Ejercicio 1.22 $p(\bar{z}) = a_n\bar{z}^n+\cdots+a_1\bar{z}+a_0 = a_n\overline{z^n}+\cdots+a_2\overline{z^2}+a_1\bar{z}+a_0$. Teniendo en cuenta que los coeficientes son reales y por tanto verifican $\bar{a_k} = a_k$ deducimos

$$p(\bar{z}) = \overline{a_n z^n + \cdots + a_1 z + a_0} = \overline{p(z)}\,.$$

Ejercicio 1.23 La satisfacción de la relación de recurrencia es una comprobación inmediata. La independencia lineal de $\{z_1^n\}$ y $\{z_2^n\}$ se reduce a la no singularidad de la matriz

$$\begin{pmatrix} 1 & z_1 \\ 1 & z_2 \end{pmatrix},$$

la cual es equivalente a $z_1 \neq z_2$.

Ejercicio 1.24

Ejercicio 1.25 $z = 1.7437 + 0.3057i$

Capítulo 2

Resolución de Sistemas de Ecuaciones Lineales

2.1 Métodos directos

La *regla de Cramer* es un método teórico de resolución directa de sistemas de ecuaciones lineales, sin embargo no lo consideramos como un método numérico por la complejidad de cálculo que conlleva. Para un sistema de orden n la regla de Cramer requiere el cálculo de $n+1$ determinantes de orden n, lo que significa realizar $(n+1)!(n-1)$ multiplicaciones: un número de operaciones excesivamente grande comparado con el requerido por los métodos que vamos a estudiar.

2.1.1 Eliminación de Gauss

El método de eliminación de Gauss es el método más sencillo de resolución de sistemas de ecuaciones lineales. La única diferencia entre este método y lo que suele hacer un estudiante principiante está en que el método de Gauss es una forma *ordenada y sistemática* de realizar la eliminación de las incógnitas, lo que da lugar a un algoritmo sencillo y eficaz, que se puede programar. Este algoritmo puede aplicarse tanto a sistemas sobredeterminados como a sistemas incompletos, pero en lo que sigue supondremos que

el sistema dado es no-singular y tiene tantas ecuaciones como incógnitas. El algoritmo consta de dos partes: (1) la *eliminación* propiamente dicha, que consiste en reducir el sistema a uno triangular superior (cuya matriz de coeficientes tiene todo ceros debajo de la diagonal); y (2) *sustitución* u obtención de la solución por el método llamado de *sustitución regresiva*. Veamos a continuación cada una de estas dos partes ilustradas con un ejemplo.

Eliminación.–

Supongamos que queremos resolver el siguiente sistema de ecuaciones lineales:

$$\begin{aligned} 10x - 7y &= 7 \\ -3x + 2y + 6z &= 4 \\ 5x - y + 5z &= 6 \end{aligned} \tag{2.1}$$

El primer paso del método de eliminación de Gauss es eliminar la primera incógnita de todas las ecuaciones que siguen a la primera, para lo cual se les resta sucesivamente la primera ecuación multiplicada por el factor adecuado a cada una (llamado *multiplicador*). Para ello, en este caso, restamos de la segunda ecuación la primera multiplicada por el *multiplicador* $-\frac{3}{10}$ y restamos de la tercera ecuación la primera multiplicada por el multiplicador $\frac{5}{10}$ con lo que el sistema se convierte en uno equivalente de la forma

$$\begin{aligned} 10x - 7y &= 7 \\ -0.1y + 6z &= 6.1 \\ 2.5y + 5z &= 2.5 \end{aligned}$$

El siguiente paso y sucesivos consisten en repetir el primer paso pero aplicado al subsistema formado por las ecuaciones que hayan sido modificadas en el paso anterior. Así pues, en nuestro ejemplo el paso siguiente es eliminar el término en y de la tercera ecuación para lo que le restamos la segunda multiplicada por $2.5/(-0.1) = -25$, de forma que nos queda

$$\begin{aligned} 10x - 7y &= 7 \\ -0.1y + 6z &= 6.1 \\ 155z &= 155\,. \end{aligned}$$

El resultado final (al cabo de $n-1$ pasos en un sistema de orden n) es haber transformado el sistema original en otro equivalente pero triangular superior.

Ejercicio 2.1
Describir un algoritmo que efectúe el proceso de eliminación que acabamos de describir para un sistema de n ecuaciones lineales con n incógnitas (de forma que lo transforme en uno equivalente pero triangular superior).

Sustitución regresiva.–

Después de haberlo reducido a forma triangular es inmediato encontrar la solución del sistema dado empezando por la última ecuación y procediendo hacia atrás. La última ecuación es ahora una ecuación lineal en una incógnita que se resuelve de forma inmediata. Llevando esta solución a la ecuación anterior podremos hallar la incógnita anterior, y así sucesivamente hasta hallar todas las incógnitas. En nuestro ejemplo,

$$\begin{aligned} z &= 1\,; \\ y &= \frac{6.1 - 6z}{-0.1} = \frac{6.1 - 6}{-0.1} = -1\,; \\ x &= \frac{7 + 7y + 0z}{10} = \frac{7 + 7(-1)}{10} = 0\,. \end{aligned}$$

Esta resolución en escalada hacia atrás se conoce como *sustitución regresiva*.

Ejercicio 2.2
Describir un algoritmo que efectúe el proceso de sustitución regresiva en un sistema de n ecuaciones lineales con n incógnitas supuesto dado en forma triangular superior.

Ejercicio 2.3
Hallar el número de operaciones (contando sólo multiplicaciones y divisiones) que es necesario realizar para resolver un sistema de n ecuaciones lineales con n incógnitas por el método de eliminación de Gauss.

Dificultades que pueden surgir en la eliminación.–

Veamos ahora dos tipos de dificultades que pueden surgir en la aplicación del método de eliminación.

En primer lugar puede ocurrir que al comienzo de alguno de los pasos el coeficiente por el que tenemos que dividir los demás de su columna para hallar los multiplicadores de ese paso sea cero, con lo cual no podremos continuar la eliminación. Por ejemplo, esto ocurre ya en el primer paso en el sistema

$$\begin{aligned} -7y + 10z &= 7 \\ 6x + 2y - 3z &= 4 \\ 5x - y + 5z &= 6 \end{aligned}$$

sin embargo es evidente que, a menos que el sistema sea singular, no pueden ser cero todos los coeficientes de la primera columna, con lo cual siempre podremos intercambiar la ecuación cuyo primer coeficiente es cero con una ecuación *posterior* cuyo primer coeficiente no sea cero (intercambio que evidentemente no afecta en nada a la solución). Este mecanismo de intercambio de ecuaciones en la búsqueda de un coeficiente adecuado para ser usado como divisor en el cálculo de los multiplicadores de un cierto paso de eliminación se llama *pivotación* y el coeficiente encontrado (que pasará a ocupar la posición $(1, 1)$ del subsistema sobre el que vamos a trabajar en ese paso) se llama el *pivote* de ese paso. Existen diversas estrategias para realizar la pivotación. La que acabamos de describir consiste en buscar el primer coeficiente no nulo en la primera columna del subsistema, estrategia conocida como *pivotación simple*. Según lo dicho,

Teorema 9 *Todo sistema no singular de n ecuaciones lineales con n incógnitas puede resolverse por el método de eliminación de Gauss con pivotación simple.* ■

Asimismo se conoce una caracterización de los sistemas que se pueden resolver por el método de eliminación de Gauss sin necesidad de realizar pivotación alguna. Esta propiedad, evidentemente, es simplemente una propiedad del orden en que decidimos tomar las ecuaciones del sistema durante el proceso de eliminación. Recordemos que los *menores principales* de una matriz son las submatrices cuadradas que se obtienen al eliminar todas las filas y columnas *a partir* de la columna k y de la fila k para un mismo k fijo.

Así, el menor principal de orden k de una matriz $A = (a_{ij})$ puede definirse como la matriz cuadrada cuyos elementos son todos los a_{ij} con $1 \leq i \leq k$ y $1 \leq j \leq k$. Dicho esto podemos enunciar:

Teorema 10 *Un sistema de n ecuaciones lineales con n incógnitas puede resolverse por el método de eliminación de Gauss sin pivotación alguna si y sólo si en la matriz de coeficientes todos los menores principales son no singulares.*

Demostración: Primeramente hay que observar que las operaciones elementales realizadas sobre la matriz de coeficientes durante el proceso de eliminación no cambian los valores de los determinantes de los menores principales. Denotemos $a_{ij}^{(k)}$ los elementos obtenidos al realizar el paso $k-1$ de eliminación. Si se puede realizar la eliminación sin pivotación entonces los elementos

$$a_{11},\ a_{22}^{(2)},\ a_{33}^{(3)},\ \ldots,\ a_{nn}^{(n)}$$

son todos distintos de cero. Además, para todo $r \in \{1, \ldots, n\}$ el menor principal de orden r, A_{rr}, tiene determinante dado por

$$\det(A_{rr}) = a_{11} \cdot a_{22}^{(2)} \cdot a_{33}^{(3)} \cdots a_{rr}^{(r)},$$

luego todos los menores principales son no singulares.

Supongamos ahora que algún menor principal sea singular. Sea r el menor de los órdenes de los menores principales que sean singulares. Entonces podemos realizar $r-1$ pasos de eliminación sin pivotación y los elementos

$$a_{11},\ a_{22}^{(2)},\ a_{33}^{(3)},\ \ldots,\ a_{r-1r-1}^{(r-1)}$$

son todos distintos de cero, pero el elemento $a_{rr}^{(r)}$ obtenido en el paso $r-1$ será necesariamente igual a cero y en consecuencia la eliminación no puede proseguir sin pivotación. ■

Evidentemente la propiedad de un sistema de ecuaciones lineales mencionada en este teorema (de poder ser resuelto por eliminación sin pivotación) no es una propiedad del sistema como tal, sino que es una propiedad del *orden en que están dadas las ecuaciones*. En todo sistema no singular pueden reordenarse las ecuaciones de tal forma que el sistema tenga dicha propiedad.

La segunda dificultad que puede surjir durante la eliminación es un poco más sutil y es debida al hecho de que los cálculos necesarios para obtener la solución se realicen en aritmética de coma flotante con una precisión prefijada. Como consecuencia de esto puede ocurrir que al comienzo de alguno de los pasos el coeficiente por el que tenemos que dividir los demás de su columna para hallar los multiplicadores de ese paso, sin ser cero, sea, en valor absoluto, tan pequeño en relación a los demás coeficientes que produzca uno o varios multiplicadores con valor absoluto excesivamente grande, resultando en graves errores de redondeo. Esto puede dar lugar a una solución completamente inaceptable como en el siguiente ejemplo en que los coeficientes y términos independientes están dados con una precisión de cuatro dígitos decimales y se usa aritmética de coma flotante con cuatro dígitos:

$$\begin{aligned} 0.0003\,x + 1.566\,y &= 1.569 \\ 0.3454\,x - 2.436\,y &= 1.018\,. \end{aligned}$$

El primer (y único) multiplicador es $\frac{0.3454}{0.0003} = 1151$, luego en el primer paso de eliminación obtenemos

$$\begin{aligned} 0.0003\,x + 1.566\,y &= 1.569 \\ -\,1804\,y &= -1805\,, \end{aligned}$$

de donde

$$\begin{aligned} y &= -1805/1804 = 1.001 \\ x &= \frac{1.569 - 1.566 \times 1.001}{0.0003} = 3.333\,. \end{aligned}$$

mientras que la solución exacta es $x = 10\,,\ y = 1$.

Una sencilla solución de esta dificultad, que simultáneamente resuelve la primera mencionada, consiste en incluir en la fase de eliminación la llamada *pivotación parcial*.

Pivotación parcial.–

El método de pivotación parcial consiste en elegir como *pivote* en cada paso de la eliminación el coeficiente de mayor valor absoluto de la primera

columna del subsistema correspondiente a ese paso. Por ejemplo, en el caso del sistema

$$\begin{aligned} -\,7y + 10z &= 7 \\ 6x + 2y - \ 3z &= 4 \\ 5x - \ y \ + \ 5z &= 6 \end{aligned}$$

en el primer paso queremos eliminar la x. La ecuación que tiene el coeficiente de x con máximo valor absoluto es la segunda, luego el pivote en este paso es 6 y comenzamos el primer paso intercambiando las ecuaciones primera y segunda quedando el sistema en la forma:

$$\begin{aligned} 6x + 2y - \ 3z &= 4 \\ -\,7y + 10z &= 7 \\ 5x - \ y \ + \ 5z &= 6 \end{aligned}$$

Con este método el cálculo de los multiplicadores a utilizar en cada paso se realiza mediante una división por el divisor de mayor valor absoluto que es posible tener sin permutar las variables, lo que resuelve en gran medida las dificultades mencionadas. Nótese que el pivote necesariamente será no nulo en cada paso si el sistema es no-singular. Este método nos ofrece, pues, un test de singularidad.

Algoritmo del método de eliminación de Gauss con pivotación parcial.–

Lo que hemos dicho hasta ahora queda reflejado en el siguiente pseudo código para resolver un sistema de ecuaciones lineales:

```
Suponemos dados: n, el orden del sistema,
  A(i, j), la matriz de coeficientes y
  B(i), la matriz de términos independientes.
La solución sera X(i).
********* BUCLE DE ELIMINACION *********
 para k = 1 hasta n-1
********* BUSQUEDA DEL pivote *********
 c = abs(A(k, k)); p = k
 para i = k+1 hasta n
  si abs(A(i, k)) > c entonces c = abs(A(i, k)) y p = i
 siguiente i
```

```
 si c = 0 entonces imprimir 'SISTEMA SINGULAR' y parar
********* INTERCAMBIO DE LA ECUACION k CON LA p *********
 para j = k hasta n
  t = A(p, j); A(p, j) = A(k, j); A(k, j) = t
 siguiente j
 t = B(p); B(p) = B(k); B(k) = t
********* PASO k DE LA ELIMINACION *********
 para i = k+1 hasta n
  m = A(i, k)/A(k, k); A(i, k) = 0
  para j = k+1 hasta n
 A(i, j) = A(i, j) - m*A(k, j)
  siguiente j
  B(i) = B(i) - m*B(k)
 siguiente i
 siguiente k
********* SUSTITUCION REGRESIVA *********
 si A(n, n) = 0 entonces imprimir 'SISTEMA SINGULAR' y parar
 X(n) = B(n)/A(n, n)
 para k = n-1 hasta 1 incremento -1
 s = 0
 para j = k+1 hasta n
 s = s + A(k, j)*X(j)
 siguiente j
 X(k) = (B(k) - s)/A(k, k)
 siguiente k
```

2.1.2 La factorización triangular realizada por la eliminación de Gauss

El proceso de transformación a que se somete un sistema de ecuaciones al resolverlo por el método de eliminación consiste en una serie de *operaciones elementales* realizadas sobre las filas de la matriz de coeficientes ampliada por la columna de términos independientes.

Recordemos que el resultado de realizar una serie de operaciones elementales sobre las filas de una matriz cualquiera es el mismo que si (1) realizamos dichas operaciones elementales sobre (las filas de) la matriz identidad que tiene el mismo número de filas que la dada y (2) multiplicamos la matriz resultante (por la izquierda) por la matriz dada.

Por ejemplo, la eliminación que hemos realizado en el sistema (2.1)

consiste en las operaciones elementales:

1. Restar de la segunda fila la primera multiplicada por $-\frac{3}{10}$,

2. Restar de la tercera fila la primera multiplicada por $\frac{5}{10}$,

3. Restar de la tercera fila la segunda multiplicada por -25.

Efectuando estas operaciones sobre la matriz identidad de orden 3 ésta se convierte sucesivamente en

$$\begin{pmatrix} 1 & 0 & 0 \\ \frac{3}{10} & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad \begin{pmatrix} 1 & 0 & 0 \\ \frac{3}{10} & 1 & 0 \\ \frac{-5}{10} & 0 & 1 \end{pmatrix}, \quad \begin{pmatrix} 1 & 0 & 0 \\ \frac{3}{10} & 1 & 0 \\ \frac{25\times 3-5}{10} & 25 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ \frac{3}{10} & 1 & 0 \\ 7 & 25 & 1 \end{pmatrix}.$$

Multiplicando ahora la matriz resultante por la matriz de coeficientes obtenemos

$$\begin{pmatrix} 1 & 0 & 0 \\ \frac{3}{10} & 1 & 0 \\ 7 & 25 & 1 \end{pmatrix} \begin{pmatrix} 10 & -7 & 0 \\ -3 & 2 & 6 \\ 5 & -1 & 5 \end{pmatrix} = \begin{pmatrix} 10 & -7 & 0 \\ 0 & -0.1 & 6 \\ 0 & 0 & 155 \end{pmatrix}$$

que es la matriz de coeficientes del sistema reducido, es decir, del obtenido tras la eliminación. Análogamente, si multiplicamos por la matriz de términos independientes obtendremos los términos independientes del sistema reducido:

$$\begin{pmatrix} 1 & 0 & 0 \\ \frac{3}{10} & 1 & 0 \\ 7 & 25 & 1 \end{pmatrix} \begin{pmatrix} 7 \\ 4 \\ 6 \end{pmatrix} = \begin{pmatrix} 7 \\ 6.1 \\ 155 \end{pmatrix}$$

La consecuencia más importante de todo esto es que el proceso de eliminación puede considerarse como la construcción de la matriz M que multiplicada por la matriz de coeficientes, A, nos da una matriz triangular superior U

$$M \cdot A = U.$$

La matriz U resultante es la matriz de los coeficientes del nuevo sistema de ecuaciones y M es la matriz que se obtiene al aplicar las operaciones elementales de eliminación a la matriz identidad. Así pues una forma sencilla de obtener la matriz M es realizando el proceso de eliminación sobre la matriz de coeficientes ampliada con la matriz identidad (igual que hacemos

en el cálculo de la matriz inversa). Cuando la matriz de coeficientes se ha puesto en forma triangular ($A \mapsto U$), la matriz identidad se ha convertido en la matriz M ($I \mapsto M$).

Una característica importante de la matriz M de las transformaciones elementales realizadas durante la eliminación es que si, como ocurre en el ejemplo, durante la eliminación no ha habido ninguna pivotación entonces M es una matriz triangular inferior con unos en la diagonal. En este caso su inversa, $L = M^{-1}$ nos proporciona la factorización triangular de la matriz de coeficientes del sistema: $A = LU$. Además esta inversa, L, tiene la propiedad de ser también una matriz triangular inferior con unos en la diagonal pero que además sus elementos bajo la diagonal son los multiplicadores utilizados en la eliminación.

El proceso de eliminación de Gauss nos proporciona, en caso de poderlo realizar sin pivotación alguna como en nuestro ejemplo, la factorización triangular (o factorización LU) de una matriz sin más que hallar $L = M^{-1}$. En nuestro ejemplo,

$$A = \begin{pmatrix} 10 & -7 & 0 \\ -3 & 2 & 6 \\ 5 & -1 & 5 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ \frac{3}{10} & 1 & 0 \\ 7 & 25 & 1 \end{pmatrix}^{-1} \begin{pmatrix} 10 & -7 & 0 \\ 0 & -0.1 & 6 \\ 0 & 0 & 155 \end{pmatrix}$$

es decir,

$$L = \begin{pmatrix} 1 & 0 & 0 \\ \frac{3}{10} & 1 & 0 \\ 7 & 25 & 1 \end{pmatrix}^{-1}, \quad U = \begin{pmatrix} 10 & -7 & 0 \\ 0 & -0.1 & 6 \\ 0 & 0 & 155 \end{pmatrix}.$$

Nótese que L resultará ser (suponiendo que no ha habido pivotaciones) la inversa de una matriz triangular inferior con unos en la diagonal, lo que justifica que ella misma es triangular inferior con unos en la diagonal.

Pero esto no es todo. El proceso de eliminación de Gauss nos proporciona directamente la matriz L sin necesidad de calcular ninguna inversa. La razón de esto es que L es la inversa de un producto de matrices elementales y por tanto se puede obtener ella misma como resultado de operaciones elementales sobre las filas de la matriz identidad. Lo único que necesitamos es realizar las inversas de las operaciones elementales usadas en la eliminación y en el orden inverso. Es decir, en nuestro ejemplo,

1. Sumar a la tercera fila la segunda multiplicada por -25.

2. Sumar a la tercera fila la primera multiplicada por $\frac{5}{10}$,

3. Sumar a la segunda fila la primera multiplicada por $-\frac{3}{10}$,

que efectuadas sucesivamente sobre la matriz identidad de orden 3 dan lugar a

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -25 & 1 \end{pmatrix}, \quad \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ \frac{5}{10} & -25 & 1 \end{pmatrix}, \quad \begin{pmatrix} 1 & 0 & 0 \\ \frac{-3}{10} & 1 & 0 \\ \frac{5}{10} & -25 & 1 \end{pmatrix} = L.$$

Ejercicio 2.4
Comprobar que

$$\begin{pmatrix} 1 & 0 & 0 \\ \frac{-3}{10} & 1 & 0 \\ \frac{5}{10} & -25 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ \frac{3}{10} & 1 & 0 \\ 7 & 25 & 1 \end{pmatrix}^{-1}.$$

Así pues, la factorización LU de nuestra matriz de coeficientes es:

$$\begin{pmatrix} 10 & -7 & 0 \\ -3 & 2 & 6 \\ 5 & -1 & 5 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ \frac{-3}{10} & 1 & 0 \\ \frac{5}{10} & -25 & 1 \end{pmatrix} \begin{pmatrix} 10 & -7 & 0 \\ 0 & -0.1 & 6 \\ 0 & 0 & 155 \end{pmatrix}.$$

Vemos que los elementos de la matriz L bajo la diagonal son precisamente los multiplicadores utilizados en la eliminación. Esto es un resultado general:

El proceso de eliminación de Gauss (sin pivotación) lleva a cabo una factorización de la matriz de coeficientes en una matriz (L) triangular inferior con unos en la diagonal y una matriz (U) triangular superior. Los elementos de la matriz L (bajo la diagonal) son los multiplicadores utilizados durante la eliminación mientras que los de U son los coeficientes del sistema reducido.

Efecto de la pivotación.–

En lo que precede hemos supuesto explícitamente que en el proceso de eliminación no se realizaba ninguna reordenación de filas, es decir, se realizaba sin pivotación. ¿Cómo se altera lo dicho en el caso de realizar la eliminación con algún tipo de pivotación?

Para contestar a esta pregunta supongamos que llevamos a cabo un proceso de eliminación con pivotación sobre un sistema de ecuaciones lineales. Esto significa que algunas de las operaciones elementales realizadas durante la eliminación pueden ser intercambios de filas. Por ejemplo, en nuestro ejemplo pivotación parcial daría lugar a las siguientes operaciones elementales:

1. Restar de la segunda fila la primera multiplicada por $-\frac{3}{10}$,

2. Restar de la tercera fila la primera multiplicada por $\frac{5}{10}$,

3. Intercambiar la segunda fila con la tercera,

4. Restar de la tercera fila la segunda multiplicada por $(-1/25)$.

Ahora bien, el intercambio de filas (o pivotación) realizado antes del paso k de la eliminación puede realizarse inmediatamente antes de realizar el paso $k-1$ sin afectar en nada al proceso (ya que es un intercambio de dos ecuaciones *posteriores* a la $k-1$, que es la ecuación pivote en el paso $k-1$). Podemos luego, para cada ecuación, calcular su multiplicador, que es independiente del orden que ocupa (aunque esto implica una "reordenación" de los multiplicadores, pero que se realiza automáticamente). Como consecuencia se tiene que todos los intercambios de ecuaciones necesarios en un proceso de eliminación con pivotación pueden realizarse juntos (pero sin alterar su orden) antes de comenzar la eliminación. En nuestro ejemplo, la eliminación puede realizarse mediante los pasos (nótese el intercambio de los dos primeros multiplicadores):

1. Intercambiar la segunda fila con la tercera,

2. Restar de la segunda fila la primera multiplicada por $\frac{5}{10}$,

3. Restar de la tercera fila la primera multiplicada por $-\frac{3}{10}$,

4. Restar de la tercera fila la segunda multiplicada por $(-1/25)$.

La consecuencia de lo que acabamos de decir es que la eliminación de Gauss con pivotación es equivalente a la eliminación sin pivotación pero realizada después de una reordenación de las ecuaciones del sistema. Esta

reordenación puede realizarse sobre la matriz de coeficientes, A, (así como sobre la matriz de términos independientes, B), mediante la multiplicación (por la izquierda) por una matriz de permutación P que se ha obtenido sometiendo la matriz identidad a los sucesivos intercambios de filas que se han de realizar sobre A. Así pues, el proceso de eliminación con pivotación es equivalente a la transformación de la matriz A en la forma

$$MPA = U$$

donde P es una matriz de permutación.

Como consecuencia de todo lo dicho tenemos:

Teorema 11 (Factorización triangular) *Para toda matriz cuadrada no singular A se puede hallar una permutación P tal que PA admite una factorización triangular, es decir, de la forma*

$$PA = LU$$

donde L una matriz triangular inferior con unos en la diagonal y U una matriz triangular superior. Las matrices L y U están completamente determinadas por A y P. Además, se podrá elegir $P = 1$ (matriz identidad) si y sólo si todos los menores principales de A son no singulares. ■

Además sabemos cómo encontrar P así como los elementos de L y U mediante el proceso de eliminación de Gauss con pivotación parcial,

Ejercicio 2.5
Adaptar el algoritmo de eliminación de Gauss con pivotación parcial para convertirlo en un algoritmo que efectúe la factorización triangular de una matriz no singular dada, es decir, que a partir de una matriz A encuentre las matrices P, L, y U del teorema anterior.

Uso de la factorización triangular en la resolución de sistemas de ecuaciones lineales.–

El mayor interés de obtener la factorización triangular mencionada más arriba está en el hecho de que permite resolver cualquier sistema de ecuaciones lineales con la misma matriz de coeficientes A mediante una sustitución progresiva seguida de una sustitución regresiva. Esto es: del sistema

$Ax = b$ pasamos a $PAx = Pb$, o sea, $LUx = b'$. Esto significa que si hemos realizado la eliminación en A obteniendo P, L y U, la resolución de un sistema $Ax = b$ se reduce a estos pasos:

1. Obtener el vector $b' = Pb$.
2. Resolver $Ly = b'$ mediante una sustitución progresiva.
3. Resolver $Ux = y$ mediante una sustitución regresiva.

Esto es especialmente útil cuando necesitamos resolver varios sistemas de ecuaciones que tienen la misma matriz de coeficientes.

Según se ha visto, al resolver un sistema de ecuaciones lineales mediante el proceso de eliminación de Gauss con pivotación parcial se lleva a cabo en realidad una factorización triangular $PA = LU$ de una permutación de la matriz de coeficientes así como una sustitución progresiva y una sustitución regresiva. En la parte de factorización no interviene la columna de términos independientes, sino solamente la matriz de coeficientes. Esto abre la posibilidad de separar las dos partes del proceso, de forma que si necesitamos resolver varios sistemas de ecuaciones lineales en los que sólo difieren los términos independientes, sólo tengamos que realizar la parte de factorización una vez. Puesto que esa parte es la más costosa (en términos del número de operaciones a realizar), esta estrategia redundará en una mayor eficacia del método.

Una de las aplicaciones de este método es el cálculo de la matriz inversa. Este cálculo es claramente equivalente a la resolución de n sistemas de ecuaciones lineales todos ellos con la matriz dada como matriz de coeficientes y teniendo como columnas de términos independientes las columnas de la matriz identidad. Es interesante observar que el cálculo de la matriz inversa de una matriz A de orden n por este método requiere realizar el mismo número de operaciones que el cálculo de A^2.

Otra aplicación de este método es el algoritmo de la mejora iterativa de las soluciones aproximadas que explicamos más adelante (página 72).

2.1.3 Método de Gauss-Jordan

Si en cada paso del proceso de eliminación

$$(\text{ecuación } i) = (\text{ecuación } i) - m_i \times (\text{ecuación } k) \quad i = k+1, \ldots n$$

aplicamos la eliminación no sólo a las ecuaciones siguientes a la ecuación k, sino también a las ecuaciones anteriores entonces estaremos convirtiendo nuestro sistema en uno diagonal, no sólo triangular. Es decir si en el paso k realizamos

$$(\text{ecuación } i) = (\text{ecuación } i) - m_i \times (\text{ecuación } k)$$

para $i = 1, \ldots, k-1, k+1, \ldots n$, nuestro sistema queda convertido en uno de n ecuaciones con una incógnita, para cuya solución no es necesario el proceso de sustitución regresiva. Solamente necesitaremos n divisiones adicionales. Este método se conoce con el nombre de *método de Gauss-Jordan*

Ejercicio 2.6
Hallar el número de operaciones (contando sólo multiplicaciones y divisiones) que es necesario realizar para resolver un sistema de n ecuaciones lineales con n incógnitas por el método de eliminación de Gauss-Jordan.

2.1.4 Otras técnicas de pivotación

Una técnica de pivotación es un criterio para la elección de la fila o ecuación *pivote* en cada paso del proceso de eliminación. Hasta ahora hemos encontrado dos técnicas de pivotación: pivotación simple y pivotación parcial, pero existen muchas otras posibilidades. Lo ideal sería descubrir una técnica sencilla de pivotación que nos asegurase que hacemos las operaciones en el orden óptimo en el sentido de que se minimizan tanto la pérdida de precisión como los errores de redondeo. En ausencia de esa técnica ideal hemos de guiarnos por la experiencia y por consideraciones generales tales como la citada más arriba. Veremos a continuación dos técnicas adicionales que suelen dar resultados adecuados.

Pivotación parcial escalada.–

Para algunos sistemas podemos obtener algún aumento en la precisión si al elegir la fila pivote tenemos en cuenta de alguna manera no tanto el tamaño del elemento pivote como su tamaño *relativo* a los demás elementos de su fila. De esta forma el proceso no quedaría afectado si se re-escala alguna de las ecuaciones del sistema, es decir, si se multiplica por un número toda una ecuación (lo cual, en principio no debería afectar a la solución).

Así pues, realizaremos una *pivotación parcial escalada* cuando en cada paso de la eliminación elijamos como pivote aquel candidato cuyo cociente por el tamaño de su fila sea máximo. Por "tamaño" de una fila se puede entender alguna norma vectorial conveniente tal como la norma del máximo, la cual se calcula con poco esfuerzo.

Para realizar la pivotación parcial escalada de forma eficaz calcularemos los tamaños de todas las filas antes de comenzar el primer paso de la eliminación,

$$d_i = \max_{1 \le j \le n} |a_{ij}|$$

Después, al comienzo del paso k para $k = 1, \ldots, n-1$, si los posibles pivotes son $a_{kk}, \ldots, a_{nk}$, elegimos como pivote aquel para el que el cociente $|a_{ik}|/d_i$ sea máximo.

Ejercicio 2.7
Modificar el algoritmo de eliminación de Gauss para que realice pivotación parcial escalada en lugar de mera pivotación parcial.

Pivotación total.–

La técnica de pivotación total es considerada como la que da mejores resultados en general, aunque su uso es bastante limitado debido al coste adicional de programación que supone. Esta técnica es una variante de la pivotación parcial en la que no sólo se barajan las ecuaciones restantes en la búsqueda del pivote sino que en cada paso se busca la incógnita más adecuada a eliminar en ese paso. Así, no sólo nos permitimos permutar las ecuaciones, sino que también nos permitimos permutar las incógnitas.

Al comienzo del paso k se elegirá como pivote aquel elemento de mayor valor absoluto en la submatriz de coeficientes con que se trabajará en ese paso. En consecuencia un algoritmo que lleve a cabo pivotación total realizará en cada paso tanto una permutación de filas (o ecuaciones) como una permutación de columnas (o variables).

2.2 Análisis de errores y Condición de un sistema de ecuaciones lineales

2.2.1 Introducción: Distintas medidas del error

En la práctica del cálculo científico raramente se conocen con exactitud los coeficientes y términos independientes del sistema de ecuaciones lineales que se quiere resolver. En general, esos coeficientes se calculan a partir de observaciones experimentales sujetas como mínimo a errores que vienen de la limitada precisión de los aparatos de medida. En consecuencia se plantea la cuestión de la *confianza* que podemos tener en los resultados de nuestros cálculos sabiendo que los propios datos están sujetos a error. En otras palabras, nos preguntamos en que medida nuestros cálculos pueden haber amplificado el error inherente a los datos y si éste se habrá mantenido dentro de límites razonables.

Normalmente la confianza que podemos tener en la solución dependerá del sistema de que se trate: distintas matrices de coeficientes darán lugar a distintos comportamientos respecto a la propagación y amplificación de errores. Nuestro objetivo es poder determinar a priori el tipo de comportamiento que una matriz dada va a tener. Para ello comenzaremos repasando varias formas de medir la exactitud o precisión de una solución aproximada, $\hat{x}$, de un sistema de ecuaciones lineales $Ax = b$.

La primera forma es mediante el **error absoluto** o diferencia $e = x - \hat{x}$ entre la solución exacta y la solución aproximada. Obviamente esto es normalmente desconocido. Otra forma de medir la exactitud de $\hat{x}$ como solución de $Ax = b$ es mediante el **error residual**, que mide cuán lejos está $\hat{x}$ de satisfacer el sistema de ecuaciones lineales. El error residual de $\hat{x}$ es

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

lo cual puede calcularse sin dificultad, pues es igual a la diferencia $b - A\hat{x}$ entre el vector de términos independientes dado y el calculado a partir de la solución aproximada.

En general el error residual no será el vector cero, aunque podemos esperar que su tamaño sea pequeño. Al hablar del "tamaño" de vectores en $\mathbf{R}^n$ nos referimos a una norma particular fija que suponemos elegida de antemano. Qué norma se use no tiene demasiada importancia mientras uno se atenga a usar siempre esa norma. Una norma muy en uso es la

"norma infinito" o del máximo ya que se calcula con muy poco esfuerzo (no requiere operaciones aritméticas). En general se puede suponer que la norma usada es una "norma p", para algún real $p \geq 1$, definida por

$$\|x\|_p = \left(|x_1|^p + \cdots + |x_n|^p\right)^{1/p}.$$

De éstas la norma del máximo es el caso límite para $p \to \infty$. Los casos más utilizados son $p = 1, 2, \infty$.

El tamaño del error residual está relacionado con el tamaño del error absoluto. Al multiplicar la matriz A por un vector columna u la norma de éste puede variar mucho, pero hay que observar, sin embargo, que el cambio *relativo* o factor por el que ha cambiado la norma (el cociente $\|Au\|/\|u\|$) no puede ser arbitrario. Necesariamente permanecerá dentro de ciertos límites que dependen de la matriz A. Concretamente $\|Au\|/\|u\|$ será necesariamente menor o igual que el *máximo factor* por el que A multiplica las normas de los vectores. Veremos a continuación que debido a esto el **error residual relativo**

$$\frac{\|b - Ax\|}{\|b\|} = \frac{\|r\|}{\|b\|} = \frac{\|Ae\|}{\|Ax\|}$$

es un buen indicador, no del error e, sino del **error relativo**

$$\frac{\|x - \hat{x}\|}{\|x\|} = \frac{\|e\|}{\|x\|}.$$

(ver fórmula (2.3)).

2.2.2 Mejora iterativa

Algunos ordenadores y lenguajes de programación admiten realizar aritmética de coma flotante en dos modos distintos de precisión: sencilla y doble, de los que el segundo modo cuesta aproximadamente el doble de tiempo-máquina que el primero pero permite reducir los errores (especialmente los de pérdida de precisión) al realizar las operaciones con aritmética de coma flotante con el doble de dígitos de precisión que en el modo de precisión sencilla. Cuando se dispone de esta posibilidad se puede reducir notablemente el error en la solución de sistemas de ecuaciones lineales sin un coste excesivo (como el que implicaría el realizar todos los cálculos con doble precisión). La técnica que permite realizar esta reducción de error se conoce como *mejora iterativa* y consiste en lo siguiente:

Supongamos que hemos utilizado el método de la factorización triangular (con precisión sencilla) para hallar la solución de un sistema de ecuaciones lineales $Ax = b$ tal como

$$\begin{pmatrix} 0.20000 & 0.16667 & 0.14286 \\ 0.16667 & 0.14286 & 0.12500 \\ 0.14286 & 0.12500 & 0.11111 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0.50953 \\ 0.43453 \\ 0.37897 \end{pmatrix}.$$

O sea, suponemos que ya hemos obtenido una factorización $A = PLU$ de la matriz de coeficientes, que en nuestro ejemplo es

$$L = \begin{pmatrix} 1 & 0 & 0 \\ 0.83335 & 1 & 0 \\ 0.71430 & 0.89673 & 1 \end{pmatrix}, \quad U = \begin{pmatrix} 0.20000 & 0.16667 & 0.14286 \\ 0 & 0.00397 & 0.00595 \\ 0 & 0 & 0.00015 \end{pmatrix},$$

sin realizar permutaciones, o sea que P es la identidad, y hemos obtenido la solución

$$x^{(1)} = \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1.0384 \\ 0.89673 \\ 1.0667 \end{pmatrix},$$

realizando las operaciones con aritmética de coma flotante con precisión de cinco dígitos. Este resultado puede mejorarse de la siguiente forma:

Primeramente calculamos el error residual $r^{(1)} = b - Ax^{(1)}$ utilizando aritmética de doble precisión. Esto significa realizar para cada $i \in \{1, \ldots, n\}$ las siguientes $(n-1)$ multiplicaciones y n sumas con doble precisión:

$$r_i^{(1)} = b - \sum_{k=1}^{n} a_{ik} x_k^{(1)},$$

lo que en nuestro ejemplo da

$$Ax^{(1)} = \begin{pmatrix} 0.5095324653 \\ 0.4345190593 \\ 0.3789619207 \end{pmatrix} \quad \text{y} \quad r^{(1)} = \begin{pmatrix} -0.24653 \\ 1.0941 \\ 0.80793 \end{pmatrix} 10^{-5}$$

Obtenida esta aproximación del vector de error residual la utilizamos para calcular una estimación $e^{(1)}$ del error absoluto e resolviendo la ecuación $Ae = r$. Esta ecuación representa un sistema de ecuaciones lineales que sólo se diferencia del dado en los términos independientes. En consecuencia, como ya disponemos de la factorización de la matriz de coeficientes la nueva resolución tendrá un coste mucho menor ya que se reduce

a una sustitución progresiva y una sustitución regresiva como las que nos llevaron de b a $x^{(1)}$, pero esta vez utilizando $r^{(1)}$ en lugar de b como vector de términos independientes. De esta forma obtenemos el vector $e^{(1)}$, que sumado a $x^{(1)}$ resultará en la mejora

$$x^{(2)} = x^{(1)} + e^{(1)} .$$

En nuestro ejemplo obtenemos

$$e^{(1)} = \begin{pmatrix} -0.03709 \\ 0.09955 \\ -0.06424 \end{pmatrix} \quad \text{y} \quad x^{(2)} = x^{(1)} + e^{(1)} = \begin{pmatrix} 1.0014 \\ 0.99628 \\ 1.0024 \end{pmatrix}$$

Repitiendo este proceso hallaríamos el resto $r^{(2)}$ asociado con $x^{(2)}$ utilizando doble precisión, después hallaríamos el 'error' $e^{(2)}$ mediante la resolución de $Ae^{(2)} = r^{(2)}$ y obtendríamos una nueva mejora $x^{(3)} = x^{(2)} + e^{(2)}$, que en nuestro ejemplo es

$$x^{(3)} = \begin{pmatrix} 1.0001 \\ 0.99986 \\ 1.0001 \end{pmatrix} .$$

Continuando de este modo se obtiene una sucesión de vectores

$$\{x^{(1)}, x^{(2)}, \dots\}$$

que son sucesivas mejoras que bajo hipótesis bastante razonables convergerán a la solución exacta x (en el ejemplo la solución exacta es

$$x = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

como se comprueba fácilmente).

Resumiendo, tenemos el siguiente algoritmo para la mejora iterativa de la solución de un sistema de ecuaciones lineales:

Algoritmo de la Mejora Iterativa

1 Factorización triangular $A = PLU$.

2 Utilizar dos sustituciones a partir de P, L, U y b para obtener x.

3 Con doble precisión calcular el resto de x, $r = b - Ax$.

4 Utilizar dos sustituciones a partir de P, L, U y r para obtener e.

5 Sumar e al viejo x para obtener el nuevo x.

6 Ir al paso **3**.

La única condición que se ha de cumplir para poder garantizar que este algoritmo converge a la solución exacta es que el proceso de resolución (una sustitución progresiva y una regresiva basadas en la factorización PLU de la matriz de coeficientes) sea equivalente a aplicar una matriz C "suficientemente cercana" a la inversa de la matriz de coeficientes, como se demuestra en el siguiente teorema:

Teorema 12 (Convergencia del método de mejora iterativa) *Sea $x^{(1)}$ un vector cualquiera y sea C una matriz cercana a la inversa de A en el sentido de que*

$$\|I - CA\| < 1 .$$

Entonces la sucesión de vectores $\{x^{(n)}\}$ definida inductivamente por

$$x^{(n+1)} = x^{(n)} + C(b - Ax^{(n)})$$

converge a la solución x del sistema de ecuaciones lineales $Ax = b$.

Demostración: Basta demostrar que $\lim_{n\to\infty} \|x^{(n)} - x\| = 0$. Ahora bien,

$$\begin{aligned} \|x^{(n)} - x\| &= \|x^{(n-1)} + C(b - Ax^{(n-1)}) - x\| \\ &= \|x^{(n-1)} - x + C(Ax - Ax^{(n-1)})\| \\ &= \|x^{(n-1)} - x - CA(x^{(n-1)} - x)\| \\ &= \|(I - CA)(x^{(n-1)} - x)\| \le \|I - CA\| \cdot \|x^{(n-1)} - x\| \end{aligned}$$

y continuando así se llega a

$$\|x^{(n)} - x\| \le \|I - CA\|^{n-1} \cdot \|x^{(1)} - x\|$$

de donde

$$0 \le \lim_{n\to\infty} \|x^{(n)} - x\| \le \|x^{(1)} - x\| \lim_{n\to\infty} \|I - CA\|^{n-1} = \|x^{(1)} - x\| \cdot 0 = 0$$

ya que por ser $\|I - CA\| < 1$, obviamente $\lim_{n\to\infty} \|I - CA\|^{n-1} = 0$. ■

Detención del algoritmo.–

Para implementar cualquier algoritmo iterativo es necesario establecer un criterio de detención, es decir, establecer una norma a seguir para determinar en qué paso conviene detener los cálculos. Un posible criterio (en ocasiones el único) es realizar un número fijo de pasos determinado a priori. Por supuesto el criterio mejor sería poder decidir cuál es el error absoluto que estamos dispuestos a aceptar y detener el algoritmo justo cuando se alcance un error inferior al que vamos a aceptar. Esto se complica por el hecho de que el error absoluto es desconocido, sin embargo en el algoritmo de la mejora iterativa que acabamos de estudiar el error absoluto puede ser acotado en términos de las sucesivas *correcciones* como indica el siguiente resultado,

Ejercicio 2.8
En el proceso de mejora iterativa definido por

$$x^{(n+1)} = x^{(n)} + C(b - Ax^{(n)})$$

el tamaño del error $x^{(n)} - x$ se puede acotar en cada paso por el siguiente múltiplo de la corrección $x^{(n+1)} - x^{(n)}$ a realizar en ese paso:

$$\|x^{(n)} - x\| \leq \|(CA)^{-1}\| \|x^{(n+1)} - x^{(n)}\| .$$

Como la matriz CA (y por tanto también $(CA)^{-1}$) es cercana a la identidad, el factor $\alpha = \|(CA)^{-1}\|$ será cercano a la unidad, por lo que podemos utilizar como criterio de detención el tamaño de las sucesivas correcciones $\|x^{(n+1)} - x^{(n)}\|$, es decir, detener el algoritmo cuando la corrección realizada tenga un tamaño menor que el error máximo admisible.

2.2.3 Normas de matrices

El máximo factor por el que A multiplica las normas de los vectores se llama la *norma de la matriz* A (relativa a la norma vectorial empleada). Es decir, definimos la norma de la matriz A como

$$\|A\| = \max_{x \neq 0} \frac{\|Ax\|}{\|x\|} \tag{2.2}$$

y en consecuencia se verifica evidentemente para todo $u \neq 0$,

$$\frac{\|Au\|}{\|u\|} \leq \|A\|,$$

y por lo tanto, suponiendo que A es no-singular, también

$$\frac{\|v\|}{\|Av\|} \leq \|A^{-1}\|,$$

para todo $v \neq 0$.

Ejercicio 2.9
Multiplicando esas dos desigualdades deducir que para cualquier matriz cuadrada no-singular A y cualesquiera dos vectores no nulos u, v se verifica

$$\frac{\|Au\|}{\|u\|} \leq \|A\|\|A^{-1}\| \cdot \frac{\|Av\|}{\|v\|} .$$

Concluir que el número $\|A\| \cdot \|A^{-1}\|$ es siempre mayor o igual que 1.

La primera consecuencia de nuestra definición de norma de una matriz es la siguiente acotación del tamaño del error absoluto en términos del tamaño del error residual:

$$\frac{\|r\|}{\|A\|} \leq \|e\| \leq \|A^{-1}\|\|r\| .$$

Ejercicio 2.10
Del resultado del ejercicio A.11 se deduce que si A es una matriz cuadrada no-singular cualquiera se verifica

$$\frac{1}{\|A\|\|A^{-1}\|} \cdot \frac{\|Av\|}{\|Au\|} \leq \frac{\|v\|}{\|u\|} \leq \|A\|\|A^{-1}\| \cdot \frac{\|Av\|}{\|Au\|}$$

para cualesquiera vectores no nulos u, v.

Ejercicio 2.11
Demostrar que la definición (2.2) de norma de una matriz es equivalente a

$$\|A\| = \max_{\|x\|=1} \|Ax\|,$$

y también a la conjunción de

$$\big(\forall x \|Ax\| \leq \|A\|\|x\|\big) \quad y \quad \exists x \neq 0 \mid \|Ax\| \geq \|A\|\|x\|.$$

Ejercicio 2.12
Demostrar que el concepto de norma de matrices definido antes satisface los axiomas usuales de normas y dos axiomas adicionales: (1) Para toda matriz identidad, I, $\|I\| = 1$, y (2) Si A y B son matrices cuadradas del mismo orden, $\|AB\| \leq \|A\| \cdot \|B\|$. Usar estas dos propiedades para probar que si A es inversible, $\|A\|\|A^{-1}\| \geq 1$.

Cálculo de algunas normas de matrices.–

El cálculo de la norma de una matriz asociada a una norma vectorial dada puede llegar a ser bastante engorroso. Un caso especial por su sencillez nos lo da la norma del máximo o "norma infinito". Obsérvese que para todo x,

$$\|Ax\|_\infty = \max_{1\leq i\leq n} \left| \sum_{j=1}^n a_{ij}x_j \right| \leq \max_{1\leq i\leq n} \sum_{j=1}^n |a_{ij}||x_j|$$
$$\leq \max_{1\leq i\leq n} \Big(\big(\max_{1\leq j\leq n} |x_j| \big) \sum_{j=1}^n |a_{ij}| \Big) = \|x\|_\infty \max_{1\leq i\leq n} \sum_{j=1}^n |a_{ij}|$$

lo cual demuestra que $\|A\|_\infty \leq \max_{1\leq i\leq n} \sum_{j=1}^n |a_{ij}|$. Por otro lado es fácil dar un vector x tal que $\|Ax\|_\infty \geq \|x\|_\infty \max_{1\leq i\leq n} \sum_{j=1}^n |a_{ij}|$. Para ello sea i_0 la fila de A para la que la suma $\sum_{j=1}^n |a_{ij}|$ es máxima, es decir, tal que

$$\sum_{j=1}^n |a_{i_0 j}| = \max_{1\leq i\leq n} \sum_{j=1}^n |a_{ij}|$$

Eligiendo x de forma que $x_j = \text{signo}(a_{i_0 j})$, tendremos (suponiendo $A \neq 0$) $\|x\|_\infty = 1$ y

$$\|Ax\|_\infty = \max_{1\leq i\leq n} \left| \sum_{j=1}^n a_{ij}x_j \right| \geq \left| \sum_{j=1}^n a_{i_0 j}x_j \right|$$
$$= \sum_{j=1}^n |a_{i_0 j}| = \|x\|_\infty \max_{1\leq i\leq n} \sum_{j=1}^n |a_{ij}|$$

En conclusión tenemos que para toda matriz cuadrada $A = (a_{ij})$,

$$\boxed{\|A\|_\infty = \max_{1\leq i\leq n} \sum_{j=1}^n |a_{ij}|}\,.$$

Por ejemplo, en el sistema de ecuaciones

$$0.0003\,x + 1.566\,y = 1.569$$
$$0.3454\,x - 2.436\,y = 1.018\,.$$

la matriz de coeficientes

$$A = \begin{pmatrix} 0.0003 & 1.566 \\ 0.3454 & -2.436 \end{pmatrix}$$

tiene norma-infinito

$$\|A\|_\infty = \max\{|0.0003| + |1.566|\,,\ |0.3454| + |-2.436|\} = 2.781\,.$$

Al extremo opuesto de la norma infinito está la norma uno que, curiosamente, su cálculo cuesta lo mismo que el de la norma infinito ya que

Ejercicio 2.13
Para toda matriz cuadrada A, su norma-uno es igual a la norma-infinito de su traspuesta

$$\|A\|_1 = \|A^t\|_\infty\,,$$

es decir, $\|A\|_1 = \max_{1\leq j\leq n} \sum_{i=1}^n |a_{ij}|$.

2.2.4 Condición de una matriz y acotación de errores

Acotación del error relativo.–

Volviendo a nuestro sistema de ecuaciones lineales $Ax = b$ y a los errores absoluto, $e = x - \hat{x}$, y residual, $r = b - A\hat{x} = Ae$, podemos interpretar el resultado del ejercicio 2.10 como estableciendo la siguiente acotación del error relativo $\|e\|/\|x\|$ en términos del error residual relativo $\|r\|/\|b\|$ y del número $c(A) = \|A\| \cdot \|A^{-1}\|$,

$$\frac{1}{c(A)} \cdot \frac{\|r\|}{\|b\|} \leq \frac{\|e\|}{\|x\|} \leq c(A) \cdot \frac{\|r\|}{\|b\|}\,. \tag{2.3}$$

Este número

$$c(A) = \|A\|\|A^{-1}\|,$$

que podemos asociar con cada matriz no singular A una vez elegida una norma, se llama *condición de la matriz A* (relativo a la norma elegida). Según la última parte del ejercicio A.11 la condición de una matriz verifica

$$c(A) \geq 1.$$

Por ejemplo, para la matriz del ejemplo anterior, con precisión de cuatro dígitos,

$$A^{-1} = \frac{1}{\det A}\begin{pmatrix} -2.436 & -1.566 \\ -0.3454 & 0.0003 \end{pmatrix} = \begin{pmatrix} 4.498 & 2.891 \\ 0.6377 & -5.539 \times 10^{-4} \end{pmatrix}$$

y por tanto

$$\begin{aligned} \|A^{-1}\|_\infty &= \max\{|4.498| + |2.891|\,,\ |0.6377| + |-0.0006|\} \\ &= 7.389\,, \end{aligned}$$

de donde obtenemos

$$c_\infty(A) = \|A\|_\infty \|A^{-1}\|_\infty = 2.781 \times 7.389 = 20.55$$

indicando un pobre condicionamiento de la matriz en cuestión.

Condición en términos de operadores lineales.–

Hay que acentuar el hecho de que la condición de una matriz está definida solamente si se ha especificado una norma vectorial, aunque algunas matrices (como por ejemplo las que son un múltiplo de la identidad) tienen la misma condición relativa a cualquier norma. En realidad el concepto de condición que acabamos de definir pertenece, rigurosamente hablando, a los operadores lineales en espacios normados. Cuando se habla de norma de una matriz se está identificando esa matriz con el operador lineal en $\mathbf{R}^n$ que la tiene por matriz en la base canónica. El siguiente ejercicio nos indica el significado geométrico de la condición de un operador lineal en un espacio normado: es una medida de la distorsión que el operador produce en las bolas.

Ejercicio 2.14
La condición de un operador lineal en un espacio normado es igual al cociente del radio máximo al radio mínimo de la imagen de la bola unitaria.

Ejercicio 2.15
La condición de una matriz A es igual a 1 *si y sólo si (el operador lineal representado por) A conserva las bolas, lo cual es decir que multiplica las normas de todos los vectores por el mismo factor.*

Dado que la condición de una matriz depende de la norma vectorial con que se calcula, se plantea la cuestión de si un sistema de ecuaciones lineales puede parecer bien condicionado en términos de una norma y mal condicionado en términos de otra distinta. Como respuesta parcial a esta cuestión tenemos:

Ejercicio 2.16
Si A es una matriz cuadrada real no singular que está bien condicionada respecto a una p-norma entonces también estará bien condicionada respecto a cualquier otra p-norma. Además, si la condición de A es igual a 1 *en relación a una p-norma para* $p \neq 2$ *entonces la condición de A también es igual a* 1 *en relación a cualquier otra p-norma. (Demuéstrese en el caso de matrices de orden* 2*.)*

Errores en los términos independientes.–

La condición de una matriz A nos sirve para estimar, mediante las desigualdades (2.3) el error relativo de una solución aproximada de cualquier sistema de ecuaciones lineales cuya matriz de coeficientes sea A. Pero además la condición de A es un indicador de la sensibilidad de la solución de dicho sistema de ecuaciones lineales a errores en el vector de términos independientes ya que las desigualdades (2.3) pueden interpretarse también de la siguiente forma: Supongamos que existe un error Δb en el vector b de términos independientes lo cual produce un error Δx en la solución, es decir, tenemos $A(x + \Delta x) = b + \Delta b$ donde x es la solución exacta de $Ax = b$. Entonces $A\Delta x = \Delta b$, $e = \Delta x$ y resulta $r = Ae = \Delta b$ en (2.3), y tenemos la siguiente acotación del error relativo resultante de los errores en los términos independientes:

$$\frac{1}{c(A)} \cdot \frac{\|\Delta b\|}{\|b\|} \leq \frac{\|\Delta x\|}{\|x\|} \leq c(A) \cdot \frac{\|\Delta b\|}{\|b\|}\,.$$

Así pues, el condicionamiento de un sistema de ecuaciones lineales está indicado por la condición de la matriz de coeficientes. Si este número es cercano a la unidad el sistema está bien condicionado porque pequeños

errores (relativos) en los datos (términos independientes) no pueden dar lugar a grandes errores (relativos) en la solución.

Errores en los coeficientes del sistema.–

Nos interesa también estudiar el efecto que tienen los errores que pueda haber en los coeficientes del sistema.

Supongamos primeramente que A y B son dos matrices de las que A es inversible y que u, v son dos vectores tales que $Au = Bv$. Entonces $u = A^{-1}Bv$ y por tanto

$$\|u\| = \|A^{-1}Bv\| \leq \|A^{-1}\|\|Bv\| \leq \|A^{-1}\|\|B\|\|v\|$$

de donde deducimos, suponiendo $v \neq 0$,

$$\|B\|\frac{\|v\|}{\|u\|} \geq \frac{1}{\|A^{-1}\|} \quad \text{o bien} \quad \frac{\|u\|}{\|v\|} \leq c(A)\frac{\|B\|}{\|A\|}\,. \tag{2.4}$$

Esta acotación puede aplicarse al caso de un sistema de ecuaciones lineales en el que los coeficientes se conozcan sólo aproximadamente como, por ejemplo, cuando se han obtenido como resultado de medidas experimentales. Tal situación es la de un sistema $Ax = b$ del que en lugar de la matriz A sólo disponemos de la matriz $\hat{A} = A + E$ donde E es una matriz de error. Si resolvemos exactamente el sistema $\hat{A}\hat{x} = b$, ¿Cuál será el error relativo en que hemos incurrido?

Nuestra situación es la de tener $\hat{A}\hat{x} = b = Ax = (\hat{A}-E)x = \hat{A}x - Ex$, de donde $\hat{A}(x - \hat{x}) = Ex$ y por tanto la acotación que acabamos de ver nos proporciona una acotación del error relativo en términos de algo que es aproximadamente el error relativo de los coeficientes,

$$\frac{\|x - \hat{x}\|}{\|x\|} \leq c(\hat{A})\frac{\|E\|}{\|\hat{A}\|}\,.$$

Alternativamente podemos poner $(A + E)\hat{x} = Ax$ para escribir $A(x - \hat{x}) = E\hat{x}$ de donde el error relativo en los coeficientes nos da una acotación de una cantidad que es aproximadamente igual al error relativo en la solución,

$$\frac{\|x - \hat{x}\|}{\|\hat{x}\|} \leq c(A)\frac{\|E\|}{\|A\|}\,.$$

2.2.5 Estimación de la condición de una matriz

El cálculo de la condición de una matriz A tiene la dificultad de requerir calcular la matriz inversa, A^{-1}, lo cual es, en general, demasiado costoso. Resulta, pues, importante encontrar métodos de estimación o acotación de la condición. Uno de tales métodos se basa en el siguiente teorema que se deduce fácilmente de (2.4),

Teorema 13 *Si A es una matriz inversible entonces para cualquier matriz singular E se verifica*

$$\|A + E\| \geq \frac{1}{\|A^{-1}\|} \quad \textit{y por tanto} \quad c(A) \geq \frac{\|A\|}{\|A + E\|}\,.$$

Demostración: Por ser E singular existe un vector no nulo x tal que $Ex = 0$. Entonces para este vector, $(A + E)x = Ax$ y la conclusión se deduce de (2.4) con $B = A + E$ y $u = v = x$. ■

Ejercicio 2.17
Usar el resultado de este teorema para demostrar que toda matriz con diagonal estrictamente dominante *es inversible, es decir, no singular.*

Corolario 5 (Condición de matrices triangulares) *Si $A = (a_{ij})$ es una matriz triangular inversible entonces para cualquier elemento a_{ii} de su diagonal se verifica*

$$c_\infty(A) \geq \frac{\|A\|_\infty}{|a_{ii}|}\,,$$

en particular,

$$c_\infty(A) \geq \frac{\|A\|_\infty}{\min\limits_{1\leq i\leq n}|a_{ii}|}\,.$$

Demostración: Sea E la matriz que resulta al sustituir el elemento a_{jj} de A por cero. Entonces E es obviamente singular y por tanto

$$c_\infty(A) \geq \frac{\|A\|_\infty}{\|A - E\|_\infty} = \frac{\|A\|_\infty}{|a_{jj}|}\,.$$ ■

Ejercicio 2.18
Demostrar que las conclusiones de este corolario siguen siendo ciertas si se sustituye la norma infinito por la norma uno, obteniéndose una fórmula

de acotación de la condición, relativa a la norma uno, de cualquier matriz triangular.

Corolario 6 (Estimación de la condición de cualquier matriz) *Si $A = (a_{ij})$ es una matriz inversible de orden n entonces para cualquier $j \in \{1, \ldots, n\}$ se verifica*

$$c_\infty(A) \geq \frac{\|A\|_\infty}{\max\limits_{1\leq i\leq n} |a_{ij}|}$$

y en particular,

$$c_\infty(A) \geq \frac{\|A\|_\infty}{\min\limits_{1\leq j\leq n} \left(\max\limits_{1\leq i\leq n} |a_{ij}| \right)}.$$

Demostración: Sea E la matriz obtenida al sustituir todos los elementos de la columna j de la matriz A por cero. Obviamente E es una matriz singular y la matriz $A - E$ es una matriz todos cuyos elementos son cero excepto que la columna j es igual a la columna j de A. Por tanto tenemos $\|A - E\|_\infty = \max\limits_{1\leq i\leq n} |a_{ij}|$. En consecuencia, según el teorema 13,

$$c_\infty(A) \geq \frac{\|A\|_\infty}{\max\limits_{1\leq i\leq n} |a_{ij}|}. \quad \blacksquare$$

Ejercicio 2.19
Establecer, por analogía con el resultado anterior, una fórmula de acotación de la condición de cualquier matriz relativa a la norma uno.

2.2.6 Condición de las matrices ortogonales y sus múltiplos

Las matrices ortogonales tienen condición óptima para la norma euclídea ya que conservan esta norma y por tanto conservan sus bolas, lo cual, según el ejercicio 2.15, implica que el número de condición es 1. Pero asimismo, cualquier múltiplo de una matriz ortogonal conservará las bolas de la norma euclídea por lo que también tendrá condición 1 en esa norma.

2.3 Métodos iterativos

Los sistemas de ecuaciones lineales que surgen de ciertos problemas (principalmente en la resolución de ecuaciones diferenciales por métodos de diferencias finitas) tienen la característica de ser de órdenes excesivamente elevados para su tratamiento por métodos directos de resolución, los cuales exigirían la utilización de gran cantidad de memoria para almacenar todos los elementos de la matriz de los coeficientes del sistema. Por otro lado los coeficientes de tales sistemas suelen venir dados por una fórmula sencilla, de forma que pueden ser generados cuando se necesiten. Esto último ocurre también con las matrices *escasas* en las que una gran mayoría de elementos son cero. Tales sistemas suelen cumplir las condiciones de convergencia de métodos iterativos como los utilizados en la determinación del punto fijo de un operador (véase (19)), métodos que se pueden aplicar sin necesidad de tener todos los coeficientes del sistema almacenados. Así pues, para la resolución de algunos sistemas de ecuaciones lineales resultan más convenientes los métodos iterativos como los que estudiamos a continuación.

2.3.1 Ejemplo introductorio del método iterativo

Supongamos que nos dan el sistema de ecuaciones lineales

$$\begin{aligned} 10x + 2y - 3z &= 5 \\ 2x + 8y + z &= 6 \\ 3x - y + 15z &= 12 \end{aligned} \tag{2.5}$$

Ahora "despejamos" x de la primera ecuación, y de la segunda y z de la tercera obteniendo

$$\begin{aligned} x &= \frac{5 - 2y + 3z}{10}, \\ y &= \frac{6 - 2x - z}{8}, \\ z &= \frac{12 - 3x + y}{15}. \end{aligned}$$

Esto obviamente no nos proporciona una solución; es simplemente una reformulación del sistema de ecuaciones lineales dado, que ha tomado la forma general

$$\mathbf{x} = f(\mathbf{x}) \tag{2.6}$$

para f dada por $f(\mathbf{x}) = G\mathbf{x} + d$ donde

$$G = \begin{pmatrix} 0 & -\frac{2}{10} & \frac{3}{10} \\ -\frac{2}{8} & 0 & -\frac{1}{8} \\ -\frac{3}{15} & \frac{1}{15} & 0 \end{pmatrix} \quad \text{y} \quad d = \begin{pmatrix} \frac{5}{10} \\ \frac{6}{8} \\ \frac{12}{15} \end{pmatrix} .$$

La reformulación (2.6) indica que la solución del sistema (2.5) es un punto fijo de la función f. ¿Qué ocurrirá si aplicamos f a un vector $\mathbf{x}$ arbitrario? Supongamos que elegimos una estimación inicial arbitraria como por ejemplo

$$x^{(0)} = 0.5\,, \quad y^{(0)} = 1\,, \quad z^{(0)} = 1\,.$$

Aplicando f a este vector obtenemos otro vector $\mathbf{x}^{(1)} = f(\mathbf{x}^{(0)}) = C\mathbf{x}^{(0)} + d$ cuyas coordenadas son

$$x^{(1)} = 0.6\,, \quad y^{(1)} = 0.5\,, \quad z^{(1)} = 0.76\,.$$

Si ahora aplicamos f al resultado que acabamos de obtener llegamos a un vector $\mathbf{x}^{(2)}$ dado por

$$x^{(2)} = 0.63\,, \quad y^{(2)} = 0.504\,, \quad z^{(2)} = 0.7133\,,$$

y continuando de esta forma en la octava iteración se obtiene

$$x^{(8)} = 0.611831\,, \quad y^{(8)} = 0.508104\,, \quad z^{(8)} = 0.711507\,,$$

lo cual está muy cerca de la solución del sistema.

Este es el procedimiento básico de los métodos iterativos: construir una función contractiva cuyo punto fijo es la solución buscada e *iterar* repetidamente dicha función hasta acercarnos suficientemente al punto fijo. Decimos que se *itera* una función cuando se evalúa ésta en el resultado de haberla evaluado previamente, es decir, la evaluación de $f(f(x))$. Si se empieza con un valor $x^{(0)}$ las sucesivas iteraciones de f se obtienen siempre evaluando f en distintos valores, pero producen los mismos resultados que si aplicamos sucesivamente las funciones f, f^2, f^3, etc. a $x^{(0)}$.

Reformulado un sistema de ecuaciones lineales como un problema de punto fijo de una función contractiva (llamada *función de iteración*), por la teoría general de las funciones contractivas sabemos que sucesivas iteraciones de la función empezando con cualquier vector dará lugar a una sucesión que converge a la solución del sistema. Veamos a continuación un método general de obtener una función de iteración para un sistema de ecuaciones lineales dado para la cual es relativamente sencillo determinar si es contractiva o no.

2.3.2 Esquema general del Método Iterativo

No es difícil inventar formas de asociar a un sistema de ecuaciones lineales una función para la que la solución del sistema sea un punto fijo. Menos trivial es hacerlo de forma que la función obtenida sea contractiva. El siguiente método general tiene la ventaja de que permite confirmar de forma sencilla si la función asociada al sistema es contractiva y por tanto si el método iterativo converge.

Sea $Ax = b$ nuestro sistema de ecuaciones. Sea $A = A_1 + A_2$ una descomposición de la matriz de coeficientes A como suma de dos matrices de forma tal que A_1 es inversible.

Ejercicio 2.20
Si calculamos

$$G = -A_1^{-1}A_2 \quad y \quad d = A_1^{-1}b$$

entonces el sistema $Ax = b$ es equivalente a

$$x = Gx + d\,.$$

En consecuencia tenemos,

Si $A = A_1 + A_2$ y $G = -A_1^{-1}A_2$, $d = A_1^{-1}b$ entonces cada solución del sistema $Ax = b$ es un punto fijo de la función

$$f(x) = Gx + d$$

y viceversa.

El método general descrito en el ejercicio 2.20 da lugar a muchos métodos particulares cuando se eligen diversas formas de descomponer la matriz de coeficientes A como suma de dos matrices una de las cuales es inversible. Una de las formas más sencillas de hacer esto es llamada el método de *Jacobi*, que estudiaremos en breve.

Ejercicio 2.21
Indicar tres formas diferentes de descomponer una matriz cuadrada A sin ceros en la diagonal como suma de dos matrices una de las cuales es inversible. Describir el método iterativo de resolución de sistemas de ecuaciones lineales a que cada una de esas formas de descomposición da lugar.

2.3.3 Convergencia del método iterativo general

Sea $Ax = b$ un sistema de ecuaciones lineales y supongamos que tenemos una descomposición $A = A_1 + A_2$ de la matriz de coeficientes tal que A_1 es inversible, de modo que el sistema original es equivalente al problema de punto fijo $x = Gx + d$ donde $G = -A_1^{-1}A_2$ y $d = A_1^{-1}b$. Nos interesa saber bajo qué condiciones se puede resolver el problema de punto fijo mediante iteraciones de la función $f(x) = Gx + d$, es decir, nos interesa conocer las condiciones que garantizan que dado un vector $x^{(0)}$, la sucesión

$$\{x^{(k)}\} = \{f(x^{(0)}),\ f(f(x^{(0)})), \ldots,\ f^k(x^{(0)}), \ldots\} \tag{2.7}$$

converge al punto fijo de f. Pero nótese que si dicha sucesión converge a un vector x, la continuidad de f implica

$$x = \lim_{k\to\infty} x^{(k)} = \lim_{k\to\infty} f(x^{(k-1)}) = f(\lim_{k\to\infty} x^{(k-1)}) = f(x),$$

o sea, x es un punto fijo de f. O sea, que basta saber que la sucesión de iteraciones converge, pues si lo hace el límite es necesariamente la solución de nuestro sistema.

Supongamos, que $\{x^{(k)}\}$ converge y que su límite es x. Entonces la sucesión de "errores" $\{x^{(k)} - x\}$ converge a cero, pero teniendo en cuenta que

$$\begin{aligned} x^{(k)} - x &= G(x^{(k-1)}) + d - (G(x) + d) = G(x^{(k-1)} - x) = \cdots \\ &= G^k(x^{(0)} - x) \end{aligned}$$

vemos que

$$\lim_{k\to\infty} G^k(x^{(0)} - x) = 0.$$

De esto se deduce que si la sucesión (2.7) converge para cualquier elección del vector inicial $x^{(0)}$ entonces la sucesión de potencias de la matriz G tiende a la matriz cero. Trivialmente se cumple también el recíproco (hay que recordar que f *tiene* un punto fijo por hipótesis), de forma que tenemos:

Teorema 14 *Un método iterativo $x = Gx + d$ de resolución de un sistema de ecuaciones lineales converge para toda elección del vector inicial $x^{(0)}$ si y sólo si $G^n \to 0$, o sea, si y sólo si la sucesión de potencias de la matriz G tiende a la matriz cero.* ■

Una forma trivial de comprobar que $G^n \to 0$ sería el comprobar que para alguna norma se verifica $\|G\| < 1$ ya que en tal caso $\|G^n\| \leq \|G\|^n \to 0$. Por tanto es evidente que una sencilla condición suficiente de convergencia es:

Condición suficiente de convergencia de un método iterativo. *Si $\|G\|_\infty < 1$ entonces el método iterativo $x = Gx + d$ converge para toda elección del vector inicial $x^{(0)}$.*

En realidad no cuesta demasiado esfuerzo adicional el dar una visión completa de la situación demostrando el siguiente teorema que da las dos condiciones necesarias y suficientes fundamentales para la convergencia del método iterativo general:

Teorema 15 (Condiciones necesarias y suficientes) *Cada una de las condiciones enunciadas a continuación es necesaria y suficiente para que un método iterativo, $x = Gx + d$, de resolución de un sistema de ecuaciones lineales converja para toda elección del vector inicial $x^{(0)}$:*

1. *Que exista una norma para la que se verifique $\|G\| < 1$,*
2. *Que todo autovalor de G tenga valor absoluto menor que* 1.

Demostración: Como ya hemos observado más arriba, la condición 1 es suficiente. Por otro lado es inmediato que la condición 2 es necesaria ya que si G tiene un autovalor, λ, cuyo valor absoluto es $|\lambda| \geq 1$, este autovalor tendrá un autovector y (para el cual $G^k y = \lambda^k y$) que hace imposible que $G^k \to 0$ (ya que para todo k, $\|G^k\| \geq \frac{\|G^k y\|}{\|y\|} = |\lambda|^k \geq 1$). Así pues, la demostración del teorema queda completa con el siguiente lema:

Lema 3 *Si todo autovalor de G tiene valor absoluto menor que* 1 *entonces existe una norma para la que se verifica $\|G\| < 1$.*

Demostración: Si todo autovalor de G tiene valor absoluto menor que 1 entonces el radio espectral, $\rho(G)$, de G (el máximo de los valores absolutos de los autovalores de G) es menor que 1, por lo tanto existe un número real, ϵ, tal que $0 < \epsilon < 1 - \rho(G)$. Sea D la matriz diagonal $D = \mathrm{diag}(1, \epsilon^{-1}, \epsilon^{-2}, ..., \epsilon^{1-m})$. Sea J la forma canónica de Jordan de G y S la matriz de semejanza tal que $J = SGS^{-1}$. Para toda matriz X definimos

$$\|X\|_T = \|TXT^{-1}\|_\infty$$

donde $T = DS$. Basta demostrar dos cosas: (1) que $\|\cdot\|_T$ es una norma, y (2) que $\|G\|_T < \rho(G) + \epsilon$. La primera se deduce de que $\|\cdot\|_T$ es la norma matricial asociada a la norma vectorial $\|x\| = \|Tx\|_\infty$, y la segunda se comprueba con un sencillo cálculo teniendo en cuenta que en DJD^{-1} en cada bloque de Jordan quedan sustituidos los unos por ϵ. ■

Veamos ahora un criterio para la detención de las iteraciones. ¿Cuándo podemos considerar que ya hemos realizado suficientes iteraciones? Para poder detener las iteraciones de una forma eficaz es necesario poder estimar el error. El resultado siguiente nos permite hallar una cota superior de la magnitud del error en cada paso en términos de la corrección realizada en ese paso:

Proposición 3 *Sea $x^{(k+1)} = Gx^{(k)} + d$ un método iterativo dado. Para cualquier norma se verifica que en cada paso de iteración*

$$\|x^{(k)} - x\| \leq \frac{\|G\|}{1 - \|G\|}\|x^{(k)} - x^{(k-1)}\| .$$

Por lo tanto, si existe alguna norma para la que $\|G\| < \frac{1}{2}$ entonces el error en cada paso (medido con esa norma) está acotado superiormente por la corrección realizada en ese paso.

Demostración: Dado que $x^{(k)} - x = -G(x^{(k)} - x^{(k-1)}) + G(x^{(k)} - x)$, tomando normas obtenemos,

$$\|x^{(k)} - x\| \leq \|G\|\|x^{(k)} - x^{(k-1)}\| + \|G\|\|x^{(k)} - x\|$$

y despejando $\|x^{(k)} - x\|$ obtenemos

$$\|x^{(k)} - x\| \leq \frac{\|G\|}{1 - \|G\|}\|x^{(k)} - x^{(k-1)}\| ,$$

como queríamos demostrar. ■

2.3.4 Método de Jacobi

Estamos ahora en situación de volver al ejemplo ilustrativo de los métodos iterativos explicado al principio y analizarlo de forma un poco más teórica.

Descrito en forma general, el método en que se basaba aquel ejemplo es el siguiente:

Dado un sistema de ecuaciones lineales $Ax = b$, la ecuación i puede escribirse

$$\sum_{j=1}^{n} a_{ij}x_j = b_i .$$

Supongamos que los elementos diagonales de A sean todos distintos de cero. Entonces podemos despejar la primera incógnita de la primera ecuación, la segunda incógnita de la segunda ecuación, y así hasta despejar la última incógnita de la última ecuación de forma que el sistema queda re-escrito como

$$x_i = \Big(b_i - \sum_{\substack{j=1 \\ j\neq i}}^{n} a_{ij}x_j \Big)\Big/ a_{ii} . \tag{2.8}$$

Esto lo podemos expresar fácilmente en forma matricial si observamos que

$$\sum_{\substack{j=1 \\ j\neq i}}^{n} a_{ij}x_j = \begin{pmatrix} 0 & a_{12} & \cdots & \cdots & a_{1n} \\ a_{21} & 0 & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & 0 & a_{n-1\,n} \\ a_{n1} & \cdots & \cdots & a_{n\,n-1} & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = A_2 x ,$$

donde A_2 es la matriz obtenida a partir de A al hacer cero todos sus elementos diagonales. Además, si A_1 es "la diagonal de A", es decir, tal que $A = A_1 + A_2$, entonces el vector de coordenadas b_i/a_{ii} es simplemente $A_1^{-1}b$ y el vector de coordenadas $\big(\sum_{\substack{j=1 \\ j\neq i}}^{n} a_{ij}x_j\big)/a_{ii}$ (para $i = 1, \ldots, n$) es simplemente $A_1^{-1}A_2x$, de forma que la expresión (2.8) en forma matricial es

$$x = A_1^{-1}b - A_1^{-1}A_2x ,$$

es decir, la expresión (2.8) tiene la forma

$$\boxed{x = Gx + d} \qquad \text{con} \quad G = -A_1^{-1}A_2 , \text{ y } d = A_1^{-1}b .$$

Esta elección de descomposición de la matriz de coeficientes de un sistema de ecuaciones lineales en "diagonal" mas "el resto" para resolverlo

por el método iterativo, se conoce con el nombre de *Método de Jacobi* o de los *desplazamientos simultáneos*. Resumiendo,

Supongamos que la matriz de coeficientes A de nuestro sistema de ecuaciones lineales no tiene ningún cero en la diagonal. En tal situación, la matriz A_1 cuya diagonal es igual a la de A y cuyos otros elementos son cero es inversible y el método iterativo basado en la descomposición $A = A_1 + A_2$ se conoce como método de Jacobi.

Convergencia del método de Jacobi.–

Según la condición suficiente de convergencia establecida justo antes del teorema 15, el método de Jacobi será aplicable siempre que la matriz A de los coeficientes del sistema de ecuaciones lineales tenga la propiedad de que $\|D^{-1}(A-D)\|_\infty = \|D^{-1}A - I\|_\infty < 1$ donde D es la matriz "diagonal de A". Ahora bien,

$$\|D^{-1}A - I\|_\infty = \max_{1\le i\le n}\left\{\Big(\sum_{\substack{j=1\\ j\neq i}}^{n}|a_{ij}|\Big)\Big/|a_{ii}|\right\}$$

luego una condición suficiente para la convergencia del método de Jacobi es:

$$\max_{1\le i\le n}\left\{\Big(\sum_{\substack{j=1\\ j\neq i}}^{n}|a_{ij}|\Big)\Big/|a_{ii}|\right\} < 1 \qquad \text{ó}\,, \qquad (\forall i)_{1\le i\le n}\ |a_{ii}| > \sum_{\substack{j=1\\ j\neq i}}^{n}|a_{ij}|\,.$$

Expresando esta condición con palabras:

Condición suficiente para la convergencia del método de Jacobi es que cada elemento de la diagonal de la matriz de coeficientes tenga un valor absoluto mayor que la suma de los valores absolutos de los demás elementos de su fila.

Las matrices descritas en la condición suficiente anterior se llaman *matrices de diagonal estrictamente dominante (por filas)*. Concluimos, pues, que el método de Jacobi es aplicable a cualquier sistema cuya matriz de coeficientes tiene la diagonal estrictamente dominante (por filas).

Obsérvese que toda matriz de diagonal estrictamente dominante (lo que implica que es no singular según el ejercicio 2.17) necesariamente tiene

todos los elementos de la diagonal distintos de cero y por tanto tiene sentido el plantearse usar el método de Jacobi.

2.3.5 Aceleración de la convergencia; método de Gauss-Seidel

La principal dificultad que plantean los métodos iterativos tales como el de desplazamientos simultáneos (o de Jacobi) es su lentitud en la convergencia. Por ello es necesario diseñar métodos de aceleración de la convergencia, El primero que estudiaremos es el de *desplazamientos sucesivos* (o de Gauss-Seidel). Más tarde estudiaremos los llamados métodos de *relajación* por su origen en el estudio de problemas mecánicos. Libros aconsejados: (17), (38) y (21).

El método de Jacobi se puede describir como basado en las ecuaciones (2.8), que nos dan las siguientes ecuaciones de iteración:

$$x_i^{(k+1)} = \Big(b_i - \sum_{\substack{j=1\\ j\neq i}}^{n} a_{ij}x_j^{(k)}\Big)\Big/ a_{ii}\,.$$

Según estas ecuaciones cada coordenada del paso $k+1$ se halla mediante las coordenadas del paso k. Pero obsérvese que habiendo calculado $x_1^{(k+1)}$, y dado que en principio $x_1^{(k+1)}$ está más cerca de la solución exacta que $x_1^{(k)}$, parece lógico pensar que sería más provechoso usar $x_1^{(k+1)}$ que $x_1^{(k)}$ en el cálculo de $x_2^{(k+1)}$. Asimismo sería más ventajoso usar $x_1^{(k+1)}$ y $x_2^{(k+1)}$ en lugar de $x_1^{(k)}$ y $x_2^{(k)}$ en el cálculo de $x_3^{(k+1)}$ y así sucesivamente. Esto sugiere el sustituir las ecuaciones de iteración anteriores por estas otras

$$x_i^{(k+1)} = \Big(b_i - \sum_{j=1}^{i-1} a_{ij}x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij}x_j^{(k)}\Big)\Big/ a_{ii}\,. \qquad (2.9)$$

en las que en cada paso se usan para cada coordenada su estimación más actual.

Este método se conoce con el nombre de *Método de Gauss-Seidel* y puede llegar a aumentar la velocidad de convergencia en algunos casos a más del doble (respecto al método de Jacobi).

Las ecuaciones (2.9) pueden expresarse en forma matricial para descubrir a qué corresponden en el esquema general del método iterativo y poder así estudiar su convergencia basados en el teorema 15. Pasando todos los términos que hay en (2.9) con coordenadas del paso $k+1$ al miembro de la izquierda después de haber multiplicado por a_{ii} las ecuaciones (2.9) quedan

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

pero teniendo en cuenta que

$$\sum_{j=1}^{i} a_{ij} x_j^{(k+1)} = \begin{pmatrix} a_{11} & 0 & \cdots & \cdots & 0 \\ a_{21} & a_{22} & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & \vdots \\ \vdots & & & a_{n-1\,n-1} & 0 \\ a_{n1} & \cdots & \cdots & a_{n\,n-1} & a_{n\,n} \end{pmatrix} \begin{pmatrix} x_1^{(k+1)} \\ x_2^{(k+1)} \\ \vdots \\ x_n^{(k+1)} \end{pmatrix}$$

y

$$\sum_{j=i+1}^{n} a_{ij} x_j^{(k)} = \begin{pmatrix} 0 & a_{12} & \cdots & \cdots & a_{1n} \\ 0 & 0 & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & \vdots \\ \vdots & & & 0 & a_{n-1\,n} \\ 0 & \cdots & \cdots & 0 & 0 \end{pmatrix} \begin{pmatrix} x_1^{(k)} \\ x_2^{(k)} \\ \vdots \\ x_n^{(k)} \end{pmatrix}$$

podemos escribir las ecuaciones anteriores como

$$A_1 x^{(k+1)} = b - A_2 x^{(k)} \qquad \text{ó} \qquad x^{(k+1)} = G x^{(k)} + d$$

donde A_1 es la matriz obtenida al hacer cero todos los elementos de A sobre la diagonal, y $A_2 = A - A_1$. En otras palabras, el método de Gauss-Seidel encaja en el esquema general de los métodos iterativos como el método obtenido con la descomposición

$$A = \begin{pmatrix} a_{11} & 0 & \cdots & \cdots & 0 \\ a_{21} & a_{22} & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & \vdots \\ \vdots & & & a_{n-1\,n-1} & 0 \\ a_{n1} & \cdots & \cdots & a_{n\,n-1} & a_{n\,n} \end{pmatrix} + \begin{pmatrix} 0 & a_{12} & \cdots & \cdots & a_{1n} \\ 0 & 0 & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & \vdots \\ \vdots & & & 0 & a_{n-1\,n} \\ 0 & \cdots & \cdots & 0 & 0 \end{pmatrix} .$$

Convergencia del método de Gauss-Seidel.–

Del teorema general de convergencia del método iterativo se deduce la siguiente condición suficiente de convergencia:

Teorema 16 (Convergencia del método de Gauss-Seidel) *El método iterativo de Gauss-Seidel para la resolución de un sistema de ecuaciones lineales converge para cualquier vector inicial siempre que la matriz de coeficientes del sistema sea de diagonal estrictamente dominante.*

Demostración: Sea $A = L + D + U$ la descomposición de la matriz de coeficientes en elementos bajo la diagonal (matriz L), elementos diagonales (matriz D) y elementos sobre la diagonal, de forma que la matriz de la función de iteración del método de Gauss-Seidel es $G = -(L+D)^{-1}U$ y la matriz de la función de iteración del método de Jacobi es $G_J = -D^{-1}(L+U)$, cumpliéndose $\|G_J\|_\infty < 1$. El teorema quedará demostrado si demostramos que $\|G\|_\infty \leq \|G_J\|_\infty$. Para ello basta demostrar que para cualquier vector x se cumple $\|Gx\|_\infty \leq \|G_J\|_\infty \|x\|_\infty$ o equivalentemente que cada componente y_k del vector $y = Gx = -(L+D)^{-1}Ux$ verifica

$$|y_k| \leq \sum_{\substack{j=1 \\ j\neq i}}^{n} \frac{|a_{ij}|}{|a_{ii}|} \|x\|_\infty . \tag{2.10}$$

Esto lo demostramos por inducción completa en las componentes. Primeramente, y teniendo en cuenta que $(L+D)y = -Ux$, expresamos el vector y en la forma $y = D^{-1}(-Ly - Ux)$ de donde se obtienen para las coordenadas de y fórmulas análogas a (2.9), a saber:

$$y_i = \Big(-\sum_{j=1}^{i-1} a_{ij} y_j - \sum_{j=i+1}^{n} a_{ij} x_j \Big) \Big/ a_{ii} .$$

De aquí se deduce (2.10) para $i = 1$ (¡ejercicio!). Supongamos ahora que $i > 1$ y que se cumple la hipótesis de inducción (2.10) para todo y_j con $j < i$. Entonces cada una de estas y_j también verifica $|y_j| \leq \|G\|_\infty \|x\|_\infty <$

$\|x\|_\infty$ y por lo tanto tenemos:

$$\begin{aligned}
|y_i| &\le \Big(\sum_{j=1}^{i-1}|a_{ij}||y_j| + \sum_{j=i+1}^{n}|a_{ij}||x_j|\Big)\Big/|a_{ii}| \\
&\le \Big(\sum_{j=1}^{i-1}|a_{ij}|\|x\|_\infty + \sum_{j=i+1}^{n}|a_{ij}|\|x\|_\infty\Big)\Big/|a_{ii}| \\
&\le \Big(\sum_{j=1}^{i-1}|a_{ij}| + \sum_{j=i+1}^{n}|a_{ij}|\Big)\|x\|_\infty\Big/|a_{ii}| \\
&\le \sum_{\substack{j=1\\ j\neq i}}^{n}\frac{|a_{ij}|}{|a_{ii}|}\|x\|_\infty
\end{aligned}$$

con lo que se completa la demostración. ■

Es interesante observar el hecho de que hay matrices (que necesariamente no son de diagonal estrictamente dominante ni por filas ni por columnas) para las que el método de Jacobi converge pero el de Gauss-Seidel no; y también hay matrices para las que el método de Gauss-Seidel converge pero el de Jacobi no.

2.3.6 Método de las relajaciones sucesivas

En los métodos iterativos que estamos estudiando, y en particular en el de Gauss-Seidel, cada paso de iteración puede considerarse como la realización de una corrección (también llamada *relajación*) a una aproximación de la solución. Visto desde este punto de vista podemos preguntarnos cuáles son las correcciones empleadas en cada paso del método de Gauss-Seidel y si hay alguna forma de mejorarlas. Evidentemente las correcciones son las diferencias $x_i^{(k+1)} - x_i^{(k)}$, que se encuentran fácilmente a partir de las fórmulas de iteración (2.9). De allí con poco esfuerzo se obtiene:

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

(nótese que el índice inferior del segundo sumatorio ha cambiado de $i+1$ a i). Ahora bien, el hecho de que estas correcciones no proporcionen de inmediato la solución exacta indica que son (en valor absoluto) insuficientes

(caso de una convergencia monótona) o excesivas (caso de una convergencia alternada), y que la corrección exacta sería de la forma

$$c_i = \omega\Big(b_i - \sum_{j=1}^{i-1}a_{ij}x_j^{(k+1)} - \sum_{j=i}^{n}a_{ij}x_j^{(k)}\Big)\Big/a_{ii}$$

para algún factor ω cercano a la unidad, que sería mayor o menor que la unidad según multiplicase a una corrección insuficiente o excesiva. Este factor se llama *coeficiente de relajación*. Si acertamos a elegir un valor apropiado para ω entonces el uso de las nuevas correcciones nos da las ecuaciones de iteración

$$x_i^{(k+1)} = x_i^{(k)} + \omega\Big(b_i - \sum_{j=1}^{i-1}a_{ij}x_j^{(k+1)} - \sum_{j=i}^{n}a_{ij}x_j^{(k)}\Big)\Big/a_{ii}$$

que pueden converger más rápidamente que (2.9). Este método de aceleración se conoce como *Método de las relajaciones sucesivas*.

El uso del término *relajación* surge del hecho de que gran parte de los trabajos iniciales sobre métodos para la solución iterativa de sistema de ecuaciones lineales estaban enfocados a la determinación de las fuerzas y momentos en estructuras mecánicas. Si se usan valores incorrectos de esas cantidades desconocidas, entonces es necesario aplicar en cada nodo de la estructura fuerzas restrictivas artificiales. A medida que uno se acerca a la solución correcta, esas fuerzas pueden ser "relajadas". Véase (17) p.92.

Ejercicio 2.22
Indicar la descomposición $A = A_1 + A_2$ con A_1 inversible que es necesario utilizar para que el esquema general de los métodos iterativos, $G = -A_1^{-1}A_2$, $d = A_1^{-1}b$, de lugar, con esta descomposición, al método de las relajaciones sucesivas, es decir, que se obtenga

$$G = (D+\omega L)^{-1}[(1-\omega)D - \omega U] \qquad y \qquad d = \omega(D+\omega L)^{-1}b$$

donde $A = L + D + U$ es la obvia descomposición en parte inferior, parte diagonal y parte superior.

La dificultad principal para poner en práctica el método de las relajaciones sucesivas consiste en realizar la elección adecuada del coeficiente ω. El valor óptimo de este coeficiente puede hallarse analíticamente sólo

en casos muy especiales y en general ha de hallarse a base de ensayos. Sin embargo es fácil demostrar, apoyados en el teorema 15 y en la expresión para G dada en el ejercicio 2.22, que los únicos valores de ω que pueden dar lugar a un método convergente son aquellos valores positivos menores que 2.

Proposición 4 *Para la matriz $G_\omega = (D+\omega L)^{-1}[(1-\omega)D-\omega U]$, donde D es diagonal y L, U son respectivamente triangulares inferior y superior con diagonal cero, se verifica que su radio espectral es mayor o igual que $|1-\omega|$ y en consecuencia para que un método iterativo de la forma*

$$x^{(k+1)} = G_\omega x^{(k)} + d$$

converja para cualquier valor inicial es condición necesaria que $|1-\omega| < 1$, es decir, $0 < \omega < 2$.

Demostración: Sean $\lambda_1, \lambda_2, \ldots, \lambda_n$ los autovalores de G_ω, entonces

$$\begin{aligned}\lambda_1 \cdots \lambda_n &= \det G_\omega = \det D^{-1} \det[(1-\omega)D] \\ &= (1-\omega)^n \det D^{-1} \det D = (1-\omega)^n\end{aligned}$$

de esto se deduce que $|1-\omega| \leq \max_i |\lambda_i|$. Como se quería demostrar. ■

Respuestas a Algunos Ejercicios del Capítulo 2

Ejercicio 2.1 Sin preocuparnos de posibles divisiones por cero, la base del algoritmo de eliminación es la siguiente

```
Datos: n, el orden del sistema,
A(i, j), la matriz de coeficientes y
B(i), la matriz de terminos independientes.
para k = 1 hasta n-1
para i = k+1 hasta n
m = A(i, k)/A(k, k)
para j = k hasta n
 A(i, j) = A(i, j) - m*A(k, j)
siguiente j
B(i) = B(i) - m*B(k)
siguiente i
siguiente k
```

Nótese que podemos ahorrarnos algunas operaciones en el bucle interno si evitamos calcular los ceros que se obtienen para $j = k$. Éstos se pueden introducir "manualmente" como en

```
para k = 1 hasta n-1
para i = k+1 hasta n
m = A(i, k)/A(k, k)
A(i, k) = 0
para j = k+1 hasta n
 A(i, j) = A(i, j) - m*A(k, j)
siguiente j
B(i) = B(i) - m*B(k)
siguiente i
siguiente k
```

Ejercicio 2.2 La sustitución regresiva se puede realizar mediante el siguiente algoritmo que se puede aplicar a sistemas singulares ya que detecta dicha condición.

```
si A(n, n) = 0 entonces
imprimir 'SISTEMA SINGULAR' y parar
X(n) = B(n)/A(n, n)
para k = n-1 hasta 1 incremento -1
si A(k, k) = 0 entonces
 imprimir 'SISTEMA SINGULAR' y parar
s = 0
para j = k+1 hasta n
 s = s + A(k, j)*X(j)
siguiente j
X(k) = (B(k) - s)/A(k, k)
siguiente k
```

Ejercicio 2.3 En el método de eliminación de Gauss hay dos partes: (1) Eliminación de los elementos bajo la diagonal y (2) sustitución regresiva.

En la primera parte, para eliminar el elemento en posición (k, l) es necesaria una división para calcular el multiplicador y, si el sistema es de orden n, $n-l+1$ multiplicaciones para multiplicar cada uno de los elementos $a_{k\,l+1}, a_{k\,l+2}, \ldots, a_{k\,n}, b_k$ por el multiplicador. Total $n-l+2$ operaciones para el elemento (k, l). Así pues, el número total de operaciones en esta parte es

$$\begin{aligned}
\sum_{l=1}^{n-1}\sum_{k=l+1}^{n}(n-l+2) &= \sum_{l=1}^{n-1}(n-l)(n-l+2)\\
&= (n-1)(n+1)+(n-2)n+(n-3)(n-1)+\cdots+1\cdot 3\\
&= 1\cdot(2+1)+2\cdot(2+2)+\cdots+(n-1)(2+n-1)\\
&= 2(1+2+\cdots+(n-1))+(1^2+2^2+\cdots+(n-1)^2)\\
&= 2\frac{n(n-1)}{2}+\tfrac{1}{6}(n-1)n(2(n-1)+1)\\
&= n^2-n+\tfrac{1}{6}(n^2-n)(2n-1)\\
&= n^2-n+\tfrac{1}{6}(2n^3-3n^2-n)\\
&= \tfrac{1}{3}n^3+\tfrac{1}{2}n^2-\tfrac{5}{6}n.
\end{aligned}$$

a esto hay que sumarle las operaciones realizadas durante la sustitución regresiva, que son: una división para la última ecuación, y en general una división y k multiplicaciones para la ecuación $n-k$ ($k=0,\ldots,n-1$), o sea $\sum_{k=0}^{n-1}(k+1)=\frac{1}{2}n(n+1)=\frac{n^2}{2}+\frac{n}{2}$. Así pues:

Operaciones en Eliminación de Gauss: $\boxed{\frac{1}{3}n^3+n^2-\frac{1}{3}n}$.

Este mismo resultado puede obtenerse de forma mucho menos trabajosa si se sabe que la función que buscamos es un polinomio en n de tercer grado, es decir de la forma $f(n)=an^3+bn^2+cn+d$. Entonces lo único que tenemos que hacer para averiguar los coeficientes a, b, c, d es resolver un sistema de cuatro ecuaciones y cuatro incógnitas que resulta al evaluar f respectivamente en 0, 1, 2, y 3: $f(0)=0$, $f(1)=1$, $f(2)=6$, $f(3)=17$, (esto se halla contando directamente el número de multiplicaciones y divisiones necesarios para resolver los sistemas de órdenes cero, uno, dos y tres) lo cual nos da el sistema de ecuaciones

$$\begin{pmatrix}0&0&0&1\\1&1&1&1\\8&4&2&1\\27&9&3&1\end{pmatrix}\begin{pmatrix}a\\b\\c\\d\end{pmatrix}=\begin{pmatrix}0\\1\\6\\17\end{pmatrix}$$

cuya solución es $a=\frac{1}{3}$, $b=1$, $c=-\frac{1}{3}$, $d=0$. Otra observación útil para este problema es que la función $f(n)$ puede definirse recursivamente a partir del valor inicial $f(0)=0$ por $f(n)=f(n-1)+(n^2+n-1)$. Con este método el cómputo se simplifica al reducirse a contar las operaciones necesarias para eliminar la primera columna y para el último paso de la sustitución regresiva (lo que nos da el sumando (n^2+n-1) de la fórmula de $f(n)$).

Ejercicio 2.4

Ejercicio 2.5 El siguiente algoritmo efectúa la descomposición triangular de una matriz. Suponemos dados: n, el orden del sistema, y $A(i, j)$, la matriz de coeficientes. La solución se almacenará en las variables $p(i)$, $L(i, j)$, y $U(i, j)$. Empezamos haciendo U igual a A y L igual a la matriz identidad.

```
para i = 1 hasta n
para j = 1 hasta n
U(i, j) = A(i, j); L(i, j) = 0
si i = j entonces L(i, j) = 1
siguiente j
siguiente i
para k = 1 hasta n-1
c = abs(U(k, k)); p(k) = k
para i = k+1 hasta n
```

```
si abs(U(i, k)) > c entonces c = abs(U(i, k)) y p(k) = i
siguiente i
si c = 0 entonces imprimir 'MATRIZ SINGULAR' y parar
para j = k hasta n
t = U(p(k), j); U(p(k), j) = U(k, j); U(k, j) = t
siguiente j
para i = k+1 hasta n
L(i, k) = U(i, k)/U(k, k); U(i, k) = 0
para j = k+1 hasta n
U(i, j) = U(i, j) - L(i, k)*U(k, j)
siguiente j
siguiente i
siguiente k
si U(n, n) = 0 entonces imprimir 'MATRIZ SINGULAR'
```

Ejercicio 2.6 En el método de Gauss-Jordan hay una parte de eliminación y una parte de sustitución. La segunda consta solamente de n divisiones mientras que en la primera se usan $n-l+2$ operaciones para eliminar el elemento (k,l) donde k toma $n-1$ valores y l varía desde la columna 1 hasta la n. Así pues, el número total de operaciones en esta parte es

$$\begin{aligned}\sum_{l=1}^{n}(n-1)(n-l+2) &= (n-1)\sum_{l=1}^{n}(n-l+2)\\ &= (n-1)[(n+1)+n+(n-1)+\cdots+2]\\ &= (n-1)(\tfrac{1}{2}(n+1)(n+2)-1)\\ &= (n-1)(\tfrac{1}{2}n^2+\tfrac{3}{2}n+1-1) = \tfrac{1}{2}n^3+n^2-\tfrac{3}{2}n.\end{aligned}$$

y sumando a éstas las n operaciones de la segunda parte, tenemos:

Operaciones en Eliminación de Gauss-Jordan: $\boxed{\frac{1}{2}n^3+n^2-\frac{1}{2}n}$.

De nuevo, al mismo resultado se llega resolviendo el sistema de ecuaciones

$$\begin{pmatrix}0&0&0&1\\1&1&1&1\\8&4&2&1\\27&9&3&1\end{pmatrix}\begin{pmatrix}a\\b\\c\\d\end{pmatrix}=\begin{pmatrix}0\\1\\7\\21\end{pmatrix}$$

satisfecho por los coeficientes de $g(n)=an^3+bn^2+cn+d$.

Ejercicio 2.7

Ejercicio 2.8 A partir de $x^{(n+1)}=x^{(n)}+C(b-Ax^{(n)})=x^{(n)}+CA(x-x^{(n)})$ se deduce $(CA)^{-1}(x^{(n+1)}-x^{(n)})=x-x^{(n)}$, de donde

$$\|x^{(n)}-x\|\le\|(CA)^{-1}\|\|x^{(n+1)}-x^{(n)}\|\,.$$

Ejercicio 2.9 Evidente. La última conclusión se obtiene tomando $u=v\neq 0$.

Ejercicio 2.10 La desigualdad de la derecha es consecuencia inmediata de la del ejercicio anterior. La desigualdad de la izquierda se obtiene de la de la derecha intercambiando u y v y tomando inversos.

Ejercicio 2.11 Si A es un operador lineal en un espacio normado, la obvia inclusión de conjuntos de números

$$E=\{\|Ax\| \mid \|x\|=1\}\subset\left\{\frac{\|Ax\|}{\|x\|}\mid x\neq 0\right\}=F$$

es en realidad una igualdad ya que si $x\neq 0$ entonces tomando $y=kx$ con $k=1/\|x\|$, su norma es $\|y\|=1$ y por lo tanto

$$\frac{\|Ax\|}{\|x\|}=k\|Ax\|=\|k\cdot Ax\|=\|A(kx)\|=\|Ay\|\in E,$$

o sea $F\subset E$. Así pues $E=F$, de donde

$$\max_{x\neq 0}\frac{\|Ax\|}{\|x\|}=\max F=\max E=\max_{\|x\|=1}\|Ax\|.$$

De las dos fórmulas dadas a continuación, la fórmula de la izquierda dice que $\|A\|$ es mayor o igual que el máximo de E, mientras que la de la derecha dice que es menor o igual que ese máximo.

Ejercicio 2.12 $\|I\|=\max_{x\neq 0}\frac{\|Ix\|}{\|x\|}=\max_{x\neq 0}\frac{\|x\|}{\|x\|}=1$. Por ejemplo el otro axioma adicional es consecuencia de que para todo x,

$$\|ABx\|\le\|A\|\|Bx\|\le\|A\|\|B\|\|x\|\,.$$

Usando estas dos propiedades, $1=\|I\|=\|AA^{-1}\|\le\|A\|\|A^{-1}\|$.

Ejercicio 2.13 Primeramente tenemos para todo vector x,

$$\begin{aligned}\|Ax\|_1&=\sum_{i=1}^{n}\Big|\sum_{j=1}^{n}a_{ij}x_j\Big|\le\sum_{i=1}^{n}\sum_{j=1}^{n}|a_{ij}||x_j|=\sum_{j=1}^{n}\Big(\sum_{i=1}^{n}|a_{ij}|\Big)|x_j|\\ &\le\Big(\max_{1\le j\le n}\sum_{i=1}^{n}|a_{ij}|\Big)\|x\|_1\,.\end{aligned}$$

Por otro lado, sea k una columna de A para la que la suma de los valores absolutos de sus elementos es máxima, es decir que

$$\max_{1\le j\le n}\sum_{i=1}^{n}|a_{ij}| = \sum_{i=1}^{n}|a_{ik}|.$$

Entonces para cualquier vector $x = (0,\dots,0,x_k,0,\dots,0)$ todas cuyas componentes sean cero excepto la del lugar k, se verifica (ya que $\|x\|_1 = |x_k|$)

$$\|Ax\|_1 = \sum_{i=1}^{n}|a_{ik}x_k| = \Big(\sum_{i=1}^{n}|a_{ik}|\Big)|x_k| = \Big(\max_{1\le j\le n}\sum_{i=1}^{n}|a_{ij}|\Big)\|x\|_1 .$$

En consecuencia

$$\|A\|_1 = \max_{1\le j\le n}\sum_{i=1}^{n}|a_{ij}| .$$

Ejercicio 2.14 Sabemos que $c(A) = \|A\|\|A^{-1}\|$. Por otro lado, según el ejercicio 2.11, $\|A\| = \max_{\|x\|=1}\|Ax\|$, y análogamente

$$\|A^{-1}\| = \max_{y\neq 0}\frac{\|A^{-1}y\|}{\|y\|} = \max_{x\neq 0}\frac{\|x\|}{\|Ax\|} = \frac{1}{\min_{x\neq 0}\frac{\|Ax\|}{\|x\|}} = \frac{1}{\min_{\|x\|=1}\|Ax\|}$$

por lo tanto

$$c(A) = \frac{\max_{\|x\|=1}\|Ax\|}{\min_{\|x\|=1}\|Ax\|}$$

Ejercicio 2.15 $c(A) = 1$ significa que

$$\max_{\|x\|=1}\|Ax\| = \min_{\|x\|=1}\|Ax\|$$

y por lo tanto $\|Ax\| = \text{const.}$ para $\|x\| = 1$. En otras palabras, A conserva las bolas de la norma en cuestión, lo cual es decir que conserva la norma salvo un factor constante.

Ejercicio 2.16 Si nos situamos en el plano es geométricamente evidente que las matrices A que conservan una p-norma para $p \neq 2$ son precisamente las matrices del grupo de simetrías del cuadrado y si $p = 2$ las del grupo de simetrías del círculo.

Por otro lado, si A tiene condicionamiento 1 respecto a la norma euclídea entonces puede no conservar las p-normas para $p \neq 2$, pero la imagen de la bola unitaria de una p-norma siempre estará contenida dentro del círculo de radio $\sqrt{2}$ y contendrá al círculo de radio $\frac{1}{2}\sqrt{2}$. En consecuencia, para cualquier x tal que $\|x\|_p = 1$ tendremos $\frac{1}{2}\sqrt{2} \le \|Ax\|_p \le \sqrt{2}$ y por lo tanto la condición de A relativa a la norma p verifica $c_p(A) \le \dfrac{\sqrt{2}}{\frac{1}{2}\sqrt{2}} = 2$.

Ejercicio 2.17 Sea A una matriz cuadrada y sea D la matriz diagonal cuya diagonal es igual a la diagonal de A. Suponemos que los elementos de esa diagonal son todos no nulos (en caso contrario A no puede ser de diagonal estrictamente dominante). Dividiendo cada elemento de A por el valor absoluto del elemento diagonal de su fila obtenemos una matriz B que es $B = D^{-1}A$, es decir, $A = DB$. Basta demostrar que si A es de diagonal estrictamente dominante entonces B es inversible. Pero decir que A es de diagonal estrictamente dominante es equivalente a decir que $\|B-I\|_\infty < 1$, lo cual, por el teorema 13 implica que $-B$ no es singular.

Ejercicio 2.18

Ejercicio 2.19

Ejercicio 2.20 $Gx + d = -A_1^{-1}A_2x + A_1^{-1}b = -A_1^{-1}(A_2x - b)$ luego $x = Gx + d$ es equivalente a $-A_1x = A_2x - b$, que a su vez es equivalente a $b = A_1x + A_2x = Ax$.

Ejercicio 2.21 Una descomposición que siempre es posible se obtiene al elegir $A_1 = I$ (identidad). En el caso de una matriz sin ceros en la diagonal también $A_1 = D = \text{diag}(A)$ es inversible, así como la matriz triangular obtenida al hacer cero en A los elementos que están encima (o debajo) de la diagonal.

Cada una de las descomposiciones anteriores da lugar a un método iterativo:

1. En este caso $G = -I^{-1}(A - I) = I - A$ y $d = I^{-1}b = b$, luego el método iterativo para resolver $Ax = b$ asociado con esta descomposición es hallar un punto fijo de la función $(I-A)x+b = x-Ax+b$. Efectivamente tal punto fijo verifica $x = x - Ax + b$ que es equivalente a $Ax = b$.

2. En el segundo caso tenemos el método de Jacobi.

3. En el tercer caso tenemos el método de Gauss-Seidel.

Ejercicio 2.22 Si tomásemos $A_1 = -(D+\omega L)$ y $A_2 = (1-\omega)D - \omega U$ tendríamos

$$A_1+A_2 = -(D+\omega L)+(1-\omega)D-\omega U = -D-\omega L+D-\omega D-\omega U = -\omega A$$

mientras que lo que queremos es $A_1 + A_2 = A$. Por tanto tenemos que dividir los valores que habíamos dado a A_1 y a A_2 por $-\omega$ y entonces tomaremos

$$A_1 = (D+\omega L)/\omega = L + \tfrac{1}{\omega}D$$

y

$$A_2 = -[(1-\omega)D - \omega U]/\omega = U - \left(\tfrac{1-\omega}{\omega}\right)D\,.$$

De esta forma tenemos $-A_1^{-1}A_2 = -\omega(D+\omega L)^{-1}(U - \left(\frac{1-\omega}{\omega}\right)D) = -(D+\omega L)^{-1}(\omega U - (1-\omega)D) = G$, y $A_1 + A_2 = A$. por lo tanto

$$d = A_1^{-1}b = \boxed{\omega(D+\omega L)^{-1}b}\,.$$

Capítulo 3

Cálculo de vectores y valores propios

El problema general de encontrar todos los valores y vectores propios de una matriz cuadrada es bastante amplio y su estudio podría llenar fácilmente un trimestre completo. En este capítulo haremos solamente una mera introducción al tema. Veremos el *método de la potencia* para el cálculo del autovalor de mayor valor absoluto de una matriz que tenga un autovalor dominante y, tras explicar la transformación, mediante matrices de Householder, de una matriz en otra semejante a ella pero de tipo Hessenberg, se explicará el llamado método de la *factorización ortogonal* (o método QR), el cual es el método general más eficaz y de mayor uso para el cálculo de todos los valores propios de una matriz arbitraria. Este método, descubierto por J. G. F. Francis y publicado en 1961, ofrece ciertas dificultades computacionales que se pueden reducir si se prepara previamente la matriz dada A reduciéndola a una forma de tipo *Hessenberg*, proceso que, por conservar la simetría, da lugar a una matriz tridiagonal semejante a A si ésta es simétrica.

3.1 Método de la potencia

El método de la potencia es un método iterativo de cálculo del autovalor de máximo valor absoluto de una matriz diagonalizable. Está basado en la siguiente observación:

Supongamos que A es una matriz diagonalizable de orden n, $x^{(0)}$ un vector arbitrario y $\{x^{(k)}\}$ la sucesión de vectores definida recurrentemente por:

$$x^{(k)} = Ax^{(k-1)} = A^k x^{(0)}.$$

Por la hipótesis de que A es diagonalizable existe una base, que denotaremos $\{u_1, u_2, \ldots, u_n\}$, formada por vectores propios de A, que podemos suponer dados en orden decreciente de magnitud de autovalores, es decir, si λ_i es el autovalor correspondiente al vector propio u_i,

$$|\lambda_1| \geq |\lambda_2| \geq \cdots \geq |\lambda_n|.$$

Entonces, si $x^{(0)} = \alpha_1 u_1 + \cdots + \alpha_n u_n$, el término general de la sucesión anterior tiene la siguiente expresión en la base $\{u_i\}$ de vectores propios de A:

$$\begin{aligned} x^{(k)} &= A^k(\alpha_1 u_1 + \alpha_2 u_2 + \cdots + \alpha_n u_n) \\ &= \alpha_1 \lambda_1^k u_1 + \alpha_2 \lambda_2^k u_2 + \cdots + \alpha_n \lambda_n^k u_n \\ &= \lambda_1^k \left(\alpha_1 u_1 + \alpha_2 \left(\tfrac{\lambda_2}{\lambda_1}\right)^k u_2 + \cdots + \alpha_n \left(\tfrac{\lambda_n}{\lambda_1}\right)^k u_n \right). \end{aligned}$$

Supongamos ahora por un momento que el vector $x^{(0)}$ no es arbitrario como hemos dicho sino que tiene la propiedad de que $\alpha_1 \neq 0$ [1]. Si, además, la matriz A es tal que tiene un *autovalor dominante* (es decir, $\lambda_1 \neq \lambda_2$, o sea, la desigualdad $|\lambda_1| \geq |\lambda_2|$ es estricta de modo que $|\lambda_1| > |\lambda_2|$) entonces tenemos

$$\lim_{k\to\infty} \frac{1}{\lambda_1^k} x^{(k)} = \alpha_1 u_1,$$

[1] Esta hipótesis no se puede comprobar de antemano porque se desconocen los vectores u_i. Sin embargo no es una hipótesis demasiado restrictiva porque, por un lado, la probabilidad de que un vector elegido al azar tenga $\alpha_1 = 0$ es cero y, por otro lado, aunque fuese $\alpha_1 = 0$, a lo largo de los cálculos los errores de redondeo acabarían por introducir una componente no nula en la dirección de u_1.

y por lo tanto para cualquier índice $i \in \{1, \ldots, n\}$,

$$1 = \lim_{k\to\infty} \left(\frac{1}{\lambda_1^k} x_i^{(k)}\right) \Big/ \left(\frac{1}{\lambda_1^{k-1}} x_i^{(k-1)}\right) = \frac{1}{\lambda_1} \lim_{k\to\infty} \frac{x_i^{(k)}}{x_i^{(k-1)}},$$

de donde,

$$\lim_{k\to\infty} \frac{x_i^{(k)}}{x_i^{(k-1)}} = \lambda_1,$$

con lo que tenemos una sucesión que converge al autovalor dominante de la matriz A.

Si ya conocemos el signo del autovalor buscado (o su argumento, en caso de que sea complejo) sólo necesitamos hallar su valor absoluto. Para esto podemos obtener una sucesión que tiene mejor velocidad de convergencia que la anterior ya que

$$\lim_{k\to\infty} \frac{\|A^k x^{(0)}\|}{\|A^{k-1} x^{(0)}\|} = \lim_{k\to\infty} \frac{\|x^{(k)}\|}{\|x^{(k-1)}\|} = |\lambda_1|.$$

Lo que acabamos de decir significa que si A es una matriz cuadrada de orden n con un autovalor dominante, éste se puede hallar mediante el siguiente proceso iterativo:

Algoritmo del Método de la Potencia

1 Elegir un vector inicial $v \neq 0$ y una tolerancia de error T. Además, inicializar la variable $\lambda_0 = 0$.
2 Calcular el producto $y = Av$.
3 Hallar j tal que $|y_j| = \max_{1\leq i\leq n} |y_i|$.
4 Asignar $S = |y_j|$; $\lambda = y_j/v_j$; $v = y/S$.
5 Si $|\lambda_0 - \lambda| \leq T$ entonces PARAR.
6 Asignar $\lambda_0 = \lambda$ e ir al paso **2**.

3.2 Método QR o de la factorización ortogonal

Cuando se desean conocer todos los autovalores de una matriz de orden grande, el cálculo de su polinomio característico y subsiguiente estimación

de raíces no es siempre el método más sencillo o eficaz. El método recomendado en general es el método de la factorización ortogonal que se basa en los siguientes teoremas:

Teorema 17 (Exist. y unicid. de la factorización ortogonal) *Para toda matriz real A existen matrices Q, ortogonal, y R, triangular superior, tales que $A = QR$. Además, si A es no singular entonces estas matrices están unívocamente determinadas por A excepto en el signo de los elementos.*

Demostración: *Existencia*: Esto se demostrará en la sección siguiente en la que se da un método de construcción de las matrices Q y R.
Unicidad: En primer lugar, la matriz identidad no admite más factorización ortogonal $I = QR$ que la trivial, o sea, aquella con $Q = I = R$ ya que $Q^{-1} = R$ nos dice que R es una matriz ortogonal triangular, lo que implica que es la identidad salvo, quizás por los signos de los elementos diagonales. En segundo lugar, si $A = QR = Q'R'$ entonces $I = Q^{-1}Q'R'R^{-1}$, de donde $R'R^{-1} = S$, una matriz de signos, que solo difiere de la identidad en los signos de sus elementos y por tanto $S = S^{-1} = Q^{-1}Q'$ lo que implica $R' = SR$ y $Q' = QS$. ■

3.2.1 La Sucesión General del Método de la Factorización Ortogonal

Una vez que determinemos un método de construir una factorización ortogonal de una matriz cuadrada cualquiera, podemos utilizarlo para construir la siguiente sucesión definida recursivamente a partir de una matriz cuadrada cualquiera A:

$$A^{(0)} = A, \quad A^{(k+1)} = R^{(k)}Q^{(k)}$$

donde $R^{(k)}$ y $Q^{(k)}$ se obtienen mediante la factorización ortogonal

$$A^{(k)} = Q^{(k)}R^{(k)}$$

de la matriz $A^{(k)}$

Llamaremos a esta sucesión *la sucesión general del método de la factorización ortogonal determinada por A*. Esta sucesión goza de algunas propiedades evidentes pero de gran importancia:

1. Dado que dos términos consecutivos están relacionados por:

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

Todas las matrices de esta sucesión son semejantes entre sí y por tanto todas tienen los mismos autovalores y éstos son también los autovalores de su límite

2. Dado que la inversa de una matriz ortogonal es su traspuesta, tenemos

$$\left(A^{(k+1)}\right)^t = \left(Q^{(k)}\right)^t\left(A^{(k)}\right)^tQ^{(k)},$$

y por tanto *si una matriz de esta sucesión es simétrica entonces todas lo son*.

Visto esto podemos enunciar el teorema en el que se basa el método de la factorización ortogonal:

Teorema 18 (Método QR para cálculo de autovalores) *Sea A una matriz real cuadrada no singular entre cuyos autovalores no hay dos con el mismo valor absoluto, es decir, que verifican*

$$|\lambda_1| > |\lambda_2| > \cdots > |\lambda_n| > 0.$$

Entonces la sucesión general del método de la factorización ortogonal determinada por A converge a una matriz triangular. Si la matriz A es simétrica, el límite es una matriz diagonal. ■

Así pues, para una matriz A cuyos autovalores tengan todos distinto valor absoluto, estos son los elementos diagonales límite de la *sucesión general del método de la factorización ortogonal* determinada por A.

Ejercicio 3.1
Demostrar que la sucesión general del método de la factorización ortogonal *determinada por $A = \left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$ no converge a una matriz triangular. ¿Cuál de las hipótesis del teorema anterior no cumple esta matriz?*

3.2.2 La factorización ortogonal de una matriz: Método de las rotaciones planas

La factorización ortogonal de una matriz puede realizarse por un método que es formalmente parecido al método de eliminación de Gauss utilizado

para realizar la factorización triangular. Al igual que en aquél la idea es aplicar sucesivas transformaciones a la matriz de partida para ir haciendo cero los elementos bajo la diagonal de forma que al final quede transformada en una matriz triangular superior. La diferencia es que si allí se utilizaban transformaciones elementales (que combinadas resultaban en una matriz triangular inferior con unos en la diagonal), aquí se utilizarán transformaciones ortogonales (concretamente rotaciones planas) que combinadas darán lugar a una matriz ortogonal.

Una rotación plana en el espacio n-dimensional queda determinada por el plano de rotación y el ángulo de rotación. Si el plano de rotación es el determinado por los ejes de coordenadas 1 y 2 y el ángulo de rotación es α entonces la matriz de rotación tiene la forma

$$P_\alpha(1,2) = \begin{pmatrix} \cos\alpha & -\operatorname{sen}\alpha & & & \\ \operatorname{sen}\alpha & \cos\alpha & & 0 & \\ & & 1 & & \\ & 0 & & \ddots & \\ & & & & 1 \end{pmatrix}$$

y en general una rotación sobre el plano de los ejes i y j (con $i > j$) tiene una matriz de la forma (escribiendo sólo los elementos no nulos)

$$\begin{array}{l} \\ \text{fila } j \to \\ \\ \\ \text{fila } i \to \\ \\ \\ \\ \end{array} \begin{pmatrix} 1 & & & & & & & & \\ & \ddots & & & & & & & \\ & & 1 & & & & & & \\ & & & c & \cdots & -s & & & \\ & & & & 1 & & & & \\ & & & \vdots & \ddots & \vdots & & & \\ & & & & & 1 & & & \\ & & & s & \cdots & c & & & \\ & & & & & & 1 & & \\ & & & & & & & \ddots & \\ & & & & & & & & 1 \end{pmatrix} = P_\alpha(i,j)$$

donde c y s son los cosenos directores de la rotación, es decir, el coseno y seno respectivamente del ángulo α de rotación.

Si queremos hallar los cosenos directores c, s de la matriz $P(i,j)$ para hacer cero el elemento (i,j) de la matriz $A = (a_{i,j})$ no tenemos más que

considerar la ecuación obtenida al multiplicar $P(i,j)$ por la columna j de A, es decir,

$$\begin{array}{l} \\ j \to \\ \\ \\ i \to \\ \\ \\ \\ \end{array} \begin{pmatrix} 1 & & & & & & & & \\ & \ddots & & & & & & & \\ & & 1 & & & & & & \\ & & & c & \cdots & -s & & & \\ & & & & 1 & & & & \\ & & & \vdots & \ddots & \vdots & & & \\ & & & & & 1 & & & \\ & & & s & \cdots & c & & & \\ & & & & & & 1 & & \\ & & & & & & & \ddots & \\ & & & & & & & & 1 \end{pmatrix} \begin{pmatrix} a_{1j} \\ \vdots \\ a_{jj} \\ \vdots \\ a_{ij} \\ \vdots \\ a_{nj} \end{pmatrix} = \begin{pmatrix} a_{1j} \\ \vdots \\ a_{j-1\,j} \\ ca_{jj} - sa_{ij} \\ a_{j+1\,j} \\ \vdots \\ a_{i-1\,j} \\ sa_{jj} + ca_{ij} \\ a_{i+1\,j} \\ \vdots \\ a_{nj} \end{pmatrix},$$

de donde se deduce que los cosenos directores c, s han de verificar

$$c^2 + s^2 = 1 \quad \text{y} \quad sa_{jj} + ca_{ij} = 0,$$

y una posible solución de esto es

$$c = \frac{a_{jj}}{\sqrt{a_{jj}^2 + a_{ij}^2}}, \qquad s = \frac{-a_{ij}}{\sqrt{a_{jj}^2 + a_{ij}^2}} \tag{3.1}$$

Teniendo en cuenta lo que acabamos de decir el proceso de eliminación para la factorización ortogonal de una matriz A se lleva a cabo de la siguiente forma: En un primer paso se hacen cero los elementos de la primera columna bajo la diagonal mediante multiplicación sucesiva de la matriz A por matrices de rotación plana sobre los ejes $(1,2), (1,3), \ldots$. Es decir obtenemos $Q_1 A$ donde $Q_1 = P(1,n)P(1,n-1)\cdots P(1,2)$. Después se hacen cero los elementos de la segunda columna bajo la diagonal multiplicando la matriz $Q_1 A$ por matrices de rotación plana sobre los ejes $(2,3), (2,4), \ldots$. Así obtenemos una matriz $Q_2 Q_1 A$ donde $Q_2 = P(2,n)P(2,n-1)\cdots P(2,3)$. Continuando de esta forma se obtiene una matriz triangular superior

$$R = Q_{n-1}\cdots Q_1 A$$

donde todas las matrices Q_i son matrices ortogonales por ser producto de matrices ortogonales. Con esto queda demostrada la existencia de la factorización ortogonal.

Ejercicio 3.2
Hallar el número de matrices de rotación plana que es necesario multiplicar para hallar la factorización de una matriz arbitraria de orden n.

3.2.3 Simplificación que ocurre para las matrices tipo Hessenberg

Supongamos que la matriz A es de tipo Hessenberg, es decir tiene ceros bajo la subdiagonal. Entonces su factorización ortogonal puede llevarse a cabo con solamente $n-1$ rotaciones planas. Las Q_i constan cada una de una sola rotación. Además:

Lema 4 *Si R es una matriz triangular superior y para cada $i \in \{1, \ldots, n-1\}$ $Q_i = P(i, i+1)$ (una rotación sobre el plano de los ejes i e $i+1$) entonces la matriz $RQ_1 \cdots Q_{n-1}$ es una matriz tipo Hessenberg.*

Demostración: Por inducción. ■

La consecuencia de este lema es que si el algoritmo QR es aplicado a una matriz Hessenberg entonces todas las matrices de la sucesión obtenida son de tipo Hessenberg. Veremos ahora cómo podemos transformar por semejanza una matriz arbitraria en otra de tipo Hessenberg.

El resultado del ejercicio 3.2 muestra el enorme coste que tendría la aplicación del método QR a una matriz arbitraria. Afortunadamente es posible reducir drásticamente ese coste gracias a dos cosas: (1) Según acabamos de ver, el algoritmo QR conserva la forma Hessenberg de una matriz (es decir, el tener ceros bajo la subdiagonal), y (2) Toda matriz puede transformarse por semejanza en otra tipo Hessenberg. Veremos cada uno de estos dos hechos a continuación.

3.2.4 Método de las reflexiones de Householder para el cálculo de una matriz Hessenberg semejante a una dada

En esta sección haremos uso del producto interior natural en $\mathbf{R}^n$. Si representamos los vectores en una base ortonormal mediante matrices columna, el producto interior de $\mathbf{x}$ e $\mathbf{y}$, normalmente denotado $\mathbf{x} \cdot \mathbf{y}$ ó $\langle \mathbf{x}, \mathbf{y} \rangle$, es igual al producto de matrices $\mathbf{x}^t\mathbf{y}$ y también igual a $\mathbf{y}^t\mathbf{x}$.

Un vector unitario $\mathbf{w}$ determina un subespacio ortogonal a él y, para cada vector $\mathbf{x}$ podemos calcular su componente en la dirección de $\mathbf{w}$ como $\mathbf{x}_\mathbf{w} = (\mathbf{x} \cdot \mathbf{w})\mathbf{w}$. Este vector (como matriz columna) es igual al producto del escalar $\mathbf{x}^t\mathbf{w}$ por el vector $\mathbf{w}$. Esto se puede expresar como el producto de matrices $\mathbf{w}(\mathbf{x}^t\mathbf{w})$ que se puede reescribir como $\mathbf{w}(\mathbf{w}^t\mathbf{x})$, y usando la propiedad asociativa,

$$\mathbf{x}_\mathbf{w} = (\mathbf{w}\mathbf{w}^t)\mathbf{x}\,.$$

Si un vector unitario $\mathbf{w}$ lo representamos como un "vector columna", la matriz $\mathbf{w}\mathbf{w}^t$ representa la aplicación lineal 'proyección ortogonal sobre la recta de dirección $\mathbf{w}$, o "*componente longitudinal* en la dirección $\mathbf{w}$". La matriz $I - \mathbf{w}\mathbf{w}^t$ representa entonces la "*componente transversal* a la dirección $\mathbf{w}$".

Si a un vector $\mathbf{x}$ le restamos su componente en la dirección de un vector $\mathbf{w}$, obtenemos la proyección ortogonal de $\mathbf{x}$ sobre el hiperplano ortogonal a $\mathbf{w}$. Si a esta proyección le restamos una vez más la componente de $\mathbf{x}$ en la dirección de $\mathbf{w}$, obtenemos la reflexión, $P_\mathbf{w}\mathbf{x}$, de $\mathbf{x}$ sobre el hiperplano ortogonal a $\mathbf{w}$. En consecuencia tenemos

$$P_\mathbf{w}\mathbf{x} = \mathbf{x} - 2\mathbf{x}_\mathbf{w} = \mathbf{x} - 2(\mathbf{w}\mathbf{w}^t)\mathbf{x} = (I - 2\mathbf{w}\mathbf{w}^t)\mathbf{x}$$

de esta forma llegamos a la expresión general de la matriz de la reflexión sobre el hiperplano ortogonal a un vector unitario $\mathbf{w}$:

$$P_\mathbf{w} = I - 2\mathbf{w}\mathbf{w}^t\,.$$

A partir de esta fórmula es sencillo demostrar que las matrices de reflexión son ortogonales:

Ejercicio 3.3
Toda matriz de la forma $P_\mathbf{w} = I - 2\mathbf{w}\mathbf{w}^t$ con $\mathbf{w}$ unitario (matriz de una reflexión referida a una base ortonormal) es simétrica ($P_\mathbf{w}^t = P_\mathbf{w}$) e involutiva ($P_\mathbf{w}^2 = I$), por lo cual es una matriz ortogonal.

Ahora observamos el hecho fundamental que nos permitirá llegar al algoritmo de Householder, esto es: que toda reflexión sobre un hiperplano queda completamente determinada por la imagen de un vector (que no esté contenido en el hiperplano de reflexión), en otras palabras:

Proposición 5 *Para cualesquiera vectores $\mathbf{x}, \mathbf{y}$ no nulos y de igual norma existe una única reflexión P tal que $P\mathbf{x} = \mathbf{y}$.*

Demostración: Primero veamos la unicidad. Esto es consecuencia de que, para toda reflexión P, la dirección $\mathbf{w}$ de (el hiperplano de) reflexión está determinada por P porque es la misma que la del vector $\mathbf{x} - P\mathbf{x}$ para cualquier $\mathbf{x}$ tal que $\mathbf{x} \neq P\mathbf{x}$. Por tanto si dos reflexiones coinciden en un tal

vector, son reflexiones sobre el mismo hiperplano y por tanto son la misma reflexión. Ahora la existencia: Dados dos vectores distintos $\mathbf{x}$ e $\mathbf{y}$ de igual norma, sea P la reflexión sobre el hiperplano ortogonal a la dirección del vector diferencia $\mathbf{x}-\mathbf{y}$ ($\neq 0$) y sea $\mathbf{w}=(\mathbf{x}-\mathbf{y})/\|\mathbf{x}-\mathbf{y}\|$ la dirección de ese hiperplano. Como $(\mathbf{x}-\mathbf{y})/2=\frac{1}{2}\|\mathbf{x}-\mathbf{y}\|\mathbf{w}$ es el centro del rombo de vértices $\mathbf{x}$, $-\mathbf{y}$, $\mathbf{x}-\mathbf{y}$ y el origen, el producto escalar $\mathbf{w}^t\mathbf{x}$ o componente de $\mathbf{x}$ en la dirección $\mathbf{w}$ es igual a la norma de $(\mathbf{x}-\mathbf{y})/2$, por tanto

$$\begin{aligned} P_{\mathbf{w}}\mathbf{x} &= (I-2\mathbf{w}\mathbf{w}^t)\mathbf{x}=\mathbf{x}-2\mathbf{w}\mathbf{w}^t\mathbf{x}=\mathbf{x}-2\mathbf{w}\|(\mathbf{x}-\mathbf{y})/2\| \\ &= \mathbf{x}-\frac{\mathbf{x}-\mathbf{y}}{\|\mathbf{x}-\mathbf{y}\|}\|\mathbf{x}-\mathbf{y}\|=\mathbf{x}-(\mathbf{x}-\mathbf{y})=\mathbf{y}\,. \end{aligned}$$

como queríamos demostrar. ■

La consecuencia de esto es que si x es un vector no nulo de coordenadas $x_1,\ldots,x_n$, y si $k\in\{1,\ldots,n\}$, eligiendo el número S de forma tal que el vector y de coordenadas $x_1,\ldots,x_k,-S,0,\ldots,0$ tenga la misma norma que x (es decir, eligiendo S tal que $S^2=x_{k+1}^2+,\ldots,+x_n^2$), existirá una (única) reflexión P tal que $Px=y$. Ésta es la reflexión de dirección

$$\mathbf{w}=(x-y)/\|x-y\|=\frac{1}{R}\begin{pmatrix}0\\ \vdots\\ 0\\ x_{k+1}+S\\ x_{k+2}\\ \vdots\\ x_n\end{pmatrix}$$

donde

$$\begin{aligned} R=\|x-y\| &= \sqrt{(x_{k+1}+S)^2+x_{k+2}^2+\cdots+x_n^2} \\ &= \sqrt{(x_{k+1}+S)^2+S^2-x_{k+1}^2} \\ &= \sqrt{2(x_{k+1}+S)S} \end{aligned}$$

Para evitar una pérdida de precisión en estos cálculos conviene elegir el signo de S de tal forma que al sumarle x_{k+1} se sumen sus valores absolutos. Esto significa elegir para S el mismo signo que tenga x_{k+1}, de forma que la fórmula de S es:

$$S=\operatorname{sgn}(x_{k+1})\sqrt{x_{k+1}^2+\cdots+x_n^2}\,.$$

Con lo dicho hasta aquí estamos preparados para demostrar el siguiente

Teorema 19 *Para toda matriz cuadrada A existe una matriz ortogonal P tal que el producto PAP^t es una matriz tipo Hessenberg.*

Demostración: Empezamos construyendo una matriz de reflexión, $P_{\mathbf{w}_1}$ que haga cero todos los elementos bajo la subdiagonal de A en la primera columna. Para ello aplicamos lo dicho más arriba al caso de que el vector x es la primera columna de A. Entonces tendremos

$$\mathbf{w}_1=\frac{1}{R_1}\begin{pmatrix}0\\ a_{21}+S_1\\ a_{31}\\ \vdots\\ a_{n1}\end{pmatrix}$$

El producto $P_{\mathbf{w}_1}A$ tendrá todo ceros bajo la subdiagonal en la primera columna. A continuación multiplicamos por la derecha de esta matriz la inversa de $P_{\mathbf{w}_1}$ (que es ella misma) para obtener la matriz $P_{\mathbf{w}_1}AP_{\mathbf{w}_1}$, semejante a A. Ahora bien, como $\mathbf{w}_1$ tiene un cero como primer elemento, la matriz $\mathbf{w}_1{\mathbf{w}_1}^t$ tiene todo ceros tanto en la primera fila como en la primera columna, lo que implica que la matriz $P_{\mathbf{w}_1}$ es de la forma

$$P_{\mathbf{w}_1}=I-2\mathbf{w}_1{\mathbf{w}_1}^t=\begin{pmatrix}1&0&\cdots&0\\ 0&\times&\cdots&\times\\ \vdots&\vdots&&\vdots\\ 0&\times&\cdots&\times\end{pmatrix}$$

Esto implica que al multiplicar $(P_{\mathbf{w}_1}A)$ por $P_{\mathbf{w}_1}$ no se altera la primera columna de $P_{\mathbf{w}_1}A$ con lo cual la matriz $P_{\mathbf{w}_1}AP_{\mathbf{w}_1}$ sigue teniendo todo ceros bajo la subdiagonal en la primera columna. Similarmente construimos una matriz de reflexión, $P_{\mathbf{w}_2}$ que haga cero todos los elementos bajo la subdiagonal de $P_{\mathbf{w}_1}AP_{\mathbf{w}_1}$ en la segunda columna y hallamos el producto $P_{\mathbf{w}_2}P_{\mathbf{w}_1}AP_{\mathbf{w}_1}P_{\mathbf{w}_2}$. Continuando de esta forma, en el paso k, puesto que las

primeras k componentes de $\mathbf{w}_k$ son cero, usamos una reflexión de la forma

$$P_{\mathbf{w}_k} = I - 2\mathbf{w}_k{\mathbf{w}_k}^t = \begin{pmatrix} 1 & & & & \\ & \ddots & & 0 & \\ & & 1 & & \\ & & & \times & \cdots & \times \\ & 0 & & \vdots & & \vdots \\ & & & \times & \cdots & \times \end{pmatrix}$$

por lo cual al multiplicar

$$P_{\mathbf{w}_k}(P_{\mathbf{w}_{k-1}} \cdots P_{\mathbf{w}_1} A P_{\mathbf{w}_1} \cdots P_{\mathbf{w}_{k-1}}) P_{\mathbf{w}_k}$$

no se alteran las k primeras columnas de

$$P_{\mathbf{w}_k}(P_{\mathbf{w}_{k-1}} \cdots P_{\mathbf{w}_1} A P_{\mathbf{w}_1} \cdots P_{\mathbf{w}_{k-1}}),$$

las cuales tienen todo ceros bajo la subdiagonal.

En consecuencia este proceso produce una matriz ortogonal $P = P_{\mathbf{w}_{n-2}} \cdots P_{\mathbf{w}_1}$ tal que PAP^t es de tipo Hessenberg. ■

Para implementar de forma eficaz el algoritmo de Householder, en cada paso del cual se efectúa un producto de la forma $P_{\mathbf{w}} A P_{\mathbf{w}}$, conviene tener en cuenta que dicho producto se puede realizar de la siguiente forma:

Proposición 6 *Si A es una matriz cuadrada y $\mathbf{w}$ es un vector unitario entonces si $\mathbf{v} = A\mathbf{w}$, $\mathbf{z}^t = \mathbf{w}^t A$, y $\lambda = \mathbf{w}^t\mathbf{v}$*

$$P_{\mathbf{w}} A P_{\mathbf{w}} = A + 2(\mathbf{q}\mathbf{w}^t + \mathbf{w}\mathbf{r}^t)$$

donde $\mathbf{q} = \lambda\mathbf{w} - \mathbf{v}$ y $\mathbf{r}^t = \lambda\mathbf{w}^t - \mathbf{z}^t$.

Demostración: Sólo hay que calcular el producto $P_{\mathbf{w}} A P_{\mathbf{w}} = (I - 2\mathbf{w}\mathbf{w}^t)A(I - 2\mathbf{w}\mathbf{w}^t)$:

$$\begin{aligned} P_{\mathbf{w}} A P_{\mathbf{w}} &= (I - 2\mathbf{w}\mathbf{w}^t)(A - 2A\mathbf{w}\mathbf{w}^t) \\ &= (I - 2\mathbf{w}\mathbf{w}^t)(A - 2\mathbf{v}\mathbf{w}^t) \\ &= A - 2\mathbf{v}\mathbf{w}^t - 2\mathbf{w}\mathbf{w}^t A + 4\mathbf{w}(\mathbf{w}^t\mathbf{v})\mathbf{w}^t \\ &= A - 2\mathbf{v}\mathbf{w}^t - 2\mathbf{w}\mathbf{z}^t + 4\lambda\mathbf{w}\mathbf{w}^t \\ &= A + 2\lambda\mathbf{w}\mathbf{w}^t - 2\mathbf{v}\mathbf{w}^t + 2\lambda\mathbf{w}\mathbf{w}^t - 2\mathbf{w}\mathbf{z}^t \\ &= A + 2(\lambda\mathbf{w} - \mathbf{v})\mathbf{w}^t + 2\mathbf{w}(\lambda\mathbf{w}^t - \mathbf{z}^t) \\ &= A + 2\mathbf{q}\mathbf{w}^t + 2\mathbf{w}\mathbf{r}^t, \end{aligned}$$

que da el resultado requerido. ■

Ejercicio 3.4
Usar el resultado anterior para demostrar que si $\mathbf{w}$ es un vector unitario, el producto $P_{\mathbf{w}} A P_{\mathbf{w}}^t$ puede calcularse mediante el siguiente algoritmo:

1. *Para cada $i \in \{1, \ldots, n\}$ hallar $v_i = \sum_j a_{ij} w_j$ y $z_i = \sum_j a_{ji} w_j$.*
2. *Hallar $\lambda = \sum_i w_i v_i$.*
3. *Para cada $i \in \{1, \ldots, n\}$ y cada $j \in \{1, \ldots, n\}$ hallar*

$$a'_{ij} = a_{ij} + 2(2\lambda w_i w_j - w_j v_i - w_i z_j)\,.$$

Ejercicio 3.5
Usando los resultados anteriores escribir un programa que transforme una matriz dada en otra semejante a ella pero de tipo Hessenberg.

3.2.5 Simplificación aplicable a matrices simétricas

En el caso de que queramos transformar una matriz simétrica a forma Hessenberg, es útil observar dos cosas. En primer lugar cada paso del algoritmo de Householder produce una matriz simétrica ya que si A es una matriz simétrica también lo es $P_{\mathbf{w}} A P_{\mathbf{w}}$ por serlo $P_{\mathbf{w}}$. Esto implica que en este caso el resultado es una matriz *tridiagonal*. En segundo lugar, los cálculos de cada paso del algoritmo pueden simplificarse ya que

Proposición 7 *Si A es una matriz simétrica y $\mathbf{w}$ es un vector unitario entonces si $\mathbf{v} = A\mathbf{w}$ y $\lambda = \mathbf{w}^t\mathbf{v}$,*

$$P_{\mathbf{w}} A P_{\mathbf{w}} = A + 2(\mathbf{q}\mathbf{w}^t + \mathbf{w}\mathbf{q}^t)$$

donde $\mathbf{q} = \lambda\mathbf{w} - \mathbf{v}$.

Demostración: Esto es consecuencia directa del resultado de la proposición 6 teniendo en cuenta que si A es simétrica entonces $\mathbf{z}^t = \mathbf{w}^t A = \mathbf{v}^t$ con lo que $\mathbf{r}^t = \lambda\mathbf{w}^t - \mathbf{z}^t = \lambda\mathbf{w}^t - \mathbf{v}^t = \mathbf{q}^t$. ■

3.2.6 Algoritmo rápido para la factorización ortogonal de matrices Hessenberg

Sea A una matriz de tipo Hessenberg de orden n. Para hallar su factorización ortogonal por medio de rotaciones planas son necesarias n rotaciones,

$Q_1, \dots Q_n$, cada una de las cuales anulará un elemento de la subdiagonal de A. Para eliminar el primer elemento usamos la rotación sobre el plano de los ejes $(1, 2)$ dada por

$$Q_1^t = \begin{pmatrix} c_1 & -s_1 & & & \\ s_1 & c_1 & & 0 & \\ & & 1 & & \\ & 0 & & \ddots & \\ & & & & 1 \end{pmatrix}$$

donde los cosenos directores de la rotación, c_1 y s_1, están determinados, como ya hemos visto en las fórmulas (3.1) por

$$c_1 = \frac{a_{11}}{\sqrt{a_{11}^2 + a_{21}^2}}, \qquad s_1 = \frac{-a_{21}}{\sqrt{a_{11}^2 + a_{21}^2}}.$$

Al multiplicar Q_1^t por A obtenemos una matriz que difiere de A a lo sumo en las dos primeras filas y cuyo elemento en posición $(2, 2)$ es $s_1 a_{12} + c_1 a_{22}$. De esta matriz eliminaremos el segundo elemento de la subdiagonal multiplicándola por

$$Q_2^t = \begin{pmatrix} 1 & & & & & \\ & c_2 & -s_2 & & & \\ & s_2 & c_2 & & & \\ & & & 1 & & \\ & & & & \ddots & \\ & & & & & 1 \end{pmatrix}$$

lo cual claramente no altera la primera columna. Continuando de esta manera llegaremos a obtener la matriz triangular superior R como el producto

$$R = Q_{n-1}^t \cdots Q_2^t Q_1^t A$$

lo que implica

$$A = QR$$

donde (teniendo en cuenta que $Q_i^t = Q_i^{-1}$ por ser ortogonal)

$$Q = Q_1 Q_2 \cdots Q_{n-1}.$$

Nos proponemos ahora describir un algoritmo que nos permite calcular la matriz Q en la forma más compacta posible. Para ello estudiamos la estructura de las matrices obtenidas al calcular los sucesivos productos $Q_0 = I, Q_1, Q_1 \cdot Q_2, \dots$ de los cuales el n-ésimo es la propia matriz Q. Las primeras dos matrices (aparte de Q_0) son:

$$Q_1 = \begin{pmatrix} c_1 & s_1 & & & \\ -s_1 & c_1 & & 0 & \\ & & 1 & & \\ & 0 & & \ddots & \\ & & & & 1 \end{pmatrix}, \quad Q_1 Q_2 = \begin{pmatrix} c_1 & s_1 c_2 & s_1 s_2 & & \\ -s_1 & c_1 c_2 & c_1 s_2 & & 0 \\ 0 & -s_2 & c_2 & & \\ & & & 1 & \\ & 0 & & & \ddots \\ & & & & 1 \end{pmatrix}.$$

Ejercicio 3.6
Comprobar que

$$\begin{pmatrix} c_1 & s_1 & 0 \\ -s_1 & c_1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & c_2 & s_2 \\ 0 & -s_2 & c_2 \end{pmatrix} = \begin{pmatrix} c_1 & s_1 c_2 & s_1 s_2 \\ -s_1 & c_1 c_2 & c_1 s_2 \\ 0 & -s_2 & c_2 \end{pmatrix}$$

En esta sucesión la matriz i-ésima (es decir el producto $Q_1 \cdots Q_i$) para $i = 1, \dots n$ se puede obtener de la anterior (el producto $Q_1 \cdots Q_{i-1}$) y de los coeficientes c_i, s_i aplicando los siguientes pasos: (Para $i = 1$ partimos de Q_0 que tomamos siempre igual a la matriz identidad.)

1. Hacer la columna $i + 1$ igual a la columna i multiplicada por s_i.
2. Multiplicar la columna i por c_i.
3. Hacer los elementos de la fila $i+1$ en posición $(i+1, i)$ e $(i+1, i+1)$ (elemento a la derecha de la diagonal y en la diagonal) iguales a $-s_i$ y c_i respectivamente.

Ejercicio 3.7
Comprobar que mediante aplicación de los pasos anteriores se obtiene la transformación (con $i = 3$):

$$\begin{pmatrix} c_1 & s_1 c_2 & s_1 s_2 & \\ -s_1 & c_1 c_2 & c_1 s_2 & \\ 0 & -s_2 & c_2 & \\ & & & 1 \end{pmatrix} \mapsto \begin{pmatrix} c_1 & s_1 c_2 & s_1 s_2 c_3 & s_1 s_2 s_3 \\ -s_1 & c_1 c_2 & c_1 s_2 c_3 & c_1 s_2 s_3 \\ 0 & -s_2 & c_2 c_3 & c_2 s_3 \\ 0 & 0 & -s_3 & c_3 \end{pmatrix}$$

y que el resultado es precisamente $Q_1 Q_2 Q_3$.

Ejercicio 3.8
Hallar el número de multiplicaciones necesarias para obtener la matriz Q (de orden n) mediante iteración de los pasos indicados más arriba a partir de los valores (supuestos conocidos) $(c_1, s_1), \ldots, (c_{n-1}, s_{n-1})$.

Estas observaciones nos llevan al siguiente algoritmo compacto para realizar la factorización ortogonal de una matriz tipo Hessenberg:

Algoritmo de factorización ortogonal de matrices Hessenberg

1 Sea A la matriz dada. Declaramos las variables Q y R como matrices del mismo orden que A.

2 Inicializamos las matrices $Q = I$ y $R = 0$.

3 Inicializamos el contador $k = 0$.

4 Incrementamos el contador $k = k + 1$.

5 Hallamos el valor provisional de la columna k de R:
Para $i = 1, \ldots, k$ **hacemos** $r_{ik} = \sum_{j=1}^{k} q_{ji} a_{jk}$.
Hallamos $r_{k+1\,k} = a_{k+1\,k}$.

6 Calculamos los cosenos directores de este paso:

$$c = r_{kk}/\sqrt{r_{kk}^2 + r_{k+1\,k}^2} \qquad s = -a_{k+1\,k}/\sqrt{r_{kk}^2 + r_{k+1\,k}^2}\,.$$

7 Corregimos la columna k de R:

$$r_{kk} = cr_{kk} - sr_{k+1\,k} \text{ y } r_{k+1\,k} = 0.$$

8 Calculamos la nueva Q haciendo
Para $j = 1, \ldots, k$ **hacemos** $q_{j\,k+1} = sq_{jk}$, $q_{jk} = cq_{jk}$.
Hecho eso $q_{k+1\,k+1} = c$, $q_{k+1\,k} = -s$.

9 Si $k < n - 1$ ir al paso **4**.

10 Si $k = n - 1$ calculamos la última columna de R:
Para $i = 1, \ldots, n$ **hacemos** $r_{in} = \sum_{j=1}^{n} q_{ji} a_{jn}$.

Nótese que este algoritmo admite una pequeña simplificación ya que el valor de $r_{k+1\,k}$ es inicialmente cero y finalmente también cero (paso 7). Ese valor sólo se usa en los pasos 6 y 7 y podemos evitar su introducción usando directamente $a_{k+1\,k}$ en esos pasos. Además debe notarse que el valor de r_{kk} calculado en el paso 7 es igual al denominador usado en el paso 6. Así pues los pasos 5 a 7 pueden sustituirse por los siguientes más eficaces:

5′. Hallamos el valor provisional de la columna k de R:
Para $i = 1, \ldots, k$ hacemos $r_{ik} = \sum_{j=1}^{k} q_{ji} a_{jk}$.

6′. Calculamos los cosenos directores de este paso:

$$t = \sqrt{r_{kk}^2 + a_{k+1\,k}^2} \qquad c = r_{kk}/t \qquad s = -a_{k+1\,k}/t\,.$$

7′. Corregimos la columna k de R: $r_{kk} = t$.

Respuestas a Algunos Ejercicios del Capítulo 3

Ejercicio 3.1 Su sucesión asociada es (por ejemplo) la sucesión constante igual a A porque A es ortogonal. No cumple la hipótesis de los autovalores porque sus autovalores son 1 y -1 (igual valor absoluto). Evidentemente ninguna matriz ortogonal cumple las hipótesis del teorema.

Ejercicio 3.2 Tantas como elementos hay bajo la diagonal, es decir la mitad del número total, n^2, menos los n de la diagonal: $\boxed{n(n-1)/2}$.

Ejercicio 3.3 Sea $P_{\mathbf{w}} = I - 2\mathbf{w}\mathbf{w}^t$. Entonces $P_{\mathbf{w}}^t = I^t - 2(\mathbf{w}\mathbf{w}^t)^t = I - 2(\mathbf{w}^t)^t\mathbf{w}^t = I - 2\mathbf{w}\mathbf{w}^t = P_{\mathbf{w}}$, luego $P_{\mathbf{w}}$ es simétrica. Además $P_{\mathbf{w}}^2 = (I - 2\mathbf{w}\mathbf{w}^t)^2 = I - 2I2\mathbf{w}\mathbf{w}^t + (2\mathbf{w}\mathbf{w}^t)^2 = I - 4\mathbf{w}\mathbf{w}^t + 4\mathbf{w}(\mathbf{w}^t\mathbf{w})\mathbf{w}^t = I - 4\mathbf{w}\mathbf{w}^t + 4\mathbf{w}\mathbf{w}^t = I$ donde se usa que $(\mathbf{w}^t\mathbf{w}) = 1$ por ser $\mathbf{w}$ un vector unitario.

Ejercicio 3.4 La respuesta es un mero e inmediato cálculo.

Ejercicio 3.5 El siguiente algoritmo transforma una matriz cualquiera en otra Hessenberg. Suponemos dados: n, el orden de la matriz, y $A(i, j)$, la matriz dada. La solución se almacenará en la misma variable A.

```
 para k = 1 hasta n-2
********* CALCULAMOS LA DIRECCION DE REFLEXION *********
 S = 0
 para i = k+1 hasta n
 S = S + A(i, k)^2
 siguiente i
 S = SGN(A(k+1, k))*SQRT(S)
 R = SQRT(2*(A(k+1, k)+S)*S)
 para j = 1 hasta k
 w(j) = 0
 siguiente j
 w(k+1) = (A(k+1, k)+S)/R
 para j = k+2 hasta n
 w(j) = A(j, k)/R
 siguiente j
********* CALCULAMOS LOS VECTORES v Y z *********
 para i = 1 hasta n
```



```
 v(i) = 0; z(i) = 0
 para j = 1 hasta n
  v(i) = v(i) + A(i, j)*w(j)
  z(i) = z(i) + A(j, i)*w(j)
 siguiente j
 siguiente i
********* CALCULAMOS EL ESCALAR lambda *********
 lambda = 0
 para j = 1 hasta n
 lambda = lambda + v(j)*w(j)
 siguiente j
********* CALCULAMOS LA NUEVA MATRIZ A *********
 para i = 1 hasta n
 para j = 1 hasta n
  t = 2*lambda*w(i)*w(j) - v(i)*w(j) - w(i)*z(j)
  A(i, j) = A(i, j) + 2*t
 siguiente j
 siguiente i
 siguiente k
```

Ejercicio 3.6

Ejercicio 3.7

Ejercicio 3.8 Las dos comprobaciones son rutinarias. El número de operaciones es la suma de las operaciones realizadas en cada paso. En el paso k, en el que se pasa de una matriz $(k-1)\times(k-1)$ a una $k\times k$ se realizan $2(k-1)$ multiplicaciones (supuesto $k \geq 3$) y por tanto el número total es

$$\sum_{k=3}^{n} 2(k-1) = 2(2+\cdots+(n-1)) = n(n-1)-2 = n^2-n-2$$

que da el valor (correcto) 0 para el menor valor posible de n (a saber: $n = 2$).