TEMA 4. MÉTODOS DIRECTOS PARA ECUACIONES LINEALES

14 de noviembre de 2002

CAPÍTULO 4

MÉTODOS DIRECTOS PARA LA RESOLUCIÓN DE ECUACIONES ALGEBRAICAS LINEALES

4.1 Aplicaciones en Ingeniería

Son muchas las aplicaciones en ingeniería de la resolución de sistemas lineales. Veamos algunos ejemplos concretos.

Circuito eléctrico pasivo. En la figura 4.1 aparece un circuito eléctrico con una fuente de tensión y tres resistencias. Aplicando las leyes de Kirchoff de los nudos y la ley de Ohm, el voltaje es el producto de la resistencia por la corriente ($V = R\,I$), obtenemos fácilmente el sistema de ecuaciones lineales

$$R_2\,I_2 = R_3\,I_3,$$

$$V - R_1\,I_1 = R_2\,I_2,$$

$$I_1 = I_2 + I_3,$$

que se puede escribir de forma matricial como

$$\begin{pmatrix} 1 & -1 & -1 \\ R_1 & R_2 & 0 \\ 0 & R_2 & -R_3 \end{pmatrix} \begin{pmatrix} I_1 \\ I_2 \\ I_3 \end{pmatrix} = \begin{pmatrix} 0 \\ V \\ 0 \end{pmatrix}.$$

Análisis estático de una estructura. En la parte izquierda de la figura 4.2 aparece una estructura formada por enlaces inextensibles unidos por bornes fijos, que se asemeja a un puente

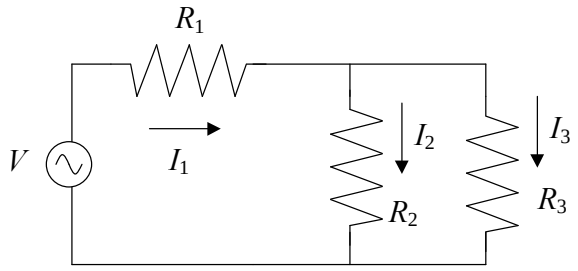

Figura 4.1. Circuito eléctrico simple sólo con resistencias.

con un extremo fijo y otro libre para realizar desplazamientos horizontales y sometida a dos fuerzas (pesos) aplicados en dos de sus nodos. Se pueden calcular las tensiones a las que están sometidos los enlaces de la estructura cuando ésta está en equilibrio aplicando la ley de Newton de acción y reacción en los nodos de la estructura, en los que sabemos que la suma de fuerzas será nula.

Considere un nodo como el que aparece en la parte derecha de la figura 4.2. El equilibrio de fuerzas verticales y horizontales conduce a las siguientes expresiones

$$\cos \tfrac{\pi}{6}\, B_1 + \cos \tfrac{\pi}{6}\, B_2 + P_1 = 0,$$

$$\sin \tfrac{\pi}{6}\, B_1 = \sin \tfrac{\pi}{6}\, B_2 + B_3,$$

respectivamente. De manera similar para los otros dos nodos obtenemos

$$\cos \tfrac{\pi}{6}\, B_4 + \cos \tfrac{\pi}{6}\, B_5 + P_2 = 0,$$

$$\sin \tfrac{\pi}{6}\, B_4 = \sin \tfrac{\pi}{6}\, B_5 + B_3,$$

$$\cos \tfrac{\pi}{6}\, B_2 + \cos \tfrac{\pi}{6}\, B_4 = 0,$$

$$\sin \tfrac{\pi}{6}\, B_2 + B_7 = \sin \tfrac{\pi}{6}\, B_4 + B_6.$$

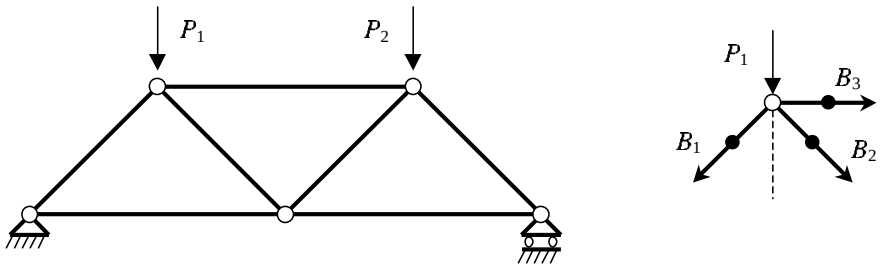

Figura 4.2. Estructura similar a un puente con un extremo libre.

De esta manera se obtiene el siguiente sistema de ecuaciones lineales

$$\begin{pmatrix} \frac{\sqrt{3}}{2} & \frac{\sqrt{3}}{2} & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{2} & -\frac{1}{2} & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{\sqrt{3}}{2} & \frac{\sqrt{3}}{2} & 0 & 0 \\ 0 & 0 & -1 & \frac{1}{2} & -\frac{1}{2} & 0 & 0 \\ 0 & \frac{\sqrt{3}}{2} & 0 & \frac{\sqrt{3}}{2} & 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & -\frac{1}{2} & 0 & -1 & 1 \end{pmatrix} \begin{pmatrix} B_1 \\ B_2 \\ B_3 \\ B_4 \\ B_5 \\ B_6 \\ B_7 \end{pmatrix} = \begin{pmatrix} -P_1 \\ 0 \\ -P_2 \\ 0 \\ 0 \\ 0 \end{pmatrix}.$$

Red de distribución de agua. En la figura 4.3 aparece una red distribución de agua de un aljibe, cuya bomba de salida tiene una presión de P_1, que está conectado a tres líneas que están sometidas a diferentes presiones (P_i). En cada nodo la suma de flujos entrantes debe ser igual a los que salen de éste. Además, suponiendo que el flujo es laminar y sigue la ley de Poiseuille, con lo que la distribución de velocidad en cada tubería es parabólica, sabemos que la diferencia de presión entre los extremos de cada tubería es proporcional a la derivada temporal de la masa $\dot{m}$. De esta forma se obtiene un sistema lineal de ecuaciones como el que sigue

$$P_1 - P_2 = \alpha_1 \dot{m}_1,$$

$$P_1 - P_3 = \alpha_2 \dot{m}_2,$$

$$P_1 - P_4 = \alpha_3 \dot{m}_3,$$

$$\dot{m}_0 = \dot{m}_1 + \dot{m}_2 + \dot{m}_3,$$

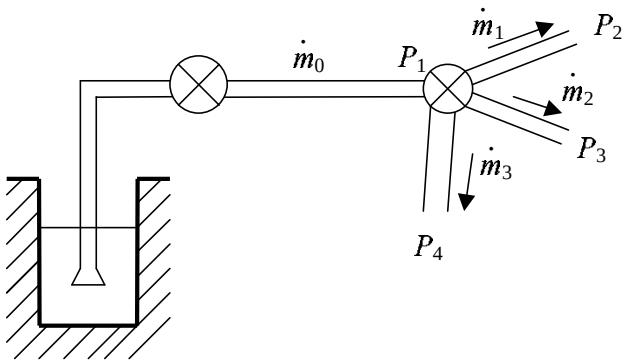

Figura 4.3. Una estructura de tuberías para flujo de Poiseuille.

que se puede escribir de forma matricial como

$$\begin{pmatrix} \alpha_1 & 0 & 0 & 0 \\ 0 & \alpha_2 & 0 & 0 \\ 0 & 0 & \alpha_3 & 0 \\ 1 & -1 & -1 & -1 \end{pmatrix} \begin{pmatrix} \dot{m}_0 \\ \dot{m}_1 \\ \dot{m}_2 \\ \dot{m}_3 \end{pmatrix} = \begin{pmatrix} P_1 - P_2 \\ P_1 - P_3 \\ P_1 - P_4 \\ 0 \end{pmatrix}.$$

En este tema vamos a presentar diferentes métodos para la resolución de este tipo de sistemas de ecuaciones cuando el número de incógnitas no es excesivamente grande o la matriz de coeficientes del sistema tiene una estructura en forma de banda. Estos métodos se denominan directos y obtienen una solución exacta en aritmética exacta, aunque en aritmética flotante están sujetos a propagación de errores cuando el número de operaciones aritméticas a realizar es muy grande. En este último caso se pueden utilizar técnicas iterativas de corrección residual que veremos en el último apartado.

4.2 Sistemas de Ecuaciones Especiales

En esta sección estudiaremos métodos para la resolución de sistemas de ecuaciones lineales que tienen pocas ecuaciones, es decir, cuya matriz de coeficientes tiene muchos ceros. Estos sistemas con estructura permiten métodos numéricos muy eficientes para su resolución.

Para cada método presentaremos un algoritmo para su resolución, contaremos el número de operaciones que involucra, y realizaremos un análisis de errores, tanto hacia atrás, como hacia

adelante. Estos análisis de errores nos permitirán introducir el concepto de condicionamiento del sistema, de gran importancia práctica.

4.2.1 Sistemas de Ecuaciones Diagonales

El sistema de ecuaciones lineales más simple de resolver es el que tiene una matriz de coeficientes diagonal,

$$D\,x = b, \qquad D \in \mathbb{R}^{n\times n}, \qquad x, b \in \mathbb{R}^n,$$

donde D es la matriz de coeficientes diagonal, x es el vector (columna) de incógnitas y b el vector no homogéneo del sistema. Este sistema tendrá solución única si todos los elementos de la diagonal principal de la matriz D son no nulos, $d_{kk} \neq 0$, $k = 1, 2, \ldots, n$.

La solución de este sistema es muy sencilla,

$$x_k = \frac{b_k}{d_{kk}}, \qquad k = 1, 2, \ldots, n.$$

Su coste computacional es exactamente de n operaciones flotantes (divisiones). Utilizando el modelo estándar modificado de la aritmética el vector flotante calculado es

$$fl(x_k) = \widehat{x}_k = \frac{b_k}{d_{kk}\,(1+\delta_k)}, \qquad b_k, d_{kk} \in \mathbb{F}, \quad |\delta_k| \leq u,$$

donde u es la unidad de redondeo (la mitad del épsilon de la máquina). Este algoritmo es estable a errores hacia atrás, ya que

$$fl(x_k) = \frac{b_k}{\widetilde{d}_{kk}}, \qquad \widetilde{d}_{kk} = d_{kk}\,(1+\delta_k),$$

resultado que se puede escribir (matricialmente) como

$$\widetilde{D}\,\widehat{x} = b, \qquad \widetilde{D} = D\,(1+\delta D), \qquad |\delta D| \leq u,$$

donde $\widehat{x} = fl(x)$, y la desigualdad se interpreta componente a componente. Utilizando errores absolutos obtenemos,

$$\widetilde{D}\,\widehat{x} = b, \qquad \widetilde{D} = D + \Delta D, \qquad |\Delta D| \leq u\,|D|.$$

Estos resultados nos confirman que este algoritmo es numéricamente estable.

Algoritmo 4.1 *Resolución de un sistema triangular superior por sustitución regresiva.*

```
function x = resuelveRegresiva (U,b)
% Resuelve U x = b donde
% U es una matriz triangular superior de n x n
% b y x son vectores de n componentes
%
  n=length(b);
  x=zeros(n,1);
  x(n)=b(n)/U(n,n);
  for k=n-1:-1:1,
    x(k)=(b(k)-U(k,k+1:n)*x(k+1:n))/U(k,k);
  end
end
```

4.2.2 Sistemas de Ecuaciones Triangulares

Un sistema de ecuaciones lineal con una matriz de coeficientes triangular superior, $U\,x = b$, toma la forma

$$\begin{aligned} u_{11}\,x_1 + u_{12}\,x_2 + \cdots + u_{1n}\,x_n &= b_1, \\ u_{22}\,x_2 + \cdots + u_{2n}\,x_n &= b_2, \\ \ddots \qquad\qquad & \vdots \\ u_{nn}\,x_n &= b_n. \end{aligned}$$

Resolver este sistema es muy fácil utilizando el algoritmo 4.1 de sustitución hacia atrás o regresiva,

$$\begin{aligned} x_n &= \frac{b_n}{u_{nn}}, \\ x_k &= \frac{1}{u_{kk}} \left(b_k - \sum_{j=k+1}^{n} u_{kj}\,x_j) \right), \qquad k = n-1, n-2, \ldots, 1. \end{aligned}$$

Un sistema de ecuaciones lineal con una matriz de coeficientes triangular inferior, $L\,x = b$, toma la forma

$$l_{11}\,x_1 \qquad\qquad\qquad\qquad = b_1,$$

Algoritmo 4.2 *Resolución de un sistema triangular inferior por sustitución progresiva.*

```
function x = resuelveProgresiva (L,b)
% Resuelve L x = b donde
% L es una matriz triangular inferior de n x n
% b y x son vectores de n componentes
%
  n=length(b);
  x=zeros(n,1);
  x(1)=b(1)/L(1,1);
  for k=1:n-1,
    x(k)=b(k)-L(k,1:k-1)*x(1:k-1);
    x(k)=x(k)/L(k,k);
  end
end
```

$$\begin{aligned} l_{21}\,x_1 + l_{22}\,x_2 \qquad\qquad &= b_2, \\ \vdots \qquad\qquad &\vdots \\ l_{n1}\,x_1 + l_{n2}\,x_2 + \cdots + l_{nn}\,x_n &= b_n. \end{aligned}$$

Resolver este sistema es muy fácil utilizando el algoritmo 4.2 de sustitución hacia adelante o progresiva,

$$\begin{aligned} x_1 &= \frac{b_1}{l_{11}}, \\ x_k &= \frac{1}{l_{kk}} \left(b_k - \sum_{j=1}^{k-1} l_{kj}\,x_j) \right), \qquad k = 2, 3, \ldots, n. \end{aligned}$$

El número de operaciones aritméticas para las sustituciones regresiva y progresiva es exactamente el mismo. Consideremos la primera. El número de divisiones necesarias para la sustitución regresiva es de $C_{dU}(n) = n$, y el número de sumas y productos coincide y es igual a[1]

$$C_{sU}(n) = C_{pU}(n) = 1 + 2 + \cdots + (n-1) = \frac{n\,(n-1)}{2}.$$

[1]El lector conoce de cursos anteriores que

$$\sum_{j=1}^{n} j = \frac{(n+1)\,n}{2}.$$

Algoritmo 4.3 *Algoritmo de sustitución progresiva.*

```
function y = sustitucionProgresiva (c,a,b)
% Evalua y=(c-a(1:k-1)*b(k-1))/b(k)
% donde a es un vector de k-1 componentes
%       b es un vector de k componentes
%       c, y son escalares
%
  k = length(b);
  s = c;
  for i=1:k-1;
     s = c - a(i)*b(i);
  end
  y = s/b(k);
end
```

Con ello, el número total de operaciones

$$C_U(n) = C_{dU}(n) + C_{sU}(n) + C_{pU}(n) = n^2 = O\big(n^2\big)\,.$$

Al estudiar la estabilidad numérica y el comportamiento ante errores de redondeo de estos algoritmos, basta estudiar uno de ellos. Consideraremos la sustitución progresiva o hacia adelante (solución de $L\,x = b$). Necesitaremos el siguiente lema.

Lema 4.1 *Considere la evaluación en aritmética flotante de*

$$y = \frac{c - \displaystyle\sum_{i=1}^{k-1} a_i\, b_i}{b_k}, \qquad c, a_i, b_i \in \mathbb{F},$$

utilizando el algoritmo 4.3. Entonces el resultado calculado cumple

$$b_k\, \widehat{y}\,(1+\theta_k) = c - \sum_{i=1}^{k-1} a_i\, b_i\,(1+\theta_i),$$

donde $|\theta_i| \leq \gamma_i = i\,u/(1-i\,u) = i\,\tilde{u}$, *$u$ la unidad de redondeo, y $\tilde{u}$ es la unidad de redondeo ajustada.*

La demostración es sencilla utilizando los resultados obtenidos en el tema 2 para el error en el producto interior de dos vectores en el marco del modelo estándar de la aritmética. Llamando

$\widehat{s}^{(i)}$ a la suma parcial en el paso i-ésimo, tenemos

$$\widehat{s}^{(0)} = c,$$

$$\widehat{s}^{(i)} = fl(\widehat{s}^{(i-1)} - fl(a_i\, b_i)) = (\widehat{s}^{(i-1)} - a_i\, b_i\, (1+\epsilon_i))\, (1+\delta_i),$$

donde $|\epsilon_i| < u$, es el error relativo en el producto y $|\delta_i| < u$, el de la suma. Iterando sucesivamente obtenemos

$$\widehat{s}^{(k-1)} = c\, (1+\delta_1)\, (1+\delta_2)\, \cdots\, (1+\delta_{k-1}) - \sum_{i=1}^{k-1} a_i\, b_i\, (1+\epsilon_i)\, (1+\delta_i)\, (1+\delta_{i+1})\, \cdots\, (1+\delta_{k-1}),$$

de donde obtenemos finalmente,

$$\widehat{y} = \widehat{s}^{(k)} = fl\left(\frac{\widehat{s}^{(k-1)}}{b_k}\right) = \frac{\widehat{s}^{(k-1)}}{b_k}\, (1+\delta_k), \qquad |\delta_k| \leq u.$$

De esta forma, dejando c sin perturbar, se tiene

$$\frac{b_k\, \widehat{y}}{(1+\delta_1)\, \cdots\, (1+\delta_k)} = c - \sum_{i=1}^{k-1} \frac{a_i\, b_i\, (1+\epsilon_i)}{(1+\delta_1)\, \cdots\, (1+\delta_{i-1})},$$

expresión, que utilizando los lemas útiles del tema 2, nos permite concluir la demostración deseada

$$b_k\, \widehat{y}\, (1+\theta_k) = c - \sum_{i=1}^{k-1} a_i\, b_i\, (1+\theta_i), \qquad |\theta_i| \leq \gamma_i.$$

El lema anterior se aplica directamente a la sustitución progresiva

$$x_k = \frac{1}{l_{kk}} \left(b_k - \sum_{j=1}^{k-1} l_{kj}\, x_j \right),$$

sustituyendo

$$y \equiv x_k, \qquad c \equiv b_k, \qquad a_j \equiv \widehat{x}_j, \qquad b_j \equiv l_{kj},$$

con lo que

$$l_{kk}\, \widehat{x}_k\, (1+\theta_k^{(k)}) = b_k - \sum_{j=1}^{k-1} l_{kj}\, \widehat{x}_j\, (1+\theta_j^{(k)}), \qquad |\theta_j^{(k)}| \leq \gamma_j = j\, \tilde{u}.$$

Hemos obtenido un análisis de errores hacia atrás, ya que perturbando los elementos de L de la forma

$$\widetilde{l}_{kj} = l_{kj}\, (1+\theta_j^{(k)}),$$

operando con aritmética exacta, obtenemos el resultado en aritmética flotante. En resumen, hemos obtenido el siguiente lema.

Lema 4.2 *La sustitución progresiva para la resolución de un sistema triangular inferior conduce a la siguiente cota de errores regresivos*

$$(L + \Delta L)\,\widehat{x} = b, \qquad |\Delta l_{ij}| \leq \gamma_j\,|l_{ij}|,$$

es decir,

$$|\Delta L| \leq \tilde{u} \begin{pmatrix} |l_{11}| & 0 & 0 & \cdots & 0 \\ |l_{21}| & 2\,|l_{22}| & 0 & & 0 \\ |l_{31}| & 2\,|l_{32}| & 3\,|l_{33}| & & 0 \\ \vdots & \vdots & & \ddots & \vdots \\ |l_{31}| & 2\,|l_{32}| & 3\,|l_{33}| & & n\,|l_{nn}| \end{pmatrix}.$$

Este resultado se cumple para la ordenación de las operaciones que aparece en el algoritmo 4.3. Cuando se utiliza una ordenación cualquiera [4], este resultado se puede reemplazar por[2] $|\Delta L| \leq \gamma_n\,|L|$.

Un análisis similar se puede realizar para la sustitución regresiva, conduciendo al siguiente lema.

Lema 4.3 *La sustitución regresiva para la resolución de un sistema triangular superior conduce a la siguiente cota de errores regresivos*

$$(U + \Delta U)\,\widehat{x} = b, \qquad |\Delta U| \leq \tilde{u} \begin{pmatrix} n\,|u_{11}| & (n-1)\,|u_{12}| & (n-2)\,|u_{13}| & \cdots & |u_{1n}| \\ 0 & (n-1)\,|u_{22}| & (n-2)\,|u_{23}| & \cdots & |u_{2n}| \\ 0 & 0 & (n-2)\,|u_{33}| & \cdots & |u_{3n}| \\ \vdots & \vdots & & \ddots & \vdots \\ 0 & 0 & 0 & & |u_{nn}| \end{pmatrix}.$$

En resumen, para una ordenación cualquiera se puede demostrar el siguiente teorema.

Teorema 4.4 *La resolución de un sistema triangular inferior $T\,x = b$, donde se utiliza sustitución progresiva si $T = L$ o regresiva si $T = U$, garantiza que existe un ΔT tal que*

$$(T + \Delta T)\,\widehat{x} = b, \qquad |\Delta T| \leq \gamma_n\,|T|.$$

[2] Pensándolo un poco, este resultado es razonable, sin embargo, una demostración rigurosa es engorrosa, aunque no difícil [4].

Este teorema nos demuestra que la resolución de sistemas triangulares es un problema numéricamente estable. Sin embargo, la estabilidad numérica no garantiza que los errores hacia adelante o progresivos de la solución sean pequeños.

4.2.3 Condicionamiento de la solución de sistemas lineales

Para analizar el comportamiento de los errores progresivos es necesario estudiar cómo afecta a la solución una perturbación, o error, en la matriz de coeficientes [5]. Consideremos un problema general $A\,x = b$ con $\det(A) \neq 0$, perturbando la matriz de coeficientes con un error relativo acotado por ϵ, tenemos

$$(A + \Delta A)\,\widehat{x} = b, \qquad \|\Delta A\| \leq \epsilon \|A\|, \qquad |\epsilon| \ll 1,$$

donde $\widehat{x} = x + \Delta x$. Como

$$b = A\,x = (A + \Delta A)\,\widehat{x}, \qquad \widehat{x} - x = A^{-1}\,\Delta A\,\widehat{x},$$

y aplicando normas

$$\|\delta\widehat{x}\| = \frac{\|\widehat{x} - x\|}{\|\widehat{x}\|} \leq \|A^{-1}\,\Delta A\| \leq \epsilon \|A^{-1}\|\,\|A\|.$$

Ahora bien, hemos acotado el error relativo de $\widehat{x}$, sin embargo, nos gustaría acotar el de x. Utilizando la desigualdad triangular inversa

$$\|\delta\widehat{x}\|\,\|\widehat{x}\| = \|\widehat{x} - x\| \geq \|\widehat{x}\| - \|x\|,$$

con lo que

$$(1 - \|\delta\widehat{x}\|)\,\|\widehat{x}\| \leq \|x\|.$$

Si tuviéramos que $\|\delta\widehat{x}\| < 1$, entonces $(1 - \|\delta\widehat{x}\|) > 0$, y podemos escribir

$$\|\delta x\| = \frac{\|\widehat{x} - x\|}{\|x\|} \leq \frac{\|\widehat{x} - x\|}{(1 - \|\delta\widehat{x}\|)\,\|\widehat{x}\|} = \frac{\|\delta\widehat{x}\|}{1 - \|\delta\widehat{x}\|}.$$

Por tanto, exigiendo que $\|A^{-1}\,\Delta A\| < 1$, lo que es cierto si ϵ es suficientemente pequeño, como $\|\delta\widehat{x}\| \leq \|A^{-1}\,\Delta A\|$, obtenemos el resultado deseado

$$\frac{\|\widehat{x} - x\|}{\|x\|} \leq \frac{\|A^{-1}\,\Delta A\|}{1 - \|A^{-1}\,\Delta A\|}.$$

Para interpretar mejor estas cotas, las debilitaremos un poco,

$$\|\delta\widehat{x}\| = \frac{\|\widehat{x} - x\|}{\|\widehat{x}\|} \leq \|A^{-1}\|\,\|\Delta A\| = \kappa(A)\,\frac{\|\Delta A\|}{\|A\|} = \kappa(A)\,\|\delta A\|,$$

donde a $\kappa(A) = \|A^{-1}\|\,\|A\|$, se le denomina número de condición[3]Informame por correo el día concreto y hora a la que vas a venir a mi despacho (por si me surge algún imprevisto):

Despacho I-323-D, sito en el Antiguo Edificio de Informática, se entra por el Edificio de la Politécnica (mirando a la plaza del ejido con la catedral a la espalda es el blanco colocado a la izquierda); se recorre un pasillo y, casi al final, a mano derecha se sale a un patio, enfrente está el edificio de Informática, de color crema; se sube cinco peldaños de escalera, se entra, a mano derecha se va hacia el ascensor, se sube a la tercera planta, se sale del ascensor; se tuerce a mano izquierda y por un pasillo al fondo, lejos del ascensor, tras doblar una esquina con tablones con notas, está el despacho I-323-D.

Pregunta en conserjería (a la entrada de la Politécnica, o a la entrada del Antiguo Edificio de Informática), por Francisco Villatoro y el despacho I-323-D, y te indicarán con más detalles.

Saludos

PACO de la matriz A. Similarmente, si $\|A^{-1}\,\Delta A\| = \epsilon\,\kappa(A) < 1$, el error relativo de la solución perturbada es

$$\|\delta x\| = \frac{\|\widehat{x} - x\|}{\|x\|} \leq \frac{\kappa(A)\,\|\delta A\|}{1 - \kappa(A)\,\|\delta A\|}.$$

El número de condición de la matriz A nos mide aproximadamente el cociente entre el error relativo del resultado $\|\delta x\|$ y el de los datos $\|\delta A\|$. El número de condición es siempre mayor que la unidad ya que

$$1 \leq \|I\| = \|A^{-1}\,A\| \leq \|A^{-1}\|\,\|A\| = \kappa(A).$$

Cuando este número es grande $\kappa(A) \gg 1$, el error relativo de la solución puede ser extremadamente grande comparado con el de los datos, por lo que se dice que matriz está mal condicionada. Para calcular este número se puede utilizar cualquier norma, dado que en $\mathbb{R}^n$ todas las normas son equivalentes y algo grande en una norma lo será también en cualquier otra. Sin embargo, normalmente se utilizan las normas infinito o uno,

$$\kappa_\infty(A) = \|A^{-1}\|_\infty\,\|A\|_\infty, \qquad \kappa_1(A) = \|A^{-1}\|_1\,\|A\|_1.$$

Si se conocen los autovalores, $A\,x = \lambda\,x$, se puede estimar el número de condición fácilmente. Como

$$\|A\,x\| = |\lambda|\,\|x\| \leq \|A\|\,\|x\|,$$
$$\|A^{-1}\,x\| = \frac{1}{|\lambda|}\,\|x\| \leq \|A^{-1}\|\,\|x\|,$$

[3]Fue introducido por Alan Turign en 1948 utilizando la norma de Frobenius.

tenemos que

$$\|A\| \geq |\lambda|, \qquad \|A^{-1}\| \geq \frac{1}{|\lambda|}, \qquad \forall\, \lambda \equiv \lambda_A,$$

y, por tanto, para los valores máximo y mínimo,

$$\kappa(A) = \|A\|\,\|A^{-1}\| \geq \rho(A)\,\rho(A^{-1}) = \frac{\max|\lambda_A|}{\min|\lambda_A|}.$$

De esta forma observamos que una matriz singular tiene un número de condición infinito.

En la práctica se puede aplicar una regla empírica que nos dice que el orden de magnitud del número de condición indica cuántos decimales de precisión perderemos en el resultado. Esta regla no se puede demostrar rigurosamente, pero se considera razonable en la práctica. Si A tiene t dígitos decimales de precisión,

$$\|\delta A\| = \frac{\|\Delta A\|}{\|A\|} = 10^{-t},$$

y el número de condición es $\kappa(A) = 10^k$, entonces el error relativo en la solución

$$\|\delta x\| = \frac{\|\widehat{x} - x\|}{\|x\|} \leq 10^{-t+k},$$

es decir, $\widehat{x}$ aproxima a la solución correcta con $t - k$ dígitos de precisión.

En el caso de un sistema triangular, el teorema 4.4 garantiza que

$$(T + \Delta T)\,\widehat{x} = b, \qquad |\Delta T| \leq \gamma_n\,|T|,$$

con lo que tenemos $\|x\| \approx \|\widehat{x}\|$,

$$\frac{\|x - \widehat{x}\|}{\|x\|} \leq \frac{\gamma_n\,\kappa(T))}{1 - \gamma_n\,\kappa(T)}$$

Se dice que la matriz está mal condicionada si su número de condición es un número muy grande respecto a la unidad. Incluso en un sistema lineal triangular la matriz de coeficientes puede estar muy mal condicionada. Por ejemplo [4], para la matriz triangular superior $U(\alpha)$ cuyos elementos están dados por

$$(U(\alpha))_{ij} = \begin{cases} 1, & i = j, \\ -\alpha, & i < j, \end{cases}$$

cuya norma es $\|A\|_\infty = 1 + n\,\alpha$, el lector puede comprobar, por inducción, que su inversa es

$$(U(\alpha)^{-1})_{ij} = \begin{cases} 1, & i = j, \\ \alpha\,(1+\alpha)^{j-i-1}, & i < j, \end{cases}$$

con norma dada por

$$\|A^{-1}\|_\infty = 1 + \alpha\,(1 + (1+\alpha) + \cdots + (1+\alpha)^{n-2}) = (1+\alpha)^{n-1}.$$

Por tanto, el número de condición es

$$\kappa_\infty(A) = \|A\|_\infty \,\|A^{-1}\|_\infty = (1 + n\,\alpha)\,(1+\alpha)^{n-1} = O(n\,\alpha^n)\,,$$

que es muy grande para $\alpha \gg 1$. Sin embargo, el lector observará que hay sistemas lineales sencillos para los que el condicionamiento de la matriz de coeficientes no influye en los errores del resultado. Por ejemplo, la solución para el problema $U(\alpha)\,x = e_n$, donde e_n es el n-ésimo vector de la base canónica de $\mathbb{R}^n$, se calcula con total exactitud $x = e_n$.

Este comportamiento, en el que un sistema aparentemente mal condicionado, realmente no lo está, se denomina mal condicionamiento artificial. De hecho, en la práctica se encuentra que este caso no es ni mucho menos excepcional. Normalmente, el número de condición matricial sobreestima pésimamente el comportamiento de los errores. De esta forma, si una matriz está bien condicionada, el problema de resolver el sistema lineal correspondiente también lo estará. Sin embargo, si la matriz está mal condicionada, el problema puede estarlo o no. De hecho, como hemos visto, en el ejemplo del párrafo anterior, el condicionamiento del problema también depende del vector no homogéneo b.

Para conseguir un número de condición que mida mejor el mal condicionamiento de la resolución de sistemas lineales, hay que introducir los números de condición componente a componente como hacemos en el siguiente teorema, que nos relaciona el error relativo en la solución del sistema lineal con una cota del error relativo en la matriz de coeficientes y el vector no homogéneo.

Teorema 4.5 *Sea el problema $A\,x = b$, cuya solución perturbada satisface*

$$(A + \Delta A)\,\widehat{x} = b, \qquad |\Delta A| \le \epsilon\,|A|,$$

asumiendo que $\epsilon\,\||A^{-1}|\,|A|\|_\infty < 1$, entonces se cumple que

$$\frac{\|\widehat{x} - x\|_\infty}{\|x\|_\infty} \le \frac{\epsilon}{1 - \epsilon\,\||A^{-1}|\,|A|\|_\infty}\,\frac{\||A^{-1}|\,|A|\,|x|\|_\infty}{\|x\|_\infty} = \frac{\epsilon\; cond(A,x)}{1 - cond(A)},$$

donde hemos definido el número de condición de Skeel[4]*, componente a componente para el problema, como*

$$cond(A,x) = \frac{\||A^{-1}|\,|A|\,|x|\|_\infty}{\|x\|_\infty},$$

y el número de condición de Skeel, componente a componente para la matriz, como

$$cond(A) = \||A^{-1}|\,|A|\|_\infty.$$

[4]Skeel los introdujo en 1979.

La demostración es sencilla, y sigue la línea de los resultados presentados previamente. El error cumple la ecuación

$$A\,\Delta x = -\Delta A\,x - \Delta A\,\Delta x,$$

donde aplicando valores absolutos componente a componente

$$\begin{aligned} |\Delta x| &= |A^{-1}\,\Delta A\,x + A^{-1}\,\Delta A\,\Delta x| \\ &\leq |A^{-1}|\,|\Delta A|\,|x| + |A^{-1}|\,|\Delta A|\,|\Delta x| \\ &\leq \epsilon\,|A^{-1}|\,|A|\,|x| + \epsilon\,|A^{-1}|\,|A|\,|\Delta x|. \end{aligned}$$

Introduciendo normas en el miembro derecho, obtenemos

$$\frac{|\widehat{x} - x|}{\|x\|_\infty} \leq \epsilon\,\frac{\||A^{-1}|\,|A|\,|x|\|_\infty}{\|x\|_\infty} + \epsilon\,\frac{\||A^{-1}|\,|A|\,|\widehat{x} - x|\|_\infty}{\|x\|_\infty},$$

que es cierta para todas las componentes en el miembro izquierdo, luego también lo será para la componente máxima, para la que

$$\frac{\|\widehat{x} - x\|_\infty}{\|x\|_\infty} \leq \frac{\||A^{-1}|\,|A|\,|x|\|_\infty}{1 - \epsilon\,\||A^{-1}|\,|A|\|}\,\frac{\epsilon}{\|x\|_\infty},$$

que es el resultado que queríamos demostrar. Este teorema no sólo es válido para la norma infinito, sino también para otras, aunque la demostración general es más larga.

Se puede demostrar que $\mathrm{cond}(A) \leq \kappa(A)$. Además, en algunos casos $\mathrm{cond}(A, x) \ll \mathrm{cond}(A)$. Tomemos el siguiente ejemplo debido a Kahan [4],

$$A = \begin{pmatrix} 2 & -1 & 1 \\ -1 & \epsilon & \epsilon \\ 1 & \epsilon & \epsilon \end{pmatrix}, \qquad A^{-1} = \begin{pmatrix} 0 & -\frac{1}{2} & \frac{1}{2} \\ -\frac{1}{2} & \frac{1}{4}\left(\frac{1}{\epsilon} - 2\right) & \frac{1}{4}\left(\frac{1}{\epsilon} + 2\right) \\ \frac{1}{2} & \frac{1}{4}\left(\frac{1}{\epsilon} + 2\right) & \frac{1}{4}\left(\frac{1}{\epsilon} - 2\right) \end{pmatrix},$$

para el que $\kappa_\infty(A) = 2\,(1 + 1/\epsilon)$, indica que la matriz está mal condicionada para $|\epsilon| \ll 1$. Además, como

$$|A^{-1}|\,|A| = \begin{pmatrix} 1 & \epsilon & \epsilon \\ 1 + \frac{1}{2\,\epsilon} & 1 & 1 \\ 1 + \frac{1}{2\,\epsilon} & 1 & 1 \end{pmatrix},$$

tenemos que $\text{cond}(A) = 3 + 1/(2\,\epsilon)$, que también indica mal condicionamiento. El lector puede comprobar fácilmente que $\text{cond}(A) \leq \kappa(A)$. Sin embargo, al resolver el sistema $A\,x = b$, con

$$b = \begin{pmatrix} 2\,(1+\epsilon) \\ -\epsilon \\ \epsilon \end{pmatrix}, \qquad x = A^{-1}\,b = \begin{pmatrix} \epsilon \\ -1 \\ 1 \end{pmatrix},$$

y $\text{cond}(A, x) = 5/2 + \epsilon$, que indica que este problema está bien condicionado, aunque la matriz de coeficientes no lo esté.

Hemos observado como el condicionamiento del problema $A\,x = b$ no sólo depende del número de condición de A, sino también del vector b, así como que este problema puede estar bien condicionado incluso si A está mal condicionada. Más aún, el condicionamiento de una matriz no está directamente relacionado con el de su traspuesta, como muestra el siguiente ejemplo,

$$T = \begin{pmatrix} 1 & 1 & 0 \\ 0 & \epsilon & \epsilon \\ 0 & 0 & 1 \end{pmatrix}, \qquad \kappa(T) = 5, \qquad \kappa(T^{\top}) = 1 + \frac{2}{\epsilon}.$$

Es decir, el problema basado en la matriz traspuesta puede estar más o menos mal condicionado que el problema original.

4.2.4 Más sobre condicionamiento. Los errores en b

En la sección anterior hemos estudiado los efectos de errores en la matriz de coeficientes A utilizando la teoría de perturbaciones. Dicha teoría también nos permite estudiar el efecto de los errores en el término no homogéneo b, y los errores simultáneos en A y en b.

Para el primer caso, consideraremos el sistema lineal perturbado

$$A\,y = b + \Delta b,$$

cuya solución perturbada es

$$\widehat{x} = x + \Delta x.$$

El error absoluto $\Delta x = \widehat{x} - x$ es solución del siguiente sistema lineal

$$A\,\Delta x = \Delta b, \qquad \Delta x = A^{-1}\,\Delta b,$$

por lo que aplicando normas a estas dos expresiones

$$\|A\|\,\|\Delta x\| \geq \|\Delta b\| \quad \Rightarrow \quad \|\Delta x\| \geq \frac{\|\Delta b\|}{\|A\|},$$

$$\|\Delta x\| = \|A^{-1}\,\Delta b\| \quad \Rightarrow \quad \|\Delta x\| \leq \|A^{-1}\|\,\|\Delta b\|,$$

y como para la solución exacta del sistema lineal se cumple que

$$\|A\|\,\|x\| \geq \|b\| \quad \Rightarrow \quad \|x\| \geq \frac{\|b\|}{\|A\|},$$

$$\|x\| = \|A^{-1}\,b\| \quad \Rightarrow \quad \|x\| \leq \|A^{-1}\|\,\|b\|,$$

obtenemos para el error relativo las siguientes acotaciones

$$\frac{\|\Delta x\|}{\|x\|} \geq \frac{\|\Delta b\|}{\|x\|\,\|A\|} \geq \frac{\Delta b}{\|A^{-1}\|\,\|A\|\,\|b\|},$$

$$\frac{\|\Delta x\|}{\|x\|} \leq \frac{\|A^{-1}\|\,\|\Delta b\|}{\|x\|} \leq \frac{\|A^{-1}\|\,\|\Delta b\|}{\|b\|/\|A\|}.$$

De esta forma hemos acotado superior e inferiormente el error relativo de la solución utilizando el error relativo en el término no homogéneo

$$\|A\|\,\|A^{-1}\|\,\frac{\|\Delta b\|}{\|b\|} \geq \frac{\|\Delta x\|}{\|x\|} \geq \frac{1}{\|A\|\,\|A^{-1}\|}\,\frac{\|\Delta b\|}{\|b\|}.$$

Esta acotación nos conduce de nuevo al número de condición de la matriz de coeficientes A,

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

Para el segundo caso, el estudio de los errores simultáneos en A y en b, podemos considerar el sistema lineal

$$(A + \Delta A)\,(x + \Delta x) = A\,x + \Delta A\,x + (A + \Delta A)\,\Delta x = b + \Delta b,$$

donde $A\,x = b$, con lo que

$$\Delta A\,x + (A + \Delta A)\,\Delta x = \Delta b. \tag{4.1}$$

Para acotar esta expresión necesitamos recurrir al siguiente lema.

Lema 4.6 *Sea A una matriz no singular, y su inversa A^{-1}. Si B es una matriz tan próxima a A como*

$$\|A - B\| < \frac{1}{\|A\|^{-1}},$$

entonces dicha matriz tiene inversa y además se cumplen las siguientes acotaciones

$$\|B^{-1}\| \leq \frac{\|A^{-1}\|}{1 - \|A^{-1}\|\,\|(A - B)\|},$$

$$\|A^{-1} - B^{-1}\| \leq \frac{\|A^{-1}\|^2\,\|A - B\|}{1 - \|A^{-1}\|\,\|A - B\|}.$$

Antes de demostrar este lema, debemos indicar que para toda A siempre existe alguna matriz B tan próxima a A como indica la hipótesis del lema. Definiendo $D = A^{-1}(A - B)$, se verifica que

$$B = A - (A - B) = A\,(I - A^{-1}\,(A - B)) = A\,(I - D).$$

Por hipótesis sabemos que

$$1 > \|A^{-1}\|\,\|A - B\| \geq \|A^{-1}\,(A - B)\| = \|D\|,$$

por lo que $1 > \|D\| \geq \rho(D)$, el radio espectral de D, y se puede probar que existe $(I - D)^{-1}$. Por reducción al absurdo, si su determinante fuera nulo, $\det(I - D) = 0$, entonces existiría al menos un $x \neq 0$ tal que $(I - D)\,x = 0$, por lo que 1 sería un autovalor de D (con $D\,x = x$) lo que es imposible ya que $\rho(D) < 1$. Por tanto, $|I - D| \neq 0$, y existe la inversa $(I - D)^{-1}$.

De esta forma, también existe la inversa de B, sea

$$B^{-1} = (I - D)^{-1}\,A^{-1},$$

que podemos fácilmente acotar como

$$\|B^{-1}\| \leq \|A^{-1}\|\,\|(I - D)^{-1}\| \leq \|A^{-1}\|\,\frac{1}{\|I - D\|} \leq \frac{\|A^{-1}\|}{1 - \|A^{-1}\,(A - B)\|},$$

donde se ha utilizado la desigualdad triangular inversa

$$\|I - D\| \geq |1 - \|D\|| = 1 - \|D\|,$$

ya que $\|D\| < 1$.

Finalmente podemos acotar la diferencia entre las dos inversas de la siguiente forma

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

$$\|A^{-1} - B^{-1}\| \leq \|A^{-1}\|\,\|A - B\|\,\|B^{-1}\| \leq \frac{\|A^{-1}\|^2\,\|A - B\|}{1 - \|A^{-1}\|\,\|A - B\|},$$

y con ello queda demostrado el lema.

Volvamos a nuestro sistema perturbado en A y en b. Asumamos que el error en la matriz de coeficientes es suficientemente pequeño como para que se pueda aplicar el lema anterior,

$$\|\Delta A\| < \frac{1}{\|A^{-1}\|},$$

entonces poniendo $B = A + \Delta A$, obtenemos por el lema anterior que existe la inversa $(A + \Delta A)^{-1}$ y su norma está acotada por

$$\|(A + \Delta A)^{-1}\| \leq \frac{\|A^{-1}\|}{1 - \|A^{-1}\|\,\|\Delta A\|}.$$

Por otro lado, la expresión (4.1), que recordamos es

$$\Delta A\, x + (A + \Delta A)\, \Delta x = \Delta b,$$

nos da para el error absoluto de la solución

$$\Delta x = (A + \Delta A)^{-1}\, (\Delta b - \Delta A\, x),$$

que ahora estamos en condiciones de acotar fácilmente

$$\begin{aligned} \|\Delta x\| &\leq \|(A + \Delta A)^{-1}\|\, \|\Delta b - \Delta A\, x\| \\ &\leq \frac{\|A^{-1}\|}{1 - \|A^{-1}\|\, \|\Delta A\|}\, (\|\Delta b\| + \|\Delta A\|\, \|x\|) \\ &= \frac{\kappa(A)}{1 - \kappa(A)\, \|\Delta A\|/\|A\|}\, \left(\frac{\|\Delta b\|}{\|A\|} + \|x\|\, \frac{\|\Delta A\|}{\|A\|} \right). \end{aligned}$$

Ya que $A\, x = b$, $\|A\| \geq \|b\|/\|x\|$, por lo que podemos acotar el error absoluto como

$$\|\Delta x\| \leq \frac{\kappa(A)}{1 - \kappa(A)\, \|\Delta A\|/\|A\|}\, \left(\|x\|\, \frac{\|\Delta b\|}{\|b\|} + \|x\|\, \frac{\|\Delta A\|}{\|A\|} \right),$$

que conduce a la siguiente cota para el error relativo

$$\begin{aligned} \frac{\|\Delta x\|}{\|x\|} &\leq \frac{\kappa(A)}{1 - \kappa(A)\, \|\Delta A\|/\|A\|}\, \left(\frac{\|\Delta b\|}{\|b\|} + \frac{\|\Delta A\|}{\|A\|} \right) \\ &= \left(\kappa(A) + O\left(\frac{\|\Delta A\|}{\|A\|} \right) \right)\, \left(\frac{\|\Delta b\|}{\|b\|} + \frac{\|\Delta A\|}{\|A\|} \right). \end{aligned}$$

Si los errores relativos en A y en b son pequeños podemos despreciar los términos de segundo orden (sus productos), y podemos escribir

$$\frac{\|\Delta x\|}{\|x\|} \leq \kappa(A)\, \left(\frac{\|\Delta b\|}{\|b\|} + \frac{\|\Delta A\|}{\|A\|} \right).$$

Es decir, el error relativo en la solución está acotado por la suma de los errores relativos de los datos (b y A) multiplicados por el número de condicionamiento de la matriz de coeficientes.

Estos resultados nos muestran que el número de condicionamiento permite cuantificar el efecto de errores no sólo en los elementos de la matriz de coeficientes, si no también en el vector no homogéneo b. Aún así, el lector no debe olvidar que el condicionamiento del problema $A\, x = b$ depende del de la matriz A pero también del vector no homogéneo b, como nos muestra el número de condición de Skeel.

4.3 Eliminación de Gauss y Factorización LU

Hay varios métodos que el alumno habrá estudiado en cursos anteriores para la resolución e sistemas lineales con matrices densas. La regla de Cramer es un algoritmo poco utilizado numéricamente dado su alto coste computacional. La eliminación de Gauss es el algoritmo más utilizado para matrices densas generales, aunque su coste limita su aplicabilidad a sistemas de como mucho miles de incógnitas, salvo que la matriz de coeficientes tenga alguna estructura especial que reduzca su coste, como que sea una matriz dispersa (*sparse*) en banda. Estudiaremos como caso especial las matrices tridiagonales por su importancia en las aplicaciones numéricas.

4.3.1 Regla de Cramer

La regla de Cramer para resolver sistemas de ecuaciones algebraicas lineales ya se ha estudiado en cursos anteriores. Para un sistema

$$A\,x = b$$

donde A es la matriz de coeficientes de orden $n \times n$, x es el vector (columna) de n incógnitas y b el vector no homogéneo. Si el determinante $\det(A) \neq 0$, A no es singular, y las n componentes x_i del vector solución x se calculan mediante

$$x_i = \frac{\det([A; b_i])}{\det(A)},$$

donde la notación $[A; b_i]$ denota la matriz que se obtiene cuando se sustituye la columna i-ésima de la matriz A por el vector b.

Para calcular estos determinantes podemos utilizar la regla de los menores que nos dice que el determinante de una matriz se obtiene multiplicando los elementos de una fila (i) o de una columna (j) por los determinantes de los menores asociados a estos elementos. El menor $A_{(ij)}$ asociado al elemento a_{ij} es una matriz de orden $(n-1) \times (n-1)$ que se obtiene de la matriz original eliminando la fila i y la columna j. Escrito en símbolos,

$$\det(A) = \sum_{k=1}^{n} a_{ik}\ \det(A_{(ik)})\,(-1)^{(i+k)} = \sum_{k=1}^{n} a_{kj}\ \det(A_{(kj)})\,(-1)^{(k+j)}.$$

En la regla de Cramer tenemos que calcular $n+1$ determinantes y realizar n cocientes, luego el coste total en número de operaciones es

$$C(n) = n + (n+1)\,C_d(n),$$

donde $C_d(n)$ es el coste de evaluar un determinante de $n \times n$. En uno de los ejercicios resueltos se demuestra que el número de operaciones en el cálculo de un determinante mediante el desarrollo en menores principales requiere un coste computacional de $C_d(n) = O(n!)$, por lo que el coste total de la regla de Cramer es de $C(n) = O((n+1)!)$, que es enorme incluso cuando n no es muy grande. Por ejemplo, para $n = 100$, $(n+1)! = 9.3 \times 10^{157}$.

Además, este gran número de operaciones puede provocar la propagación "perniciosa" de gran número de errores, especialmente por el gran número de posibles cancelaciones catastróficas que pueden aparecer en la evaluación de los determinantes. Estos errores pueden conducir a un resultado completamente sin sentido.

Estos resultados desaconsejan el uso de la regla de Cramer para resolver sistemas lineales con muchas ecuaciones e incógnitas.

4.3.2 Eliminación de Gauss

En cursos anteriores, para la resolución de sistemas lineales, el alumno ha estudiado el algoritmo de eliminación de Gauss, que consiste en transformar la matriz de coeficientes del sistema lineal en una matriz triangular superior realizando combinaciones lineales adecuadas de las ecuaciones del sistema. En esta sección estudiaremos dicho algoritmo desde el punto de vista numérico.

Sea el sistema lineal $A\,x = b$ escrito de la forma

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

Para transformar la matriz A en una matriz triangular superior, se procederá paso a paso, columna a columna. Para la columna i-ésima se realizarán combinaciones lineales entre la ecuación i-ésima y cada una de las restantes ecuaciones j-ésimas, con $j = i+1, i+2, \ldots, n$, de tal forma que se hagan ceros los elementos a_{ji}, correspondientes a la columna i-ésima. Estas combinaciones lineales de ecuaciones afectan tanto a la matriz A como al vector b.

Veamos el proceso en detalle. Como primer paso ($k = 1$), partimos de la matriz original $A^{(1)} = A$, y $b^{(1)} = b$, es decir,

$$a_{ij}^{(1)} = a_{ij}, \qquad b_i^{(1)} = b_i, \qquad 1 \leq i, j \leq n.$$

El segundo paso ($k = 2$) es hacer ceros en los elementos subdiagonales de la primera columna, obteniendo el sistema $A^{(2)}\, x = b^{(2)}$,

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

Para ello, dejamos la primera fila inalterada,

$$a_{1j}^{(2)} = a_{1j}^{(1)}, \qquad 1 \leq j \leq n, \qquad b_1^{(2)} = b_1^{(1)};$$

para la i-ésima fila, $i \geq 2$, la multiplicamos por

$$m_{i1} = \frac{a_{i1}^{(1)}}{a_{11}^{(1)}}, \qquad 2 \leq i \leq n,$$

y se la restamos a la primera fila

$$\begin{aligned} a_{i1}^{(2)} &= a_{i1}^{(1)} - m_{i1}\, a_{11}^{(1)} = 0, \qquad 2 \leq i \leq n, \\ a_{ij}^{(2)} &= a_{ij}^{(1)} - m_{i1}\, a_{1j}^{(1)}, \qquad 2 \leq i, j \leq n, \\ b_i^{(2)} &= b_i^{(1)} - m_{i1}\, b_1^{(1)}, \qquad 2 \leq i \leq n. \end{aligned}$$

A los m_{ik} se les denomina multiplicadores y los elementos $a_{kk}^{(k)}$ pivotes. Para que este procedimiento sea aplicable, los pivotes han de ser no nulos, $a_{kk}^{(k)} \neq 0$.

Supongamos que tras el k-ésimo paso hemos hecho cero en los elementos subdiagonales de

las primeras $(k-1)$ columnas, obteniendo la matriz $A^{(k)}$ cuyos elementos $a_{ij}^{(k)}$ son

$$\begin{pmatrix} a_{11}^{(k)} & \cdots & \cdots & a_{1k}^{(k)} & \cdots & a_{1j}^{(k)} & \cdots & a_{1n}^{(k)} \\ 0 & \ddots & & \vdots & & \vdots & & \vdots \\ \vdots & & \ddots & \vdots & & \vdots & & \vdots \\ 0 & \cdots & 0 & a_{kk}^{(k)} & \cdots & a_{kj}^{(k)} & \cdots & a_{kn}^{(k)} \\ \vdots & & \vdots & \vdots & & \vdots & & \vdots \\ 0 & \cdots & 0 & a_{ik}^{(k)} & \cdots & a_{ij}^{(k)} & \cdots & a_{in}^{(k)} \\ \vdots & & \vdots & \vdots & & \vdots & & \vdots \\ 0 & \cdots & 0 & a_{nk}^{(k)} & \cdots & a_{nj}^{(k)} & \cdots & a_{nn}^{(k)} \end{pmatrix}.$$

Para hacer ceros en la k-ésima columna operaremos como sigue

$$m_{ik} = \frac{a_{ik}^{(k)}}{a_{kk}^{(k)}}, \qquad k+1 \leq i \leq n.$$

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

$$b_{i}^{(k+1)} = \begin{cases} b_{i}^{(k)}, & 1 \leq 1 \leq k, \\ b_{i}^{(k)} - m_{ik}\, b_{1}^{(k)}, & k+1 \leq i \leq n, \end{cases}$$

Finalmente, en el paso $k = n$, la matriz resultante $U \equiv A^{(n)}$, $u_{ij} = a_{ij}^{(n)}$, será una matriz triangular superior. Una vez que tenemos un sistema lineal con una matriz triangular superior U es fácil obtener la solución del sistema $U\,x = b^{(n)}$ por sustitución hacia atrás o regresiva.

Podemos contar fácilmente el número de operaciones aritméticas realizadas en la eliminación de Gauss, que sumaremos a las $O\big(n^2\big)$ de la eliminación hacia atrás para resolver el sistema $U\,x = b$. En la eliminación de Gauss tenemos

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

divisiones entre los $n-1$ pivotes, y

$$C_p(n) = C_s(n) = n\,(n-1) + (n-1)\,(n-2) + \cdots + 6 + 2 = \sum_{k=1}^{n-1} k\,(k+1),$$

productos y sumas, respectivamente, donde también se han tenido en cuenta las operaciones realizadas sobre el vector b. Sumando estas expresiones[5]

$$C_d(n) = \frac{n\,(n-1)}{2},$$

$$C_s(n) = C_p(n) = \frac{n\,(n-1)}{2} + \frac{n\,(n-1)\,(2\,n-1)}{6} = \frac{(n+1)\,n\,(n-1)}{3},$$

es decir, el número total de operaciones es

$$C_T(n) = C_d(n) + C_s(n) + C_p(n) = \frac{n\,(n-1)\,(4\,n+7)}{6} = O\left(\frac{2\,n^3}{3}\right),$$

para $n \gg 1$. El coste de la eliminación hacia atrás, $O\big(n^2\big)$, es despreciable por lo que el número de operaciones de la eliminación de Gauss es del orden de

$$C(n) = O\left(\frac{2\,n^3}{3}\right).$$

El coste computacional cúbico, aunque mucho menor que el orden de complejidad de la regla de Cramer, es relativamente alto y limita las aplicaciones de la eliminación de Gauss para matrices densas de no más de 1000 incógnitas.

El procedimiento de eliminación de Gauss que hemos descrito tiene dos problemas básicamente. El primero es la aparición de pivotes nulos, $a_{kk}^{(k)} = 0$, o muy pequeños $O(u)$, que pueden incurrir en la aparición de *overflow*. El segundo es la aparición de multiplicadores m_{ik} muy grandes, que provocan una cancelación de dígitos significativos en $a_{ij}^{(k)} - m_{ik}\,a_{kj}^{(k)}$, ya que los dígitos menos significativos de $a_{ij}^{(k)}$ se pueden perder. La pérdida de estos dígitos puede suponer un cambio relativo muy grande en los coeficientes de la matriz A. El ejemplo más simple es la matriz

$$A = A^{(1)} = \begin{pmatrix} \epsilon & 1 \\ 1 & 1 \end{pmatrix}, \qquad A^{(2)} = \begin{pmatrix} \epsilon & 1 \\ 0 & 1 - 1/\epsilon \end{pmatrix},$$

[5]En cualquier libro de tablas y fórmulas matemáticas el lector puede encontrar que

$$\sum_{j=1}^{n} j^2 = \frac{n\,(n+1)\,(2\,n+1)}{6},$$

lo que puede verificar fácilmente por inducción.

en la que para $\epsilon < u$, $fl(a_{22}^{(2)}) = -1/\epsilon$, que implica resolver la matriz que tiene $a_{22} = 0$, en lugar de $a_{22} = 1$.

Estos problemas se pueden minimizar si se utiliza una estrategia de pivotaje. Si en el paso k-ésimo, en lugar de elegir como pivote al elemento $a_{kk}^{(k)}$, intercambiamos las filas k-ésima y r-ésima, tomando

$$|a_{rk}^{(k)}| = \max_{k \leq i \leq n} |a_{ik}^{(k)}|$$

entonces aseguramos que

$$|m_{ik}| \leq 1, \qquad i = k+1, k+2, \ldots, n.$$

A esta técnica se le denomina pivotaje parcial y requiere $O(n^2)$ operaciones. Se puede utilizar también una técnica más costosa, el pivotaje completo, en el que se intercambian tanto filas como columnas, buscando el mejor pivote posible. En el paso k-ésimo, intercambiamos las filas k y r, y las columnas k y s, donde

$$|a_{rs}^{(k)}| = \max_{k \leq i,j \leq n} |a_{ij}^{(k)}|.$$

Como esta técnica requiere $O(n^3)$ operaciones, se utiliza mucho menos en la práctica. Sin embargo, como veremos más adelante, tiene varias ventajas desde el punto de vista de la estabilidad numérica.

El algoritmo 4.4 muestra un algoritmo para la resolución de un sistema lineal utilizando eliminación de Gauss con pivotaje parcial.

4.3.3 Factorización LU de Doolittle y Crout

El procedimiento de eliminación de Gauss se puede escribir de forma matricial, lo que nos permite interpretarlo como una factorización de la matriz A en el producto de una matriz triangular inferior L por una triangular superior U.

La eliminación de los ceros en la primera columna de $A^{(1)} = A$, se puede obtener mediante el producto matricial

$$A^{(2)} = L_1 \, A^{(1)},$$

Algoritmo 4.4 *Resolución del sistema lineal $A\,x = b$ mediante eliminación de Gauss con pivotaje parcial.*

```
function x = GEPPsol (A,b)
% Resuelve A x = b mediante eliminacion de Gauss
% A es una matriz de orden n x n
% b y x son vectores de n x 1
%   Si A es singular, retorna x=NaN
%
  eps = 2^(-52);   % epsilon de la máquina
  n=length(b);
  x=zeros(n,1);    % declara x como vector columna
  Ab = [A b];      % matriz aumentada [A|b]
  for k=1:n-1,
    mayor=abs(Ab(k,k));  % >Cuál es el pivote en columna k-ésima?
    kpivote=k;
    for r=k+1:n,
      if (abs(Ab(r,k))>mayor), mayor=abs(Ab(r,k)); kpivote=r; end
    end
    swap=Ab(k,:);         % Intercambia filas k- y kpivote-ésima
    Ab(k,:)=Ab(kpivote,:);
    Ab(kpivote,:)=swap;
    if (abs(Ab(k,k)) <= n*eps),  % >Es el pivote nulo?
      x=NaN;                     % Devuelve: matriz singular
    else
      for i=k+1:n,
        m=Ab(i,k)/Ab(k,k); [i k],              % multiplicador (elemento de L)
        Ab(i,k:n+1)=Ab(i,k:n+1)-m*Ab(k,k:n+1); % calcula fila de la matriz U
      end
    end
  end
  x(n)=Ab(n,n+1)/Ab(n,n);       % Resolucion del sistema tringular U x = P b
  for k=n-1:-1:1,
    x(k)=(Ab(k,n+1)-Ab(k,k+1:n)*x(k+1:n))/Ab(k,k);
  end
```

donde la matriz triangular inferior L_1 toma la forma

$$L_1 = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ -m_{21} & 1 & 0 & \cdots & 0 \\ -m_{31} & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & & \ddots & \vdots \\ -m_{n1} & 0 & 0 & \cdots & 1 \end{pmatrix}, \qquad m_{i1} = \frac{a_{i1}^{(1)}}{a_{11}^{(1)}}.$$

Llamando $m_1 \in \mathbb{R}^n$ al vector cuyas componentes son $(0, m_{21}, m_{31}, \ldots, m_{n1})^\top$, y a e_1 al primer vector de la base canónica de $\mathbb{R}^n$, podemos escribir,

$$L_1 = I - m_1\, e_1^\top.$$

De igual forma, para el siguiente paso, tenemos

$$A^{(3)} = L_2\, A^{(2)},$$

donde

$$L_2 = I - m_2\, e_2^\top = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & -m_{32} & 1 & \cdots & 0 \\ \vdots & \vdots & & \ddots & \vdots \\ 0 & -m_{n2} & 0 & \cdots & 1 \end{pmatrix}, \qquad m_2 = \begin{pmatrix} 0 \\ 0 \\ m_{32} \\ \vdots \\ m_{n2} \end{pmatrix}, \qquad m_{i2} = \frac{a_{i2}^{(2)}}{a_{22}^{(2)}}.$$

y e_2 es el segundo vector de la base canónica.

De esta manera, la etapa k-ésima corresponde a hacer

$$A^{(k)} = L_k\, A^{(k-1)}, \qquad L_k = I - m_k\, e_k^\top, \qquad m_k = (0, \ldots, 0, m_{k+1,k}, \ldots, m_{nk})^\top,$$

y la matriz triangular superior U obtenida tras el último paso es

$$U = L_{n-1} \cdots L_2\, L_1\, A = (I - m_{n-1}\, e_{n-1}^\top) \cdots (I - m_2\, e_2^\top)\, (I - m_1\, e_1^\top)\, A.$$

La matriz A se puede escribir como el producto de matrices

$$A = L_1^{-1}\, L_2^{-1} \cdots L_{n-1}^{-1}\, U.$$

La inversa de las matrices triangulares L_k es fácil de obtener, basta invertir el signo de los multiplicadores, $L_k^{-1} = I + m_k\, e_k^\top$, ya que $e_k^\top\, m_k = 0$ y por tanto

$$(I - m_k\, e_k^\top)\,(I + m_k\, e_k^\top) = I - m_k\, e_k^\top\, m_k\, e_k^\top = I.$$

Más aún, como $e_i^\top\, m_k = 0$ para $k \geq i$, el producto de las inversas de las matrices triangulares es una matriz triangular

$$L = (I + m_1\, e_1^\top)\,(I + m_2\, e_2^\top)\,\cdots\,(I + m_{n-1}\, e_{n-1}^\top) = I + \sum_{i=1}^{n-1} m_i\, e_i^\top,$$

que en componentes se escribe fácilmente como

$$L = \begin{pmatrix} 1 & 0 & \cdots & 0 & 0 \\ m_{21} & 1 & \ddots & 0 & 0 \\ m_{31} & m_{32} & \ddots & \ddots & 0 \\ \vdots & \vdots & & \ddots & \vdots \\ m_{n1} & m_{n2} & \cdots & m_{n,n-1} & 1 \end{pmatrix}, \qquad m_{ik} = \frac{a_{ik}^{(k)}}{a_{kk}^{(k)}}.$$

El lector observará que nuestro desarrollo tiene el siguiente corolario. La inversa de una matriz triangular inferior (superior) es también una matriz triangular inferior (superior). Además, si la matriz triangular era (con diagonal) unitaria, su inversa también lo será.

En resumen, hemos demostrado por construcción que la eliminación de Gauss equivale a factorizar la matriz de coeficientes $A = L\,U$ como el producto de una matriz triangular inferior unitaria L y una triangular superior U. A esta factorización LU se la denomina de Doolittle. También se puede factorizar la matriz $A = \widetilde{L}\,\widetilde{U}$ en una matriz triangular inferior $\widetilde{L}$ y una matriz triangular superior unitaria $\widetilde{U}$, procedimiento que se denomina factorización LU de Crout. Dejamos al lector interesado los detalles de esta última factorización que equivale a una eliminación de Gauss, pero con ligeros cambios. Finalmente, se puede factorizar la matriz como $A = \widehat{L}\,\widehat{D}\,\widehat{U}$ donde $\widehat{D}$ es una matriz diagonal, y $\widehat{L}$ y $\widehat{U}$ son sendas matrices triangulares inferiores y superiores, respectivamente, unitarias.

La resolución del sistema de ecuaciones $A\,x = b$ por eliminación de Gauss matricialmente corresponde a resolver el sistema triangular

$$U\,x = L^{-1}\,b, \qquad U = L^{-1}\,A,$$

aunque cuando se conoce la factorización LU es más eficiente resolver los dos sistemas triangulares con matrices de coeficientes L y U, en este orden, es decir,

$$L\,U\,x = b \quad \Rightarrow \quad L\,y = b, \qquad U\,x = y.$$

Ejemplo 4.7 *Para el sistema $A\,x = b$, donde*

$$A = \begin{pmatrix} 4 & -9 & 2 \\ 2 & -4 & 4 \\ -1 & 2 & 2 \end{pmatrix}, \qquad b = \begin{pmatrix} 2 \\ 3 \\ 1 \end{pmatrix},$$

aplicaremos el procedimiento de eliminación de Gauss y su equivalencia con la factorización LU.

El primer paso en el procedimiento de eliminación de Gauss es multiplicar la primera fila por $-2/4$ y sumarla a la segunda, y la primera por $1/4$ y sumarla a la tercera. Definiendo

$$L_1 = \begin{pmatrix} 1 & 0 & 0 \\ -0.5 & 1 & 0 \\ +0.25 & 0 & 1 \end{pmatrix},$$

tenemos que

$$A_1 = L_1\,A = \begin{pmatrix} 4 & -9 & 2 \\ 0 & 0.5 & 3 \\ 0 & -0.25 & 2.5 \end{pmatrix}, \qquad b_1 = L_1\,b = \begin{pmatrix} 2 \\ 2 \\ 1.5 \end{pmatrix}.$$

En el segundo paso del método de eliminación de Gauss multiplicamos la segunda fila por $1/2$ y la sumamos a la tercera, por lo que tenemos

$$L_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & +0.5 & 1 \end{pmatrix},$$

tenemos que

$$A_2 = L_2\,A_1 = \begin{pmatrix} 4 & -9 & 2 \\ 0 & 0.5 & 3 \\ 0 & 0 & 4 \end{pmatrix}, \qquad b_2 = L_2\,b = \begin{pmatrix} 2 \\ 2 \\ 2.5 \end{pmatrix}.$$

Con lo que hemos obtenido una matriz triangular superior

$$U = L_2\, A_1 = L_2\, L_1\, A.$$

La matriz triangular inferior

$$L_0 = L_2\, L_1 = \begin{pmatrix} 1 & 0 & 0 \\ -0.5 & 1 & 0 \\ +0.25 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & +0.5 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ -0.5 & 1 & 0 \\ 0 & 0.5 & 1 \end{pmatrix},$$

tiene como inversa

$$L = L_0^{-1} = L_1^{-1}\, L_2^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ 0.5 & 1 & 0 \\ -0.25 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -0.5 & 1 \end{pmatrix}$$

$$= \begin{pmatrix} 1 & 0 & 0 \\ 0.5 & 1 & 0 \\ -0.25 & -0.5 & 1 \end{pmatrix}.$$

De esta forma hemos obtenido una descomposición LU de la matriz A,

$$A = L_0^{-1}\, U = L\, U.$$

4.3.4 Cálculo directo de la factorización LU

El procedimiento que hemos seguido, que utiliza una eliminación de Gauss previa a la factorización se puede evitar realizando ésta directamente.

Consideremos, primero, la factorización LU de Doolittle, en la que L tiene diagonal unitaria, $l_{ii} = 1$,

$$A = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ l_{21} & 1 & 0 & \cdots & 0 \\ l_{31} & l_{32} & 1 & \cdots & 0 \\ \vdots & \vdots & & \ddots & \vdots \\ l_{n1} & l_{n2} & l_{n3} & \cdots & 1 \end{pmatrix} \begin{pmatrix} u_{11} & u_{12} & u_{13} & \cdots & u_{1n} \\ 0 & u_{22} & u_{23} & \cdots & u_{2n} \\ 0 & 0 & u_{33} & \cdots & u_{3n} \\ \vdots & \vdots & & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & u_{nn} \end{pmatrix}.$$

Para determinar directamente la expresión de los coeficientes de L y de U en función de los de A basta multiplicar dichas matrices y comparar sus coeficientes con los de A. Multiplicando la primera fila de L por todas y cada una de las columnas de U se obtiene

$$u_{1j} = a_{1j}, \qquad 1 \leq j \leq n.$$

Multiplicando todas las filas de L por la primera columna de U, obtenemos

$$l_{i1}\, u_{11} = a_{i1} \quad \Rightarrow \quad l_{i1} = \frac{a_{i1}}{u_{11}} = \frac{a_{i1}}{a_{11}}, \qquad 2 \leq i \leq n.$$

Multiplicando la segunda fila de L por las columnas de U, y utilizando los coeficientes ya calculados, se obtiene

$$l_{j1}\, u_{1j} + u_{2j} = a_{2j} \quad \Rightarrow \quad u_{2j} = a_{2j} - l_{j1}\, u_{1j} = a_{2j} - \frac{a_{j1}}{a_{11}}\, a_{1j}, \qquad 2 \leq j \leq n.$$

Multiplicando las filas de L por la segunda columna de U, y utilizando los coeficientes ya calculados, se obtiene

$$l_{i1}\, u_{1i} + l_{i2}\, u_{2i} = a_{i2}, \quad \Rightarrow \quad l_{i1} = \frac{1}{u_{2i}}\,(a_{i2} - l_{i1}\, u_{1i}), \qquad 3 \leq j \leq n.$$

Los resultados que se van obteniendo son muy similares a los obtenidos por el procedimiento de eliminación de Gauss. Operando sucesivamente de la misma forma se llega a la siguiente expresión general para calcular los coeficientes de L y U,

$$u_{ij} = a_{ij} - \sum_{k=1}^{i-1} l_{ik}\, u_{kj}, \qquad 1 \leq i \leq j \leq n, \tag{4.2}$$

$$l_{ij} = \frac{1}{u_{ii}} \left(a_{ij} - \sum_{k=1}^{j-1} l_{ik}\, u_{kj} \right), \qquad 1 \leq j < i \leq n-1, \tag{4.3}$$

que se aplicará alternativamente, primero para la fila i de U y luego para la columna j de L.

Consideremos ahora, la factorización LU de Crout, en la que U tiene diagonal unitaria, $u_{ii} = 1$,

$$A = \begin{pmatrix} l_{11} & 0 & 0 & \cdots & 0 \\ l_{21} & l_{22} & 0 & \cdots & 0 \\ l_{31} & l_{32} & l_{33} & \cdots & 0 \\ \vdots & \vdots & & \ddots & \vdots \\ l_{n1} & l_{n2} & l_{n3} & \cdots & l_{nn} \end{pmatrix} \begin{pmatrix} 1 & u_{12} & u_{13} & \cdots & u_{1n} \\ 0 & 1 & u_{23} & \cdots & u_{2n} \\ 0 & 0 & 1 & \cdots & u_{3n} \\ \vdots & \vdots & & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix}.$$

Como antes, determinar la expresión de los coeficientes de L y de U es relativamente fácil. Para la primera columna de L,

$$l_{i1} = a_{i1}, \qquad 1 \leq i \leq n,$$

y para la primera fila de U,

$$l_{11}\, u_{1j} = a_{1j}, \quad \Rightarrow \quad u_{1j} = \frac{a_{1j}}{l_{11}} = \frac{a_{1j}}{a_{11}}, \qquad 2 \leq j \leq n.$$

Para las demás columnas de L y filas de U el procedimiento general que se obtiene es

$$l_{ij} = a_{ij} - \sum_{k=1}^{j-1} l_{ik}\, u_{kj}, \qquad 1 \leq j \leq i \leq n,$$

$$u_{ij} = \frac{1}{l_{ii}} \left(a_{ij} - \sum_{k=1}^{i-1} l_{ik}\, u_{kj} \right), \qquad 2 \leq j < i \leq n.$$

Como hemos visto, la factorización LU equivale al procedimiento de eliminación de Gauss y para la resolución de sistemas lineales requiere, de hecho, el mismo coste computacional $O\big(2\,n^3/3\big)$, aunque no detallaremos esta demostración. Sin embargo, la ventaja fundamental de esta factorización aparece cuando hay que resolver muchos sistemas lineales con la misma matriz de coeficientes A y diferentes términos no homogéneos, b, en cuyo caso basta almacenar las matrices L y U una sola vez, siendo el cálculo de las sucesivas soluciones realizado con el coste de la resolución de dos sistemas triangulares, que es del orden de $O\big(n^2\big)$, muy inferior al de la factorización.

Hemos de indicar que algunos autores utilizan una factorización $A = L\,D\,U$, en la que las matrices L y U tienen diagonal unitaria y D es un matriz diagonal. No presentaremos más detalles de este tipo de factorización, que por otra parte no reviste mayor dificultad.

También hemos de indicar que normalmente las matrices L y U no se almacenan con sus elementos nulos, sino que se almacenan las dos matrices en una única matriz LU con objeto de reducer el coste en memoria. Además, se puede observar que en la iteración k-ésima del algoritmo no se requiere el conocimiento de los valores de a_{ij} con $1 < i, j < k$, por lo que los valores ya calculados de las matrices L y U se pueden sobreescribir sobre los valores originales de la matriz A. Eso sí, siempre y cuando no importe que los valores originales de la matriz A se pierdan, lo que se puede aprovechar en la mayoría de las ocasiones. De esta forma se reduce el coste en memoria a la mitad.

El algoritmo 4.5 muestra un algoritmo para la resolución de un sistema lineal utilizando factorización LU con pivotaje parcial.

Algoritmo 4.5 *Resolución del sistema lineal $A\,x = b$ mediante factorización LU con pivotaje parcial.*

```
function x = LUPPsol (A,b)
% Resuelve A x = b mediante factorizacion L U =P A
% A es una matriz de orden n x n (contendra L y U)
% b y x son vectores de n x 1
%   Si A es singular, retorna x=NaN
%
  eps = 2^(-52);   % epsilon maquina
  n=length(b);
  kr=1:n;          % vector que representa P
  for k=1:n-1,
    mayor=abs(A(k,k));    % >Cuál es el pivote en col. k-ésima?
    kpivote=k;
    for r=k+1:n,
      if (abs(A(r,k))>mayor), mayor=abs(A(r,k)); kpivote=r, end
    end
    swapk=kr(k);        swap=A(k,:);         % Intercambia filas
    kr(k)=kr(kpivote);  A(k,:)=A(kpivote,:);
    kr(kpivote)=swapk;  A(kpivote,:)=swap;
    if (abs(A(kr(k),kr(k))) <= n*eps),       % >Es el pivote nulo?
      x=NaN;                                 % Matriz singular
    else
      for i=k+1:n,
        A(i,k)=A(i,k)/A(k,k);                    % Elementos de L
        A(i,k+1:n)=A(i,k+1:n)-A(i,k)*A(k,k+1:n);  % Elementos de U
      end
    end
  end
  y=zeros(n,1); y(1)=b(kr(1));         % Resuelve L y = P*b
  for k=2:n,
    y(k)=b(kr(k))-A(k,1:k-1)*y(1:k-1);
  end
  x=zeros(n,1); x(n)=y(n)/A(n,n);      % Resuelve U x = y
  for k=n-1:-1:1,
    x(k)=(y(k)-A(k,k+1:n)*x(k+1:n))/A(k,k);
  end
```

El lector se puede preguntar bajo qué condiciones matemáticas una matriz A cuadrada tiene una factorización LU única. El siguiente teorema contesta dicha pregunta.

Teorema 4.8 *La matriz $A \in \mathbb{R}^{n\times n}$ tiene factorización LU de Doolittle única si y sólo si sus menores principales A_k, para $k = 1, 2, \ldots, n-1$, son no singulares. Si alguno de estos menores es singular, la factorización puede existir, pero entonces no es única. El mismo resultado se obtiene para la factorización de Crout y la LDU.*

Concretarermos la demostración para el caso de la factorización de Doolittle. Los menores principales A_k son las submatrices $A(1:k, 1:k)$. La demostración por inducción es sencilla. El caso base $k = 1$ es trivial. Supongamos que A_{k-1} tiene factorización LU única, sea $A_{k-1} = L_{k-1}\, U_{k-1}$, entonces podemos escribir el menor principal A_k de la forma

$$A_k = \begin{pmatrix} A_{k-1} & b \\ c^\top & a_{kk} \end{pmatrix} = \begin{pmatrix} L_{k-1} & 0 \\ l^\top & 1 \end{pmatrix} \begin{pmatrix} U_{k-1} & u \\ 0 & u_{kk} \end{pmatrix} = L_k\, U_k,$$

que será cierta si se cumplen las siguientes ecuaciones

$$L_{k-1}\, u = b, \qquad c^\top = l^\top\, U_{k-1}, \qquad a_{kk} = l^\top\, u + u_{kk}.$$

Las matrices L_{k-1} y U_{k-1} son no singulares ya que $0 \neq \det(A_{k-1}) = \det(L_{k-1})\, \det(U_{k-1})$, luego las ecuaciones anteriores tienen solución única. Esto completa la demostración del "si".

Hay que probar la parte del "sólo si". Lo más fácil, es el caso en que A es no singular. Supongamos que la factorización LU existe. Entonces $A_k = L_k\, U_k$, $k = 1, 2, \ldots, n$, lo que nos indica que

$$\det(A_k) = \det(U_k) = u_{11}\, u_{22} \cdots u_{kk},$$

pero como $\det(A) = u_{11} \cdots u_{nn} \neq 0$, automáticamente $\det(A_k) \neq 0$, para $k = 1, 2, \ldots, n-1$. El caso de que A sea singular se lo dejamos al lector.

La no unicidad de la factorización cuando la condición anterior no se cumple es fácil de comprobar, por ejemplo,

$$\begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ \alpha & 1 \end{pmatrix} \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}, \qquad \forall \alpha,$$

y la existencia de matrices sin factorización, también, por ejemplo,

$$\begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ \alpha & 1 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 0 & * \end{pmatrix}, \qquad \alpha \cdot 0 \neq 1.$$

4.3.5 Técnicas de pivotaje y factorización LU

Como ya hemos indicado previamente, el gran problema de la eliminación de Gauss es la aparición de pivotes nulos o pequeños, efecto que se puede minimizar si se utiliza una técnica de pivotaje.

En el pivotaje parcial, en el k-ésimo paso de eliminación, en lugar de elegir directamente el elemento $a_{kk}^{(k)}$, buscamos el elemento de dicha columna de mayor módulo, sea el r-ésimo,

$$|a_{rk}^{(k)}| = \max_{k \leq i \leq n} |a_{ik}^{(k)}|$$

e intercambiamos las filas k-ésima y r-ésima, de tal manera que dicho elemento pasa a ocupar la posición $a_{kk}^{(k)}$, y será usado como pivote. El pivotaje parcial, que requiere $O\big(n^2\big)$ operaciones de comparación, nos garantiza que los multiplicadores son menores que la unidad,

$$|m_{ik}| \leq 1, \qquad i = k+1, k+2, \ldots, n.$$

La técnica de pivotaje parcial también se puede aplicar al algoritmo directo de factorización LU de Doolittle (Crout), de forma que evitemos la aparición de elementos u_{kk} (l_{kk}) nulos o pequeños. Para ello, intercambiaremos, en el caso de Doolittle, las filas k-ésima y r-ésima, donde

$$|u_{rk}| = \max_{k \leq i \leq n} |u_{ik}|,$$

(de forma similar en el caso de Crout).

El intercambio de las filas k-ésima y j-ésima de una matriz se realiza mediante la pre-multiplicación por una matriz de permutación $P = P_{kj}$, que es la matriz identidad con las filas

k-ésima y j-ésima intercambiadas, es decir,

$$P_{kj} = \begin{pmatrix}
1 & 0 & & (k) & & \cdots & & (j) & & 0 \\
0 & \ddots & & & & & & & & \\
(k) & & & 0 & 0 & \cdots & & 1 & & \\
 & & & 0 & 1 & & & & & \\
\vdots & & & \vdots & & \ddots & & \vdots & & \vdots \\
 & & & & & & 1 & 0 & & \\
(j) & & & 1 & & \cdots & 0 & 0 & & \\
 & & & & & & & & \ddots & 0 \\
0 & & & & & \cdots & & & 0 & 1
\end{pmatrix}.$$

Las matrices de permutación son ortogonales, $P^{-1} = P^\top$, y cumplen $P^{-1} = P$. Además, tienen determinante unitario, $\det(P_{kj}) = 1$.

En el algoritmo de eliminación de Gauss con pivotaje parcial, antes de premultiplicar por las matrices L_k de multiplicadores, es necesario realizar una permutación de filas P_k, en la que se permutan la fila k-ésima y la r_k-ésima, por lo que obtenemos

$$U = L_{n-1}\, L_{n-2}\, P_{n-2} \cdots L_2\, P_2\, L_1\, P_1\, A,$$

donde hemos omitido $P_{n-1} = I$. Desde el punto de vista del análisis numérico de este método es necesario notar que se pueden realizar todas las permutaciones antes de aplicar el algoritmo. Veamos en particular el caso $n = 4$, para el que

$$U = L_4\, L_3\, P_3\, L_2\, P_2\, L_1\, P_1\, A,$$

se puede escribir de la forma

$$U = L_4\, L_3\, \underbrace{P_3\, L_2\, P_3}_{\widetilde{L}_2}\, \underbrace{P_3\, P_2\, L_1\, P_2\, P_3}_{\widetilde{L}_1}\, \underbrace{P_3\, P_2\, P_1\, A}_{P\,A},$$

es decir,

$$U = L_4\, L_3\, \widetilde{L}_2\, \widetilde{L}_1\, P\, A,$$

donde $\widetilde{L}_i$ son matrices triangulares con ceros en los mismos elementos que L_i, y $P = P_3\, P_2\, P_1$ es el producto de varias matrices de permutación, que es también una matriz de permutación

$P^{-1} = P$. En este sentido la factorización LU con pivotaje parcial se puede interpretar como la factorización LU normal aplicada a la matriz permutada $P\,A$, es decir, $P\,A = L\,U$. Por supuesto, de antemano no conocemos la matriz de permutación P adecuada, por lo que esta interpretación sólo tiene utilidad desde el punto de vista teórico.

El lector notará que para almacenar la matriz P_k basta dar el número r_k de la fila a permutar con la k-éxima, ya que ésta última está implícita. Por ello, la matriz de permutación P producto de las P_k algunas veces se representa mediante un vector cuyas componentes son los valores r_k de las filas permutadas por la k-ésima en la matriz P_k.

Ejemplo 4.9 *Apliquemos eliminación de Gauss con pivotaje parcial a la resolución del sistema*

$$\begin{aligned} x_2 + x_3 &= 1, \\ x_1 + 2\,x_2 + 3\,x_3 &= 0, \\ x_1 + x_2 + x_3 &= 2. \end{aligned}$$

Obviamente, el primer pivote es nulo, aunque el determinante de la matriz de coeficientes sea 1 y el sistema tenga solución. Intercambiando las filas 1 y 2, mediante la siguiente matriz de intercambio

$$P_{12} = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

obtenemos el sistema

$$P_{12}\,A = \begin{pmatrix} 1 & 2 & 3 \\ 0 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = P_{12}\,b = \begin{pmatrix} 0 \\ 1 \\ 2 \end{pmatrix},$$

por lo que aplicando eliminación de Gauss

$$\left(\begin{array}{ccc|c} 1 & 2 & 3 & 0 \\ 0 & 1 & 1 & 1 \\ 1 & 1 & 1 & 2 \end{array}\right) \rightarrow \left(\begin{array}{ccc|c} 1 & 2 & 3 & 0 \\ 0 & 1 & 1 & 1 \\ 0 & -1 & -2 & 2 \end{array}\right) \rightarrow \left(\begin{array}{ccc|c} 1 & 2 & 3 & 0 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & -1 & 3 \end{array}\right),$$

y, finalmente, las soluciones al sistema son

$$x_3 = -3, \qquad x_2 = 1 - x_3 = 4, \qquad x_1 = -3\,x_3 - 2\,x_2 = 1.$$

Si hubiéramos aplicado factorización LU de Doolittle con pivotaje parcial, obtendríamos $P\,A = L\,U$, donde $P = P_{12}$,

$$L = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & -1 & 1 \end{pmatrix}, \qquad U = \begin{pmatrix} 1 & 2 & 3 \\ 0 & 1 & 1 \\ 0 & 0 & -1 \end{pmatrix}.$$

Obviamente, la solución del sistema es la misma de antes.

La técnica de pivotaje completo consiste en buscar el pivote de la etapa k-ésima de eliminación entre los valores más grandes de la submatriz que resta por considerar de la matriz $A^{(k)}$. Para ello se intercambian tanto filas como columnas. En el paso k-ésimo, intercambiamos las filas k y r_k, y las columnas k y s_k, donde

$$|a^{(k)}_{r_k s_k}| = \max_{k \le i,j \le n} |a^{(k)}_{ij}|.$$

Esta técnica tiene un alto coste computacional, y requiere $O\big(n^3\big)$ operaciones de comparación, por ello se utiliza mucho menos en la práctica. Además, aunque, como veremos más adelante, tiene varias ventajas desde el punto de vista de la estabilidad numérica, en la práctica las dos técnicas se comportan de forma parecida.

Para intercambiar dos columnas, sean la k-ésima y la j-ésima, se multiplica por la derecha (o post-multiplica) por la matriz de permutación de filas P_{kj}, como es fácil de comprobar. De esta manera, el intercambio simultáneo de filas y columnas se realiza de la forma $P_{ki}\,A\,P_{kj}$ y el sistema resultante queda

$$P_{ki}\,A\,P_{kj}\,P_{kj}^{-1}\,x = P_{ki}\,b.$$

Es importante notar que la aplicación de pivotaje completo a la factorización directa de Doolittle y Crout no es ni mucho menos obvia, y requiere cambiar el sustancialmente dichos algoritmos.

Como el pivotaje parcial, el completo se puede interpretar como una permutación inicial de la matriz A de la forma $P\,A\,Q$, donde P y Q son matrices de permutación para las filas y las columnas, respectivamente.

Ejemplo 4.10 *Aplique eliminación de Gauss con pivotaje completo al sistema*

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

Se deben intercambiar las filas 1 y 2, y las columnas 1 y 3, para obtener como pivote el mayor valor posible ($a_{23} = 3$). El intercambio de filas lo haremos multiplicando por la izquierda por P_{12} y el de columnas multiplicando por la derecha por P_{13}, donde

$$P_{12} = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \qquad P_{13} = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix},$$

y el sistema se escribe

$$P_{12}\, A\, P_{13}\, y = P_1\, b, \qquad y = P_{13}^{-1}\, x,$$

es decir,

$$P_{12}\, A\, P_{13} = A_1 = \begin{pmatrix} 3 & 2 & 1 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{pmatrix}, \quad P_{12}\, b = b_1 = \begin{pmatrix} 0 \\ 1 \\ 2 \end{pmatrix}, \quad P_{13}^{-1}\, x = \begin{pmatrix} x_3 \\ x_2 \\ x_1 \end{pmatrix}.$$

Aplicando eliminación de Gauss obtenemos

$$\left(\begin{array}{c|c} A_1 & b_1 \end{array} \right) = \left(\begin{array}{ccc|c} 3 & 2 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 2 \end{array} \right) \rightarrow \left(\begin{array}{ccc|c} 3 & 2 & 1 & 0 \\ 0 & 1 & -1 & 3 \\ 0 & 1 & 2 & 6 \end{array} \right) = \left(\begin{array}{c|c} A_2 & b_2 \end{array} \right).$$

Ahora la técnica de pivotaje completo indica que se deben intercambiar las filas 2 y 3, y las columnas 2 y 3, para poner el elemento más grande ($a_{22} = 3$) como segundo pivote. El intercambio de filas lo haremos multiplicando por la izquierda por P_{23} y el de columnas multiplicando por la derecha por P_{23}, donde

$$P_{23} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix},$$

y el sistema se escribe

$$A_3\, z = P_{23}\, A_2\, P_{23}\, z = P_{23}\, b = b_3, \qquad z = P_{23}^{-1}\, y = P_{23}^{-1}\, P_{13}^{-1}\, x,$$

es decir,

$$A_3 = \begin{pmatrix} 3 & 2 & 1 \\ 0 & 2 & 1 \\ 0 & -1 & 1 \end{pmatrix}, \quad b_3 = \begin{pmatrix} 0 \\ 6 \\ 3 \end{pmatrix}, \qquad z = \begin{pmatrix} x_3 \\ x_1 \\ x_2 \end{pmatrix}.$$

Aplicando eliminación de Gauss obtenemos

$$\left(\begin{array}{c|c} A_3 & b_3 \end{array} \right) = \left(\begin{array}{ccc|c} 3 & 2 & 1 & 0 \\ 0 & 2 & 1 & 6 \\ 0 & -1 & 1 & 3 \end{array} \right) \rightarrow \left(\begin{array}{ccc|c} 3 & 2 & 1 & 0 \\ 0 & 2 & 1 & 6 \\ 0 & 0 & 3 & 12 \end{array} \right) = \left(\begin{array}{c|c} A_4 & b_4 \end{array} \right),$$

con lo que obtenemos el sistema triangular superior

$$\begin{pmatrix} 3 & 2 & 1 \\ 0 & 2 & 1 \\ 0 & 0 & 3 \end{pmatrix} \begin{pmatrix} z_1 \\ z_2 \\ z_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 6 \\ 12 \end{pmatrix},$$

cuya solución es

$$z_3 = 12/3 = 4, \qquad z_2 = (6 - z_3)/2 = 1, \qquad z_1 = (-2\, z_3 - x_2)/3 = -3,$$

y la solución original $x = P_{13}\, P_{23}\, z$ es

$$x_1 = z_2 = 1, \qquad x_2 = z_3 = 4, \qquad x_3 = z_1 = -3.$$

4.3.6 Análisis de errores regresivos

El análisis de errores de la eliminación de Gauss es una combinación de los análisis ya presentados para el producto interior y para la resolución de sistemas triangulares[6]. La observación más importante a tener en cuenta es que todas las variantes de este algoritmo son matemáticamente

[6]La eliminación de Gauss fue el primer algoritmo numérico para el que se hizo un análisis de propagación de errores. El norteamericano Harold Hotelling (1895–1973) en 1943 hizo un análisis muy pesimista, muchas de cuyas conclusiones eran erróneas. El húngaro Johann Louis von Neumann (1903–1957), quien cambió su nombre

equivalentes, realizan las mismas operaciones aunque en un orden diferente, y por tanto satisfacen una cota de error común. Por ello, nos limitaremos a analizar los errores en el algoritmo de factorización LU de Doolittle. Más aún, basta analizar el método sin pivotaje, ya que éste, tanto parcial como completo, equivale a aplicar el método a una matriz convenientemente permutada.

Las asignaciones (4.2) y (4.3) toman la forma del algoritmo de sustitución progresiva o hacia adelante cuyos errores se analizaron en el teorema 4.4. Aplicando este teorema directamente, sin importar el orden con el que se evalúen los productos interiores, la matrices $\widehat{L}$ (con $\widehat{l}_{kk} = 1$) y $\widehat{U}$ calculadas satisfacen

$$\left| a_{kj} - \sum_{i=1}^{k-1} \widehat{l}_{ki}\, \widehat{u}_{ij} - \widehat{u}_{kj} \right| \leq \gamma_k \sum_{i=1}^{k} |\widehat{l}_{ki}|\, |\widehat{u}_{ij}|, \qquad j \geq k,$$

$$\left| a_{ik} - \sum_{j=1}^{k} \widehat{l}_{ij}\, \widehat{u}_{jk} \right| \leq \gamma_k \sum_{j=1}^{k} |\widehat{l}_{ij}|\, |\widehat{u}_{jk}|, \qquad i > k,$$

donde $\gamma_k = k\,\tilde{u}$, como siempre. Estas desigualdades nos dan directamente el análisis de errores hacia atrás para la factorización LU.

Teorema 4.11 *La factorización LU de la matriz $A \in \mathbb{R}^{n\times n}$ tanto por eliminación de Gauss como por el algoritmo directo de Doolittle o Crout, si acaba con éxito, conduce a dos factores $\widehat{L}$ y $\widehat{U}$ que satisfacen la cota de error regresivo*

$$\widehat{L}\,\widehat{U} = A + \Delta A, \qquad |\Delta A| \leq \gamma_n\, |\widehat{L}|\, |\widehat{U}|.$$

Para obtener una análisis de errores hacia atrás para la resolución de un sistema lineal por eliminación de Gauss o por factorización LU basta sumar al resultado de este teorema los errores implicados en la inversión de los dos sistemas triangulares. Utilizando los teoremas 4.4 y 4.11 obtenemos

$$\widehat{L}\,\widehat{U} = A + \Delta A_1, \qquad |\Delta A_1| \leq \gamma_n\, |\widehat{L}|\, |\widehat{U}|,$$

$$(\widehat{L} + \Delta L)\,\widehat{y} = b, \qquad |\Delta L| \leq \gamma_n\, |\widehat{L}|,$$

$$(\widehat{U} + \Delta U)\,\widehat{x} = \widehat{y}, \qquad |\Delta U| \leq \gamma_n\, |\widehat{U}|,$$

a John, y el norteamericano Hermann Heine Goldstine (1913–) en 1947 estudiaron el caso de matrices simétricas definidas positivas. El inglés Alan Mathison Turing (1912–1954) hizo el primer análisis de errores general en 1948, donde introdujo el concepto de número de condición. El inglés James Hardy Wilkinson (1919–1986) estudió empíricamente en 1954 el comportamiento del factor de crecimiento, lo que le llevó siguiendo ideas de Turing a introducir el análisis de errores hacia atrás y atacar con él la eliminación de Gauss en 1961. Los resultados presentados en este capítulo ya se conocían en 1963.

con lo que

$$(A+\Delta A)\,\widehat{x} = b = (\widehat{L}+\Delta L)\,(\widehat{U}+\Delta U)\,\widehat{x},$$
$$\Delta A = \Delta A_1 + \widehat{L}\,\Delta U + \Delta L\,\widehat{U} + \Delta L\,\Delta U,$$

que se acota fácilmente como

$$|\Delta A| \leq 3\,\gamma_n\,|\widehat{L}|\,|\widehat{U}| + \gamma_n^2\,|\widehat{L}|\,|\widehat{U}|.$$

Hemos obtenido como cota de error $c_n\,|\widehat{L}|\,|\widehat{U}|$, donde la constante c_n es de orden lineal, $O(n)$, que es el mejor resultado que podíamos esperar dado que cada elemento de A sufre del orden de $O(n)$ operaciones. El valor exacto de la constante de error no es importante. De hecho se puede demostrar fácilmente que se puede reducir a $c_n = 2\,\gamma_n$. En lugar de aplicar el teorema 4.4, podemos aplicar el lemas 4.1, que es más preciso, obteniendo

$$|\Delta L| \leq \mathrm{diag}(\gamma_{i-1})\,|\widehat{L}|, \qquad |\Delta U| \leq \mathrm{diag}(\gamma_{n-i+1})\,|\widehat{U}|,$$

donde $\mathrm{diag}(x_j)$ representa la matriz diagonal cuya diagonal toma los valores $d_{jj} = x_j$. De esta forma,

$$|\widehat{L}\,\Delta U + \Delta L\,\widehat{U} + \Delta L\,\Delta U| \leq \mathrm{diag}(\gamma_{i-1} + \gamma_{n-i+1} + \gamma_{i-1}\,\gamma_{n-i+1})\,|\widehat{L}|\,|\widehat{U}| \leq \gamma_n\,|\widehat{L}|\,|\widehat{U}|,$$

donde hemos usado que $\gamma_k + \gamma_j + \gamma_k\,\gamma_j \leq \gamma_{k+j}$, propiedad que presentamos en el tema 2. En resumen, hemos demostrado el siguiente teorema.

Teorema 4.12 *La resolución por eliminación de Gauss, o por factorización LU de Doolittle o Crout, si acaba con éxito, conduce a una solución cuyo error cumple la siguiente cota de error regresivo*[7]

$$(A+\Delta A)\,\widehat{x} = b, \qquad |\Delta A| \leq 2\,\gamma_n\,|\widehat{L}|\,|\widehat{U}|.$$

Este teorema se interpreta fácilmente. Lo ideal sería obtener una cota $|\Delta A| \leq u\,|A|$, sin embargo, como cada elemento de A sufre del orden de $O(n)$ operaciones no podemos esperar una cota mejor que $|\Delta A| \leq c_n\,u\,|A|$, donde $c_n = O(n)$. Una cota de este tipo se obtiene si $|\widehat{L}|\,|\widehat{U}| = |\widehat{L}\,\widehat{U}|$, es decir, cuando los elementos de $\widehat{L}$ y $\widehat{U}$ son no negativos, ya que entonces

$$|\widehat{L}|\,|\widehat{U}| = |\widehat{L}\,\widehat{U}| = |A+\Delta A| \leq |A| + \gamma_n\,|\widehat{L}|\,|\widehat{U}|,$$

con lo que obtenemos

$$(A+\Delta A)\,\widehat{x} = b, \qquad |\Delta A| \leq \frac{2\,\gamma_n}{1-\gamma_n}\,|A|, \qquad (\widehat{L},\widehat{U} \geq 0).$$

[7]Si el lector lo prefiere, sustituya $2\,\gamma_n$ por $3\,\gamma_n + \gamma_n^2$. Lo único importante para nuestro análisis es que la constante $c_n = O(\gamma_n)$.

Como observamos de los teoremas anteriores, la estabilidad de la eliminación de Gauss está determinada por el tamaño de la matriz $|\widehat{L}|\,|\widehat{U}|$ y no por el tamaño de los multiplicadores. Esta matriz puede ser pequeña cuando los multiplicadores $(\widehat{l}_{ij})$ son grandes, y grande cuando éstos son del orden de la unidad (como cuando usamos pivotaje).

La estabilidad numérica de la eliminación de Gauss requiere que $\||\widehat{L}|\,|\widehat{U}|\|/\|A\|$ sea pequeña. Sin embargo, cuando no se utiliza pivotaje este cociente puede ser arbitrariamente grande; por ejemplo, para la matriz

$$\begin{pmatrix} \epsilon & 1 \\ 1 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ \frac{1}{\epsilon} & 1 \end{pmatrix} \begin{pmatrix} \epsilon & 0 \\ 1 & 1-\frac{1}{\epsilon} \end{pmatrix},$$

el cociente $\||\widehat{L}|\,|\widehat{U}|\|/\|A\| = O(1/\epsilon)$.

Sin embargo, cuando se utiliza pivotaje parcial, $|l_{ij}| \leq 1$, $\forall i \geq j$, y además

$$|u_{ij}| \leq 2^{i-1} \max_{k\leq i} |a_{kj}|,$$

que se prueba fácilmente por inducción. El caso base, $|u_{1j}| = |a_{1j}|$. Aplicando la hipótesis de inducción, obtenemos

$$|u_{ij}| \leq |a_{ij}| + \sum_{k=1}^{i-1} |l_{ik}|\,|u_{kj}| \leq |a_{ij}| + \sum_{k=1}^{i-1} 2^{k-1} \max_{l\leq k} |a_{lj}|,$$

que conduce al resultado buscado

$$|u_{ij}| \leq \max_{k\leq i} |a_{ij}| \left(1 + \sum_{k=1}^{i-1} 2^{k-1}\right) = 2^{i-1} \max_{k\leq i} |a_{kj}|.$$

En resumen, al utilizar pivotaje parcial $|L|$ es pequeña y U está acotada relativamente con A, con lo que $\||\widehat{L}|\,|\widehat{U}|\|/\|A\|$ está acotado. Ello garantiza la estabilidad numérica en la práctica de la eliminación de Gauss con pivotaje parcial. Cuando se utiliza pivotaje completo, la estabilidad numérica se puede demostrar de forma rigurosa, aunque omitiremos dicha demostración.

4.3.7 Errores y el factor de crecimiento

En el análisis de la propagación de errores es clásico partir del siguiente teorema debido a Wilkinson, que introduce el importante concepto de factor de crecimiento.

Teorema 4.13 *(Wilkinson) Sea $A \in \mathbb{R}^{n\times n}$, entonces la solución del problema $A\,x = b$ mediante eliminación de Gauss con pivotaje parcial conduce a la siguiente cota de error*

$$(A + \Delta A)\,\widehat{x} = b, \qquad \|\Delta A\|_\infty \leq 2\,n^2\,\gamma_n\,\rho_n\,\|A\|_\infty,$$

donde el factor de crecimiento

$$\rho_n = \frac{\max\limits_{i,j,k} |a_{ij}^{(k)}|}{\max\limits_{i,j} |a_{ij}|},$$

cuyo cálculo requiere los valores $a_{ij}^{(k)}$ *de las matrices intermedias que aparecen en la eliminación de Gauss.*

No demostraremos este teorema, pero observaremos que es razonable para la eliminación de Gauss con pivotaje parcial utilizando el teorema 4.12, ya que con pivotaje parcial $|l_{ij}| \leq 1$ y

$$|u_{ij}| = |a_{ij}^{(i)}| \leq \rho_n \max_{i,j} |a_{ij}|.$$

La demostración este teorema para la eliminación de Gauss sin pivotaje es mucho más difícil y engorrosa, por lo que la omitiremos [**?**].

Cuando se usa pivotaje parcial es fácil comprobar que $\rho_n \leq 2^{n-1}$. Esta cota superior se alcanza para las matrices con la forma, en el caso de $n = 4$,

$$\begin{pmatrix} 1 & 0 & 0 & 1 \\ -1 & 1 & 0 & 1 \\ -1 & -1 & 1 & 1 \\ -1 & -1 & -1 & 1 \end{pmatrix}.$$

Aunque muy excepcionalmente el factor de crecimiento puede ser muy grande, incluso crecer exponencialmente, en la práctica siempre es pequeño. Estudios realizados con matrices aleatorias indican que el factor de crecimiento es $O\big(n^{2/3}\big)$, con pivotaje parcial, y $O\big(n^{1/2}\big)$, con pivotaje completo. Uno de los problemas aún sin resolver en análisis numérico es saber el porqué de este comportamiento en la práctica.

Para algunas clases de matrices el factor de crecimiento ha sido estudiado con toda generalidad. Entre ellas destacan las matrices diagonalmente dominantes por filas (o por columnas)

$$|a_{ii}| \geq \sum_{j \neq i} |a_{ij}|, \qquad \forall i.$$

Estas matrices surgen en muchos problemas prácticos y por ello presentaremos, sin demostración, el siguiente teorema, que nos indica que la eliminación de Gauss es perfectamente estable para estas matrices.

Teorema 4.14 *(Wilkinson) Si $A \in \mathbb{C}^{n\times n}$ es diagonalmente dominante por filas, entonces tiene factorización LU sin necesidad de pivotaje, y además el factor de crecimiento está acotado $\rho_n \leq 2$ [4].*

4.4 Sistemas de Ecuaciones Tridiagonales

Un sistema lineal de ecuaciones tridiagonal tiene la forma

$$\begin{aligned}
d_1\, x_1 + c_1\, x_2 &= b_1,\\
a_2\, x_1 + d_2\, x_2 + c_2\, x_3 &= b_2,\\
a_3\, x_2 + d_3\, x_3 + c_3\, x_4 &= b_3,\\
\ddots \qquad \ddots &\qquad \ddots\\
a_n\, x_{n-1} + d_n\, x_n &= b_n,
\end{aligned}$$

que podemos escribir como $A\, x = b$, donde A es la matriz tridiagonal que tiene como valores

$$\begin{aligned}
A_{k,k-1} &= a_k, && k = 2, \ldots, n,\\
A_{k,k} &= d_k, && k = 1, 2, \ldots, n,\\
A_{k,k+1} &= c_k, && k = 1, 2, \ldots, n-1,\\
A_{k,j} &= 0, && j \notin \{k-1, k, k+1\}.
\end{aligned}$$

La eliminación de Gauss nos permite obtener un sistema bidiagonal superior si hacemos ceros en la subdiagonal principal mediante combinaciones lineales de las ecuaciones del sistema. Por ejemplo, para eliminar el término multiplicado por a_2 le restaremos a la segunda ecuación la primera multiplicada por a_2/d_1. Procediendo de esta manera con las sucesivas ecuaciones obtendremos tras $(n-1)$ pasos, el sistema

$$\begin{aligned}
\alpha_1\, x_1 + \gamma_1\, x_2 &= \beta_1,\\
\alpha_2\, x_2 + \gamma_2\, x_3 &= \beta_2,\\
\ddots \qquad &\qquad \ddots\\
\alpha_n\, x_n &= \beta_n.
\end{aligned}$$

El cálculo de estos coeficientes se denomina iteración hacia adelante o progresiva. Estudiaremos en detalle este cálculo. La primera fila ha de quedar inalterada

$$\alpha_1 = d_1, \qquad \gamma_1 = c_1, \qquad \beta_1 = b_1.$$

Supongamos que hemos realizado $(k-1)$ pasos de eliminación de elementos subdiagonales. En el k-ésimo, tendremos las siguientes ecuaciones en las filas $(k-1)$ y k-ésima,

$$\alpha_{k-1}\, x_{k-1} + \gamma_{k-1}\, x_k = \beta_{k-1},$$

$$a_k\, x_{k-1} + d_k\, x_k + c_k\, x_{k+1} = b_k.$$

Eliminaremos el elemento subdiagonal a_k si le restamos a la k-ésima ecuación la $(k-1)$-ésima multiplicada por a_k/α_{k-1}, lo que conduce a

$$\alpha_k = d_k - \frac{a_k\,\gamma_{k-1}}{\alpha_{k-1}}, \qquad \gamma_k = c_k, \qquad \beta_k = b_k - \frac{a_k\,\beta_{k-1}}{\alpha_{k-1}}.$$

Aplicaremos este procedimiento para $k = 2, 3, \ldots, n$.

La solución del sistema la podemos obtener resolviendo el sistema bidiagonal superior obtenido mediante sustitución hacia atrás o regresiva,

$$x_n = \frac{\beta_n}{\alpha_n},$$

$$x_k = \frac{\beta_k - \gamma_k\, x_{k+1}}{\alpha_k},$$

para $k = n-1, n-2, \ldots, 1$.

El número de operaciones realizadas en la iteración hacia adelante es de $2\,(n-1)$ divisiones, $2\,(n-1)$ productos y $2\,(n-1)$ sumas, y en la iteración hacia atrás 1 división (si los números $1/\alpha_k$ se guardan durante la iteración hacia adelante), $2\,n+1$ productos y $n-1$ sumas. Es decir, n divisiones, $5\,n-2$ productos y $3\,(n-1)$ sumas de bloques. En total se requieren $O(8\,n)$ operaciones aritméticas. Esta expresión es lineal en el tamaño de la matriz n.

En cuanto a almacenamiento, necesitamos solamente 4 vectores ($4\,n$ elementos) ya que en el paso hacia adelante podemos almacenar el vector α_k en el vector a_k y el β_k en el b_k. Además, es usual almacenar $1/\alpha_k$ en lugar de α_k con objeto de evitar calcular dos veces los números $1/\alpha_k$, hecho que hemos utilizado en el cálculo del número de operaciones previamente presentado.

También podemos resolver un sistema tridiagonal por factorización LU, por ejemplo, de

Doolittle. Escribiendo

$$L = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ l_2 & 1 & 0 & & \vdots \\ 0 & l_3 & 1 & & \\ \vdots & & \ddots & \ddots & \\ 0 & \cdots & & l_n & 1 \end{pmatrix}, \qquad U = \begin{pmatrix} u_1 & c_1 & 0 & \cdots & 0 \\ 0 & u_2 & c_2 & & \vdots \\ & & \ddots & \ddots & \\ \vdots & & & u_{n-1} & c_{n-1} \\ 0 & \cdots & & 0 & u_n \end{pmatrix}.$$

Comparando el producto de L y U con A se obtienen fácilmente las relaciones de recurrencia

$$u_1 = d_1, \qquad l_i = \frac{c_i}{u_{i-1}}, \qquad u_i = d_i - l_i\, e_{i-1}, \qquad i = 2, 3, \ldots, n.$$

Una vez obtenidas las matrices bidiagonales L y U, para obtener la solución de $A\,x = b$, se resuelven los sistemas $L\,y = b$, y $U\,x = y$. Este procedimiento es completamente equivalente a la eliminación de Gauss previamente presentada y su coste es exactamente el mismo.

Un análisis de errores hacia atrás para este procedimiento es sencillo. Aplicando el modelo estándar de la aritmética y el modelo modificado,

$$(1 + \epsilon_i)\, \widehat{l}_i = \frac{c_i}{\widehat{u}_{i-1}}, \qquad |\epsilon_i| \leq u,$$

$$(1 + \tau_i)\, \widehat{u}_i = d_i - \widehat{u}_i\, e_{i-1}(1 + \delta_i), \qquad |\tau_i|, |\delta_i| \leq u,$$

y por tanto se cumplen las cotas de error

$$|c_i - \widehat{l}_i\, \widehat{u}_{i-1}| \leq u\, |\widehat{l}_i\, \widehat{u}_{i-1}|,$$

$$|d_i - \widehat{l}_i\, e_{i-1} - \widehat{u}_i| \leq u\, \left(|\widehat{l}_i\, e_{i-1}| + |\widehat{u}_i| \right).$$

Este resultado de análisis de errores regresivo se escribe matricialmente de la forma siguiente

$$A + \Delta A = \widehat{L}\, \widehat{U}, \qquad |\Delta A| \leq u\, |\widehat{L}|\, |\widehat{U}|.$$

La solución del sistema tridiagonal $A\,x = b$ mediante factorización LU conduce a una solución numérica que cumple las siguientes cotas de error regresivo

$$(\widehat{L} + \Delta L)\, (\widehat{U} + \Delta U)\, \widehat{x} = b, \qquad |\Delta L| \leq u\, |\widehat{L}|, \qquad |\Delta U| \leq (2\, u + u^2)\, |\widehat{U}|,$$

de forma que

$$(A + \Delta A)\, \widehat{x} = b, \qquad |\Delta A| \leq c(u)\, |\widehat{L}|\, |\widehat{U}|, \qquad c(u) = 4\, u + 3\, u^2 + u^3.$$

La demostración, sencilla por otro lado, de estas expresiones se deja como ejercicio para el lector.

Se pueden estudiar las matrices tridiagonales para las que la cota de error regresivo es de la forma $|\Delta A| \leq c(u)\,|A|$, es decir, para las que $|\widehat{L}|\,|\widehat{U}| \leq c\,|\widehat{L}\widehat{U}|$, donde c es una constante. Entre ellas destacan las matrices tridiagonales simétricas definidas positivas ($c = 1$) y las diagonalmente dominantes por filas ($c = 3$) [4].

4.5 Factorización de Cholesky

En muchos problemas aplicados surgen matrices simétricas. Cuando se aplica la factorización a LU a una matriz simétrica, la matriz U no corresponde a $L^\top$, obviamente. Si eliminamos la condición de que L o U sean de diagonal unitariaLos elementos de la diagonal de A,

Toda matriz real A que tenga factorización LU única, y que sea simétrica y definida positiva, entonces tiene una factorización única de la forma $A = L\,L^\top$, donde L es una matriz triangular inferior con diagonal positiva, y se dice que A tiene factorización de Cholesky.

Demostración: Partiendo de la unicidad de la factorización LU Doolittle (o Crout), tenemos que

$$A = L\,U = A^\top = U^\top\,L^\top,$$

donde L y U son matrices no singulares, por lo que tienen inversa y podemos escribir

$$U\,(L^\top)^{-1} = L^{-1}\,U^\top = D,$$

que es una matriz diagonal porque el primer término es una triangular superior y el segundo una inferior. Entonces

$$U = D\,L^\top, \qquad A = L\,D\,L^\top.$$

Por otro lado, D debe ser definida positiva (y por tanto tener diagonal positiva), ya que A es definida positiva,

$$0 \leq \langle x, A\,x\rangle = \langle x, L\,D\,L^\top\,x\rangle = \langle L^\top\,x, D\,L^\top\,x\rangle = \langle y, D\,y\rangle$$

donde $y = L^\top\,x$. Si la expresión anterior es cierta para todo $x \neq 0$ entonces también será cierta para todo $y \neq 0$, ya que $L^\top$ es inversible. Escribiendo ahora

$$\tilde{L} = L\,D^{1/2}, \qquad A = \tilde{L}\,\tilde{L}^\top,$$

con $\tilde{L}$ triangular inferior, **cqd**.

Dada una matriz compleja A que tenga factorización LU única, y que sea hermítica y definida positiva, entonces tiene una factorización única de la forma $A = L\,L^*$, donde L es una matriz triangular inferior con diagonal positiva, y se dice que A tiene factorización de Cholesky. La demostración de este resultado es del todo similar al caso real y se deja al lector.

Para determinar la factorización de Cholesky compararemos uno a uno los elementos de la matriz A y del producto $L\,L^\top$, es decir,

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

Como antes, determinar la expresión de los coeficientes de L es relativamente fácil. Para la primera columna de L,

$$a_{11} = l_{11}^2 \quad \Rightarrow \quad l_{11} = +\sqrt{a_{11}},$$
$$a_{i1} = l_{i1}\,l_{11} \quad \Rightarrow \quad l_{i1} = \frac{a_{i1}}{l_{11}},$$

para la segunda columna de L,

$$a_{22} = l_{21}^2 + l_{22}^2 \quad \Rightarrow \quad l_{22} = +\sqrt{a_{22} - l_{21}^2},$$

$$a_{i2} = l_{i1}\,l_{21} + l_{i2}\,l_{22} \quad \Rightarrow \quad l_{i2} = \frac{1}{l_{22}}\left(a_{i2} - l_{21}\,l_{i1}\right),$$

y, en general, para la j-ésima columna,

$$l_{jj} = +\sqrt{a_{jj} - \sum_{k=1}^{j-1} l_{jk}^2},$$

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

para $1 \leq i < j \leq n$.

Es interesante notar que $l_{jj} > 0$ implica que

$$a_{jj} = \sum_{k=1}^{j} l_{jk}^2 \geq l_{jk}^2,$$

por lo que

$$|l_{jk}| \leq \sqrt{a_{jj}}, \qquad 1 \leq j \leq k \leq n,$$

es decir, la raíz cuadrada de los elementos de la diagonal de A son cotas superiores de los elementos de las correspondientes columnas de L.

Actualmente, el coste de calcular una raíz cuadrada es el mismo que calcular una suma o un producto, por lo que su aparición en el algoritmo no introduce ninguna dificultad. Sin embargo, si se desea, éstas se pueden evitar mediante la factorización de Cholesky modificada,

$$A = L\,D\,L^{\top},$$

donde es una matriz diagonal (de elementos positivos) y L es triangular inferior con diagonal unitaria ($l_{ii} = 1$).

De esta forma

$$A = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ l_{21} & 1 & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ l_{n1} & l_{n2} & \cdots & 1 \end{pmatrix} \begin{pmatrix} d_{11} & 0 & \cdots & 0 \\ 0 & d_{22} & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & d_{nn} \end{pmatrix} \begin{pmatrix} 1 & l_{21} & \cdots & l_{n1} \\ 0 & 1 & \cdots & l_{n2} \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}$$

$$= \begin{pmatrix} 1 & 0 & \cdots & 0 \\ l_{21} & 1 & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ l_{n1} & l_{n2} & \cdots & 1 \end{pmatrix} \begin{pmatrix} d_{11} & d_{11}\,l_{21} & \cdots & d_{11}\,l_{n1} \\ 0 & d_{22} & \cdots & d_{22}\,l_{n2} \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & d_{nn} \end{pmatrix},$$

con lo que igualando coeficientes hemos eliminado las raíces cuadradas. Los detalles se relegan al lector.

Una análisis de la complejidad computacional de la factorización (o factorización modificada) de Cholesky indica que el número de operaciones es del orden de la mitad que en la factorización LU, es decir, $O\big(n^3/3\big)$, lo que supone su ventaja computacional más importante. Además, el almacenamiento de la matrices L y U se reduce a almacenar sólo los elementos de la primera, con lo que el espacio de memoria requerido también se reduce a la mitad.

4.6 Análisis de errores y número de condicionamiento

Para realizar un análisis de errores del problema $A\,x = b$, podemos considerar de forma separada la presencia de errores en el término no homogéneo b, en la matriz de coeficientes A o en ambos.

4.6.1 Errores en el cálculo de la inversa

Tanto el procedimiento de eliminación de Gauss como las factorizaciones LU de Doolittle y Crout pueden ser usadas para calcular la inversa de una matriz. De hecho basta resolver n sistemas lineales cuyos términos no homogéneos sean los vectores de la base canónica (las columnas de la matriz unitaria I) y entonces las soluciones serán las columnas de la inversa. Estudiaremos los errores al calcular dicha inversa.

Sea C una inversa aproximada de A. Llamemos residuo a $R = I - C\,A \equiv 1 - D$. Si el residuo es pequeño $\|R\| < 1$, por la demostración del lema de la sección anterior, existe la inversa de $D \equiv C\,A$ y además

$$\|(C\,A)^{-1}\| = \|(I-R)^{-1}\| \leq \frac{1}{1-\|R\|}.$$

Como existe la inversa de $C\,A$, entonces $|C\,A| = |C|\,|A| \neq 0$ y por tanto los determinantes $|C| \neq 0$ y $|A| \neq 0$, y existen las inversas de C y de A.

Como $R = I - C\,A = (A^{-1} - C)\,A$, se obtiene

$$\frac{\|R\|}{\|A\|\,\|C\|} \leq \frac{\|A^{-1} - C\|}{\|C\|},$$

y como $I - R = C\,A$, tenemos que

$$(I-R)^{-1} = A^{-1}\,C^{-1}, \qquad A^{-1} = (I-R)^{-1}\,C,$$

$$A^{-1} - C = ((I-R)^{-1} - I)\,C,$$

además

$$A^{-1} - C = (I - C\,A)\,A^{-1} = R\,A^{-1} = R\,(I-R)^{-1}\,C,$$

que permite obtener

$$\|A^{-1} - C\| \leq \|R\|\,\|(I-R)^{-1}\|\,\|C\| \leq \frac{\|R\|\,\|C\|}{1-\|R\|},$$

por lo que

$$\frac{\|A^{-1} - C\|}{\|C\|} \leq \frac{\|R\|}{1-\|R\|}.$$

Finalmente podemos acotar el error relativo en la inversa en función de la norma del residuo como

$$\frac{\|R\|}{\|A\|\,\|C\|} \leq \frac{\|A^{-1}-C\|}{\|C\|} \leq \frac{\|R\|}{1-\|R\|}.$$

También podemos acotar el error absoluto de la solución de un sistema lineal en función de la inversa y del residuo de dicho sistema lineal. Sea $\hat{x}$ una solución aproximada del sistema lineal $A\,x = b$, entonces se define el residuo como $r = b - A\,\hat{x}$. Si el residuo es nulo, entonces $\hat{x}$ es la solución exacta. Escribamos

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

$$x - \hat{x} = A^{-1}\,r = (I-(I-C\,A))^{-1}\,C\,r,$$

donde C es una aproximación a la inversa de A. Entonces el error absoluto de la solución se acota por

$$\|x-\hat{x}\| \leq \frac{\|C\,r\|}{1-\|I-C\,A\|} \leq \|C\|\,\|r\|\,(1+O(\|I-C\,A\|)).$$

Por tanto, que el residuo r sea pequeño no indica que el error en la solución sea pequeño, ya que al estar multiplicado por una inversa aproximada de A, si la norma de ésta es muy grande, el error en la solución puede ser mucho más grande que el residuo.

4.7 Sistemas de ecuaciones mal condicionadas

Si la matriz de coeficientes de un sistema lineal de ecuaciones es mucho mayor de la unidad, este sistema está mal condicionado. En esta sección presentaremos ejemplos sencillos de los efectos del mal condicionamiento en la solución del sistema. En algunos de estos ejemplos será posible corregir o al menos reducir los efectos de este comportamiento.

Ejemplo. Sea la matriz

$$A = \begin{pmatrix} 1 & 0 \\ 0 & 10^{-10} \end{pmatrix}, \qquad |A| = 10^{-10},$$

cuyo determinante es muy pequeño, luego debe estar mal condicionada. Estudiemos su número de condicionamiento, para lo cual tenemos que calcular su inversa

$$A^{-1} = \frac{1}{10^{-10}} \begin{pmatrix} 10^{-10} & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 10^{10} \end{pmatrix}.$$

Calculando el número de condicionamiento mediante la norma 1,

$$\|A\|_1 = \max\{1, 10^{-10}\} = 1, \qquad \|A^{-1}\|_1 = \max\{1, 10^{10}\} = 10^{10},$$

por lo que

$$\mathrm{cond}(A) = \|A\|_1 \, \|A^{-1}\|_1 = 10^{10},$$

lo que indica que esta matriz está muy mal condicionada. Este comportamiento se puede corregir fácilmente si multiplicamos la segunda fila por 10^{10}, con lo que se obtiene

$$B = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},$$

que es una matriz que está muy bien condicionada $\kappa(B) = 1$.

Ejemplo. Otro ejemplo de sistema lineal mal condicionado es

$$\begin{pmatrix} 7 & 10 \\ 5 & 7 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 0.7 \end{pmatrix}.$$

La inversa de la matriz de coeficientes es

$$A^{-1} = \begin{pmatrix} -7 & 10 \\ 5 & -7 \end{pmatrix}$$

y por lo tanto la solución del sistema es $x_1 = 0$, $x_2 = 0.1$. El número de condicionamiento de este sistema en norma 1 es

$$\|A\|_1 = 17, \qquad \|A^{-1}\|_1 = 17, \qquad \mathrm{cond}(A) = 17^2 = 289,$$

que indica que la matriz está mal condicionada y un pequeño error en los datos, por ejemplo,

$$\begin{pmatrix} 7 & 10 \\ 5 & 7 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} 1.01 \\ 0.69 \end{pmatrix},$$

conduce a un gran error en la nueva solución $y_1 = -0.17$ e $y_2 = 0.22$ (un error relativo superior al 100%).

Ejemplo. Hay muchas familias de matrices mal condicionadas que surgen en determinados problemas y que exigen técnicas específicas de resolución. Un ejemplo clásico es la matriz de Hilbert

$$H = \begin{pmatrix} 1 & \frac{1}{2} & \frac{1}{3} & \cdots & \frac{1}{n} \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{4} & \cdots & \frac{1}{n+1} \\ \vdots & & & \ddots & \vdots \\ \frac{1}{n} & \frac{1}{n+1} & \cdots & & \frac{1}{2n+1} \end{pmatrix}, \qquad h_{ij} = \frac{1}{i+j-1},$$

cuya inversa se puede calcular de forma exacta. Por ejemplo, en Matlab, `hilb(n)` nos da la matriz de Hilbert de orden `n` e `invhilb(n)` nos da su inversa.

4.7.1 Precondicionado y reescalado

Con objeto de reducir el número de condición de una matriz se utilizar la técnica de precondicionado, en la que se multiplica la matriz de coeficientes por una matriz especialmente elegida. De esta forma, en lugar de resolver $A\,x = b$ se resuelve el sistema $(C\,A)\,x = C\,b$ donde C se elige para que $\mathrm{cond}(C\,A) \ll \mathrm{cond}(A)$. Normalmente C es una aproximación a la inversa de A, por ejemplo, la inversa de la diagonal de A (que conduce a la técnica de reescalado) o la inversa de la parte tridiagonal de A (que es fácil de calcular).

Entre las técnicas de precondicionamiento también se puede utilizar una postmultiplicación, de la forma $A\,C\,C^{-1}\,x \equiv B\,y = b$, donde $\mathrm{cond}(A\,C) \ll \mathrm{cond}(A)$ y $x = C\,y$. En este caso también se recomienda que C sea una aproximación a la inversa de A.

Un estudio detallado de las técnicas de precondicionamiento (como la factorización LU incompleta) está fuera de los objetivos de este curso en el que nos limitaremos a la técnica de reescalado.

Una de las razones por las que un sistema está mal condicionado, y en su solución ocurrirán pérdidas importantes de dígitos significativos, es que los elementos de la matriz A tengan magnitudes muy diferentes. Para evitarlo se puede reescalar la matriz multiplicándola por matrices diagonales D_i, tanto por la izquierda como por la derecha, sea $B = D_1\,A\,D_2$, de forma que resulta el sistema lineal

$$D_1\,A\,D_2\,D_2^{-1}\,x = B\,D_2^{-1}\,x = D_1\,b,$$

que se puede resolver en dos pasos

$$B\,y = D_1\,b, \qquad x = D_2\,y.$$

Una posible técnica de reescalado es intentar conseguir que

$$\max_{1\le j\le n} |b_{ij}| \approx 1,$$

para lo que multiplimos las filas de A por una matriz diagonal de forma que $B = D_1\, A$ nos da

$$b_{ij} = \frac{a_{ij}}{s_i}, \qquad s_i = \max_{1\le j\le n} |a_{ij}|.$$

Ejemplo. Una señal de que un sistema está mal condicionado es que su determinante sea pequeño. Consideremos, por ejemplo, el sistema

$$\left.\begin{aligned} x_1 + 2\,x_2 &= 10 \\ 1.1\,x_1 + 2\,x_2 &= 10.4 \end{aligned}\right\} \quad 0.1\,x_1 = 0.4, \quad x_1 = 4, \quad x_2 = 3,$$

no está bien condicionado porque su determinante vale -0.2, como podemos comprobar introduciendo un pequeño error,

$$\left.\begin{aligned} x_1 + 2\,x_2 &= 10 \\ 1.05\,x_1 + 2\,x_2 &= 10.4 \end{aligned}\right\} \quad 0.05\,x_1 = 0.4, \quad x_1 = 8, \quad x_2 = 1,$$

que ha provocado un gran error en la solución. El sistema perturbado tiene como determinante -0.1, por lo que también está mal condicionado.

El reescalado es una técnica que permite incrementar el tamaño del determinante y, por tanto, reducir el mal condicionamiento del sistema. Multipliquemos las dos fila por 10, lo que da el sistema

$$\left.\begin{aligned} 10\,x_1 + 20\,x_2 &= 100 \\ 11\,x_1 + 20\,x_2 &= 104 \end{aligned}\right\}$$

cuyo determinante es -20. Si introducimos la misma perturbación que antes, $a_{21} \to a_{21} - 0.05$, el nuevo sistema

$$\left.\begin{aligned} 10\,x_1 + 20\,x_2 &= 100 \\ 10.95\,x_1 + 20\,x_2 &= 104 \end{aligned}\right\} \quad x_1 = 4.21, \quad x_2 = 2.89,$$

el error obtenido en la solución es mucho más pequeño. El lector puede verificar que se ha reducido significativamente el número de condicionamiento de la matriz.

4.8 Métodos de corrección residual

Debido al gran número de operaciones que requieren los métodos directos para resolver un sistema lineal, incluso con pivotaje, reescalado, etc., pueden conducir a grandes errores. Estos errores se pueden reducir mediante técnicas iterativas de corrección del error residual.

Sea $x^{(0)}$ la solución de la ecuación $A\,x = b$ obtenida mediante un método directo. Definimos el residuo y el error de dicha solución como

$$r^{(0)} = b - A\,x^{(0)}, \qquad e^{(0)} = x - x^{(0)},$$

respectivamente, por lo que podemos estimar el error cometido resolviendo el siguiente sistema lineal

$$A\,e^{(0)} = r^{(0)}.$$

Si se utiliza un método de factorización LU, se deberían de haber almacenado las matrices L y U con objeto de no tenerlas que volver a recalcular. De esta forma obtenemos una nueva solución aproximada

$$x^{(1)} = x^{(0)} + e^{(0)}.$$

Este procedimiento se puede iterar sucesivamente,

$$A\,e^{(m)} = b - A\,x^{(m)}, \qquad x^{(m+1)} = x^{(m)} + e^{(m)}.$$

Para estudiar la convergencia de la sucesión $x^{(m)}$ que se obtiene de esta forma, consideraremos que el método directo utilizado equivale al cálculo de una matriz inversa aproximada C, de tal forma que

$$x^{(0)} = C\,b.$$

De esta forma, las iteraciones sucesivas nos dan

$$x^{(1)} = x^{(0)} + C\,r^{(0)}, \qquad r^{(0)} = b - A\,x^{(0)},$$

$$x^{(m+1)} = x^{(m)} + C\,r^{(m)}, \qquad r^{(m)} = b - A\,x^{(m)},$$

con lo que el error $e^{(m+1)}$ obtenido es

$$\begin{aligned} e^{(m+1)} &= x - x^{(m+1)} = x - x^{(m)} - C\,r^{(m)} \\ &= x - x^{(m)} - C\,(b - A\,x^{(m)}) \\ &= x - x^{(m)} - C\,(A\,x - A\,x^{(m)}) \\ &= (I - C\,A)\,(x - x^{(m)}), \end{aligned}$$

con lo que

$$\|x - x^{(m+1)}\| = \|e^{(m+1)}\| \leq \|I - C\,A\|^m \, \|x - x^{(0)}\|.$$

Si la inversa aproximada C es suficientemente próxima a la inversa de A, es decir, si

$$\|I - C\,A\| < 1$$

entonces

$$\lim_{m\to\infty} \|I - C\,A\|^m = 0,$$

y tenemos que la sucesión de soluciones aproximadas converge

$$\lim_{m\to\infty} x^{(m)} = x.$$

El orden de convergencia obtenido es lineal, es decir, existe una constante c tal que

$$\|x - x^{(m+1)}\| \leq c\,\|x - x^{(m)}\|.$$

Seguidamente mostraremos que dicha constante vale

$$c \approx \max \frac{\|x^{(m+2)} - x^{(m+1)}\|}{\|x^{(m+1)} - x^{(m)}\|}.$$

Aplicando normas a la ecuación

$$x - x^{(m+1)} = (I - C\,A)\,(x - x^{(m)}),$$

obtenemos

$$\|x - x^{(m+1)}\| \leq c\,\|x - x^{(m)}\|, \qquad c = \|I - C\,A\|.$$

Restando las dos ecuaciones siguientes

$$x - x^{(m+2)} = (I - C\,A)\,(x - x^{(m+1)}),$$

$$x - x^{(m+1)} = (I - C\,A)\,(x - x^{(m)}),$$

se obtiene

$$x^{(m+2)} - x^{(m+1)} = (I - C\,A)\,(x^{(m+1)} - x^{(m)}),$$

por lo que aplicando normas

$$\|x^{(m+2)} - x^{(m+1)}\| \leq \|I - C\,A\|\,\|x^{(m+1)} - x^{(m)}\|,$$

por lo que

$$c = \|I - C\,A\| \geq \frac{\|x^{(m+2)} - x^{(m+1)}\|}{\|x^{(m+1)} - x^{(m)}\|}.$$

BIBLIOGRAFÍA

[1] Granero Rodríguez, Francisco, "Álgebra y geometría analítica," McGraw-Hill / Interamericana de España, 1985. [FTM-4-c/GRA/alg (5)]

[2] Hernández Rodríguez, Eugenio, "Álgebra y geometría," (2ª ed.), Addison-Wesley Iberoamericana España, 1998. [FTM-4/HER (5)]

[3] Burgos Román, Juan de, "Álgebra lineal," McGraw-Hill / Interamericana de España, 1993. [FTM-4-c/B (9)]

[4] Nicholas J. Higham, "*Accuracy and Stability of Numerical Algorithms*," SIAM, Philadelphia (1996).

[5] G.W. Stewart, "*Matrix Algorithms. Volume 1, Basic Decompositions*," SIAM, Philadelphia (1998).