Cuaderno de Investigación:
Métodos Iterativos para Sistemas Lineales

Carlos Daniel Acosta Medina
Departamento de Matemáticas y Estadística
Universidad Nacional de Colombia Sede Manizales

Carlos Enrique Mejía Salazar
Escuela de Matemáticas
Universidad Nacional de Colombia Sede Medellín

Dirección de Investigación Sede Manizales
Universidad de Nacional de Colombia
Abril de 2005

Contenido

Introducción

En este trabajo presentamos conceptos y técnicas del álgebra lineal numérica que hemos venido estudiando en los últimos meses. El tema central es la solución por métodos iterativos de sistemas de ecuaciones lineales, que es un tópico indispensable cuando en Matemática Aplicada, Ciencias e Ingeniería, se requiere resolver un problema lineal o no lineal. Esperamos por tanto que estas notas puedan resultar útiles a personas de diversas profesiones interesadas en resolver problemas de forma numérica.

Además de teoría básica de álgebra lineal numérica, incluimos explicaciones detalladas de métodos específicos, por ejemplo, los métodos iterativos LSQR y CG y precondicionamiento por factorización incompleta y por un método recientemente desarrollado que es especial para matrices 2×2 por bloques.

En nuestro concepto, lo más destacado del trabajo es el tratamiento de la factorización incompleta y el capítulo de resultados numéricos. La factorización incompleta la explicamos a partir de la fuente original [13] y del libro reciente [20]. Nos parece que en esta parte logramos simultáneamente originalidad y claridad.

El capítulo de resultados numéricos incluye una amplia combinación de métodos aplicados a problemas de diversa índole y procedencia. Consideramos entre otros, discretizaciones de elementos finitos, problemas en los que no se almacena la matriz sino que se dispone solamente de su acción por medio de una rutina, etc.

Todo el trabajo fue redactado por nosotros con base en fuentes modernas o clásicas, que en ocasiones, son también fuentes originales. También es trabajo nuestro la preparación de todos los programas de computador que utilizamos. Todos ellos son hechos para MATLAB y aprovechan la calidad de las rutinas que trae MATLAB para la solución de sistemas por medio de métodos iterativos de la familia de los subespacios de Krylov.

El trabajo consta de dos capítulos teóricos y un capítulo de resultados numéricos. En el primer capítulo teórico presentamos los métodos itera-

tivos no estacionarios, algunos con bastante detalle. En el segundo capítulo, también teórico, estudiamos el precondicionamiento incluido el precondicionamiento de un tipo especial de matrices 2×2 por bloques. El tercer capítulo sirve para ilustrar todos los métodos descritos en los dos capítulos previos por medio de ejemplos cuidadosamente escogidos.

Capítulo 1

Métodos No Estacionarios

1.1 Gradiente Conjugado

1.1.1 Problema

Sea $A \in \mathbb{R}^{n\times n}$ matriz real simétrica y definida positiva, en adelante *s.d.p.* Es decir, para todo $x \in \mathbb{R}^n$, $x \neq 0$,

$$\begin{aligned} A &= A^T, \\ x^T A x &> 0. \end{aligned} \tag{1.1}$$

Sea ahora $b \in \mathbb{R}^n$, el problema propuesto es resolver el sistema de ecuaciones lineales

$$Ax = b. \tag{1.2}$$

Debe notarse que como A es *s.d.p.*, entonces es invertible y la solución buscada es

$$\tilde{x} = A^{-1}b. \tag{1.3}$$

Ahora como A es definida positiva, entonces induce una norma, a saber la dada por:

$$\|x\|_A = \left(x^T A x\right)^{1/2}.$$

Entonces, el problema original es equivalente a resolver:

$$Hallar\ el\ \tilde{x}\ que\ hace\ \|A\tilde{x} - b\|^2_{A^{-1}} = 0. \tag{1.4}$$

1.1.2 Formulación Variacional

Nótese ahora que $Ax = b$ si y sólo si para todo v en $\mathbb{R}^n$

$$v^T Ax = v^T b.$$

Si definimos $a : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ y $L : \mathbb{R}^n \to \mathbb{R}$ por:

$$a(v, u) = v^T Au,$$
$$L(v) = v^T b.$$

Entonces nuestro problema puede formularse como: Hallar $\tilde{x} \in \mathbb{R}^n$ tal que para todo $v \in \mathbb{R}^n$,

$$a(v, \tilde{x}) = L(v). \tag{1.5}$$

Nótese que el funcional a es una forma bilineal simétrica y definida positiva, gracias a las propiedades de A. Note también que L es un operador lineal.

1.1.3 Formulación Funcional

Si definimos ahora $\phi : \mathbb{R}^n \to \mathbb{R}$ por

$$\phi(x) = \frac{1}{2} x^T Ax - x^T b. \tag{1.6}$$

Entonces

$$\nabla\phi(x) = Ax - b,$$
$$H\phi(x) = A.$$

Así, el único punto crítico de ϕ es $\tilde{x}$ y es un punto mínimo pues el Hessiano de ϕ es A, la cual tiene valores propios positivos por ser $s.d.p.$

Por tanto nuestro problema es equivalente a:

Hallar el $\tilde{x}$ que minimiza ϕ sobre todo $\mathbb{R}^n$

1.1.4 Subespacios de Krylov

Sea x_0 una primera aproximación de la solución, definimos su residual por

$$r_0 = b - Ax_0,$$

y definimos el subespacio de Krylov de orden k de A y respecto de r_0 por

$$K_k(A, r_0) = gen\left\{r_0, Ar_0, A^2r_0, \cdots, A^{k-1}r_0\right\}.$$

Definimos también la $k - ésima$ iteración de gradiente conjugado como el x_k que minimiza ϕ sobre el espacio afín

$$x_0 + K_k(A, r_0).$$

O equivalentemente, definimos x_k como el único elemento de $V_k = x_0 + K_k(A, r_0)$, tal que

$$\|x_k - \tilde{x}\|_A = \|Ax_k - b\|_{A^{-1}} = \min_{x \in V_k} \|Ax - b\|_{A^{-1}}.$$

Revisamos enseguida que x_k esté bien definido.

1.1.5 Existencia y Unicidad

Lema 1.1.1 *(Proyección sobre Subespacios) Sean S un subespacio de $\mathbb{R}^n$ y $y_0 \in \mathbb{R}^n$. Sea $\langle \cdot, \cdot \rangle$ un producto interior en $\mathbb{R}^n$ y $\|\cdot\|$ la norma inducida por este producto interior $\left(\|\cdot\|^2 = \langle \cdot, \cdot \rangle\right)$. Existe un único $y \in S$ tal que*

$$\|y_0 - y\| = \min_{z \in S} \|y_0 - z\|.$$

Prueba. Empezamos notando que como S es finito dimensional, entonces S es un subconjunto cerrado de $\mathbb{R}^n$. Escogemos ahora una sucesión $\{z_k\}$ de puntos de S tal que

$$d_k = \|y_0 - z_k\| \ \rightarrow \ \inf_{z \in S} \|y_0 - z\| = d.$$

Note que este d es la distancia de y_0 a S. Note también que si $\{z_k\}$ es convergente con límite y, tal límite debe ser un elemento de S y cumple la propiedad requerida.

Mostraremos que $\{z_k\}$ es una sucesión de Cauchy con lo que ha de ser convergente en S por la cerradura de S y la completez de $\mathbb{R}^n$.

En efecto, como $\|\cdot\|$ es norma inducida por producto interior, se cumple la ley del paralelogramo que establece que para todo $x, y \in \mathbb{R}^n$

$$\|x + y\|^2 + \|x - y\|^2 = 2(\|x\|^2 + \|y\|^2).$$

Sean ahora para $k, m \in \mathbb{Z}^+$, con $k \geq m$. Entonces haciendo $x = y_0 - z_k$, $y = y_0 - z_m$ se tiene que

$$\|2y_0 - (z_k + z_m)\|^2 + \|z_k - z_m\|^2 = 2\left(d_k^2 + d_m^2\right),$$
$$\left\|y_0 - \frac{(z_k + z_m)}{2}\right\|^2 + \left\|\frac{z_k - z_m}{2}\right\|^2 = \frac{1}{2}\left(d_k^2 + d_m^2\right),$$

y como $\frac{z_k+z_m}{2} \in S$ se tiene $d^2 \leq \left\|y_0 - \frac{z_k+z_m}{2}\right\|^2$. Así,

$$\left\|\frac{z_k - z_m}{2}\right\|^2 \leq \frac{1}{2}\left(d_k^2 + d_m^2\right) - d^2$$

luego $\|z_k - z_m\|^2 \to 0$, si $k, m \to +\infty$. Lo que prueba que $\{z_k\}$ es de Cauchy.

Para mostrar ahora la unicidad de y, supongamos que $y_1, y_2 \in S$ son tales que

$$\|y_0 - y_1\| = \|y_0 - y_2\| = \min_{z \in S} \|y_0 - z\| = d.$$

Entonces, de la ley del paralelogramo con $x = y_0 - y_1$, $y = y_0 - y_1$

$$\left\|y_0 - \frac{(y_1 + y_2)}{2}\right\|^2 + \left\|\frac{y_1 - y_2}{2}\right\|^2 = d^2,$$

pero $d^2 \leq \left\|y_0 - \frac{y_1+y_2}{2}\right\|^2$. Así

$$\left\|\frac{y_1 - y_2}{2}\right\|^2 \leq 0.$$

O sea, $y_1 = y_2$.

Teorema 1.1.2 *(Existencia y Unicidad de Iteración CG) Existe un único elemento $x_k \in V_k = x_0 + K_k(A, r_0)$, tal que*

$$\|x_k - \tilde{x}\|_A = \|Ax_k - b\|_{A^{-1}} = \min_{x \in V_k} \|Ax - b\|_{A^{-1}}.$$

Prueba: Este resultado es una aplicación directa del lema anterior haciendo: $\|\cdot\| = \|\cdot\|_A$, $S = K_k(A, r_0)$ y $y_0 = \tilde{x} - x_0$. Observando que,

$$\begin{aligned}\min_{x \in V_k} \|Ax - b\|_{A^{-1}} &= \min_{x \in V_k} \|x - \tilde{x}\|_A \\ &= \min_{x \in V_k} \|(x - x_0) - (\tilde{x} - x_0)\|_A \\ &= \min_{z \in S} \|z - y_0\|_A .\end{aligned}$$

Entonces $x_k = y + x_0$ es el vector requerido, donde y es el elemento de S garantizado por el lema.

1.1.6 Implementación CG

Nos concentramos ahora en la forma de calcular x_k, empezamos por observar que para todo v en $K_k(A, r_0)$ la función

$$f(t) = \phi(x_k + tv)$$

tiene un mímimo en $t = 0$. Pero,

$$f'(t) = \nabla\phi(x_k + tv)^T v.$$

Entonces,

$$\begin{aligned} 0 &= f'(0) \\ &= \nabla\phi(x_k)^T v \\ &= (Ax_k - b)^T v \\ &= -r_k^T v. \end{aligned}$$

O sea

$$r_k \perp K_k. \tag{1.7}$$

Hacemos

$$x_{k+1} = x_k + \alpha_{k+1} p_{k+1},$$

donde $p_{k+1} \in K_{k+1}$ y α_{k+1} es un escalar apropiado. Entonces, $Ap_{k+1} \in K_{k+2}$ y

$$\begin{aligned} b - Ax_{k+1} &= b - Ax_k - \alpha_{k+1} Ap_{k+1}, \\ r_{k+1} &= r_k - \alpha_{k+1} Ap_{k+1}. \end{aligned}$$

Con ello

$$r_k \in K_{k+1}.$$

Asi,

$$K_k = gen\left\{r_0, r_1, \cdots, r_{k-1}\right\}.$$

Ahora, para todo v en K_{k+1}

$$\begin{aligned} \nabla\phi(x_k + \alpha_{k+1} p_{k+1})^T v = 0, \\ \left(Ax_k + \alpha_{k+1} Ap_{k+1} - b\right)^T v = 0, \\ \left(Ax_k - b\right)^T v + \left(\alpha_{k+1} Ap_{k+1}\right)^T v = 0. \end{aligned}$$

Entonces, para todo v en K_k

$$\alpha_{k+1} p_{k+1}^T A v = 0. \tag{1.8}$$

Afirmamos que existe un escalar β_{k+1} que nos permite escribir

$$p_{k+1} = r_k + \beta_{k+1} p_k.$$

En efecto, debe ser

$$p_{k+1}^T A r_j = 0,\ j = 1, 2, \cdots, k.$$

O sea,

$$\begin{aligned} p_{k+1}^T A r_j &= \left(r_k + \beta_{k+1} p_k\right)^T A r_j \\ &= r_k^T A r_j + \beta_{k+1} p_k^T A r_j \\ &= 0. \end{aligned}$$

para $j = 1, 2, \cdots, k-1$.

Pero para $j = 1, 2, \cdots, k-2$

$$\begin{aligned} r_k^T A r_j &= 0, \\ p_k^T A r_j &= 0. \end{aligned}$$

Sólo falta tener,

$$r_k^T A r_{k-1} + \beta_{k+1} p_k^T A r_{k-1} = 0,$$

es decir,

$$\beta_{k+1} = \frac{-r_k^T A r_{k-1}}{p_k^T A r_{k-1}}. \tag{1.9}$$

Ahora, por definición la función

$$g(\alpha) = \phi(x_k + \alpha p_{k+1}),$$

tiene mínimo en $\alpha = \alpha_{k+1}$, entonces $g'(\alpha_{k+1}) = 0$, pero

$$\begin{aligned} g'(\alpha_{k+1}) &= \nabla\phi(x_k + \alpha_{k+1} p_{k+1})^T p_{k+1} \\ &= \left(A\left(x_k + \alpha_{k+1} p_{k+1}\right) - b\right)^T p_{k+1} \\ &= \left(A x_k - b\right)^T p_{k+1} + \left(\alpha_{k+1} A p_{k+1}\right)^T p_{k+1} \\ &= -r_k^T p_{k+1} + \alpha_{k+1} p_{k+1}^T A p_{k+1}. \end{aligned}$$

Entonces,

$$\alpha_{k+1} = \frac{r_k^T p_{k+1}}{p_{k+1}^T A p_{k+1}}. \tag{1.10}$$

Antes de escribir un algoritmo hacemos las siguientes afirmaciones

$$\alpha_{k+1} = \frac{\|r_k\|^2}{p_{k+1}^T A p_{k+1}},$$
$$\beta_{k+1} = \frac{\|r_k\|^2}{\|r_{k-1}\|^2}.$$

En efecto,

$$r_k^T p_{k+1} = r_k^T (r_k + \beta_{k+1} p_k) = \|r_k\|^2.$$

Ahora, por construcción

$$p_{k+1}^T A p_k = 0,$$

o sea

$$\begin{aligned} 0 &= (r_k + \beta_{k+1} p_k)^T A p_k \\ &= r_k^T A p_k + \beta_{k+1} p_k^T A p_k \end{aligned}$$

luego

$$\beta_{k+1} = \frac{-r_k^T A p_k}{p_k^T A p_k}.$$

Pero

$$\begin{aligned} p_k^T A p_k &= p_k^T A (r_{k-1} + \beta_k p_{k-1}) \\ &= p_k^T A r_{k-1} + \beta_k p_k^T A p_{k-1} \\ &= p_k^T A r_{k-1}. \end{aligned}$$

Entonces,

$$\beta_{k+1} = \frac{-r_k^T A p_k}{p_k^T A r_{k-1}}.$$

Ahora,

$$\begin{aligned} 0 &= r_k^T r_{k-1} \\ &= (r_{k-1} - \alpha_k A p_k)^T r_{k-1} \\ &= \|r_{k-1}\|^2 - \alpha_k p_k^T A r_{k-1}, \end{aligned}$$

entonces

$$\beta_{k+1} = \frac{-r_k^T A p_k}{\|r_{k-1}\|^2} \alpha_k.$$

Además,

$$\begin{aligned} \|r_k\|^2 &= r_k^T r_k \\ &= (r_{k-1} - \alpha_k A p_k)^T r_k \\ &= -\alpha_k p_k^T A r_k. \end{aligned}$$

Así,

$$\beta_{k+1} = \frac{\|r_k\|^2}{\|r_{k-1}\|^2}.$$

1.1.7 Algoritmo CG

Supongamos que se ha calculado x_k y hemos salvado p_k y r_{k-1}, podemos entonces calcular en su orden

$$\begin{aligned} r_k &= b - A x_k, \\ \beta_{k+1} &= \frac{\|r_k\|^2}{\|r_{k-1}\|^2}, \\ p_{k+1} &= r_k + \beta_{k+1} p_k, \\ \alpha_{k+1} &= \frac{\|r_k\|^2}{p_{k+1}^T A p_{k+1}}, \\ x_{k+1} &= x_k + \alpha_{k+1} p_{k+1}. \end{aligned}$$

Podemos postular entonces el siguiente seudo código,

```
Input A, x, b, tol, M
r = b − A ∗ x;
rho_old = norm(r)^2;
rho_new = rho_old;
for k = 1 : M
    if k == 1
        p = r;
    else
        betha = rho_new/rho_old;
        p = r + betha ∗ p;
    end
    w = A ∗ p;
```

```
        alpha = rho_new/(p' * w);
        x = x + alpha * p;
        r = r - alpha * w;
        rho_old = rho_new;
        rho_new = norm(r)^2;
        if sqrt(rho_new) < tol * norm(b)
            break
        end
    end
```

1.1.8 Análisis de Convergencia

Encontramos ahora las características de convergencia del algoritmo CG. Empezamos por notar que si $x \in V_k$, entonces existen escalares $\alpha_1, \alpha_2, \cdots, \alpha_k$ tales que:

$$\begin{aligned}
\tilde{x} - x &= \tilde{x} - x_0 - \sum_{i=0}^{k-1} \alpha_i A^i r_0 \\
&= e_0 - \sum_{i=0}^{k-1} \alpha_i A^{i+1} e_0 \\
&= (I - \sum_{i=0}^{k-1} \alpha_i A^{i+1}) e_0 \\
&= p(A) e_0.
\end{aligned}$$

Donde $p(x) = 1 - \sum_{i=0}^{k-1} \alpha_i x^{i+1}$, note que p es polinomio de orden k tal que $p(0) = 1$. Por conveniencia construimos el conjunto $\wp_k$, como

$$\wp_k = \{p \text{ polinomios de orden } k \text{ tales que } p(0) = 1\}.$$

Note que cada elemento de V_k induce un único elemento de $\wp_k$ y viceversa. Entonces, si notamos p_k el polinomio asociado a x_k, de la definición de la iteración CG se sigue que :

$$\begin{aligned}
\|e_k\|_A &= \|\tilde{x} - x_k\|_A \\
&= \|p_k(A) e_0\|_A \\
&= \min_{p \in \wp_k} \|p(A) e_0\|_A .
\end{aligned}$$

Ahora, como A real simétrica y definida positiva existe una base ortogonal de $\mathbb{R}^n$ constituida por vectores propios de A y todos su valores propios son positivos. Sean $0 < \lambda_n \leq \lambda_{n-1} \leq \cdots \leq \lambda_2 \leq \lambda_1$ los valores propios de A y $\{v_1, v_1, \cdots, v_n\}$ una base tal, con v_i asociado a λ_i.

Escribimos ahora e_0 como combinación lineal de los elementos de esta base,

$$e_0 = \sum_{i=1}^{n} \beta_i v_i.$$

Entonces

$$\|e_0\|_A^2 = \sum_{i=1}^{n} \beta_i^2 \lambda_i$$

y

$$\begin{aligned} \|p(A)e_0\|_A^2 &= \left\| \sum_{i=1}^{n} \beta_i p(\lambda_i) v_i \right\|_A^2 \\ &= \sum_{i=1}^{n} \beta_i^2 \lambda_i \left(p(\lambda_i)\right)^2 \\ &\leq \|e_0\|_A^2 \max_{1 \leq i \leq n} \left(p(\lambda_i)\right)^2. \end{aligned}$$

Así,

$$\frac{\|p(A)e_0\|_A^2}{\|e_0\|_A^2} \leq \max_{1 \leq i \leq n} \left(p(\lambda_i)\right)^2,$$

para todo $p \in \wp_k$, entonces

$$\inf_{p \in \wp_k} \frac{\|p(A)e_0\|_A}{\|e_0\|_A} \leq \inf_{p \in \wp_k} \max_{1 \leq i \leq n} |p(\lambda_i)|.$$

Tenemos el siguiente:

Teorema 1.1.3 (Convergencia CG) *Si $r_{k-1} \neq 0$, entonces en el $k-$ésimo paso de la iteración de CG se tiene que:*

$$\begin{aligned} \frac{\|e_k\|_A}{\|e_0\|_A} &= \frac{\|p_k(A)e_0\|_A}{\|e_0\|_A} \\ &= \inf_{p \in \wp_k} \frac{\|p(A)e_0\|_A}{\|e_0\|_A} \\ &\leq \inf_{p \in \wp_k} \max_{\lambda \in \sigma(A)} |p(\lambda)|. \end{aligned}$$

Donde $\sigma(A)$ representa el espectro de A.

Corolario 1.1.4 *Si A tiene sólo m valores propios distintos, entonces la iteración CG coverge en a lo más m pasos.*

Prueba:

Basta considerar $p(x) = \prod_{i=1}^{m}(1-x/\lambda_i) \in \wp_m$, para el cual $\max_{\lambda\in\sigma(A)} |p(\lambda)| = 0$. Así, $\|e_m\|_A = 0$.

Obtenemos ahora otro estimado para la razón de convergencia dependiente del número de condición en 2-norma de A, $\kappa = \lambda_{\max}/\lambda_{\min}$.

Teorema 1.1.5 *Si κ es el número de condición de la matriz s.d.p. A, entonces los errores de la iteración CG aplicada al problema $Ax = b$ satisfacen*

$$\begin{aligned}\frac{\|e_m\|_A}{\|e_0\|_A} &\leq 2/\left[\left(\frac{\sqrt{\kappa}+1}{\sqrt{\kappa}-1}\right)^m + \left(\frac{\sqrt{\kappa}+1}{\sqrt{\kappa}-1}\right)^{-m}\right] \\ &\leq 2\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^m.\end{aligned}$$

Prueba:

Del teorema de Convergencia de CG, basta conseguir un polinomio p cuyo máximo valor para $\lambda \in [\lambda_{\min,}\lambda_{\max}]$ sea la expresión de en medio.

Antes de indicar el polinomio adecuado, recordamos la recurrerencia que define los polinomios de Chebyshev y sus principales propiedades.

Los polinomios de Chebyshev T_k pueden generarse por la recurrencia: $T_0(x) = 1$, $T_1(x) = x$,

$$T_k(x) = 2xT_{k-1}(x) - T_{k-2}(x),\ k = 2, 3, \cdots$$

Estos polinomios satisfacen:

- El coeficiente de x^m de $T_m(x)$ es 2^{m-1} para $m \geq 1$.
- T_{2m} es par y T_{2m+1} es impar.
- $T_m(x) = \cos(m \arccos(x))$, para $-1 \leq x \leq 1$.
- $T_m(x)$ tiene m ceros distintos todos en $[-1, 1]$ y son

$$p_k = \cos\left(\frac{(2k+1)\,\pi}{2m}\right),\ k = 0, 1, 2, \cdots, m-1.$$

- $|T_m(x)| \leq 1$, para $-1 \leq x \leq 1$.

- Si $x = \frac{1}{2}\left(z + z^{-1}\right)$, con $z \in \mathbb{R}$ entonces $T_m(x) = \frac{1}{2}\left(z^m + z^{-m}\right)$.

La demostración de estas propiedades se omiten por no ser de nuestro interés, su obtención se basa en la recurrencia y en propiedades de las funciones trigonométricas.

Sea ahora

$$\gamma = \frac{\lambda_{\max} + \lambda_{\min}}{\lambda_{\max} - \lambda_{\min}} = \frac{\kappa + 1}{\kappa - 1},$$

entonces para $x \in [\lambda_{\min,}\lambda_{\max}]$ se tiene que:

$$\gamma - \frac{2x}{\lambda_{\max} - \lambda_{\min}} \in [-1, 1].$$

Definimos ahora

$$p(x) = T_m\left(\gamma - \frac{2x}{\lambda_{\max} - \lambda_{\min}}\right) / T_m(\gamma).$$

Entonces,

$$\max_{\lambda \in \sigma(A)} |p(\lambda)| \leq \frac{1}{|T_m(\gamma)|}.$$

Ahora escogemos z de modo que $\gamma = \frac{1}{2}\left(z + z^{-1}\right)$, así

$$z^2 - 2\gamma z + 1 = 0.$$

Tomamos entonces la solución a la ecuación cuadrática

$$\begin{aligned} z &= \left(\gamma + \sqrt{\gamma^2 - 1}\right) \\ &= \left(\frac{\kappa + 1}{\kappa - 1} + \sqrt{\left(\frac{\kappa + 1}{\kappa - 1}\right)^2 - 1}\right) \\ &= \left((\kappa + 1) + \sqrt{(\kappa + 1)^2 - (\kappa - 1)^2}\right) / (\kappa - 1) \end{aligned}$$

Así,

$$\begin{aligned} z &= \left((\kappa + 1) + \sqrt{4\kappa}\right) / (\kappa - 1) \\ &= \frac{\left(\sqrt{\kappa} + 1\right)^2}{\left(\sqrt{\kappa} + 1\right)\left(\sqrt{\kappa} - 1\right)} \\ &= \left(\frac{\sqrt{\kappa} + 1}{\sqrt{\kappa} - 1}\right). \end{aligned}$$

Para este valor de z tenemos

$$\begin{aligned}\max_{\lambda\in\sigma(A)} |p(\lambda)| &\leq \frac{1}{|T_m(\gamma)|} \\ &= 2/\left[\left(\frac{\sqrt{\kappa}+1}{\sqrt{\kappa}-1}\right)^m + \left(\frac{\sqrt{\kappa}+1}{\sqrt{\kappa}-1}\right)^{-m}\right].\end{aligned}$$

Que es el resultado deseado.

1.1.9 Los Métodos CGNR y CGNE

Cuando se trabaja con matrices no singulares no simétricas es posible seguir utilizando gradiente conjugado a través de un cambio en el sistema a resolver por uno equivalente cuya matriz de coeficientes sea *s.d.p.* Existen dos formas clásicas de llevar esto a cabo. La primera se conoce como CGNR y consiste en remplazar el sistema $Ax = b$ por el sistema de ecuaciones normales

$$A^T Ax = A^T b$$

y resolver éste usando CG. Es importante notar que

$$\begin{aligned}\|\tilde{x} - x\|^2_{A^TA} &= (\tilde{x} - x)^T A^T A (\tilde{x} - x) \\ &= (A\tilde{x} - Ax)^T (A\tilde{x} - Ax) \\ &= (b - Ax)^T (b - Ax) \\ &= \|r\|_2^2 .\end{aligned}$$

Se sigue, de la propiedad de minimización de CG, que CGNR escoge x_k en $x_0 + \mathcal{K}_k(A^TA, r_0)$ de modo que se minimiza el residual $\|b - Ax\|_2$. De aquí el nombre CGNR que significa *CG* aplicado a ecuaciones *Normales* para minimizar el *Residual.*

La segunda forma de abordar problemas lineales no singulares no simétricos es la conocida como CGNE, que significa *CG* aplicado a ecuaciones *Normales* para minimizar el *Error.* En este caso el sistema que se resuelve por CG es el sistema

$$AA^T y = b,$$

y luego se hace $x = A^T y$. La naturaleza del nombre de la estrategia se debe

a que en este caso la propiedad de minimización es de la forma

$$\begin{aligned}\|\tilde{y}-y\|_{AA^T}^2 &= (\tilde{y}-y)^T AA^T (\tilde{y}-y)\\ &= (A\tilde{y}-Ay)^T (A\tilde{y}-Ay)\\ &= (\tilde{x}-x)^T (\tilde{x}-x)\\ &= \|e\|_2^2 .\end{aligned}$$

1.2 Método de Residuos Mínimos Generalizado

En contraste con CG, *Residuos Mínimos Generalizado (GMRES)* es usado para sistemas no simétricos. El proposito de GMRES es el de minimizar la $2-norma$ del residuo sobre $\mathcal{K}_k$. Es decir la GMRES iteración esta caracterizada por

$$x_k \in x_0 + \mathcal{K}_k \ ; \ \|b - Ax_k\|_2 = \min_{z\in\mathcal{K}_k} \|r_0 - Az\|_2 \tag{1.11}$$

Esto es, GMRES pretende resolver un problema de mínimos cuadrados sobre el espacio $\mathcal{K}_k$, [21]. El análisis de este método se llevará a cabo suponiendo que no ha sido dada solución inicial o equivalentemente que la solución inicial es el vector nulo. Si el método debe aplicarse a un problema con solución inicial, basta aplicarlo al problema $Az = r_0$ y hacer $x = z + x_0$.

1.2.1 Existencia y Unicidad.

Mientras que CG utiliza como base para $\mathcal{K}_k$ el conjuto de residuos r_k, GMRES utiliza una base ortonormal que se forma aplicando directamente el proceso de ortonormalización de *Gram-Schmidt modificado* [21] a la sucesión $b, Ab, A^2b, ...$ El algoritmo correspondiente es conocido como *Algoritmo de Arnoldi* y los vectores (v_j) que obtienen son llamados *Vectores de Arnoldi* y constituyen una base ortonormal para $\mathcal{K}_k$. El algoritmo de Arnoldi es también utilizado para calcular los valores propios de una matriz [21].

Algoritmo de Arnoldi

$v_1 = \frac{r_0}{\|r_0\|_2}$
$for\ k = 1 : M$
    $w = Av_k$
    $for\ j = 1 : k$
        $h_{jk} = \langle v_j, w\rangle$

$w = w - h_{jk}v_j$
end
$h_{k+1,k} = \|w\|$
if $h_{k+1,k} \neq 0$,
$v_{k+1} = \frac{w}{h_{k+1,k}}$
else
break
end
end

Nótese que para el cálculo de v_{k+1} es necesario que

$$Av_k - \sum_{j=1}^{k} \langle v_j, Av_k \rangle v_j \neq 0$$

Si esto no ocurre entonces $Av_k \in \mathcal{K}_k = gen\,(v_1, v_2, ..., v_k) = gen\,(b, Ab, A^2b, ..., A^{k-1}b)$, $A^kb \in \mathcal{K}_k$ y por tanto $\mathcal{K}_{k+1} = gen\,(b, Ab, A^2b, ..., A^{k-1}b, A^kb) \subseteq \mathcal{K}_k$. De donde se sigue que $\mathcal{K}_k = \mathcal{K}_{k+1} = \mathcal{K}_{k+2} = ...$ En consecuencia, la información obtenida por el algoritmo de Arnoldi es suficiente para encontrar la base ortonormal deseada. Además, en tal paso k se puede encontrar la solución exacta al sistema, esto se prueba en el siguiente lema.

Lema 1.2.1 *Sean A una matriz real $n \times n$ no singular y $b \in \mathbb{R}^n$. Suponga que los vectores de Arnoldi $v_1, v_2, ..., v_k$ han sido calculados y que*

$$Av_k = \sum_{j=1}^{k} \langle v_j, Av_k \rangle v_j$$

entonces $x = A^{-1}b \in \mathcal{K}_k$.

Prueba. Puesto que $V_k = [v_1, v_2, ..., v_k]$ es una matriz ortogonal $n \times k$, entonces $H = V_k^T AV_k$ es una matriz $k \times k$ no singular. Defínase $\beta = \|b\|_2$ y sea $y \in \mathbb{R}^k$ cualquiera. Así, $V_k y \in \mathcal{K}_k$ y

$$\|b - AV_k y\|_2 = \|\beta v_1 - V_k Hy\|_2 = \|V_k(\beta e_1 - Hy)\|_2$$

donde $e_1 = (1, 0, ..., 0)^T \in \mathbb{R}^k$. Luego, tomando $y = \beta H^{-1}e_1$ se obtiene $b - AV_k y = 0$ y así $x = A^{-1}b = V_k y$.

Es importante notar que este lema muestra que el algoritmo de Arnoldi calcula v_{k+1} mientras que $x = A^{-1}b \notin \mathcal{K}_k$.

Teorema 1.2.2 *(Existencia y Unicidad de Iteración GMRES) Supóngase que $r_j \neq 0$ para $0 \leq j \leq k-1$. Existe un único $x_k \in \mathcal{K}_k$ que satisface (1.11).*

Prueba. Cómo $\min_{y\in\mathcal{K}_{k-1}} \|b-Ay\|_2 = r_{k-1} \neq 0$, entonces $x \notin \mathcal{K}_{k-1}$. Sean $V_k = [v_1, v_2, ..., v_k]$, donde $\{v_1, v_2, ..., v_k\}$ es la base de Arnoldi para $\mathcal{K}_k$. Haciendo $S = Rango(AV_k)$, $y_0 = b$ y $\|\cdot\| = \|\cdot\|_2$ en el Lema 1.1.1, se tiene que existe un único $y_k \in \mathbb{R}^k$ tal que

$$\|b - AV_k y_k\|_2 = \min_{z\in Rango(AV_k)} \|b-z\|_2 \tag{1.12}$$

$$= \min_{y\in\mathbb{R}^k} \|b - AV_k y_k\|_2 \tag{1.13}$$

Así, $x_k = V_k y_k$ es el único elemento de $\mathcal{K}_k$ que satisface (1.11).

1.2.2 Implementación

En primer lugar obsérvese que en el Algoritmo de Arnoldi se construye una matriz *Hessenberg Superior* H_k de orden $(k+1)\times k$ con la propiedad

$$Av_k = h_{1k}v_1 + ... + h_{kk}v_k + h_{k+1,k}v_{k+1}$$

es decir H_k esta caracterizada por la identidad

$$AV_k = V_{k+1}H_k. \tag{1.14}$$

Además, como V_{k+1} es ortogonal, es isometría y tal que $V_{k+1}e_1 = \frac{b}{\beta}$ con $\beta = \|b\|_2$, se sigue entonces que para $y \in \mathbb{R}^k$ se tiene

$$\|b - AV_k y\|_2 = \|V_{k+1}(\beta e_1 - H_k y)\|_2 = \|\beta e_1 - H_k y\|_2$$

donde $e_1 = (1, 0, ..., 0)^T \in \mathbb{R}^{k+1}$. Entonces,

$$\|b - Ax_k\|_2 = \min_{x\in\mathcal{K}_k} \|b - Ax\|_2 \tag{1.15}$$

$$= \min_{y\in\mathbb{R}^k} \|b - AV_k y\|_2 \tag{1.16}$$

$$= \min_{y\in\mathbb{R}^k} \|\beta e_1 - H_k y\|_2 \tag{1.17}$$

$$= \|\beta e_1 - H_k y_k\|_2 . \tag{1.18}$$

Lo que hemos hecho es reducir nuestro problema de minimización original a uno en el que la matriz de coeficientes tiene estructura Hessenberg lo cual fácilita el proceso de cálculo de y_k. Por esta razón se usa (1.15) en lugar de (1.12) en la implementación del algoritmo de GMRES. Podemos entonces plantear ya un primer algoritmo para GMRES el cual lleva a cabo el proceso de simplificación hasta ahora descrito.

Algoritmo GMRES Básico

$v_1 = \frac{b}{\|b\|}$
for $k = 1 : M$,
    $w = Av_k$
    for $j = 1 : k$,
        $h_{jk} = \langle v_j, w \rangle$
        $w = w - h_{jk} v_j$
    end
    $h_{k+1,k} = \|w\|$
    $v_{k+1} = \frac{w}{h_{k+1,k}}$
    {*Resolver para* y_k, *por mínimos cuadrados*,
     $H_k y_k = \beta e_1$}
    $x_k = V_k y_k$
end

Obviamente este algoritmo no es del todo completo al no incluir una forma explícita de encontrar los x_k por no proporcionar un procedimiento concreto para resolver el problema de mínimos cuadrados que se plantea. Se desarrolla enseguida un procedimiento presentado por Saad en [18] para lo obtención de un algoritmo explícito a partir de la factorización QR de H_k.

Factorización QR y Mínimos Cuadrados

La factorización QR permite escribir una matriz dada A de orden $n \times m$ como el producto de una matriz ortogonal Q de orden $n \times n$ por una matriz triangular superior R de orden $n \times m$. Si se impone a R la condición de que su diagonal tenga entradas no negativas, esta descomposición es única. Para la obtención de estas matrices nos basamos en la igualdad $Q^T A = R$ como nuestro objetivo y hallamos Q^T como la multiplicación sucesiva de matrices unitarias diseñadas para ir sucesivamente anulando las componentes debajo de la diagonal. Con ello Q será el producto de matrices de la forma

$$\begin{bmatrix} I_k & 0 \\ 0 & I_{n-k} - vv^T \end{bmatrix}.$$

Estas matrices suelen llamarse reflexiones o transformaciones de Householder, ver ([11] KINCAID D. AND W. CHENEY 1994) para detalles.

Nos concentramos ahora en la factorización QR de H_k**.** Empezamos por notar que dado que la matriz triangular superior a obtener debe tener el mismo rango de H_k y por tanto debe entonces tener su última fila nula, podemos entonces escribir la factorización QR de $H_k$como

$$Q_k^T H_k = \begin{bmatrix} R_k \\ 0^T \end{bmatrix}$$

donde R_k es cuadrada de orden y rango k y triangular superior. Denotamos sus componentes así:

$$R_k = \begin{bmatrix} r_{11} & r_{12} & r_{13} & \cdots & r_{1k} \\ & r_{22} & r_{23} & \cdots & r_{2k} \\ & & \ddots & \ddots & \\ & & & r_{k-1,k-1} & r_{k-1,k} \\ & & & & r_{kk} \end{bmatrix}.$$

Una forma muy eficiente de llevar a cabo esta factorización en el caso que nos interesa, es multiplicar de manera sucesiva a H_k por las matrices

$$P_j = \begin{bmatrix} I_{j-1} & 0 & & 0 \\ 0 & c_j & s_j & 0 \\ & s_j & -c_j & \\ 0 & 0 & & I_{k-j} \end{bmatrix},$$

donde la parte no trivial de P_j hace el siguiente trabajo

$$\begin{bmatrix} c_j & s_j \\ s_j & -c_j \end{bmatrix} \begin{bmatrix} \bar{r}_{jj} & \bar{r}_{j,j+1} \\ h_{j+1,j} & h_{j+1,j+1} \end{bmatrix} = \begin{bmatrix} r_{jj} & r_{j,j+1} \\ 0 & \bar{r}_{j+1,j+1} \end{bmatrix}$$

y además observa las propiedades de ortogonalidad requeridas. Entonces,

$$P_k P_{k-1} \cdots P_2 P_1 H_k = \begin{bmatrix} R_k \\ 0^T \end{bmatrix}.$$

Así $Q_k^T = P_k P_{k-1} \cdots P_2 P_1$ es la transpuesta de la matriz requerida en la factorización QR.

Consideramos ahora sí el sistema a resolver

$$H_k y_k = \beta e_1.$$

Multiplicando a izquierda por Q_k^T tenemos

$$\begin{aligned} Q_{kk}^T H y_k &= Q_{kk}^T \beta e_1, \\ Q_k^T H y_k &= Q_k^T \beta e_1, \\ \begin{bmatrix} R_k \\ 0^T \end{bmatrix} y_k &= \begin{bmatrix} g_k \\ \bar{\varsigma}_{k+1} \end{bmatrix}. \end{aligned}$$

Donde $g_k \in \mathbb{R}^k$ y

$$\begin{bmatrix} g_k \\ \bar{\varsigma}_{k+1} \end{bmatrix} = Q_k^T \beta e_1.$$

Entonces dado que la última fila de $\begin{bmatrix} R_k \\ 0^T \end{bmatrix}$ es nula, resolver por mínimos cuadrados

$$H_k y_k = \beta e_1.$$

es equivalente a resolver

$$R_k y_k = g_k.$$

Una vez hallado y_k se sigue el procedimiento de nuestro primer algoritmo para encontrar el x_k correspondiente. Sin embargo haremos una observación adicional que permite evitar tener que calcular x_k desde y_k hasta tanto se tenga garantizada la convergencia. Resumimos todos estos resultados en el siguiente teorema.

Teorema 1.2.3 *Si V_k y H_k son las matrices generadas luego de k pasos de GMRES y*

$$H_k = Q_k \begin{bmatrix} R_k \\ 0^T \end{bmatrix}$$

es la factorización QR de H_k, entonces

- *El rango de AV_k es igual al rango de R_k. En particular si $r_{kk} = 0$, entonces A es singular.*
- *El vector y_k que minimiza $\|H_k y_k - \beta e_1\|_2$ está dado por $y_k = R_k^{-1} g_k$.*
- *El vector residual a los k pasos satisface*

$$\begin{aligned} b - Ax_k &= V_{k+1}(\beta e_1 - H_k y_k) \\ &= V_{k+1} Q_k (\bar{\varsigma}_{k+1} e_{k+1}). \end{aligned}$$

Por tanto,

$$\|b - Ax_k\|_2 = |\bar{\varsigma}_{k+1}|.$$

Prueba:
En efecto, por (1.14)

$$\begin{aligned} AV_k &= V_{k+1} H_k \\ &= V_{k+1} Q_k Q_k^T H_k \\ &= V_{k+1} Q_k R_k. \end{aligned}$$

Como $V_{k+1}Q_k$ es ortogonal entonces AV_k y R_k tienen el mismo rango.

Obsérvese ahora que,

$$\begin{aligned}(H_k y_k - \beta e_1) &= \left(Q_k \begin{bmatrix} R_k \\ 0^T \end{bmatrix} y_k - Q_k \begin{bmatrix} g_k \\ \bar{\varsigma}_{k+1} \end{bmatrix} \right) \\ &= Q_k \left(\begin{bmatrix} R_k \\ 0^T \end{bmatrix} y_k - \begin{bmatrix} g_k \\ \bar{\varsigma}_{k+1} \end{bmatrix} \right).\end{aligned} \tag{1.19}$$

Por tanto, como Q_k es ortogonal, es isometría y se tiene

$$\|H_k y_k - \beta e_1\|_2 = \left\| \begin{bmatrix} R_k \\ 0^T \end{bmatrix} y_k - \begin{bmatrix} g_k \\ \bar{\varsigma}_{k+1} \end{bmatrix} \right\|_2$$

Así, la solución por mínimos cuadrados de $H_k y_k = \beta e_1$ debe ser igual a la solución por mínimos cuadrados de

$$\begin{bmatrix} R_k \\ 0^T \end{bmatrix} y_k = \begin{bmatrix} g_k \\ \bar{\varsigma}_{k+1} \end{bmatrix},$$

la cual obviamente es $y_k = R_k^{-1} g_k$.

Por último, notamos que por (1.19)

$$\begin{aligned} b - Ax_k &= V_{k+1}(\beta e_1 - H_k y_k) \\ &= V_{k+1} Q_k \left(\begin{bmatrix} g_k \\ \bar{\varsigma}_{k+1} \end{bmatrix} - \begin{bmatrix} R_k \\ 0^T \end{bmatrix} y_k \right) \\ &= V_{k+1} Q_k \left(\begin{bmatrix} g_k \\ \bar{\varsigma}_{k+1} \end{bmatrix} - \begin{bmatrix} g_k \\ 0 \end{bmatrix} \right) \\ &= V_{k+1} Q_k \left(\bar{\varsigma}_{k+1} e_{k+1} \right). \end{aligned}$$

El último resultado de este teorema es de especial importancia en la medida que permite conocer el residual que generará x_k justo después de efectuado el proceso de factorización QR, evitando así tener que calcular x_k hasta no tener garantizada la convergencia.

Recogemos todas estas estrategias en el siguiente

Algoritmo GMRES basado en QR

$v_1 = \frac{b}{\|b\|}$
$for\ k = 1 : M,$
    $\%Proceso\ de\ Arnoldi$
    $w = Av_k$

$for\ j = 1 : k,$
$\quad h_{jk} = \langle v_j, w \rangle$
$\quad w = w - h_{jk} v_j$
end
$h_{k+1,k} = \|w\|$
$v_{k+1} = \frac{w}{h_{k+1,k}}$
$\%Se\ inicia\ la\ factorización\ QR\ con$
$\%las\ Transformaciones\ de\ Householder$
$for\ j = 1 : k - 1$
$\quad aux = c_j h_{jk} - s_j h_{j+1,k}$
$\quad h_{j+1,k} = s_j h_{jk} + c_j h_{j+1,k}$
$\quad h_{jk} = aux$
end
$u = sqrt(h_{k+1,k}^2 + h_{kk}^2)$
$c_k = h_{kk}/u$
$s_k = -h_{k+1,k}/u;$
$h_{kk} = c_k h_{kk} - s_k h_{k+1,k}$
$h_{k+1,k} = 0$
$g_{k+1} = s_k g_k;$
$g_k = c_k g_k;$
$\%Se\ Chequea\ la\ convergencia$
$rho = abs(g_{k+1});$
$if\ rho <= tolerance$
$\quad y = h(1 : k, 1 : k) \backslash g(1 : k)';$
$\quad x = x + v(:, 1 : k) * y;$
$\quad return$
end
end

Este nuevo algoritmo aunque completo tiene la dificultad de que en su ejecución se requiere mantener almacenada la matriz H, lo cual puede requerir gran capacidad de almacenamiento para sistemas de gran tamaño. Esta dificultad, aunque importante, es fácil de salvar si se escoge un M que sea pequeño respecto del orden de la matriz y se ejecuta este algoritmo de modo iterativo reiniciando con la mejor aproximación de los M pasos dados previamente. Otra observación crucial es que si la matriz A es una matriz con estructura, como por ejemplo matrices Toeplitz, tridiagonales, etcétera, se puede remplazar en el algoritmo el cálculo Av_k por la acción de la matriz A sobre el vector, evitando así tener que almacenar A. Entonces, teniendo en cuenta estas dos estrategias, la de reinicialización y la de utilización de

acciones en lugar de almacenar la matriz de coeficientes, habilitan a GMRES como un eficiente e importante método para la solución de sistemas lineales de gran tamaño con estructura. Por último, una ganancia adicional del uso de GMRES es que el algoritmo no requiere del uso de acciones de la matriz A^T.

1.2.3 Análisis de convergencia

Puesto que A es no simétrica, no se puede contar con una base ortonormal de vectores propios y por lo tanto no se puede obtener una desigualdad análoga a la del teorema (2.3.1). Sin embargo, se da una igualdad análoga a la que aparece en este teorema y se consigna en el siguiente resultado. La prueba es completamente similar y debido a esto no se presenta.

Teorema 1.2.4

$$\|r_k\|_2 = \min_{p\in \mathrm{P}_k} \|p(A)b\|_2 \leq \left(\min_{p\in \mathrm{P}_k} \|p(A)\|_2\right) \|b\|$$

Corolario 1.2.5 *GMRES calcula la solución exacta, a lo más en n iteraciones.*

Prueba. Sea $q(t) = \det(A - tI)$ el polinomio característico de A. Nótese que $q(0) = \det(A) \neq 0$, puesto que A es no singular. Defínase

$$p(t) = \frac{q(t)}{q(0)} \in P_n$$

Obsérvese que $p(A) = 0$ y aplíquese el teorema anterior.

Para el caso en que A es una matriz diagonalizable, y basándonos en el teorema anterior, se puede obtener una cota para la norma residual que depende del número de condicion espectral de la matriz diagonalizante. Este es el objeto del siguiente teorema.

Teorema 1.2.6 *Sea $A = V\Lambda V^{-1}$ una matriz no singular diagonalizable, con Λ diagonal. Para cualquier polinomio $p \in \mathrm{P}_k$, se tiene*

$$\|r_k\|_2 \leq \kappa \left(\max_{\lambda\in\sigma(A)} |p(\lambda)|\right) \|b\|$$

donde $\kappa = \|V\|_2 \left\|V^{-1}\right\|_2$.

Prueba. Basta observar que

$$\begin{aligned}\|p(A)\|_2 &= \left\|Vp(\Lambda)V^{-1}\right\|_2 \\ &\leq \|V\|_2 \ \|p(\Lambda)\|_2 \ \left\|V^{-1}\right\|_2 = \kappa\left(\max_{\lambda\in\sigma(A)}|p(\lambda)|\right)\end{aligned}$$

y utilizar el teorema (3.3.1). ■

Corolario 1.2.7 *Si A es una matriz no singular diagonalizable, con exactamente k valores propios distintos, entonces GMRES calcula la solución exacta a lo más en k iteraciones.*

1.3 LSQR

1.3.1 Problema

Nuevamente se pretende dar solución al problema

$$Ax = b. \tag{1.20}$$

Aquí A es rala (o esparcida), de rango columna completo, y de orden $n \times m$ con n y m grandes y b es un vector de $\mathbb{R}^n$. El problema se resuelve en el sentido de mínimos cuadrados, es decir, remplazamos el problema (1.20) por el problema: Hallar $x \in \mathbb{R}^m$ tal que

$$\min\left\{\|Ay - b\|_2 / y \in R^m\right\} = \|Ax - b\|_2 \tag{1.21}$$

Presentamos el Algoritmo LSQR ([17] PAIGE, C. AND M. SAUNDERS 1982) como un procedimiento para resolver (1.21) que requiere de poco esfuerzo computacional y ofrece resultados de calidad.

1.3.2 Introducción

El algoritmo LSQR es un procedimiento para la solución de sistemas de ecuaciones lineales desarrollado por Paige y Saunders en ([17] PAIGE, C. AND M. SAUNDERS 1982) y que se basa en el proceso de bidiagonalización de una matriz. La bidiagonalización de una matriz utilizada en LSQR fue desarrollada por Golub y Kahan en ([5] GOLUB, G. AND W. KAHAN 1965) para el cálculo de valores singulares. A su vez la bidiagonalización se desarrollo originalmente desde el proceso de Tridiagonalización desarrollado por Lanczos ([12] LANCZOS, C 1950) en la generación de métodos iterativos para el cálculo de valores propios. Por esta razón, la bidiagonalización

es atribuida a Lanczos. Es importante resaltar que estas factorizaciones se originaron resolviendo problemas de valores propios y no problemas de mínimos cuadrados.

Luego, en ([16] PAIGE, C. AND M. SAUNDERS 1975) los autores establecen una relación entre la solución de sistemas por métodos clásicos como Gradiente Conjugado y MINRES y la solución de los sistemas inducidos por la tridiagonalización. Para una referencia más reciente ver ([6] GOLUB, G. AND C. VAN LOAN 1996). Más tarde, como mencionamos arriba, en ([17] PAIGE, C. AND M. SAUNDERS 1982) se presenta el método LSQR y se muestra su equivalencia con CGLS (Gradiente Conjugado para Mínimos Cuadrados, sobre la ecuación normal). Por último, es bueno anotar que en ([1] BENBOW, S 1997), Benbow extiende el algoritmo LSQR a fin de generar un LSQR precondicionado.

En seguida presentamos la Bidiagonalización de Lanczos, se deduce el algoritmo y se establece la relación con problemas de mínimos cuadrados y la relación con subespacios de Krylov. En la Sección 1.3.4 se obtiene el algoritmo LSQR mostrando primero la relación entre factorización QR y bidiagonales, luego entre QR y mínimos cuadrados y por último la deducción del algoritmo.

1.3.3 Bidiagonalización de Lanczos

La bidiagonalización de Lanczos de una matriz A de rango columna completo, y de orden $n \times m$ proporciona matrices U, V y B tales que U es $n \times n, V$ es $m \times n, B$ es $n \times n$, $U^T U = I_n$, $V^T V = I_n$, $U^T AV = B$ y B es de la forma

$$B = \begin{bmatrix} \alpha_1 & & & & & \\ \beta_2 & \alpha_2 & & & & \\ & \beta_3 & \alpha_3 & & & \\ & & \ddots & \ddots & & \\ & & & \beta_{n-1} & \alpha_{n-1} & \\ & & & & \beta_n & \alpha_n \end{bmatrix}.$$

Por tanto se satisfacen se satisfacen las dos condiciones

$$AV = UB,$$
$$A^T U = VB^T.$$

La primera se escribe como:

$$
\begin{aligned}
& A \begin{bmatrix} v_1 & v_2 & \cdots & v_n \end{bmatrix} \\
& = \begin{bmatrix} u_1 & u_2 & \cdots & u_n \end{bmatrix}
\begin{bmatrix}
\alpha_1 & & & & \\
\beta_2 & \alpha_2 & & & \\
& \beta_3 & \ddots & & \\
& & \ddots & \alpha_{n-1} & \\
& & & \beta_n & \alpha_n
\end{bmatrix}
\end{aligned}
$$

y conduce a las relaciones

$$
\begin{aligned}
& Av_i = \alpha_i u_i + \beta_{i+1} u_{i+1},\ 1 \leq i \leq n-1, \\
& Av_n = \alpha_n u_n.
\end{aligned} \tag{1.22}
$$

De otra parte, la condición $A^T U = V B^T$ se escribe como:

$$
\begin{aligned}
& A^T \begin{bmatrix} u_1 & u_2 & \cdots & u_n \end{bmatrix} \\
& = \begin{bmatrix} v_1 & v_2 & \cdots & v_n \end{bmatrix}
\begin{bmatrix}
\alpha_1 & \beta_2 & & & \\
& \alpha_2 & \beta_3 & & \\
& & \ddots & \ddots & \\
& & & \alpha_{n-1} & \beta_n \\
& & & & \alpha_n
\end{bmatrix}
\end{aligned}
$$

lo que conduce a las relaciones:

$$
\begin{aligned}
& A^T u_1 = \alpha_1 v_1, \\
& A^T u_i = \alpha_i v_i + \beta_i v_{i-1},\ 2 \leq i \leq n.
\end{aligned} \tag{1.23}
$$

Ahora las relaciones establecidas en (1.22) y (1.23) permiten notar que conocido un vector incial para determinar el u_1 se encuentra fácil el v_1 y para el resto de los vectores se puede hacer uso de las relaciones

$$
\begin{aligned}
& \beta_{j+1} u_{j+1} = Av_j - \alpha_j u_j,\ j = 1, 2, ... \\
& \alpha_{j+1} v_{j+1} = A^T u_j - \beta_{j+1} v_j,\ j = 1, 2, ...
\end{aligned} \tag{1.24}
$$

Una vez efectuado el proceso l veces se cuenta con matrices

$$
\begin{aligned}
U_{l+1} &= \begin{bmatrix} u_1 & u_2 & \cdots & u_{l+1} \end{bmatrix}, \\
V_l &= \begin{bmatrix} v_1 & v_2 & \cdots & v_l \end{bmatrix},
\end{aligned}
$$

tales que, gracias a las relaciones en (1.24), satisfacen:

$$\begin{aligned} U_{l+1}(\beta_1 e_1) &= b, \\ AV_l &= U_{l+1}B_l, \\ A^T U_{l+1} &= V_l B_l^T + \alpha_{l+1} v_{l+1} e_{l+1}^T. \end{aligned} \tag{1.25}$$

Estas relaciones concuerdan con nuestro objetivo de bidiagonalización y junto con (1.24) nos habilitan para plantear el siguiente algoritmo.

Algoritmo

Input b
$\beta_1 u_1 = b$
$\alpha_1 v_1 = A^T u_1$
For $j = 1 : l$
$\beta_{j+1} u_{j+1} = A v_j - \alpha_j u_j$
$\alpha_{j+1} v_{j+1} = A^T u_{j+1} - \beta_{j+1} v_j$
end

donde los $\alpha_j, \beta_j > 0$ y se escogen de modo tal que $\|u_j\|_2 = \|v_j\|_2 = 1$.

Bidiagonalización y Mínimos Cuadrados

Retomamos ahora nuestro problema inicial (1.20),

$$Ax = b.$$

Supongamos que se han llevado a cabo l pasos del algoritmo de bidiagonalización tomando como vector de entrada el lado derecho b. Nos concentramos en aproximar la solución mediante combinaciones lineales de las columnas de V_l. Nos planteamos entonces el problema de: Hallar $t_l \in \mathbb{R}^l$ tal que $x_l = V_l t_l$ satisfaga

$$\min\left\{\|AV_l t - b\|_2 : t \in \mathbb{R}^l\right\} = \|Ax_l - b\|_2 . \tag{1.26}$$

Pero de (1.25),

$$\begin{aligned} \|AV_l t - b\| &= \|U_{l+1}B_l t - b\| \\ &= \|U_{l+1}B_l t - \beta_1 u_1\| \\ &= \|U_{l+1}\left(B_l t - \beta_1 e_1\right)\| \\ &= \|B_l t - \beta_1 e_1\| . \end{aligned}$$

La última igualdad se sigue de la ortogonalidad de las columnas de U_{l+1}. Entonces el problema (1.26) se reduce a: Hallar $t_l \in \mathbb{R}^l$ tal que

$$\min\left\{\|B_l t - \beta_1 e_1\| : t \in \mathbb{R}^l\right\} = \|B_l t_l - \beta_1 e_1\|. \tag{1.27}$$

En este punto nos encontramos en una situación completamente similar a la que nos encontramos en el caso GMRES luego de aplicar el algoritmo de Arnoldi, razón por la cual nos sentimos fuertemente tentados a efectuar una factorización QR a B_l y obtener asi un algoritmo explícito. En efecto esto es lo que haremos para obtener el algoritmo LSQR, pero lo aplazamos un poco para establecer otra importante relación de LSQR, pero esta vez con CG.

Relación con Espacios de Krylov

La razón de buscar aproximar la solución del problema (1.20) con combinaciones lineales de columnas de V_l la indica el siguiente teorema que indica la naturaleza de ese espacio.

Teorema 1.3.1 *Si* $U_{l+1} = [u_1\ u_2\ \ldots\ u_{l+1}]$, $V_l = [v_1\ v_2\ \ldots\ v_l]$ *son las matrices obtenidas en el proceso de bidiagonalización de Lanczos, entonces*
$u_j \in K_j(AA^T, b)$,
$v_j \in K_j(A^TA, A^Tb)$.

Prueba: Para $j = 1$, el resultado es inmediato de las igualdades

$$\beta_1 u_1 = b,$$
$$\alpha_1 v_1 = A^T u_1.$$

Supongamos que se cumple para j. Veamos que se cumple para $j+1$.

Como se cumple para j, entonces existen escalares $\varphi_i, \lambda_i \in \mathbb{R}$, $i = 1, 2, ..., j$ tales que

$$u_j = \sum_{i=1}^{j} \lambda_i (AA^T)^{i-1} b,$$
$$v_j = \sum_{i=1}^{j} \varphi_i (A^TA)^{i-1} A^T b.$$

Entonces del algoritmo de bidiagonalización,

$$\begin{aligned}
&\beta_{j+1}u_{j+1}\\
&= Av_j - \alpha_j u_j\\
&= A\sum_{i=1}^{j}\varphi_i(A^TA)^{i-1}A^Tb - \alpha_j\sum_{i=1}^{j}\lambda_i(AA^T)^{i-1}b\\
&= \sum_{i=1}^{j}\varphi_i(AA^T)^{i}b - \alpha_j\sum_{i=1}^{j}\lambda_i(AA^T)^{i-1}b\\
&= -\alpha_j\lambda_1 b + \sum_{i=2}^{j}(\varphi_{i-1}+\lambda_i)(AA^T)^{i-1}b + \varphi_j(AA^T)^j b.
\end{aligned}$$

De donde se sigue que $u_{j+1} \in K_{j+1}(AA^T, b)$, podemos entonces garantizar la existencia de escalares σ_i tales que

$$u_{j+1} = \sum_{i=1}^{j+1}\sigma_i(AA^T)^{i-1}b.$$

Con lo que a partir del algoritmo obtenemos que:

$$\begin{aligned}
&\alpha_{j+1}v_{j+1}\\
&= A^Tu_{j+1} - \beta_{j+1}v_j\\
&= A^T\sum_{i=1}^{j+1}\sigma_i(AA^T)^{i-1}b - \beta_{j+1}\sum_{i=1}^{j}\varphi_i(A^TA)^{i-1}A^Tb\\
&= \sum_{i=1}^{j+1}\sigma_i(A^TA)^{i-1}A^Tb - \beta_{j+1}\sum_{i=1}^{j}\varphi_i(A^TA)^{i-1}A^Tb\\
&= \left(\sum_{i=1}^{j}(\sigma_i - \beta_{j+1}\varphi_i)(A^TA)^{i-1}A^Tb\right) + \sigma_{j+1}(A^TA)^jA^Tb.
\end{aligned}$$

Así, $v_{j+1} \in K_{j+1}(A^TA, A^Tb)$.
En el objetivo del método LSQR es:
Hallar $x_l \in K_l(A^TA, A^Tb)$ tal que

$$\min\left\{\|Ax-b\| / x \in K_l(A^TA, A^Tb)\right\} = \|Ax_l - b\|. \tag{1.28}$$

En tanto que el objetivo del método de gradiente conjugado (CG), aplicado al sistema $\left(A^TA\right)x = \left(A^Tb\right)$, es:
Hallar $x_l \in K_l(A^TA, A^Tb)$ tal que

$$\min\left\{\left\|A^TAx - A^Tb\right\|_{(A^TA)^{-1}} : x \in K_l(A^TA, A^Tb)\right\} \tag{1.29}$$

$$= \left\|A^TAx_l - A^Tb\right\|_{(A^TA)^{-1}}. \tag{1.30}$$

Donde $\|z\|^2_{(A^TA)^{-1}} = z^T(A^TA)^{-1}z$.

Sea ahora $\hat{x}$ la solución óptima por mínimos cuadrados del problema (1.20), entonces $A^TA\hat{x} = A^Tb$.

Luego

$$\begin{aligned}
&\left\|A^TAx - A^Tb\right\|_{(A^TA)^{-1}} \\
&= \left(A^TAx - A^Tb\right)^T (A^TA)^{-1} \left(A^TAx - A^Tb\right) \\
&= \left(A^TAx - A^TA\hat{x}\right)^T (A^TA)^{-1} \left(A^TAx - A^TA\hat{x}\right) \\
&= (x - \hat{x})^T (A^TA) (x - \hat{x}) .
\end{aligned}$$

Nótese que $\hat{b} = A\hat{x}$ es la proyección ortogonal de b sobre el espacio generado por las columnas de A. Entonces,

$$\begin{aligned}
\left\|A^TAx - A^Tb\right\|^2_{(A^TA)^{-1}} &= (x - \hat{x})^T (A^TA) (x - \hat{x}) \\
&= \left(Ax - \hat{b}\right)^T \left(Ax - \hat{b}\right) \\
&= \left\|Ax - \hat{b}\right\|^2 .
\end{aligned} \tag{1.31}$$

Pero,

$$\begin{aligned}
\|Ax - b\|^2 &= \left\|Ax - \hat{b} + \hat{b} - b\right\|^2 \\
&= \left\|Ax - \hat{b}\right\|^2 + \left\|\hat{b} - b\right\|^2 .
\end{aligned} \tag{1.32}$$

La ultima igualdad se debe a que $(Ax - \hat{b})$ está en el rango de A y $(\hat{b} - b)$ está en su complemento ortogonal, es decir en el espacio nulo de A^T.

Con esto la minimización indicada en (1.28) es la misma indicada en (1.29).

Por esta razon puede afirmarse que LSQR, bajo aritmética exacta, es equivalente a CGLS.

1.3.4 El Algoritmo LSQR

Una vez establecido con claridad el objetivo de LSQR tenemos por delante la tarea de diseñar un algoritmo eficiente para calcular x_l . Aunque la respuesta inmediata en este punto es resolver $B_l t_l = \beta_1 e_1$ y usar entonces la relación $x_l = V_l t_l$. Esta resulta ser nada atractiva ya que obligaría a mantener en memoria los vectores v_j, lo cual definitivamente no es nada económico.

Para resolver esto utilizaremos la estrategia empleada por Paige y Saunders en ([17] PAIGE, C. AND M. SAUNDERS 1982), que saca provecho del proceso de factorización QR para matrices bidiagonales que requiere tan

sólo de una actualización muy sencilla a la hora de incrementar el orden. Con ello Paige y Saunders consiguieron un algoritmo con poco consumo de memoria y usando sólo una multiplicación matriz-vector con A y una con A^T, en cada iteración. Esto último facilita su implementación en problemas donde A es grande y con estructura.

Estudiaremos primero algunas propiedades útiles de la factorización QR de una matriz bidiagonal, para luego pasar a la deducción del algoritmo.

Factorización QR y Bidiagonales

Usaremos nuevamente la factorización QR como herramienta eficiente para resolver el problema de mínimos cuadrados al que hemos reducido nuestro problema inicial gracias a la bidiagonalización. Empezamos por observar que la factorización QR de la bidiagonal ha de seguirse de modo muy similar a la efectuada en el caso de la matriz de Hessenberg obtenida en el caso GMRES, pero como veremos tendremos la ventaja adicional de que la matriz R que se obtiene será también bidiagonal simplificando aún más el problema de minimización.

La implementación de esta factorización se basa nuevamente en transformaciones Householder aplicadas sucesivamente para obtener Q^T. Nos concentramos ahora en los detalles de la factorización QR de B_l. Empezamos por notar que dado que la matriz triangular inferior a obtener debe tener el mismo rango de B_l y por tanto debe entonces tener su última fila nula, podemos entonces escribir la factorización QR de B_l como

$$Q_l^T B_l = \begin{bmatrix} R_l \\ 0^T \end{bmatrix}$$

donde R_l es cuadrada de orden y rango l y bidiagonal superior. Denotamos sus componentes así:

$$R_l = \begin{bmatrix} \rho_1 & \theta_2 & & & \\ & \rho_2 & \theta_3 & & \\ & & \ddots & \ddots & \\ & & & \rho_{l-1} & \theta_l \\ & & & & \rho_l \end{bmatrix}.$$

A fin de efectuar de modo eficiente el proceso, en el caso que nos interesa, multiplicamos de manera sucesiva a B_{l+1} por las matrices

$$P_j = \begin{bmatrix} I_{j-1} & 0 & & 0 \\ 0 & \begin{matrix} c_j & s_j \\ s_j & -c_j \end{matrix} & & 0 \\ 0 & 0 & & I_{l+1-j} \end{bmatrix},$$

donde, en este caso, la parte no trivial de P_j hace el siguiente trabajo

$$\begin{bmatrix} c_j & s_j \\ s_j & -c_j \end{bmatrix} \begin{bmatrix} \bar{\rho}_j & 0 \\ \beta_{j+1} & \alpha_{j+1} \end{bmatrix} = \begin{bmatrix} \rho_j & \theta_{j+1} \\ 0 & \bar{\rho}_{j+1} \end{bmatrix}$$

y conserva las propiedades de ortogonalidad requeridas. Entonces,

$$P_l P_{l-1} \cdots P_2 P_1 B_{l+1} = \begin{bmatrix} R_l & 0 \\ & \theta_{l+1} \\ 0^T & \bar{\rho}_{l+1} \\ 0^T & \beta_{l+2} \end{bmatrix}.$$

Sea ahora $Q^T = P_l P_{l-1} \cdots P_2 P_1$, entonces si usamos la notación de Matlab, podemos afirmar que $Q_l^T = (Q(1 : l+1, 1 : l+1))^T$.

Factorización QR y Mínimos Cuadrados

Retomamos ahora el problema (1.27) de resolver por mínimos cuadrados la ecuación $B_l t = \beta_1 e_1$. Como Q_l es ortogonal

$$\begin{aligned} \|B_l t - \beta_1 e_1\| &= \|Q_l B_l t - Q_l(\beta_1 e_1)\| \\ &= \left\| \begin{bmatrix} R_l \\ 0^T \end{bmatrix} t - \begin{bmatrix} f_l \\ \bar{\varphi}_{l+1} \end{bmatrix} \right\|, \end{aligned}$$

donde $Q_l(\beta_1 e_1) = \begin{bmatrix} f_l \\ \bar{\varphi}_{l+1} \end{bmatrix}$ y podemos denotar $f_l = \begin{bmatrix} \varphi_1 & \varphi_2 & \cdots & \varphi_l \end{bmatrix}^T$.

Claramente la mejor solución al sistema está dada por $t_l = R_l^{-1} f_l$. Por tanto la mejor solución a nuestro sistema (1.20) es

$$x_l = V_l t_l = (V_l R_l^{-1}) f_l.$$

Sea ahora $D_l = (V_l R_l^{-1})$ y digamos $D_l = [d_1 d_2 ... d_l]$, entonces

$$x_l = \sum_{j=1}^{l} \varphi_j d_j$$

Pero $D_l R_l = V_l$, entonces si convenimos $d_0 = 0$, encontramos que

$$d_j = \frac{1}{\rho_j}(v_j - \theta_j d_{j-1}).$$

Además, como

$$R_{l+1} = \begin{bmatrix} R_l & 0 \\ & \theta_{l+1} \\ 0^T & \rho_{l+1} \end{bmatrix},$$

entonces $D_{l+1} = [D_l, d_{l+1}]$ con

$$d_{l+1} = \frac{1}{\rho_{l+1}}(v_{l+1} - \theta_{l+1} d_l).$$

Así, $x_{l+1} = x_l + \varphi_{l+1} d_{l+1}$. Lo que nos indica que la siguiente aproximación puede obtenerse de la anterior sólo mediante una actualización que no requiere mucho esfuerzo si recordamos que se debe cumplir que:

$$\begin{bmatrix} c_{l+1} & s_{l+1} \\ s_{l+1} & -c_{l+1} \end{bmatrix} \begin{bmatrix} \bar{\rho}_{l+1} & 0 \\ \beta_{l+2} & \alpha_{l+2} \end{bmatrix} = \begin{bmatrix} \rho_{l+1} & \theta_{l+2} \\ 0 & \bar{\rho}_{l+2} \end{bmatrix},$$
$$\begin{bmatrix} c_{l+1} & s_{l+1} \\ s_{l+1} & -c_{l+1} \end{bmatrix} \begin{bmatrix} \bar{\varphi}_{l+1} \\ 0 \end{bmatrix} = \begin{bmatrix} \varphi_{l+1} \\ \bar{\varphi}_{l+2} \end{bmatrix}.$$

Esto nos deja completamente habilitados para proponer un algoritmo para LSQR.

Construcción del Algoritmo LSQR

Como vimos arriba para obtener x_{l+1} a partir de x_l se requiere calcular d_{l+1} y φ_{l+1}. Pero d_{l+1} requiere v_{l+1}, por tanto se hace preciso que se ejecute un nuevo paso de bidiagonalización de Lanczos. Ahora, d_{l+1} requiere también ρ_{l+1}, así es preciso llevar a cabo un nuevo paso de factorización QR y por último es necesario usar los coeficientes de la ortogonalización para encontrar φ_{l+1}.

Presentamos dos algoritmos, el primero conservando los nombres y subindices de la teoría y está basado en el obtenido por Benbow en ([1] BENBOW, S. 1997).

Algoritmo LSQR

Input A, b

$$
\begin{aligned}
y_0 &= 0,\\
\beta_1 u_1 &= b,\\
\alpha_1 v_1 &= A^T u_1,\\
d_1 &= v_1,\\
\bar{\varphi}_1 &= \beta_1,\\
\bar{\rho}_1 &= \alpha_1,
\end{aligned}
$$

For $j = 1 : M$

%Paso de Bidiagonalización.

$$
\begin{aligned}
\beta_{j+1} u_{j+1} &= A v_j - \alpha_j u_j\\
\alpha_{j+1} v_{j+1} &= A^T u_{j+1} - \beta_{j+1} v_j
\end{aligned}
$$

%Paso de transformación ortogonal

$$
\begin{aligned}
\rho_j &= \left(\bar{\rho}_j^2 + \beta_{j+1}^2\right)^{1/2}\\
c_j &= \bar{\rho}_j / \rho_j\\
s_j &= \beta_{j+1} / \rho_j\\
\theta_{j+1} &= s_j \alpha_{j+1}\\
\bar{\rho}_{j+1} &= -c_j \alpha_{j+1}\\
\varphi_j &= c_j \bar{\varphi}_j\\
\bar{\varphi}_{j+1} &= s_j \bar{\varphi}_j
\end{aligned}
$$

%Paso de actualización

$$
\begin{aligned}
y_j &= y_{j-1} + (\varphi_j / \rho_j)\, d_j\\
d_{j+1} &= v_{j+1} - (\theta_{j+1} / \rho_j)\, d_j
\end{aligned}
$$

end

Es claro que dado que solo se requiere efectuar algunas multiplicaciones por A y A^T, no necesariamente hay que cargar A a memoria, siempre y cuando se disponga de una subrutina que efectúe este trabajo.

Capítulo 2

Precondicionamiento

2.1 Introducción

Queremos resolver el sistema lineal $Ax = b$, con A matriz no singular real. En software comercial, no es raro que para este problema se prefieran los métodos directos a los iterativos. El motivo principal es que los primeros se conocen mejor y que los segundos tienen mayor probabilidad de dar resultados desalentadores por demorados o por incorrectos. En este capítulo se reportan estudios recientes que buscan acabar con ambas debilidades. Se trata de métodos de precondicionamiento, que no solo buscan acelerar los métodos iterativos sino también hacerlos más confiables. Para que estos métodos sean verdaderamente útiles, es conveniente que la matriz A sea *rala*, es decir, que la mayor parte de sus elementos sean cero. Es muy conveniente también que la matriz sea estructurada, por ejemplo, una matriz en la que los elementos no nulos están situados únicamente en unas cuantas diagonales, llamadas bandas. De esta manera, no es necesario almacenar la matriz en memoria.

La idea del precondicionamiento es la reducción del número de iteraciones requerido para la convergencia, transformando el sistema original $Ax = b$ en un sistema $\widetilde{A}x = \widetilde{b}$, de tal forma que se satisfagan las siguientes propiedades:

- Resolver $\widetilde{A}x = \widetilde{b}$ no debe incrementar considerablemente el número de operaciones que se requieren para resolver $Ax = b$.
- $Ax = b$ y $\widetilde{A}x = \widetilde{b}$ tienen la misma solución, es decir, $\widetilde{A}^{-1}\widetilde{b} = A^{-1}b$.

La matriz $\widetilde{A}$ y el vector $\widetilde{b}$ se consiguen por premultiplicación o posmultiplicación por matrices llamadas *precondicionadores* que deben ser fácilmente

invertibles. Por medio de un precondicionador se transforma el sistema $Ax = b$ en otro equivalente con condiciones espectrales más favorables y se reduce el número de iteraciones requeridas para la convergencia, sin incrementar significativamente la cantidad de cálculos por iteración. Cuando esto se hace, hay que poner en una balanza los costos y los beneficios; es decir, el costo de construir y aplicar un precondicionador versus la ganancia en rapidez de convergencia. Ciertos precondicionadores necesitan una pequeña fase de construcción pero otros pueden necesitar un trabajo sustancial. En el segundo caso la ganancia puede estar en el uso repetido del mismo precondicionador en múltiples sistemas lineales.

Hay varias opciones para construir el nuevo sistema. Por ejemplo, tomar una matriz no singular M y formar el sistema

$$M^{-1}Ax = M^{-1}b \tag{2.1}$$

o la que se usa mucho más, tomar $M = M_1M_2$, con M_1 y M_2 matrices no singulares y formar el sistema

$$M_1^{-1}AM_2^{-1}\left(M_2x\right) = M_1^{-1}b.$$

A las matrices M_1 y M_2 se les llama precondicionadores a izquierda y derecha respectivamente. En este caso, el esquema para precondicionar un método iterativo es como sigue:

1. Hacer $b \leftarrow M_1^{-1}b$.
2. Aplicar el método iterativo sin precondicionamiento al sistema

$$M_1^{-1}AM_2^{-1}y = b.$$

3. Calcular $x = M_2^{-1}y$.

Nótese que si A es simétrica y definida positiva (spd) y además, $M_1 = M_2^T$, entonces la matriz transformada también es spd. Por tanto esta manera de hacer precondicionamiento es preferida a simplemente multiplicar por una matriz como en (2.1). Al fin y al cabo, si A y M son spd, nada se puede decir de $M^{-1}A$. Afortunadamente, aunque se trabaje con un precondicionador de la forma $M = M_1M_2$, de todas maneras hay formas de escribir el algoritmo del método iterativo precondicionado de tal forma que la única adición al algoritmo previo sea un paso de la forma

$$\text{Resolver para } z \text{ en } Mz = v.$$

El capítulo está organizado asi: Los métodos iterativos básicos se describen en la sección 2. En la sección 3 se tratan métodos de precondicionamiento

basados en los métodos iterativos básicos y en factorización LU incompleta. En la sección 4, se mencionan otros aspectos sobre precondicionamiento, como el tratamiento de matrices 2×2 por bloques de la forma KKT y unos comentarios acerca de la computación en paralelo.

2.2 Métodos Iterativos Estacionarios

Sea A una matriz cuadrada de orden $n \times n$ sobre $\mathbb{C}$ y sean x, b elementos de $\mathbb{C}^n$ tales que

$$Ax = b.$$

Entonces para cada matriz Q de orden $n \times n$ se tiene que

$$Qx = (Q - A)x + b. \tag{2.2}$$

Ahora, si Q es invertible se tiene también

$$x = (I - Q^{-1}A)x + Q^{-1}b. \tag{2.3}$$

Lo cual indica que x es un punto fijo de la aplicación $F(x) = (I - Q^{-1}A)x + Q^{-1}b$. Además la ecuaciones (2.2) y (2.3) inducen las iteraciones

$$\begin{aligned} Qx^{(k)} &= (Q - A)x^{(k-1)} + b \text{ y} \\ x^{(k)} &= (I - Q^{-1}A)x^{(k-1)} + Q^{-1}b, \end{aligned} \tag{2.4}$$

que a su vez generan una sucesión $\left\{x^{(k)}\right\}$ de aproximaciones a x.

Note que si $x^{(k)} \to z$ entonces

$$z = (I - Q^{-1}A)z + Q^{-1}b.$$

O sea, z sería también un punto fijo de la aplicación F.

Nótese además que

$$\begin{aligned} \left\|x^{(k)} - x\right\| &= \left\|(I - Q^{-1}A)x^{(k-1)} + Q^{-1}b \right. \\ &\quad \left. - \left((I - Q^{-1}A)x + Q^{-1}b\right)\right\| \\ &= \left\|(I - Q^{-1}A)\left(x^{(k-1)} - x\right)\right\| \\ &\leq \left\|I - Q^{-1}A\right\| \left\|x^{(k-1)} - x\right\|. \end{aligned}$$

Y aplicando esta desigualdad de modo iterado encontramos que

$$\left\|x^{(k)} - x\right\| \leq \left\|I - Q^{-1}A\right\|^k \left\|x^{(0)} - x\right\|,$$

donde $x^{(0)}$ es una aproximación inicial dada.

Por tanto, si $\left\|I - Q^{-1}A\right\| < 1$ para alguna norma subordinada, entonces $\left\|I - Q^{-1}A\right\|^k \to 0$ cuando $k \to +\infty$ y con ello $\left\{x^{(k)}\right\}$ converge y lo hace al punto x. Tenemos entonces el siguiente

Teorema 2.2.1 *Si $\left\|I - Q^{-1}A\right\| < 1$ para alguna norma subordinada, entonces la sucesión $\left\{x^{(k)}\right\}$ definida en (2.4) converge a x para cualquier vector inicial $x^{(0)}$.*

Definición 2.2.2 *Todo procedimiento iterativo de la forma (2.4) se conoce como método iterativo estacionario para $Ax = b$. Se dice que un método iterativo estacionario es convergente si la sucesión obtenidad converge a la solución del sistema para todo punto inicial $x^{(0)}$.*

2.2.1 Radio Espectral y Convergencia

Definición 2.2.3 *Dada una matriz A se define su radio espectral, notado $\rho(A)$, por:*

$$\rho(A) = \max\left\{|\lambda| : \det\left(A - \lambda I\right) = 0\right\}.$$

Es decir, $\rho(A)$ es el máximo de los módulos de los valores propios de A.

A fin de establecer una relación directa entre el recién definido radio espectral y la convergencia de métodos iterativos estacionarios presentamos el siguiente

Teorema 2.2.4 *La función radio espectral satisface la ecuación*

$$\rho(A) = \inf_{\|\cdot\|} \|A\|,$$

en la cual el ínfimo es tomado sobre todas las normas matriciales subordinadas.

Antes de probar este resultado probaremos el conocido Lema de Schur que establece que toda matriz es unitariamente semejante a una matriz triangular. Aquí debemos recordar que una matriz A es semejante a una matriz T si existe una matriz invertible P tal que

$$A = PTP^{-1}.$$

Es fácil notar que la semejanza de matrices es una relación de equivalencia, es decir, es reflexiva, simétrica y transitiva. Además, matrices semejantes

tienen el mismo polinomio característico y por ende los mismos valores propios y el mismo radio espectral.

Ahora, una matriz P se dice unitaria si es invertible y su inversa es su hermitiana, la traspuesta conjugada $P^H = [\overline{p_{ji}}]$. Entonces A es unitariamente semejante a T si existe P unitaria tal que

$$A = PTP^H.$$

Probaremos también otro Lema auxiliar que utiliza el Lema de Schur para encontrar una matriz triangular semejante a A y en la que el tamaño de las entradas fuera de la diagonal se puede controlar.

Lema 2.2.5 (de Shur) *Toda matriz cuadrada es unitariamente semejante a una matriz triangular superior.*

Prueba al Lema de Schur

Procedemos por inducción sobre el orden n de la matriz. En todo caso A es una matriz cuadrada de entradas complejas y orden n.

Para $n = 2$. Sea λ un valor propio de A y x_1 un vector propio de A asociado a λ, tal que $x_1^T x_1 = 1$, esto es de norma uno. Escogemos ahora x_2 vector de $\mathbb{C}^2$ tal que $\{x_1, x_2\}$ sea base ortonormal de $\mathbb{C}^2$.

Por tanto, $U = \begin{bmatrix} x_1 & x_2 \end{bmatrix}$ es matriz unitaria tal que

$$\begin{aligned} U^H AU &= U^H \begin{bmatrix} Ax_1 & Ax_2 \end{bmatrix} \\ &= U^H \begin{bmatrix} \lambda x_1 & Ax_2 \end{bmatrix} \\ &= \begin{bmatrix} x_1^H \\ x_2^H \end{bmatrix} \begin{bmatrix} \lambda x_1 & Ax_2 \end{bmatrix} \\ &= \begin{bmatrix} x_1^H \lambda x_1 & x_1^H Ax_2 \\ x_2^H \lambda x_1 & x_2^H Ax_2 \end{bmatrix} \\ &= \begin{bmatrix} \lambda & x_1^H Ax_2 \\ 0 & x_2^H Ax_2 \end{bmatrix}. \end{aligned}$$

Entonces, A es unitariamente semejante a la matriz triangular superior

$$\begin{bmatrix} \lambda & x_1^H Ax_2 \\ 0 & x_2^H Ax_2 \end{bmatrix}.$$

Supongamos ahora que la proposición es cierta para $n \in \mathbb{Z}$ y veamos que se cumple para $n + 1$.

En efecto, sea A cuadrada de orden $n + 1$ y sean además λ un valor propio de A y x_1 un vector propio de A asociado a λ, tal que $x_1^T x_1 = 1$.

Escogemos ahora $a_1, a_2, \cdots, a_n$ en $\mathbb{C}^{n+1}$ tales que $\{x_1, a_1, a_2, \cdots, a_n\}$ sea base ortonormal de $\mathbb{C}^{n+1}$. Entonces, $\tilde{U}_1 = \begin{bmatrix} a_1 & a_2 & \cdots & a_n \end{bmatrix}$ es matriz de orden $(n+1) \times n$ tal que $U_1 = \begin{bmatrix} x_1 & \tilde{U} \end{bmatrix}$ es unitaria y además

$$\begin{aligned} U_1^H A U_1 &= U_1^H \begin{bmatrix} Ax_1 & A\tilde{U}_1 \end{bmatrix} \\ &= U_1^H \begin{bmatrix} \lambda x_1 & A\tilde{U}_1 \end{bmatrix} \\ &= \begin{bmatrix} x_1^H \\ \tilde{U}_1^H \end{bmatrix} \begin{bmatrix} \lambda x_1 & A\tilde{U}_1 \end{bmatrix} \\ &= \begin{bmatrix} x_1^H \lambda x_1 & x_1^H A\tilde{U}_1 \\ \tilde{U}_1^H \lambda x_1 & \tilde{U}_1^H A\tilde{U}_1 \end{bmatrix} \\ &= \begin{bmatrix} \lambda & x_1^H A\tilde{U}_1 \\ 0_{n\times 1} & \tilde{U}_1^H A\tilde{U}_1 \end{bmatrix}. \end{aligned}$$

Donde hemos usado los hecho $x_1^T x_1 = 1$ y que las columnas de U_1 son ortogonales.

Ahora, $B = \tilde{U}_1^H A\tilde{U}_1$ es cuadrada de orden n. Así, por hipótesis de inducción existe una matriz unitaria $\tilde{U}_2$ tal que

$$\tilde{U}_2^H B\tilde{U}_2 = T,$$

con T cierta triangular de orden n. Hacemos ahora

$$U_2 = \begin{bmatrix} 1 & 0_{1\times n} \\ 0_{n\times 1} & \tilde{U}_2 \end{bmatrix}.$$

Entonces

$$\begin{aligned} U_2^H U_1^H A U_1 U_2 &= U_2^H \begin{bmatrix} \lambda & x_1^H A\tilde{U}_1 \\ 0_{n\times 1} & B \end{bmatrix} U_2 \\ &= U_2^H \begin{bmatrix} \lambda & x_1^H A\tilde{U}_1 \\ 0_{n\times 1} & B \end{bmatrix} \begin{bmatrix} 1 & 0_{1\times n} \\ 0_{n\times 1} & \tilde{U}_2 \end{bmatrix} \\ &= U_2^H \begin{bmatrix} \lambda & x_1^H A\tilde{U}_1\tilde{U}_2 \\ 0_{n\times 1} & B\tilde{U}_2 \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0_{1\times n} \\ 0_{n\times 1} & \tilde{U}_2^H \end{bmatrix} \begin{bmatrix} \lambda & x_1^H A\tilde{U}_1\tilde{U}_2 \\ 0_{n\times 1} & B\tilde{U}_2 \end{bmatrix} \\ &= \begin{bmatrix} \lambda & x_1^H A\tilde{U}_1\tilde{U}_2 \\ 0_{n\times 1} & \tilde{U}_2^H B\tilde{U}_2 \end{bmatrix} \\ &= \begin{bmatrix} \lambda & x_1^H A\tilde{U}_1\tilde{U}_2 \\ 0_{n\times 1} & T \end{bmatrix}. \end{aligned}$$

Entonces, $(U_1U_2)^H A (U_1U_2)$ es triangular superior y es fácil ver que U_2 es unitaria y con ello también lo es (U_1U_2). Por tanto, A unitariamente semejante a una matriz triangular.

Lema 2.2.6 *Toda matriz cuadrada es semejante a una matriz triangular superior con entradas fuera de la diagonal tan pequeñas como se desee.*

Prueba. Sea A una matriz cuadrada de orden n. Por el Lema de Schur, A es unitariamente semejante a una matriz triangular superior, es decir existe una matriz U unitaria y una matriz triangular superior T tales que

$$U^H AU = T.$$

Sea ahora ϵ real positivo tal que $0 < \epsilon < 1$, construimos la matriz diagonal $D = diag(\epsilon, \epsilon^2, \cdots, \epsilon^n)$. Entonces D es invertible y además

$$D^{-1}TD$$

tiene elementos de la forma $t_{ij}\epsilon^{j-i}$. Pero como T triangular superior, los elementos de este producto con $j < i$ son todos nulos y para los demas se tiene que

$$\left|t_{ij}\epsilon^{j-i}\right| \leq \epsilon \left|t_{ij}\right|.$$

Por tanto, tomando un valor apropiado para ϵ,las entradas de la matriz triangular $D^{-1}TD$, la cual es semejante a A por transitividad, serán tan pequeños como se desee.

Prueba del Teorema 2.2.4. Probaremos ahora sí la relación

$$\rho(A) = \inf_{\|\cdot\|} \|A\|,$$

donde el ínfimo se toma sobre todas las normas subordinadas y $\rho(A)$ se entiende como el máximo de los módulos de los valores propios de A.

Empezamos probando que

$$\rho(A) \leq \inf_{\|\cdot\|} \|A\|,$$

para ello basta observar que si λ es un valor propio de A y x es vector propio asociado, entonces para cada norma subordinada se tiene que

$$\begin{aligned} |\lambda| \|x\| &= \|Ax\| \\ &\leq \|A\| \|x\|. \end{aligned}$$

Entonces, $|\lambda| \leq \|A\|$ para todo valor propio de A y toda norma subordinada. De donde,

$$\max\{|\lambda| : \det(A - \lambda I) = 0\} \leq \inf_{\|\cdot\|} \|A\|,$$
$$\rho(A) \leq \inf_{\|\cdot\|} \|A\|.$$

Antes de probar la otra desigualdad notamos dos hechos referentes a las normas de matrices. El primero es que si $\|\cdot\|$ es una norma subordinada para matrices de orden n y S es matriz cuadrada invertible de orden n, entonces la relación

$$\|B\|' = \left\|S^{-1}BS\right\|$$

define también una norma subordinada. Para probar esto basta notar que dada una norma $\|\cdot\|$ para elementos x en $\mathbb{R}^n$, entonces $\|x\|' = \left\|S^{-1}x\right\|$ es también una norma y que para esta norma

$$\begin{aligned}
\|B\|' &= \sup_{\|x\|'\neq 0} \frac{\|Bx\|'}{\|x\|'} \\
&= \sup_{\|S^{-1}x\|\neq 0} \frac{\left\|S^{-1}Bx\right\|}{\|S^{-1}x\|} \\
&= \sup_{\|y\|\neq 0} \frac{\left\|S^{-1}BSy\right\|}{\|y\|} \\
&= \left\|S^{-1}BS\right\|.
\end{aligned}$$

Aquí se utiliza el cambio de variable $y = S^{-1}x$.

La segunda observación es que para la norma $\|\cdot\|_\infty$ que se define en $\mathbb{R}^n$ por

$$\|x\|_\infty = \max\{|x_i| : 1 \leq i \leq n\},$$

es tal que su norma matricial subordinada asociada satisface

$$\|B\|_\infty = \max_{1\leq i\leq n} \sum_{j=1}^{n} |b_{ij}|.$$

Procedemos ahora sí a probar la desigualdad faltante. Sea $\epsilon > 0$, en este punto usamos el lema 2.2.6 y las observaciones anteriores para concluir que existe una matriz invertible S tal que $S^{-1}AS = D+T$, donde D es diagonal y T es estrictamente triangular superior de entradas suficientemente pequeñas como para que

$$\|T\|_\infty < \epsilon,$$

entonces

$$\begin{aligned}\left\|S^{-1}AS\right\|_\infty &= \|D+T\|_\infty \\ &\leq \|D\|_\infty + \|T\|_\infty \\ &\leq \rho(A)+\epsilon.\end{aligned}$$

Así,

$$\|A\|'_\infty \leq \rho(A)+\epsilon. \tag{2.5}$$

En resumen, dado $\epsilon > 0$ existe una norma matricial subordinada tal que se cumple (2.5). Por tanto, para todo $\epsilon > 0$

$$\inf_{\|\cdot\|} \|A\| \leq \rho(A)+\epsilon.$$

Luego,

$$\inf_{\|\cdot\|} \|A\| \leq \rho(A).$$

Teorema 2.2.7 (Convergencia) *El método iterativo definido por (2.4) converge a la solución de $Ax = b$, para cualquier vector inicial $x^{(0)}$, siempre que $\rho(I - Q^{-1}A) < 1$.*

Prueba.
Como

$$\begin{aligned}&\rho(I-Q^{-1}A) < 1 \ y \\ &\rho(I-Q^{-1}A) = \inf_{\|\cdot\|} \left\|I-Q^{-1}A\right\|,\end{aligned}$$

donde el ínfimo se toma sobre todas las normas matriciales subordinadas, se tiene que existe una norma matricial subordinada $\|\cdot\|$ tal que $\left\|I-Q^{-1}A\right\| < 1$ y así por Teorema 2.2.1 el método iterativo converge para cualquier vector inicial $x^{(0)}$.

2.2.2 Métodos Estacionarios Clásicos

En esta sección nos basaremos en la presentación hecha en [11] para introducir los métodos estacionarios más importantes y utilizados. Se presentarán resultados básicos de convergencia para estos métodos. Para una exposición más completa ver [15]. Nos concentramos en los métodos de Jacobi, Gauss-Seidel y SOR.

En todos los casos caracterizaremos el método por la descripción de la escogencia de la matriz de iteración Q en (2.4). Para fines de notación la matriz A la descomponemos así:

$$A = D + C_U + C_L,$$

donde D denota la matriz diagonal cuyas entradas son las de la diagonal de A, es decir $D = diag(diag(A))$, C_L denota la parte estrictamente triangular inferior de A y C_U la parte estrictamente triangular superior.

Método de Jacobi

Consiste en tomar la matriz de iteración Q como la matriz diagonal D. Nótese entonces que si $a_{ii} \neq 0$ para $i = 1, 2, \cdots, n$ la iteración toma la forma

$$x^{(k+1)} = D^{-1}(D - A)x^{(k)} + D^{-1}b, \tag{2.6}$$

lo cual en las componentes se escribe como

$$x_i^{(k+1)} = \frac{1}{a_{ii}} \left(b_i - \sum_{j \neq i} a_{ij} x_j^{(k)} \right). \tag{2.7}$$

Ahora, como vimos arriba para garantizar la convergencia del método de Jacobi para cualquier solución inicial debemos tener que $\rho\left(I - D^{-1}A\right) < 1$. En este punto recordamos la norma matricial subordinada inducida por la norma infinito en $\mathbb{R}^n$ para la cual

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

Entonces como la diagonal de $I - D^{-1}A$ es nula, se sigue que

$$\left\|I - D^{-1}A\right\|_\infty = \max_{1 \leq i \leq n} \frac{1}{|a_{ii}|} \sum_{j \neq i} |a_{ij}|.$$

Por tanto, $\left\|I - D^{-1}A\right\|_\infty$ será menor que uno siempre que para todo $i = 1, 2, \cdots, n$ se cumpla que

$$\sum_{j \neq i} |a_{ij}| < |a_{ii}|. \tag{2.8}$$

Las matrices que cumplen (2.8) se dicen Diagonalmente Dominantes. Hemos probado el siguiente

Teorema 2.2.8 *Si la matriz A es diagonalmente dominante entonces el método de Jacobi converge a la solución de $Ax = b$ para cualquier vector inicial $x^{(0)}$.*

Método de Gauss - Seidel

Si el método de Jacobi es aplicado de modo secuencial en el tiempo, respecto de las componentes de $x^{(k+1)}$, entonces al calcular $x_i^{(k+1)}$ se habrán calculado ya $x_1^{(k+1)}, x_2^{(k+1)}, \cdots, x_{i-1}^{(k+1)}$ podemos entonces usarlas para el cálculo de la siguiente componente remplazando la expresión (2.7) por la expresión

$$x_i^{(k+1)} = \frac{1}{a_{ii}} \left(b_i - \sum_{j<i} a_{ij} x_j^{(k+1)} - \sum_{j>i} a_{ij} x_j^{(k)} \right). \tag{2.9}$$

Esta es la iteración de Gauss-Seidel. A nivel matricial esto implica tomar Q como la parte triangular inferior de A, o sea $Q = D + C_L$. En este caso la condición de convergencia no cambia y la convergencia está garantizada por el siguiente

Teorema 2.2.9 *Si la matriz A es diagonalmente dominante entonces el método de Gauss-Seidel converge a la solución de $Ax = b$ para cualquier vector inicial $x^{(0)}$.*

Prueba: Necesitamos probar que $\rho\left(I - Q^{-1}A\right) < 1$. Para tal efecto sea λ valor propio de $I - Q^{-1}A$ y sea x vector propio asociado con $\|x\|_\infty = 1$. Entonces

$$\begin{aligned} \left(I - Q^{-1}A\right) x &= \lambda x, \\ (Q - A)\, x &= \lambda Q x, \\ -C_U &= \lambda (D + C_L) x. \end{aligned}$$

Por tanto, se tiene que para $1 \leq i \leq n$

$$\begin{aligned} -\sum_{j=i+1}^{n} a_{ij} x_j &= \lambda \sum_{j=1}^{i} a_{ij} x_j, \\ \lambda a_{ii} x_i &= -\lambda \sum_{j=1}^{i-1} a_{ij} x_j - \sum_{j=i+1}^{n} a_{ij} x_j. \end{aligned} \tag{2.10}$$

Escogemos ahora i de modo que $|x_i| = \|x\|_\infty = 1$. Entonces a partir de (2.10) se tiene que para este valor de i,

$$|\lambda|\,|a_{ii}| \leq |\lambda| \sum_{j=1}^{i-1} |a_{ij}| + \sum_{j=i+1}^{n} |a_{ij}|.$$

Ahora como A es diagonalmente dominante se tiene que

$$\sum_{j\neq i} |a_{ij}| < |a_{ii}|\,,$$
$$\sum_{j=i+1}^{n} |a_{ij}| < |a_{ii}| - \sum_{j=1}^{i-1} |a_{ij}|$$

Por tanto,

$$|\lambda| \left(|a_{ii}| - \sum_{j=1}^{i-1} |a_{ij}| \right) \leq \sum_{j=i+1}^{n} |a_{ij}|\,,$$

y así

$$|\lambda| \leq \left(|a_{ii}| - \sum_{j=1}^{i-1} |a_{ij}| \right)^{-1} \sum_{j=i+1}^{n} |a_{ij}| < 1.$$

Como λ se tomó arbitrario se tiene que $\rho\left(I - Q^{-1}A\right) < 1$.

Método SOR

La sigla SOR viene del inglés *Successive Overrelaxation* que podriamos traducir como Sobrerrelajación Sucesiva. La iteración SOR puede pensarse como una variación paramétrica de Gauss-Seidel consistente en lo siguiente. Partimos de la iteración Gauss-Seidel original dada por

$$\tilde{x}_i^{(k+1)} = \frac{1}{a_{ii}} \left(b_i - \sum_{j<i} a_{ij} x_j^{(k+1)} - \sum_{j>i} a_{ij} x_j^{(k)} \right), \tag{2.11}$$

pero tomamos el nuevo valor de $x_i^{(k+1)}$ como

$$x_i^{(k+1)} = x_i^{(k)} + w \left(\tilde{x}_i^{(k+1)} - x_i^{(k)} \right), \tag{2.12}$$

donde w es un parámetro de valor real. Nótese que $w = 1$ lleva a Gauss-Seidel.

Ahora, remplazando (2.11) en (2.12) se obtiene la expresión matricial equivalente que nos devela la matriz de iteración involucrada en el proceso. En efecto se obtiene,

$$(D + wC_L)\, x^{(k+1)} = ((1-w)D - wC_U)\, x^{(k)} + wb. \tag{2.13}$$

De donde, $Q = (D + wC_L)$ y se debe chequear bajo qué condiciones

$$\rho\left((D + wC_L)^{-1}\left((1-w)D - wC_U\right)\right) < 1.$$

En este caso se tiene el siguiente

Teorema 2.2.10 *Si A es hermitiana y definida positiva, Q no singular, D definida positiva y $\alpha = w^{-1} > 1/2$ $(0 < w < 2)$, entonces el método SOR converge.*

Prueba:

Empezamos recordando que la recurrencia SOR está dada por las ecuaciones

$$(D + wC_L)\, x^{(k+1)} = \left((1-w)D - wC_U\right) x^{(k)} + wb,$$

O equivalentemente,

$$(\alpha D + C_L)\, x^{(k+1)} = \left((\alpha - 1)D - C_L^H\right) x^{(k)} + b, \tag{2.14}$$

$$Q_\alpha x^{(k+1)} = (Q_\alpha - A)x^{(k)} + b, \tag{2.15}$$

donde $Q_\alpha = (\alpha D + C_L)$ y hemos utilizado el hecho que A es hermitiana para obtener $C_U = C_L^H$.

Nos concentramos en (2.14) y encontraremos el radio espectral de $G = \left(I - Q_\alpha^{-1}A\right)$ es menor que uno cuando se cumple la condición indicada en el teorema. Sea λ un valor propio de G y x un vector propio de G asociado a λ. Definimos de modo auxiliar $y = (I - G)x$, se sigue que

$$\begin{aligned} y &= Q_\alpha^{-1}Ax, \\ Q_\alpha y &= Ax. \end{aligned} \tag{2.16}$$

Además,

$$\begin{aligned} (Q_\alpha - A)\, y &= Q_\alpha y - Ay \\ &= Ax - Ay \\ &= A\left(x - Q_\alpha^{-1}Ax\right) \\ &= A\left(I - Q_\alpha^{-1}A\right) x \\ &= AGx. \end{aligned} \tag{2.17}$$

Luego, de la definición de Q_α y de (2.16) y 2.17) se siguen las relaciones

$$\begin{aligned} \left((\alpha - 1)\, D - C_L^H\right) y &= AGx, \\ (\alpha D + C_L)\, y &= Ax. \end{aligned} \tag{2.18}$$

Al hacer producto interno con y se llega a

$$\begin{aligned} (\alpha - 1) \langle Dy, y\rangle - \left\langle C_L^H y, y\right\rangle &= \langle AGx, y\rangle , \\ \alpha \langle y, Dy\rangle + \langle y, C_L y\rangle &= \langle y, Ax\rangle . \end{aligned} \tag{2.19}$$

Sumando estas igualdades tenemos

$$(2\alpha - 1) \langle Dy, y\rangle = \langle AGx, y\rangle + \langle y, Ax\rangle . \tag{2.20}$$

Aquí se usan los hechos que

$$\begin{aligned} \left\langle C_L^H y, y\right\rangle &= \langle y, C_L y\rangle \text{ y} \\ \langle Dy, y\rangle &= \left\langle y, D^H y\right\rangle = \langle y, Dy\rangle . \end{aligned}$$

De otra parte como x es vector propio de G asociado a λ, se sigue que

$$\begin{aligned} \langle y, Ax\rangle &= \langle y, Ax\rangle \\ &= \langle x - Gx, Ax\rangle \\ &= \langle x - \lambda x, Ax\rangle \\ &= (1 - \lambda) \langle x, Ax\rangle . \end{aligned} \tag{2.21}$$

Ahora

$$\begin{aligned} \langle AGx, y\rangle &= \langle A\lambda x, y\rangle \\ &= \lambda \langle Ax, x - \lambda x\rangle \\ &= \lambda \left(1 - \bar{\lambda}\right) \langle Ax, x\rangle \\ &= \lambda \left(1 - \bar{\lambda}\right) \langle x, Ax\rangle , \end{aligned} \tag{2.22}$$

donde la última igualdad se debe a que A es hermitiana. Finalmente, notamos que

$$\begin{aligned} \langle Dy, y\rangle &= \langle D(1 - \lambda) x, (1 - \lambda) x\rangle \\ &= |1 - \lambda|^2 \langle Dx, x\rangle . \end{aligned} \tag{2.23}$$

Usando (2.21), (2.22) y (2.23) podemos reescribir (2.20) como

$$\begin{aligned} (2\alpha - 1) |1 - \lambda|^2 \langle Dx, x\rangle &= \lambda \left(1 - \bar{\lambda}\right) \langle x, Ax\rangle + (1 - \lambda) \langle x, Ax\rangle \\ &= \left(1 - |\lambda|^2\right) \langle x, Ax\rangle . \end{aligned} \tag{2.24}$$

Nótese que si $\lambda \neq 1$, entonces al ser $\alpha > 1/2$ y D definida positiva el lado izquierdo de (2.24) es estrictamente positivo. Pero como A definida positiva

y x no es el vector nulo, se tiene $\langle x, Ax\rangle > 0$ y por tanto $\left(1 - |\lambda|^2\right) > 0$. Así, $|\lambda| < 1$ que era lo que queríamos demostrar.

Ahora $\lambda = 1$ implicaría, $y = (1-\lambda)\,x = 0$, que a su vez conlleva a $Ax = Q_\alpha y = 0$, por lo que tendríamos $\langle x, Ax\rangle = 0$ con x no nulo, lo que contradice el hecho A definida positiva. Así, debe ser $\lambda \neq 1$. Por tanto, $\rho\left(I - Q_\alpha^{-1}A\right) < 1$.

Es importante notar que nuestra prueba sigue siendo válida para cualquier escritura de A hermitiana y definida positiva en la forma

$$A = D + C + C^H,$$

con D hermitiana y definida positiva. También debe notarse que como el método Gauss-Seidel es SOR con $w = 1$, este teorema es un nuevo resultado para convergencia de Gauss-Seidel. Por supuesto la pregunta que nos aborda ahora es qué valor de w nos garantiza la convergencia más rápida? Una respuesta a esta pregunta es la escogencia

$$w_0 = \frac{2}{1 + \sqrt{1 - \rho^2(J)}},$$

donde $\rho(J)$ denota el radio espectral de la matriz de iteración de Jacobi para A. Las condiciones en las que esta escogencia es óptima son presentadas en [15] y no las presentamos aquí por no ser el objetivo central de este documento, máxime cuando en la mayoría de las aplicaciones prácticas tal w_0 resulta tan difícil de encontrar como resolver el problema $Ax = b$, que es nuestro objetivo.

2.3 Factorización Incompleta

Una clase importante de precondicionadores se obtiene de la factorización LU, que sabemos trabaja para matrices arbitrarias. Sin embargo, existen algunas variantes que explotan la estructura especial que pueda tener la matriz: simétrica, definida positiva, ralas (sparse), etc. Para el caso de las matrices con un patrón de dispersión o ralas, se tiene un procedimiento denominado factorización incompleta, es decir: Si A es una matriz rala, los factores L y U usualmente no tienen el mismo patrón de dispersión de la matriz A, y como la idea es conservar este patrón, entonces se descartan los elementos diferentes de cero en aquellas posiciones donde la matriz A tenía un cero. Tales posiciones se denominan **fill-elements** (posiciones rellenadas). Así se obtendría una factorización aproximada: $A \approx LU$.

La idea de generar factorizaciones aproximadas ha sido estudiada por muchos investigadores desde 1960. Posteriormente J. Meijerink y H. Van der Vorst en 1977 [13] la hicieron popular cuando la usaron para generar precondicionadores para el gradiente conjugado y otros métodos iterativos.

En esta sección obtendremos un algoritmo para llevar a cabo la factorización incompleta de una matriz dado cierto patrón de dispersión y se prueba la posibilidad de ejecutar el algoritmo cuando la matriz A es diagonalmente dominante. Para llevar esto a cabo nos basamos en la presentación hecha en [19]. Empezaremos precisando qué entenderemos por factorización incompleta, para luego obtener un algoritmo a partir de esta descripción y por último estudiaremos para qué tipos de matrices nuestro algoritmo ha de converger.

2.3.1 Conjunto de Dispersión

A fin de controlar los patrones de dispersión presentes en las matrices L y U notamos por $\mathcal{Z}$ el subconjunto de $\{(i,j) : j \neq i, 1 \leq i,j \leq n\}$ constituido por las entradas en las que deseamos que L y U posean entradas nulas. Se trata entonces de escribir la matriz A en la forma

$$A = LU + R, \tag{2.25}$$

donde las matrices L, U y R satisfacen las siguientes condiciones

1. $l_{ii} = 1$ para $i = 1, 2, \cdots, n$.
2. Si $(i,j) \in \mathcal{Z}$ e $i > j$, entonces $l_{ij} = 0$.
3. Si $(i,j) \in \mathcal{Z}$ e $i < j$, entonces $u_{ij} = 0$.
4. Si $(i,j) \notin \mathcal{Z}$, entonces $r_{ij} = 0$.

Una escritura de este tipo se dice una factorización incompleta de A asociada al patrón de dispersión $\mathcal{Z}$.

Es importante observar que estas condiciones implican que las matrices L y U han de respetar el patrón de dispersión postulado por $\mathcal{Z}$ y hacen que la aproximación de A por LU sea exacta por fuera de $\mathcal{Z}$.

2.3.2 Algoritmo de Factorización Incompleta

Obtendremos ahora un algoritmo para llevar a cabo la factorización incompleta de una matriz A, supuesta que dicha factorización existe. Esto probará que supuesta la existencia de la factorización incompleta, ésta ha de

ser única. Las condiciones para la existencia las estudiaremos más adelante en la sección 2.3.5.

Obtendremos las matrices L y U mediante un proceso inductivo de fila a fila. Supongamos que han sido calculadas las primeras $k-1$ filas de L y de U, encontremos la $k-ésima$. Para tal efecto escribimos las k primeras filas de (2.25) en la forma

$$\begin{bmatrix} A_{11} & A_{1k} \\ A_{k1}^T & A_{kk}^T \end{bmatrix} = \begin{bmatrix} L_{11} & 0 \\ L_{k1}^T & e_{1,n-k+1}^T \end{bmatrix} \begin{bmatrix} U_{11} & U_{1k} \\ 0 & U_{kk}^T \end{bmatrix} + \begin{bmatrix} R_{11} & R_{1k} \\ R_{k1}^T & R_{kk}^T \end{bmatrix}. \quad (2.26)$$

Donde, en notación de Matlab,

$$\begin{aligned} A_{11} &= A(1:k-1, 1:k-1), \\ A_{1k} &= A(1:k-1, k:n), \\ A_{k1}^T &= A(k, 1:k-1), \\ A_{kk}^T &= A(k, k:n), \end{aligned}$$

$e_{1,n-k+1}$ es el vector de $\mathbb{R}^{n-k+1}$ con primera entrada uno y las demás nulas y para las demás matrices la interpretación es análoga.

Para nuestro algoritmo el interés se centra en encontrar los vectores L_{k1}^T y U_{kk}^T. Pero de (2.26) se tiene que

$$L_{k1}^T U_{11} + R_{k1}^T = A_{k1}^T, \quad (2.27)$$

y

$$U_{kk}^T + R_{kk}^T = A_{k1}^T - L_{k1}^T U_{1k}. \quad (2.28)$$

Encontraremos los vectores deseados de modo secuencial, siguiendo el orden

$$l_{k1}, l_{k2}, \cdots, l_{k,k-1}, u_{k,k}, u_{k,k+1}, \cdots, u_{k,n}.$$

Supongamos que han sido calculados $l_{k1}, l_{k2}, \cdots, l_{k,j-1}$, debemos hallar ahora $l_{k,j}$. Si $(k,j) \in \mathcal{Z}$, hacemos $l_{kj} = 0$. En caso contrario, $r_{kj} = 0$ y entonces de (2.27) se tiene que

$$\sum_{i=1}^{j} l_{ki} u_{ij} = a_{kj.}$$

Luego,

$$l_{kj} = \frac{a_{kj} - \sum_{i=1}^{j-1} l_{ki} u_{ij}}{u_{jj}}. \quad (2.29)$$

En este punto es importante observar que para $(k,j) \notin \mathcal{Z}$ no interesa cómo hallan sido encontrados los valores previos para l_{ki} y u_{ij}, si l_{kj} se

calcula utilizando (2.29) la componente (k, j) del producto LU ha de ser a_{kj}. Por tanto, el hacer $l_{kj} = 0$ para los $(k, j) \in \mathcal{Z}$ no ha de afectar los valores de LU fuera del patrón de dispersión.

Similarmente, si hemos calculado $u_{kk}, u_{k,k+1}, \cdots, u_{k,j-1}$, debemos hallar ahora $u_{k,j}$. Si $(k, j) \in \mathcal{Z}$, hacemos $u_{kj} = 0$. En caso contrario, $r_{kj} = 0$ y entonces de (2.28) se tiene que

$$u_{kj} = a_{kj} - \sum_{i=1}^{k-1} l_{ki} u_{ij}.$$

Tenemos el siguiente algoritmo

$L(1,1) = 1,$
$for\ j = 1 : n$
    $if\ (1, j) \in \mathcal{Z}$
        $U(1, j) = 0;$
    $else$
        $U(1, j) = A(1, j);$
    end
end
$for\ k = 2 : n$
    $for\ j = 1 : k - 1$
        $if\ (k, j) \in \mathcal{Z}$
            $L(k, j) = 0;$
        $else$
            $L(k, j) = (A(k, j) - L(k, 1 : j - 1) * U(1 : j - 1, j)) / U(j, j);$
        end
    end
    $L(k, k) = 1;$
    $for\ j = k : n$
        $if\ (k, j) \in \mathcal{Z}$
            $U(k, j) = 0;$
        $else$
            $U(k, j) = A(k, j) - L(k, 1 : k - 1) * U(1 : k - 1, j);$
        end
    end
end

Obsérvese que la inicialización para la primera fila de L y U garantiza cumplir las condiciones deseadas, independientemente de lo que vaya a colocarse en las entradas de las siguientes filas. Por supuesto, este proceso se

completa siempre que no se anule ningún $U(j,j)$. Veremos ahora en más detalle bajo qué condiciones este algoritmo funciona.

2.3.3 Más de Matrices Diagonalmente Dominantes

Las matrices diagonalmente dominantes han mostrado ser de gran importancia en el estudio de métodos iterativos. Recordemos que para matrices no singulares diagonalmente dominantes los teoremas 2.2.8 y 2.2.9 garantizan la convergencia de los métodos iterativos estacionarios de Gauss-Seidel y de Jacobi. En esta sección encontraremos otras propiedades importantes de estas matrices que nos permitirán probar la existencia de la factorización LU incompleta para ellas. Esencialmente encontraremos algunas condiciones de singularidad y procedimientos que permiten encontrar nuevas matrices no singulares y diagonalmente dominantes a partir de una primera dada.

Empezamos recordando que una matriz A de orden n se dice diagonalmente dominante si

$$|a_{ii}| \geq \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}|\,, \quad i = 1, 2, \cdots, n \tag{2.30}$$

A se dice estrictamente diagonalmente dominante si la desigualdad estricta se cumple en (2.30) para todo i.

Teorema 2.3.1 *Sea A diagonalmente dominante y x vector tal que*

$$Ax = 0.$$

Si

$$|x_i| = \max_j \{|x_j|\} > 0,$$

entonces

$$a_{ij} \neq 0 \Rightarrow |x_i| = |x_j|\,,\ y$$
$$|a_{ii}| = \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}|\,.$$

Prueba. Como $Ax = 0$ y $|x_i| = \max_j \{|x_j|\} > 0$, se tiene que

$$0 = a_{ii} + \sum_{j\neq i} a_{ij}\frac{x_j}{x_i}.$$

De donde,

$$
\begin{aligned}
0 &= \left| a_{ii} + \sum_{j \neq i} a_{ij} \frac{x_j}{x_i} \right| \qquad (2.31) \\
&\geq |a_{ii}| - \left| \sum_{j \neq i} a_{ij} \frac{x_j}{x_i} \right| \\
&\geq |a_{ii}| - \sum_{j \neq i} |a_{ij}| \, \frac{|x_j|}{|x_i|} \\
&\geq |a_{ii}| - \sum_{j \neq i} |a_{ij}| \, .
\end{aligned}
$$

Pero por dominancia diagonal,

$$
0 \leq |a_{ii}| - \sum_{j \neq i} |a_{ij}| \, .
$$

Así debe tener la igualdad para la $i - ésima$ fila. Ahora, si $a_{ij} \neq 0$ y $|x_i| \neq |x_j|$, se tendría $|x_j| / |x_i| < 1$ y la última desigualdad en (2.31) sería estricta contradiciendo la dominancia diagonal. ■

Corolario 2.3.2 *Una matriz estrictamente diagonalmente dominante es no singular.*

Prueba. Veamos que no existe vector x no nulo tal que $Ax = 0$. En efecto, supongamos que $Ax = 0$ y $x \neq 0$. Entonces en virtud del teorema anterior las filas de A correspondientes a entradas de x con absoluto maximal no satisfacen dominancia diagonal estricta, lo que contradice la hipótesis. ■

Es importante notar que aunque la dominancia estricta garantiza la no singularidad, la sola dominancia no lo hace. Por ejemplo, la matriz

$$
A = \begin{bmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{bmatrix},
$$

es no singular, diagonalmente dominante pero sin dominancia estricta. Esto se nota por simple inspección y la no singularidad por ser $\det A = 4$, por indicar algún método.

De otra parte la matriz

$$
B = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 2 & 4 \end{bmatrix},
$$

es claramente singular, dos filas iguales, siendo diagonalmente dominante.

La siguiente proposición establece un procedimiento que permite obtener nuevas matrices no singulares y diagonalmente dominantes a partir de una primera dada.

Proposición 2.3.3 *Sea A no singular y diagonalmente dominante. Si $\tilde{A}$ es obtenida de A disminuyendo la magnitud de un conjunto (quizá vacío) de elementos fuera de la diagonal principal e incrementando la magnitud de otro conjunto (quizá vacío) de elementos de la diagonal, entonces $\tilde{A}$ es no singular diagonalmente dominante.*

Prueba. La dominancia diagonal es obvia si se tiene en cuenta que las entradas $\tilde{a}_{ij}$ de $\tilde{A}$ satisfacen

$$\begin{aligned} |\tilde{a}_{ij}| &\leq |a_{ij}|\,,\ i \neq j, \\ |\tilde{a}_{ii}| &\geq |a_{ii}|\,. \end{aligned}$$

Y como A diagonalmente dominante se tiene

$$\begin{aligned} |\tilde{a}_{ii}| &\geq |a_{ii}| \\ &\geq \sum_{j\neq i} |a_{ij}| \\ &\geq \sum_{j\neq i} |\tilde{a}_{ij}|\,. \end{aligned}$$

Veamos ahora que $\tilde{A}$ es no singular. Supongamos que $\tilde{A}x = 0$ con $x \neq 0$. Ahora, si intercambiamos dos filas de $\tilde{A}$ y las correspondientes dos columnas, la nueva matriz conservará la dominancia diagonal. Respecto del sistema $\tilde{A}x = 0$ estos intercambios corresponderían a intercambio de filas y reordenamientos de variables. Podemos hacer entonces estos cambios, las veces que se requiera de modo que se obtenga una expresión de la forma

$$\begin{bmatrix} \tilde{A}_{11} & \tilde{A}_{12} \\ \tilde{A}_{21} & \tilde{A}_{22} \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix},$$

donde z_1 agrupa todas las entradas de x de máxima magnitud. Obsérvese ahora que ningún elemento de $\tilde{A}_{12}$ puede ser no nulo pues en virtud del teorema 2.3.1 eso implicaría que alguna componente almacenada en z_2 tendría norma máxima contradiciendo la escogencia de z_1. Tenemos entonces una expresión de la forma

$$\begin{bmatrix} \tilde{A}_{11} & 0 \\ \tilde{A}_{21} & \tilde{A}_{22} \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}. \tag{2.32}$$

Además, por el mismo teorema, las filas asociadas a z_1 satisfacen la igualdad en la dominancia diagonal. Ahora para regresar a la matriz A original requerimos disminuir la magnitud de algunos elementos de la diagonal de la nueva matriz y aumentar la magnitud de elementos fuera de la diagonal. Siendo obligatorio que alguno de estos cambios tenga lugar en la parte superior de (2.32) para recuperar la no singularidad original de A, pero esto conduciría a que A no satisfaga la dominancia diagonal. Por tanto, debe ser $\tilde{A}$ no singular. ■

Es importante notar que esta proposición habla de cambio de magnitud en las entradas más sin embargo no de cambio de signo de las entradas. Por supuesto, si cambiamos las magnitudes como permite el teorema y cambiamos los signos no salimos de la hipótesis. Pero un mero cambio en los signos de las entradas puede alterar la no singularidad como lo muestran el par de matrices

$$\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}, \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}.$$

Corolario 2.3.4 *Cualquier submatriz principal de una matriz A no singular diagonalmente dominante, es una matriz no singular diagonalmente dominante.*

Prueba. Para notar esto escribimos A en la forma

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$$

y nos concentramos en probar que A_{11} es no singular pues la dominancia diagonal es inmediata.

Ahora, de la proposición anterior se sigue que la matriz

$$\tilde{A} = \begin{bmatrix} A_{11} & 0 \\ 0 & A_{22} \end{bmatrix},$$

obtenida desde A por la disminución de magnitud de entradas fuera de la diagonal, es una matriz no singular. Pero,

$$\det \tilde{A} = \det A_{11} * \det A_{22}.$$

Así, A_{11} no singular. ■

2.3.4 Complemento de Schur y Eliminación Gaussiana

Establezcamos ahora una relación entre el proceso de eliminación gaussiana y la matrices no singulares diagonalmente dominantes. Como concepto auxiliar pero muy importante en este punto y en desarrollos posteriores, presentamos el llamado Complemento de Schur.

Empezamos considerando una matriz A no singular diagonalmente dominante, particionada en la forma

$$A = \begin{bmatrix} a_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix},$$

entonces por el corolario anterior $a_{11} \neq 0$ y podemos usar eliminación gaussiana para anular las componentes en A_{21}. Para ello debemos multiplicar a izquierda por la matriz

$$\begin{bmatrix} 1 & 0 \\ -A_{21}a_{11}^{-1} & 1 \end{bmatrix},$$

para obtener

$$\begin{bmatrix} 1 & 0 \\ -A_{21}a_{11}^{-1} & 1 \end{bmatrix} \begin{bmatrix} a_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} = \begin{bmatrix} a_{11} & A_{12} \\ 0 & A_{22} - A_{21}a_{11}^{-1}A_{12} \end{bmatrix}.$$

La matriz $A_{22} - A_{21}a_{11}^{-1}A_{12}$ se dice el complemento de Schur de a_{11}. Si pudieramos garantizar que $A_{22} - A_{21}a_{11}^{-1}A_{12}$ es nuevamente no singular diagonalmente dominante, podríamos garantizar el poder hacer un nuevo paso de eliminación gaussiana sin problema alguno. Empezamos por la no singularidad.

Definición 2.3.5 *Supongamos que la matriz $A \in \mathbb{R}^{n\times n}$ se puede escribir de la forma $A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}$, donde A_{11} es $r \times r$. Si se supone que A_{11} es no singular, la matriz $S = A_{22} - A_{21}A_{11}^{-1}A_{12}$ se denomina el* ***Complemento de Schur*** *de A_{11} en A.*

Observamos que si A es no singular también S es no singular pues

$$A = \begin{bmatrix} I & 0 \\ A_{21}A_{11}^{-1} & I \end{bmatrix} \begin{bmatrix} A_{11} & A_{12} \\ 0 & S \end{bmatrix}.$$

Tenemos entonces garantizada la no singularidad de $A_{22} - A_{21}a_{11}^{-1}A_{12}$ cuando A es no singular diagonalmente dominante. De hecho el corolario (2.3.4) nos garantiza que el complemento de Schur a cualquier submatriz principal de una matriz A no singular y diagonalmente dominante es no singular. Veamos que la dominancia diagonal también es preservada.

Teorema 2.3.6 *Si A es una matriz no singular diagonalmente dominante, entonces el complemento de Schur de a_{11} es no singular diagonalmente dominante.*

Prueba. Por las observaciones de arriba nos falta sólo probar la dominancia diagonal. Empezamos particionando la matriz A en la forma

$$A = \begin{bmatrix} a_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$$

y escribiendo A_{22} como $A_{22} = D + B$ donde D representa su diagonal y B consta de los elementos no diagonales de A_{22}. Si notamos por $|A|$ la matriz cuyas entradas son las magnitudes de las entradas de A y por e el vector de $\mathbb{R}^n$ con todas sus entradas iguales a uno, podemos entonces expresar la dominancia diagonal de A_{22} por la condición equivalente

$$|D|\, e - |B|\, e \geq 0.$$

Sea $S = A_{22} - A_{21}a_{11}^{-1}A_{12}$ el complemento de Schur de a_{11}. Podemos ahora descomponer S como $S = D_S + B_S$. Note que la presencia de $-A_{21}a_{11}^{-1}A_{12}$ en S pueden generar una disminución en la magnitud de la diagonal D y un incremento en la magnitud de los elementos no diagonales contenidos en B. En todo caso, estas alteraciones podemos acotarlas así:

$$\begin{aligned} |D_S|\, e - |B_S|\, e &\geq |D|\, e - |B|\, e - \left|a_{11}^{-1}A_{21}A_{12}\right| e \\ &\geq |D|\, e - |B|\, e - \left|a_{11}^{-1}\right| |A_{21}A_{12}|\, e \\ &\geq |D|\, e - |B|\, e - \left|a_{11}^{-1}\right| |A_{21}|\, |A_{12}|\, e. \end{aligned}$$

La última desigualdad se debe a la desigualdad triangular para valor absoluto aplicada a la definición de producto de matrices. Ahora como A es diagonalmente dominante

$$\begin{aligned} |a_{11}| &\geq |A_{12}|\, e, \\ 1 &\geq \left|a_{11}^{-1}\right| |A_{12}|\, e, \end{aligned}$$

por tanto,

$$|D_S|\, e - |B_S|\, e \geq |D|\, e - |B|\, e - |A_{21}|\, e.$$

Pero por la dominancia diagonal de A aplicada a las últimas $n-1$ filas se sigue que la parte derecha de esta última desigualdad es no negativa y por tanto ha de ser

$$|D_S|\, e - |B_S|\, e \geq 0.$$

Con lo que queda establecida la dominancia diagonal de S. ■

Este resultado junto con los comentarios previos a la definición 2.3.5 de complemento de Schur nos permiten garantizar que el proceso de eliminación gaussiana estándar se puede llevar a cabo hasta su completación en matrices no singulares diagonalmente dominantes.

2.3.5 Existencia de la Factorización LU Incompleta

Con todas estas herramientas disponibles abordamos ahora sí de modo directo la prueba de la existencia de la factorización LU incompleta para matrices no singulares diagonalmente dominantes. El resultado es establecido por el siguiente:

Teorema 2.3.7 *Si A es una matriz no singular diagonalmente dominante, entonces A posee una descomposición LU para cada patrón de dispersión dado $\mathcal{Z}$.*

Prueba. La demostración consiste esencialmente en apreciar que los procedimientos llevados a cabo en el algoritmo expuesto arriba llevan a cabo, bajo las hipótesis del teorema, un proceso de eliminación gaussiana a una sucesión de matrices no singulares diagonalmente dominantes. Eliminación gaussiana tal que está garantizada por los resultados de las secciones previas.

Procederemos por inducción sobre el orden n de la matriz A. Supongamos que el resultado es cierto para matrices de orden $n-1$. Particionamos la matriz A como sigue

$$A = \begin{bmatrix} a_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$$

y descompongamos A_{12} y A_{21} de la forma

$$A_{21} = \tilde{A}_{21} + R_{21} \text{ and } A_{12} = \tilde{A}_{12} + R_{12}, \tag{2.33}$$

donde $\tilde{A}_{21}$ y $\tilde{A}_{12}$ son obtenidas haciendo cero las entradas dentro del patrón de dispersión. Entonces, R_{21} y R_{12} son cero por fuera del patrón de dispersión y la matriz

$$\tilde{A} = \begin{bmatrix} a_{11} & \tilde{A}_{12} \\ \tilde{A}_{21} & A_{22} \end{bmatrix}$$

se obtiene de A por disminución de la magnitud de entradas fuera de la diagonal y en virtud de la proposición (2.3.3) $\tilde{A}$ ha de ser no singular diagonalmente dominante. Efectuamos ahora un paso de factorización incompleta,

haciendo

$$L_1 = \begin{bmatrix} 1 & 0 \\ a_{11}^{-1}\tilde{A}_{21} & 0 \end{bmatrix} \text{ y}$$
$$U_1 = \begin{bmatrix} a_{11} & \tilde{A}_{12} \\ 0 & 0 \end{bmatrix},$$

para obtener

$$\begin{aligned} A - L_1U_1 &= \begin{bmatrix} a_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} - \begin{bmatrix} 1 & 0 \\ a_{11}^{-1}\tilde{A}_{21} & 0 \end{bmatrix} \begin{bmatrix} a_{11} & \tilde{A}_{12} \\ 0 & 0 \end{bmatrix} \\ &= \begin{bmatrix} a_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} - \begin{bmatrix} a_{11} & \tilde{A}_{12} \\ \tilde{A}_{21} & a_{11}^{-1}\tilde{A}_{21}\tilde{A}_{12} \end{bmatrix} \\ &= \begin{bmatrix} 0 & R_{12} \\ R_{21} & A_{22} - a_{11}^{-1}\tilde{A}_{21}\tilde{A}_{12} \end{bmatrix}. \end{aligned}$$

Pero, $\tilde{A}_{22} = A_{22} - a_{11}^{-1}\tilde{A}_{21}\tilde{A}_{12}$ es el complemento de Schur de a_{11} en $\tilde{A}$ y por tanto es no singular y diagonalmente dominante. Así, por hipótesis de inducción $\tilde{A}_{22}$ posee una fatorización LU incompleta que respeta el patrón de dispersión correspondiente a $\mathcal{Z}$ y podemos escribir

$$\tilde{A}_{22} = L_{22}U_{22} + R_{22}.$$

Ahora podemos hacer,

$$L = \begin{bmatrix} 1 & 0 \\ a_{11}^{-1}\tilde{A}_{21} & L_{22} \end{bmatrix} \text{ y}$$
$$U = \begin{bmatrix} a_{11} & \tilde{A}_{12} \\ 0 & U_{22} \end{bmatrix},$$

y obtener

$$\begin{aligned} A - LU &= \begin{bmatrix} a_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} - \begin{bmatrix} 1 & 0 \\ a_{11}^{-1}\tilde{A}_{21} & L_{22} \end{bmatrix} \begin{bmatrix} a_{11} & \tilde{A}_{12} \\ 0 & U_{22} \end{bmatrix} \\ &= \begin{bmatrix} a_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} - \begin{bmatrix} a_{11} & \tilde{A}_{12} \\ \tilde{A}_{21} & a_{11}^{-1}\tilde{A}_{21}\tilde{A}_{12} + L_{22}U_{22} \end{bmatrix} \\ &= \begin{bmatrix} 0 & R_{12} \\ R_{21} & R_{22} \end{bmatrix}. \end{aligned}$$

Hemos obtenido entonces una factorización incompleta de A de modo que el patrón de dispersión $\mathcal{Z}$ es respetado y de la forma $A = LU + R$, donde obviamente tomamos

$$R = \begin{bmatrix} 0 & R_{12} \\ R_{21} & R_{22} \end{bmatrix}.$$

■

Definición 2.3.8 *Si A y B son matrices de igual orden escribimos $A \leq B$ si las entradas de A y B satisfacen*

$$a_{ij} \leq b_{ij}.$$

Corolario 2.3.9 *Si A es una matriz estrictamente diagonalmente dominante tal que*

$$a_{ii} > 0,\ a_{ij} \leq 0 \text{ si } i \neq j \text{ y } A^{-1} \geq 0$$

entonces para cada patrón de dispersión dado $\mathcal{Z}$ la matriz A posee una descomposición incompleta LU tal que

$$A \leq LU \text{ y } (LU)^{-1} \geq 0.$$

Prueba. Este resultado se obtiene por simple observación de los pasos en la prueba del teorema anterior. Primero notamos que por contrucción R_{12} y R_{21} en (2.33) son respectivamente vectores columna y fila de $n-1$ entradas y que bajo las hipótesis del corolario se hacen no positivos. Ahora, para aplicar de modo apropiado el proceso inductivo debemos verificar que la matriz

$$\tilde{A}_{22} = A_{22} - a_{11}^{-1}\tilde{A}_{21}\tilde{A}_{12},$$

satisface las condiciones del corolario. En efecto, al ser un complemento de Schur de una matriz no singular diagonalmente dominante $\tilde{A}_{22}$ es no singular diagonalmente dominante. Además, por dominancia diagonal de A, para todo $i > 1$

$$\left|a_{11}^{-1}\tilde{a}_{i1}\tilde{a}_{1i}\right| < |\tilde{a}_{i1}| < a_{ii}.$$

Así, la diagonal de $\tilde{A}$ es positiva. Ahora, como tanto los elementos de A_{22} y como los de $-a_{11}^{-1}\tilde{A}_{21}\tilde{A}_{12}$ son no positivos, entonces las entradas no diagonales de $\tilde{A}_{22}$ han de ser no positivas. Nos resta mostrar que $\left(\tilde{A}_{22}\right)^{-1} \geq 0$, para estar en las condiciones de la hipótesis de inducción. Para ello nos valdremos de la factorización,

$$\begin{bmatrix} 1 & 0 \\ -A_{21}a_{11}^{-1} & I \end{bmatrix}\begin{bmatrix} a_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} = \begin{bmatrix} a_{11} & A_{12} \\ 0 & \tilde{A}_{22} \end{bmatrix},$$

que implica

$$\begin{aligned}\begin{bmatrix} a_{11} & A_{12} \\ 0 & \tilde{A}_{22} \end{bmatrix}^{-1} &= A^{-1}\begin{bmatrix} 1 & 0 \\ A_{21}a_{11}^{-1} & I \end{bmatrix}, \\ &= \begin{bmatrix} b_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix}\begin{bmatrix} 1 & 0 \\ A_{21}a_{11}^{-1} & I \end{bmatrix},\end{aligned}$$

donde hemos particionado A^{-1} en la forma

$$A^{-1} = \begin{bmatrix} b_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix}.$$

Entonces, b_{11}, B_{12}, B_{21} y B_{22} son de entradas no negativas. Pero,

$$\begin{bmatrix} a_{11} & A_{12} \\ 0 & \tilde{A}_{22} \end{bmatrix}^{-1} = \begin{bmatrix} a_{11}^{-1} & -a_{11}^{-1} A_{12} \left(\tilde{A}_{22}\right)^{-1} \\ 0 & \left(\tilde{A}_{22}\right)^{-1} \end{bmatrix},$$

Así, $\left(\tilde{A}_{22}\right)^{-1} = B_{22} \geq 0$. Entonces, por hipótesis de inducción $\tilde{A}_{22}$ admite una descomposición incompleta LU que respeta el correspondiente patrón de dispersión heredado desde $\mathcal{Z}$ y tal que

$$\begin{aligned} \tilde{A}_{22} &= L_{22}U_{22} + R_{22}, \\ \tilde{A}_{22} &\leq L_{22}U_{22}, \\ 0 &\leq (L_{22}U_{22})^{-1}. \end{aligned}$$

Ahora, procedemos como en el teorema anterior y hacemos

$$\begin{aligned} L &= \begin{bmatrix} 1 & 0 \\ a_{11}^{-1}\tilde{A}_{21} & L_{22} \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0 \\ a_{11}^{-1}\tilde{A}_{21} & I \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & L_{22} \end{bmatrix} \\ &\text{y} \\ U &= \begin{bmatrix} a_{11} & \tilde{A}_{12} \\ 0 & U_{22} \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0 \\ 0 & U_{22} \end{bmatrix} \begin{bmatrix} a_{11} & \tilde{A}_{12} \\ 0 & I \end{bmatrix}, \end{aligned}$$

Entonces, como $\tilde{A}_{22} \leq L_{22}U_{22}$ se tiene $A_{22} \leq a_{11}^{-1}\tilde{A}_{21}\tilde{A}_{12} + L_{22}U_{22}$ y por tanto

$$A - LU = \begin{bmatrix} 0 & R_{12} \\ R_{21} & R_{22} \end{bmatrix} \leq 0.$$

Ahora,

$$LU = \begin{bmatrix} 1 & 0 \\ a_{11}^{-1}\tilde{A}_{21} & I \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & L_{22}U_{22} \end{bmatrix} \begin{bmatrix} a_{11} & \tilde{A}_{12} \\ 0 & I \end{bmatrix},$$

entonces

$$(LU)^{-1} = \begin{bmatrix} a_{11} & -\tilde{A}_{12} \\ 0 & I \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & (L_{22}U_{22})^{-1} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -a_{11}^{-1}\tilde{A}_{21} & I \end{bmatrix}.$$

Así, $(LU)^{-1}$ es producto de matrices no negrativas y por ende $(LU)^{-1} \geq 0$. ■

2.3.6 M-Matrices.

El principal objetivo de esta sección es introducir el concepto de M-Matriz y demostrar la existencia de la factorización incompleta para este tipo de matrices. El teorema que hace este trabajo fue primeramente probado en [13], siendo allí la primera vez que se hablaba de factorización incompleta. Presentamos a continuación la definición de M-Matriz y el teorema mencionado. Estableceremos dos demostraciones, una basada en el Corolario 2.3.9 recien demostrado que aunque no aparece demostrado en [20] como ya vimos se basa en la presentación allí hecha para matrices diagonalmente dominantes y los comentarios para M-Matrices que allí aparecen. La segunda la prueba es esencialmente la misma presentada en [13] y la incluimos no sólo por razones históricas sino porque se basa en mostrar un par de famosos resultados conocidos como el teorema de Perron-Frobenius y un lema debido a Varga [22], resultados estos que relacionan el radio espectral de matrices y la calidad de M-Matriz con la relación de orden parcial "$\leq$" recientemente introducida para matrices.

Definición 2.3.10 *Una matriz A de orden n se dice que es una $M-Matriz$ si satisface las siguientes condiciones*

$$\begin{aligned} &i)\ a_{ii} > 0,\ i = 1, 2, \cdots, n \\ &ii)\ a_{ij} \leq 0,\ i = 1, 2, \cdots, n,\ i \neq j \\ &iii)\ A \text{ es no singular y } A^{-1} \geq 0 \end{aligned}$$

Aunque la condición i) puede obtenerse de las otras dos la conservaremos en la definición para usarla libremente sin prueba.

Teorema 2.3.11 (Meijerink y Van der Vorst) *Si A es $M-Matriz$, entonces para todo patrón de dispersión $\mathcal{Z}$ la matriz A posee una factorización LU incompleta asociada a $\mathcal{Z}$ y que además satisface $LU \geq A$ y $(LU)^{-1} \geq 0$.*

Antes de probar el teorema estableceremos relaciones entre diagonalmente dominante y $M-Matrices$ a fin de usar los resultados anteriores.

M-Matrices, Jacobi, Gauss-Seidel y Factorización Incompleta

Empezamos estableciendo una relación entre las $M-Matrices$ y las matrices diagonalmente dominantes

Lema 2.3.12 *Sea M una $M-Matriz$, entonces M es diagonalmente semejante a una matriz A estrictamente diagonalmente dominante. Además, la matriz de semejanza diagonal D puede escogerse de modo que $D \geq 0$.*

Prueba. En efecto, sea d la solución al sistema $Md = e$, donde e es el vector con todas sus entradas iguales a uno. Entonces, $d = M^{-1}e$ y d_i sería la suma de las entradas de la $i-ésima$ fila de M^{-1}. Pero $M^{-1} \geq 0$, así $d > 0$.

Ahora, $Md = e > 0$, por tanto para cada $i = 1, 2, \cdots, n$

$$\begin{aligned} \sum_{j=1}^{n} m_{ij}d_j &> 0, \\ -\sum_{j\neq i} m_{ij}d_j &< m_{ii}d_i, \\ \sum_{j\neq i} \left|d_i^{-1}m_{ij}d_j\right| &< \left|d_i^{-1}m_{ii}d_i\right|, \end{aligned} \tag{2.34}$$

donde se ha usado que las entradas de d son positivas y que las m_{ij} son no positivas para $i \neq j$. Sea ahora $D = diag(d)$ la matriz diagonal con entradas en la diagonal principal iguales a las entradas de d. Entonces $D \geq 0$ y por la última desigualdad en (2.34) se tiene $A = D^{-1}MD$ es estrictamente diagonalmente dominante y diagonalmente semejante a M con matriz de semejanza $D \geq 0$. ■

Teorema 2.3.13 *Si la matriz M es una $M-Matriz$, entonces los métodos de Jacobi y Gauss-Seidel convergen a la solución de $Mx = b$ para cualquier solución inicial $x^{(0)}$.*

Prueba. Para mostrar la convergencia de estos métodos para la $M-matriz$ simplemente mostraremos que la correspondiente matriz de iteración Q satisface que $\rho(I - Q^{-1}M) < 1$.

Por el Lema 2.3.12 existe una matriz diagonal e invertible $D \geq 0$ tal que $A = D^{-1}MD$ es estrictamente diagonalmente dominante. Entonces, dado que matrices semejantes tienen el mismo polimomio característico han de

tener el mismo radio espectral y por tanto

$$\begin{aligned}\rho(I - Q^{-1}M) &= \rho(I - Q^{-1}DAD^{-1}) \\ &= \rho\left(DD^{-1} - DD^{-1}Q^{-1}DAD^{-1}\right) \\ &= \rho\left(D\left(I - D^{-1}Q^{-1}DA\right)D^{-1}\right) \\ &= \rho\left(I - \left(D^{-1}QD\right)^{-1}A\right).\end{aligned}$$

Ahora, es claro de la relación $A = D^{-1}MD$ que si Q es la matriz de iteración de Jacobi o Gauss-Seidel para M, entonces $\tilde{Q} = D^{-1}QD$ será respectivamente la matriz de iteración de Jacobi o de Gauss-Seidel para A. Pero como A es estrictamente diagonalmente dominante se sigue de los teoremas 2.2.8 y 2.2.9 que $\rho\left(I - \tilde{Q}^{-1}A\right) = \rho(I - Q^{-1}M) < 1$. ■

Prueba del Teorema de Meijerink y Van der Vorst. Sea M una $M - Matriz$ entonces M es diagonalmente semejante a una matriz A estrictamente diagonalmente dominante con matriz de semejanza $D \geq 0$. Entonces, A satisface las condiciones del corolario 2.3.9 y por tanto A posee una factorización LU incompleta que respeta el patrón de dispersión $\mathcal{Z}$ y tal que $A \leq LU$ y $(LU)^{-1} \geq 0$.

Ahora como

$$D^{-1}MD = A = LU + R,$$

se sigue que

$$M = \left(DLD^{-1}\right)\left(DUD^{-1}\right) + \left(DRD^{-1}\right),$$

es una factorización LU incompleta que respeta el patrón de dispersión y tal que

$$M \leq \left(DLD^{-1}\right)\left(DUD^{-1}\right) \text{ y}$$

y también

$$\begin{aligned}\left[\left(DLD^{-1}\right)\left(DUD^{-1}\right)\right]^{-1} &= \left[DLUD^{-1}\right]^{-1} \\ &= D\left[LU\right]^{-1}D^{-1} \geq 0.\end{aligned}$$

■

Teoría de Perron-Frobenius

Hasta el momento hemos trabajado con las llamadas normas matriciales subordinadas que son las que definimos a partir de una norma vectorial determinada. Sin embargo, es posible tener normas matriciales no provenientes de normas vectoriales, para poder utilizarlas concretaremos qué entenderemos entonces por norma matricial.

Definición 2.3.14 *Una norma matricial es una función* $\|\cdot\|$ *que a cada matriz cuadrada* A *le asigna un número real y para todo par de matrices cuadradas* A *y* B *satisface las siguientes condiciones*

1. $\|A\| \geq 0$ *y* $\|A\| = 0$ *si y sólo si* $A = 0$.
2. $\|cA\| = |c| \|A\|$, *para todo escalar* c.
3. $\|A + B\| \leq \|A\| + \|B\|$
4. $\|AB\| \leq \|A\| \|B\|$

En esta sección suponemos que $\|\cdot\|$ usado sobre matrices representa una norma matricial cualquiera, cuando se trate de subordinadas se indicará explícitamente. Además, $\|\cdot\|_F$ representará la conocida como norma de Frobenius para matrices la cual no es subordinada y está dada por

$$\|A\|_F = \left(\sum_{1 \leq i,j \leq n} |a_{ij}|^2 \right)^{1/2}.$$

Debe observarse que esta es la norma usual de $\mathbb{R}^{n^2}$ si consideramos A como un "*vector largo*" de $\mathbb{R}^{n^2}$. Es fácil probar que las normas matriciales y la norma de Frobenius recién definida son en efecto normas matriciales.

Empezamos con el siguiente

Teorema 2.3.15 *Sea* A *matriz de orden* n. *Entonces,* $\lim_{k \to +\infty} A^k = 0$ *si y sólo si* $\rho(A) < 1$.

Prueba. Supongamos primero que $\lim_{k \to +\infty} A^k = 0$ y sean λ un valor propio de A y v un vector propio asociado a λ. Entonces

$$A^k v = \lambda^k v \to 0, \text{ cuando } k \to +\infty.$$

Por tanto, debe ser $|\lambda| < 1$. Así, $\rho(A) < 1$.

Supongamos ahora que $\rho(A) < 1$, entonces por Teorema (2.2.4) existe una norma matricial subordinada $\|\cdot\|$ tal que

$$\|A\| < 1.$$

Para tal norma,

$$\left\| A^k \right\| \leq \|A\|^k \to 0, \text{ cuando } k \to +\infty.$$

Así, $\lim_{k \to +\infty} A^k = 0$. Debe recordarse que como todas las normas son equivalentes en $\mathbb{R}^{n^2}$, basta entonces mostrar el resultado para una norma cualquiera. ■

Corolario 2.3.16 *Sea $\|\cdot\|$ una norma matricial cualquiera (no necesariamente subordinada). Entonces para toda matriz A de orden n,*

$$\rho(A) = \lim_{k\to+\infty} \left\|A^k\right\|^{1/k}.$$

Prueba. Empezamos mostrando que

$$\rho(A) \leq \|A\|.$$

En efecto, sean λ un valor propio de A y v un vector propio asociado a λ. Construimos ahora la matriz V tal que todas sus columnas coinciden con v, entonces

$$\begin{aligned} |\lambda| \|V\| = \|\lambda V\| &= \|AV\| \\ &\leq \|A\| \|V\|. \end{aligned}$$

Así, $\rho(A) \leq \|A\|$.

Ahora dado que $\rho(A)^k = \rho(A^k)$ y por la propiedad recién probada se tiene que

$$\begin{aligned} \rho(A)^k &\leq \left\|A^k\right\|, \\ \rho(A) &\leq \left\|A^k\right\|^{1/k} \end{aligned}$$

para todo $k \in \mathbb{Z}^+$.

Sea ahora $\epsilon > 0$, entonces $\tilde{A} = [\rho(A) + \epsilon]^{-1} A$ satisface

$$\rho(\tilde{A}) = [\rho(A) + \epsilon]^{-1} \rho(A) < 1.$$

Por tanto, $\lim_{k\to+\infty} \tilde{A}^k = 0$ y así,

$$\left\|\tilde{A}^k\right\| \to 0, \text{ cuando } k \to +\infty.$$

Así, existe $K \in \mathbb{Z}^+$ tal que

$$\left\|\tilde{A}^k\right\| < 1, \text{ cuando } k \geq K.$$

Lo que implica,

$$\left\|A^k\right\| \leq [\rho(A) + \epsilon]^k, \text{ cuando } k \geq K.$$

O sea,

$$\rho(A) \leq \left\|A^k\right\|^{1/k} \leq \rho(A) + \epsilon, \text{ cuando } k \geq K.$$

■

Teorema 2.3.17 (Series de Neumann) *Si $\rho(A) < 1$, entonces $I - A$ es invertible y*

$$(I - A)^{-1} = \sum_{k=0}^{+\infty} A^k.$$

Prueba. Como $\rho(A) < 1$ existe una norma matricial subordinada tal que $\|A\| < 1$, escogemos una norma tal. Veamos primero que $I-A$ es no singular. De ser no invertible existiría un vector x tal que $\|x\| = 1$ y $(I - A)x = 0$, para tal x se tiene

$$1 = \|x\| = \|Ax\| \leq \|A\| \, \|x\| = \|A\| \,,$$

lo que contradice $\|A\| < 1$.

Veamos ahora que las sumas parciales de la serie de Neumman convergen a $(I - A)^{-1}$. Es decir, debemos probar que

$$\sum_{k=0}^{m} A^k \to (I - A)^{-1} \,, \text{ cuando } m \to +\infty.$$

O sea probar que

$$\left\| \sum_{k=0}^{m} A^k - (I - A)^{-1} \right\| \to 0, \text{ cuando } m \to +\infty,$$

Pero,

$$\begin{aligned}
\left\| \sum_{k=0}^{m} A^k - (I - A)^{-1} \right\| &\leq \left\| (I - A)^{-1} \right\| \left\| (I - A) \sum_{k=0}^{m} A^k - I \right\| \\
&\leq \left\| (I - A)^{-1} \right\| \left\| \sum_{k=0}^{m} \left(A^k - A^{k+1} \right) - I \right\| \\
&\leq \left\| (I - A)^{-1} \right\| \left\| \left(I - A^{m+1} \right) - I \right\| \\
&\leq \left\| (I - A)^{-1} \right\| \left\| A^{m+1} \right\| \\
&\leq \left\| (I - A)^{-1} \right\| \|A\|^{m+1}
\end{aligned}$$

y $\|A\|^{m+1} \to 0$ cuando $m \to +\infty$ por ser $\|A\| < 1$. ■

En este punto debemos recordar que por $|A|$ entendemos la matriz cuyas entradas son las magnitudes de la matriz A. Es muy fácil chequear por

inducción que para todo $k \in \mathbb{Z}^+$

$$\left|A^k\right| \leq |A|^k, \tag{2.35}$$
$$0 \leq A^k \leq B^k, \text{ si } 0 \leq A \leq B.$$

Usaremos estas propiedades para probar algunos de los resultados de la llamada teoría de Perron-Frobenius ver [7] ó [18] para más detalles.

Teorema 2.3.18 (Perron-Frobenius) *Si A y B son matrices de orden n tales que $|A| \leq B$, entonces $\rho(A) \leq \rho(|A|) \leq \rho(B)$.*

Prueba. De (2.35) es claro que $\left|A^k\right| \leq |A|^k \leq B^k$ por tanto la norma de Frobenius de estas matrices satisfacen

$$\left\|A^k\right\|_F^{1/k} \leq \left\||A|^k\right\|_F^{1/k} \leq \left\|B^k\right\|_F^{1/k},$$

entonces por (2.3.16) tomando límite tenemos

$$\rho(A) \leq \rho(|A|) \leq \rho(B).$$

■

Corolario 2.3.19 *Si A y B son matrices de orden n tales que $0 \leq A \leq B$, entonces $\rho(A) \leq \rho(B)$.*

Corolario 2.3.20 *Si A y B son matrices de orden n tales que $0 \leq A < B$, entonces $\rho(A) < \rho(B)$.*

Prueba. Como $0 \leq A < B$, entonces existe un real positivo α tal que $\alpha > 1$ y $0 \leq A \leq \alpha A < B$. Entonces,

$$\rho(A) < \alpha\rho(A) \leq \rho(B)$$

si $\rho(A) \neq 0$. Si $\rho(A) = 0$, el resultado es trivial. ■

Teorema 2.3.21 (Perron) *Si A es de orden n tal que $A \geq 0$, entonces $\rho(A)$ es un valor propio de A y existe un vector no negativo $v \geq 0$, con $\|v\| = 1$ tal que*

$$Av = \rho(A)v.$$

Los lemas siguientes los tomamos de [7].

Lema 2.3.22 (Fan) *Si $M = [m_{ij}]$ es una $M-Matriz$, entonces la matriz $M^{(1)}$ que se obtiene del primer paso de eliminación gaussiana eliminando la primera columna de M con su primera fila es nuevamente una $M-Matriz$.*

Prueba. Procedemos como en la prueba del corolario 2.3.9. El primer paso de eliminación gaussiana está dado por

$$\begin{bmatrix} 1 & 0 \\ -M_{21}m_{11}^{-1} & I \end{bmatrix} \begin{bmatrix} m_{11} & M_{12} \\ M_{21} & M_{22} \end{bmatrix} = \begin{bmatrix} m_{11} & M_{12} \\ 0 & M_{22} - M_{21}m_{11}^{-1}M_{12} \end{bmatrix} = M^{(1)}.$$

Es claro que por ser M una $M-Matriz$ las entradas no diagonales de $\tilde{M}_{22} = M_{22} - M_{21}m_{11}^{-1}M_{12}$ son no positivas. Además,

$$\begin{aligned} \left(M^{(1)}\right)^{-1} &= \begin{bmatrix} m_{11} & M_{12} \\ 0 & \tilde{M}_{22} \end{bmatrix}^{-1} \\ &= \begin{bmatrix} m_{11}^{-1} & -m_{11}^{-1}M_{12}\tilde{M}_{22}^{-1} \\ 0 & \tilde{M}_{22}^{-1} \end{bmatrix} \\ &= \begin{bmatrix} w_{11} & W_{12} \\ W_{21} & W_{22} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ M_{21}m_{11}^{-1} & I \end{bmatrix}, \end{aligned}$$

donde hemos escrito M^{-1} como

$$M^{-1} = \begin{bmatrix} w_{11} & W_{12} \\ W_{21} & W_{22} \end{bmatrix}.$$

Entonces, $\tilde{M}_{22}^{-1} = W_{22} \geq 0$ y así, $\left(M^{(1)}\right)^{-1} \geq 0$. Por tanto, $M^{(1)}$ es $M-Matriz$. ■

Lema 2.3.23 (Varga) *Si M y N son matrices de orden n tales que M es $M-Matriz$ y las entradas de N satisfacen*

$$0 < m_{ii} \leq n_{ii},\ m_{ij} \leq n_{ij} \leq 0 \text{ para } i \neq j,$$

Entonces N es también una $M-Matriz$.

Prueba. De la definición de $M-Matriz$ sólo nos falta verificar que $N^{-1} \geq 0$. Empezamos escribiendo M y N como,

$$\begin{aligned} M &= D_M - C_M \\ N &= D_N - C_N \end{aligned}$$

donde D_N y D_M son respectivamente las diagonales de M y de N. Entonces D_M, D_N, C_M y C_N tienen entradas no negativas y de la hipótesis satisfacen

$$0 \leq D_N^{-1} C_N \leq D_M^{-1} C_M.$$

Entonces en virtud del teorema (2.3.18) de Perron-Frobenius se tiene

$$\begin{aligned} \rho\left(D_N^{-1} C_N\right) &\leq \rho\left(D_M^{-1} C_M\right) \\ &= \rho\left(D_M^{-1}\left(D_M - M\right)\right) \\ &= \rho\left(I - D_M^{-1} M\right) \\ &< 1. \end{aligned}$$

La última igualdad se sigue de la prueba al teorema (2.3.13) de convergencia de los métodos de Jacobi y Gauss-Seidel para $M - Matrices$. Entonces, por el teorema (2.3.17) de Series de Neumann se sigue que $I - D_N^{-1} C_N$ es no singular y

$$\left(I - D_N^{-1} C_N\right)^{-1} = \sum_{k=0}^{+\infty} \left(D_N^{-1} C_N\right)^k \geq 0,$$

al ser cada sumando no negativo. Entonces, $N = D_N\left(I - D_N^{-1} C_N\right)$ es no singular y su inversa es $N^{-1} = \left(I - D_N^{-1} C_N\right)^{-1} D_N^{-1} \geq 0$. ■

Daremos ahora la prueba clásica al teorema de Meijerink y Van der Vorst.

Prueba Clásica del Teorema de Meijerink y Van der Vorst. Debemos motrar que si M es una $M - Matriz$ y $\mathcal{Z}$ es un patrón de dispersión dado entonces M posee una factorización incompleta que respeta el patrón de dispersión y tal que $M \leq LU$ y $(LU)^{-1} \geq 0$. Empezamos escribiendo $M^{(0)} = M$,

$$M^{(1/2)} = M - R^{(0)},$$

donde $M^{(1/2)}$ se obtiene haciendo cero las entradas de M de la primera fila y de la primera columna que estén en el patrón de dispersión $\mathcal{Z}$. Entonces, $M^{(0)}$ satisface las hipótesis del Lema de Varga y es por ende $M - Matriz$. Aplicamos ahora un paso de eliminación gaussiana como en el Lema de Fan y obtenemos una matriz $M^{(1)}$ que es nuevamente $M - Matriz$. Entonces,

$$M^{(1)} = \begin{bmatrix} 1 & 0 \\ -M_{21}^{(1/2)} \left(m_{11}^{(1/2)}\right)^{-1} & I \end{bmatrix} M^{(1/2)}.$$

Escribimos ahora,

$$M^{(1+1/2)} = M^{(1)} - R^{(1)}$$

y ahora $M^{(1+1/2)}$ se obtiene haciendo ceros las entradas no nulas de $M^{(1)}$ que estén en el patrón de dispersión y estén sobre la segunda fila o sobre la segunda columna. Nuevamente por Lema de Varga $M^{(1+1/2)}$ es una $M - Matriz$ y del Lema de Fan se sigue que si hacemos eliminación gaussiana de la segunda columna obtendremos una $M - Matriz$, así

$$M^{(2)} = \begin{bmatrix} 1 & & \\ & 1 & \\ & -M_{32}^{(1+1/2)} \left(m_{22}^{(1+1/2)}\right)^{-1} & I \end{bmatrix} M^{(1+1/2)}.$$

Este proceso puede claramente continuarse, teniendose en el paso k,

$$M^{(k-1/2)} = M^{(k-1)} - R^{(k-1)},$$

con $M^{(k-1/2)}$ obtenido haciendo cero las entradas de la $k - ésima$ fila y la $k - ésima$ columna de $M^{(k-1)}$ que estén en el patrón de dispersión y luego

$$M^{(k)} = L^{(k)} M^{(k-1/2)}, \text{ con}$$

$$L^{(k)} = \begin{bmatrix} I_{k-1} & & \\ & 1 & \\ & -M_{k+1,k}^{(k-1/2)} \left(m_{kk}^{(k-1/2)}\right)^{-1} & I_{n-k} \end{bmatrix}.$$

En este punto hacemos varias observaciones. Primero que despues de $n-1$, $M^{(n-1)}$ será triangular superior y respetará el patrón de dispersión. Ahora,

$$\begin{aligned} M^{(k)} &= L^{(k)} M^{(k-1/2)} \\ &= L^{(k)} \left(M^{(k-1)} - R^{(k-1)}\right) \\ &= L^{(k)} L^{(k-1)} M^{(k-1-1/2)} - L^{(k)} R^{(k-1)} \\ &= L^{(k)} L^{(k-1)} \left(M^{(k-2)} - R^{(k-2)}\right) - L^{(k)} R^{(k-1)} \\ &= \left(\prod_{i=0}^{k-1} L^{(k-i)}\right) M - \sum_{j=0}^{k-1} \left(\prod_{i=0}^{k-1-j} L^{(k-i)}\right) R^{(j)}. \end{aligned}$$

De otra parte es fácil comprobar por simple cálculo que $L^{(i)} R^{(j)} = R^{(j)}$ si $i \leq j$. Esto se debe a que $L^{(i)}$ se comporta en este caso como una identidad dada su similitud estructural con esta matriz y a lo rala (esparcida) que es la matriz $R^{(j)}$ al tener pocas entradas no nulas localizadas en la parte derecha

de la $j-ésima$ fila y la parte inferior de la $j-ésima$ columna. Podemos entonces escribir,

$$\begin{aligned} M^{(k)} &= \left(\prod_{i=0}^{k-1} L^{(k-i)}\right) M - \sum_{j=0}^{k-1} \left(\prod_{i=0}^{k-1-j} L^{(k-i)}\right) R^{(j)} \\ &= \left(\prod_{i=0}^{k-1} L^{(k-i)}\right) \left(M - \sum_{j=0}^{k-1} R^{(j)}\right). \end{aligned}$$

Ahora hacemos $U = M^{(n-1)}$, $R = \sum_{j=0}^{n-2} R^{(j)}$ tenemos que U y R respetan el patron de dispersión $R \leq 0$ y

$$U = \left(\prod_{i=0}^{n-2} L^{(n-1-i)}\right)(M - R).$$

Definimos ahora,

$$L = \left(\prod_{i=0}^{n-2} L^{(n-1-i)}\right)^{-1}, \text{ entonces}$$

$$L = \prod_{i=0}^{n-2} \left(L^{(i+1)}\right)^{-1}.$$

Además, es fácil chequear que

$$\begin{aligned} &\left(L^{(k)}\right)^{-1} \left(L^{(k+1)}\right)^{-1} \\ &= \begin{bmatrix} I_{k-1} & & & \\ & 1 & & \\ & & 1 & \\ & M_{k+1,k}^{(k-1/2)} \left(m_{kk}^{(k-1/2)}\right)^{-1} & M_{k+2,k+1}^{(k+1/2)} \left(m_{k+1,k+1}^{(k+1/2)}\right)^{-1} & I_{n-k-1} \end{bmatrix}. \end{aligned}$$

Por tanto, L satisface el patrón de dispersión. Ahora, como $U = M^{(n-1)}$, entonces U es $M-Matriz$ por Lema de Fan, además cada $L^{(k)}$ es de entradas no negativas y $R \leq 0$ se sigue entonces que:

$$\begin{aligned} (LU)^{-1} = U^{-1}L^{-1} &\geq 0, \\ M = LU + R \text{ y} \\ M &\leq LU. \end{aligned}$$

■

Corolario 2.3.24 (Factorización Incompleta de Cholesky) *Si M es una $M-Matriz$ simétrica, entonces para cada patrón de dispersión $\mathcal{Z}$ tal que $(i,j) \in \mathcal{Z}$ implica $(j,i) \in \mathcal{Z}$, existe una única matriz triangular inferior L que satisface $l_{ij} = 0$ si $(i,j) \in \mathcal{Z}$ y tal que $M = LL^T + R$ donde $r_{ij} = 0$ si $(i,j) \notin \mathcal{Z}$, $M \leq LL^T$ y $(LL^T)^{-1} \geq 0$.*

Prueba. Por teorema anterior M posee una única factorización LU incompleta. Es decir, existe matrices L_0, U_0 y R_0 únicas tales que L_0 es triangular inferior, U_0 triangular superior, respetan el patrón de dispersión y

$$\begin{aligned} M &= L_0U_0 + R_0 \\ M &\leq L_0U_0, \ \ (L_0U_0)^{-1} \geq 0. \end{aligned}$$

Sea ahora $D = diag(d_1, \cdots, d_n)$ la matriz diagonal con elementos en la diagonal principal los de la diagonal principal de U_0. Claramente D es invertible, pues al serlo L_0 y L_0U_0 también lo es U_0. Además la simetría del patrón de dispersión y de la matriz M implica que R_0 es simétrica y con ello

$$\begin{aligned} M = M^T &= (L_0U_0)^T + R_0 \\ &= \left(L_0DD^{-1}U_0\right)^T + R_0 \\ &= \left(D^{-1}U_0\right)^T (L_0D)^T + R_0. \end{aligned}$$

En este punto basta observar que gracias a la simetría del patrón de dispersión las matrices $L_1 = \left(D^{-1}U_0\right)^T$ y $U_1 = (L_0D)^T$ satisfaces las condiciones de la factorización LU incompleta, por lo que debe ser

$$\begin{aligned} L_0 &= \left(D^{-1}U_0\right)^T \\ &= U_0^T \left(D^{-1}\right)^T \\ &= U_0^T \left(D^{-1}\right). \end{aligned}$$

Esta iguadad implica que para cada pareja (i,j) se tenga

$$l_{ij}d_i = u_{ji}.$$

De donde,

$$0 < m_{ii} = \sum_{j=1}^{n} l_{ij}u_{ji} = d_i \sum_{j=1}^{n} l_{ij}^2.$$

Así, cada d_i es positivo y por tanto podemos escribir

$$\begin{aligned} M &= \left(D^{-1}U_0\right)^T (L_0 D)^T + R_0 \\ &= \left(D^{-1}U_0\right)^T (L_0 D)^T + R_0 \\ &= (U_0)^T \left(D^{-1}\right)^T U_0 + R_0 \\ &= (U_0)^T \left(D^{-1/2}\right)^T D^{-1/2} U_0 + R_0 \\ &= \left(D^{-1/2}U_0\right)^T \left(D^{-1/2}U_0\right) + R_0. \end{aligned}$$

Tomando entonces $L = \left(D^{-1/2}U_0\right)^T$ se obtiene la buscada Factorización de Cholesky. ■

2.4 Precondicionadores por Bloques

En las aplicaciones, frecuentemente las matrices deben considerarse por bloques. De hecho, las estrategias anteriores para producir precondicionadores por factorización incompleta y por los métodos iterativos clásicos se pueden adaptar para cuando la matriz A se particiona por bloques. Una fuente de precondicionadores con estructura de bloque se origina cuando la matriz del sistema tiene la forma denominada **KKT** (Karush - Kuhn - Tucker) o de punto de ensilladura. Es decir, matrices *no singulares* que tienen la siguiente estructura:

$$\begin{pmatrix} A & B^T \\ C & 0 \end{pmatrix}, \tag{2.36}$$

donde $A \in \mathbb{R}^{n\times n}$ y $B, C \in \mathbb{R}^{m\times n}$ con $n \geq m$. En muchas aplicaciones A es simétrica y $B = C$, en cuyo caso (2.36) sería simétrica; en todo caso, si A es o no simétrica, la matriz (2.36) es generalmente indefinida, es decir, las partes reales de sus valores propios son positivas o negativas. Matrices con esta estructura se presentan frecuentemente en aplicaciones: problemas de optimización, problemas de fluidos, problemas de estática magnética, etc. Murphy, Golub y Wathen muestran en [14] un precondicionador bastante eficiente para sistemas lineales cuya matriz de coeficientes tiene la forma (2.36). Dicho precondicionador tiene la siguiente estructura cuando la matriz A es invertible:

$$P = \begin{pmatrix} A & 0 \\ 0 & CA^{-1}B^T \end{pmatrix} \tag{2.37}$$

Este precondicionador fue generalizado posteriormente por Ipsen en [8] para matrices *no singulares* de la forma $\begin{pmatrix} A & B^T \\ C & D \end{pmatrix}$, donde $A \in \mathbb{R}^{n\times n}$, B y $C \in \mathbb{R}^{m\times n}$ y $D \in \mathbb{R}^{m\times m}$con $n \geq m$. Este precondicionador fue utilizado recientemente en la solución de sistemas lineales provenientes de la discretización de las ecuaciones linealizadas de Navier-Stokes [9] con muy buenos resultados.

Tanto en [14] como en [8], los autores dejan sin demostrar varias proposiciones. En [2] se pueden encontrar pruebas completas de todas ellas. Enseguida presentamos los aspectos más relevantes de este estudio utilizando las pautas propuestas en [3]. Empezamos con un teorema que establece una importante relación entre la dimensión del subespacio de Krylov y el grado del polinomio minimal de la matriz del sistema.

Teorema 2.4.1 *Sea A una matriz de orden n, m el grado del polinomio minimal de A y $K_k(A, r_0) = gen\left\{r_0, Ar_{0,}..., A^{k-1}r_0\right\}$. Entonces la dimensión de $K_k(A, r_0)$ es $min\{m, k\}$.*

Prueba

Supongamos que $m < k$. Entonces $A^m r_0$ es el primer término de la sucesión $r_0, Ar_{0,}..., A^{k-1}r_0$ que es combinación lineal de los anteriores. Por tanto

$$\left\{r_0, Ar_{0,}..., A^{m-1}r_0\right\}$$

es linealmente independiente.

Ahora probemos que

$$gen\left\{r_0, Ar_0, ..., A^{k-1}r_0\right\} = gen\left\{r_0, Ar_0, ..., A^{m-1}r_0\right\}$$

En efecto: como $m < k$ es evidente que

$$gen\left\{r_0, Ar_0, ..., A^{m-1}r_0\right\} \subset gen\left\{r_0, Ar_0, ..., A^{k-1}r_0\right\}$$

Si $y \in gen\left\{r_0, Ar_0, ..., A^{k-1}r_0\right\}$ entonces $y = \alpha_0 r_0 + \alpha_1 Ar_0 + ... + \alpha_{k-1}A^{k-1}r_0$ pero todos los términos de la forma $A^l r_0$, con $l > m-1$, por el razonamiento anterior, son combinación lineal de $r_0, Ar_0, ..., A^{m-1}r_0$. Entonces

$$gen\left\{r_0, Ar_0, ..., A^{k-1}r_0\right\} \subset gen\left\{r_0, Ar_0, ..., A^{m-1}r_0\right\}$$

Por tanto

$$dim\left(K_k\left(A, r_0\right)\right) = m$$

Si $m \geq k$, entonces $\{r_0, Ar_0, ..., A^{k-1}r_0\}$ es linealmente independiente y por consiguiente

$$dim\left(K_k\left(A, r_0\right)\right) = k$$

La primera proposición que presentamos afirma que el precondicionador a izquierda (2.37) genera una matriz precondicionada con polinomio minimal de grado a lo más 4.

Proposición 2.4.2 *Si la matriz* $\mathcal{A} = \begin{pmatrix} A & B^T \\ C & 0 \end{pmatrix}$, *donde* $A \in \mathbb{R}^{n\times n}$ *es no singular y* $B, C \in \mathbb{R}^{m\times n}$ *con* $n \geq m$, *se precondiciona con (2.37), entonces la matriz precondicionada* $T = P^{-1}\mathcal{A}$ *satisface*

$$T\left(T - I\right)\left(T^2 - T - I\right) = 0.$$

Prueba

Como $P = \begin{pmatrix} A & 0 \\ 0 & CA^{-1}B^T \end{pmatrix}$, entonces $P^{-1} = \begin{pmatrix} A^{-1} & 0 \\ 0 & (CA^{-1}B^T)^{-1} \end{pmatrix}$ y

$$T = P^{-1}\mathcal{A} = \begin{pmatrix} I & A^{-1}B^T \\ (CA^{-1}B^T)^{-1}C & 0 \end{pmatrix}.$$

Además,

$$\left(T - \tfrac{1}{2}I\right)^2 - \tfrac{1}{4}I = \begin{pmatrix} A^{-1}B^T(CA^{-1}B^T)^{-1}C & 0 \\ 0 & I \end{pmatrix}.$$

Como la matriz en la posición $(1,1)$ por bloques es idempotente, se tiene que

$$\left(\left(T - \frac{1}{2}I\right)^2 - \frac{1}{4}I\right)^2 = \left(T - \frac{1}{2}I\right)^2 - \frac{1}{4}I$$

y asi se llega a

$$T\left(T - I\right)\left(T - \frac{1+\sqrt{5}}{2}I\right)\left(T - \frac{1-\sqrt{5}}{2}I\right) = 0.$$

La segunda proposición afirma algo similar del polinomio minimal de la matriz precondicionada cuando el precondicionamiento se toma a derecha o a derecha e izquierda.

Proposición 2.4.3 *Si* $T = \mathcal{A}P^{-1}$ *o si* $T = P_1^{-1}\mathcal{A}P_2^{-1}$ *con* $P_1P_2 = P$ *y* $\mathcal{A}$ *como en la proposición anterior, entonces también se satisface que*

$$T\left(T-I\right)\left(T^2-T-I\right)=0.$$

La proposición que se presenta a continuación es referenciada sin demostración en [14]. Indica que métodos iterativos basados en subespacios de Krylov, como *minres*, convergen en a lo más 4 iteraciones, suponiendo que se puede trabajar con *matemática exacta*.

Proposición 2.4.4 *Para cualquier vector* r*, el subespacio de Krylov*

$$gen\{r,\ Tr,\ T^2r,\ T^3r,....\}$$

con T *como en la proposición anterior, es a lo más de dimensión 3 si* T *es no singular (o* 4 *si* T *es singular).*

Prueba

Sin pérdida de generalidad supondremos que T es singular. Como

$$Q\left(t\right)=t\left(t-1\right)\left(t-\frac{1+\sqrt{5}}{2}\right)\left(t-\frac{1-\sqrt{5}}{2}\right) \tag{2.38}$$

es un polinomio mónico que anula a T, entonces el polinomio minimal de T tiene a lo más grado 4. Usando el Teorema 2.4.1, obtenemos que la dimensión de

$$gen\{r,\ Tr,\ T^2r,\ T^3r,....\} \tag{2.39}$$

es a lo sumo 4 . Si T es no singular, el polinomio (2.38) tendría a lo sumo grado 3 y por tanto el subespacio en (2.39) tiene a lo más dimensión 3.

El precondicionador trabajado en las proposiciones anteriores puede ser extendido a matrices no singulares de la forma

$$\mathcal{A}=\begin{pmatrix} A & B^T \\ C & D \end{pmatrix}, \tag{2.40}$$

donde la matriz A es invertible.

Proposición 2.4.5 *Sea* $\mathcal{A}=\begin{pmatrix} A & B^T \\ C & D \end{pmatrix}$*, donde* $A\in\mathbb{R}^{n\times n}$ *es invertible,* B *y* $C\in\ \mathbb{R}^{m\times n}, D\in\mathbb{R}^{m\times m}$*,* $n\geq m$*,* $C\neq 0$ *y* $P=\begin{pmatrix} A & B^T \\ 0 & S \end{pmatrix}$*, donde* $S=D-CA^{-1}B^T$ *es el Complemento de Schur de* A*. Entonces* $\mathcal{A}P^{-1}$ *y* $P^{-1}\mathcal{A}$ *tienen como polinomio minimal a* $Q\left(t\right)=\left(t-1\right)^2$*.*

Prueba

Como $P = \begin{pmatrix} A & B^T \\ 0 & S \end{pmatrix}$ entonces $P^{-1} = \begin{pmatrix} A^{-1} & -A^{-1}B^TS^{-1} \\ 0 & S^{-1} \end{pmatrix}$ y se sigue que

$$\mathcal{A}P^{-1} - I = \begin{pmatrix} 0 & 0 \\ CA^{-1} & 0 \end{pmatrix},$$
$$\left(\mathcal{A}P^{-1} - I\right)^2 = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}.$$

Nótese que $T = \mathcal{A}P^{-1}$ es distinta de la identidad pues CA^{-1} no es la matriz nula. Luego su polinomio minimal no es $(t-1)$ sino que debe ser $Q(t) = (t-1)^2$.

Por otro lado

$$\begin{aligned} &\left(P^{-1}\mathcal{A} - I\right)^2 \\ &= \left(P^{-1}\mathcal{A}P^{-1}P - P^{-1}P\right)^2 \\ &= \left[P^{-1}\left(\mathcal{A}P^{-1} - I\right)P\right]^2 \\ &= P^{-1}\left(\mathcal{A}P^{-1} - I\right)^2 P \\ &= P^{-1}0P = 0 \end{aligned}$$

De nuevo observamos que $P^{-1}\mathcal{A} - I$ es distinta de cero pues

$$P^{-1}\mathcal{A} - I = P^{-1}\left(\mathcal{A}P^{-1} - I\right)P$$

y la matriz $\mathcal{A}P^{-1} - I$ no es nula. Por tanto el polinomio minimal de $T = P^{-1}\mathcal{A}$ es $Q(t) = (t-1)^2$.

La siguiente proposición, también de [8], es una generalización de la proposición 2.4.3 propuesta arriba.

Proposición 2.4.6 *Si* $P_1 = \begin{pmatrix} I & 0 \\ CA^{-1} & -I \end{pmatrix}$, $P_2 = \begin{pmatrix} A & B^T \\ 0 & S \end{pmatrix}$ *y* $\mathcal{A}$ *son como en la proposición anterior, entonces* $P_1^{-1}\mathcal{A}P_2^{-1} = \begin{pmatrix} I & 0 \\ 0 & -I \end{pmatrix}$. *Además la matriz* $P = P_1P_2 = \begin{pmatrix} A & B^T \\ C & D - 2S \end{pmatrix}$ *es tal que* $P^{-1}\mathcal{A}$ *tiene como polinomio minimal a* $Q(t) = (t-1)(t+1)$.

La proposición que se presenta a continuación aparece referenciada en [8] pero sin demostración.

Proposición 2.4.7 *Si* $\mathcal{A} = \begin{pmatrix} A & B^T \\ C & D \end{pmatrix}$, *con* A, B, C *y* D *como en la proposición 2.4.5, entonces el precondicionador* $P = \begin{pmatrix} A & 0 \\ 0 & -S \end{pmatrix}$, *donde* S *es el complemento de Schur de* A, *es tal que la matriz precondicionada es* $T = \mathcal{A}P^{-1} = \begin{pmatrix} I & -B^T S^{-1} \\ CA^{-1} & -DS^{-1} \end{pmatrix}$. *Si* $\mathcal{A}$ *es de la forma KKT, es decir* $D = 0$, *entonces* $T^2 - T = \begin{pmatrix} -B^T S^{-1} C A^{-1} & 0 \\ 0 & I \end{pmatrix}$. *Puesto que* $\left(T^2 - T\right)^2 = T^2 - T$, *entonces el polinomio minimal de* T *es a lo más de grado 4.*

Las proposiciones anteriores permiten concluir que para matrices de la forma $\mathcal{A} = \begin{pmatrix} A & B^T \\ C & 0 \end{pmatrix}$, donde $A \in \mathbb{R}^{n\times n}$ y $B, C \in \mathbb{R}^{m\times n}$ con $n \geq m$ que por lo general son indefinidas, los precondicionadores $P = \begin{pmatrix} A & 0 \\ 0 & \pm CA^{-1}B^T \end{pmatrix}$ y $P = \begin{pmatrix} A & B^T \\ 0 & \pm CA^{-1}B^T \end{pmatrix}$logran que el polinomio minimal de $P^{-1}\mathcal{A}$ tenga a lo más 4 raíces distintas, lo cual implica que, bajo aritmética exacta, métodos Krylov como el **minres** al ser aplicados al sistema lineal precondicionado converja en a lo sumo 4 iteraciones.

Conclusión similar se puede obtener para matrices de la forma $\mathcal{A} = \begin{pmatrix} A & B^T \\ C & D \end{pmatrix}$, donde $A \in \mathbb{R}^{n\times n}$, B y $C \in \mathbb{R}^{m\times n}$.y $D \in \mathbb{R}^{m\times m}$con $n \geq m$, $C \neq 0$ con precondicionadores $P = \begin{pmatrix} A & B^T \\ 0 & \pm S \end{pmatrix}$, $P = \begin{pmatrix} A & B^T \\ C & D - 2S \end{pmatrix}$ donde $S = D - CA^{-1}B^T$.

Capítulo 3

Resultados Numéricos

3.1 Introducción

En este capítulo se muestra una selección de problemas en la que se resuelven diversos sistemas lineales de interés dada su alta ocurrencia en problemas de ciencia e ingeniería. Los ejemplos buscan ilustrar las dos más grandes fuentes de problemas lineales propicios para la aplicación de métodos iterativos gracias a los patrones de dispersión de las matrices y su gran tamaño: estamos hablando de la solución numérica de ecuaciones diferenciales.

El primer ejemplo ilustra una de esas fuentes, la estrategia en este caso consiste en aproximar la solución al problema como una combinación lineal de un conjunto finito de n funciones base. Esto reduce entonces el problema a encontrar un número finito de coeficientes para la combinación. Luego se utilizan las condiciones de contorno y/o iniciales para obtener una cantidad finita, digamos m, de condiciones adicionales que debe cumplir la solución buscada. Se obtiene así un problema con n incógnitas y m condiciones. Dentro de esta técnica se encuentran los Elementos Finitos que es el procedimiento utilizado en el primer ejemplo para resolver un problema lineal elíptico en una región circular.

El segundo ejemplo muestra una forma distinta de abordar el problema y consiste en renunciar a encontrar la solución a la ecuación y se concentra en aproximar la solución al problema en un número finito n de puntos en el dominio de definición. Luego los operadores de diferenciación son remplazados por versiones discretas por diferencias finitas. Aplicando las condiciones de contorno y/o iniciales se consiguen el número de restricciones adicionales que garantizan la existencia de solución. Este procedimiento, llamado solución por diferencias finitas, es mostrado en el segundo ejemplo en la solución de

un problema lineal elíptico en una región rectangular.

El tercer y último ejemplo muestra una implementación de las técnicas de precondicionamiento por bloques. Se hacen comparaciones contra métodos sin precondicionamiento y se estudia la relación entre los tamaños de los bloques interiores y el desempeño final del algoritmo.

Todas las pruebas fueron efectuadas utilizando *Matlab 7.0 Service Pack 1* bajo *Windows XP Service Pack 2,* corriendo en un equipo con procesador *Pentium IV* de 3.06 GHz con Hyper Threading y 512 Mb de Memoria RAM. Se emplearon directamente las rutinas precompiladas que trae Matlab para los métodos de Krylov aquí estudiados. Los métodos estacionarios si fue preciso programarlos por parte de los autores, pero siempre bajo el mismo entorno de Matlab. En las tablas de los ejemplos la columna Flg muestra la razón de terminación del algoritmo así: 0 significa convergencia obtenida, 1 significa máximo de iteraciones alcanzado sin obtener convergencia, 3 significa que dos iteraciones consecutivas son muy cercanas y se suspende el trabajo. Todas las matrices involucradas en los procesos se manejaron en tipo $Sparse$ de Matlab, excepto en el caso por bloques donde se preparó una rutina que evalúa la acción de la matriz a partir de los bloques.

3.2 Elementos Finitos

En este ejemplo resolvemos un problema elíptico en el disco unidad utilizando elementos finitos. Para esto nos basamos en adaptaciones propias al conjunto de programas ofrecidos en [4]. El paquete en mención permite a partir de una triangulación de la región que debe ser programada por el usuario obtener el sistema lineal asociado a problemas elípticos tipicos junto con rutinas que permita una buena visualización de los resultados obtenidos. Nuestra adaptación consistió en escribir una triangulación para el círculo, que si bien no cumple criterios de optimalidad conocidos para la generación de mallas es útil como ejemplo de la versatilidad del método. También fué preciso adaptar las rutinas de Gockenbach a la generalidad de nuestro problema que es del tipo

$$\begin{cases} -\nabla \cdot (a(x,y)\nabla u) = f, & (x,y) \in \Omega \\ u = g, & (x,y) \in \partial\Omega \end{cases} \tag{3.1a}$$

con $a(x,y)$ positiva en el dominio Ω.

El problema específico que consideraremos es

$$\begin{cases} -\nabla \cdot \left(\left(1+x^2\right)\nabla u\right) = -2x(2xy+y^2) - 2(1+x^2)y - 2(1+x^2)x, & (x,y) \in \mathbb{D} \\ u = x^2y + xy^2, & (x,y) \in \partial\mathbb{D} \end{cases}$$

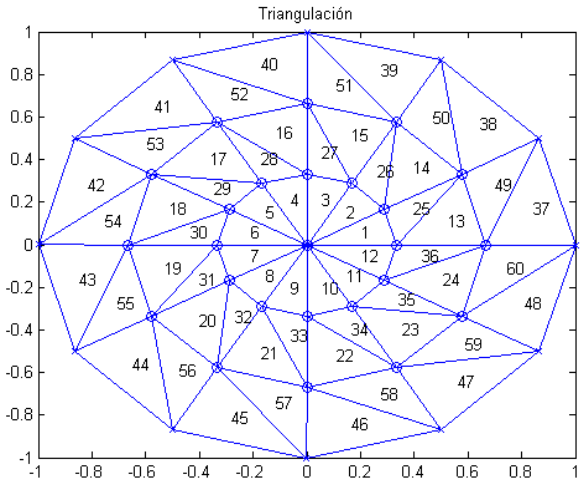

Figura 3.1: Triangulación de Círculo para $n = 3$.

Se trata de un problema fabricado a partir del conocimiento de u. Su solución exacta es $u(x, y) = x^2y + xy^2$. El objetivo es poder contar con una expresión que permita evaluar el error cometido. Por su puesto esto no es posible en los problemas reales, pero este es sólo un ejemplo ilustrativo.

La triangulación del círculo que utilizamos se basa en la escogencia de un entero positivo n que se utiliza para subdividir el radio en n partes iguales, luego cada cuadrante se divide en n sectores circulares. Obteniendose así $4n$ triangulos con vértice común en el origen y $4n(n-1)$ sectores trapezoidales. Luego cada uno de estos sectores los dividimos en dos triangulos, para un total de $4n(2n-1)$ triangulos. Obtenemos entonces un triangulación de un poligono regular de $4n$ lados inscrito en el círculo. En esta triangulación, el número total de nodos es $4n^2+1$ de los cuales $4n$ están en frontera y resto son nodos libres. Por tanto el tamaño del sistema resultante será $4n(n-1)+1$. Ilustramos la triangulación para el caso $n = 3$ en la figura 3.1.

La solución obtenida y error cometido en la utilización de esta triangulación para resolver el problema con $n = 20$ se muestra en las figuras 3.2 y 3.3.

Dado que en el caso $n = 20$ el sistema es de sólo 1521 incognitas y a fín de comparar el desempeño de los métodos estudiados, se consideró el problema anterior para distintos valores de n. En las tablas 3.1, 3.2 y 3.3 se resumen los resultados obtenidos por los distintos métodos para $n = 50$. En

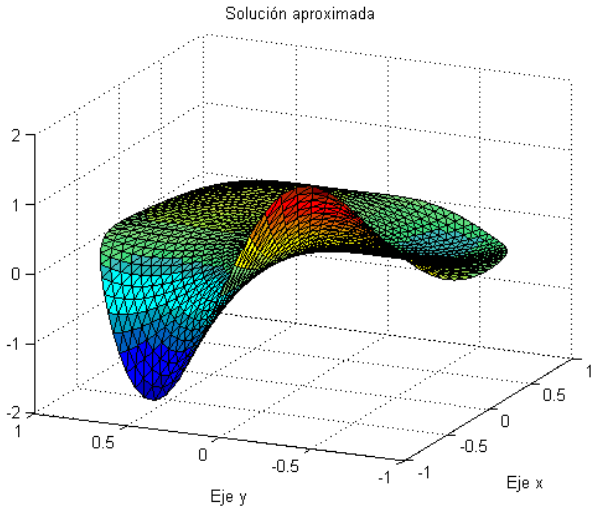

Figura 3.2: Solución Encontrada con $n = 128$.

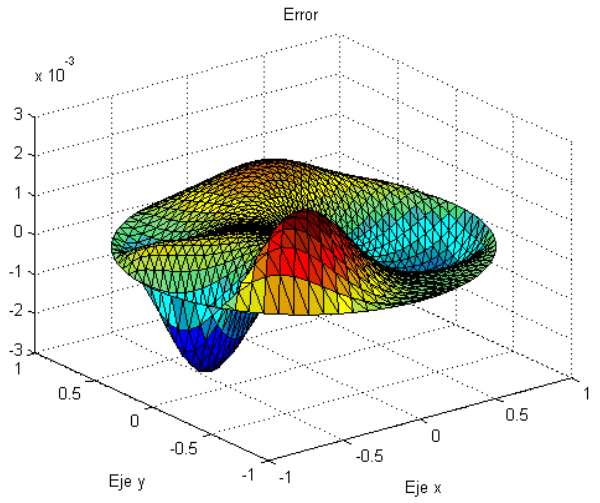

Figura 3.3: Error en la solución encontrada para $n = 128$.

Método	*Iter*	*Res Rel*	*Err Rel*	*Err Inf*	*Flg*	*t(seg)*
Jacobi	2000	1.7264e-04	2.1113e-02	1.7038e-02	1	25.61
$Gauss - Seidel$	2000	3.0178e-05	4.1254e-03	3.0417e-03	1	35.79
$Sor(w = 0.1)$	2000	9.5615e-03	4.4043e-01	4.7402e-01	1	36.00
$Sor(w = 0.5)$	2000	5.3328e-04	5.0172e-02	4.4510e-02	1	36.12
$Sor(w = 1.5)$	1383	3.2581e-06	3.8917e-04	3.9932e-04	3	25.35
$Sor(w = 1.7)$	848	2.1114e-06	3.3173e-04	3.7504e-04	3	15.28

Tabla 3.1: Mtodos Estacionarios en Ejemplo 01 con Tolerancia $1e-6$

Método	*Iter*	*Res Rel*	*Err Rel*	*Err Inf*	*Flg*	*t(seg)*
CG	996	9.5583e-09	2.9832e-04	3.5710e-04	0	6.25
LSQR	2000	2.1226e-02	7.2658e-01	9.2435e-01	1	26.28
$GMRES(10)$	200 10	3.1088e-04	3.4304e-02	2.9670e-02	1	26.14
$GMRES(50)$	40 50	1.0752e-06	3.2757e-04	3.9541e-04	1	55.17
$GMRES(100)$	20 59	9.9607e-09	2.9783e-04	3.5724e-04	0	82.64

Tabla 3.2: Mtodos No Estacionarios en Ejemplo 01 con Tolerancia $1e-8$

tal caso el tamaño del sistema se incrementa a 9801 incognitas y ecuaciones. El título de cada tabla indica la tolerancia y el tipo de métodos a listar.

3.3 Diferencias Finitas

El problema a resolver es propuesto como ejercicio en [10] y pide utilizar una discretización del dominio de interés junto con diferencias finitas centradas para reducir a un sistema de ecuacione lineales el problema

$$\begin{cases} -u_{xx} - u_{yy} + e^{x+y}u = 1, & 0 < x, y < 1, \\ u(x,0) = u(x,1) = u(1,0) = 0, u(0,y) = 1 & 0 < x, y < 1. \end{cases} \tag{3.2}$$

Pide además utilizar el sistema lineal resultante como problema de prueba para comparar el desempeño de gradiente conjugado con y sin precondi-

Método	*Iter*	*Res Rel*	*Err Rel*	*Err Inf*	*Flg*	*t(seg)*
CG	996	9.5583e-09	2.9832e-04	3.5710e-04	0	6.25
PCG, *Jacobi*	436	9.7786e-09	2.9836e-04	3.5703e-04	0	4.73
PCG, $Chol(6)$	2	1.0278e-10	2.9836e-04	3.5703e-04	0	7.65

Tabla 3.3: CG Precondicionado en Ejemplo 01 con Tolerancia $1e-8$

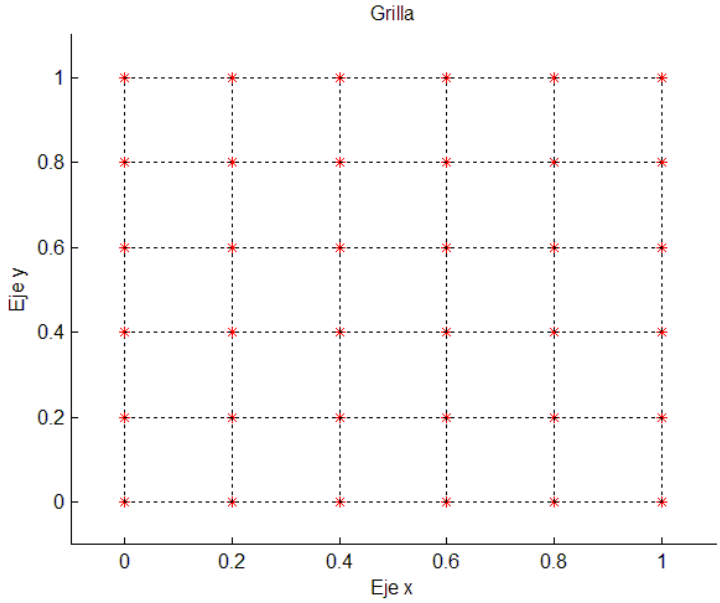

Figura 3.4: Grilla para $n = 4$

cionamiento. Nosotros la emplearemos no sólo para tal comparación sino también para revisar el desempeño de los métodos estacionarios no como precondicionadores si no como inversores lineales en sí.

Empezamos la solución numérica escogiendo un entero positivo n, hacemos $h = 1/(n+1)$ y particionamos el cuadrado $[0,1] \times [0,1]$ usando los nodos

$$x_i = ih,\ 0 \leq i \leq (n+1),$$
$$y_j = jh,\ 0 \leq j \leq (n+1).$$

La figura 3.4 ilustra el caso $n = 4$.

En este punto renunciamos a la posibilidad de conocer u sobre toda la región y nos concentramos en poder estimar su valor en los puntos de la malla. Ahora, obsérvese que gracias a las condiciones de contorno dadas el número de puntos en los que es necesario estimar el valor de u es n^2 y no $(n+1)^2$. Se trata precisamente de los nodos interiores. En efecto, para cada nivel de altura y_j se desconocen n valores de la la función u correspondiente a los n puntos $(x_1, y_j), (x_2, y_j), \cdots, (x_n, y_j)$. Nuestra estimación del valor de u en el punto (x_i, y_j) lo notaremos v_i^j.

Discretizado el dominio, procedemos a discretizar la ecuación diferencial para llevar esto a cabo nos valemos de las conocida fórmula de diferencia centrada que establece que

$$\begin{aligned}\frac{f(x+h)-2f(x)+f(x-h)}{h^2} &= f''(x)+\frac{h^2}{12}f^{(4)}(\zeta)\\ &= f''(x)+O\left(h^2\right).\end{aligned}$$

Siempre que f tenga cuarta derivada continua en una vecindad de x y h sea suficientemente pequeño. Entonces la versión discreta de nuestro problema queda

$$-\frac{v_{i+1}^j-2v_i^j+v_{i-1}^j}{h^2}-\frac{v_i^{j+1}-2v_i^j+v_i^{j-1}}{h^2}+\exp(x_i+y_j)v_i^j=1,$$

o equivalentemente

$$-v_i^{j-1}-v_{i-1}^j+\left(4+h^2\exp(x_i+y_j)\right)v_i^j-v_{i+1}^j-v_i^{j+1}=h^2. \tag{3.3}$$

Esta discretización suele llamarse discretización a cinco puntos. Convenimos ahora pensar en v como un vector que empieza en v_1^1 y al llegar a v_n^1 la numeración continúa en el siguiente nivel de altura en y. Así hasta llegar a v_n^n. La matriz de coeficientes A, inducida por (3.3), es una matriz tridiagonal por bloques de la forma

$$A=\begin{bmatrix} T_1 & -I & & & \\ -I & T_2 & -I & & \\ & \ddots & \ddots & \ddots & \\ & & -I & T_{n-1} & -I \\ & & & -I & T_n \end{bmatrix}.$$

Cada bloque es de orden n, así A es de orden n^2. Cada T_j es a su vez tridiagonal de la forma $T_j = trid([-1,\left(4+h^2\exp(x_i+y_j)\right),-1], i=1:n)$. El nombre discretización a cinco puntos proviene precisamente del hecho que A tiene a lo más cinco entradas no nulas por fila.

De otra parte, el lado derecho b inducido por (3.3) es un vector columna de n^2 entradas, casi todas iguales a h^2 excepto aquellas que son 1 o n módulo n. En estas hay que tener en cuenta las condiciónes de contorno, estas son precisamente las filas en las que A no tiene cinco entradas no nulas. Tenemos el sistema lineal

$$Av=b. \tag{3.4}$$

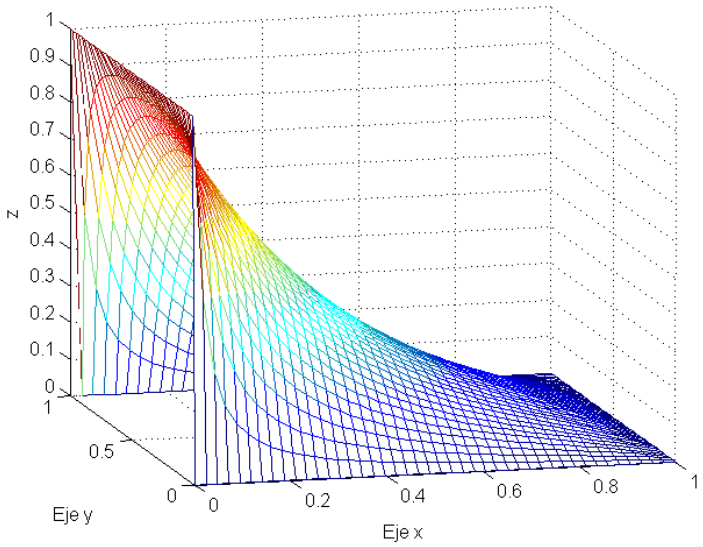

Figura 3.5: Solución Ejemplo 01 con $n = 128$

La figura 3.5 muestra la solución computada con $n = 128$. Es importante notar la buena concordancia con las condiciones exigidas a pesar de la discontinuidad inherente exigida por el problema a la solución buscada.

Las tablas 3.4, 3.5 y 3.6 muestran los resultados encontrados al resolver el sistema (3.4) por los distintos métodos estudiados, con $n = 128$. Note que este valor de n implica un tamaño de sistema de 16384 incognitas y ecuaciones. Nuevamente, el título de cada tabla indica la tolerancia y el tipo de métodos a listar.

Método	*Iter*	*Res Rel*	*Flg*	$t(seg)$
$Jacobi$	2000	2.1082e-03	1	31.95
$Gauss - Seidel$	2000	8.9498e-04	1	56.81
$Sor(w = 0.1)$	2000	1.4772e-02	1	56.65
$Sor(w = 0.5)$	2000	3.1165e-03	1	56.75
$Sor(w = 1.5)$	2000	5.5335e-05	1	68.54
$Sor(w = 1.7)$	1987	1.4707e-06	3	67.87

Tabla 3.4: Mtodos Estacionarios en Ejemplo 02 con Tolerancia $1e - 6$

Método	*Iter*	*Res Rel*	*Flg*	*t(seg)*
CG	396	9.9033e-09	0	5.64
LSQR	2000	5.4551e-03	1	34.98
GMRES(10)	200 10	4.3691e-06	1	53.28
GMRES(50)	20 13	9.8876e-09	0	57.75
GMRES(100)	7 18	9.8827e-09	0	65.89

Tabla 3.5: Mtodos No Estacionarios en Ejemplo 02 con Tolerancia $1e-8$

Método	*Iter*	*Res Rel*	*Flg*	*t(seg)*
CG	396	9.9033e-09	0	6.25
PCG, *Jacobi*	396	9.9028e-09	0	4.73
PCG, *Chol*(6)	3	1.2813e-09	0	7.65

Tabla 3.6: CG Precondicionado en Ejemplo 02 con Tolerancia $1e-8$

3.4 Trabajo por Bloques (MGW)

En este ejemplo consideramos un sistema de la forma

$$\mathcal{A}x = b. \tag{3.5}$$

Donde la matriz de coeficientes se deja escribir en la forma

$$\mathcal{A} = \begin{bmatrix} A & B^T \\ C & D \end{bmatrix}. \tag{3.6}$$

Resolveremos el sistema (3.5) utilizando las dos últimas técnicas de precondicionamento estudiadas para este tipo de matriz de coeficientes. Recordemos que el primero de ellos se basaba en tomar

$$P = \begin{bmatrix} A & B^T \\ 0 & D - CA^{-1}B^T \end{bmatrix}$$

y resolver entonces el sistema

$$Tx = \tilde{b}. \tag{3.7}$$

Donde, $T = P^{-1}\mathcal{A}$ y $\tilde{b} = P^{-1}b$. Entonces la proposición 2.4.5 establece que el polinomio minimal de T es de orden 2 y así métodos Krylov que basen su convergencia en el espectro de T deberán converger en *dos* (2) iteraciones. Note que este resultado es aplicable a GMRES sobre el sistema 3.7. Sin

embargo, este resultado no es aplicable a LSQR pues la convergencia de LSQR la domina el espectro de T^tT. En las tablas la utilización de esta estrategia la notaremos con $GMRES + MGW$.

El segundo procedimiento se basa en hacer no precondicionamiento a izquierda como arriba sino un precondicionamiento escalado a izquierda y derecha. Se consideran

$$\begin{aligned} P_1 &= \begin{bmatrix} I & 0 \\ CA^{-1} & -I \end{bmatrix}, \\ P_2 &= \begin{bmatrix} A & B^T \\ 0 & D - CA^{-1}B^T \end{bmatrix} \text{ y} \\ P &= P_1 P_2 \\ &= \begin{bmatrix} A & B^T \\ C & 2CA^{-1}B^T - D \end{bmatrix}. \end{aligned}$$

Se hace $T = P_1^{-1}\mathcal{A}P_2^{-1}$, $\tilde{b} = P_1^{-1}b$ y se considera el sistema

$$Tz = \tilde{b}.$$

Resuelto este sistema para z se encuentra x por $x = P_2^{-1}z$. Es importante anotar que en este ejemplo a fin de chequear la estabilidad computacional del resultado teórico no se explota la condición

$$P_1^{-1}\mathcal{A}P_2^{-1} = \begin{bmatrix} I & \\ & -I \end{bmatrix},$$

para resolver el problema de modo inmediato, en lugar de ello se usan los métodos estudiados hallando la acción de T sobre un vector z por $P_1^{-1}\left(\mathcal{A}\left(P_2^{-1}z\right)\right)$. Obsérvese que para este caso si es posible utilizar $LSQR$ pues las buenas características espectrales de T son conservadas por T^tT. La combinación de esta estrategia con $LSQR$ para resolver (3.5) la notamos en las tablas por $LSQR + IPSEN$ y la combinación con $GMRES$ es notada como $GMRES + IPSEN$.

Presentamos aquí el estudio de dos casos, en ambos A se escoge de la galeria de matrices de Matlab como la matriz de Poisson que es la tridiagonal por bloques con diagonales constantes que viene de la discretización a cinco puntos del problema de Dirichlet homogeneo. Las matrices B, C y D se generan como matrices aleatorias de entradas uniformemente distribuidas entre 0 y 1.

La tabla 3.7 muestra los resultados de generar A de orden $n = 10000$ y las otras matrices con $m = 100$, entonces el sistema resultante será de

Método	*Iter*	*Res Rel*	*Err Rel*	*Err Inf*	*Flg*	*t(seg)*
GMRES	1 500	5.5774e-07	1.3047e-02	5.1615e-02	1	189.31
GMRES+MGW	1 2	8.5995e-12	8.1283e-04	1.4432e-03	0	10.59
GMRES+IPSEN	1 2	8.4849e-15	6.6089e-07	1.3624e-06	0	11.09
LSQR	500	3.8963e-07	3.9658e-03	1.4444e-02	1	276.00
LSQR+IPSEN	2	3.9139e-11	7.4352e-03	2.8309e-02	0	22.06

Tabla 3.7: Precondicionamiento por bloques en Ejemplo 03

A	$\mathcal{A}$	*Iter*	*Res Rel*	*Err Rel*	*Err Inf*	*Flg*	*t(seg)*
1225	3200	1 2	2.7049e-12	1.7379e-05	4.8941e-05	0	55.25
1600	3200	1 2	7.1285e-10	6.2599e-02	1.5736e-01	0	42.42
2025	3200	1 2	3.5646e-12	2.8393e-05	6.2508e-05	0	32.06
2500	3200	1 2	5.7244e-12	2.6170e-04	5.1842e-04	0	19.34
3025	3200	1 2	4.7754e-12	5.1054e-05	9.2116e-05	0	4.93

Tabla 3.8: Precondicionamiento por bloques en Ejemplo 03 con tamao variable

orden 10100. Se tomó un vector x con todas sus entradas iguales a uno para generar un lado derecho y poder luego comparar la salida de los programas con la solución exacta. En la tabla en mención está la comparación de los métodos sin precondicionar con las formas precondicionadas.

El otro experimento realizado consistió en dejar fijo el orden de $\mathcal{A}$ en 3200 y variar el orden de su componente principal A, para revisar el comportamiento de la convergencia para diferentes razones entre el tamaño de A y el de $\mathcal{A}$. Los resultados encontrados se muestran en la tabla 3.8, donde siempre fue usado como método de solución $GMRES + MGW$.

.

Bibliografía

[1] S. Benbow, *Extending lsqr to generalized least-squares and schur complement problems without resorting to cholesky decompositions.*, (Available at http://citeseer.nj.nec.com/benbow97extending.html), (1997).

[2] G. Cervantes, *Introducción a los precondicionadores*, 2003.

[3] G. Cervantes and C. Meja, *Precondicionamiento de métodos iterativos*, Rev. Acad. Colomb. Cienc., 28 (2004), pp. 49–55.

[4] M. Gockenbach, *Partial Differential Equations: Analytical and Numerical Methods*, SIAM, 2002.

[5] G. Golub and W. Kahan, *Calculating the singular values and pseudoinverse of a matrix*, SIAM J. Numer. Anal., (1965), pp. 205–224.

[6] G. Golub and C. VanLoan, *Matrix Computations*, Johns Hopkins, 1996.

[7] A. Greenbaum, *Iterative Methods for Solving Linear Systems*, Siam, 1997.

[8] I. Ipsen, *A note on preconditioning nonsymmetric matrices*, SIAM J. SCI. Compt., 23 (2001), pp. 1050–1051.

[9] D. Kay, D. Loghin, and A. Wathen, *A preconditioner for steady-state navier-stokes equations*, SIAM J. SCI. Compt., 24 (2002), pp. 237–256.

[10] C. Kelley, *Iterative Methods for Linear and Nonlinear Equations*, SIAM, 1995.

[11] D. Kincaid and W. Cheney, *Análisis Numérico*, Addison-Wesley Iberoamericana, 1994.

[12] C. Lanczos, *An iteration method for the solution of the eigenvalue problem of linear differential and integral operators*, J. Res. N.B.S., (1950), pp. 255–282.

[13] J. A. Meijerink and H. A. van der Vorst, *An iterative solution method for linear systems of which the coefficient matrix is a m-matrix*, Math. Comput., 31 (1977), pp. 148–162.

[14] M. Murphy, G. Golub, and A. Wathen, *A note on preconditioning for indefinite linear systems*, SIAM J. SCI. Compt., 21 (2000), pp. 1969–1972.

[15] D. Ortega, *Numerical Analysis: A Second Course*, Siam, 1990.

[16] C. Paige and M. Saunders, *Solution of sparse indefinite systems of linear equations*, SIAM J. Num. Anal., (1975), pp. 617–629.

[17] ——, *Lsqr: An algorithm for sparse linear equations and sparse least squares*, ACM Trans. Math. Softw., (1982), pp. 43–71.

[18] Y. Saad, *Iterative Methods for Sparse Linear Systems*, Siam, second ed., 2000.

[19] G. W. Stewart, *Afternotes in Numerical Analysis*, Siam, 1996.

[20] ——, *Afternotes goes to Graduate School*, Siam, 1998.

[21] L. Trefheten and D. Bau, *Numerical Linear Algebra*, Siam, 1997.

[22] R. Varga, *Factorization and normalized iterative methods*, Boundary Problems in Differential Equations, (1960), pp. 121–142.