TEMA 3: RESOLUCION DE SISTEMAS DE ECUACIONES LINEALES Y NO LINEALES

Abordaremos en este tema la resolución de Sistemas de Ecuaciones Lineales (por diferentes métodos directos, iterativos y de descenso) y la resolución de Sistemas de Ecuaciones No Lineales (por métodos de iteración funcional y de Newton).

Antes de plantear los métodos de resolución de S.E.L. estudiaremos su condicionamiento, esto es, la sensibilidad de la solución exacta del sistema a las pequeñas variaciones en los datos.

1 CONDICIONAMIENTO DE UN S.E.L.

Consideremos un sistema de n ecuaciones lineales con n incógnitas $Ax = b$ con $A \in M_{n\times n}$ inversible (esto es, un sistema compatible determinado con solución única).

Ejemplo 1 .- *Sea el S.E.L.*

$$\begin{pmatrix} 10 & 7 & 8 & 7 \\ 7 & 5 & 6 & 5 \\ 8 & 6 & 10 & 9 \\ 7 & 5 & 9 & 10 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 32 \\ 23 \\ 33 \\ 31 \end{pmatrix} \quad \Rightarrow \quad \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}$$

Si perturbamos ligeramente b:

$$\begin{pmatrix} 10 & 7 & 8 & 7 \\ 7 & 5 & 6 & 5 \\ 8 & 6 & 10 & 9 \\ 7 & 5 & 9 & 10 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 32.1 \\ 22.9 \\ 33.1 \\ 30.9 \end{pmatrix} \quad \Rightarrow \quad \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 9.2 \\ -12.6 \\ 4.5 \\ -1.1 \end{pmatrix}$$

Si perturbamos ligeramente A:

$$\begin{pmatrix} 10 & 7 & 8.1 & 7.2 \\ 7.08 & 5.04 & 6 & 5 \\ 8 & 5.98 & 9.89 & 9 \\ 6.99 & 4.99 & 9 & 9.98 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 32 \\ 23 \\ 33 \\ 31 \end{pmatrix} \quad \Rightarrow \quad \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} -81 \\ 137 \\ -34 \\ 22 \end{pmatrix}$$

Por tanto, frente a pequeñas variaciones de A o de b, la solución x varía mucho. Diremos en este caso que el sistema está **mal condicionado**.

Para medir la variación de la solución respecto a las variaciones de A o b (*condicionamiento del sistema*) se introduce el **número de condición de** A:

$$cond_*(A) = \|A\|_*.\|A^{-1}\|_*$$

para una norma matricial subordinada $\|.\|_*$

Nosotros utilizaremos usualmente alguna de las normas matriciales subordinadas siguientes:

$$\begin{aligned} \|A\|_1 &= \max_{1\leq j\leq n} \sum_{i=1}^{n} |a_{ij}| \\ \|A\|_\infty &= \max_{1\leq i\leq n} \sum_{j=1}^{n} |a_{ij}| \\ \|A\|_2 &= \sqrt{\rho(A^t\, A)} \end{aligned}$$

Todas las normas matriciales subordinadas verifican, además de las propiedades características de toda norma:

$$\begin{aligned} &\|A\|_* \geq 0 \\ &\|A\|_* = 0 \;\Leftrightarrow A = 0 \\ &\|A+B\|_* \leq \|A\|_* + \|B\|_* \\ &\|c.A\|_* = |c|.\|A\|_* \end{aligned}$$

las siguientes propiedades:

$$\begin{aligned} &\|A\|_* \geq \rho(A) \\ &\|A.B\|_* \leq \|A\|_*.\|B\|_* \\ &\|A^{-1}\|_* \geq {\|A\|_*}^{-1} \\ &\|I\|_* = 1 \end{aligned}$$

Como consecuencia de estas propiedades y de la propia definición de número de condición, este verifica:

1. $cond(A) \geq 1, \quad \forall A \in M_{n\times n}$
2. $cond(A) = cond(A^{-1}), \quad \forall A \in M_{n\times n}$
3. $cond(cA) = cond(A), \quad \forall A \in M_{n\times n}, \ \forall c \neq 0$

Se tienen entonces los siguientes resultados:

Teorema 1 .- *Se consideran, para A inversible y para $b \neq 0$, los sistemas:*

$$Ax = b,$$

$$A\left(x + \delta x\right) = b + \delta b.$$

Entonces:

$$\frac{\|\delta x\|}{\|x\|} \leq cond(A)\, \frac{\|\delta b\|}{\|b\|}.$$

Teorema 2 .- *Se consideran, para A inversible y para $b \neq 0$, los sistemas:*

$$Ax = b,$$

$$\left(A + \delta A\right)\left(x + \delta x\right) = b.$$

Entonces:

$$\frac{\|\delta x\|}{\|x + \delta x\|} \leq cond(A)\, \frac{\|\delta A\|}{\|A\|}.$$

Observación 1 .- *Por tanto, cuanto mayor sea el número de condición peor será el condicionamiento del sistema.*

En el ejemplo anterior, el mal condicionamiento es debido a que $cond_2(A) = \|A\|_2.\|A^{-1}\|_2 = 2984$.

Para resolver los sistemas mal condicionados se usan las *técnicas de precondicionamiento*:

Dado un sistema $Ax = b$ mal condicionado (es decir, con $cond(A)$ grande), se busca una matriz inversible P tal que $cond(PA)$ sea pequeño. Entonces, se resuelve el sistema equivalente $PAx = Pb$ que ya es bien condicionado.

El caso más favorable sería $P = A^{-1}$, pues entonces $cond(PA) = cond(I) = 1$.
Las técnicas de precondicionamiento buscan una matriz P "próxima" a A^{-1}, pero fácilmente calculable. (La estrategia más favorable consiste en calcular directamente PA y Pb.)
Las técnicas más comunes son la factorización incompleta de Gauss, factorización incompleta de Choleski,...

2 METODOS DIRECTOS PARA S.E.L.

Los **métodos directos** se caracterizan por el hecho de que, si no hubiera errores de redondeo, se obtendría la solución exacta del sistema en un número finito de operaciones.

Presentamos los siguientes métodos:

2.1 Método de Gauss.-

Dado un S.E.L. $Ax = b$ con $A \in M_{n\times n}$ inversible, el principio que rige el **método de Gauss** para la resolución del sistema se puede resumir en "la determinación de una matriz inversible M tal que la matriz MA sea triangular superior". Este es el proceso llamado de *eliminación*. Una vez finalizado este proceso se resolverá el sistema triangular equivalente $MAx = Mb$ mediante el método de sustitución retrógrada.

En la práctica no se calcula M, sino directamente los productos MA y Mb.

El método de Gauss se realiza en tres bloques:

1. Proceso de eliminación sucesiva de incógnitas, que equivale a la determinación de una matriz M tal que MA sea triangular superior.

2. Cálculo del vector Mb, que se suele realizar simultáneamente al bloque 1.

3. Resolución de sistema triangular $MAx = Mb$ por sustitución retrógrada.

Se pueden distinguir básicamente tres versiones del método de Gauss:

- Gauss normal
- Gauss con pivote parcial
- Gauss con pivote total

y un cuarto algoritmo:

- factorización LU

que es la interpretación matricial de la primera versión.

Gauss normal:
El proceso de eliminación se realiza en $(n-1)$ etapas: en cada etapa k-ésima se obtienen ceros en la columna k por debajo de la diagonal principal. Así, partiendo de $A_1 = A$, en la etapa k-ésima se construye A_{k+1} a partir de $A_k = (a_{ij}^k)$.
Para poder realizar cada etapa k-ésima se exigirá (y esta es la característica esencial de Gauss normal) que:

$$a_{kk}^k \neq 0, \quad \forall k = 1, \ldots, n.$$

Etapa k-ésima :
Se hacen ceros en la columna k por debajo de la diagonal principal restando a las filas $i = k+1, \ldots, n$, la fila k multiplicada por $\dfrac{a_{ik}^k}{a_{kk}^k}$.
Matricialmente, esto corresponde a hacer $A_{k+1} = E_k A_k$ con:

$$E_k = \begin{pmatrix} 1 & & & & & 0 \\ & \ddots & & & & \\ & & 1 & & & \\ & & -\frac{a_{k+1,k}^k}{a_{kk}^k} & 1 & & \\ & & \vdots & & \ddots & \\ 0 & & -\frac{a_{nk}^k}{a_{kk}^k} & 0 & & 1 \end{pmatrix} \qquad (det(E_k) = 1).$$

Una vez realizadas las $(n-1)$ etapas se tiene:

$$A_n = \underbrace{E_{n-1} \ldots E_2 E_1}_{M} A = MA = U,$$

donde U es una matriz triangular superior, y simultáneamente:

$$E_{n-1} \ldots E_2 E_1 b = Mb.$$

Factorización LU:
Es la interpretación matricial de método Gauss normal. Teniendo en cuenta que todas las E_k son matrices triangulares inferiores con elementos diagonales iguales a uno (por tanto, inversibles), se tiene que la matriz:

$$L = M^{-1} = E_1^{-1} E_2^{-1} \dots E_{n-1}^{-1}$$

es también triangular inferior con unos en la diagonal. Por tanto:

$$MA = U \;\Leftrightarrow\; A = M^{-1}U = LU.$$

Teorema 3 .- *(Factorización LU)*
Sea $A = (a_{ij}) \in M_{n\times n}$ *tal que las* n *submatrices principales:*

$$\Delta_k = \begin{pmatrix} a_{11} & \dots & a_{1k} \\ \vdots & & \vdots \\ a_{k1} & \dots & a_{kk} \end{pmatrix}, \qquad k = 1, \dots, n,$$

sean inversibles (esto es, $det(\Delta_k) \neq 0$, $\forall k = 1, \dots, n$*). Entonces,* $a_{kk}^k \neq 0$, $\forall k = 1, \dots, n$, *y, por tanto, existe una matriz triangular inferior* $L = (l_{ij})$ *con elementos diagonales* $l_{ii} = 1$, $\forall i = 1, \dots, n$, *y una matriz triangular superior* $U = (u_{ij})$ *tales que* $A = LU$*. Además, esta factorización es única.*

Por tanto, si la eliminación gaussiana es "normal" la resolución del sistema $Ax = b$ se puede sustituir por la resolución de dos sistemas con matriz triangular (por tanto, fácilmente resolubles por sustitución):

$$Ax = b \;\Leftrightarrow\; L\underbrace{Ux}_{y} = b \;\Leftrightarrow\; \begin{cases} Ly = b \\ Ux = y \end{cases}$$

Fórmulas para el cálculo de L y U :

$$\begin{cases} u_{1j} = a_{1j}, \quad j = 1, \ldots, n. \\ \\ l_{j1} = \dfrac{a_{j1}}{u_{11}}, \quad j = 2, \ldots, n. \end{cases}$$

$$\text{Para } \; i = 2, \ldots, n :$$

$$\begin{cases} u_{ij} = a_{ij} - \sum_{k=1}^{i-1} l_{ik} u_{kj}, \quad j = i, \ldots, n. \\ \\ l_{ji} = \dfrac{a_{ji} - \sum_{k=1}^{i-1} l_{jk} u_{ki}}{u_{ii}}, \quad j = i+1, \ldots, n. \end{cases}$$

Observación 2 *.- A la hora de implementar el método en el ordenador, esta estructura de cálculo permite almacenar cada u_{ij} en la posición de a_{ij} y cada l_{ji} en la correspondiente posición de a_{ji}, con el consiguiente ahorro de memoria.*

Como aplicación directa del método LU se tiene que:

$$det(A) = \underbrace{det(L)}_{=1} det(U) = det(U) = u_{11} u_{22} \ldots u_{nn}.$$

Gauss con pivote parcial:

No es difícil encontrar ejemplos de sistemas donde el método de Gauss normal no puede llevarse a la práctica, puesto que algún elemento a_{kk}^k resulta nulo. En este

caso será necesario, antes del proceso de eliminación, un cambio de orden en las ecuaciones del sistema a fin de obtener un elemento diagonal no nulo. En esto consiste en método de Gauss con pivote parcial, que también se desarrolla en $(n-1)$ etapas:

Etapa k-ésima : Cada etapa se realiza en dos pasos:

PASO 1: Se elige el *pivote parcial* y se lleva a la diagonal, esto es, se intercambia la fila k con la fila $i \geq k$ tal que el pivote a_{ik}^k sea no nulo y verifique:

$$|a_{ik}^k| = \max_{k \leq p \leq n} |a_{pk}^k|$$

Si no existe dicho elemento, se pasa a la etapa $(k+1)$. Matricialmente, este paso equivale a multiplicar $P_k A_k$, con:

$$P_k = \begin{cases} I, & \text{si } \ i = k, \\ V_{ki}, & \text{si } \ i \neq k, \end{cases}$$

donde V_{ki} es la matriz de permutación resultante de intercambiar en la identidad las filas i y k:

$$V_{ki} = \begin{pmatrix} 1 & & & & & & 0 \\ & \ddots & & & & & \\ & & 0 & & 1 & & \\ & & & \ddots & & & \\ & & 1 & & 0 & & \\ & & & & & \ddots & \\ 0 & & & & & & 1 \end{pmatrix} \qquad (det(V_{ki}) = -1).$$

Por tanto:

$$det(P_k) = \begin{cases} 1, & \text{si } \ i = k, \\ -1, & \text{si } \ i \neq k, \end{cases}$$

PASO 2: Se hacen ceros en la columna k por debajo de la diagonal principal como en el caso de Gauss normal. Matricialmente, corresponde a hacer $A_{k+1} = E_k P_k A_k$.

Una vez realizadas las $(n-1)$ etapas se tiene la matriz triangular superior:

$$A_n = \underbrace{E_{n-1}P_{n-1}\ldots E_2P_2E_1P_1}_{M} A = MA = U,$$

y simultáneamente:

$$E_{n-1}P_{n-1}\ldots E_2P_2E_1P_1b = Mb.$$

Teorema 4 .- *Sea $A \in M_{n\times n}$ cualquiera. Existe, al menos, una matriz M inversible tal que MA es triangular superior. Además:*

$$det(M) = (-1)^{\text{número de cambios de filas}} = \pm 1.$$

Por tanto:

$$det(A) = (-1)^{\text{número de cambios de filas}}\, u_{11}u_{22}\ldots u_{nn}.$$

Observación 3 .- *Las estrategias de pivote, además de hacer factible el proceso de eliminación, evitan la división por números muy pequeños con los consiguientes errores de redondeo.*

Gauss con pivote total:
Constituye una nueva estrategia para evitar en la medida de lo posible los errores de redondeo de los métodos de Gauss normal y con pivote parcial.
La idea consiste en elegir como *pivote total* el elemento de mayor valor absoluto, no sólo en la columna k, sino en todas las columnas j con $k \leq j \leq n$. Esto introduce una nueva dificultad, ya que el cambio de orden en las columnas corresponde a un cambio de orden en las incógnitas, hecho que debe ser tenido en cuenta a la hora de la resolución de sistema triangular.
El método de Gauss con pivote total se realiza en $(n-1)$ etapas:

Etapa k-ésima : Cada etapa se realiza en dos pasos:

PASO 1: Se elige el *pivote total* $a_{ij}^k \neq 0$ tal que:

$$|a_{ij}^k| = \max_{k \leq p,q \leq n} |a_{pq}^k|$$

y se intercambian la fila k con la fila i y la columna k con la columna j. Este paso equivale a multiplicar $P_k A_k Q_k$,

con:

$$P_k = \begin{cases} I, & \text{si } i = k, \\ V_{ki}, & \text{si } i \neq k, \end{cases} \qquad Q_k = \begin{cases} I, & \text{si } j = k, \\ V_{kj}, & \text{si } j \neq k. \end{cases}$$

PASO 2: Se hacen ceros en la columna k por debajo de la diagonal principal como en el caso de Gauss normal. Matricialmente, corresponde a hacer $A_{k+1} = E_k P_k A_k Q_k$.

Una vez realizadas las $(n-1)$ etapas se tiene la matriz triangular superior:

$$A_n = \underbrace{E_{n-1}P_{n-1}\ldots E_2P_2E_1P_1}_{M} A \underbrace{Q_1Q_2\ldots Q_{n-1}}_{N} = MAN = U,$$

y simultáneamente:

$$E_{n-1}P_{n-1}\ldots E_2P_2E_1P_1b = Mb.$$

Teorema 5 .- *Sea $A \in M_{n\times n}$ cualquiera. Existen, al menos, dos matrices M y N inversibles tales que MAN es triangular superior. Además:*

$$det(M) = (-1)^{\text{número de cambios de filas}} = \pm 1.$$

$$det(N) = (-1)^{\text{número de cambios de columnas}} = \pm 1.$$

Por tanto:

$$det(A) = (-1)^{\text{número de cambios de filas y columnas}}\, u_{11}u_{22}\ldots u_{nn}.$$

Observación 4 .- *A la hora de plantear el sistema triangular se tiene:*

$$Ax = b \;\Leftrightarrow\; \underbrace{MAN}_{U}\underbrace{N^{-1}x}_{y} = Mb \;\Leftrightarrow\; Uy = Mb.$$

Al resolver el sistema se obtiene y*, que es simplemente una reordenación de las coordenadas de la solución buscada* x*, pues:*

$$y = N^{-1}x = Q_{n-1}^{-1}\ldots Q_2^{-1}Q_1^{-1}x = Q_{n-1}\ldots Q_2Q_1x.$$

2.2 Variantes al método de Gauss.-

Dentro de cada una de las versiones del método de Gauss (normal, LU, con pivote parcial o total) se pueden considerar diferentes variantes. Por ejemplo:

Método de Gauss-Jordan:

El método de G.-J. se basa en encontrar una matriz inversible $\tilde{M}$ tal que $\tilde{M}A$ sea diagonal y, a continuación, resolver el sistema diagonal equivalente $\tilde{M}Ax = \tilde{M}b$. (Consiste simplemente en dividir las componentes de $\tilde{M}b$ por los correspondientes elementos diagonales de $\tilde{M}A$.)

Aunque se puede hablar de Gauss-Jordan normal, con pivote parcial o total, sólo estudiaremos el caso de pivote parcial. El algoritmo de eliminación se desarrolla en n etapas: en cada etapa k-ésima se hacen ceros en toda la columna k excepto en el elemento diagonal.

Etapa k-ésima : Cada etapa se realiza en dos pasos:

PASO 1: Se elige el *pivote parcial*, igual que en Gauss, y se lleva a la diagonal. Matricialmente, este paso equivale a multiplicar $P_k A_k$.

PASO 2: Se hacen ceros en toda la columna k excepto el elemento diagonal, restando a todas las filas, excepto la k, un múltiplo adecuado de la fila k. Esto corresponde a hacer $A_{k+1} = \tilde{E}_k P_k A_k$, con:

$$\tilde{E}_k = \begin{pmatrix} 1 & & -\frac{a_{1k}^k}{a_{kk}^k} & & 0 \\ & \ddots & \vdots & & \\ & & 1 & & \\ & & -\frac{a_{k+1,k}^k}{a_{kk}^k} & 1 & \\ & & \vdots & & \ddots \\ 0 & & -\frac{a_{nk}^k}{a_{kk}^k} & 0 & 1 \end{pmatrix} \qquad (det(\tilde{E}_k) = 1).$$

Una vez realizadas las n etapas se tiene la matriz diagonal:

$$A_{n+1} = \underbrace{\tilde{E}_n P_n \ldots \tilde{E}_2 P_2 \tilde{E}_1 P_1}_{\tilde{M}} A = \tilde{M} A = D,$$

y simultáneamente:

$$\tilde{E}_n P_n \ldots \tilde{E}_2 P_2 \tilde{E}_1 P_1 b = \tilde{M} b.$$

Teorema 6 .- *Sea $A \in M_{n\times n}$ inversible. Existe, al menos, una matriz $\tilde{M}$ inversible tal que $\tilde{M}A$ es diagonal. Además:*

$$det(\tilde{M}) = (-1)^{\text{número de cambios de filas}} = \pm 1.$$

Por tanto:

$$det(A) = (-1)^{\text{número de cambios de filas}}\, d_{11}d_{22}\ldots d_{nn}.$$

Observación 5 .- *El método de G.-J. está especialmente indicado para el cálculo de inversas (aunque se puede usar cualquier otro método para la resolución de S.E.L.).*
Dada una matriz $A \in M_{n\times n}$ inversible, si denotamos

$$A^{-1} = (u_1|u_2|\ldots|u_n), \quad I = (e_1|e_2|\ldots|e_n),$$

entonces:

$$AA^{-1} = I \;\Leftrightarrow\; A(u_1|u_2|\ldots|u_n) = (e_1|e_2|\ldots|e_n)$$

$$\Leftrightarrow\; (Au_1|Au_2|\ldots|Au_n) = (e_1|e_2|\ldots|e_n)$$

$$\Leftrightarrow\; Au_i = e_i, \quad \forall i = 1,\ldots,n.$$

Por tanto, cada columna de la matriz inversa A^{-1} es la solución de un sistema lineal con matriz A y segundo miembro el correspondiente vector de la base canónica.

Utilizando el método de G.-J., para calcular A^{-1} basta resolver los n sistemas diagonales:

$$\tilde{M}Au_i = \tilde{M}e_i, \quad \forall i = 1, \ldots, n.$$

Método de Crout:
Se basa en el siguiente resultado:

Teorema 7 .- *(Factorización de Crout)*
Sea $A = (a_{ij}) \in M_{n\times n}$ tal que las n submatrices principales sean inversibles.
Entonces existe una matriz triangular inferior $L = (l_{ij})$ con elementos diagonales $l_{ii} = 1,\ \forall i = 1, \ldots, n$, una matriz diagonal $D = (d_{ij})$ y una matriz triangular superior $U = (u_{ij})$ con elementos diagonales $u_{ii} = 1,\ \forall i = 1, \ldots, n$, tales que $A = LDU$.
Además, esta factorización es única.

El método tiene espacial interés cuando la matriz del sistema es simétrica, pues entonces:

Corolario 1 .- *Sea $A = (a_{ij}) \in M_{n\times n}$ simétrica tal que las n submatrices principales sean inversibles.*
Entonces existe una matriz triangular inferior $L = (l_{ij})$ con elementos diagonales $l_{ii} = 1,\ \forall i = 1, \ldots, n$, y una matriz diagonal $D = (d_{ij})$ tales que $A = LDL^t$.
Además, esta factorización es única.

Por tanto, en el caso simétrico, la resolución del sistema $Ax = b$ se puede sustituir por la resolución de tres sistemas sencillos (dos con matriz triangular y uno con matriz diagonal):

$$Ax = b \Leftrightarrow \underbrace{L\, D\, \underbrace{L^t x}_{y}}_{z} = b \Leftrightarrow \begin{cases} Lz = b \\ Dy = z \\ L^t x = y \end{cases}$$

Fórmulas para el cálculo de L y D :

$$\begin{cases} d_{11} = a_{11} \\ \\ l_{i1} = \dfrac{a_{i1}}{d_{11}}, \quad i = 2, \ldots, n. \end{cases}$$

Para $j = 2, \ldots, n$:

$$\begin{cases} d_{jj} = a_{jj} - \displaystyle\sum_{k=1}^{j-1} l_{jk}^2 d_{kk} \\ \\ l_{ij} = \dfrac{a_{ij} - \displaystyle\sum_{k=1}^{j-1} l_{ik} u_{jk} d_{kk}}{d_{jj}}, \quad i = j+1, \ldots, n. \end{cases}$$

Como aplicación directa del método de Crout se tiene que:

$$det(A) = det(D) = d_{11} d_{22} \ldots d_{nn}.$$

Método de Choleski:
Si A es una matriz simétrica y definida positiva está en las condiciones del teorema de factorización LU y por tanto admite una factorización de ese tipo. Sin embargo, es posible encontrar otra factorización aún más simple:

Teorema 8 .- *(Factorización de Choleski)*
Sea $A = (a_{ij}) \in M_{n\times n}$ simétrica y definida positiva. Entonces existe una matriz triangular inferior $B = (b_{ij})$ tal que $A = BB^t$.
Además, se puede imponer que los elementos diagonales $b_{ii} > 0,\ \forall i = 1, \ldots, n$. En este caso, la factorización es única.

Corolario 2 .- *Sea $A = (a_{ij}) \in M_{n\times n}$ simétrica. Entonces son equivalentes:*

1. *A es definida positiva ($x^t Ax > 0, \quad \forall x \neq 0$).*
2. *Las submatrices principales verifican:*

$$det(\Delta_k) > 0, \quad \forall k = 1, \ldots, n.$$

3. *A admite factorización de Choleski.*
4. *$Sp(A) \subset (0, \infty)$.*

Por tanto, en el caso definido positivo, la resolución del sistema $Ax = b$ se puede sustituir por la resolución de dos sistemas triangulares:

$$Ax = b \;\Leftrightarrow\; B\underbrace{B^t x}_{y} = b \;\Leftrightarrow\; \begin{cases} By = b \\ B^t x = y \end{cases}$$

Fórmulas para el cálculo de B :

$$\begin{cases} b_{11} = \sqrt{a_{11}} \\ \\ b_{i1} = \dfrac{a_{i1}}{b_{11}}, \quad i = 2, \ldots, n. \end{cases}$$

Para $j = 2, \ldots, n$:

$$\begin{cases} b_{jj} = \sqrt{a_{jj} - \sum_{k=1}^{j-1} b_{jk}^2} \\ \\ b_{ij} = \dfrac{a_{ij} - \sum_{k=1}^{j-1} b_{ik} b_{jk}}{b_{jj}}, \quad i = j+1, \ldots, n. \end{cases}$$

Como aplicación directa del método de Choleski se tiene que:

$$det(A) = det(B)\, det(B^t) = [det(B)]^2 = b_{11}^2 b_{22}^2 \ldots b_{nn}^2.$$

2.3 Factorización QR y método de Householder.-

Dado un vector $v \in R^n,\ v \neq 0$, se llama **matriz de Householder** a una matriz de la forma:

$$H(v) = I - 2\,\frac{vv^t}{v^t v} = I - \frac{2}{\|v\|_2^2}\,vv^t \in M_{n\times n}(R).$$

(Por convenio se suele considerar $H(0) = I$.)

Se puede comprobar que toda matriz de Householder es simétrica y ortogonal, por tanto, ella es su propia inversa.

Teorema 9 .- *(Householder)*
Sea $a = (a_i) \in R^n$ *un vector tal que* $\sum_{i=2}^{n} |a_i| > 0$. *Entonces existen, al menos, dos matrices de Householder tal que las* $(n-1)$ *últimas componentes del vector resultante de multiplicarlas por* a *son nulos. De forma más precisa:*

$$H(a + \|a\|_2 e_1)a = -\|a\|_2 e_1,$$
$$H(a - \|a\|_2 e_1)a = \|a\|_2 e_1.$$

Observación 6 .-

1. *Si* $\sum_{i=2}^{n} |a_i| = 0$, *entonces basta tomar* I.
2. *En todos los casos se puede encontrar una matriz de Householder tal que la primera componente del producto sea positiva.*

3. *En la práctica no se calcula $H(v)$ sino directamente $H(v)a$:*

$$H(v)a = Ia - \frac{2}{\|v\|_2^2} vv^t a = a - 2\,\frac{v^t a}{\|v\|_2^2}\, v.$$

4. *A la hora de la programación en ordenador, la elección del signo de $v = a \pm \|a\|_2 e_1$ se suele hacer adecuada para evitar divisiones por números demasiado pequeños:*

$$v = a + \|a\|_2 e_1, \qquad si \;\; a_1 > 0,$$

$$v = a - \|a\|_2 e_1, \qquad si \;\; a_1 < 0.$$

El **método de Householder** para resolver un sistema $Ax = b$ consiste en encontrar $(n-1)$ matrices de Householder $H_1, H_2, \ldots, H_{n-1}$ tales que la matriz $H_{n-1} \ldots H_2 H_1 A = R$ sea triangular superior.
A continuación se resuelve el sistema triangular equivalente:

$$H_{n-1} \ldots H_2 H_1 Ax = H_{n-1} \ldots H_2 H_1 b$$

por sustitución retrógrada.

Entonces, partiendo de $A_1 = A$, en cada etapa k-ésima, $k = 1, \ldots, n-1,$ se van calculando las matrices $A_{k+1} = H_k A_k$ de forma que se hagan ceros en la columna k por debajo de la diagonal.

El paso de la matriz A_k a la matriz A_{k+1} se realiza entonces de la siguiente manera:
Se toma el vector $a_k \in R^{n-k+1}$ formado por los elementos de la columna k de A_k a partir del diagonal (inclusive). Se elige la matriz de Householder $H(\tilde{v}_k) \in M_{(n-k+1)\times(n-k+1)}(R)$ tal que $H(\tilde{v}_k)a_k = \|a_k\|_2 e_1$.
Se construye entonces la nueva matriz de Householder:

$$H_k = \left(\begin{array}{c|c} I_{k-1} & 0 \\ \hline 0 & H(\tilde{v}_k) \end{array}\right) = H(v_k), \quad \text{con } v_k = \begin{pmatrix} 0 \\ \tilde{v}_k \end{pmatrix} \in R^n.$$

Matricialmente, la etapa k-ésima equivale a multiplicar $A_{k+1} = H_k A_k$. Entonces la matriz A_{k+1} tiene ceros en la columna k por debajo de la diagonal.

Una vez realizadas las $(n-1)$ etapas se tiene la matriz triangular superior:

$$A_n = \underbrace{H_{n-1} \dots H_2 H_1}_{Q^t} A = Q^t A = R,$$

y, simultáneamente, el vector:

$$H_{n-1} \dots H_2 H_1 b = Q^t b.$$

Dado que:

$$Q = (H_{n-1} \dots H_2 H_1)^t = H_1^t H_2^t \dots H_{n-1}^t = H_1 H_2 \dots H_{n-1},$$

Q es ortogonal (por ser producto de matrices ortogonales), esto es, $Q^{-1} = Q^t$. En consecuencia:

$$Q^t A = R \;\Leftrightarrow\; A = QR.$$

Todo lo anterior puede resumirse en el siguiente resultado:

Teorema 10 .- *(Factorización QR)*
Dada una matriz $A \in M_{n\times n}(R)$ inversible existen, al menos, una matriz ortogonal Q y una matriz triangular superior R tales que $A = QR$.
Además, se puede elegir R con los elementos diagonales $r_{ii} > 0,\ \forall i = 1, \ldots, n$. En este caso, la factorización es única.

Observación 7 .- *Como una primera consecuencia de la factorización QR se tiene la siguiente aplicación para el cálculo de determinantes:*

$$det(A) = \underbrace{det(Q)}_{\pm 1} det(R) = \pm\, det(R) = \pm\, r_{11} r_{22} \ldots r_{nn}.$$

El método de Householder para la resolución del S.E.L. $Ax = b$ consiste entonces en calcular la factorización QR de A (mediante las matrices de Householder) y, teniendo en cuenta las equivalencias:

$$Ax = b \ \Leftrightarrow\ Q^t Ax = Q^t b \ \Leftrightarrow\ Rx = Q^t b$$

resolver el sistema triangular equivalente $Rx = Q^t b$.

3 METODOS ITERATIVOS PARA S.E.L.

En general, los sistemas lineales de gran tamaño $Ax = b$, con $A \in M_{n\times n}(R)$ inversible, que surgen en la práctica suelen tener la matriz con muchos ceros (matriz hueca o *sparse*). Los métodos directos de resolución no suelen resultar ventajosos en este caso ya que a lo largo del proceso de eliminación muchos de los coeficientes nulos de A dejan de serlo, elevando notablemente el gasto de memoria en el ordenador.

La siguiente tabla da el desarrollo a lo largo del tiempo de lo que era considerado el tamaño límite "más grande" que podía ser tratado por los métodos directos:

año	tamaño
1950	20
1965	200
1980	2000
1995	20000

En contraste con estos métodos existen otros que sólo hacen uso de la matriz original A, son los llamados **métodos iterativos** (o indirectos), y que se caracterizan por construir, a partir de un vector inicial x_0 arbitrario, una sucesión de vectores $\{x_k\}_{k\in N}$ destinada a converger a la solución del sistema.

3.1 Generalidades.-

Los métodos iterativos que vamos a estudiar son los llamados **métodos lineales** que construyen la sucesión $\{x_k\}_{k\in N}$ mediante el siguiente esquema lineal:

$$\begin{cases} x_0 \quad \text{arbitrario,} \\ x_{k+1} = Bx_k + c, \quad k = 0, 1, \ldots \end{cases}$$

donde la matriz $B \in M_{n\times n}(R)$ (*matriz del método*) y el vector $c \in R^n$ (*vector del método*) se eligen a partir de los datos A y b.

Un método lineal se dice **convergente** si cualquiera que sea el vector inicial $x_0 \in R^n$ la sucesión es convergente, esto es:

$$\exists \lim_{k\to\infty} x_k = x.$$

Supongamos que el método es convergente, entonces tomando límites en la expresión del método se tiene:

$$x = \lim_{k\to\infty} x_{k+1} = \lim_{k\to\infty} (Bx_k + c) = Bx + c,$$

esto es:

$$(I - B)x = c.$$

Por tanto, B y c deben elegirse de tal modo que el sistema $(I - B)x = c$ tenga solución (es decir, $(I - B)$ sea inversible) y los sistemas $Ax = b$ e $(I - B)x = c$ sean equivalentes.

Tenemos entonces la siguiente relación:

$$x_k - x = (Bx_{k-1} + c) - (Bx + c) = B(x_{k-1} - x)$$
$$= B^2(x_{k-2} - x) = \ldots = B^k(x_0 - x).$$

Como consecuencia se tiene el siguiente resultado de caracterización de la convergencia para los métodos iterativos lineales:

Teorema 11 .- *Son equivalentes:*

1. *El método lineal es convergente.*
2. $\rho(B) < 1$.
3. $\|B\| < 1$, *para, al menos, alguna norma matricial subordinada.*

La cuestión es ahora la siguiente: ¿qué método iterativo (es decir, qué matriz B) elegir para resolver el sistema $Ax = b$?

Si la matriz B es normal ($BB^t = B^tB$) entonces se tiene que $\|B\|_2 = \rho(B)$. Por tanto:

$$\|x_k - x\|_2 = \|B^k(x_0 - x)\|_2 \leq \|B^k\|_2\|x_0 - x\|_2$$
$$\leq \|B\|_2^k\|x_0 - x\|_2 = [\rho(B)]^k\|x_0 - x\|_2$$

Se observa que en este caso el método convergerá más rápidamente cuanto menor sea $\rho(B) < 1$.

(La conclusión es similar para matrices generales).

En resumen, en el estudio de los métodos iterativos lineales para la resolución de un sistema $Ax = b$ se debe:

1. Comprobar que $(I - B)$ es inversible y que los sistemas $Ax = b$ e $(I - B)x = c$ son equivalentes.
   (*Método bien construido*).
2. Comprobar que $\rho(B) < 1$ ó que $\|B\| < 1$.
   (*Método convergente*).
3. Elegir entre los métodos posibles el que tenga menor $\rho(B)$.
   (*Velocidad de convergencia alta*).

3.2 Métodos iterativos clásicos.-

Los métodos clásicos que vamos a describir son los representantes más característicos de una clase general de métodos basados en una descomposición de la matriz A en la forma $A = M - N$, donde M es una matriz "fácilmente" inversible (triangular, diagonal ...)

Este hecho es fundamental a la hora de llevar el método a la práctica pues, aunque no calcularemos M^{-1}, sí resolveremos sistemas lineales con matriz M.

Se tienen las siguientes equivalencias:

$$Ax = b \Leftrightarrow (M - N)x = b \Leftrightarrow Mx = Nx + b$$

$$\Leftrightarrow x = M^{-1}Nx + M^{-1}b$$

Comparando con la expresión general:

$$x = Bx + c$$

esto nos induce a elegir:

$$B = M^{-1}N, \qquad c = M^{-1}b.$$

Además, en este caso la matriz:

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

$$= M^{-1}(M - N) = M^{-1}A$$

es inversible, por serlo M y A.

Esto da lugar al método lineal general:

$$\begin{cases} x_0 \quad \text{arbitrario,} \\ x_{k+1} = M^{-1}Nx_k + M^{-1}b, \quad k = 0, 1, \ldots \end{cases}$$

o equivalentemente (y tal como haremos en la práctica):

$$\begin{cases} x_0 \quad \text{arbitrario,} \\ Mx_{k+1} = Nx_k + b, \quad k = 0, 1, \ldots \end{cases}$$

Esto es, para calcular un iterante a partir del anterior no se calculan inversas o productos de matrices, sino que simplemente se debe resolver un sistema lineal sencillo.

Veamos a continuación los métodos de Jacobi, de Gauss-Seidel y de Relajación. Para ello, descomponemos la matriz del sistema en la forma $A = D - E - F$, con:

$$D = \begin{pmatrix} a_{11} & & & 0 \\ & a_{22} & & \\ & & \ddots & \\ 0 & & & a_{nn} \end{pmatrix},$$

$$E = \begin{pmatrix} 0 & & & 0 \\ -a_{21} & 0 & & \\ \vdots & \ddots & \ddots & \\ -a_{n1} & \ldots & -a_{n,n-1} & 0 \end{pmatrix}, \quad F = \begin{pmatrix} 0 & -a_{12} & \ldots & -a_{1n} \\ & \ddots & \ddots & \vdots \\ & & 0 & -a_{n-1,n} \\ 0 & & & 0 \end{pmatrix}.$$

Supondremos que D es inversible, esto es, que se verifica $a_{ii} \neq 0, \ \forall i = 1, \ldots, n$. (Esta hipótesis no es excluyente, pues siempre se puede encontrar una reordenación de las filas de A que la verifique).

Método de Jacobi:
Se elige:

$$M = D, \qquad \text{(matriz diagonal)}$$

$$N = E + F.$$

La matriz del método es:

$$J = D^{-1}(E + F) \quad \text{(Matriz de Jacobi)}$$

y el vector del método es:

$$c = D^{-1}b.$$

Por tanto, el método quedará:

$$x_{k+1} = D^{-1}(E + F)x_k + D^{-1}b, \quad k = 0, 1, \ldots$$

o equivalentemente:

$$Dx_{k+1} = (E + F)x_k + b, \quad k = 0, 1, \ldots$$

Si denotamos por x_k^i la coordenada i-ésima del iterante x_k, entonces se tiene la expresión:

$$x_{k+1}^i = \frac{1}{a_{ii}}[-\sum_{j=1}^{i-1} a_{ij}x_k^j - \sum_{j=i+1}^{n} a_{ij}x_k^j + b_i],$$
$$i = 1, \ldots, n.$$
$$k = 0, 1, \ldots$$

Método de Gauss-Seidel:
Se elige:

$$M = D - E, \qquad \text{(matriz triagular inferior)}$$

$$N = F.$$

La matriz del método es:

$$\mathcal{L}_1 = (D - E)^{-1}F \quad \text{(Matriz de Gauss-Seidel)}$$

y el vector del método es:

$$c_1 = (D - E)^{-1}b.$$

Por tanto, el método quedará:

$$x_{k+1} = (D - E)^{-1}Fx_k + (D - E)^{-1}b, \quad k = 0, 1, \ldots$$

o equivalentemente:

$$(D - E)x_{k+1} = Fx_k + b, \quad k = 0, 1, \ldots$$

Se tiene entonces la expresión en coordenadas:

$$x_{k+1}^i = \frac{1}{a_{ii}}[-\sum_{j=1}^{i-1} a_{ij}x_{k+1}^j - \sum_{j=i+1}^{n} a_{ij}x_k^j + b_i],$$

$$i = 1, \ldots, n.$$

$$k = 0, 1, \ldots$$

Método de Relajación:

Se introduce el parámetro real $\omega \neq 0$ para hacer una descomposición $A = M_\omega - N_\omega$ de modo que la matriz del método tenga el menor radio espectral posible:

$$M_\omega = \frac{1}{\omega}D - E, \qquad \text{(matriz triagular inferior)}$$
$$N_\omega = \frac{1-\omega}{\omega}D + F.$$

La matriz y el vector del método son:

$$\mathcal{L}_\omega = (\frac{1}{\omega}D - E)^{-1}(\frac{1-\omega}{\omega}D + F) = (D - \omega E)^{-1}[(1-\omega)D + \omega F],$$
$$c_\omega = (\frac{1}{\omega}D - E)^{-1}b = (D - \omega E)^{-1}\omega b.$$

Por tanto, el método quedará:

$$x_{k+1} = (D - \omega E)^{-1}[(1-\omega)D + \omega F]x_k + (D - \omega E)^{-1}\omega b, \quad k = 0, 1, \ldots$$

o equivalentemente:

$$(D - \omega E)x_{k+1} = [(1-\omega)D + \omega F]x_k + \omega b, \quad k = 0, 1, \ldots$$

Se tiene entonces la expresión en coordenadas:

$$x^i_{k+1} = (1-\omega)x^i_k + \frac{\omega}{a_{ii}}[-\sum_{j=1}^{i-1} a_{ij}x^j_{k+1} - \sum_{j=i+1}^{n} a_{ij}x^j_k + b_i],$$
$$i = 1, \ldots, n.$$
$$k = 0, 1, \ldots$$

(Nótese que Gauss-Seidel es Relajación para $\omega = 1$.)

3.3 Convergencia de los métodos clásicos.-

Por el teorema general de convergencia, los métodos clásicos convergen si y sólo si $\rho(M^{-1}N) < 1$, esto es, si todas las raíces de $det(M^{-1}N - \lambda I) = 0$ tienen módulo menor que uno.

Teniendo en cuenta que $det(M) \neq 0$, esto es equivalente a que todas las raíces de $det(\lambda M - N) = 0$ tengan módulo menor que uno, que es una condición más sencilla de comprobar.

Aún así, esta condición no es aplicable en la práctica por sus dificultades de cálculo. Nos interesa, por tanto, obtener condiciones de convergencia más fáciles de comprobar (y a ser posible, sobre la matriz A del sistema).

Teorema 12 .- *(Fröbenius-Mises)*
Una condición suficiente para la convergencia del método de Jacobi es que se verifique una de las dos desigualdades siguientes:

1. $$\sum_{\substack{j=1 \\ j\neq i}}^{n} |\frac{a_{ij}}{a_{ii}}| < 1, \quad \forall i = 1, \ldots, n.$$

2. $$\sum_{\substack{i=1 \\ i\neq j}}^{n} |\frac{a_{ij}}{a_{ii}}| < 1, \quad \forall j = 1, \ldots, n.$$

Teorema 13 .- *(Geiringer)*
Una condición suficiente para la convergencia del método de Gauss-Seidel es que se verifique una de las dos desigualdades siguientes:

1. $$\sum_{\substack{j=1 \\ j\neq i}}^{n} |\frac{a_{ij}}{a_{ii}}| < 1, \quad \forall i = 1, \ldots, n.$$

2. $$\sum_{\substack{i=1 \\ i\neq j}}^{n} |\frac{a_{ij}}{a_{ii}}| < 1, \quad \forall j = 1, \ldots, n.$$

Observación 8 .- *La condición 1 puede escribirse también en la forma equivalente:*

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

es decir, A es una matriz de diagonal estrictamente dominante.
Entonces, para toda matriz A de diagonal estrictamente dominante (por tanto inversible) los métodos de Jacobi y Gauss-Seidel son convergentes.

Teorema 14 .- *(Kahan)*
El radio espectral de la matriz del Relajación verifica:

$$\rho(\mathcal{L}_\omega) \geq |1 - \omega|.$$

Por tanto, una condición necesaria para la convergencia del método de Relajación es que $\omega \in (0, 2)$.

Teorema 15 .- *(Ostrowski-Reich)*
Si A es simétrica y definida positiva, una condición necesaria y suficiente para la convergencia del método de Relajación es que $\omega \in (0, 2)$.
(Por tanto, el método de Gauss-Seidel es simpre convergente para una matriz A simétrica y definida positiva).

Observación 9 .- *Este último resultado es consecuencia de un resultado más general según el cual, si consideramos la descomposición $A = M - N$ de una matriz A simétrica y definida positiva, una condición suficiente para la convergencia del método asociado es que la matriz $(M^t + N)$ sea también simétrica y definida positiva.*

Para finalizar el estudio de los métodos iterativos, veremos algunos resultados para la comparación de la velocidad de convergencia de los métodos clásicos.
Para ello comenzaremos restringiéndonos al caso de una matriz A tridiagonal (caso frecuente en la práctica), y veremos a continuación otro resultado para matrices más generales:

Teorema 16 .- *Si A es una matriz tridiagonal tal que la matriz de Jacobi J tiene todos sus autovalores reales, entonces los métodos de Jacobi y de Relajación para $\omega \in (0,2)$ convergen o divergen simultáneamente.*
En caso de convergencia, existe un parámetro óptimo de Relajación:

$$\omega_0 = \frac{2}{1+\sqrt{1-[\rho(J)]^2}} \in (1,2)$$

tal que:

$$\rho(\mathcal{L}_{\omega_0}) = \omega_0 - 1 < \rho(\mathcal{L}_1) = [\rho(J)]^2 < \rho(J) < 1$$

Teorema 17 .- *(Stein-Rosenberg)*
Sea A una matriz tal que la matriz de Jacobi J tiene todos sus coeficientes no negativos. Entonces se cumple una de las cuatro condiciones siguientes:

1. $\rho(\mathcal{L}_1) = \rho(J) = 0$
2. $0 < \rho(\mathcal{L}_1) < \rho(J) < 1$
3. $\rho(J) = \rho(\mathcal{L}_1) = 1$
4. $1 < \rho(J) < \rho(\mathcal{L}_1)$

(Esto es, los métodos de Jacobi y de Gauss-Seidel convergen o divergen simultáneamente. En caso de convergencia, el método de G.-S. es más rápido).

4 METODOS BASADOS EN ALGORITMOS DE DESCENSO

Para resolver el sistema de ecuaciones lineales $Ax = b$ con $A \in M_{n\times n}$ simétrica y definida positiva, se introduce la función:

$$J : y \in R^n \to J(y) = \frac{1}{2}\, y^t A y - y^t b \in R,$$

que verifica $\nabla J(y) = Ay - b, \quad \forall y \in R^n$.

Teorema 18 .-

1. *J es estictamente convexa en R^n.*
2. *J admite un único mínimo x en R^n.*
3. $\nabla J(x) = 0$.
4. *x es la única solución del sistema $Ax = b$.*

Por tanto, resolver el S.E.L. es equivalente a resolver el problema de minimización:

$$(P) \quad \begin{cases} \text{Hallar } \ x \in R^n \ \text{ tal que:} \\ J(x) < J(y), \quad \forall y \in R^n, \ y \neq x. \end{cases}$$

Vamos entonces a comentar brevemente los *métodos de descenso* para resolver el problema (P):
Dado $y \in R^n$, se llama **dirección de descenso en** y a un vector $d \in R^n$ tal que $J(y + \rho d) < J(y)$, para un **paso** ρ suficientemente pequeño.

Una condición suficiente para que d sea una dirección de descenso es que:

$$d^t \nabla J(y) < 0.$$

Los algoritmos de descenso se caracterizan fundamentalmente por la construcción de una sucesión de iterantes donde, para pasar de un iterante x^k al siguiente x^{k+1}, se eligen una dirección de descenso d^k y un paso ρ^k que aseguren un decrecimiento de J, esto es, tales que para $x^{k+1} = x^k + \rho^k d^k$ se verifique $J(x^{k+1}) < J(x^k)$.

Esquema general de los algoritmos de descenso:

1. Elegir $x^0 \in R^n$. $k = 0$.
2. Si $\nabla J(x^k) = 0$: Parar ($\Rightarrow$ $x = x^k$).
3. Elegir una dirección de descenso d^k.
4. Elegir un paso ρ^k tal que $J(x^k + \rho^k d^k) < J(x^k)$.
5. Actualizar $x^{k+1} = x^k + \rho^k d^k$. $k \to k+1$.
6. Test de parada (por ej., $\|x^{k+1} - x^k\| < \varepsilon$).
   (a) Si se satisface: Parar ($\Rightarrow$ $x \simeq x^{k+1}$).
   (b) Si no se satisface: Volver a 2.

Observación 10 .- *Ya hemos visto un criterio que nos permite elegir la dirección de descenso. Para la elección del paso se tienen las reglas de Cauchy, de Armijo, de Goldstein, ...*
Todas ellas se basan en calcular ρ^k *como el mínimo de la función real* $\phi^k(\rho) = J(x^k + \rho d^k)$, *esto es, calcular* ρ^k *tal que sea solución de la ecuación:*

$$\frac{d\phi^k}{d\rho}(\rho^k) = 0.$$

Las distintas reglas proporcionan diferentes aproximaciones a esta solución.

Veamos a continuación las dos clases de métodos de descenso más simples:

4.1 Métodos de gradiente.-

Los métodos de gradiente se caracterizan por tomar como dirección de descenso:

$$d^k = -\nabla J(x^k) = b - Ax^k,$$

pues es ese caso:

$$(d^k)^t \nabla J(x^k) = -\|\nabla J(x^k)\|_2^2 < 0.$$

Para la elección del paso ρ^k existen distintas posibilidades. Veremos lo que se conoce como **método de máximo descenso** que consiste en tomar ρ^k como la

solución exacta de la ecuación, que en este caso es fácil de calcular:

$$\phi^k(\rho) = J(x^k + \rho d^k) = \frac{1}{2}(x^k)^t A x^k + \frac{1}{2}\rho (x^k)^t A d^k$$

$$+\frac{1}{2}\rho (d^k)^t A x^k + \frac{1}{2}\rho^2 (d^k)^t A d^k - (x^k)^t b - \rho (d^k)^t b.$$

$$\Rightarrow \frac{d\phi^k}{d\rho}(\rho) = \frac{1}{2}(x^k)^t A d^k + \frac{1}{2}(d^k)^t A x^k + \rho (d^k)^t A d^k - (d^k)^t b.$$

Entonces:

$$\frac{d\phi^k}{d\rho}(\rho^k) = 0 \Leftrightarrow \rho^k = -\frac{\frac{1}{2}(x^k)^t A d^k + \frac{1}{2}(d^k)^t A x^k - (d^k)^t b}{(d^k)^t A d^k}.$$

Como A es simétrica: $(x^k)^t A d^k = [(x^k)^t A d^k]^t = (d^k)^t A x^k$ y por tanto:

$$\rho^k = -\frac{(d^k)^t A x^k - (d^k)^t b}{(d^k)^t A d^k} = -\frac{(d^k)^t [A x^k - b]}{(d^k)^t A d^k} = \frac{(d^k)^t d^k}{(d^k)^t A d^k}.$$

Así pues, el método de máximo descenso consiste en, partiendo de un iterante inicial $x^0 \in R^n$, construir la sucesión $\{x^k\}_{k \in N}$ de la forma:

$$x^{k+1} = x^k + \rho^k d^k, \quad k = 0, 1, \ldots$$

tomando:

$$d^k = b - A x^k, \qquad \rho^k = \frac{(d^k)^t d^k}{(d^k)^t A d^k}.$$

4.2 El método de gradiente conjugado.-

Los métodos de gradiente convergen, en general, de manera muy lenta; por lo cual en la práctica se suelen utilizar diferentes modificaciones basadas en esos métodos. La más sencilla es el **método de gradiente conjugado** donde, partiendo de un iterante inicial $x^0 \in R^n$, se minimiza J siguiendo n direcciones $d^0, d^1, \ldots, d^{n-1}$ linealmente independientes y mutuamente A-conjugadas, es decir:

$$(d^j)^t A d^i = 0, \quad \forall j \neq i.$$

Las direcciones d^k se construyen, como veremos más adelante, combinando las direcciones previas con el gradiente de J en el punto x^k.
Calculando ρ^k de forma que se minimice la función ϕ^k, se definen:

$$x^{k+1} = x^k + \rho^k d^k, \quad k = 0, 1, \ldots, n-1.$$

Dado que las n direcciones de descenso son linealmente independientes forman una base de R^n y, por tanto, siguiendo esas n direcciones podemos alcanzar el mínimo de J.
Puede probarse entonces que el iterante x^n minimiza J, esto es:

$$A\, x^n = b.$$

Por tanto, si no hubiese errores de redondeo, se alcanzaría la solución exacta en n iteraciones.

El esquema del método es como sigue:

$$x^0 \in R^n \quad \text{arbitrario,}$$

$$\begin{cases} r^0 = b - Ax^0 \\ d^0 = r^0 \\ \rho^0 = \dfrac{(r^0)^t r^0}{(d^0)^t A d^0} \\ \\ x^1 = x^0 + \rho^0 d^0 \end{cases}$$

$$\text{Para} \;\; k = 1, \ldots, n-1 \;\; :$$

$$\begin{cases} r^k = r^{k-1} - \rho^{k-1} A d^{k-1} \\ \beta^k = \dfrac{(r^k)^t r^k}{(r^{k-1})^t r^{k-1}} \\ \\ d^k = r^k + \beta^k d^{k-1} \\ \\ \rho^k = \dfrac{(r^k)^t r^k}{(d^k)^t A d^k} \\ \\ x^{k+1} = x^k + \rho^k d^k \end{cases}$$

Observación 11 .- *Aún teniendo convergencia en n iteraciones, si n es grande el resultado no es muy satisfactorio.*
En estos casos el método de gradiente conjugado, sin ser estrictamente un método iterativo (ya hemos visto que puede encuadrarse dentro de los que denominamos métodos directos), puede considerarse como un método iterativo mediante la introducción de un test de parada que nos permita finalizar el proceso antes de la iteración n siempre que el error sea suficientemente pequeño.

4.3 El método de gradiente conjugado precondicionado.-

Para acelerar la convergencia del método de gradiente conjugado en la resolución de $Ax = b$ se pueden utilizar técnicas de precondicionamiento.
Para ello, tomaremos como precondicionador una matriz S simétrica y definida positiva (puede ser elegida a partir de A por diferentes métodos). Entonces, se resuelve utilizando gradiente conjugado el sistema equivalente:

$$(SAS)(S^{-1}x) = Sb$$

que tiene también matriz simétrica y definida positiva.
La elección de un buen precondicionador puede tener efectos espectaculares sobre la velocidad de convergencia del algoritmo.

En la práctica los cálculos no se hacen con la matriz S, sino con la matriz $M = S^2$, también definida positiva.

El esquema del método sería de la siguiente forma:

$$x^0 \in R^n \quad \text{arbitrario,}$$

$$\begin{cases} r^0 = b - Ax^0 \\ z^0 = Mr^0 \\ d^0 = z^0 \\ \rho^0 = \dfrac{(r^0)^t z^0}{(d^0)^t A d^0} \\ \\ x^1 = x^0 + \rho^0 d^0 \end{cases}$$

$$\text{Para} \;\; k = 1, \ldots, n-1 \;\; :$$

$$\begin{cases} r^k = r^{k-1} - \rho^{k-1} A d^{k-1} \\ z^k = Mr^k \\ \beta^k = \dfrac{(r^k)^t z^k}{(r^{k-1})^t z^{k-1}} \\ \\ d^k = z^k + \beta^k d^{k-1} \\ \\ \rho^k = \dfrac{(r^k)^t z^k}{(d^k)^t A d^k} \\ \\ x^{k+1} = x^k + \rho^k d^k \end{cases}$$

5 METODOS PARA SISTEMAS NO LINEALES

En general se trata de resolver el problema $F(x) = 0$, donde:

$$F : D \subset R^n \to R^n$$

es una función arbitraria. Esto es, el problema es hallar $\alpha \in D$ tal que $F(\alpha) = 0$.

Ejemplo 2 .- *Supongamos el sistema no lineal:*

$$\begin{cases} 4x_1^4 - 27x_1x_2^2 - sen(x_2) &= 4 \\ x_1x_2 + 13\,log(x_1) - x_2^2 &= 0 \end{cases}$$

Entonces, trabajando con:

$$F : (x_1, x_2) \in (0, \infty) \times R \subset R^2 \to F(x_1, x_2) \in R^2,$$

donde:

$$F(x_1, x_2) = (4x_1^4 - 27x_1x_2^2 - sen(x_2) - 4, x_1x_2 + 13\,log(x_1) - x_2^2)$$

el problema se puede escribir:

$$F(x_1, x_2) = (0, 0).$$

Se utilizan métodos que generalizan los expuestos para el caso de una única ecuación. Veremos los dos métodos más simples:

5.1 Algoritmos de iteración funcional.-

Se basan en escribir la ecuación $F(x) = 0$ en la forma equivalente $x = f(x)$ y considerar un algoritmo iterativo de búsqueda de puntos fijos, consistente en la construcción de una sucesión de vectores $\{x^k\}_{k\in N}$ en la forma siguiente:

$$x^0 \quad \text{dado},$$
$$x^{n+1} = f(x^n), \quad n = 0, 1, 2, \ldots$$

Tenemos el siguiente resultado de convergencia global:

Teorema 19 .- *Sea D un abierto acotado de R^n y sea $f : \bar{D} \to R^n$ verificando:*

1. *$f(\bar{D}) \subset \bar{D}$.*
2. *$\exists L \in [0, 1) \,/\, \|f(x)-f(y)\| \leq L\|x-y\|$, $\forall x, y \in \bar{D}$, esto es, f es contractiva en $\bar{D}$ de constante L.*

Entonces, f tiene un único punto fijo α en $\bar{D}$.

Además, cualquier sucesión $\{x^k\}_{k\in N}$ definida por:

$$x^0 \in \bar{D},$$
$$x^{n+1} = f(x^n), \quad n = 0, 1, \ldots$$

converge a α y el error en el iterante n-ésimo verifica la siguiente acotación:

$$\|x^n - \alpha\| \leq \frac{L^n}{1-L}\|x^1 - x^0\|, \quad n = 1, 2, \ldots$$

Para asegurar la contractividad de f se tiene el siguiente resultado:

Teorema 20 .- *Sea D un abierto acotado y convexo de R^n y sea $f : \bar{D} \to R^n$ verificando:*

1. *f derivable en $\bar{D}$.*
2. $\dfrac{\partial f_i}{\partial x_j}$ *continuas en D,* $\quad \forall i, j = 1, \ldots, n.$
3. $\|Df(x)\| \leq L, \quad \forall x \in D,$ *siendo el jacobiano:*

$$Df(x) = \begin{pmatrix} \frac{\partial f_1}{\partial x_1}(x) & \ldots & \frac{\partial f_1}{\partial x_n}(x) \\ \ldots & \ldots & \ldots \\ \frac{\partial f_n}{\partial x_1}(x) & \ldots & \frac{\partial f_n}{\partial x_n}(x) \end{pmatrix}.$$

Entonces, f es contractiva en $\bar{D}$ de constante L.

5.2 Método de Newton.-

Para resolver la ecuación $F(x) = 0$, se construye una sucesión de vectores $\{x^k\}_{k \in N}$ mediante:

$$x^0 \quad \text{dado},$$
$$x^{n+1} = x^n - [DF(x^n)]^{-1} F(x^n), \quad n = 0, 1, \ldots$$

Esta expresión es, en principio, bastante costosa desde el punto de vista numérico, pues implica la inversión de una matriz diferente para el cálculo de cada iterante.

Pero debemos tener en cuenta que no es necesario calcular la inversa del jacobiano, pues se puede utilizar la siguiente estrategia:

$$\Delta x^n = x^{n+1} - x^n = -[DF(x^n)]^{-1} F(x^n)$$

$$\Leftrightarrow \; DF(x^n)\, \Delta x^n = -F(x^n).$$

Entonces, dado x^n se calcula Δx^n resolviendo el S.E.L. anterior (por cualquiera de los métodos ya vistos en los apartados anteriores) y a continuación se actualiza x^{n+1} mediante la expresión:

$$x^{n+1} = x^n + \Delta x^n.$$

Para finalizar, daremos el siguiente resultado de convergencia local para el método de Newton:

Teorema 21 .- *Sea* $F : \bar{D} \subset R^n \to R^n$ *y sea* $\alpha \in \bar{D}$ *una raíz de la ecuación* $F(x) = 0$.

1. *Si* DF *es continua en* α *y* $DF(\alpha)$ *es inversible, entonces para* x^0 *suficientemente próximo a* α *se tiene que la sucesión* $\{x^k\}_{k\in N}$ *construida por el método de Newton converge a* α, *y verifica:*

$$\lim_{n\to\infty} \frac{\|x^{n+1} - \alpha\|}{\|x^n - \alpha\|} = 0.$$

   (Por tanto, el orden de convergencia es $p > 1$.*)*

2. *Además, si existe una constante* λ *tal que:*

$$\|DF(x) - DF(\alpha)\| \leq \lambda \|x - \alpha\|$$

   para todo x *en un entorno de* α, *entonces existe otra constante* c *tal que:*

$$\|x^{n+1} - \alpha\| \leq c \|x^n - \alpha\|^2, \quad \forall n \geq n_0.$$

   (Por tanto, el orden de convergencia es 2.)

Ejemplo 3 .- *Sea el sistema no lineal:*

$$\begin{cases} x_1 - 0.3\,\mathrm{sen}(x_1) - 0.4\cos(x_2) &= 0 \\ x_2 - 0.3\cos(x_1) + 0.4\,\mathrm{sen}(x_2) &= 0 \end{cases}$$

Entonces, para $F : (x_1, x_2) \in R^2 \to F(x_1, x_2) \in R^2$, *dada por:*

$$F(x_1, x_2) = (x_1 - 0.3\,\mathrm{sen}(x_1) - 0.4\cos(x_2),\, x_2 - 0.3\cos(x_1) + 0.4\,\mathrm{sen}(x_2))$$

el problema se puede escribir $F(x_1, x_2) = (0, 0)$.

1. *Para resolver el sistema mediante iteración funcional, consideramos*

$$f(x_1, x_2) = (0.3\,\mathrm{sen}(x_1) + 0.4\cos(x_2),\, 0.3\cos(x_1) - 0.4\,\mathrm{sen}(x_2))$$

y expresamos el problema como $(x_1, x_2) = f(x_1, x_2)$.
Sea $D = \{(x_1, x_2) \in R^2 \ : \ |x_1| < 1,\ |x_2| < 1\}$, *entonces*

$$f(x_1, x_2) \in D, \quad \forall (x_1, x_2) \in D.$$

Esto es, se cumple que $f(\bar{D}) \subset \bar{D}$.
Además, se tiene que

$$Df(x_1, x_2) = \begin{pmatrix} 0.3\cos(x_1) & -0.4\,\mathrm{sen}(x_2) \\ -0.3\,\mathrm{sen}(x_1) & -0.4\cos(x_2) \end{pmatrix} \in C^\infty(R^2),$$

$$\|Df(x_1, x_2)\|_\infty$$
$$= \max\{|0.3\cos(x_1)| + |0.4\,\mathrm{sen}(x_2)|, |0.3\,\mathrm{sen}(x_1)| + |0.4\cos(x_2)|\}$$
$$\leq \max\{0.7, 0.7\} = 0.7, \quad \forall (x_1, x_2) \in D.$$

Esto es, f es contractiva en $\bar{D}$ de constante $L = 0.7 < 1$.

Por tanto, el método de iteración funcional

$$x^0 \in \bar{D},$$

$$x^{n+1} = f(x^n), \quad n = 0, 1, 2, \ldots$$

es globalmente convergente.

2. *Para resolver el sistema mediante Newton, se tiene en cuenta que*

$$DF(x_1, x_2) = \begin{pmatrix} 1 - 0.3\cos(x_1) & 0.4\,\mathrm{sen}(x_2) \\ 0.3\,\mathrm{sen}(x_1) & 1 + 0.4\cos(x_2) \end{pmatrix} \in C^\infty(R^2),$$

$$\begin{aligned} &|\det(DF(x_1, x_2))| = |1 - 0.3\cos(x_1) + 0.4\cos(x_2) \\ &-0.12\,[\cos(x_1)\cos(x_2) - \mathrm{sen}(x_1)\,\mathrm{sen}(x_2)]| \\ &= |1 - 0.3\cos(x_1) + 0.4\cos(x_2) - 0.12\cos(x_1 - x_2)| \\ &\geq 1 - 0.3 - 0.4 - 0.12 = 0.18 > 0, \quad \forall (x_1, x_2) \in R^2. \end{aligned}$$

Así pues, $DF(x_1, x_2)$ es inversible siempre, en particular en la solución.

Por tanto, el método de iteración funcional

$$x^0 \quad \textit{dado},$$

$$x^{n+1} = x^n - [DF(x^n)]^{-1} F(x^n), \quad n = 0, 1, 2 \ldots$$

converge localmente (esto es, para x^0 suficientemente próximo a la solución) con orden de convergencia 2.

.