_{i1} = l_{i1}u_{11},\ i \geq 2 \end{array}\right\} \quad \Longrightarrow \quad a_{1j} = 0, \forall j \quad \text{ó} \quad a_{i1} = 0, \forall i.$$

Un criterio que nos permite asegurar cuando dicha descomposición es posible es el siguiente resultado:

Teorema 3.9. *Siempre que la matriz A es no singular, existe una permutación de las filas de A tal que la matriz resultante es descomponible en la forma LU, con $l_{ii} = 1$ y $u_{ii} \neq 0$, para todo valor de i.*

Dicho de otra manera, si A es no singular, existe una matriz de permutación P tal que

$$PA = LU$$

donde L es una matriz triangular inferior con $l_{ii} = 1$, $\forall i$ y U es una matriz triangular superior no singular.

3.3. Método de Cholesky

Si en la factorización $A = LU$ se exige que $u_{ji} = l_{ij}$, $\forall i, j$, es decir, que $U = L^t$ y además que $l_{ii} > 0$ para todo valor de i, podría ser que la descomposición no existiera. Sin embargo podemos asegurar que si la matriz A es simétrica y definida positiva tal descomposición es posible, y se conoce como descomposición de Cholesky.

Definición 3.10 (Matriz simétrica y definida positiva). *Una matriz $A \in \mathcal{M}_n(\mathbb{R})$ se dice que es simétrica y definida positiva si*

(i) $A = A^t$, *es decir, A es simétrica, y*

(ii) $x^t Ax > 0, \forall x \neq 0$.

Teorema 3.11 (Caracterización de matrices simétricas y definidas positivas). *Sea $A \in \mathcal{M}_n(\mathbb{R})$. Entonces:*

(i) A es simétrica y definida positiva si y sólo si $\exists A^{-1}$ y es simétrica y definida positiva.

(ii) A es simétrica y definida positiva si y sólo si todas las submatrices principales de A son simétricas y definidas positivas.

(iii) A es simétrica y definida positiva si y sólo si todos sus menores principales son positivos y $A = A^t$.

(iv) A es simétrica y definida positiva si y sólo si todos los autovalores de A son positivos y $A = A^t$.

(v) A es simétrica y definida positiva si y sólo si $\exists P \in \mathcal{M}_n(\mathbb{R})$ no singular tal que $A = P\,P^t$.

Veamos a continuación cómo se puede plantear el cálculo de los elementos de la matriz L para que $LL^t = A$, es decir,

$$\begin{pmatrix} l_{11} & 0 & \cdots & \cdots & \cdots & 0 \\ l_{21} & l_{22} & 0 & \cdots & \cdots & 0 \\ l_{31} & l_{32} & l_{33} & \cdots & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ l_{n1} & l_{n2} & \cdots & \cdots & \cdots & l_{nn} \end{pmatrix} \begin{pmatrix} l_{11} & l_{21} & l_{31} & \cdots & \cdots & l_{n1} \\ 0 & l_{22} & l_{32} & \cdots & \cdots & l_{n2} \\ 0 & 0 & l_{33} & l_{34} & \cdots & l_{n3} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \cdots & \cdots & l_{nn} \end{pmatrix} = A.$$

Multiplicando la primera fila de L por la primera columna de L^t obtenemos a_{11}, por lo que

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

A partir del valor de l_{11} podemos calcular los elementos de la primera columna de L, sin más que multiplicar la fila i de L por la columa 1 de L^t:

$$l_{i1}l_{11} = a_{i1} \implies l_{i1} = \frac{a_{i1}}{l_{11}}.$$

Trabajando de manera análoga se obtendrán el resto de los elementos de L. Si los l_{ij}, $\forall i, j \leqslant k-1$ son conocidos, la columna k de L se calcula con las relaciones:

$$a_{kk} = \sum_{p=1}^{k-1} l_{kp}l_{kp} + l_{kk}l_{kk} \implies l_{kk} = \sqrt{a_{kk} - \sum_{p=1}^{k-1} l_{kp}^2}\,,$$

$$a_{ik} = \sum_{p=1}^{k-1} l_{ip}l_{kp} + l_{ik}l_{kk} \implies l_{ik} = \frac{1}{l_{kk}}\left(a_{ik} - \sum_{p=1}^{k-1} l_{ip}l_{kp}\right), \quad k+1 \leqslant i \leqslant n.$$

El proceso queda reflejado en el siguiente algoritmo:

ALGORITMO de CHOLESKY:

`Para` $k = 1$ `hasta` n `hacer`
  $l_{kk} = (a_{kk} - \sum_{r=1}^{k-1} l_{kr}^2)^{1/2}$
  `Para` $j = k+1$ `hasta` n `hacer`
    $l_{jk} = (a_{jk} - \sum_{r=1}^{k-1} l_{jr}l_{kr})/l_{kk}$
  `Fin j`
`Fin k`

Observaciones:

(O1) En primer lugar este método nos permite saber si una matriz simétrica es definida positiva o no, según que admita la descomposición $A = LL^t$ o no.

(O2) La descomposición de Cholesky también puede ser utilizada para calcular el determinante de la matriz A, ya que

$$A = LL^t \implies |A| = |L||L^t| = |L|^2 = \prod_{i=1}^{n} l_{ii}^2.$$

(O3) Esta factorización sirve para resolver sistemas de ecuaciones cuya matriz de coeficientes sea A. Cuando utilizamos tal descomposición para resolver el sistema de partida se tiene

$$Ax = b \longrightarrow (LL^t)x = b \longrightarrow \begin{cases} Ly = b \\ L^t x = y \end{cases}$$

El coste del método desde el punto de vista computacional, considerando también el coste asociado a la resolución de los dos sistemas triangulares, es $(2n^3 + 15n^2 + n)/6$ que es menor que el del método de Gauss (aproximadamente la mitad). Además un análisis de errores nos permite afirmar que el método de Cholesky es interesante por su poca sensibilidad a los errores de redondeo.

(O4) Si A es simétrica y definida positiva, entonces pueden existir varias matrices L, triangulares inferiores (e inversibles), tales que $A = LL^t$, pero solo una de ellas es de elementos diagonales positivos. Ejemplo:

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

Descomposición de Cholesky de A:

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

(O5) Si A es simétrica y definida positiva, pueden existir descomposiciones de A de la forma $A = LU$, donde L (U) es una matriz trangular inferior (superior) y $U \neq L^t$. Ejemplo:

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

Descomposición de Cholesky de A:

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

Ejemplo 3.12. *Halle la decomposición de Cholesky de la matriz:*

$$A = \begin{pmatrix} 2 & 2 & -1 & 0 \\ 2 & 5 & 1 & 1 \\ -1 & 1 & 14 & 5 \\ 0 & 1 & 5 & 3 \end{pmatrix}.$$

Resolución. En primer lugar veamos si A es simétrica y definida positiva. A es claramente simétrica. Hallemos a continuación sus menores principales:

$$\Delta_1 = |2| = 2 > 0, \qquad \Delta_2 = \begin{vmatrix} 2 & 2 \\ 2 & 5 \end{vmatrix} = 6 > 0,$$

$$\Delta_3 = \begin{vmatrix} 2 & 2 & -1 \\ 2 & 5 & 1 \\ -1 & 1 & 14 \end{vmatrix} = 73 > 0, \qquad \Delta_4 = |A| = 82 > 0.$$

Como A es simétrica y todos sus menores pincipales son positivos, concluimos que A es simétrica ydefinida positiva y, por tanto, admite factorización de Cholesky.

Tenemos que calcular una matriz triangular inferior $L \in \mathcal{M}_4(\mathbb{R})$ tal que $A = LL^t$, es decir:

$$\begin{pmatrix} 2 & 2 & -1 & 0 \\ 2 & 5 & 1 & 1 \\ -1 & 1 & 14 & 5 \\ 0 & 1 & 5 & 3 \end{pmatrix} = \begin{pmatrix} l_{11} & 0 & 0 & 0 \\ l_{21} & l_{22} & 0 & 0 \\ l_{31} & l_{32} & l_{33} & 0 \\ l_{41} & l_{42} & l_{43} & l_{44} \end{pmatrix} \begin{pmatrix} l_{11} & l_{21} & l_{31} & l_{41} \\ 0 & l_{22} & l_{32} & l_{42} \\ 0 & 0 & l_{33} & l_{43} \\ 0 & 0 & 0 & l_{44} \end{pmatrix}.$$

Cálculo de los elementos de la primera columna de L:

$$\begin{aligned} 2 = l_{11}^2 &\implies l_{11} = \sqrt{2}, \\ 2 = l_{21} l_{11} &\implies l_{21} = 2/l_{11} = 2/\sqrt{2}, \\ -1 = l_{31} l_{11} &\implies l_{31} = -1/l_{11} = -1/\sqrt{2}, \\ 0 = l_{41} l_{11} &\implies l_{41} = 0. \end{aligned}$$

Cálculo de los elementos no nulos de la segunda columna de L:

$$\begin{aligned} 5 = l_{21}^2 + l_{22}^2 &\implies l_{22} = \sqrt{5 - l_{21}^2} = \sqrt{3}, \\ 1 = l_{31} l_{21} + l_{32} l_{22} &\implies l_{32} = \frac{1}{l_{22}}(1 - l_{31} l_{21}) = 2/\sqrt{3}, \\ 1 = l_{41} l_{21} + l_{42} l_{22} &\implies l_{42} = \frac{1}{l_{22}}(1 - l_{41} l_{21}) = 1/\sqrt{3}. \end{aligned}$$

Cálculo de los elementos no nulos de la tercera columna de L:

$$\begin{aligned} 14 = l_{31}^2 + l_{32}^2 + l_{33}^2 &\implies l_{33} = \sqrt{14 - l_{31}^2 - l_{32}^2} = \sqrt{73/6}, \\ 5 = l_{41} l_{31} + l_{42} l_{32} + l_{43} l_{33} &\implies l_{43} = \frac{1}{l_{33}}(5 - l_{41} l_{31} - l_{42} l_{32}) = 13\sqrt{2/219}. \end{aligned}$$

Cálculo de los elementos no nulos de la cuarta columna de L:

$$3 = l_{41}^2 + l_{42}^2 + l_{43}^2 + l_{44}^2 \implies l_{44} = \sqrt{3 - l_{41}^2 - l_{42}^2 - l_{43}^2} = \sqrt{82/73}.$$

Por lo tanto, tenemos que $A = LL^t$, donde:

$$L = \begin{pmatrix} \sqrt{2} & 0 & 0 & 0 \\ \frac{2}{\sqrt{2}} & \sqrt{3} & 0 & 0 \\ \frac{-1}{\sqrt{2}} & \frac{2}{\sqrt{3}} & \sqrt{\frac{73}{6}} & 0 \\ 0 & \frac{1}{\sqrt{3}} & 13\sqrt{\frac{2}{219}} & \sqrt{\frac{82}{73}} \end{pmatrix}. \quad \blacktriangleleft$$

3.4. Análisis del error

Los métodos llamados directos proporcionan un valor exacto de la solución después de un número finito de operaciones elementales. ¿Qué sentido tiene entonces hablar de error?

Evidentemente, todo proceso que utiliza precisión finita está sujeto a un error de redondeo lo cual hace que normalmente no se obtenga la solución exacta, sino una solución aproximada que denotamos como $\hat{x}$, cometiéndose una desviación de la solución exacta x que denominaremos error, $e = x - \hat{x}$. Obviamente el error no se puede determinar en general de forma exacta ya que para ello necesitaríamos conocer el valor de x.

Buscaremos una forma de medir este error en la práctica, para lo que introduciremos el concepto de vector residuo. Llamaremos vector residuo al vector r dado por la expresión

$$r = Ax - A\hat{x} = A(x - \hat{x}) = Ae,$$

valor que sí podemos calcular pues sabemos que $Ax = b$, con lo que

$$r = b - A\hat{x}.$$

Podemos admitir r como una medida de hasta que punto $\hat{x}$ satisface el sistema $Ax = b$. Ahora bien, ¿hasta qué punto es fiable el residuo como medida del error? Los siguientes ejemplos ponen de manifiesto que la respuesta a esta pregunta depende del sistema que se maneje.

Sea el sistema de ecuaciones

$$\begin{aligned} 1.01x_1 + 0.99x_2 &= 2, \\ 0.99x_1 + 1.01x_2 &= 2, \end{aligned}$$

cuya solución exacta es $x_1 = x_2 = 1$.

En primer lugar, si consideramos la siguiente aproximación

$$\hat{x} = \begin{pmatrix} 1.01 \\ 1.01 \end{pmatrix} \quad \Longrightarrow \quad e = \begin{pmatrix} -0.01 \\ -0.01 \end{pmatrix}, \quad r = \begin{pmatrix} -0.02 \\ -0.02 \end{pmatrix},$$

tenemos un residuo "pequeño" y un error "pequeño". Sin embargo, también podemos considerar que la solución aproximada es

$$\hat{x} = \begin{pmatrix} 2 \\ 0 \end{pmatrix} \quad \Longrightarrow \quad e = \begin{pmatrix} -1 \\ 1 \end{pmatrix}, \quad r = \begin{pmatrix} -0.02 \\ 0.02 \end{pmatrix},$$

donde el residuo es "pequeño" pero el error es "grande".

Por otro lado, consideremos el sistema

$$\begin{aligned} 1.01x_1 &+ 0.99x_2 = 2, \\ 0.99x_1 &+ 1.01x_2 = -2, \end{aligned}$$

cuya solución exacta es $x_1 = 100$, $x_2 = -100$. Resolviendo el sistema obtenemos

$$\hat{x} = \begin{pmatrix} 101 \\ -99 \end{pmatrix} \quad \Longrightarrow \quad e = \begin{pmatrix} -1 \\ -1 \end{pmatrix}, \quad r = \begin{pmatrix} -2 \\ -2 \end{pmatrix},$$

en cuyo caso tenemos un residuo "grande" y un error "grande".

Los ejemplos anteriores pone de manifiesto que el tamaño del vector residuo no es siempre un indicador fiable del tamaño del error.

Para formalizar de algún modo estos conceptos, es necesario introducir el estudio de las normas sobre el espacio vectorial de las matrices.

Definición 3.13 (Norma matricial). *Una norma matricial es una aplicación*

$$\|\cdot\| : \mathcal{M}_n(\mathbb{R}) \longrightarrow \mathbb{R}$$

que verifica las siguientes propiedadades:

(P1) $\|A\| \geqslant 0, \quad \forall A \in \mathcal{M}_n(\mathbb{R}),$

(P2) $\|A\| = 0 \Leftrightarrow A = 0,$

(P3) $\|A + B\| \leqslant \|A\| + \|B\|, \quad \forall A, B \in \mathcal{M}_n(\mathbb{R}),$

(P4) $\|\lambda A\| = |\lambda|\|A\|, \quad \forall \lambda \in \mathbb{R}, \quad \forall A \in \mathcal{M}_n(\mathbb{R}),$

(P5) $\|AB\| \leqslant \|A\|\|B\|, \quad \forall A, B \in \mathcal{M}_n(\mathbb{R}).$

Obsérvese que las cuatro primeras propiedades caracterizan a las normas vectoriales mientras que la propiedad 5 es propia de las normas matriciales.

Definición 3.14. *Una norma matricial $\|\cdot\|_M$ en $\mathcal{M}_n(\mathbb{R})$ y una norma vectorial $\|\cdot\|_v$ en $\mathbb{R}^n$ se dice que son compatibles si se verifica:*

$$\|A\,x\|_v \leqslant \|A\|_M \|x\|_v, \ \forall A \in \mathcal{M}_n(\mathbb{R}), \ \forall x \in \mathbb{R}^n.$$

Definición 3.15 (Norma matricial subordinada). *Sea $\|\cdot\|_v$ una norma vectorial en $\mathbb{R}^n$. Se llama norma matricial subordinada a esa norma o norma matricial inducida, a la norma matricial definida así*

$$\|A\|_M = \max_{x \neq 0} \frac{\|A\,x\|_v}{\|x\|_v} = \max_{\|x\|_v = 1} \|A\,x\|_v.$$

Teorema 3.16. *Una norma matricial subordinada es compatible con la norma vectorial dada.*

Definición 3.17 (Radio espectral). *Dada $A \in \mathcal{M}_n(\mathbb{C})$, se define el radio espectral de A, y se representa por $\rho(A)$, como el máximo de los módulos de los autovalores de A, esto es*

$$\rho(A) = \text{máx}\{|\lambda| \ / \ \lambda \text{ es autovalor de } A\}.$$

A las normas matriciales más habituales se las designa del mismo modo que a las normas vectoriales a la que están subordinadas. Ejemplos de normas matriciales inducidas:

✯ $\|A\|_1 = \max\limits_{1\leqslant j\leqslant n}\left\{\sum\limits_{i=1}^{n}|a_{ij}|\right\}$ inducida por $\|x\|_1 = \sum\limits_{i=1}^{n}|x_i|$.

✯ $\|A\|_2 = \rho(A^tA)^{1/2}$ inducida por $\|x\|_2 = \left(\sum\limits_{i=1}^{n}|x_i|^2\right)^{1/2}$.

✯ $\|A\|_\infty = \max\limits_{1\leqslant i\leqslant n}\left\{\sum\limits_{j=1}^{n}|a_{ij}|\right\}$ inducida por $\|x\|_\infty = \max\limits_{1\leqslant i\leqslant n}|x_i|$.

Con estas nociones sobre normas podemos comparar de algún modo el tamaño del error relativo, $\|e\|/\|x\|$, y del residuo relativo, $\|r\|/\|b\|$, como se pone de manifiesto en el siguiente

Teorema 3.18. *Sea A una matriz no singular y $Ax = b$ un sistema lineal de ecuaciones. Entonces:*

$$\frac{1}{\kappa(A)}\frac{\|r\|}{\|b\|} \leqslant \frac{\|e\|}{\|x\|} \leqslant \kappa(A)\frac{\|r\|}{\|b\|}, \quad \kappa(A) \equiv \|A\|\|A^{-1}\|$$

donde las normas vectorial y matricial han de ser compatibles.

Al número real $\kappa(A)$ se le llama condicionamiento o número de condición de A.

El teorema anterior refleja claramente que la relación entre $\|e\|/\|x\|$ y $\|r\|/\|b\|$ depende del número $\|A\|\|A^{-1}\|$, que es llamado número de condición de la matriz A y se denota por $\kappa(A)$.

Observaciones:

(O1) Se cumple que $\kappa(A)$ es un número mayor o igual que uno, ya que

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

Nota. Veremos más adelante que el radio espectral de una matriz es siempre menor o igual que cualquiera de sus normas.

(O2) Si $\kappa(A) \gtrsim 1$, el residuo relativo es una buena estimación del error relativo, ya que $\kappa(A)$ y $1/\kappa(A)$ serán muy próximos. Se dice en este caso que el sistema está bien condicionado.

(O3) Si $\kappa(A) \gg 1$, $\kappa(A)$ y $1/\kappa(A)$ serán muy dferentes. En este caso no podemos asegurar que el residuo relativo sea una buena estimación del error relativo. Se dice en este caso que el sistema está mal condicionado.

Por lo dicho anteriormente, es conveniente que $\kappa(A)$ sea próximo a 1 como ocurre con las matrices ortogonales ($A^t = A^{-1}$), para las cuales $\kappa(A) = 1$ cuando se utiliza la norma matricial $\|\cdot\|_2$. Un ejemplo de matrices mal condicionadas son las matrices de Hilbert, definidas como:

$$A = (a_{ij}), \quad a_{ij} = \frac{1}{i+j-1}, \quad 1 \leqslant i, j \leqslant n.$$

4. Métodos iterativos para sistemas lineales

En la sección anterior hemos estudiado métodos directos para la resolución de sistemas de ecuaciones lineales. El número de operaciones de estos métodos es razonable para valores de n moderados pero para el tamaño de sistemas que aparece, por ejemplo, en la resolución aproximada de ecuaciones diferenciales (matrices de orden 1000 o más en determinados problemas) el número de operaciones es ya muy elevado y los errores de redondeo acumulados pueden anular la validez de los resultados obtenidos. Además su aplicación a matrices dispersas no siempre puede aprovechar la gran cantidad de ceros de estas matrices y lo que esto supone en el almacenamiento, ya que a lo largo del proceso pueden aparecer elementos no nulos en posiciones ocupadas anteriormente por elementos nulos.

Las razones anteriores son las que nos llevan a diseñar métodos alternativos, los llamados métodos iterativos, en los que se construyen sucesiones vectoriales que convergen a la solución del sistema. Dichos métodos son menos sensibles a los errores de redondeo en el sentido de que la falta de precisión se puede corregir aumentando el número de iteraciones. Por otra parte, estos métodos aprovechan las características de las matrices dispersas para almacenarlas en memoria de forma adecuada ya que la matriz no se modifica a lo largo del proceso.

4.1. Iteración del punto fijo. Condiciones de convergencia

La aplicación de los métodos de iteración del punto fijo a la resolución de ecuaciones ya fue tratado en el Tema 2. En nuestro caso la ecuación vectorial que debemos resolver es $F(x) = 0$, donde $F(x) = Ax - b$ y la función de iteración la elegiremos del tipo $G(x) = Bx + c$, donde B es una matriz cuadrada de orden n y c un vector de $\mathbb{R}^n$. Obviamente, la elección de B y c debe realizarse de modo que la solución de $G(x) = x$ coincida con la de $F(x) = 0$.

Como en todo proceso de iteración del punto fijo, consideraremos la sucesión vectorial

$$x^{(m+1)} = G(x^{(m)}) = Bx^{(m)} + c, \quad m = 0, 1, 2, \ldots$$

y analizaremos su convergencia a la solución del sistema para un vector inicial $x^{(0)}$ adecuado. Recordemos que la convergencia de $x^{(m)}$ a x equivale a la convergencia a 0 de la sucesión de números reales $\|x^{(m)} - x\|$.

Puesto que la solución de $F(x) = 0$ es $x = A^{-1}b$, la función G elegida debe verificar que $G(A^{-1}b) = A^{-1}b$, es decir, $A^{-1}b = BA^{-1}b + c$, y, por consiguiente,

$$c = (I - B)A^{-1}b. \tag{7}$$

Definición 3.19 (Método iterativo consistente). *Si B y c verifican la condición* (7) *se dice que el método iterativo $x^{(m+1)} = Bx^{(m)} + c$ es consistente para la resolución del sistema de ecuaciones $Ax = b$.*

Obviamente solo nos interesan los métodos consistentes y, supuesta esa condición, debemos determinar condiciones que garanticen la convergencia. El teorema siguiente proporciona un criterio para dicha convergencia que además es independiente del vector inicial elegido.

Teorema 3.20 (Primer criterio de convergencia). *Si B es una matriz cuadrada de orden n de manera que existe una norma matricial $\|\cdot\|_M$ tal que $\|B\|_M < 1$ entonces $Bx + c = x$ tiene una única solución y además la sucesión vectorial $(x^{(m)})$ definida por $x^{(m+1)} = Bx^{(m)} + c$ converge a dicha solución para cualquier vector inicial $x^{(0)}$.*

Recíprocamente, si el método $x^{(m+1)} = Bx^{(m)} + c$ es convergente, existe una norma matricial $\|\cdot\|_M$ tal que $\|B\|_M < 1$.

Observaciones:

(O1) El concepto de convergencia de un proceso de iteración del punto fijo en el ámbito de los sistemas de ecuaciones lineales se asocia siempre a que dicha convergencia no dependa del vector inicial elegido y, en lo que sigue, la palabra converge tendrá siempre ese significado.

(O2) El recíproco del teorema anterior, es decir, el hecho de que la convergencia de un método iterativo $x^{(m+1)} = Bx^{(m)} + c$ implique la existencia de una norma matricial $\|\cdot\|_M$ tal que $\|B\|_M < 1$, no es un resultado de utilidad práctica para descartar la convergencia de un método iterativo debido a la existencia de infinitas normas matriciales. Es decir, el hecho de que $\|B\| \geqslant 1$ para varias normas matriciales, no implica necesariamente que $(x^{(m)})$ sea divergente.

4.1.1. Condición alternativa de convergencia

Los problemas que presenta el **Teorema 3.20** como condición suficiente de convergencia (véase la observación (O2) del Teorema) nos motiva a encontrar una condición necesaria y suficiente alternativa aplicable en la práctica. En ella juegan un papel importante el radio espectral (y por tanto los autovalores) de la matriz B del método iterativo.

Recordemos que un número λ (real ó complejo) es un autovalor (o valor propio) de una matriz cuadrada de orden n, A, si existe un vector x (de $\mathbb{R}^n$ o de $\mathbb{C}^n$) no nulo de modo que $Ax = \lambda x$ y que a cada vector en esas condiciones se le denomina vector propio o autovector de A asociado al autovalor λ. Es un resultado conocido del Álgebra Lineal que los valores propios de una matriz A coinciden con las raíces del polinomio característico de A definido como $|A - \lambda I|$, donde se representa por I a la

matriz identidad de orden n. Asociado a los valores propios se encuentra el concepto de radio espectral de una matriz, que fue introducido en la **Definición 3.17**.

El radio espectral de una matriz guarda estrecha relación con el valor de las normas matriciales de dicha matriz como se pone de manifiesto en el siguiente resultado.

Teorema 3.21.

(i) Si $A \in M_n(\mathbb{R})$, entonces para toda norma matricial $\|\cdot\|$ se verifica que $\rho(A) \leqslant \|A\|$.

(ii) Dada $A \in M_n(\mathbb{R})$ y cualquier número real positivo ε existe una norma matricial $\|\cdot\|_\varepsilon$ tal que $\|A\|_\varepsilon \leqslant \rho(A) + \varepsilon$.

Del Teorema anterior se obtiene inmediatamente el siguiente

Corolario 3.22. *Para toda matriz $A \in \mathcal{M}_n(\mathbb{R})$ se tiene:*

$$\rho(A) = \text{ínf}\{\|A\| \ / \ \|\cdot\| \text{ es norma matricial}\}. \tag{8}$$

Ahora estamos en condiciones de presentar una condición necesaria y suficiente para la convergencia de un método iterativo en términos del radio espectral de la matriz de iteración B.

Teorema 3.23 (Segundo criterio de convergencia). *El método de iteración del punto fijo definido por $x^{(m+1)} = Bx^{(m)} + c$ es convergente si y sólo si $\rho(B) < 1$.*

Demostración. Es consecuencia de (8). Así, si $\rho(B) < 1$, debe existir una norma matricial $\|\cdot\|$ tal que $\|B\| < 1$ y, por el **Teorema 3.20**, el método es convergente. Recíprocamente, si el método converge entonces existe una norma matricial tal que $\|B\| < 1$ y así, por definición de ínfimo, se tiene que $\rho(B) < 1$. □

Observaciones:

(O1) La condición de convergencia que proporciona el **Teorema 3.23** tiene utilidad práctica en el caso en que se puedan calcular o estimar los autovalores de la matriz lo cual no siempre es simple.

(O2) Puede ocurrir que una matriz tenga algunas normas mayores que 1 mientras que su radio espectral sea menor que 1. Ejemplo:

$$B = \begin{pmatrix} 0 & 1/2 \\ 3/2 & 0 \end{pmatrix}, \quad \|B\|_1 = \|B\|_2 = \|B\|_\infty = 3/2 > 1,$$
$$\rho(B) = \sqrt{3}/2 < 1.$$

Interpretación del radio espectral

El radio espectral es un indicador de la velocidad de convergencia de la sucesión $(x^{(m)})$: cuanto más pequeño sea $\rho(B)$ más rápido converge el método iterativo (asintóticamente). Para ilustrar esta afirmación, consideremos el caso en que B es diagonalizable, por lo que existe una matriz no singular P tal que $D = P^{-1}BP$ es una matriz diagonal. Entonces, tomando una norma vectorial y otra matricial compatibles, se tiene

que:

$$\begin{aligned} x^{(m)} - s &= B^m(x^{(0)} - s), \\ \left\|x^{(m)} - s\right\| &\leqslant \|P\|\,\|D\|^m\,\|P^{-1}\|\,\|x^{(0)} - s\| \\ &= \kappa(P)\|D\|^m\left\|x^{(0)} - s\right\| \\ &= K\|D\|^m, \quad K \equiv \kappa(P)\|x^{(0)} - s\|. \end{aligned}$$

La cantidad $\kappa(P) = \|P\|\|P^{-1}\|$ es el número de condición de la matriz P, el cual fue introducido en la sección 3.4.

Restringiéndonos a las normas matriciales usuales vistas en la sección 3.4, se verifica que $\|D\| = \rho(D) = \rho(B)$, por lo que

$$\|x^{(m)} - s\| \leqslant K\rho(B)^m. \tag{9}$$

Si $\rho(B) < 1$ (caso en el que hay convergencia), cuanto más pequeño sea $\rho(B)$, más rápido convergerá hacia cero ha sucesión $(\rho(B)^m)$ y, por (9), más rápido convergerá hacia cero la sucesión $(\|x^{(m)} - s\|)$.

Ejemplo 3.24. *Consideremos dos métodos iterativos de la forma $x^{(m+1)} = Bx^{(m)} + c$ cuyas matrices B y vectores c son los siguientes:*

$$B_1 = \begin{pmatrix} 0 & -\frac{1}{3} & 0 & \frac{1}{3} \\ -\frac{23}{20} & \frac{1}{3} & \frac{23}{20} & \frac{17}{30} \\ \frac{1}{4} & -\frac{1}{3} & -\frac{1}{4} & \frac{1}{3} \\ -\frac{23}{20} & 0 & \frac{23}{20} & \frac{9}{10} \end{pmatrix}, \quad c_1 = \begin{pmatrix} \frac{1}{3} \\ -\frac{97}{30} \\ \frac{17}{6} \\ -\frac{19}{10} \end{pmatrix}, \quad \rho(B_1) = 0.9,$$

$$B_2 = \begin{pmatrix} 0 & -\frac{1}{3} & 0 & \frac{1}{3} \\ -\frac{3}{4} & \frac{1}{3} & \frac{3}{4} & \frac{1}{6} \\ \frac{1}{4} & -\frac{1}{3} & -\frac{1}{4} & \frac{1}{3} \\ -\frac{3}{4} & 0 & \frac{3}{4} & \frac{1}{2} \end{pmatrix}, \quad c_2 = \begin{pmatrix} \frac{1}{3} \\ -\frac{5}{6} \\ \frac{17}{6} \\ \frac{1}{2} \end{pmatrix}, \quad \rho(B_2) = 0.5.$$

La solución exacta de $x = B_i x + c_i$, $i = 1, 2$, es $s = (1, 2, 3, 4)$. En la siguiente tabla se muestran los primeros elementos de las sucesiones $(x^{(m)})$ para ambos métodos tomando como $x^{(0)}$ el vector nulo:

	$\rho = 0.5$		$\rho = 0.9$	
m	$x^{(m)}$	$\|x^{(m)} - s\|_\infty$	$x^{(m)}$	$\|x^{(m)} - s\|_\infty$
0	$(0, 0, 0, 0.)$	4.	$(0, 0, 0, 0.)$	4.
1	$(0.333, -0.833, 2.83, 0.5)$	3.5	$(0.333, -3.23, 2.83, -1.90)$	5.90
2	$(0.777, 0.847, 2.65, 2.62)$	1.37	$(0.777, -2.51, 2.65, -0.735)$	4.73
3	$(0.925, 1.29, 2.95, 3.21)$	7.81×10^{-1}	$(0.925, -2.33, 2.95, -0.405)$	4.40
4	$(0.975, 1.65, 2.96, 3.63)$	3.67×10^{-1}	$(0.975, -1.90, 2.96, 0.0712)$	3.92
5	$(0.991, 1.81, 2.99, 3.81)$	1.89×10^{-1}	$(0.991, -1.53, 2.99, 0.455)$	3.54
6	$(0.997, 1.90, 2.99, 3.90)$	9.32×10^{-2}	$(0.997, -1.18, 2.99, 0.811)$	3.18
7	$(0.999, 1.95, 2.99, 3.95)$	4.70×10^{-2}	$(0.999, -0.869, 2.99, 1.13)$	2.87
8	$(0.999, 1.97, 3., 3.97)$	2.34×10^{-2}	$(0.999, -0.582, 3., 1.41)$	2.58
9	$(0.999, 1.98, 3., 3.98)$	1.17×10^{-2}	$(0.999, -0.324, 3., 1.67)$	2.32
10	$(1., 1.99, 3., 3.99)$	5.85×10^{-3}	$(1., -0.0920, 3., 1.90)$	2.09
11	$(1., 1.99, 3., 3.99)$	2.93×10^{-3}	$(1., 0.117, 3., 2.11)$	1.88
12	$(1., 1.99, 3., 3.99)$	1.46×10^{-3}	$(1., 0.305, 3., 2.30)$	1.69
13	$(1., 1.99, 3., 3.999)$	7.32×10^{-4}	$(1., 0.474, 3., 2.47)$	1.52
14	$(1., 2., 3., 4.)$	3.66×10^{-4}	$(1., 0.627, 3., 2.62)$	1.37
15	$(1., 2., 3., 4.)$	1.83×10^{-4}	$(1., 0.764, 3., 2.76)$	1.23

4.1.2. Cotas del error

Vamos a obtener acotaciones del error que se comete al sustituir la solución exacta por el valor de $x^{(m)}$.

Proposición 3.25. *Dado el método iterativo $x^{(m+1)} = Bx^{(m)} + c$ y una norma matricial $\|\cdot\|_M$ tal que $\|B\|_M < 1$, entonces se verifica:*

(i) $$\|x^{(m)} - s\|_v \leqslant \frac{\|B\|_M}{1 - \|B\|_M} \|x^{(m)} - x^{(m-1)}\|_v,$$

(ii) $$\|x^{(m)} - s\|_v \leqslant \frac{\|B\|_M^m}{1 - \|B\|_M} \|x^{(1)} - x^{(0)}\|_v,$$

donde s es la solución de la ecuación $x = Bx + c$ y $\|\cdot\|_v$ es una norma vectorial compatible con la norma matricial $\|\cdot\|_M$.

Observaciones:

(O1) La diferencia en norma entre dos aproximaciones consecutivas pueden usarse para establecer un criterio de parada en el proceso iterativo de obtención de la solución de un sistema $A\,x = b$.

$$\|x^{(m)} - x^{(m-1)}\|_v < \delta \implies \|x^{(m)} - s\|_v \leqslant \frac{\|B\|_M}{1 - \|B\|_M}\,\delta.$$

Si $0 < \delta < \dfrac{1 - \|B\|_M}{\|B\|_M}\,\varepsilon$, entonces $\|x^{(m)} - s\| < \varepsilon$.

(O2) Si $\|B\|_M$ es próximo a 1, entonces el cociente $\frac{\|B\|_M}{1-\|B\|_M}$ es mucho mayor que uno. En este caso hemos de exigir un valor muy pequeño de $\|x^{(m)} - x^{(m-1)}\|_v$ para asegurar un error pequeño.

Ejemplo 3.26. *Dado un vector inicial $x^{(0)}$, halle un número de iteraciones que garantice un error absoluto menor que ε al tomar $x^{(m)}$ como solución aproximada del sistema $Ax = b$.*

Resolución. Se trata de encontrar el menor entero m satisfaciendo

$$\frac{\|B\|^m}{1-\|B\|}\|x^{(1)} - x^{(0)}\| < \varepsilon,$$

de donde se deduce que

$$\|B\|^m < \frac{1-\|B\|}{\|x^{(1)} - x^{(0)}\|}\,\varepsilon \quad \Longrightarrow \quad m\log\|B\| < \log\left[\frac{(1-\|B\|)\varepsilon}{\|x^{(1)} - x^{(0)}\|}\right].$$

Como $\|B\| < 1$, entonces $\log\|B\| < 0$, por lo que al aislar m se modifica el sentido de la desigualdad:

$$m > \frac{\log\left[\frac{(1-\|B\|)\varepsilon}{\|x^{(1)}-x^{(0)}\|}\right]}{\log\|B\|} \equiv \alpha \quad \Longrightarrow \quad N = \min\{m \;/\; m > \alpha\} = E(\alpha) + 1.$$

La función $E(x)$ se denomina función parte entera, y se define como el único número entero que verifica $E(x) \leqslant x < E(x) + 1$. Por ejemplo, $E(2) = 2$, $E(3.71) = 3$. ◀

4.2. Métodos iterativos usuales

Dado un sistema de ecuaciones lineales $Ax = b$, la aplicación de la iteración del punto fijo requiere, en primer lugar, la obtención de una matriz B y un vector c que verifiquen la condición de consistencia $c = (I - B)A^{-1}b$. En esta sección analizaremos formas sistemáticas de construcción de métodos iterativos consistentes a partir de una matriz A arbitraria.

Así, si se conoce una descomposición de A en la forma $A = M - N$ con M inversible, podemos escribir el sistema $Ax = b$ en la forma $(M - N)x = b$, es decir, $Mx = Nx + b$, o equivalentemente $x = M^{-1}Nx + M^{-1}b$, lo que determina un método de iteración del punto fijo donde $B = M^{-1}N$ y $c = M^{-1}b$. Por la construcción efectuada ya se deduce que el método es consistente. No obstante podemos comprobar fácilmente la condición $c = (I - B)A^{-1}b$ teniendo en cuenta que

$$I - B = I - M^{-1}N = M^{-1}M - M^{-1}N = M^{-1}A,$$

lo que implica que

$$(I - B)A^{-1}b = M^{-1}AA^{-1}b = M^{-1}b = c.$$

Acabamos de demostrar el resultado siguiente:

Teorema 3.27. *Sea $A = M - N$, donde M es inversible. Si $B = M^{-1}N$ y $c = M^{-1}b$, el método iterativo $x^{(m+1)} = Bx^{(m)} + c$ es consistente para la resolución del sistema $Ax = b$.*

Según el Teorema anterior, si $A = M - N$ con M inversible, el método iterativo

$$x^{(m+1)} = M^{-1}Nx^{(m)} + M^{-1}b, \quad m = 0, 1, 2, \ldots \tag{10}$$

es consistente para la resolución del sistema $Ax = b$. Habrá convergencia si y solo si $\rho(M^{-1}N) < 1$ o, si existe una norma matricial tal que $\|M^{-1}N\| < 1$. Para hacer cálculos es más conveniente escribir el método (10) en la forma

$$Mx^{(m+1)} = Nx^{(m)} + b, \quad m = 0, 1, 2, \ldots$$

Entonces se ve que interesa tomar *M fácilmente inversible* (por ejemplo, diagonal o triangular).

Con el fin de encontrar descomposiciones del tipo $A = M - N$ con M inversible, observemos que si $A = (a_{ij})_{1 \leqslant i,j \leqslant n}$ podemos escribir $A = D - L - U$, donde las matrices D (diagonal), L (triangular inferior) y U (triangular superior) están dadas por:

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

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

4.2.1. El método de Jacobi

Una primera elección de M y N asociada a la relación $A = D - L - U$ consiste en tomar $M = D$ y $N = L + U$. Al método iterativo resultante se le denomina método de Jacobi.

Partiendo de un sistema $Ax = b$ de manera que la matriz A no tenga elementos nulos en la diagonal (pues eso garantiza que D es inversible) se llega al sistema equivalente $x = D^{-1}(L + U)x + D^{-1}b$. A partir de aquí obtenemos el proceso iterativo de Jacobi que se expresa en la forma

$$x^{(m+1)} = B_J x^{(m)} + c_J,$$

donde $B_J = D^{-1}(L + U)$ y $c_J = D^{-1}b$.

Por la naturaleza de D, L y U se deduce fácilmente la expresión de los elementos de

B_J y c_J obteniéndose que

$$B_J = \begin{pmatrix} 0 & -a_{12}/a_{11} & -a_{13}/a_{11} & \dots & -a_{1n}/a_{11} \\ -a_{21}/a_{22} & 0 & -a_{23}/a_{22} & \dots & -a_{2n}/a_{22} \\ \dots & \dots & \dots & \dots & \dots \\ -a_{n1}/a_{nn} & -a_{n2}/a_{nn} & -a_{n3}/a_{nn} & \dots & 0 \end{pmatrix}, \tag{11}$$

$$c_J = (b_1/a_{11}, b_2/a_{22}, \dots, b_n/a_{nn})^t.$$

A partir de esta expresión de B_J podemos determinar las ecuaciones componente a componente del método de Jacobi. Puesto que

$$\begin{pmatrix} x_1^{(m+1)} \\ x_2^{(m+1)} \\ \dots \\ x_n^{(m+1)} \end{pmatrix} = B_J \begin{pmatrix} x_1^{(m)} \\ x_2^{(m)} \\ \dots \\ x_n^{(m)} \end{pmatrix} + c_J,$$

obtenemos las ecuaciones

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

Estas ecuaciones permiten diseñar de forma inmediata un algoritmo para el método de Jacobi. Se puede utilizar como test de parada la diferencia en norma entre dos iteraciones consecutivas. Señalemos que la estructura del método requiere la utilización permanente de dos vectores puesto que para el cálculo de todas las componentes del vector $x^{(m+1)}$ se utilizan siempre las componentes de $x^{(m)}$ independientemente de que algunas de ellas ya tengan un valor más actualizado.

4.2.2. El método de Gauss-Seidel

Otra posible elección de M y N a partir de la relación $A = D - L - U$ consiste en tomar $M = D - L$ y $N = U$. Al método iterativo resultante se le denomina método de Gauss-Seidel.

Nota. La elección $M = D - U$, $N = L$ es igualmente válida, pero por costumbre se elige la anterior.

Partiendo de un sistema $Ax = b$ de manera que la matriz A no tenga elementos nulos en la diagonal (pues eso garantiza que $D - L$ es inversible) se llega al sistema equivalente $x = (D - L)^{-1} Ux + (D - L)^{-1} b$. A partir de aquí obtenemos el proceso iterativo de Gauss-Seidel que se expresa en la forma

$$x^{(m+1)} = B_G x^{(m)} + c_G,$$

donde $B_G = (D-L)^{-1}U$ y $c_G = (D-L)^{-1}b$.

A diferencia del método de Jacobi, la expresión explícita de la matriz del método B_G no es fácil de determinar pues requiere el cálculo de la inversa de la matriz triangular $D-L$. No obstante, las ecuaciones del método componente a componente se deducen fácilmente cuando se expresa en la forma

$$(D-L)x^{(m+1)} = Ux^{(m)} + b,$$

es decir,

$$\begin{pmatrix} a_{11} & 0 & \dots & 0 \\ a_{21} & a_{22} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{pmatrix} \begin{pmatrix} x_1^{(m+1)} \\ x_2^{(m+1)} \\ \vdots \\ x_n^{(m+1)} \end{pmatrix} = \begin{pmatrix} 0 & -a_{12} & \dots & -a_{1n} \\ 0 & 0 & \dots & -a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 0 \end{pmatrix} \begin{pmatrix} x_1^{(m)} \\ x_2^{(m)} \\ \vdots \\ x_n^{(m)} \end{pmatrix} + \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix},$$

de donde se sigue que

$$\begin{aligned} a_{11}x_1^{(m+1)} &= -a_{12}x_2^{(m)} - a_{13}x_3^{(m)} - \cdots - a_{1n}x_n^{(m)} + b_1, \\ a_{21}x_1^{(m+1)} + a_{22}x_2^{(m+1)} &= -a_{23}x_3^{(m)} - \cdots - a_{2n}x_n^{(m)} + b_2, \\ &\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots \\ a_{n1}x_1^{(m+1)} + a_{n2}x_2^{(m+1)} + \cdots + a_{nn}x_n^{(m+1)} &= b_n, \end{aligned}$$

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

Estas ecuaciones permiten diseñar de forma inmediata un algoritmo para el método de Gauss-Seidel, para el que se puede utilizar el mismo test de parada del método de Jacobi. En este caso, a diferencia del método de Jacobi, en el cálculo de $x_i^{(m+1)}$ se utilizan los valores de $x_1^{(m+1)}, x_2^{(m+1)}, \dots, x_{i-1}^{(m+1)}, x_{i+1}^{(m)}, \dots, x_n^{(m)}$ y, por tanto, es suficiente el almacenamiento en memoria de un único vector.

4.2.3. Convergencia de los métodos

Puesto que los métodos de Jacobi y Gauss-Seidel son casos particulares de los métodos de iteración del punto fijo para sistemas de ecuaciones lineales, podemos aplicar en el estudio de su convergencia los criterios generales relativos a la existencia de una norma matricial tal que $\|B_J\| < 1$ (respectivamente, $\|B_G\| < 1$) o que el radio espectral de la matriz correspondiente sea menor que 1. Además de estas condiciones, vamos a comprobar que en el caso de los métodos de Jacobi y Gauss-Seidel es posible obtener condiciones sobre la matriz A que garantizan la convergencia de los métodos sin necesidad de construir explícitamente la matriz de iteración.

Definición 3.28 (Matriz estrictamente diagonal dominante por filas). *Diremos que una matriz $A = (a_{ij}) \in \mathcal{M}_n(\mathbb{R})$ es estrictamente diagonal dominante por filas si se verifica*

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

Teorema 3.29. *Si A es una matriz estrictamente diagonal dominante por filas, entonces los métodos de Jacobi y Gauss-Seidel son convergentes cuando se aplican a la resolución de cualquier sistema de la forma $Ax = b$.*

Otras matrices para las que existen resultados sobre la convergencia son las simétricas y definidas positivas o aquellas tales que $D + L + U$ es simétrica y definida positiva.

Teorema 3.30. *Si A es una matriz tal que $D + L + U$ es simétrica y definida positiva, entonces el método de Jacobi es convergente cuando se aplica a la resolución de cualquier sistema de la forma $Ax = b$.*

Teorema 3.31. *Si A es una matriz simétrica y definida positiva, entonces el método de Gauss-Seidel es convergente cuando se aplica a la resolución de cualquier sistema de la forma $Ax = b$.*

Reseñamos a continuación algunos resultados que permiten comparar la velocidad de convergencia de los métodos de Jacobi y de Gauss-Seidel. En este sentido, debemos señalar que la medida más común para estimar la velocidad de convergencia de un método iterativo es el valor de $\rho(B)$.

Teorema 3.32. *Si A es una matriz tridiagonal se verifica que $\rho(B_G) = \rho(B_J)^2$ y, por tanto, ambos métodos son simultáneamente convergentes o divergentes y cuando ambos convergen el método de Gauss-Seidel es asintóticamente más rápido.*

Teorema 3.33 (Stein-Rosenberg). *Si A es tal que B_J es no negativa entonces se verifica alguna de las posibilidades siguientes:*

(i) $\rho(B_J) = \rho(B_G) = 0$.

(ii) $0 < \rho(B_G) < \rho(B_J) < 1$.

(iii) $\rho(B_J) = \rho(B_G) = 1$.

(iv) $1 < \rho(B_J) < \rho(B_G)$.

Así pues, para este tipo de matrices los métodos de Jacobi y de Gauss-Seidel convergen o divergen simultáneamente y en caso de convergencia lo hace más rápido el método de Gauss-Seidel. La demostración del Teorema de Stein-Rosenberg está basada en la teoría espectral de Perron-Frobenius y rebasa los objetivos de este tema.