2. ÁLGEBRA LINEAL

2.1 Definiciones

Una matriz **A** = (a_{ij}), de orden *n* x *m*, es un conjunto de números dispuestos en *n* filas y *m* columnas.

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & a_{13} & .... & a_{1m} \\ a_{21} & a_{22} & a_{23} & .... & a_{2m} \\ a_{31} & a_{32} & a_{33} & .... & a_{3m} \\ ..... & & & & \\ a_{n1} & a_{n2} & a_{n3} & .... & a_{nm} \end{pmatrix}$$

Un elemento, a_{ij}, se identifica por dos sub – índices, el primero de los cuales denota la fila y el segundo la columna. Si *m* = 1 se tiene una *matriz columna* o "vector" de dimensión *n*:

$$\mathbf{b} = \begin{Bmatrix} b_1 \\ b_2 \\ \vdots \\ b_3 \end{Bmatrix}$$

Si en cambio *n* = 1, se tiene una *matriz fila*: $\mathbf{c} = \begin{bmatrix} c_1 & c_2 & \dots & c_m \end{bmatrix}$. Si *n* = *m* se dice que la matriz es cuadrada (de orden *n*). Por ejemplo:

$$\mathbf{A} = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 1 & 4 & 9 & 16 \\ 1 & 8 & 27 & 64 \\ 1 & 16 & 81 & 256 \end{pmatrix} \qquad \mathbf{D} = \begin{pmatrix} d_1 & 0 & 0 & 0 \\ 0 & d_2 & 0 & 0 \\ 0 & 0 & d_3 & 0 \\ 0 & 0 & 0 & d_4 \end{pmatrix} \qquad \mathbf{I_n} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

A, **D** e $\mathbf{I_n}$ son matrices cuadradas. La matriz $\mathbf{D} = diag\begin{bmatrix} d_1 & d_2 & \dots & d_n \end{bmatrix}$ es una *matriz diagonal*, cuyos elementos son todos cero, excepto aquellos ubicados en la diagonal principal (de la esquina superior izquierda a la inferior derecha). Un caso particular es el de $\mathbf{I_n} = diag\begin{bmatrix} 1 & 1 & \dots & 1 \end{bmatrix} = \left(\delta_{ij}\right)$, que es una *matriz unidad* (o identidad) de orden *n*. La matriz identidad tiene en el álgebra matricial un papel similar al uno en álgebra común. Por otro lado, el equivalente del cero es una matriz nula (no necesariamente cuadrada), cuyos elementos son todos ceros.

Las matrices cuadradas cuyos elementos tienen simetría conjugada: $a_{ij} = a_{ji}^{\ *}$ (donde * indica conjugada compleja) se denominan *Hermitianas*. Por ejemplo:

$$\mathbf{H} = \begin{pmatrix} 1 & 2+i & 3-2i & 0 \\ 2-i & 5 & 1-i & 1+i \\ 3+2i & 1+i & 3 & 2-3i \\ 0 & 1-i & 2+3i & 4 \end{pmatrix} \qquad i = \sqrt{-1}$$

es una matriz Hermitiana. Si todos los elementos de una matriz Hermitiana son reales, es decir $a_{ij} = a_{ji}$, se tiene una *matriz simétrica*.

Una matriz cuadrada en la que la mayor parte de los elementos son ceros y los elementos con valor significativo están agrupados alrededor de la diagonal principal se denomina *matriz banda*. Por ejemplo:

$$\mathbf{B}=\begin{pmatrix} 1 & -1 & & & \\ -1 & 2 & -1 & & \\ & -1 & 2 & -1 & \\ & & -1 & 2 & -1 \\ & & & -1 & 1 \end{pmatrix}$$

Las líneas paralelas a la diagonal principal se llaman *codiagonales*. El número total de diagonal y codiagonales con elementos significativos en el *ancho de banda* (3 en este ejemplo). Para matrices simétricas puede también hablarse de un *ancho de semi – banda*; que incluye a la diagonal principal (2 en el ejemplo precedente). Una matriz banda tiene baja *densidad*. Por densidad se entiende la razón entre el número de elementos con valor significativo y el número total de elementos.

Si en una matriz cuadrada todos los elementos por encima (o por debajo) de la diagonal principal son cero se dice que ésta es una *matriz triangular* inferior (superior):

$$\mathbf{L}=\begin{pmatrix} \ell_{11} & 0 & 0 & \cdots & 0 \\ \ell_{21} & \ell_{22} & 0 & \cdots & 0 \\ \ell_{31} & \ell_{32} & \ell_{33} & \cdots & 0 \\ \cdots & & & & \\ \ell_{n1} & \ell_{n2} & \ell_{n3} & \cdots & \ell_{nm} \end{pmatrix} \qquad \mathbf{U}=\begin{pmatrix} u_{11} & u_{12} & u_{13} & \cdots & u_{1n} \\ 0 & u_{22} & u_{23} & \cdots & u_{2n} \\ 0 & 0 & u_{33} & \cdots & u_{3n} \\ \cdots & & & & \\ 0 & 0 & 0 & \cdots & u_{nm} \end{pmatrix}$$

En lo que sigue se usan letras **negritas** para denotar matrices. Para las matrices columna y para las matrices filas se usan minúsculas, mientras que para las matrices rectangulares (incluyendo las matrices cuadradas) se usan mayúsculas. En todos los casos, los elementos de una matriz se indican en minúsculas.

2.2 Operaciones Básicas con Matrices

Subdivisión o partición. El conjunto de elementos de una matriz $\mathbf{A}$ puede ser dividido en otros más pequeños mediante líneas horizontales y/o verticales. Las distintas partes, $\mathbf{A}_{11}$, $\mathbf{A}_{12}$, etc. son *submatrices* de la matriz $\mathbf{A}$. Las submatrices pueden tratarse como elementos comunes de una matriz, excepto que deben operarse según las reglas del álgebra matricial.

Igualdad. Dos matrices, $\mathbf{A}$, $\mathbf{B}$, del mismo orden, son iguales si cada elemento de una es igual al correspondiente elemento de la otra. $\mathbf{A} = \mathbf{B}$ implica $a_{ij} = b_{ij}$ para todo i, j.

Suma (resta). La suma (o diferencia) de dos matrices $\mathbf{A}$, $\mathbf{B}$ del mismo orden es una tercera matriz del mismo orden, cuyos elementos se obtienen sumando (restando) algebraicamente los correspondientes elementos de las dos matrices originales:

$$\mathbf{A} \pm \mathbf{B} = \mathbf{C} \qquad a_{ij} \pm b_{ij} = c_{ij}$$

La suma (resta) de matrices es asociativa y conmutativa:

$$(\mathbf{A} + \mathbf{B}) + \mathbf{C} = \mathbf{A} + (\mathbf{B} + \mathbf{C}) \qquad \mathbf{A} + \mathbf{B} = \mathbf{B} + \mathbf{A}$$

Derivada e integral. Análogamente, puede definirse la derivada de una matriz:

$$\frac{\partial \mathbf{A}}{\partial \alpha} = \mathbf{B} \quad \Rightarrow \quad \frac{\partial a_{ij}}{\partial \alpha} = b_{ij}$$

y la integral de una matriz en forma similar.

Multiplicación por un escalar. El producto de una matriz por un escalar es otra matriz del mismo orden cuyos elementos son los de la matriz original multiplicados por el escalar:

$$\alpha \mathbf{A} = \mathbf{B} \quad \Rightarrow \quad \alpha\, a_{ij} = b_{ij}$$

Multiplicación de dos matrices. Dos matrices, $\mathbf{A}$ (m x p) y $\mathbf{B}$ (p x n) pueden ser multiplicadas en el orden $\mathbf{A}\,\mathbf{B}$ sólo si son *conformables* para el producto, es decir, si el número de columnas de $\mathbf{A}$ es igual al número de filas de $\mathbf{B}$. El producto $\mathbf{C}$ (m x n) es una matriz cuyos elementos se obtienen de:

$$c_{ij} = \sum_{k=1}^{p} a_{ik} \cdot b_{kj} \qquad i = 1, m \quad j = 1, n$$

Por ejemplo, si:

$$\mathbf{A} = \begin{pmatrix} 5 & 3 & 1 \\ 4 & 6 & 2 \\ 10 & 3 & 4 \end{pmatrix} \qquad \mathbf{B} = \begin{pmatrix} 1 & 5 \\ 2 & 4 \\ 3 & 2 \end{pmatrix} \qquad \mathbf{C} = \mathbf{A} \cdot \mathbf{B}$$

$$\begin{aligned} c_{11} &= 5 \cdot 1 + 3 \cdot 2 + 1 \cdot 3 = 14 \\ c_{21} &= 4 \cdot 1 + 6 \cdot 2 + 2 \cdot 3 = 22 \\ &\dots \\ c_{32} &= 10 \cdot 5 + 3 \cdot 4 + 4 \cdot 2 = 70 \end{aligned} \quad \Rightarrow \quad \mathbf{C} = \begin{pmatrix} 14 & 39 \\ 22 & 48 \\ 28 & 70 \end{pmatrix}$$

La multiplicación de matrices es asociativa y distributiva, pero en general no es conmutativa:

$$\mathbf{A}(\mathbf{B} \cdot \mathbf{C}) = (\mathbf{A} \cdot \mathbf{B})\mathbf{C} \qquad \mathbf{A}(\mathbf{B} + \mathbf{C}) = \mathbf{A}\mathbf{B} + \mathbf{A}\mathbf{C} \qquad \mathbf{A}\mathbf{B} \neq \mathbf{B}\mathbf{A}$$

Siendo el orden de multiplicación importante, es frecuente enfatizarlo, diciendo por ejemplo que en el producto $\mathbf{AB}$ la matriz $\mathbf{A}$ premultiplica a $\mathbf{B}$, o bien que $\mathbf{B}$ postmultiplica a $\mathbf{A}$. En algunos casos $\mathbf{AB} = \mathbf{BA}$; se dice entonces que $\mathbf{A}$ y $\mathbf{B}$ son conmutables.

Es fácil verificar que el producto de dos matrices triangulares inferiores (superiores) es otra matriz triangular inferior (superior).

Transposición. La transpuesta $\mathbf{A}^T$ de una matriz $\mathbf{A}$ es aquella cuyas filas son las columnas de $\mathbf{A}$ (y viceversa). Si $\mathbf{A}^T = \mathbf{B} = (b_{ij})$, entonces $b_{ij} = a_{ji}$:

$$\mathbf{A} = \begin{pmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{pmatrix} \qquad \mathbf{A}^T = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{pmatrix}$$

La transpuesta de una matriz simétrica es obviamente la matriz original. Productos del tipo $\mathbf{A}^T\mathbf{A}$ resultan siempre en matrices simétricas. Lo mismo puede decirse de productos $\mathbf{A}^T\mathbf{S}\mathbf{A}$ si $\mathbf{S}$ es simétrica.

Cuando se transpone un producto matricial la secuencia de los factores debe invertirse:

$$(\mathbf{AB}\ldots\mathbf{C})^T = \mathbf{C}^T \ldots \mathbf{B}^T \mathbf{A}^T$$

Determinante de una matriz cuadrada. Es un número que resulta de:

$$\det \mathbf{A} = |\mathbf{A}| = \sum_{n!} \pm a_{1i} a_{2j} a_{3k} \ldots a_{nr}$$

Donde cada término de la suma incluye un solo elemento de cada fila y de cada columna. Si en estos productos se considera a los elementos ordenados por filas 1, 2, .. *n*, los índices de las columnas en cada término de la suma pueden ser obtenidos como permutación del orden normal. Según el número de cambios requeridos para esta permutación sea par o impar se asigna al producto correspondiente el signo **+** o **-**. La suma incluye las *n*! permutaciones posibles.

Las siguientes propiedades facilitan el cómputo de la determinante de una matriz cuadrada A cualquiera:

- Si se intercambian dos filas (columnas) la determinante cambia de signo.
- La determinante de una matriz, $|\mathbf{A}|$, es igual a la determinante de su transpuesta.
- El valor de la determinante de una matriz $\mathbf{A}$ no se altera si una columna (fila) multiplicada por un escalar se suma algebraicamente a otra columna (fila):

$$\det\begin{pmatrix} a & b \\ c & d \end{pmatrix} = \det\begin{pmatrix} a & b \\ 0 & d - \dfrac{bc}{a} \end{pmatrix} = ad - bc$$

- En consecuencia, la determinante de una matriz con dos filas (o columnas) iguales (o proporcionales) es cero. Más aún,si dos o más columnas (filas) de una matriz $\mathbf{A}$ son linealmente dependientes, es decir $\alpha_1\mathbf{a}_1 + \alpha_2\mathbf{a}_2 + \alpha_3\mathbf{a}_3 + \ldots + \alpha_{n-1}\mathbf{a}_{n-1} + \alpha_n\mathbf{a}_n = 0$ para un conjunto de coeficientes α_i de los que por lo menos uno es distinto de cero, la determinante es cero. Se dice entonces que la matriz $\mathbf{A}$ es singular. Considérese por ejemplo el caso:

$$\mathbf{A} = \left(\begin{array}{c|c|c} 1 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 1 & 1 \end{array}\right)$$

$\mathbf{A}$ es singular puesto que: $(1)\begin{Bmatrix}1\\1\\0\end{Bmatrix} + (-1)\begin{Bmatrix}1\\2\\1\end{Bmatrix} + (1)\begin{Bmatrix}0\\1\\1\end{Bmatrix} = \begin{Bmatrix}0\\0\\0\end{Bmatrix}$

- La determinante de una matriz triangular es igual al producto de los elementos de su diagonal principal.
- Para un producto matricial se cumple que:

$$\det(\mathbf{A} \cdot \mathbf{B} \ldots \mathbf{C}) = \det(\mathbf{A}) \cdot \det(\mathbf{B}) \ldots \det(\mathbf{C})$$

Así, por ejemplo, si:

$$\mathbf{A} = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 1 & 4 & 9 & 16 \\ 1 & 8 & 27 & 64 \\ 1 & 16 & 81 & 256 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 3 & 1 & 0 \\ 1 & 7 & 6 & 1 \end{pmatrix} \begin{pmatrix} 1 & 2 & 3 & 4 \\ 0 & 2 & 6 & 12 \\ 0 & 0 & 6 & 24 \\ 0 & 0 & 0 & 24 \end{pmatrix}$$

entonces: $\det(\mathbf{A}) = (1) \cdot (1 \cdot 2 \cdot 6 \cdot 24) = 288$

Inversa de una matriz. Si una matriz $\mathbf{A}$ es no singular, es posible obtener su "inversa", $\mathbf{A}^{-1}$, que satisface:

$$\mathbf{A}^{-1}\mathbf{A} = \mathbf{A}\mathbf{A}^{-1} = \mathbf{I}_n \qquad \left(\mathbf{A}^{-1}\right)^{-1} = \mathbf{A}$$

Obviamente $\mathbf{I}_n{}^{-1} = \mathbf{I}_n$. La inversa de una matriz diagonal es otra matriz diagonal, cuyos elementos son inversas de los elementos de la matriz original. La inversa de una matriz triangular (inferior o superior) es otra matriz triangular del mismo tipo.

La inversión de matrices permite efectuar la operación equivalente a la división del álgebra común.

$$\mathbf{AB} = \mathbf{C} \quad \Rightarrow \quad \mathbf{B} = \mathbf{A}^{-1}\mathbf{C}$$ (véanse los comentarios del ítem 2.5.5)

Para la inversa de un producto matricial se cumple:

$$\left(\mathbf{AB}\ldots\mathbf{C}\right)^{-1} = \mathbf{C}^{-1}\ldots\mathbf{B}^{-1}\mathbf{A}^{-1}$$

Una matriz $\mathbf{Q}$ se denomina ortogonal si: $\mathbf{Q}\mathbf{Q}^T = \mathbf{I}_n$. Particularmente, si $\mathbf{Q}$ es una matriz cuadrada se tiene entonces que $\mathbf{Q}^{-1} = \mathbf{Q}^T$. Por ejemplo:

$$\mathbf{R} = \begin{pmatrix} \cos\theta & -\operatorname{sen}\theta \\ \operatorname{sen}\theta & \cos\theta \end{pmatrix}$$

es ortogonal, puesto que:

$$\mathbf{R}^{-1} = \begin{pmatrix} \cos\theta & \operatorname{sen}\theta \\ -\operatorname{sen}\theta & \cos\theta \end{pmatrix} = \mathbf{R}^T$$

.

Refiriéndose a una matriz con coeficientes complejos, $\mathbf{U}$, se dice que ésta es unitaria si $\mathbf{U}\,\mathbf{U}^* = \mathbf{I}$

2.3 Espacios y Subespacios Vectoriales

Una matriz columna de orden n es un conjunto números que pueden ser interpretados como componentes de un vector en un espacio de dimensión n.

Se dice que un conjunto de vectores $\mathbf{v}_1$ $\mathbf{v}_2$ $\mathbf{v}_3$ .... $\mathbf{v}_5$ son linealmente dependientes si existen números α_1 α_2 α_3 .... α_5, no todos cero, tales que:

$$\alpha_1\mathbf{v}_1 + \alpha_2\mathbf{v}_2 + \alpha_3\mathbf{v}_3 + \ldots + \alpha_5\mathbf{v}_5 = 0$$

Alternativamente, puede decirse que los vectores son linealmente dependientes si uno cualquiera de ellos puede expresarse como combinación lineal de los otros: $\mathbf{v}_r = \sum_{i\neq r} c_i\mathbf{v}_i$ (y linealmente independientes si esto no es posible).

p vectores linealmente independientes de orden n ($n \geq p$) conforman una base de un espacio vectorial de dimensión p. Por otro lado, q vectores, de los que p ($p \leq q$) son linealmente independientes, están contenidos en un espacio de dimensión p.

Si los vectores linealmente independientes $\mathbf{x}_1$ $\mathbf{x}_2$ .... $\mathbf{x}_p$ constituyen una base de un espacio vectorial de dimensión p, un sub – conjunto de estos puede considerarse como base de un sub – espacio contenido en el espacio vectorial original.

Las columnas (o filas) de una matriz rectangular $\mathbf{A}$ pueden tratarse como vectores. El número de vectores linealmente independientes define el "rango" de la matriz. Una matriz cuadrada es no singular si su rango es igual al orden de la matriz, es decir si todas las columnas son linealmente independientes. Lo contrario implica que una o más columnas (filas) pueden obtenerse como combinación lineal de las otras y la determinante es cero.

2.4 Sistemas de Ecuaciones Lineales

Se ha estimado que un 75% de los problemas de ingeniería se presenta, en alguna etapa del trabajo, la solución de un sistema de ecuaciones lineales:

$$\begin{aligned}
&a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + \ldots + a_{1n}x_n = b_1 \\
&a_{21}x_1 + a_{22}x_2 + a_{23}x_3 + \ldots + a_{2n}x_n = b_2 \\
&a_{31}x_1 + a_{32}x_2 + a_{33}x_3 + \ldots + a_{3n}x_n = b_3 \\
&\ldots \\
&a_{n1}x_1 + a_{n2}x_2 + a_{n3}x_3 + \ldots + a_{nn}x_n = b_n
\end{aligned} \qquad (2.1a)$$

o bien: $\mathbf{Ax} = \mathbf{b}$

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} & \ldots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \ldots & a_{2n} \\ a_{31} & a_{32} & a_{33} & \ldots & a_{3n} \\ \ldots & & & & \\ a_{n1} & a_{n2} & a_{n3} & \ldots & a_{nn} \end{pmatrix} \begin{Bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{Bmatrix} = \begin{Bmatrix} b_1 \\ b_2 \\ b_3 \\ \vdots \\ b_n \end{Bmatrix} \qquad (2.1b)$$

En las secciones siguientes se supone que el sistema de ecuaciones tiene solución única, es decir, que $\det(\mathbf{A}) \neq 0$.

La solución de sistemas de ecuaciones es un buen ejemplo de las diferencias entre las matemáticas "clásicas" y los métodos numéricos modernos. Así, la *Regla de Cramer*:

$$x_j = \frac{\det \begin{pmatrix} a_{11} & a_{12} & \ldots & b_1 & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & b_2 & \ldots & a_{2n} \\ a_{31} & a_{32} & \ldots & b_3 & \ldots & a_{3n} \\ \ldots & & & & & \\ a_{n1} & a_{n2} & \ldots & b_n & \ldots & a_{nn} \end{pmatrix}}{\det \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1j} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2j} & \ldots & a_{2n} \\ a_{31} & a_{32} & \ldots & a_{3j} & \ldots & a_{3n} \\ \ldots & & & & & \\ a_{n1} & a_{n2} & \ldots & a_{nj} & \ldots & a_{nn} \end{pmatrix}} \qquad (2.2)$$

si bien proporciona fórmulas explícitas es tremendamente ineficiente cuando se trata de resolver sistemas con más de 3 incógnitas (excepto para casos muy especiales de la matriz de coeficientes).

Muchos métodos frecuentemente utilizados en ingeniería, como por ejemplo los métodos de elementos finitos para la solución de ecuaciones en derivadas parciales, resultan en

el planteamiento de grandes sistemas de ecuaciones lineales. El costo de análisis y en muchos casos la factibilidad de un modelo suficientemente preciso dependen en gran medida de la forma de almacenamiento de las ecuaciones y de la eficiencia del algoritmo utilizado en su solución.

2.5 Métodos Directos para la Solución de Sistemas de Ecuaciones Lineales

Este acápite considera métodos que, de no haber errores de redondeo, producen la solución exacta en un número finito de pasos. Para sistemas $\mathbf{Ax} = \mathbf{b}$, en los que $\mathbf{A}$ es de alta densidad, los métodos directos son en general los más eficientes (para las computadoras actualmente utilizadas). Sin embargo, cuando un gran número de elementos de $\mathbf{A}$ son cero, y en especial cuando $\mathbf{A}$ es definida positiva ($\mathbf{x}^T\mathbf{Ax} > 0$ para cualquier $\mathbf{x} \neq 0$), puede ser más conveniente utilizar un método iterativo en que se obtiene una secuencia de soluciones aproximadas que convergen a la solución exacta.

2.5.1. Sistemas Triangulares

La solución de sistemas de ecuaciones lineales es particularmente simple cuando la matriz de coeficientes es triangular. Por ejemplo, considérese un sistema $\mathbf{Ux} = \mathbf{b}$ en el que $\mathbf{U}$ es triangular superior:

$$\begin{aligned} u_{11}x_1 + u_{12}x_2 + u_{13}x_3 + \ldots + u_{1n}x_n &= b_1 \\ u_{22}x_2 + u_{23}x_3 + \ldots + u_{2n}x_n &= b_2 \\ u_{33}x_3 + \ldots + u_{3n}x_n &= b_3 \\ \ldots\ \ldots & \\ u_{n-1,n-1}x_{n-1} + u_{nn}x_n &= b_n \\ u_{nn}x_n &= b_n \end{aligned} \qquad (2.3)$$

Si $\mathbf{U}$ es no singular ($u_{ii} \neq 0$ para todo i), las incógnitas pueden evaluarse en el orden: n, n-1, n-2, n-3, ... 2, 1:

$$x_n = \frac{b_n}{u_{nn}} \qquad (2.4a)$$

$$x_i = \frac{1}{u_{ii}}\left(b_i - \sum_{k=i+1}^{n} u_{ik}x_k\right) \qquad (2.4b)$$

Este proceso se denomina “sustitución inversa”. Análogamente, para un sistema $\mathbf{Lx} = \mathbf{b}$, en el que $\mathbf{L}$ es una matriz triangular inferior no singular ($l_{ii} \neq 0$ para todo i), puede utilizarse una sustitución directa o “reducción”:

$$x_1 = \frac{b_1}{l_{11}} \qquad (2.5a)$$

$$x_i = \frac{1}{l_{ii}}\left(b_i - \sum_{k=1}^{i-1} l_{ik}x_k\right) \qquad (2.5b)$$

En ambos casos, la solución del sistema requiere n divisiones y $\frac{1}{2}n(n-1)$ operaciones de multiplicación y suma (casi lo mismo que para multiplicar una matriz triangular por un vector).

2.5.2 Método de Gauss

Éste es el más importante de los métodos directos para la solución de sistemas de ecuaciones lineales. La idea básica está en combinar las distintas ecuaciones para ir eliminando incógnitas en forma sistemática y obtener finalmente un sistema triangular, fácil de resolver. Considérese el sistema de orden n:

$$\begin{aligned}
&a_{11}^{(1)} x_1 + a_{12}^{(1)} x_2 + a_{13}^{(1)} x_3 + \ldots + a_{1n}^{(1)} x_n = b_1^{(1)} \\
&a_{21}^{(1)} x_1 + a_{22}^{(1)} x_2 + a_{23}^{(1)} x_3 + \ldots + a_{2n}^{(1)} x_n = b_2^{(1)} \\
&a_{31}^{(1)} x_1 + a_{32}^{(1)} x_2 + a_{33}^{(1)} x_3 + \ldots + a_{3n}^{(1)} x_n = b_3^{(1)} \\
&\ldots\ldots \\
&a_{n1}^{(1)} x_1 + a_{n2}^{(1)} x_2 + a_{n3}^{(1)} x_3 + \ldots + a_{nn}^{(1)} x_n = b_n^{(1)}
\end{aligned} \qquad (2.6)$$

o en forma compacta: $\mathbf{Ax} = \mathbf{b}$. En lo que sigue se supone que $\mathbf{A}$ es no singular. Supóngase también que $a_{11} \neq 0$. Puede entonces eliminarse x_1 de la ecuación i si de ésta se resta la ecuación 1 multiplicada por:

$$l_{i1} = \frac{a_{i1}^{(1)}}{a_{11}^{(1)}} \qquad (2.7a)$$

Con ello se obtiene:

$$\begin{aligned}
a_{11}^{(1)} x_1 + a_{12}^{(1)} x_2 + a_{13}^{(1)} x_3 + \ldots + a_{1n}^{(1)} x_n &= b_1^{(1)} \\
a_{22}^{(2)} x_2 + a_{23}^{(2)} x_3 + \ldots + a_{2n}^{(2)} x_n &= b_2^{(2)} \\
a_{32}^{(2)} x_2 + a_{33}^{(2)} x_3 + \ldots + a_{3n}^{(2)} x_n &= b_3^{(2)} \\
\ldots\ldots & \\
a_{n2}^{(2)} x_2 + a_{n3}^{(2)} x_3 + \ldots + a_{nn}^{(2)} x_n &= b_n^{(2)}
\end{aligned} \qquad (2.7b)$$

donde

$$\begin{aligned}
a_{ij}^{(2)} &= a_{ij}^{(1)} - l_{i1}\, a_{1j}^{(1)} \\
b_i^{(2)} &= b_i^{(1)} - l_{i1}\, b_1^{(1)}
\end{aligned} \qquad (2.7c)$$

En forma similar, puede eliminarse x_2 de las ecuaciones $i = 3,4,..n$ restando de la ecuación i la ecuación 2 multiplicada por:

$$l_{i2} = \frac{a_{i2}^{(2)}}{a_{22}^{(2)}}$$

y así sucesivamente hasta obtener el sistema triangular:

$$\begin{aligned}
a_{11}^{(1)} x_1 + a_{12}^{(1)} x_2 + a_{13}^{(1)} x_3 + \ldots + a_{1n}^{(1)} x_n &= b_1^{(1)} \\
a_{22}^{(2)} x_2 + a_{23}^{(2)} x_3 + \ldots + a_{2n}^{(2)} x_n &= b_2^{(2)} \\
a_{33}^{(3)} x_3 + \ldots + a_{3n}^{(3)} x_n &= b_3^{(3)} \\
\ldots\ldots & \\
a_{nn}^{(n)} x_n &= b_n^{(n)}
\end{aligned} \qquad (2.8)$$

o en notación matricial: $\mathbf{Ux} = \mathbf{b}$.

Los elementos $a_{11}^{(1)}, a_{22}^{(2)}, a_{33}^{(3)} \ldots a_{n-1,n-1}^{(n-1)}$ que se usan como divisores en esta reducción se llaman “pivotes”. El proceso – tal como ha sido planteado hasta el momento – falla si alguno de estos es cero. Esto en general no ocurre si la matriz $\mathbf{A}$ tiene diagonal

dominante (es decir, si $|a_{ii}| > \sum_{j \neq i} |a_{ij}|$) o si $\mathbf{A}$ es simétrica ($\mathbf{A}^T = \mathbf{A}$) y definida positiva ($\mathbf{v}^T\mathbf{A}\mathbf{v} > 0$ para $\mathbf{v}$ arbitrario).

El siguiente ejemplo ilustra el proceso:

$$\begin{array}{ll} & \begin{pmatrix} 1 & 2 & 3 & 4 \\ 1 & 4 & 9 & 16 \\ 1 & 8 & 27 & 64 \\ 1 & 16 & 81 & 256 \end{pmatrix} \begin{Bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{Bmatrix} = \begin{Bmatrix} 2 \\ 10 \\ 44 \\ 190 \end{Bmatrix} \\ (1) & \\ (1) & \\ (1) & \end{array}$$

Los números indicados a la izquierda (entre paréntesis) son los factores l_{i1} por los que es necesario multiplicar la ecuación 1 antes de restarla de la ecuación i, para lograr el objetivo de eliminar x_1 de la segunda y las siguientes ecuaciones.

$$\begin{array}{ll} & \begin{pmatrix} 1 & 2 & 3 & 4 \\ 0 & 2 & 6 & 12 \\ 0 & 6 & 24 & 60 \\ 0 & 14 & 78 & 252 \end{pmatrix} \begin{Bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{Bmatrix} = \begin{Bmatrix} 2 \\ 8 \\ 42 \\ 188 \end{Bmatrix} \\ (3) & \\ (7) & \end{array}$$

Análogamente:

$$\begin{array}{ll} & \begin{pmatrix} 1 & 2 & 3 & 4 \\ 0 & 2 & 6 & 12 \\ 0 & 0 & 6 & 24 \\ 0 & 0 & 36 & 168 \end{pmatrix} \begin{Bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{Bmatrix} = \begin{Bmatrix} 2 \\ 8 \\ 18 \\ 132 \end{Bmatrix} \\ (6) & \end{array}$$

$$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 0 & 2 & 6 & 12 \\ 0 & 0 & 6 & 24 \\ 0 & 0 & 0 & 24 \end{pmatrix} \begin{Bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{Bmatrix} = \begin{Bmatrix} 2 \\ 8 \\ 18 \\ 24 \end{Bmatrix}$$

finalmente:

$$\begin{array}{ll} 24\,x_4 = 24 & x_4 = 1 \\ 6\,x_3 + 24\,x_4 = 18 & x_3 = -1 \\ 2\,x_2 + 6\,x_3 + 12\,x_4 = 8 & x_2 = 1 \\ x_1 + 2\,x_2 + 3\,x_3 + 4\,x_4 = 2 & x_1 = -1 \end{array}$$

Para estimar el esfuerzo de cómputo es habitual referirse al número de "operaciones" requeridas. La costumbre es contar como una operación a la combinación de una suma (o resta, o simplemente una copia) con una multiplicación (o división). Esta práctica proviene de las épocas en que el tiempo requerido para efectuar una multiplicación o una división era un orden de magnitud mayor que el necesario para una suma o una resta, pudiendo despreciarse estas últimas. La reducción de la matriz de coeficientes requiere de un número de operaciones de orden $\frac{1}{3}n^3$. La reducción del segundo miembro y la sustitución inversa requieren aproximadamente n^2 operaciones. Si se tuvieran varios sistemas de ecuaciones con la misma matriz de coeficientes: $\mathbf{Ax} = \mathbf{b}1$, $\mathbf{Ay} = \mathbf{b}2$, ... sólo se requeriría efectuar la reducción de $\mathbf{A}$ una vez, por lo que el número de operaciones sería siempre aproximadamente $\frac{1}{3}n^3$. Más precisamente, se hacen $\frac{1}{3}n^3 + 2n^2 + \frac{2}{3}n$

operaciones para resolver un sistema de n ecuaciones lineales, pero si n es grande sólo el primer término es importante.

El proceso antes descrito falla cuando se presenta un pivote, $a_{ii}^{(i)}$, igual a cero. Un ejemplo simple de tal situación es el siguiente:

$$\begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 2 \\ 1 & 2 & 2 \end{pmatrix} \begin{Bmatrix} x_1 \\ x_2 \\ x_3 \end{Bmatrix} = \begin{Bmatrix} 1 \\ 2 \\ 1 \end{Bmatrix}$$

La matriz de coeficientes no es singular y el sistema tiene una solución única $\mathbf{x} = \begin{pmatrix} 1 & -1 & 1 \end{pmatrix}^T$. Sin embargo, después del primer paso (efectuado en el orden indicado anteriormente), se obtiene:

$$\begin{pmatrix} 1 & 1 & 1 \\ 0 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix} \begin{Bmatrix} x_1 \\ x_2 \\ x_3 \end{Bmatrix} = \begin{Bmatrix} 1 \\ 1 \\ 0 \end{Bmatrix}$$

y siendo $a_{22}^{(2)} = 0$, no es posible proseguir como habitualmente. La solución es en este caso obvia: intercambiar las ecuaciones (filas) 2 y 3. En general, si $a_{ii}^{(i)} = 0$, algún otro elemento de la misma columna, $a_{ji}^{(i)}$, debe ser distinto de cero (lo contrario implicaría una dependencia lineal de por lo menos dos de las ecuaciones, es decir la singularidad de $\mathbf{A}$). Intercambiando las filas j e i puede entonces continuarse la reducción. Dados los elementos $a_{ji}^{(i)}$ de la columna i, es conveniente escoger como pivote aquel de máximo valor absoluto, puesto que el uso de pivotes pequeños introduce fuertes errores en la solución. El ejemplo siguiente es ilustrativo:

$$\begin{pmatrix} 3\times10^{-11} & 1 \\ 1 & 1 \end{pmatrix} \begin{Bmatrix} x_1 \\ x_2 \end{Bmatrix} = \begin{Bmatrix} 7 \\ 9 \end{Bmatrix}$$

Trabajando con 10 cifras significativas se obtiene:

$$\begin{pmatrix} 3.000\,000\,000\times10^{-11} & 1 \\ 0 & -3.333333333\times10^{10} \end{pmatrix} \begin{Bmatrix} x_1 \\ x_2 \end{Bmatrix} = \begin{Bmatrix} 7 \\ 7\left(-3.333333333\times10^{10}\right) \end{Bmatrix}$$

de donde: $x_2 = 7$

$x_1 = 0$

La solución correcta es, sin embargo, $x_1 = 2$. Es fácil comprobar que no se presenta este problema si se evita el pivote pequeño intercambiando previamente las ecuaciones:

$$\begin{pmatrix} 1 & 1 \\ 3\times10^{-11} & 1 \end{pmatrix} \begin{Bmatrix} x_1 \\ x_2 \end{Bmatrix} = \begin{Bmatrix} 9 \\ 7 \end{Bmatrix}$$

El intercambio de filas al que se ha hecho referencia se denomina “intercambio parcial”. Alternativamente, puede pensarse en un “intercambio completo”, en que se selecciona el siguiente pivote como el elemento de máximo valor absoluto entre todos los elementos de la sub matriz por reducirse. Se intercambian entonces filas (ecuaciones) y columnas (incógnitas) para continuar el proceso como se ha descrito.

El intercambio parcial es generalmente satisfactorio, desde el punto de vista de la estabilidad numérica, y requiere bastante menos trabajo que el proceso con intercambio total.

2.5.3 Descomposición $\mathbf{A = LU}$

Supóngase que $\mathbf{A}$ es tal que el proceso de reducción del método de Gauss puede efectuarse sin necesidad de intercambiar filas o columnas. En tal caso, la descomposición $\mathbf{A} = \mathbf{LU}$ donde $\mathbf{L}$ es una matriz triangular inferior con $l_{ii} = 1$ y $\mathbf{U}$ es una matriz triangular superior, es única. Esto puede probarse fácilmente por inducción. Para el caso del primer ejemplo:

$$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 1 & 4 & 9 & 16 \\ 1 & 8 & 27 & 64 \\ 1 & 16 & 81 & 256 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 3 & 1 & 0 \\ 1 & 7 & 6 & 1 \end{pmatrix} \begin{pmatrix} 1 & 2 & 3 & 4 \\ 0 & 2 & 6 & 12 \\ 0 & 0 & 6 & 24 \\ 0 & 0 & 0 & 24 \end{pmatrix}$$

Los elementos de $\mathbf{L}$ son justamente los coeficientes l_{ij} usados durante la reducción; $\mathbf{U}$ es en cambio ¡la matriz $\mathbf{A}$ reducida!

Se ha mencionado anteriormente que varios sistemas de ecuaciones con la misma matriz de coeficientes pueden ser resueltos simultáneamente. Sin embargo, no siempre se conocen desde un principio todos los vectores de coeficientes del segundo miembro. Por ejemplo, puede querer resolverse $\mathbf{Ax}_1 = \mathbf{b}$ y $\mathbf{Ax}_2 = \mathbf{x}_1$. Aún en este caso, al resolver el segundo sistema no es necesario volver a reducir la matriz $\mathbf{A}$ como al inicio. El sistema $\mathbf{Ax} = \mathbf{b}$ es equivalente a $\mathbf{LUx} = \mathbf{b}$, o bien a los dos sistemas triangulares: $\mathbf{Ly} = \mathbf{b}$, $\mathbf{Ux} = \mathbf{y}$. Siendo $\mathbf{L}$ y $\mathbf{U}$ conocidos, estos dos sistemas pueden resolverse en O(n^2) operaciones. $\mathbf{L}$ y $\mathbf{U}$ pueden almacenarse en las mismas posiciones de memoria que en la matriz $\mathbf{A}$: Como $l_{ki} = a_{ki}^{(i)} \big/ a_{ii}^{(i)}$ se determina con el objeto de hacer $a_{ki}^{(i+1)} = 0$, l_{ki} puede almacenarse en las posición de a_{ki}. Por otro lado, no es necesario almacenar los elementos de la diagonal de $\mathbf{L}$ (que son todos iguales a 1). Dado que los elementos de $\mathbf{U}$ son aquellos de la matriz reducida, el efecto de la reducción o descomposición en la distribución de memoria es de la forma:

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} & \dots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \dots & a_{2n} \\ a_{31} & a_{32} & a_{33} & \dots & a_{3n} \\ \dots & & & & \\ a_{n1} & a_{n2} & a_{n3} & \dots & a_{nn} \end{pmatrix} \Rightarrow \begin{pmatrix} u_{11} & u_{12} & u_{13} & \dots & u_{1n} \\ l_{21} & u_{22} & u_{23} & \dots & u_{2n} \\ l_{31} & l_{32} & u_{33} & \dots & u_{3n} \\ \dots & & & & \\ l_{n1} & l_{n2} & l_{n3} & \dots & u_{nn} \end{pmatrix}$$

Para el ejemplo precedente:

$$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 1 & 4 & 9 & 16 \\ 1 & 8 & 27 & 64 \\ 1 & 16 & 81 & 256 \end{pmatrix} \Rightarrow \begin{pmatrix} 1 & 2 & 3 & 4 \\ 1 & 2 & 6 & 12 \\ 1 & 3 & 6 & 24 \\ 1 & 7 & 6 & 24 \end{pmatrix}$$

En los casos en los que se efectúan intercambios de filas y/o columnas es siempre posible (si $\mathbf{A}$ no es singular) obtener factores triangulares $\mathbf{L}$ y $\mathbf{U}$ tales que $\mathbf{LU} = \mathbf{A}$',

donde $\mathbf{A}$' es la matriz que resulta de efectuar los intercambios mencionados en la matriz original $\mathbf{A}$.

2.5.4 Otros Métodos Directos

Todos los métodos tratados en esta sección pueden considerarse como variantes del método de Gauss.

Una posible alternativa es la de calcular los elementos de $\mathbf{L}$ y $\mathbf{U}$ mediante las fórmulas:

$$u_{kj} = a_{kj} - \sum_{p=1}^{k-1} l_{kp}\, u_{pj} \qquad j = k, k+1, \cdots n \tag{2.9a}$$

$$l_{ik} = \frac{1}{u_{kk}} \left(a_{ik} - \sum_{p=1}^{k-1} l_{ip}\, u_{pk} \right) \qquad i = k+1, \cdots n \tag{2.9b}$$

en lugar de efectuar "reducciones" como anteriormente. Esta modificación (*Doolitle*) es conveniente cuando se usan calculadoras manuales, ya que evita la escritura de muchos resultados intermedios. Su uso en computadoras es ventajoso si las operaciones se hacen con una precisión mayor que aquella con la que se almacenan los resultados.

El método de *Crout* efectúa la factorización $\mathbf{A} = \mathbf{LDR}$, donde $\mathbf{L}$ es la misma matriz triangular inferior obtenida durante el proceso de Gauss, $\mathbf{D}$ es una matriz diagonal y $\mathbf{R}$ es una matriz triangular superior con coeficientes 1 en su diagonal principal. $\mathbf{D}$ y $\mathbf{R}$ están relacionados con la $\mathbf{U}$ de Gauss.

$$\begin{aligned} d_{ii} &= u_{ii} \\ r_{ij} &= \frac{u_{ij}}{d_{ii}} \qquad j > i \end{aligned} \tag{2.10}$$

En particular, para $\mathbf{A}$ simétrica: $\mathbf{R} = \mathbf{L}^T$. Este método no posee ventajas ni desventajas con relación al de Gauss, bien sea en cuanto a estabilidad numérica y precisión, como en el número de operaciones necesarias.

Si durante el proceso de reducción se usa la ecuación i para eliminar x_i, no sólo de las ecuaciones que siguen a la i sino también de la ecuaciones precedentes, se tiene el método de *Gauss – Jordan*. Para el ejemplo antes considerado:

$$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 1 & 4 & 9 & 16 \\ 1 & 8 & 27 & 64 \\ 1 & 16 & 81 & 256 \end{pmatrix} \begin{Bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{Bmatrix} = \begin{Bmatrix} 2 \\ 10 \\ 44 \\ 190 \end{Bmatrix}$$

$$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 0 & 2 & 6 & 12 \\ 0 & 6 & 24 & 60 \\ 0 & 14 & 78 & 252 \end{pmatrix} \begin{Bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{Bmatrix} = \begin{Bmatrix} 2 \\ 8 \\ 42 \\ 188 \end{Bmatrix}$$

$$\begin{pmatrix} 1 & 0 & -3 & -8 \\ 0 & 2 & 6 & 12 \\ 0 & 0 & 6 & 24 \\ 0 & 0 & 36 & 168 \end{pmatrix} \begin{Bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{Bmatrix} = \begin{Bmatrix} -6 \\ 8 \\ 18 \\ 132 \end{Bmatrix}$$

Nótese que se utilizó la segunda ecuación para reducir no solamente las ecuaciones 3 y 4, sino también la ecuación 1. Análogamente:

$$\begin{pmatrix} 1 & 0 & 0 & 4 \\ 0 & 2 & 0 & -12 \\ 0 & 0 & 6 & 24 \\ 0 & 0 & 0 & 24 \end{pmatrix} \begin{Bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{Bmatrix} = \begin{Bmatrix} 3 \\ -10 \\ 18 \\ 24 \end{Bmatrix}$$

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 6 & 0 \\ 0 & 0 & 0 & 24 \end{pmatrix} \begin{Bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{Bmatrix} = \begin{Bmatrix} -1 \\ 2 \\ -6 \\ 24 \end{Bmatrix}$$ de donde se obtiene fácilmente la solución.

El método de Gauss- Jordan es más simple de programar, pero requiere casi 1.5 veces el número de operaciones del método de Gauss tradicional.

Finalmente, para concluir esta sección, debe mencionarse que el método de Gauss es aplicable también a sistemas de ecuaciones con coeficientes complejos. Por ejemplo:

$$\begin{pmatrix} 2 & 1-i & 0 \\ 1+i & 2 & 1+i \\ 0 & 1-i & 3 \end{pmatrix} \begin{Bmatrix} x_1 \\ x_2 \\ x_3 \end{Bmatrix} = \begin{Bmatrix} 4-2i \\ 8+4i \\ 11-2i \end{Bmatrix}$$

$$\begin{pmatrix} 2 & 1-i & 0 \\ 0 & 1 & 1+i \\ 0 & 1-i & 3 \end{pmatrix} \begin{Bmatrix} x_1 \\ x_2 \\ x_3 \end{Bmatrix} = \begin{Bmatrix} 4-2i \\ 5+3i \\ 11-2i \end{Bmatrix}$$

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

de donde:

$$x_3 = 3$$
$$x_2 = (5+3i) - 3(1+i) = 2$$
$$x_1 = \tfrac{1}{2}\left[(4-2i) - 2(1-i)\right] = 1$$

2.5.5 Inversión de Matrices

Si la inversa, $\mathbf{A}^{-1}$, de una matriz $\mathbf{A}$ se conoce, la solución de un sistema $\mathbf{Ax} = \mathbf{b}$ puede escribirse $\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$. Podría entonces parecer conveniente determinar $\mathbf{A}^{-1}$, en especial si se tienen varios sistemas de ecuaciones con la misma matriz de coeficientes. Sin embargo, la solución puede ser obtenida con mucho menos operaciones – y en general con mucha más precisión – utilizando la descomposición $\mathbf{A} = \mathbf{LU}$. La solución de los dos sistemas triangulares $\mathbf{Ly} = \mathbf{b}$ y $\mathbf{Ux} = \mathbf{y}$ requiere sólo O(n^2) operaciones (por cada columna de $\mathbf{b}$ ó $\mathbf{x}$). Por otro lado, la multiplicación $\mathbf{A}^{-1}\mathbf{b}$ también demanda O(n^2)

operaciones. Sin embargo, la determinación de $\mathbf{A}^{-1}$ requiere aproximadamente el triple de trabajo que para obtener $\mathbf{L}$ y $\mathbf{U}$. El número de operaciones necesarias para obtener la inversa de una matriz cuadrada (no simétrica) de orden n es $n^3 + 2n^2 - n + 1$.

No obstante esto, en algunos casos se necesita la inversa en forma explícita. La inversa puede obtenerse de un modo eficiente resolviendo n sistemas de ecuaciones lineales: $\mathbf{AX} = \mathbf{I}_n$, donde $\mathbf{X} = \mathbf{A}^{-1}$. El siguiente ejemplo utiliza una variante del método de Gauss con este objeto:

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

En la columna de la izquierda se tienen la matriz $\mathbf{A}$ y sus sucesivas modificaciones. A la derecha se presentan la matriz $\mathbf{I}$ y las modificaciones obtenidas efectuando sobre las filas las mismas operaciones que en $\mathbf{A}$:

$$\begin{pmatrix} 1 & 1 & 1 \\ 2 & 1 & 3 \\ 3 & 1 & 4 \end{pmatrix} \qquad \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

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

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

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

Alternativamente, si la descomposición $\mathbf{A} = \mathbf{LU}$ de una matriz $\mathbf{A}$ se conoce, la inversa puede obtenerse de $\mathbf{A}^{-1} = \mathbf{U}^{-1}\mathbf{L}^{-1}$, también en O($n^2$) operaciones. Si en los cómputos para $\mathbf{L}$ y $\mathbf{U}$ se hacen intercambios de filas, el producto $\mathbf{U}^{-1}\mathbf{L}^{-1}$ resulta la inversa de una cierta matriz $\mathbf{A}$'. La matriz $\mathbf{A}^{-1}$ puede obtenerse a partir de $(\mathbf{A}')^{-1}$ intercambiando columnas en secuencia inversa a los cambios de fila durante el proceso.

Para la matriz antes considerada:

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

$$\mathbf{A} = \mathbf{LU}$$

La inversa de una matriz triangular es otra matriz del mismo tipo, fácil de determinar.

Para una matriz triangular inferior, $\mathbf{L}$, cada columna de la matriz inversa $\mathbf{L}^{-1}$ puede ser obtenida por sustitución directa o "reducción": $\mathbf{LY} = \mathbf{I}_n$.

$$y_{ij} = 0 \qquad i < j \qquad \text{(2.11a)}$$

$$y_{ij} = \frac{1}{l_{ii}}\left(\delta_{ij} - \sum_{k=j}^{i-1} l_{ik} y_{kj}\right) \qquad i \geq j \tag{2.11b}$$

En forma análoga, la inversa, $\mathbf{U}^{-1}$, de una matriz triangular superior, $\mathbf{U}$, es también una matriz triangular superior. Cada fila i, puede determinarse mediante $\mathbf{UZ} = \mathbf{I}_n$:

$$z_{ij} = \frac{1}{u_{jj}}\left(\delta_{ij} - \sum_{k=i}^{j-1} z_{ik} u_{kj}\right) \qquad i \leq j \tag{2.12a}$$

$$z_{ij} = 0 \qquad i > j \tag{2.12b}$$

Para las matrices $\mathbf{L}$ y $\mathbf{U}$ del ejemplo considerado:

$$\mathbf{L}^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ 1 & -2 & 1 \end{pmatrix} \qquad \mathbf{U}^{-1} = \begin{pmatrix} 1 & 1 & 2 \\ 0 & -1 & -1 \\ 0 & 0 & -1 \end{pmatrix}$$

$$\mathbf{A}^{-1} = \mathbf{U}^{-1}\mathbf{L}^{-1} = \begin{pmatrix} 1 & -3 & 2 \\ 1 & 1 & -1 \\ -1 & 2 & -1 \end{pmatrix}$$

2.5.6 Casos Especiales

Matrices Simétricas Definidas Positivas.

Para una matriz simétrica: $a_{jk}^{(1)} = a_{kj}^{(1)}$. Si se efectúa la reducción de Gauss sin intercambio de filas y/o columnas se tiene también que: $a_{jk}^{(i)} = a_{kj}^{(i)}$ para $i < j$, $k \leq n$. En otras palabras, la sub – matriz que debe aún reducirse en un paso dado es también simétrica. Esto puede probarse por inducción, teniendo en cuenta las condiciones iniciales de simetría y además que:

$$a_{kj}^{(i+1)} = a_{kj}^{(i)} - l_{ki} a_{ij}^{(i)} = a_{kj}^{(i)} - \frac{a_{ki}^{(i)}}{a_{ii}^{(i)}} a_{ij}^{(i)} \tag{2.13a}$$

$$a_{jk}^{(i+1)} = a_{jk}^{(i)} - l_{ji} a_{ik}^{(i)} = a_{jk}^{(i)} - \frac{a_{ji}^{(i)}}{a_{ii}^{(i)}} a_{ik}^{(i)} \tag{2.13b}$$

Puede observarse que, si los coeficientes en la etapa i son simétricos, aquellos en la etapa $i+1$ también lo son, puesto que se obtienen operando del mismo modo con números iguales.

Considérese, por ejemplo, el sistema de ecuaciones con coeficientes simétricos:

$$\begin{pmatrix} 5 & -4 & 1 & 0 \\ -4 & 6 & -4 & 1 \\ 1 & -4 & 6 & -4 \\ 0 & 1 & -4 & 5 \end{pmatrix} \begin{Bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{Bmatrix} = \begin{Bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{Bmatrix}$$

En las sucesivas etapas del proceso de eliminación, las sub matrices que quedan por reducir siguen siendo simétricas:

$$\begin{pmatrix} 5 & -4 & 1 & 0 \\ 0 & \frac{14}{5} & -\frac{16}{5} & 1 \\ 0 & -\frac{16}{5} & \frac{29}{5} & -4 \\ 0 & 1 & -4 & 5 \end{pmatrix} \begin{Bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{Bmatrix} = \begin{Bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{Bmatrix}$$

$$\begin{pmatrix} 5 & -4 & 1 & 0 \\ 0 & \frac{14}{5} & -\frac{16}{5} & 1 \\ 0 & 0 & \frac{15}{7} & -\frac{20}{7} \\ 0 & 0 & -\frac{20}{7} & \frac{65}{14} \end{pmatrix} \begin{Bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{Bmatrix} = \begin{Bmatrix} 0 \\ 1 \\ \frac{8}{7} \\ -\frac{5}{14} \end{Bmatrix}$$

$$\begin{pmatrix} 5 & -4 & 1 & 0 \\ 0 & \frac{14}{5} & -\frac{16}{5} & 1 \\ 0 & 0 & \frac{15}{7} & -\frac{20}{7} \\ 0 & 0 & 0 & \frac{5}{6} \end{pmatrix} \begin{Bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{Bmatrix} = \begin{Bmatrix} 0 \\ 1 \\ \frac{8}{7} \\ \frac{7}{6} \end{Bmatrix} \quad \text{de donde} \quad \begin{Bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{Bmatrix} = \frac{1}{5} \begin{Bmatrix} 8 \\ 13 \\ 12 \\ 7 \end{Bmatrix}$$

La simetría de la matriz por reducirse permite hacer: $l_{ki} = a_{ik}^{(i)} \big/ a_{ii}^{(i)}$ (utilizando $a_{ik}^{(i)}$ en lugar de $a_{ki}^{(i)}$) y restringir los cálculos de: $a_{kj}^{(i+1)} = a_{kj}^{(i)} - l_{ki} a_{ij}^{(i)}$ a las columnas $k \le j \le n$, en lugar de $i \le j \le n$. El número de operaciones para la reducción es entonces $O(\frac{1}{6} n^2)$, aproximadamente la mitad que para el caso general.

También los requerimientos de memoria pueden reducirse, almacenando los coeficientes de la matriz en un arreglo monodimensional. Para el caso de una matriz simétrica de alta densidad el siguiente esquema de numeración de los coeficientes es apropiado:

$$\begin{pmatrix} 1 & 2 & 4 & 7 & 11 & \vdots & \vdots \\ & 3 & 5 & 8 & 12 & \vdots & \vdots \\ & & 6 & 9 & 13 & \vdots & \vdots \\ & & & 10 & 14 & \vdots & \vdots \\ & & & & 15 & \vdots & \vdots \\ & & & & & \vdots & \vdots \\ & & & & & & \frac{1}{2} n (n+1) \end{pmatrix}$$

Es evidente que intercambios de filas y columnas destruyen la simetría, a menos que se tome siempre como pivote un elemento de la diagonal principal. Tales intercambios no son necesarios si la matriz es definida positiva ($\mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x} > 0$ para $\mathbf{x}$ arbitraria, no nula), ya que en tal caso:

$$\begin{aligned} & a_{ii}^{(k)} > 0 && i \ge 1,\ k \le n \\ & \left| a_{ij}^{(k)} \right|^2 \le a_{ii}^{(k)} a_{jj}^{(k)} && k \le i,\ j \le n \\ & a_{ii}^{(k+1)} \le 2 a_{ii}^{(k)} && k < i \le n \end{aligned} \qquad (2.14)$$

Estas condiciones garantizan que no se presentan pivotes pequeños.

Para el caso de matrices simétricas definidas positivas puede también utilizarse el método de *Cholesky*. Éste método efectúa la descomposición $\mathbf{A} = \mathbf{R}^{\mathrm{T}}\mathbf{R}$, donde $\mathbf{R}$ es una matriz triangular superior cuyos elementos pueden obtenerse (por filas) de:

$$r_{ii} = \left(a_{ii} - \sum_{p=1}^{i-1} r_{pi}^2 \right)^{\frac{1}{2}} \tag{2.15a}$$

$$r_{ij} = \frac{1}{r_{ii}} \left(a_{ij} - \sum_{p=1}^{i-1} r_{pi} r_{pj} \right) \qquad j = i+1, i+2, \cdots \tag{2.15b}$$

Para el ejemplo anterior se obtiene:

$$r_{11} = (a_{11})^{\frac{1}{2}} = 2.2360$$

$$r_{12} = a_{12} / r_{11} = -1.7888$$

$$r_{13} = a_{13} / r_{11} = 0.44721$$

$$r_{14} = a_{14} / r_{11} = 0$$

$$r_{22} = (a_{22} - r_{12}^2)^{\frac{1}{2}} = 1.6733$$

$$r_{23} = (a_{23} - r_{12} r_{13}) / r_{22} = -1.9123$$

$$r_{24} = (a_{24} - r_{12} r_{14}) / r_{22} = 0.5976$$

$$r_{33} = (a_{33} - r_{13}^2 - r_{23}^2)^{\frac{1}{2}} = 1.4639$$

$$r_{34} = (a_{34} - r_{13} r_{14} - r_{23} r_{24}) / r_{33} = -1.9518$$

$$r_{44} = (a_{44} - r_{14}^2 - r_{24}^2 - r_{34}^2)^{\frac{1}{2}} = 0.9129$$

es decir:

$$\mathbf{R} = \begin{pmatrix} 2.2360 & -1.7888 & 0.4472 & 0 \\ 0 & 1.6733 & -1.9123 & 0.5976 \\ 0 & 0 & 1.4639 & -1.9518 \\ 0 & 0 & 0 & 0.9129 \end{pmatrix}$$

El sistema $\mathbf{Ax} = \mathbf{b}$ puede entonces rescribirse como $\mathbf{R}^T\mathbf{Rx} = \mathbf{b}$ o bien $\mathbf{R}^T\mathbf{y} = \mathbf{b}$; $\mathbf{Rx} = \mathbf{y}$

Resolviendo el primer sistema triangular:

$$\mathbf{y} = \begin{Bmatrix} 0 \\ 0.5976 \\ 0.7808 \\ 1.2781 \end{Bmatrix}$$

y finalmente:

$$\mathbf{x} = \frac{1}{5} \begin{Bmatrix} 8 \\ 13 \\ 12 \\ 7 \end{Bmatrix}$$

Puede anotarse que $\mathbf{R}$ está relacionada con las $\mathbf{L}$ y $\mathbf{U}$ de Gauss mediante $\mathbf{R}^T = \mathbf{LD}$; $\mathbf{R} = \mathbf{D}^{-1}\mathbf{U}$; donde $\mathbf{D} = \text{diag}\left(\sqrt{u_{11}} \quad \sqrt{u_{22}} \quad \cdots \quad \sqrt{u_{nn}}\right)$.

Matrices Banda.

Los sistemas de ecuaciones en que los coeficientes forman matrices banda son frecuentes. Tales sistemas se resuelven eficientemente por el método de Gauss y otros similares, ya que éstos conservan la estructura de banda de las matrices: $\mathbf{A} = \mathbf{LU}$:

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

Nótese que $\mathbf{A}^{-1} = \mathbf{U}^{-1}\mathbf{L}^{-1}$ no es una matriz banda:

$$\mathbf{A}^{-1} = \begin{pmatrix} 5 & 4 & 3 & 2 & 1 \\ 4 & 4 & 3 & 2 & 1 \\ 3 & 3 & 3 & 2 & 1 \\ 2 & 2 & 2 & 2 & 1 \\ 1 & 1 & 1 & 1 & 1 \end{pmatrix}$$

y por lo tanto no conviene hallar $\mathbf{A}^{-1}$ en forma explícita.

Particularmente simples de tratar son los sistemas con matrices banda simétricas y definidas positivas (no se requieren intercambios de filas y/o columnas). Dos posibles esquemas para almacenar los coeficientes en un arreglo monodimensional son en este caso:

$$\mathbf{A} = \begin{pmatrix} 1 & 8 & 15 & & & & \\ \cdot & 2 & 9 & 16 & & & \\ \cdot & \cdot & 3 & 10 & 17 & & \\ & \cdot & \cdot & 4 & 11 & 18 & \\ & & \cdot & \cdot & 5 & 12 & 19 \\ & & & \cdot & \cdot & 6 & 13 \\ & & & & \cdot & \cdot & 7 \end{pmatrix} \begin{matrix} \\ \\ \\ \\ \\ (20) \\ (14) \end{matrix} \quad (21)$$

Las posiciones tales como 14, 20 y 21 no se usan, pero se requieren para tener un número fijo de coeficientes en cada *codiagonal*, lo que facilita la programación. Siendo el ancho de la *semibanda*, *m*, mucho menor que el número de ecuaciones, *n*, las posiciones de memoria "perdidas" son despreciables. Este esquema de almacenamiento (y su variante por filas) es apropiado cuando el ancho de banda es aproximadamente constante.

Otra posibilidad es:

$$\mathbf{A} = \begin{pmatrix} 1 & 2 & 4 & & & & \\ \cdot & 3 & 5 & 7 & & 13 & \\ \cdot & \cdot & 6 & 8 & 10 & 14 & \\ & \cdot & \cdot & 9 & 11 & 15 & \\ & & \cdot & \cdot & 12 & 16 & 18 \\ & \cdot & \cdot & \cdot & \cdot & 17 & 19 \\ & & & & \cdot & \cdot & 20 \end{pmatrix}$$

Esta forma de almacenamiento es más eficiente cuando el ancho de banda es variable (como ocurre en la mayor parte de los problemas reales). Se guardan los coeficientes por columnas, desde el "perfil" superior a la diagonal principal. Se requiere un arreglo de apuntadores o índices que indican las posiciones ocupadas por los coeficientes de la diagonal principal (que son los más referidos durante el proceso de solución). Nótese que al aplicar el método de Gauss (o cualquier variante de ese procedimiento) no se producen valores significativos por encima del perfil original y por lo tanto no se requiere más memoria que aquella en la situación inicial.

Se necesitan nm posiciones de memoria (donde n es el orden de la matriz y $m << n$ el ancho de semibanda), mucho menos que las n^2 posiciones para la matriz completa o las $\frac{1}{2}n(n+1)$ para una matriz simétrica de alta densidad. Por otro lado, la reducción de la matriz de coeficientes demanda sólo $O\left(\frac{1}{2}nm^2\right)$ operaciones, ya que:

$$l_{ij} = 0 \quad \text{excepto para} \quad j \le i < j + m$$

$$u_{ij} = 0 \quad \text{excepto para} \quad i \le j < i + m$$

Esto debe compararse con $O\left(\frac{1}{6}n^3\right)$ operaciones para reducir una matriz simétrica de alta densidad. La reducción del segundo miembro y la sustitución inversa requieren $O(2nm)$, en lugar de $O\left(n^2\right)$ operaciones. En la práctica, rara vez se tiene un ancho de banda constante, pero aún así estos estimadores son útiles, si se considera m como la media cuadrática de los anchos de semibanda en las ecuaciones.

Un caso especial es aquel en que la matriz de coeficientes es "tridiagonal":

$$\begin{pmatrix} a_1 & b_1 & & & & & \\ b_1 & a_2 & b_2 & & & & \\ & b_2 & a_3 & b_3 & & & \\ & & & \ddots & & & \\ & & & b_{n-2} & a_{n-1} & b_{n-1} & \\ & & & & b_{n-1} & a_n & \end{pmatrix} \begin{Bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_{n-1} \\ x_n \end{Bmatrix} = \begin{Bmatrix} c_1 \\ c_2 \\ c_3 \\ \vdots \\ c_{n-1} \\ c_n \end{Bmatrix} \tag{2.16}$$

Los únicos coeficientes significativos son aquellos de la diagonal principal y de dos *codiagonales*, es decir, dos líneas paralelas a la referida diagonal.

Se observa que al descomponer la matriz de coeficientes, $\mathbf{A}$, en sus factores triangulares $\mathbf{LU}$ los factores mantienen la estructura banda:

$$\begin{pmatrix} a_1 & b_1 & & & & & \\ b_1 & a_2 & b_2 & & & & \\ & b_2 & a_3 & b_3 & & & \\ & & & \ddots & & & \\ & & & b_{n-2} & a_{n-1} & b_{n-1} & \\ & & & & b_{n-1} & a_n & \end{pmatrix} = \begin{pmatrix} 1 & & & & & & \\ q_1 & 1 & & & & & \\ & q_2 & 1 & & & & \\ & & \ddots & \ddots & & & \\ & & & q_{n-2} & 1 & & \\ & & & & q_{n-1} & 1 & \end{pmatrix} \begin{pmatrix} r_1 & b_1 & & & & & \\ & r_2 & b_2 & & & & \\ & & r_3 & b_3 & & & \\ & & & r_4 & \ddots & & \\ & & & & \ddots & b_{n-1} & \\ & & & & & r_n & \end{pmatrix}$$

La determinación de los q_i y r_i es muy simple:

$$
\begin{aligned}
& r_1 = a_1 \\
& q_i = b_i / a_i \qquad i = 1,2,\cdots n-1 \\
& r_{i+1} = a_{i+1} - q_i\, b_i
\end{aligned}
\tag{2.17a}
$$

y, considerando $\mathbf{L}\,\mathbf{y} = \mathbf{c}$:

$$
\begin{aligned}
& y_1 = c_1 \\
& y_{i+1} = c_{i+1} - q_i\, y_i \qquad i = 1,2,\cdots n-1
\end{aligned}
\tag{2.17b}
$$

de donde se obtiene $\mathbf{x}$ resolviendo $\mathbf{U}\,\mathbf{x} = \mathbf{y}$:

$$
\begin{aligned}
& x_n = y_n / r_n \\
& x_i = (\, y_i - b_i\, x_{i+1}\,) / r_i \qquad i = n-1,\cdots 2,1
\end{aligned}
$$

Para resolver un sistema de *n* ecuaciones lineales con matriz de coeficientes tridiagonal se requieren sólo $5n-4$ operaciones. Como se indicó anteriormente, se cuenta como una operación la combinación de una multiplicación o división con una suma, resta o almacenamiento del resultado.

Grandes sistemas de ecuaciones lineales

(con matrices de coeficientes banda, simétricas y definidas positivas).

Cuando la memoria de la computadora es insuficiente para almacenar todos los coeficientes del sistema de ecuaciones, se recurre al disco. El acceso a este medio es (en términos relativos) muy lento y en lo posible debe tratar de minimizarse su uso.

Es frecuente subdividir la información de sistemas de ecuaciones excesivamente grandes en "bloques" de una o más ecuaciones (o columnas).

Los datos de cada bloque se almacenan en disco. Éstos son leídos a la memoria principal conforme van siendo utilizados y regrabados en la memoria auxiliar una vez operados. La solución del sistema de ecuaciones por el método de Gauss (u otro similar) requiere mantener en memoria principal la información de por lo menos dos bloques en forma simultánea. Así por ejemplo, durante el proceso de reducción, las ecuaciones del bloque *k* deben ser utilizadas para reducir ecuaciones del mismo bloque y de los bloques sucesivos *k*+1, *k*+2, ..., *k*+*n* (*n* en general es pequeña), lo que implica que, estando el bloque *k* en memoria, los bloques sucesivos deben ser leídos, parcialmente reducidos, y regrabados en secuencia. Algo similar ocurre con el proceso de sustitución inversa.

2.6. Errores en la Solución de Sistemas de Ecuaciones Lineales

En la solución práctica de grandes sistemas de ecuaciones lineales se realizan millones de operaciones y en cada una ocurren errores de redondeo, ¿Cómo afectan estos errores a los resultados? ¿Cómo puede estimarse la magnitud del error en la solución?

Podría pensarse que, habiendo resuelto el sistema $\mathbf{A}\,\mathbf{x} = \mathbf{b}$, la magnitud del residuo $\mathbf{r} = \mathbf{b} - \mathbf{A}\,\mathbf{x}$ sea una buena medida del error introducido en $\mathbf{x}$. !Esto es falso! Considérese por ejemplo:

$$
\mathbf{A} = \begin{pmatrix} 0.780 & 0.563 \\ 0.913 & 0.659 \end{pmatrix} \qquad \mathbf{b} = \begin{Bmatrix} 0.217 \\ 0.254 \end{Bmatrix}
$$

Y supóngase que se ha resuelto $\mathbf{A}\,\mathbf{x} = \mathbf{b}$ obteniendo $\mathbf{x}_1 = (0.341 \quad -0.087)^T$ ¿Qué tan buena es esta solución?

$$\mathbf{r}_1 = \mathbf{b} - \mathbf{A}\,\mathbf{x}_1 = (10^{-6} \quad 0)^T$$

Por otro lado si se afirma que la solución es $\mathbf{x}_2 = (0.999 \quad -1.001)^T$ se obtiene el residuo.

$$\mathbf{r}_2 = \mathbf{b} - \mathbf{A}\;\mathbf{x}_2 = (1.343\text{x}10^{-3} \quad 1.572\text{x}10^{-3})^T$$

¿Es $\mathbf{x}_1$ mejor solución que $\mathbf{x}_2$? No. La solución exacta es $\mathbf{x} = (1 \quad -1)^T$.

Aunque la magnitud del vector residuo $\mathbf{r} = \mathbf{b} - \mathbf{A}\,\mathbf{x}$ no da una indicación directa del error en $\mathbf{x}$, es posible utilizar residuos para estimar el error e incluso para corregir la solución. Esto se discute más adelante.

2.6.1 Normas de Vectores y Matrices

Con el propósito de discutir los errores al resolver sistemas de ecuaciones lineales, se define como *norma* (o medida) de un vector:

$$\left\| \mathbf{x} \right\|_p = (x_1^p + x_2^p + \ldots)^{1/p} \qquad 1 \le p \le \infty \tag{2.18a}$$

Dos casos particulares son de interés:

$$\left\| \mathbf{x} \right\|_2 = (x_1^2 + x_2^2 + \ldots)^{1/2} \qquad \text{(norma Euclidiana)} \tag{2.18b}$$

$$\left\| \mathbf{x} \right\|_\infty = máx\left| x_i \right| \qquad \text{(máximo valor absoluto)} \tag{2.18c}$$

Es relativamente fácil probar que:

$$\left\| \mathbf{x} \right\| \ge 0 \qquad \text{sólo hay igualdad si } \mathbf{x} = \mathbf{0}$$

$$\left\| a\,\mathbf{x} \right\| = \left| a \right| \left\| \mathbf{x} \right\| \tag{2.19}$$

$$\left\| \mathbf{x} + \mathbf{y} \right\| \le \left\| \mathbf{x} \right\| + \left\| \mathbf{y} \right\|$$

Estas propiedades son familiares en relación a la norma Euclidiana o “longitud” de un vector.

La norma de una matriz cuadrada, $\mathbf{A}$, puede ser definida en forma consistente con la definición de norma de un vector:

$$\left\| \mathbf{A} \right\|_p = máx \frac{\left\| \mathbf{A}\,\mathbf{x} \right\|_p}{\left\| \mathbf{x} \right\|_p} \qquad (\mathbf{x} \ne \mathbf{0}) \tag{2.20a}$$

La norma $\left\| \mathbf{A} \right\|_2$ es $\overline{\lambda}_{\text{máx}}^{1/2}$, donde $\overline{\lambda}_{\text{máx}}$ es el máximo valor característico de $\mathbf{A}^T\mathbf{A}$ (ver capítulo 3). Por otro lado:

$$\left\| \mathbf{A} \right\|_\infty = \underset{i}{máx} \sum_{j=1}^{n} \left| a_{ij} \right| \tag{2.20b}$$

Estas normas satisfacen condiciones similares a las normas de vectores. Además:

$$\left\| \mathbf{A}\,\mathbf{B} \right\| \le \left\| \mathbf{A} \right\| \left\| \mathbf{B} \right\| \tag{2.21}$$

2.6.2 Condicionamiento de una matriz:

En esta ecuación se analizan los efectos de una pequeña perturbación $\delta\mathbf{A}$ en la matriz $\mathbf{A}$, o de una perturbación $\delta\mathbf{b}$ en $\mathbf{b}$.

Si $\mathbf{x}$ es la solución exacta de $\mathbf{A}\,\mathbf{x} = \mathbf{b}$, cuando se considera la matriz de coeficientes $\mathbf{A} + \delta\mathbf{A}$ la solución resulta $\mathbf{x} + \delta\mathbf{x}$:

$$(\mathbf{A} + \delta\mathbf{A})\,(\mathbf{x} + \delta\mathbf{x}) = \mathbf{b} \tag{2.22}$$

de donde:

$$\delta\mathbf{x} = -\mathbf{A}^{-1}\delta\mathbf{A}\,(\mathbf{x} + \delta\mathbf{x})$$

tomando normas:

$$\| \delta\mathbf{x} \| \le \| \mathbf{A}^{-1} \| \| \delta\mathbf{A} \| \| \mathbf{x} + \delta\mathbf{x} \|$$

y dividiendo entre $\| \mathbf{x} + \delta\mathbf{x} \|$:

$$\frac{\| \delta\mathbf{x} \|}{\| \mathbf{x} + \delta\mathbf{x} \|} \le \mathrm{K}(\mathbf{A}) \frac{\| \delta\mathbf{A} \|}{\| \mathbf{A} \|} \tag{2.23}$$

donde $K(\mathbf{A}) = \| \mathbf{A} \| \| \mathbf{A}^{-1} \|$ (2.24)

es el *número de condicionamiento* de la matriz $\mathbf{A}$. Dado que $\| \mathbf{A}^{-1} \|_2 = \overline{\lambda}_{mín}^{-1/2}$, donde $\overline{\lambda}_{min}$ es el menor valor característico de la matriz $\mathbf{A}^{\mathrm{T}}\mathbf{A}$, puede escribirse:

$$K_2(\mathbf{A}) = \left(\overline{\lambda}_{máx} / \overline{\lambda}_{mín} \right)^{1/2} \tag{2.25}$$

Por otro lado: para una perturbación $\delta\mathbf{b}$ en $\mathbf{b}$:

$$\mathbf{A}\,(\mathbf{x} + \delta\mathbf{x}) = \mathbf{b} + \delta\mathbf{b} \tag{2.26}$$

de donde:

$$\delta\mathbf{x} = \mathbf{A}^{-1}\delta\mathbf{b}$$

$$\| \delta\mathbf{x} \| \le \| \mathbf{A}^{-1} \| \| \delta\mathbf{b} \|$$

y dado que $\mathbf{b} = \mathbf{A}\,\mathbf{x}$, lo que implica $\| \mathbf{x} \| \ge \dfrac{\| \mathbf{b} \|}{\| \mathbf{A} \|}$

se obtiene:

$$\frac{\| \delta\mathbf{x} \|}{\| \mathbf{x} \|} \le K(\mathbf{A}) \frac{\| \delta\mathbf{b} \|}{\| \mathbf{b} \|} \tag{2.27}$$

Las ecuaciones (2.23) y (2.27) indican que, si $K(\mathbf{A})$ es grande, pequeños cambios en $\mathbf{A}$ o en $\mathbf{b}$ pueden originar cambios importantes en la solución.

Si se tienen errores relativos de orden $\in$ tanto en $\mathbf{A}$ como en $\mathbf{b}$, (2.23) y (2.27) pueden combinarse, para escribir:

$$\| \delta\mathbf{x} \| \le 2 \in K(\mathbf{A}) \| \mathbf{x} \| \tag{2.28}$$

Los errores de redondeo introducidos en el proceso de solución pueden ser considerados como equivalentes a perturbaciones en las matrices **A** y **b** iniciales. $K(\mathbf{A})$ es también un buen indicador de los efectos de los errores de redondeo en la solución.

La expresión (3) implica que si **A** y **b** están dadas con *t* cifras significativas, el número de cifras que puede esperarse sean correctas en la solución, *s*, puede estimarse mediante:

$$s \geq t - \log_{10}[K(\mathbf{A})] \tag{2.29}$$

Para el ejemplo precedente: $\| \mathbf{A} \|_\infty = 0.913 + 0.659 = 1.572$

además: $\mathbf{A}^{-1} = 10^6 \begin{pmatrix} 0.659 & -0.563 \\ -0.913 & 0.780 \end{pmatrix}$

de donde $\| \mathbf{A}^{-1} \|_\infty = 0.913 \text{ x } 10^6 + 0.780 \text{ x } 10^6 = 1.693 \text{ x } 10^6$

$$K_\infty(\mathbf{A}) = \| \mathbf{A} \|_\infty \| \mathbf{A}^{-1} \|_\infty = 1.572 \text{ x } 1.693 \text{ x } 10^6 = 2.7 \text{ x } 10^6$$

Alternativamente, trabajando con normas Euclidianas:

$$\mathbf{A}^T \mathbf{A} = \begin{pmatrix} 1.441969 & 1.0040807 \\ 1.040807 & 0.751250 \end{pmatrix}$$

cuyos valores característicos son $\bar{\lambda}_{\text{máx}} = 2.1932$, $\bar{\lambda}_{\text{mín}} = 4.56 \text{ x } 10^{-13}$

de donde $K_2(\mathbf{A}) = \left(\bar{\lambda}_{\text{máx}} / \bar{\lambda}_{\text{mín}} \right)^{1/2} = 2.2 \text{ x } 10^6$

Ambos resultados indican un mal condicionamiento de la matriz **A**.

Note que en el ejemplo anterior la matriz **A** no era simétrica, por lo que fue necesario evaluar los valores característicos de $\mathbf{A}^T \mathbf{A}$. Si **A** fuera simétrica, los valores característicos de $\mathbf{A}^T \mathbf{A}$ serían exactamente los cuadrados de los valores característicos de **A**.

2.6.3 Errores de redondeo en la solución de sistemas de ecuaciones lineales por el método de Gauss (y otros métodos de eliminación similares)

Las relaciones teóricas utilizadas en la reducción son:

$$\begin{aligned} l_{ki} &= a_{ki}^{(i)} / a_{ii}^{(i)} \\ a_{kj}^{(i+1)} &= a_{kj}^{(i)} - l_{ki} a_{ij}^{(i)} \\ b_k^{(i+1)} &= b_k^{(i)} - l_{ki} b_i^{(i)} \end{aligned} \tag{2.30}$$

Sin embargo, como resultado de los errores de redondeo, los valores calculados (aquí indicados en barras) satisfacen:

$$\begin{aligned} \overline{l_{ki}} &= (\bar{a}_{ki}^{(i)} / \bar{a}_{ii}^{(i)})(1+\delta_1) \\ \bar{a}_{kj}^{(i+1)} &= (\bar{a}_{kj}^{(i)} - \overline{l_{ki}}\, \bar{a}_{ij}^{(i)} (1+\delta_2))(1+\delta_3) \\ \bar{b}_k^{(i+1)} &= (\bar{b}_k^{(i)} - \overline{l_{ki}}\, \bar{b}_i^{(i)} (1+\delta_4))(1+\delta_5) \end{aligned} \tag{2.31}$$

donde $|\delta_i| \leq \in$, siendo $\in$ el máximo error relativo de redondeo. Alternativamente puede escribirse:

$$\overline{\overline{l_{ki}}} = (\overline{a}_{ki}^{(i)} + e_{ki}^{(i)}) / \overline{a}_{ii}^{(i)}$$

$$\overline{a}_{kj}^{(i+1)} = \overline{a}_{kj}^{(i)} - \bar{l}_{ki}\,\overline{a}_{ij}^{(i)} + e_{kj}^{(i)} \qquad (2.32)$$

$$\overline{b}_{k}^{(i+1)} = \overline{b}_{k}^{(i)} - \bar{l}_{ki}\,\overline{b}_{i}^{(i)} + e_{k}^{(i)}$$

y puede probarse que:

$$\left|e_{ki}^{(i)}\right| \le \in \left|\overline{a}_{ki}^{(i)}\right|$$

$$\left| e_{kj}^{(i)}\right| \le 3 \in .máx\left(\left|\overline{a}_{kj}^{(i)}\right|, \left|\overline{a}_{kj}^{(i+1)}\right| \right) \qquad (2.33)$$

$$\left| c_{k}^{(i)}\right| \le 3 \in .máx\left(\left|\overline{b}_{k}^{(i)}\right|, \left|\overline{b}_{k}^{(i+1)}\right| \right)$$

Por otro lado, considerando que $\overline{a}_{kj}^{(1)} = a_{kj}$, $l_{kk} = 1$, pueden utilizarse las expresiones precedentes para escribir a_{kj} en función de los $\bar{l}_{ki}, \overline{a}_{ij}^{(1)}$. (es decir los elementos de las matrices **L** y **U**). Se obtiene así:

$$a_{kj} + \sum_{i=1}^{r} e_{kj}^{(i)} = \sum_{i=1}^{s} \bar{l}_{ki}\,\overline{a}_{ij}^{(i)} \qquad (2.34a)$$

donde *r = min (k-1,j), s = min (k, j).* Por otro lado, teniendo en cuenta que $\overline{b}_{k}^{(1)} = b_k$, se obtiene:

$$b_k + \sum_{i=1}^{k-1} c_{k}^{(i)} = \sum_{i=1}^{k} \bar{l}_{ki}\,\overline{b}_{i}^{(i)} \qquad (2.34b)$$

Esto demuestra que las matrices calculadas:

$$\overline{\mathbf{L}} = (\bar{l}_{ki}) \qquad \overline{\mathbf{U}} = (\overline{a}_{ij}^{(i)}) \qquad \mathbf{y} = (b_i^{(i)})$$

No son factores exactos de **A** y **b** sino de **A** + Δ**A** y **b** + Δ**b**:

$$\mathbf{A} + \Delta\mathbf{A} = \overline{\mathbf{L}}\,\overline{\mathbf{U}}$$

$$\mathbf{b} + \Delta\mathbf{b} = \overline{\mathbf{L}}\,\overline{\mathbf{y}}$$

Los elementos de Δ**A** son sumatorias de los $e_{kj}^{(i)}$; los elementos de Δ**b** son sumatorias de los $c_k^{(i)}$. Las expresiones (4) dan una medida de estas perturbaciones. Obsérvese que las expresiones (2.23) y (2.27) son aplicables también en este caso, y un valor de $K(\mathbf{A})$ alto indica que los errores de redondeo tendrán efectos importantes en la solución.

Por otro lado, las expresiones (2.33) y (2.34) indican que es conveniente limitar el crecimiento de los $a_{kj}^{(i)}$, $b_k^{(i)}$. Este es el propósito al realizar intercambios de filas y/o columnas.

Finalmente, debe mencionarse que en el proceso de sustitución inversa, para obtener **x** resolviendo $\overline{\mathbf{U}}\,\mathbf{x} = \overline{\mathbf{y}}$, los errores acumulados son despreciables en términos relativos a los que resultan de la reducción.

Las ecuaciones precedentes permiten una estimación a-posteriori de la magnitud del error. A-priori puede establecerse [1]:

$$g_n = \frac{\underset{i,j,k}{máx}\left| a_{ij}^{(k)} \right|}{\underset{i,j}{máx}\left| a_{ij}^{(1)} \right|} \qquad (2.35)$$

teniendo que:

$g_n \leq 2^{n-1}$ para intercambio parcial (filas)

$g_n \leq 1.8\, n^{0.25 Ln\, n}$ para intercambio total.

Estos límites son teóricos. Nótese por ejemplo que para un sistema de orden 100 se tendría $g_n \leq 6.3x10^{29}$ para intercambio parcial y $g_n \leq 18$ para intercambio completo, lo que justificaría el trabajo adicional necesario para la segunda alternativa. Sin embargo, en la práctica rara vez se observa un g_n mayor que 10, aún con intercambio parcial. Para matrices simétricas definidas positivas se tiene que $g_n \leq 1$.

2.6.4 Algunas consideraciones relativas a unidades. Equilibrio de las ecuaciones.

En un sistema de ecuaciones $\mathbf{A}\,\mathbf{x} = \mathbf{b}$... los a_{ij}, b_i, x_j pueden expresarse en diversos sistemas de unidades. Un cambio de unidades equivale a considerar $\mathbf{b} = \mathbf{D}_1\,\mathbf{b}$'; $\mathbf{x} = \mathbf{D}_2\,\mathbf{x}$' y por lo tanto $(\mathbf{D}_1\,\mathbf{A}\,\mathbf{D}_2)\,\mathbf{x}$' == $\mathbf{D}_1\,\mathbf{b}$'. En estas expresiones las matrices $\mathbf{D}_1$ y $\mathbf{D}_2$ son diagonales. Puede demostrarse que, si se utilizan los mismos pivotes y las $\mathbf{D}_1$ y $\mathbf{D}_2$ solo contienen potencias enteras de la base del sistema de numeración utilizado, los resultados son los mismos (habida cuenta de los cambios de unidades).

Sin embargo las unidades utilizadas pueden afectar la selección de pivotes, especialmente si sólo se hace intercambio parcial.

En tal caso, es recomendable *equilibrar* las ecuaciones. Para las incógnitas deben seleccionarse escalas que reflejen su importancia relativa. Las ecuaciones deben multiplicarse por factores $\mathbf{D}_1$ tales que:

$$\underset{1\leq j\leq n}{máx}\left| a_{ij} \right| = 1 \quad i=1,2,3,...n$$

2.6.5 Método iterativo para mejorar la solución

Considérese el sistema de ecuaciones $\mathbf{Ax} = \mathbf{b}$ para el que se tiene la solución aproximada $\mathbf{x}^{(0)}$. Si $\bar{\mathbf{x}}$ es la solución exacta, se tiene que:

$$\bar{\mathbf{x}} = \mathbf{x}^{(0)} + \Delta\mathbf{x}^{(0)}$$

y entonces:

$$\mathbf{A}\,\Delta\mathbf{x}^{(0)} = \mathbf{r}^{(0)}$$

donde: $\mathbf{r}^{(0)} = \mathbf{b} - \mathbf{A}\mathbf{x}^{(0)}$

Al determinar . $\mathbf{x}^{(0)}$. se obtienen los factores triangulares aproximados $\bar{\mathbf{L}}$ y $\bar{\mathbf{U}}$ tales que $\bar{\mathbf{L}}\,\bar{\mathbf{U}} = \mathbf{A} + \Delta\mathbf{A}$, siendo $\Delta\mathbf{A}$ pequeño. Esta descomposición requiere aproximadamente $O\left(\tfrac{1}{3}n^3\right)$ operaciones.

A partir de $\mathbf{x}^{(0)}$ puede determinarse $\mathbf{r}^{(0)}$ en $O\left(n^2\right)$ operaciones y resolverse:

$$\overline{\mathbf{L}}\,\mathbf{z} = \mathbf{r}$$

$$\overline{\mathbf{U}}\,\Delta\mathbf{x} = \mathbf{z}$$

también en $O\left(n^2\right)$ operaciones. Dado que $\overline{\mathbf{L}}$ y $\overline{\mathbf{U}}$ no son los factores exactos de $\mathbf{A}$, y además se introducen nuevos errores de redondeo, es necesario iterar:

$$\begin{aligned}
&\mathbf{r}^{(i)} = \mathbf{b} - \mathbf{A}\,\mathbf{x}^{(i)} \\
&\overline{\mathbf{L}}\,\mathbf{z}^{(i)} = \mathbf{r}^{(i)} \\
&\overline{\mathbf{U}}\,\Delta\mathbf{x}^{(i)} = \mathbf{z}^{(i)} \\
&\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \Delta\mathbf{x}^{(k)}
\end{aligned} \qquad (2.36)$$

Pero nada se ganaría si las operaciones se hicieran siempre con el mismo número de cifras significativas empleadas en los cómputos originales. Si los a_{ij} b_i x_i están dados con t dígitos, el cómputo de los residuos:

$$r_i^{(k)} = b_i - \sum_{j=1}^{n} a_{ij} x_j^{(k)}$$

debe hacerse con $2t$ dígitos (para minimizar errores de cancelación). Sin embargo, el almacenamiento de los resultados puede hacerse en precisión simple, es decir, con t dígitos.

Los vectores $\Delta\mathbf{x}^{(1)}$ y $\mathbf{x}^{(2)}$ permiten también estimar el número de condicionamiento:

$$\kappa(\mathbf{A}) \leq \frac{1}{n\varepsilon} \frac{\left\| \Delta\mathbf{x}^{(1)} \right\|}{\left\| \mathbf{x}^{(2)} \right\|} \qquad (2.37)$$

donde n es el orden del sistema y $\in$ es el máximo error relativo de redondeo (al operar en precisión simple). Si $\left\| \Delta\mathbf{x}^{(1)} \right\|$ no es mucho menor que $\left\| \mathbf{x}^{(1)} \right\|$, o lo que es lo mismo, si $\kappa(\mathbf{A})\, n\, \varepsilon$ no es mucho menor que 1, el proceso iterativo no es adecuado. En tal caso, la única alternativa sería operar con mayor precisión en toda la solución.

Considérese, por ejemplo, el sistema de ecuaciones:

$$\begin{pmatrix} 5 & 7 & 3 \\ 7 & 11 & 2 \\ 3 & 2 & 6 \end{pmatrix} \begin{Bmatrix} x_1 \\ x_2 \\ x_3 \end{Bmatrix} = \begin{Bmatrix} 0. \\ -1 \\ 0. \end{Bmatrix}$$

y supóngase que la computadora opera en base 10 con 3 cifras significativas. La factorización de la matriz de coeficientes, $\mathbf{A} = \overline{\mathbf{L}}\,\overline{\mathbf{U}}$, resulta en:

$$\begin{pmatrix} 5 & 7 & 3 \\ 7 & 11 & 2 \\ 3 & 2 & 6 \end{pmatrix} = \begin{pmatrix} 1.00 & & \\ 1.40 & 1.00 & \\ 0.60 & -1.83 & 1.00 \end{pmatrix} \begin{pmatrix} 5.00 & 7.00 & 3.00 \\ & 1.20 & -2.20 \\ & & 0.17 \end{pmatrix}$$

De la reducción del segundo miembro, es decir la solución de $\overline{\mathbf{L}}\,\mathbf{y} = \mathbf{b}$ se obtiene:

$$\mathbf{y} = \begin{pmatrix} 0.00 & -1.00 & -1.83 \end{pmatrix}^T$$

Finalmente por sustitución inversa, es decir resolviendo $\overline{\mathbf{U}}\,\mathbf{x} = \mathbf{y}$, se determina

$$x^{(1)} = \begin{pmatrix} 35.3 & -20.6 & -10.8 \end{pmatrix}^T$$

Para esta solución aproximada se tiene el residuo:

$$\mathbf{r}^{(1)} = \mathbf{b} - \mathbf{A}\,\mathbf{x}^{(1)} = \begin{pmatrix} 0.100 & 0.100 & 0.100 \end{pmatrix}^T$$

El cómputo de los $b_i - \sum a_{ij}\, x_j$ deben hacerse en doble precisión, almacenándose los resultados r_i en precisión simple.

Resolviendo los dos sistemas triangulares: $\overline{\mathbf{L}}\,\mathbf{z} = \mathbf{r}^{(1)}$ y $\overline{\mathbf{U}}\,\Delta\mathbf{x}^{(1)} = \mathbf{z}$ se obtiene:

$$\Delta\mathbf{x}^{(1)} = \begin{pmatrix} 0.685 & -0.391 & -0.195 \end{pmatrix}^T$$

Y entonces:

$$\mathbf{x}^{(2)} = \mathbf{x}^{(1)} + \Delta\mathbf{x}^{(1)} = \begin{pmatrix} 36.0 & -21.0 & -11.0 \end{pmatrix}^T$$

(redondeado a 3 cifras significativas). Este resultado es mejor que $x^{(1)}$ (en este caso el resultado es exacto, aunque debería decirse que por accidente).

Puede verificarse fácilmente que la matriz $\mathbf{A}$ del ejemplo anterior es bien condicionada.

Por otro lado, considérese nuevamente el sistema:

$$\begin{pmatrix} 0.780 & 0.563 \\ 0.913 & 0.659 \end{pmatrix} \begin{Bmatrix} x_1 \\ x_2 \end{Bmatrix} = \begin{Bmatrix} 0.217 \\ 0.254 \end{Bmatrix}$$

para el cual se obtuvo anteriormente $\kappa(\mathbf{A})$ de orden 2×10^6. Supóngase que se opera en base 10 con 6 cifras significativas:

$$\begin{pmatrix} 0.780 & 0.563 \\ 0.913 & 0.659 \end{pmatrix} = \begin{pmatrix} 1.00\,000 & \\ 1.17\,051 & 1.00\,000 \end{pmatrix} \begin{pmatrix} 0.780\,000 & 0.563\,000 \\ & 3 \cdot 10^{-6} \end{pmatrix}$$

se pierden cifras significativas en el elemento a_{22} de esta última matriz al restar dos números que solo difieren en la última cifra almacenada). De aquí resultan:

$$\mathbf{x}^{(1)} = \begin{pmatrix} 0.518\,803 & -0.333\,333 \end{pmatrix}^T$$

$$\mathbf{r}^{(1)} = \begin{pmatrix} 0.139 \cdot 10^{-6} & 0.692 \cdot 10^{-6} \end{pmatrix}^T$$

No obstante ser este residuo “pequeño”, se obtiene la corrección:

$$\Delta\mathbf{x}^{(1)} = \begin{pmatrix} -0.127\,348 & 0.176\,433 \end{pmatrix}^T$$

$$\mathbf{x}^{(2)} = \mathbf{x}^{(1)} + \Delta\mathbf{x}^{(1)} = \begin{pmatrix} 0.391\,455 & -0.156\,900 \end{pmatrix}^T$$

y es obvio que este resultado difiere más de la solución exacta $\mathbf{x} = \begin{pmatrix} 1 & -1 \end{pmatrix}^T$ que la aproximación $\mathbf{x}^{(1)}$ antes obtenida. ¡Para resolver este sistema de ecuaciones se requiere trabajar con un mínimo de 8 cifras significativas!

2.7. Métodos Iterativos para la Solución de Sistemas de Ecuaciones Lineales

En los acápites siguientes se tratan dos tipos distintos de métodos iterativos. Estos procesos pueden ser muy eficientes cuando la matriz de coeficientes, $\mathbf{A}$, es de baja densidad, más aún si la evaluación de productos de la forma $\mathbf{Av}$ no requiere la previa determinación y el almacenamiento de $\mathbf{A}$ en forma explícita.

2.7.1 Métodos de Relajación

Estos procedimientos son adecuados sólo cuando la diagonal principal de la matriz de coeficientes es dominante. En general, se considera una aproximación inicial, tal como $\mathbf{x}^{(0)} = \mathbf{0}$, y ésta es sucesivamente mejorada hasta obtener una solución suficientemente precisa.

Considérese el sistema de orden n: $\mathbf{A}\,\mathbf{x} = \mathbf{b}$, con $a_{ii} \neq 0$ para todo i. En el método de *Jacobi* se calculan las aproximaciones $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \mathbf{x}^{(3)} \cdots$ mediante:

$$x_i^{(k+1)} = \frac{1}{a_{ii}} \left(b_i - \sum_{j \neq i} a_{ij}\, x_j^{(k)} \right) \tag{2.38}$$

La aproximación es arbitraria; con frecuencia $\mathbf{x}^{(0)} = \mathbf{0}$. Si los $x_i^{(k+1)}$ se determinan en el orden habitual, al determinar $x_r^{(k+1)}$ ya se han previamente obtenido las nuevas aproximaciones $x_1^{(k+1)}, x_2^{(k+1)} \cdots x_{r-1}^{(k+1)}$. Sin embargo, en el método de Jacobi no se hace uso de estas nuevas aproximaciones hasta la iteración siguiente, difiriendo en esto del método de *Gauss - Seidel*:

$$x_i^{(k+1)} = \frac{1}{a_{ii}} \left(b_i - \sum_{j=1}^{i-1} a_{ij}\, x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij}\, x_j^{(k)} \right) \tag{2.39}$$

Nótese que sólo se requiere almacenar las últimas aproximaciones a los x_i.

En el ejemplo siguiente se usan las dos alternativas:

$$\begin{pmatrix} 5 & -1 & -1 & 0 \\ -1 & 5 & 0 & -1 \\ -1 & 0 & 5 & -1 \\ 0 & -1 & -1 & 5 \end{pmatrix} \begin{Bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{Bmatrix} = \begin{Bmatrix} 1 \\ 2.75 \\ -1 \\ -2.75 \end{Bmatrix}$$

La solución exacta es

$$\mathbf{x} = \begin{pmatrix} 0.25 & 0.50 & -0.25 & -0.50 \end{pmatrix}^T$$

Con el método de Jacobi se obtienen las sucesivas aproximaciones:

k	$x_1^{(k)}$	$x_2^{(k)}$	$x_3^{(k)}$	$x_4^{(k)}$
0	0	0	0	0
1	0.2	0.55	-0.2	-0.55
2	0.27	0.48	-0.27	-0.48
3	0.242	0.508	-0.242	-0.508
4	0.2532	0.4968	-0.2532	-0.4968
5	0.24872	0.50128	-0.24872	-0.50128
6	0.250512	0.499488	-0.250512	-0.499488
7	0.249795	0.500205	-0.249795	-0.500205
8	0.250082	0.499918	-0.250082	-0.499918
9	0.249967	0.500033	-0.249967	-0.500033
10	0.250013	0.499987	-0.250013	-0.499987
11	0.249995	0.500005	-0.249995	-0.500005
12	0.250002	0.499998	-0.250002	-0.499998

k	$x_1^{(k)}$	$x_2^{(k)}$	$x_3^{(k)}$	$x_4^{(k)}$
13	0.249999	0.500001	-0.249999	-0.500001
14	0.250000	0.500000	-0.250000	-0.500000
15	0.250000	0.500000	-0.250000	-0.500000

La convergencia es mejor con el método de Gauss – Seidel:

k	$x_1^{(k)}$	$x_2^{(k)}$	$x_3^{(k)}$	$x_4^{(k)}$
0	0	0	0	0
1	0.2	0.59	-0.16	-0.464
2	0.286	0.5144	-0.2356	-0.49424
3	0.255760	0.502304	-0.247696	-0.499078
4	0.250922	0.500369	-0.249631	-0.499853
5	0.250147	0.500059	-0.249941	-0.499976
6	0.250024	0.500009	-0.249991	-0.499996
7	0.250004	0.500002	-0.249998	-0.499999
8	0.250001	0.500000	-0.250000	-0.500000
9	0.250000	0.500000	-0.250000	-0.500000

En algunos casos la convergencia puede acelerarse con *sobrerelajación*:

$$x_i^{(k+1)} = x_i^{(k)} + \beta\, r_i^{(k)} \tag{2.40a}$$

$$r_i^{(k)} = \frac{1}{a_{ii}}\left(b_i - \sum_{j=1}^{i-1} a_{ij}\, x_j^{(k+1)} - \sum_{j=i}^{n} a_{ij}\, x_j^{(k)} \right) \tag{2.40b}$$

El valor óptimo de β depende de $\mathbf{A}$ e incluso de la aproximación $\mathbf{x}^{(k)}$. Cuanto mayores sean los valores absolutos de los términos de la diagonal principal, respecto a la suma de los valores absolutos de los restantes coeficientes de la misma fila, más se aproxima β a 1. Para el ejemplo precedente, utilizando $\beta = 1.05$ se obtienen:

k	$x_1^{(k)}$	$x_2^{(k)}$	$x_3^{(k)}$	$x_4^{(k)}$
0	0	0	0	0
1	0.210000	0.621600	-0.165900	-0.481803
2	0.295197	0.507233	-0.240892	-0.497478
3	0.251172	0.500414	-0.249680	-0.499972
4	0.250096	0.500005	-0.249990	-0.499998
5	0.249998	0.500000	-0.250000	-0.500000
6	0.250000	0.500000	-0.250000	-0.500000

Estos métodos no son necesariamente más precisos que los procesos de eliminación. El ejemplo al inicio de la sección 2.6 muestra que si el sistema es mal condicionado puede aceptarse como correcta una solución totalmente equivocada, pero con la que se tiene un residuo “pequeño”.

2.7.2 Convergencia

En esta sección se analiza la convergencia de los métodos de relajación. Un paso típico en la solución de $\mathbf{A}\,\mathbf{x} = \mathbf{b}$ puede escribirse como:

$$\mathbf{x}^{(k+1)} = \mathbf{G}\,\mathbf{x}^{(k)} + \mathbf{f} \tag{2.41}$$

Esto puede verse más fácilmente si se considera la descomposición:

$$\mathbf{A} = \mathbf{D}\left(\mathbf{T}_i + \mathbf{I} + \mathbf{T}_s\right) \tag{2.42}$$

donde $\mathbf{D}$ es una matriz diagonal, con elementos a_{ii}; $\mathbf{T}_i$ y $\mathbf{T}_s$ son matrices triangulares, inferior y superior respectivamente, con ceros en la diagonal principal, cuyos coeficientes son los a_{ij}/a_{ii}. Por ejemplo:

$$\begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix} = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}\left[\begin{pmatrix} 0 & 0 \\ -\frac{1}{2} & 0 \end{pmatrix} + \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + \begin{pmatrix} 0 & -\frac{1}{2} \\ 0 & 0 \end{pmatrix}\right]$$

Con esta notación, para el método de Jacobi se tiene:

$$\mathbf{x}^{(k+1)} = -\left(\mathbf{T}_i + \mathbf{T}_s\right)\mathbf{x}^{(k)} + \mathbf{D}^{-1}\mathbf{b} \tag{2.43a}$$

es decir: $\mathbf{G} = -\left(\mathbf{T}_i + \mathbf{T}_s\right)$ (2.43b)

mientras que para el método de Gauss-Seidel puede escribirse:

$$\mathbf{x}^{(k+1)} = -\mathbf{T}_i\mathbf{x}^{(k+1)} - \mathbf{T}_s\mathbf{x}^{(k)} + \mathbf{D}^{-1}\mathbf{b} \tag{2.44a}$$

y por lo tanto: $\mathbf{G} = -\left(\mathbf{I} + \mathbf{T}_i\right)^{-1}\mathbf{T}_s$ (2.44b)

De modo similar, para el método de sobre relajación se tiene:

$$\mathbf{G} = \left(\mathbf{I} + \beta\,\mathbf{T}_i\right)^{-1}\left[(1-\beta)\mathbf{I} - \beta\,\mathbf{T}_s\right] \tag{2.45}$$

Por otro lado, dado, que la solución exacta, $\bar{\mathbf{x}}$, debe cumplir la ecuación (2.41), se tiene que:

$$\bar{\mathbf{x}} = \mathbf{G}\,\bar{\mathbf{x}} + \mathbf{f} \tag{2.46}$$

y restando (2.46) de (2.41):

$$\left(\mathbf{x}^{(k+1)} - \bar{\mathbf{x}}\right) = \mathbf{G}\left(\mathbf{x}^{(k)} - \bar{\mathbf{x}}\right) \tag{2.47a}$$

de donde:

$$\left(\mathbf{x}^{(k+1)} - \bar{\mathbf{x}}\right) = \mathbf{G}\left(\mathbf{x}^{(k)} - \bar{\mathbf{x}}\right) = \mathbf{G}^2\left(\mathbf{x}^{(k-1)} - \bar{\mathbf{x}}\right) = \cdots = \mathbf{G}^{k+1}\left(\mathbf{x}^{(0)} - \bar{\mathbf{x}}\right) \tag{2.47b}$$

Además, si $\phi_1, \phi_2, \phi_3 \cdots \phi_n$ son los vectores característicos de la matriz $\mathbf{G}$, a los que corresponden los valores característicos $\lambda_1, \lambda_2, \lambda_3 \cdots \lambda_n$, puede escribirse:

$$\left(\mathbf{x}^{(0)} - \bar{\mathbf{x}}\right) = \alpha_1\phi_1 + \alpha_2\phi_2 + \alpha_3\phi_3 + \cdots + \alpha_n\phi_n$$

ya que los vectores característicos constituyen una base completa. Es relativamente fácil probar que:

$$\left(\mathbf{x}^{(k)} - \bar{\mathbf{x}}\right) = \mathbf{G}^k\left(\mathbf{x}^{(0)} - \bar{\mathbf{x}}\right) = \alpha_1\lambda_1^k\phi_1 + \alpha_2\lambda_2^k\phi_2 + \alpha_3\lambda_3^k\phi_3 + \cdots + \alpha_n\lambda_n^k\phi_n \tag{2.47c}$$

Para tener convergencia:

$$\lim_{k\to\infty}\left(\mathbf{x}^{(k)} - \bar{\mathbf{x}}\right) = \mathbf{0} \tag{2.48a}$$

y por tanto se requiere $|\lambda_i| < 1$ para todo i, o lo que es lo mismo:

$$\rho(\mathbf{G}) = \max_i |\lambda_i| \leq 1 \qquad (2.48b)$$

$\rho(\mathbf{G})$ se denomina el *radio espectral* de la matriz $\mathbf{G}$.

Para k suficientemente grande el error se multiplica por $\rho(\mathbf{G})$ en cada paso, es decir se tiene aproximadamente $-\log_{10}[\rho(\mathbf{G})]$ cifras decimales exactas adicionales en cada paso.

No es práctico determinar con gran precisión los valores característicos de $\mathbf{G}$ (esto significaría más trabajo que resolver el sistema de ecuaciones), pero ciertos límites pueden ser fácilmente establecidos.

Para el método de Jacobi: $g_{ij} = -a_{ij}/a_{ii}$ si $i \neq j$ (2.49a)

$$g_{ii} = 0$$

y utilizando el teorema de Gerschgorin (véase el capítulo relativo a la evaluación de valores y vectores característicos):

$$\rho(\mathbf{G}) = \max_i |\lambda_i| \leq \max_i \sum_j |g_{ij}| \quad \text{o bien} \quad \max_j \sum_i |g_{ji}| \qquad (2.49b)$$

con lo que la condición de convergencia $\rho(\mathbf{G}) \leq 1$ puede rescribirse:

$$|a_{jj}| > \sum_{\substack{i=1 \\ i \neq j}}^{n} |a_{ij}| \qquad |a_{ii}| > \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}| \qquad (2.49c)$$

Estas son condiciones suficientes pero no necesarias. La convergencia es más rápida cuanto más fuertes son las desigualdades.

Para el método de Gauss – Seidel $\rho(\mathbf{G}) = \max_i [r_i/(1-s_i)]$ (2.50a)

donde:

$$r_i > \sum_{j=i+1}^{n} \frac{|a_{ij}|}{|a_{ii}|} \qquad s_i > \sum_{j=1}^{i-1} \frac{|a_{ij}|}{|a_{ii}|} \qquad (2.50b)$$

y finalmente se concluye que las condiciones para la convergencia son las mismas que para el método de Jacobi (aunque en general el método de Gauss -Seidel converge más rápidamente).

Un análisis similar del método de sobre relajación permite establecer la condición adicional: $0 < \beta \leq 2$

2.7.3 Métodos de Máxima Gradiente y de Gradiente Conjugada

En la primera parte de esta sección se consideran métodos para la solución de sistemas de ecuaciones $\mathbf{A}\,\mathbf{x} = \mathbf{b}$ con matriz $\mathbf{A}$ simétrica y definida positiva, es decir, $\mathbf{v}^T \mathbf{A}\,\mathbf{v} > 0$ para todo vector $\mathbf{v}$ no nulo.

Considérese la función:

$$f(\mathbf{x}) = \tfrac{1}{2}\mathbf{x}^T \mathbf{A}\,\mathbf{x} - \mathbf{x}^T \mathbf{b} \qquad (2.51)$$

Si $\bar{\mathbf{x}}$ es la solución exacta de $\mathbf{A}\,\bar{\mathbf{x}} = \mathbf{b}$ se tiene que:

$$f(\mathbf{x}) - f(\bar{\mathbf{x}}) = \left(\tfrac{1}{2}\mathbf{x}^T\mathbf{A}\,\mathbf{x} - \mathbf{x}^T\mathbf{b}\right) - \left(\tfrac{1}{2}\bar{\mathbf{x}}^T\mathbf{A}\,\bar{\mathbf{x}} - \bar{\mathbf{x}}^T\mathbf{b}\right)$$
$$= \tfrac{1}{2}\left(\mathbf{x}-\bar{\mathbf{x}}\right)^T\mathbf{A}\left(\mathbf{x}-\bar{\mathbf{x}}\right)$$

Pero, siendo $\mathbf{A}$ definida positiva: $\tfrac{1}{2}\left(\mathbf{x}-\bar{\mathbf{x}}\right)^T\mathbf{A}\left(\mathbf{x}-\bar{\mathbf{x}}\right) \geq 0$

Y por lo tanto $f(\mathbf{x}) - f(\bar{\mathbf{x}}) \geq 0$, es decir, $f(\mathbf{x}) \geq f(\bar{\mathbf{x}})$ (2.52)

La igualdad solo se da si $\mathbf{x} = \bar{\mathbf{x}}$. La solución de $\mathbf{A}\,\mathbf{x} = \mathbf{b}$ es entonces equivalente a una minimización de $f(\mathbf{x})$.

Dada la aproximación inicial $\mathbf{x}^{(0)}$, a la que corresponden el residuo $\mathbf{r}^{(0)} = \mathbf{b} - \mathbf{A}\,\mathbf{x}^{(0)}$ y el valor $f(\mathbf{x})$, debe determinarse una nueva aproximación, $\mathbf{x}^{(1)}$, tal que $f(\mathbf{x}^{(1)}) < f(\mathbf{x}^{(0)})$. Para reducir el valor de $f(\mathbf{x})$ lo más rápidamente posible, la corrección debe hacerse en la dirección de máxima gradiente. Debe entonces determinarse esta dirección, $\mathbf{z}$, tal que:

$$\frac{d}{d\alpha} f(\mathbf{x}^{(0)} + \alpha\,\mathbf{z})\bigg|_{\alpha=0}$$

sea máxima (en valor absoluto). Siendo $f(\mathbf{x}) = \tfrac{1}{2}\mathbf{x}^T\mathbf{A}\,\mathbf{x} - \mathbf{x}^T\mathbf{b}$, puede escribirse:

$$f(\mathbf{x}^{(0)} + \alpha\,\mathbf{z}) = \tfrac{1}{2}(\mathbf{x}^{(0)} + \alpha\,\mathbf{z})^T\mathbf{A}\,(\mathbf{x}^{(0)} + \alpha\,\mathbf{z}) - (\mathbf{x}^{(0)} + \alpha\,\mathbf{z})^T\mathbf{b}$$
$$= \tfrac{1}{2}\alpha^2\mathbf{z}^T\mathbf{A}\,\mathbf{z} - \alpha\,\mathbf{z}^T\mathbf{r}^{(0)} + f(\mathbf{x}^{(0)}) \qquad (2.53a)$$

de donde:

$$\frac{d}{d\alpha} f(\mathbf{x}^{(0)} + \alpha\,\mathbf{z})\bigg|_{\alpha=0} = -\mathbf{z}^T\mathbf{r}^{(0)}$$

Esto significa que debe tomarse la dirección $\mathbf{z} = \mathbf{r}^{(0)}$ (2.53b)

Ahora puede determinarse α_0 de modo que $f(\mathbf{x}^{(0)} + \alpha_0\,\mathbf{r}^{(0)})$ sea un mínimo. Rescribiendo (2.53a) con $\mathbf{z} = \mathbf{r}^{(0)}$ y derivando con respecto a α:

$$\frac{d}{d\alpha} f(\mathbf{x}^{(0)} + \alpha\,\mathbf{r}^{(0)}) = \alpha\,\mathbf{r}^{(0)^T}\mathbf{A}\,\mathbf{r}^{(0)} - \mathbf{r}^{(0)^T}\mathbf{r}^{(0)} = 0$$

de donde:

$$\alpha_0 = \frac{\mathbf{r}^{(0)^T}\mathbf{r}^{(0)}}{\mathbf{r}^{(0)^T}\mathbf{A}\,\mathbf{r}^{(0)}}$$

(dado que $\mathbf{A}$ es definida positiva, nunca se presenta el caso $\mathbf{r}^{(0)^T}\mathbf{A}\,\mathbf{r}^{(0)} = 0$)

Finalmente:

$$\mathbf{x}^{(1)} = \mathbf{x}^{(0)} + \alpha_0\,\mathbf{r}^{(0)}$$

El proceso puede repetirse en sucesivos ciclos:

$$\mathbf{r}^{(k)} = \mathbf{b} - \mathbf{A}\,\mathbf{x}^{(k)} \qquad (2.54a)$$

$$\alpha_k = \frac{\mathbf{r}^{(k)^T}\mathbf{r}^{(k)}}{\mathbf{r}^{(k)^T}\mathbf{A}\,\mathbf{r}^{(k)}} \qquad (2.54b)$$

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \alpha_k\,\mathbf{r}^{(k)} \qquad (2.54c)$$

Este método es siempre convergente, pero no puede conocerse a priori en cuantos ciclos se tendrá la precisión requerida.

En los párrafos siguientes se estudia una modificación de este proceso, el método de *Gradiente Conjugada*, para el que – al menos en teoría - puede garantizarse la convergencia en un número de pasos igual o inferior al orden del sistema de ecuaciones.

Considérese el sistema de ecuaciones de orden n, $\mathbf{A}\,\mathbf{x} = \mathbf{b}$. Dada una solución aproximada, $\mathbf{x}^{(0)}$, la solución exacta, $\bar{\mathbf{x}}$, puede escribirse como:

$$\bar{\mathbf{x}} = \mathbf{x}^{(0)} + \Delta\mathbf{x}$$

$\Delta\mathbf{x}$ puede expresarse como combinación lineal de n vectores linealmente independientes. En particular, si se consideran vectores $\mathbf{s}_0, \mathbf{s}_1, \mathbf{s}_2 \cdots \mathbf{s}_{n-2}, \mathbf{s}_{n-1}$, que satisfacen las relaciones de ortogonalidad:

$$\mathbf{s}_i^T \mathbf{A}\,\mathbf{s}_j = c_i \delta_{ij}$$

puede escribirse:

$$\mathbf{x}^{(1)} = \mathbf{x}^{(0)} + \alpha_0\,\mathbf{s}_0$$
$$\mathbf{x}^{(2)} = \mathbf{x}^{(1)} + \alpha_1\,\mathbf{s}_1$$
$$\cdots\cdots$$
$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \alpha_k\,\mathbf{s}_k$$
$$\cdots\cdots$$
$$\bar{\mathbf{x}} = \mathbf{x}^{(n)} = \mathbf{x}^{(n-1)} + \alpha_{n-1}\,\mathbf{s}_{n-1}$$

alternativamente:

$$\bar{\mathbf{x}} = \mathbf{x}^{(0)} + \sum_{k=0}^{n-1} \alpha_k\,\mathbf{s}_k \tag{2.55}$$

Suponiendo que los vectores $\mathbf{s}_k$ son conocidos, los coeficientes α_k pueden obtenerse utilizando las relaciones de ortogonalidad ya mencionadas. Dado que:

$$\mathbf{r}^{(i)} = \mathbf{b} - \mathbf{A}\,\mathbf{x}^{(i)} = \mathbf{A}\left(\bar{\mathbf{x}} - \mathbf{x}^{(i)}\right) = \sum_{k=i}^{n-1} \alpha_k\,\mathbf{A}\,\mathbf{s}_k \tag{2.56}$$

premultiplicando por $\mathbf{s}_j^T$ se obtiene:

$$\mathbf{s}_j^T \mathbf{r}^{(i)} = \sum_{k=i}^{n-1} \alpha_k\,\mathbf{s}_j^T \mathbf{A}\,\mathbf{s}_k = 0 \qquad \text{si } j < i \tag{2.57}$$
$$= \alpha_j\,\mathbf{s}_j^T \mathbf{A}\,\mathbf{s}_j \qquad \text{si } j \geq i$$

de donde puede escribirse:

$$\alpha_j = \frac{\mathbf{s}_j^T \mathbf{r}^{(j)}}{\mathbf{s}_j^T \mathbf{A}\,\mathbf{s}_j} \tag{2.58a}$$

Alternativamente, puede utilizarse

$$\alpha_j = \frac{\mathbf{r}^{(j)^T} \mathbf{r}^{(j)}}{\mathbf{s}_j^T \mathbf{A}\,\mathbf{s}_j} \tag{2.58b}$$

La expresión alternativa $\alpha_j = (\mathbf{s}_j^T \mathbf{r}^{(0)}) \big/ (\mathbf{s}_j^T \mathbf{A}\,\mathbf{s}_j)$ no es conveniente, por la acumulación de errores de redondeo.

Dado que los $\mathbf{s}_0, \mathbf{s}_1, \mathbf{s}_2 \cdots \mathbf{s}_{n-2}, \mathbf{s}_{n-1}$ son n vectores linealmente independientes en un espacio n-dimensional, el error siempre puede ser expresado como una combinación lineal de estos vectores , es decir el proceso debería llegar a la solución exacta (salvo errores de redondeo) en n pasos.

El vector $\mathbf{s}_{k+1}$ se obtiene eliminando de $\mathbf{r}^{(k+1)}$ la componente según $\mathbf{A}\,\mathbf{s}_k$:

$$\mathbf{s}_{k+1} = \mathbf{r}^{(k+1)} - \beta_k\, \mathbf{s}_k \tag{2.59}$$

donde:

$$\beta_k = \frac{\mathbf{s}_k^T \mathbf{A}\,\mathbf{r}^{(k+1)}}{\mathbf{s}_k^T \mathbf{A}\,\mathbf{s}_k} \tag{2.60}$$

En el proceso de determinación de pueden tenerse errores de cancelación importantes si son aproximadamente paralelos.

Es relativamente fácil probar que si $\mathbf{s}_0, \mathbf{s}_1, \mathbf{s}_2 \cdots \mathbf{s}_k$ son $\mathbf{A}$-ortogonales, entonces $\mathbf{s}_{k+1}$ calculado con (2.59) resulta también $\mathbf{A}$-ortogonal a todos los vectores previamente hallados. Para empezar, con $\mathbf{s}_k$:

$$\mathbf{s}_{k+1}^T \mathbf{A}\,\mathbf{s}_k = \mathbf{s}_k^T \mathbf{A}\left(\mathbf{r}^{(k+1)} - \beta_k\, \mathbf{s}_k\right) = \mathbf{s}_k^T \mathbf{A}\,\mathbf{r}^{(k+1)} - \frac{\mathbf{s}_k^T \mathbf{A}\,\mathbf{r}^{(+1)}}{\mathbf{s}_k^T \mathbf{A}\,\mathbf{s}_k}\left(\mathbf{s}_k^T \mathbf{A}\,\mathbf{s}_k\right) = 0$$

Por otro lado, de (2.57) se concluye que:

$$\mathbf{s}_{k+1}^T \mathbf{A}\,\mathbf{s}_j = \mathbf{r}^{(k)^T} \mathbf{A}\,\mathbf{s}_j$$

y

$$\mathbf{A}\,\mathbf{s}_j = \frac{1}{\alpha_j}\left(\mathbf{r}^{(j-1)} - \mathbf{r}^{(j)}\right)$$

y por lo tanto, para $j < k$:

$$\mathbf{s}_{k+1}^T \mathbf{A}\,\mathbf{s}_j = \frac{1}{\alpha_j}\left(\mathbf{r}^{(k)^T} \mathbf{r}^{(j-1)}\right) - \frac{1}{\alpha_j}\left(\mathbf{r}^{(k)^T} \mathbf{r}^{(j)}\right) = 0$$

El método de gradiente conjugada puede resumirse en los pasos siguientes:

Dado $\mathbf{x}^{(0)}$, determinar $\mathbf{r}^{(0)} = \mathbf{s}_0 = \mathbf{b} - \mathbf{A}\,\mathbf{x}^{(0)}$

Y luego para $k = 0, 1, 2 \cdots n-1$:

$$\mathbf{q}_k = \mathbf{A}\,\mathbf{s}_k \qquad \text{(no se requiere } \mathbf{A} \text{ en forma explícita)}$$

$$\alpha_k = \frac{\mathbf{r}^{(k)^T} \mathbf{r}^{(k)}}{\mathbf{s}_k^T\, \mathbf{q}_k}$$

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \alpha_k\, \mathbf{s}_k \tag{2.61}$$

$$\mathbf{r}^{(k+1)} = \mathbf{r}^{(k)} - \alpha_k\, \mathbf{q}_k$$

$$\beta_k = \frac{\mathbf{r}^{(k+1)^T} \mathbf{q}_k}{\mathbf{s}_k^T \mathbf{q}_k} = -\frac{\mathbf{r}^{(k+1)^T} \mathbf{r}^{(k+1)}}{\mathbf{r}^{(k)^T} \mathbf{r}^{(k)}}$$

$$\mathbf{s}_{k+1} = \mathbf{r}^{(k+1)} - \beta_k \, \mathbf{s}_k$$

Como ejemplo, considérese la solución del sistema de ecuaciones $Ax = b$ definido por:

$$\mathbf{A} = \begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{pmatrix} \qquad \mathbf{b} = \begin{Bmatrix} 0 \\ 0 \\ 4 \end{Bmatrix}$$

Con la aproximación inicial $\mathbf{x}^{(0)} = 0$ se obtienen:

$\mathbf{r}^{(0)} = \mathbf{s}_0 = \mathbf{b} - \mathbf{A}\,\mathbf{x}^{(0)}$	$\begin{pmatrix} 0 & 0 & 4 \end{pmatrix}^T$
$\mathbf{q}_0 = \mathbf{A}\,\mathbf{s}_0$	$\begin{pmatrix} 0 & -4 & 8 \end{pmatrix}^T$
$\alpha_0 = (\mathbf{r}^{(0)^T}\,\mathbf{r}^{(0)}) \big/ (\mathbf{s}_0^T\,\mathbf{q}_0)$	$^1/_2$
$\mathbf{x}^{(1)} = \mathbf{x}^{(0)} + \alpha_0\,\mathbf{s}_0$	$\begin{pmatrix} 0 & 0 & 2 \end{pmatrix}^T$
$\mathbf{r}^{(1)} = \mathbf{r}^{(0)} - \alpha_0\,\mathbf{q}_0$	$\begin{pmatrix} 0 & 2 & 0 \end{pmatrix}^T$
$\beta_0 = (\mathbf{r}^{(1)^T}\,\mathbf{q}_0) \big/ (\mathbf{s}_0^T\,\mathbf{q}_0)$	$^1/_4$
$\mathbf{s}_1 = \mathbf{r}^{(1)} - \beta_0\,\mathbf{s}_0$	$\begin{pmatrix} 0 & 2 & 1 \end{pmatrix}^T$
$\mathbf{q}_1 = \mathbf{A}\,\mathbf{s}_1$	$\begin{pmatrix} -2 & 3 & 0 \end{pmatrix}^T$
$\alpha_1 = (\mathbf{r}^{(1)^T}\,\mathbf{r}^{(1)}) \big/ (\mathbf{s}_1^T\,\mathbf{q}_1)$	$^2/_3$
$\mathbf{x}^{(2)} = \mathbf{x}^{(1)} + \alpha_1\,\mathbf{s}_1$	$\begin{pmatrix} 0 & 4/3 & 8/3 \end{pmatrix}^T$
$\mathbf{r}^{(2)} = \mathbf{r}^{(1)} - \alpha_1\,\mathbf{q}_1$	$\begin{pmatrix} 4/3 & 0 & 0 \end{pmatrix}^T$
$\beta_1 = (\mathbf{r}^{(2)^T}\,\mathbf{q}_1) \big/ (\mathbf{s}_1^T\,\mathbf{q}_1)$	$^4/_9$
$\mathbf{s}_2 = \mathbf{r}^{(2)} - \beta_1\,\mathbf{s}_1$	$\begin{pmatrix} 4/3 & 8/9 & 4/9 \end{pmatrix}^T$
$\mathbf{q}_2 = \mathbf{A}\,\mathbf{s}_2$	$\begin{pmatrix} 16/9 & 0 & 0 \end{pmatrix}^T$
$\alpha_2 = (\mathbf{r}^{(2)^T}\,\mathbf{r}^{(2)}) \big/ (\mathbf{s}_2^T\,\mathbf{q}_2)$	$^3/_4$
$\mathbf{x}^{(3)} = \mathbf{x}^{(2)} + \alpha_2\,\mathbf{s}_2$	$\begin{pmatrix} 1 & 2 & 3 \end{pmatrix}^T$

El método de gradiente conjugada puede ser generalizado para resolver cualquier sistema de ecuaciones $\mathbf{A}\,\mathbf{x} = \mathbf{b}$ (con $\mathbf{A}$ no singular):

Con $\mathbf{x}^{(0)}$ arbitrario, se obtiene $\mathbf{r}^{(0)} = \mathbf{s}_0 = \mathbf{b} - \mathbf{A}\,\mathbf{x}^{(0)}$

Y luego para $k = 0, 1, 2 \cdots n-1$:

$$\mathbf{q}_k = \mathbf{A}^T\,\mathbf{s}_k \qquad \text{(no se requiere } \mathbf{A} \text{ en forma explícita)}$$

$$\alpha_k = \frac{\mathbf{r}^{(k)^T}\,\mathbf{r}^{(k)}}{\mathbf{q}_k^T\,\mathbf{q}_k}$$

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \alpha_k\,\mathbf{q}_k \qquad (2.62)$$

$$\mathbf{r}^{(k+1)} = \mathbf{r}^{(k)} - \alpha_k \, \mathbf{A}\,\mathbf{q}_k$$

$$\beta_k = -\frac{\mathbf{r}^{(k+1)^T}\mathbf{r}^{(k+1)}}{\mathbf{r}^{(k)^T}\mathbf{r}^{(k)}}$$

$$\mathbf{s}_{k+1} = \mathbf{r}^{(k+1)} - \beta_k \, \mathbf{s}_k$$

2.8 Sistemas Sobre-Determinados de Ecuaciones Lineales

El problema de determinación de los parámetros de un modelo lineal para aproximar un conjunto de datos es frecuente. A fin de reducir la influencia de errores de medición, es habitual hacer más mediciones que las estrictamente necesarias, de donde resultan más ecuaciones que incógnitas.

Dada una matriz $\mathbf{A}$ de orden $m \times n$ ($m > n$) y un vector $\mathbf{b}$ de orden m, se requiere determinar $\mathbf{x}$ de modo tal que $\mathbf{A}\,\mathbf{x}$ sea la mejor aproximación posible a $\mathbf{b}$.

Un proceso simple (y muy adecuado si los errores en los b_i son estadísticamente independientes) es el método de *mínimos cuadrados*, que consiste en minimizar la magnitud del residuo $\mathbf{r} = \mathbf{b} - \mathbf{A}\,\mathbf{x}$ (o minimizar $|\mathbf{r}|^2 = \mathbf{r}^T\mathbf{r}$) con respecto a las $\mathbf{x}$. Dado que:

$$f = \mathbf{r}^T\mathbf{r} = \mathbf{b}^T\mathbf{b} - 2\mathbf{x}^T\mathbf{A}^T\mathbf{b} + \mathbf{x}^T\mathbf{A}^T\mathbf{A}\,\mathbf{x} \tag{2.63}$$

y por lo tanto:

$$\frac{\partial f}{\partial \mathbf{x}} = -2\,\mathbf{A}^T\,\mathbf{b} + 2\,\mathbf{A}^T\mathbf{A}\,\mathbf{x} = \mathbf{0}$$

el método de mínimos cuadrados puede formularse como la solución del sistema de *ecuaciones normales*:

$$\left(\mathbf{A}^T\mathbf{A}\right)\mathbf{x} = \mathbf{A}^T\mathbf{b} \tag{2.64}$$

Si $\mathbf{A} = \begin{pmatrix}\mathbf{a}_1 & \mathbf{a}_2 & \mathbf{a}_3 & \cdots & \mathbf{a}_n\end{pmatrix}$, la matriz simétrica $\mathbf{C} = \mathbf{A}^T\mathbf{A}$ tiene elementos $c_{ij} = \mathbf{a}_i^T\mathbf{a}_j$. La matriz $\mathbf{C}$ es no singular sólo si todas las columnas $\mathbf{a}_k$ de la matriz $\mathbf{A}$ son linealmente independientes.

Para formar las ecuaciones normales se requieren $\frac{1}{2}mn(n+3)$ operaciones. Para resolver el sistema $O\left(\frac{1}{6}n^3\right)$ operaciones. La mayor parte del trabajo está en formar las ecuaciones normales.

Considérese por ejemplo

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ -1 & 1 & 0 \\ 0 & -1 & 1 \\ 1 & 0 & 1 \end{pmatrix} \begin{Bmatrix} x_1 \\ x_2 \\ x_3 \end{Bmatrix} = \begin{Bmatrix} 1 \\ 2 \\ 3 \\ 1 \\ 2 \\ 1 \end{Bmatrix}$$

las ecuaciones normales son en este caso:

$$\begin{pmatrix} 3 & -1 & -1 \\ -1 & 3 & -1 \\ -1 & -1 & 3 \end{pmatrix} \begin{Bmatrix} x_1 \\ x_2 \\ x_3 \end{Bmatrix} = \begin{Bmatrix} -1 \\ 1 \\ 6 \end{Bmatrix}$$

de donde:

$$\mathbf{x} = \begin{pmatrix} 1.25 & 1.75 & 3 \end{pmatrix}^T$$

Un método alternativo (y numéricamente mejor condicionado) se basa en al descomposición de la matriz de coeficientes, $\mathbf{A}$, en el producto de una matriz ortogonal, $\mathbf{Q}$, y una matriz triangular superior, $\mathbf{R}$ (en el capítulo relativo a valores y vectores característicos se describen procedimientos que pueden ser empleados para esto).

Al tenerse $\mathbf{A} = \mathbf{Q}\,\mathbf{R}$ (2.65)

las ecuaciones normales $\left(\mathbf{A}^T\mathbf{A}\right)\mathbf{x} = \mathbf{A}^T\mathbf{b}$ pueden rescribirse:

$$\mathbf{A}^T\left(\mathbf{b} - \mathbf{A}\mathbf{x}\right) = 0$$
$$\mathbf{R}^T\mathbf{Q}^T\left(\mathbf{b} - \mathbf{Q}\,\mathbf{R}\,\mathbf{x}\right) = 0$$

y dado que $\mathbf{Q}^T\mathbf{Q} = \mathbf{I}$ se obtiene:

$$\mathbf{R}^T\left(\mathbf{Q}^T\mathbf{b} - \mathbf{R}\,\mathbf{x}\right) = 0$$

La matriz $\mathbf{R}$ no es singular y por tanto:

$$\mathbf{R}\,\mathbf{x} = \mathbf{Q}^T\mathbf{b} \qquad (2.66)$$

La matriz $\mathbf{R}$ es la misma que se obtendría al descomponer $\mathbf{A}^T\mathbf{A}$ en dos factores triangulares por el método de Cholesky. Para el ejemplo precedente:

$$\mathbf{A} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ -1 & 1 & 0 \\ 0 & -1 & 1 \\ 1 & 0 & 1 \end{pmatrix} = \mathbf{QR} = \begin{pmatrix} 0.5774 & 0.2041 & 0.3536 \\ 0 & 0.6124 & 0.3536 \\ 0 & 0 & 0.7071 \\ -0.5774 & 0.4082 & 0 \\ 0 & -0.6124 & 0.3536 \\ -0.5774 & -0.2041 & 0.3536 \end{pmatrix} \begin{pmatrix} 1.7321 & -0.5774 & -0.5774 \\ 0 & 1.6330 & -0.8165 \\ 0 & 0 & 1.4142 \end{pmatrix}$$

de donde:

$$\mathbf{R}\,\mathbf{x} = \begin{pmatrix} 1.7321 & -0.5774 & -0.5774 \\ 0 & 1.6330 & -0.8165 \\ 0 & 0 & 1.4142 \end{pmatrix} \begin{Bmatrix} x_1 \\ x_2 \\ x_3 \end{Bmatrix} = \mathbf{Q}^T\mathbf{b} = \begin{Bmatrix} -0.5774 \\ 0.4082 \\ 4.2426 \end{Bmatrix}$$

y finalmente: $\mathbf{x} = \begin{pmatrix} 1.25 & 1.75 & 3 \end{pmatrix}^T$