Universidad de Valladolid

Facultad de Ciencias

Trabajo Fin de Grado

Grado en Matemáticas

Curso 2018-2019

Métodos iterativos para sistemas lineales provenientes de problemas de punto de silla

Autor: Alba Crespo Martínez

Tutor: Luis Mª Abia Llera

Dpto. Matemática Aplicada

Prefacio

El objetivo de este trabajo Fin de Grado es generalizar algunos resultados sobre el método iterativo de sobrerrelajación (SOR) de parámetro ω, para sistemas lineales aumentados. El interés de estos sistemas estaría ya justificado porque surgen de forma natural en la discretización del problema de Stokes en mecánica de fluidos, de gran interés para las aplicaciones. Este tipo de sistemas aparecen también en problemas punto de silla en optimización.

El punto de partida son los resultados básicos de convergencia de los métodos iterativos basados en escisiones de la matriz A estudiados en el Grado de Matemáticas. En el capítulo primero ampliamos la teoría de convergencia del método SOR para sistemas lineales generales con vistas a proporcionar resultados relativos a la elección del parámetro óptimo. Esto nos lleva a introducir clases de matrices para las que es factible esta optimización del parámetro.

El capítulo 2 aporta extensiones del método SOR para tratar los sistemas aumentados. Constatamos en los artículos manejados una diversidad de extensiones del método SOR para sistemas aumentados, cada una con su propia teoría de convergencia y de optimización del parámetro. Hemos seleccionado dos análisis: uno se refiere al método SOR con un parámetro (con y sin preacondicionamiento), y el segundo al que se denomina GSOR (SOR generalizado), que depende de dos parámetros.

El capítulo 3 considera métodos iterativos en los que la escisión de la matriz del sistema es diferente a la de los métodos SOR. Se basan en escisiones de la matriz del sistema en una parte Hermítica y otra antihermítica, con desplazamiento de la diagonal. Desarrollamos una teoría de convergencia para sistemas generales, señalando las modificaciones que permiten aplicar los resultados a sistemas aumentados.

El énfasis de la memoria es en el análisis antes que en un estudio comparativo de la eficiencia de las distintas propuestas.

En Valladolid, a 15 de julio de 2019

Índice general

Capítulo 1

Métodos iterativos para sistemas lineales

1.1. Introducción

Sean $A \in \mathbb{C}^{n\times n}$ una matriz cuadrada compleja y $b \in \mathbb{C}^n$ un vector complejo dado. Consideramos el sistema lineal

$$Ax = b. \tag{1.1}$$

La solución $x \in \mathbb{C}^n$ existe y es única si y solo si A es no singular. Entonces, $x = A^{-1}b$. Supodremos a partir de aquí que A es no singular y todas las entradas de la diagonal son números complejos distintos del cero.

Estudiaremos ahora distintos métodos iterativos. Estos métodos iterativos no calculan, en general, la solución exacta del sistema, sino que a partir de una aproximación inicial $x^{(0)}$ van obteniendo sucesivas aproximaciones a la solución hasta que se alcanza la precisión deseada. La convergencia de las aproximaciones así obtenidas a la solución exacta puede garantizarse bajo diferentes condiciones que estudiaremos. El número de iteraciones necesarias para alcanzar una precisión prefijada dependerá, en general, del método utilizado.

Aunque el punto de partida de este trabajo es el análisis del método SOR y la optimización de su parámetro, en la siguiente sección revisamos sucintamente la teoría general de convergencia de métodos iterativos para (1.1) ligados a escisiones $A = M - N$ de la matriz, para lo que nos hemos basado principalmente en el libro de Varga [6] y el de Young [7].

1.2. Teoría general de convergencia

Consideramos la siguiente escisión para la matriz A

$$A = M - N$$

y podemos reescribir (1.1) como

$$Mx = Nx + b \tag{1.2}$$

Entonces, para un iterante inicial $x^{(0)}$, el método iterativo genera una sucesión definida recurrentemente por

$$Mx^{(m+1)} = Nx^{(m)} + b, \qquad m = 0, 1, 2, ... \tag{1.3}$$

Para que el método sea útil en la práctica necesitaremos que resolver el sistema lineal (1.3) con matriz M sea mucho más barato computacionalmente que resolver el sistema inicial (1.1) de matriz A. Además para que (1.3) defina el iterante $x^{(m+1)}$ de forma única es necesario que M sea regular.

En ese caso, podemos escribir (1.3) como

$$x^{(m+1)} = Gx^{(m)} + d, \qquad m = 0, 1, 2, ... \tag{1.4}$$

donde $G = M^{-1}N$ se llama la **matriz de iteración** del método iterativo y $d = M^{-1}b$.

En particular, si x^* es la solución de (1.1), restando de (1.4) la identidad

$$x^* = Gx^* + d,$$

se obtiene la siguiente recurrencia para el error $e^{(m)} = x^{(m)} - x^*$,

$$e^{(m+1)} = Ge^{(m)}, \quad m = 0, 1, 2, \ldots$$

Obviamente, $e^{(m)} = G^m e^{(0)}$, y se tiene que la sucesión $\{e^{(m)}\}_{m=0}^{\infty}$ converge a cero, cualquiera que sea $e^{(0)}$, si y solo si la matriz G^m converge a cero.

1.1 Definición. Sea $G \in \mathbb{C}^{n\times n}$. Se dice que G **es convergente** (a cero) si la secuencia de matrices $G, G^2, G^3, \ldots$converge a la matriz nula O, y se dice que G es **divergente** en caso contrario.

1.1 Teorema. *Sea $G \in \mathbb{C}^{n\times n}$. Entonces, G es convergente si y solo si el radio espectral (recordemos que el radio espectral es el mayor de entre los valores absolutos de los autovalores de G, $\rho(G) := \max_i(|\lambda_i|)$) satisface $\rho(G) < 1$.*

Demostración. Recordemos que para una matriz G dada, existe una matriz $P \in \mathbb{C}^{n\times n}$ con la cual podemos reducir G a su forma canónica de Jordan, es decir,

$$PGP^{-1} = J = \begin{pmatrix} J_1 & & & \\ & J_2 & & \\ & & \ddots & \\ & & & J_r \end{pmatrix}, \tag{1.5}$$

donde cada matriz $J_i \in \mathbb{C}^{n_i\times n_i}$ es de la forma

$$J_i = \begin{pmatrix} \lambda_i & 1 & & & & \\ & \lambda_i & 1 & & & \\ & & \lambda_i & 1 & & \\ & & & \ddots & \ddots & \\ & & & & & 1 \\ & & & & & \lambda_i \end{pmatrix}.$$

Como cada matriz J_i es triangular superior, también lo es J. Luego, $\{\lambda_i\}_{i=1}^r$ incluye todos los autovalores de G y J (que tienen los mismos autovalores por ser matrices semejantes). De (1.5) se tiene

$$J^m = \begin{pmatrix} J_1^m & & & \\ & J_2^m & & \\ & & \ddots & \\ & & & J_n^m \end{pmatrix}, \quad m \geq 1 \tag{1.6}$$

y por la forma que tienen las J_i,

$$J_i^2 = \begin{pmatrix} \lambda_i^2 & 2\lambda_i & 1 & & & \\ & \lambda_i^2 & 2\lambda_i & 1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & \lambda_i^2 & 2\lambda_i & 1 \\ & & & & \lambda_i^2 & 2\lambda_i \\ & & & & & \lambda_i^2 \end{pmatrix}.$$

En general, $J_i^m = \left(d_{k,l}^{(m)}(i)\right)$ para $1 \leq k, l \leq n_i$, donde

$$d_{k,l}^{(m)}(i) = \begin{cases} 0 & \text{si } l < k \\ \begin{pmatrix} m \\ l-k \end{pmatrix} \lambda_i^{m-l+k} & \text{si } k \leq l \leq \text{mín}(n_i, m+k) \\ 0 & \text{si } m+k < l \leq n_i \end{cases} \tag{1.7}$$

es decir,

$$J_i^m = \begin{pmatrix} \lambda_i^m & \begin{pmatrix} m \\ 1 \end{pmatrix} \lambda_i^{m-1} & \cdots & \begin{pmatrix} m \\ n_i - 1 \end{pmatrix} \lambda_i^{m-(n_i-1)} \\ 0 & \lambda_i^m & \cdots & \begin{pmatrix} m \\ n_i - 2 \end{pmatrix} \lambda_i^{m-(n_i-2)} \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_i^m \end{pmatrix}.$$

Supongamos entonces que G es convergente, entonces $G^m \to O$ cuando $m \to \infty$. Como $J^m = PG^mP^{-1}$, entonces $J^m \to O$ cuando $m \to \infty$. Entonces, cada $J_i^m \to O$ cuando $m \to \infty$, luego las entradas diagonales λ_i de J_i

deben satisfacer $|\lambda_i| < 1$ para $i = 1, \ldots, r$. Luego,

$$\rho(G) = \rho(J) = \max_{1 \leq i \leq r} |\lambda_i| < 1.$$

Recíprocamente, si $\rho(G) = \rho(J) < 1$, entonces, $|\lambda_i| < 1$ para $i = 1, \ldots, r$. De (1.7) se deduce que

$$\lim_{m \to \infty} d_{k,l}^{(m)}(i) = 0, \quad \forall k, l = 1, \ldots, n_i$$

de forma que cada J_i es convergente. Luego, de (1.6), se deduce que J también es convergente y como $J^m = PG^mP^{-1}$, también G es convergente. ■

Para cuantificar la **velocidad asintótica de convergencia** de la sucesión de errores $\{e^{(m)}\}_{m=0}^{\infty}$, observemos que si G es diagonalizable, $v_1, \ldots, v_n$ es una base de autovectores, $Gv_i = \lambda_i v_i$ para $i = 1, \ldots, n$; y si

$$e^{(0)} = \alpha_1 v_1 + \ldots + \alpha_n v_n,$$

entonces

$$G^m e^{(0)} = \alpha_1 \lambda_1^m v_1 + \ldots + \alpha_n \lambda_n^m v_n.$$

Observemos ahora que si $|\lambda_1| > |\lambda_2| \geq \ldots \geq |\lambda_n| > 0$, entonces cuando $m \to \infty$ el término dominante en $G^m e^{(0)}$ es el primer sumando del miembro de la derecha, que en cada iteración se reduce en norma por un factor $|\lambda_1| = \rho(G)$. Así, el valor $\rho(G)$ cuantifica la velocidad asintótica de convergencia del método iterativo. Será nuestro parámetro principal en el análisis de convergencia de los diferentes métodos iterativos.

1.3. El método SOR de parámetro ω

Describimos primero los métodos de Jacobi y Gauss-Seidel, que sirven para motivar la iteración SOR.

El **método de Jacobi** corresponde a la escisión

$$A = D - E - F = D(I - D^{-1}E - D^{-1}F) = D(I - L - U),$$

para $L := D^{-1}E$ y $U := D^{-1}F$, donde D es la matriz diagonal cuyas entradas son las entradas diagonales de A, E es una matriz triangular inferior estricta y F es una matriz triangular superior estricta. La iteración es entonces

$$x^{(m+1)} = D^{-1}(E + F)x^{(m)} + D^{-1}b;$$

de forma que $M_J = D^{-1}(E + F) = L + U$ es la correspondiente matriz de iteración.

El **método de Gauss-Seidel** corresponde a la iteración

$$(D-E)x^{(m+1)} = Fx^{(m)} + b,$$

cuya matriz de iteración es $M_{GS} = (D-E)^{-1}F = (I-L)^{-1}U$.

Introducimos ya el **método SOR** . Está muy relacionado con el método de Gauss-Seidel, ya que cada componente del nuevo iterante es una media ponderada de la componente que obtendríamos en el método de Gauss-Seidel y la misma componente del iterante anterior. Formalmente viene dado por:

$$\begin{aligned} &x_i^{(m+1)} = x_i^{(m)} + \omega\left\{\tilde{x}_i^{(m+1)} - x_i^{(m)}\right\} = (1-\omega)x_i^{(m)} + \omega\tilde{x}_i^{(m+1)}, \\ &1 \le i \le n,\ m \ge 0, \end{aligned} \tag{1.8}$$

donde los vectores auxiliares $\tilde{x}_i^{(m)}$ quedan definidos por

$$a_{i,i}\tilde{x}_i^{(m+1)} = -\sum_{j=1}^{i-1} a_{i,j}x_j^{(m+1)} - \sum_{j=i+1}^{n} a_{i,j}x_j^{(m)} + b_i, \quad 1 \le i \le n,\ m \ge 0. \tag{1.9}$$

Podemos combinar (1.8) y (1.9) en una sola ecuación para obtener

$$\begin{aligned} &a_{i,i}x_i^{(m+1)} = a_{i,i}x_i^{(m)} + \omega\left\{-\sum_{j=1}^{i-1} a_{i,j}x_j^{(m+1)} - \sum_{j=i+1}^{n} a_{i,j}x_j^{(m)} + b_i - a_{i,i}x_i^{(m)}\right\} \\ &1 \le i \le n,\ m \ge 0. \end{aligned} \tag{1.10}$$

Entonces, las componentes de cada iterante se definen como

$$\begin{aligned} &x_i^{(m+1)} = \frac{\omega}{a_{i,i}}\left(b_i - \sum_{j=1}^{i-1} a_{i,j}x_j^{(m+1)} - \sum_{j=i+1}^{n} a_{i,j}x_j^{(m)}\right) + (1-\omega)x_i^{(m)}, \\ &1 \le i \le n,\ m \ge 0. \end{aligned}$$

De (1.8) sabemos que $x_i^{(m+1)}$ es una media ponderada de $x_i^{(m)}$ y $\tilde{x}_i^{(m+1)}$ donde los pesos dependen solo de ω. Este parámetro ω se llama factor de relajación y el método se conoce como método de sobrerrelajaciones sucesivas (SOR). Observamos que para el caso $\omega = 1$ este método coincide exactamente con el método de Gauss-Seidel.

Para escribir el método SOR en forma matricial recurrimos de nuevo a la representación $A = D - E - F$. De esta forma (1.10) se convierte en

$$(D-\omega E)x^{(m+1)} = \left\{(1-\omega)D + \omega F\right\}x^{(m)} + \omega b, \qquad m \ge 0.$$

Como $D - \omega E$ es no singular para cualquier elección de ω, entonces llamando como antes $L := D^{-1}E$ y $U := D^{-1}F$ tenemos

$$x^{(m+1)} = (I-\omega L)^{-1}\left\{(1-\omega)I + \omega U\right\}x^{(m)} + \omega(I-\omega L)^{-1}D^{-1}b, \qquad m \ge 0.$$

La matriz de iteración será $\mathcal{L}_\omega := (I-\omega L)^{-1}\left\{(1-\omega)I + \omega U\right\}$.

1.4. Convergencia del método SOR

1.2 Teorema. (Kahan, 1958, [6]) *Sea* $\mathcal{L}_\omega := (I - \omega L)^{-1}\{\omega U + (1-\omega)I\}$ *la matriz del método SOR, entonces para cualquier* ω *real o complejo,*

$$\rho(\mathcal{L}_\omega) \geq |\omega - 1| \tag{1.11}$$

donde la igualdad se da solo si todos los autovalores de $\mathcal{L}_\omega$ *tienen módulo* $|\omega - 1|$.

Demostración. Llamemos $\phi(\lambda) = \det(\lambda I - \mathcal{L}_\omega)$ al polinomio característico de $\mathcal{L}_\omega$. Observemos por otra parte que como L es estrictamente triangular inferior, entonces $I - \omega L$ es no singular y además $\det(I - \omega L) = 1$. Por tanto,

$$\begin{aligned}
&\phi(\lambda) = \det(\lambda I - \mathcal{L}_\omega) = \det(I - \omega L)\det(\lambda I - \mathcal{L}_\omega) \\
&= \det\{(I - \omega L)(\lambda I - \mathcal{L}_\omega)\} = \det\{\lambda I - \omega\lambda L - (I - \omega L)\mathcal{L}_\omega\} \\
&= \det\{\lambda I - \omega\lambda L - (I - \omega L)(I - \omega L)^{-1}\{\omega U + (1-\omega)I\}\} \\
&= \det\{(\lambda + \omega - 1)I - \omega\lambda L - \omega U\}.
\end{aligned}$$

Denotemos por μ al término constante del polinomio $\phi(\lambda)$, que es el producto de los autovalores de $\mathcal{L}_\omega$ cambiados de signo y se obtiene haciendo $\lambda = 0$ en la expresión anterior. Por tanto,

$$\mu = \prod_{i=1}^{n}(-\lambda_i(\omega)) = \det\{(\omega - 1)I - \omega U\} = (\omega - 1)^n$$

donde la última igualdad es cierta ya que U es estrictamente triangular superior. Por tanto,

$$\rho(\mathcal{L}_\omega) = \max_{1 \leq i \leq n} |\lambda_i(\omega)| \geq |\omega - 1|,$$

donde la igualdad se verifica solo si todos los autovalores de $\mathcal{L}_\omega$ tienen módulo $|\omega - 1|$.

■

En el caso del método SOR nuestro objetivo es encontrar el valor real de ω que minimiza $\rho(\mathcal{L}_\omega)$. Para ello, el teorema anterior nos muestra que solo es necesario considerar $0 < \omega < 2$, valores para los que $\mathcal{L}_\omega$ es convergente. Cuando $0 < \omega < 1$ llamamos al método de subrrelajación y para $1 < \omega < 2$, de sobrerrelajación.

1.3 Teorema. (Householder-John) *Sea* A *una matriz simétrica definida positiva y sea* $A = M - N$, *una escisión de* A, *con* M *regular y cuyo método iterativo asociado es*

$$Mx^{(n+1)} = Nx^{(n)} + b$$

para un $x^{(0)}$ *fijo. Si la matriz simétrica* $M^T + N$ *es definida positiva entonces* $\rho(M^{-1}N) < 1$.

Demostración. Como $A = M - N$, entonces $M^{-1}A = I - M^{-1}N$, luego podemos escribir $M^{-1}N = I - M^{-1}A$. Sea ahora λ un autovalor de $I - M^{-1}A$ y sea $x \in \mathbb{C}^n$, $x \neq 0$, un autovector asociado a λ. Entonces,

$$\begin{aligned} &\lambda x = (I - M^{-1}A)x \Rightarrow \lambda x = Ix - M^{-1}Ax \Rightarrow M^{-1}Ax = Ix - \lambda x \Rightarrow \\ &M^{-1}Ax = (1-\lambda)x \Rightarrow Ax = (1-\lambda)Mx. \end{aligned} \tag{1.12}$$

Luego $\lambda \neq 1$. Consideremos el producto escalar por el vector $x = u + iv$ y tomemos el vector transpuesto conjugado. Entonces,

$$x^H Ax = (1-\lambda)x^H Mx = (1-\bar{\lambda})x^H M^T x,$$

donde denotamos x^H al vector transpuesto conjugado de x y $\bar{\lambda}$ es el conjugado de λ. De aquí deducimos

$$\left(\frac{1}{1-\lambda} + \frac{1}{1-\bar{\lambda}} - 1\right) x^H Ax = x^H (M^T + M - A)x.$$

Tomamos ahora la parte real en la expresión anterior

$$\frac{1-|\lambda|^2}{|1-\lambda|^2}(u^T Au + v^T Av) = u^T(N + N^T - A)u + v^T(N + N^T - A)v.$$

Tanto A como $M + M^T - A$ son definidas positivas, luego de lo anterior se deduce que $|\lambda| < 1$. Esto es, $\rho(M^{-1}N) < 1$.

■

1.4 Teorema. *Sea A una matriz simétrica y definida positiva. El método SOR de parámetro ω para A converge si y solo si $0 < \omega < 2$.*

Demostración. Ya vimos que $0 < \omega < 2$ es una condición necesaria para la convergencia del método para matrices regulares. Veamos entonces que también es una condición suficiente. Para ello es suficiente con aplicar el teorema de Householder-John. Tenemos que ver cuándo es $M^T + M - A$ definida positiva. Esta condición queda reducida a que $(2\omega^{-1} - 1)D$ sea definida positiva, lo cual ocurre si y solo si $0 < \omega < 2$.

■

Teorema de Ostrowski- Reich

Consideremos ahora que la matriz $A \in \mathbb{C}^{n\times n}$ del sistema inicial $Ax = b$ es una matriz hermítica, es decir, $A = A^H$ donde A^H es la matriz transpuesta conjugada de A. Supongamos también que A puede escribirse como $A = D - E - E^H$, donde D y E son matrices $n \times n$ y D es hermítica y definida

positiva. Supongamos, por último, que también $D-\omega E$ es no singular para todo $0 \leq \omega \leq 2$. Entonces,

$$Ax = b \implies (D - E - E^H)x = b \implies$$
$$(D-E)x = E^H x + b \implies \omega(D-E)x = \omega E^H x + \omega b$$

sumando $(1-\omega)Dx$ a ambos lados,

$$(1-\omega)Dx + \omega(D-E)x = (1-\omega)Dx + \omega E^H x + \omega b \implies$$
$$Dx - \omega Ex = (1-\omega)Dx + \omega E^H x + \omega b \implies$$
$$(D-\omega E)x = \{\omega E^H + (1-\omega)D\}x + \omega b$$

De manera que tenemos el método iterativo siguiente

$$(D-\omega E)x^{(m+1)} = \{\omega E^H + (1-\omega)D\}x^{(m)} + \omega b, \quad m \geq 0 \tag{1.13}$$

que puede llevarse acabo ya que habíamos supuesto que $D-\omega E$ era no singular para $0 \leq \omega \leq 2$

Llamemos entonces $L := D^{-1}E$, $U := D^{-1}E^H$, de forma que (1.13) se puede reescribir como

$$x^{(m+1)} = \mathcal{L}_\omega x^{(m)} + \omega(D-\omega E)^{-1}k, \quad m \geq 0$$

donde $\mathcal{L}_\omega = (I-\omega L)^{-1}\{\omega U + (1-\omega)I\}$, de tal forma que el método anterior es el método SOR de relajaciones sucesivas. Sin embargo, aquí L y U no tienen que ser necesariamente estrictamente inferiores y superiores.

1.5 Teorema. (Ostrowski, 1954, [6]) *Sea $A = D-E-E^H$ una matriz Hermítica $n \times n$, donde D es Hermítica y definida positiva y $D-\omega E$ es no singular para $0 \leq \omega \leq 2$. Entonces $\rho(\mathcal{L}_\omega) < 1$ si y solo si A es positiva definida y $0 < \omega < 2$.*

Demostración. Sea $e^{(0)}$ el vector del error inicial. Los vectores para los errores del método SOR cumplen: $e^{(m+1)} = \mathcal{L}_\omega e^{(m)}$ para $m \geq 0$. Equivalentemente,

$$(D-\omega E)e^{(m+1)} = (\omega E^H + (1-\omega)D)e^{(m)}, \quad m \geq 0 \tag{1.14}$$

Denotemos $\delta^{(m)} := e^{(m)} - e^{(m+1)}$ para $m \geq 0$.

Entonces, de la definición de $A = D - E - E^H$ podemos escribir (1.14) como

$$(D-\omega E)\delta^{(m)} = \omega A e^{(m)}, \quad m \geq 0 \tag{1.15}$$

y también como

$$\omega A e^{(m+1)} = (1-\omega)D\delta^{(m)} + \omega E^H \delta^{(m)} \quad m \geq 0 \tag{1.16}$$

Operando, podemos combinar (1.15) y (1.16) en una sola ecuación para dar

$$(2-\omega)(\delta^{(m)})^H D\delta^{(m)} = \omega\{(e^{(m)})^H A e^{(m)} - (e^{(m+1)})^H A e^{(m+1)}\}, \quad m \geq 0 \tag{1.17}$$

Supongamos que A es definida positiva y que $0 < \omega < 2$ y elegimos $e^{(0)}$ tal que sea un autovector de $\mathcal{L}_\omega$ con autovalor asociado λ. Luego, $e^{(1)} = \mathcal{L}_\omega e^{(0)} = \lambda e^{(0)}$, y $\delta^{(0)} = (1-\lambda)e^{(0)}$. Entonces, (1.17) se traduce en este caso particular en

$$\left(\frac{2-\omega}{\omega}\right)|1-\lambda|^2(e^{(0)})^H De^{(0)} = (1-|\lambda|^2)(e^{(0)})^H Ae^{(0)} \tag{1.18}$$

Veamos que λ no puede ser igual a 1, pues entonces $\delta^{(0)}$ sería nulo. Si fuera $\lambda = 1$, de (1.15) se deduciría que $Ae^{(0)} = 0$ y como $e^{(0)}$ es no nulo por definición, entonces esto contradiría que A es definida positiva. Como $0 < \omega < 2$, el lado izquierdo de (1.18) es positivo, y como $(e^{(0)})^H Ae^{(0)} > 0$, entonces debe ser $|\lambda| < 1$. Por tanto, $\mathcal{L}_\omega$ es convergente.

Supongamos ahora que para $0 < \omega < 2$, $\mathcal{L}_\omega$ es convergente. Tenemos que ver que entonces A es definida positiva. Como $\mathcal{L}_\omega$ es convergente entonces cuando m aumenta $e^{(m)}$ tiende a cero para cualquier vector inicial $e^{(0)}$. Luego, $(e^{(m)})^H Ae^{(m)}$ tiende a cero cuando m aumenta. Por otro lado, como D es definida positiva, $(\delta^{(m)})^H D\delta^{(m)} \geq 0$. De (1.17) se tiene que

$$\begin{aligned}(e^{(m)})^H Ae^{(m)} &= (e^{(m+1)})^H Ae^{(m+1)} + \left(\tfrac{2-\omega}{\omega}\right)(\delta^{(m)})^H D\delta^{(m)} \\ &\geq (e^{(m+1)})^H Ae^{(m+1)}, \quad m \geq 0.\end{aligned} \tag{1.19}$$

Si A no fuera definida positiva, podríamos encontrar un vector no nulo $e^{(0)}$ tal que $(e^{(0)})^H Ae^{(0)} \leq 0$. Como ningún autovalor de $\mathcal{L}_\omega$ puede ser igual a la unidad , entonces no puede ser el vector nulo $\delta^{(0)} = (I - \mathcal{L}_\omega)e^{(0)}$. Entonces, de (1.18) y $(\delta^{(0)})^H D\delta^{(0)} > 0$ se deduce que $(e^{(1)})^H Ae^{(1)} < (e^{(0)})^H Ae^{(0)} \leq 0$. Pero entonces, (1.19) contradice el hecho de que $(e^{(m)})^H Ae^{(m)}$ tiende a cero cuando m aumenta. Por tanto, A debe ser definida positiva.

■

1.1 Corolario. (Reich, 1949, [6]) *Sea $A = D - E - E^H$ una matriz Hermítica donde D es hermítica y positiva definida y $(D-E)$ es no singular. Entonces, el método de Gauss-Seidel es convergente si y solo si A es positiva definida.*

Demostración. Se deduce inmediatamente el teorema de Ostrowski, ya que el método de Gauss-Seidel es un caso particular de (1.14) con $\omega = 1$.

■

1.5. Optimización del parámetro para el método SOR

Consideremos para el sistema (1.1) , $Ax = b$, con $A = (a_{i,j}) \in \mathbb{C}^{n\times n}$, $n \geq 2$ la siguiente partición para A,

$$\begin{pmatrix} A_{1,1} & A_{1,2} & \dots & A_{1,N} \\ A_{2,1} & A_{2,2} & \dots & A_{2,N} \\ \vdots & & & \vdots \\ A_{N,1} & A_{N,2} & \dots & A_{N,N} \end{pmatrix}, \tag{1.20}$$

donde los bloques diagonales $A_{i,i}$ para $1 \leq i \leq N$ son matrices cuadradas y no vacías. Cada matriz $A_{i,i}$ es de dimensión $n_i \times n_i$ con $n_i \geq 1$. Asumimos que cada submatriz diagonal es no singular, de forma que

$$D = \begin{pmatrix} A_{1,1} & & & \\ & A_{2,2} & & \\ & & \ddots & \\ & & & A_{N,N} \end{pmatrix} \tag{1.21}$$

es no singular. La matriz $B \in \mathbb{C}^{n \times n}$

$$B := -D^{-1}A + I \tag{1.22}$$

es la matriz de Jacobi por bloques asociada a la partición (1.20) de A.

1.2 Definición. Sea $A \in \mathbb{C}^{n \times n}$ (no necesariamente no negativa o irreducible). Se dice que A es **débilmente cíclica de índice** $k > 1$ si existe una matriz de permutación P tal que PAP^T es de la forma

$$PAP^T = \begin{pmatrix} 0 & 0 & \dots & 0 & A_{1,k} \\ A_{2,1} & 0 & \dots & 0 & 0 \\ 0 & A_{3,2} & \dots & 0 & 0 \\ \vdots & & \ddots & & \vdots \\ 0 & 0 & \dots & A_{k,k-1} & 0 \end{pmatrix}, \tag{1.23}$$

donde las submatrices diagonales nulas son cuadradas.

La teoría de matrices no negativas de Perron-Frobenius establece la invariancia frente a rotaciones de ángulos $2\pi/k$ de matrices no negatives irreducibles que tienen la forma (1.23). Esta invariancia se puede establecer directamente a partir de la forma (1.23) para matrices generales, un resultado que se debe a Romanovsky [6]

1.6 Teorema. *(1936, [6]) Sea $A = (a_{ij})$ una matriz $n \times n$ débilmente k-cíclica. Entonces*

$$\phi(t) = det\ (tI - A) = t^m \prod_{i=1}^{r} (t^k - \sigma_i^k),$$

donde $m + rk = n$.

1.3 Definición. Si la matriz de Jacobi por bloques, B, de (1.22) para la matriz A de (1.20) es débilmente cíclica de índice $p \geq 2$, entonces A es **p-cíclica** relativa a la partición de (1.20).

Las siguientes matrices tienen un especial interés:

$$A_1 = \begin{pmatrix} A_{1,1} & 0 & 0 & \dots & 0 & A_{1,p} \\ A_{2,1} & A_{2,2} & 0 & \dots & 0 & 0 \\ \vdots & \ddots & \ddots & & & \vdots \\ \vdots & & \ddots & \ddots & & \vdots \\ \vdots & & & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & A_{p,p-1} & A_{p,p} \end{pmatrix}, \quad p \geq 2 \tag{1.24}$$

$$A_2 = \begin{pmatrix} A_{1,1} & A_{1,2} & & & & \\ A_{2,1} & A_{2,2} & A_{2,3} & & & \\ & \ddots & \ddots & \ddots & & \\ & & \ddots & \ddots & A_{N-1,N} & \\ & & & A_{N,N-1} & A_{N,N} & \end{pmatrix}. \tag{1.25}$$

A las matrices de la forma de A_2 se les llama matrices tridiagonales por bloques. De acuerdo con (1.22), las matrices A_1 y A_2 dan lugar, respectivamente, a las siguientes matrices de Jacobi:

$$B_1 = \begin{pmatrix} 0 & 0 & 0 & \dots & 0 & B_{1,p} \\ B_{2,1} & 0 & 0 & \dots & 0 & 0 \\ 0 & B_{3,2} & 0 & \dots & 0 & 0 \\ \vdots & & \ddots & \ddots & & \vdots \\ \vdots & & & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & B_{p,p-1} & 0 \end{pmatrix}, \quad p \geq 2 \tag{1.26}$$

$$B_2 = \begin{pmatrix} 0 & B_{1,2} & & & & \\ B_{2,1} & 0 & B_{2,3} & & & \\ & \ddots & \ddots & \ddots & & \\ & & \ddots & \ddots & B_{N-1,N} & \\ & & & B_{N,N-1} & 0 & \end{pmatrix}. \tag{1.27}$$

Observemos que la matriz B_1 es una matriz débilmente cíclica de índice p. Por tanto, por definición, se tiene que la matriz A_1 es p-cíclica. Por otro lado, los bloques de la matriz B_2 se pueden permutar para ver que B_2 es una matriz débilmente cíclica de índice 2. Por tanto, la matriz A_2 es una matriz 2-cíclica, es decir, las matrices tridiagonales por bloques son un caso importante de matrices 2-cíclicas.

1.4 Definición. Si la matriz A de (1.20) es p-cíclica, entonces se dice que A es **consistentemente ordenada** si todos los autovalores de la matriz

$$B(\alpha) := \alpha L + \alpha^{-(p-1)} U,$$

que deriva de la matriz de Jacobi $B = L + U$, son independientes de α, para $\alpha \neq 0$. En ese caso también diremos que B es consistentemente ordenada. En caso contrario diremos que A y B son inconsistentemente ordenadas.

Veamos que las matrices A_1 de (1.24) y A_2 de (1.25) son matrices consistentemente ordenadas y, por tanto, lo son también B_1 de (1.26) y B_2 de (1.27).

Para B_1 consideremos

$$B_1(\alpha) := \begin{pmatrix} 0 & 0 & 0 & \dots & 0 & \alpha^{-(p-1)} B_{1,p} \\ \alpha B_{2,1} & 0 & 0 & \dots & 0 & 0 \\ 0 & \alpha B_{3,2} & 0 & \dots & 0 & 0 \\ \vdots & & \ddots & \ddots & & \vdots \\ \vdots & & & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \alpha B_{p,p-1} & 0 \end{pmatrix}$$

y es fácil comprobar que $B_1^p(\alpha) = B_1^p$, para cualquier $\alpha \neq 0$. Por tanto, los autovalores de $B_1(\alpha)$ son independientes de α. Así pues, las matrices A_1 y B_1 son consistentemente ordenadas.

Centrémonos ahora en la matriz B_2 y consideremos los autovalores λ de la matriz $B_2(\alpha)$, es decir,

$$B_2(\alpha)x = \lambda x,$$

donde $x \neq 0$. De acuerdo con la partición (1.27), podemos tomar una partición también para x, de forma que se tiene

$$\alpha B_{j,j-1} X_{j-1} + \frac{1}{\alpha} B_{j,j+1} X_{j+1} = \lambda X_j, \quad 1 \leq j \leq N,$$

donde tanto $B_{1,0}$ como $B_{N,N+1}$ se definen como matrices nulas. Definamos entonces $Z_j := (1/\alpha^{j-1}) X_j$, para $1 \leq j \leq N$, de forma que la ecuación anterior es

$$B_{j,j-1} Z_{j-1} + B_{j,j+1} Z_{j+1} = \lambda Z_j, \quad 1 \leq j \leq N.$$

Luego todo autovalor de $B_2(\alpha)$, para $\alpha \neq 0$ es también un autovalor de B_2. Luego, B_2 y A_2 son ambas matrices consistentemente ordenadas. Esto es, cualquier matriz tridiagonal por bloques es una matriz consistentemente ordenada.

1.5 Definición. Se dice que una matriz A de orden n tiene la **propiedad A de Young** si existen dos subconjuntos disjuntos S_1 y S_2 que particionan $W = \{1, 2, \ldots, n\}$ y tales que si $i \neq j$ y si o bien $a_{ij} \neq 0$ o $a_{ji} \neq 0$ entonces $i \in S_1$ y $j \in S_2$ o bien $i \in S_2$ y $j \in S_1$. Cuando para un par (i, j) se tiene que o bien $a_{ij} \neq 0$ o bien $a_{ji} \neq 0$ se dice que los índices i y j están **asociados**.

En esta definición puede suceder que S_1 o S_2 sea vacío, en cuyo caso la matriz A es diagonal. La definición anterior es equivalente a la siguiente:

1.1 Proposición. *Una matriz A tiene la propiedad A si y sólo si A es una matriz diagonal o existe una matriz de permutación P tal que PAP^T tiene la forma*

$$P^T AP = \begin{pmatrix} D_1 & B \\ C & D_2 \end{pmatrix} \tag{1.28}$$

donde D_1 y D_2 son matrices diagonales cuadradas y no necesariamente del mismo orden.

Demostración. Si A tiene la propiedad A, sean S_1 y S_2 los conjuntos que especifican la definición. Si S_1 o S_2 son vacíos entonces A es diagonal. En otro caso, denotemos con $s1$ y $s2$ el número de elementos de S_1 y S_2, respectivamente, y denotemos los índices de S_k por $i_1^k < i_2^k < \ldots < i_{sk}^k$, $k = 1, 2$. Construyamos la permutación σ definida mediante

$$\begin{aligned} &\sigma(i_1^1) = 1, \quad \sigma(i_2^1) = 2, \ldots, \quad \sigma(i_{s1}^1) = s_1, \\ &\sigma(i_1^2) = s_1 + 1, \quad \sigma(i_2^2) = s_1 + 2, \ldots, \quad \sigma(i_{s2}^2) = s_1 + s_2, \end{aligned}$$

Demostraremos que si P es la matriz de permutación asociada a σ, entonces $A' = P^T AP$ tiene la forma (1.28). Sea $T_1 = \{1, 2, \ldots, s_1\}$ y $T_2 = \{s_1 + 1, \ldots, s_1 + s_2\}$. Basta demostrar que si $a'_{ij} \neq 0$ e $i \neq j$ entonces $j \in T_2$ si $i \in T_1$, y $j \in T_1$ si $i \in T_2$. Si $a'_{ij} \neq 0$ e $i \neq j$, entonces $a_{\sigma^{-1}(i),\sigma^{-1}(j)} \neq 0$ y por tanto, $\sigma^{-1}(i)$ y $\sigma^{-1}(j)$ están asociados. Puesto que A tiene la propiedad A y puesto que $\sigma^{-1}(i) \neq \sigma^{-1}(j)$ se tiene que o bien $\sigma^{-1}(i) \in S_1$ y $\sigma^{-1}(j) \in S_2$ o $\sigma^{-1}(i) \in S_2$ y $\sigma^{-1}(j) \in S_1$. Por la construcción de σ entonces o bien $i = \sigma(\sigma^{-1}(i)) \in T_1$ y $j = \sigma(\sigma^{-1}(j)) \in T_2$ o bien $i \in T_2$ y $j \in T_1$. Por tanto A' tiene la forma que postula el teorema si T_1 y T_2 son ambos no vacíos o A' es diagonal.

Recíprocamente, si para alguna permutación P, la matriz $A' = P^T AP$ tiene la forma (1.28), entonces A' tiene la propiedad A puesto que A' es una matriz tridiagonal por bloques y, por tanto, está consistentemente ordenada. ■

1.7 Teorema. *Sea A de la forma* (1.20) *una matriz p-cíclica consistentemente ordenada cuyas submatrices diagonales $A_{i,i}$ para $1 \leq i \leq N$ son regulares.*

Si $\omega \neq 0$, si λ es un autovalor no nulo de la matriz $\mathcal{L}_\omega = (I - \omega E)^{-1}\{\omega F + (1-\omega)I\}$ y si μ satisface

$$(\lambda + \omega - 1)^p = \lambda^{p-1}\omega^p \mu^p, \tag{1.29}$$

entonces μ es un autovalor de la matriz por bloques de Jacobi $B = L + U$. Recíprocamente, si μ es un autovalor de B y λ satisface (1.29)*, entonces λ es un autovalor de $\mathcal{L}_\omega$.*

Demostración. Los autovalores de $\mathcal{L}_\omega$ son las raíces del polinomio característico

$$\det(\lambda I - \mathcal{L}_\omega) = 0. \tag{1.30}$$

Como $I - \omega L$ es no singular y $\det(I - \omega L) = 1$,

$$\det(\lambda I - \mathcal{L}_\omega) = \det(I - \omega L)\det(\lambda I - \mathcal{L}_\omega) = \det\{(\lambda + \omega - 1)I - \omega\lambda L - \omega U\}.$$

Entonces, llamando $\phi(\lambda) = \det\{(\lambda + \omega - 1)I - \omega\lambda L - \omega U\}$, (1.30) es equivalente a

$$\phi(\lambda) = \det\{(\lambda + \omega - 1)I - \omega\lambda L - \omega U\} = 0.$$

Se puede probar que

$$\phi(\lambda) = \det\{(\lambda + \omega - 1)I - \lambda^{(p-1)/p}\omega B\}. \tag{1.31}$$

Como A es p-cíclica, entonces B es débilmente cíclica de índice p y, por tanto, $\lambda^{(p-1)/p}\omega B$ es débilmente cíclica de índice p. Aplicando el teorema de Romanovsky 1.6,

$$\phi(\lambda) = (\lambda + \omega - 1)^m \prod_{i=1}^{r}\{(\lambda + \omega - 1)^p - \lambda^{p-1}\omega^p \mu_i^p\}, \tag{1.32}$$

donde los μ_i son no nulos si $r \geq 1$.

Veamos en primer lugar la segunda implicación del teorema. Supongamos que μ es un autovalor de B y que λ satisface (1.29). Entonces, uno de los factores de (1.32) desaparece y tendríamos que $\phi(\lambda) = 0$ lo que implica directamente que λ es autovalor de $\mathcal{L}_\omega$.

Veamos ahora la primera parte del teorema. Sea $\omega \neq 0$ y sea λ un autovalor no nulo de $\mathcal{L}_\omega$. Entonces, $\phi(\lambda) = 0$, luego al menos uno de los factores de (1.32) ha de ser cero. Si $\mu \neq 0$ y μ satisface (1.29), entonces $\lambda + \omega - 1 \neq 0$. Luego debe ser $(\lambda + \omega - 1)^p = \lambda^{p-1}\omega^p \mu_i^p$, para algún i, $1 \leq i \leq r$, donde μ_i es no nulo. Combinando esto con (1.29) tenemos que $\lambda^{p-1}\omega^p(\mu^p - \mu_i^p) = 0$ y como λ y μ son no nulos, debe ser $\mu^p = \mu_i^p$. Tomando raíces p-ésimas, $\mu = \mu_i e^{2\pi i r/p}$ con r entero que satisface $0 \leq r < p$. Como B es débilmente cíclica, entonces $\mu \neq 0$ es autovalor de B. Si $\mu = 0$, $\omega \neq 0$ y λ es un autovalor no nulo de $\mathcal{L}_\omega$, como μ satisface (1.29) entonces de (1.31) se deduce que $\phi(\lambda) = \det B = 0$, luego $\mu = 0$ es autovalor de B.

■

1.2 Corolario. *Sea la matriz A con la partición* (1.20) *una matriz p-cíclica y consistentemente ordenada cuyas submatrices diagonales son no singulares. Si μ es un autovalor de la matriz de Jacobi por bloques $B = L+U$, entonces μ^p es un autovalor de $\mathcal{L}_1$. Recíprocamente, si λ es un autovalor no nulo de $\mathcal{L}_1$ y $\mu^p = \lambda$, entonces μ es un autovalor de B. Por tanto, el método iterativo de Jacobi por bloques converge si y solo si el método iterativo de Gauss-Seidel por bloques converge, y si ambos convergen, entonces*

$$\rho(\mathcal{L}_1) = (\rho(B))^p < 1.$$

Supongamos que la matriz A es una matriz p-cíclica y consistentemente ordenada de la forma (1.20) y cuyas submatrices diagonales son no singulares. Supongamos también que la matriz de Jacobi $B = L+U$ es convergente. Por el corolario anterior, también la matriz $\mathcal{L}_1$ es convergente y, por continuidad, también es convergente en un intervalo de ω que contenga al 1.

Buscamos ω_b tal que $\rho(\mathcal{L}_{\omega_b}) = \text{mín}_{\omega\in\mathbb{R}}\, \rho(\mathcal{L}_\omega)$.

En concreto, si los autovalores de B^p son reales y no negativos, el valor de ω_b que minimiza $\rho(\mathcal{L}_\omega)$ de forma única es la única raiz real y positiva de la ecuación

$$(\rho(B)\omega_b)^p = [p^p(p-1)^{1-p}](\omega_b - 1)$$

donde $\rho(B)$ denota el radio espectral de la matriz de Jacobi por bloques.

1.8 Teorema. *Sea la matriz A de la forma* (1.20) *una matriz p-cíclica consistentemente ordenada cuyas submatrices diagonales, $A_{i,i}$ para $1 \le i \le N$ son regulares. Si todos los autovalores de la p-ésima potencia de la matriz B de Jacobi por bloques son reales y no negativos y $0 \le \rho(B) < 1$ entonces para ω_b, la única raiz real , positiva y menor que $p/(p-1)$ de*

$$(\rho(B)\omega_b)^p = [p^p(p-1)^{1-p}](\omega_b - 1)$$

se tiene que

1. $\rho(\mathcal{L}_{\omega_b}) = (\omega_b - 1)(p-1)$;

2. $\rho(\mathcal{L}_\omega) > \rho(\mathcal{L}_{\omega_b}), \quad \omega \neq \omega_b$.

Además, la matriz $\mathcal{L}_\omega$ del método SOR es convergente para todo ω con $0 < \omega < p/(p-1)$.

Demostración. Para $p = 2$ la raíz de la ecuación anterior que nos interesa se puede expresar como

$$\omega_b = \frac{2}{1+\sqrt{1-\rho^2(B)}} = 1 + \left(\frac{\rho(B)}{1+\sqrt{1-\rho^2(B)}}\right)^2. \tag{1.33}$$

Veamos que, efectivamente, la expresión (1.33) minimiza $\rho(\mathcal{L}_\omega)$ para $p = 2$ (la misma prueba, con las correspondientes modificaciones, sería válida para el caso general). Si los autovalores de B^2 son números reales no

negativos, entonces como B es débilmente cíclica de índice 2, los autovalores no nulos de B , μ_i vienen dados en pares de autovalores con signos opuestos. Por tanto, $-\rho(B) \leq \mu_i \leq \rho(B) < 1$. Por el teorema anterior sabemos que los autovalores de la matriz para el método SOR, λ, y los autovalores de la matriz de Jacobi, μ, satisfacen $(\lambda + \omega - 1)^2 = \lambda\omega^2\mu^2$. Operando tenemos

$$\frac{\lambda + \omega - 1}{\omega} = \pm\lambda^{1/2}\mu.$$

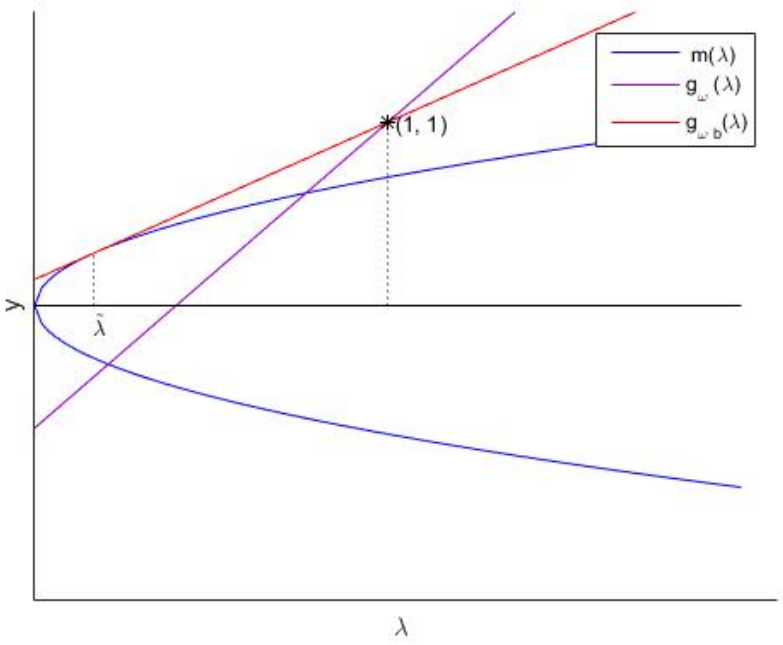
Definamos entonces

$$g_\omega(\lambda) := \frac{\lambda + \omega - 1}{\omega}, \quad \omega \neq 0$$

y

$$m(\lambda) := \lambda^{1/2}\mu, \quad 0 \leq \mu \leq \rho(B) < 1.$$

Figura 1.1: Optimización del parámetro para el método SOR



Como se puede observar en la Figura 1.1 , $g_\omega(\lambda)$ es una recta que pasa por el punto $(1, 1)$ y cuya pendiente decrece al aumentar el valor de ω. Como teníamos $g_\omega(\lambda) = m(\lambda)$, se puede interpretar como la intersección de estas dos funciones. Además, el valor óptimo para ω se obtendrá cuando $g_\omega(\lambda)$ sea tangente a $m(\lambda)$, ya que el mayor de los dos puntos de intersección alcanzará

su valor mínimo cuando esto ocurra. En dicho caso será

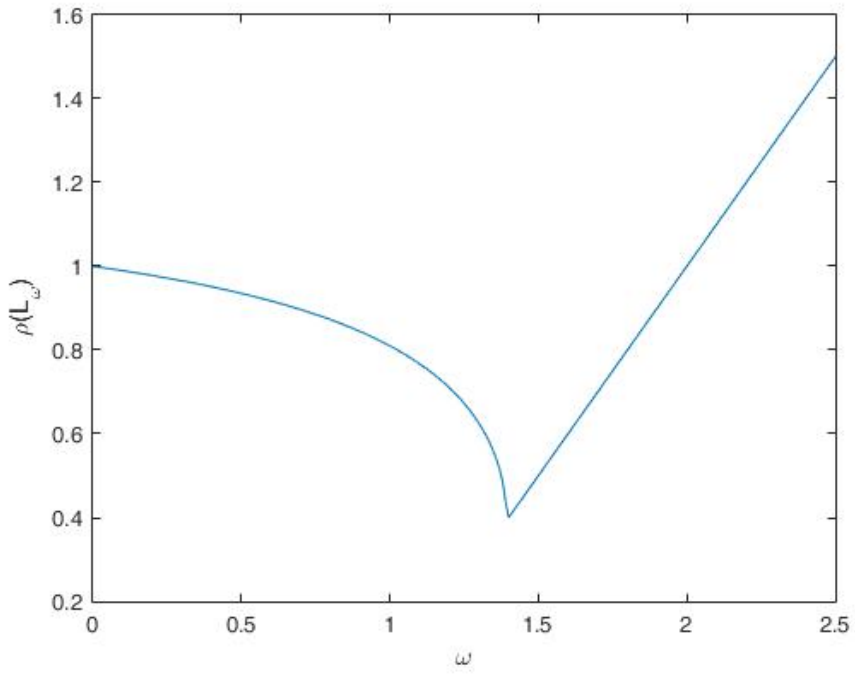
$$\omega_b = \frac{2}{1+\sqrt{1-\mu^2}}.$$

Entonces, el valor de la abscisa en este punto es $\tilde{\lambda} = \omega_b - 1$.Por tanto,

$$\min_{\omega \in \mathbb{R}} \rho(\mathcal{L}_\omega) = \rho(\mathcal{L}_{\omega_b}) = \omega_b - 1.$$

■

Figura 1.2: Optimización del parámetro para el método SOR



Como podemos apreciar en la Figura 1.2, donde el pico corresponde al valor óptimo ω_b , es mejor sobreestimar el valor óptimo en una pequeña cantidad que subestimar en esta misma cantidad.

Capítulo 2

Métodos SOR para sistemas aumentados

2.1. Los sistemas aumentados y su interés

Sean $A \in \mathbb{R}^{m\times m}$ una matriz simétrica y definida positiva y $B \in \mathbb{R}^{m\times n}$. Consideramos el siguiente sistema lineal aumentado

$$\begin{pmatrix} A & B \\ B^T & 0 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} b \\ q \end{pmatrix} \tag{2.1}$$

donde $b \in \mathbb{R}^m$ y $q \in \mathbb{R}^n$ son dos vectores dados. En ocasiones, denotamos a la matriz del sistema (2.1) con la letra $\mathcal{A}$.

Cuando en (2.1) las matrices A y B son grandes y dispersas, es cuando los métodos iterativos que vamos a tratar toman mayor relevancia, puesto que resolverán el sistema con mayor eficiencia que los métodos directos.

El sistema aumentado (2.1) aparece en diversos problemas como pueden ser problemas de optimización sujetos a restricciones, el método de los elementos finitos para resolver las ecuaciones de Navier-Stokes, problemas de elasticidad, problemas de mínimos cuadrados o problemas de puntos de silla.

Muchas veces, por simplicidad, reescribiremos el sistema (2.1) como

$$\begin{pmatrix} A & B \\ -B^T & 0 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} b \\ -q \end{pmatrix} \tag{2.2}$$

Para sistemas aumentados, la matriz del segundo bloque diagonal es nula, por lo que el método SOR estándar no es aplicable. Desarrollaremos a continuación el método SOR para sistemas aumentados como (2.1), basándonos en el artículo [5]. Además, estudiaremos en este capítulo la teoría del método SOR generalizado para sistemas aumentados utilizando [2] como referencia.

2.2. El método SOR para sistemas aumentados

Estudiamos el sistema aumentado (2.2), para el cual ya hemos dicho que no podemos aplicar el método SOR estándar por ser cero uno de sus bloques diagonales.

Consideramos entonces la siguiente escisión para la matriz del sistema aumentado

$$\begin{pmatrix} A & B \\ -B^T & 0 \end{pmatrix} = \mathcal{D} - \mathcal{L} - \mathcal{U}$$

donde

$$\mathcal{D} = \begin{pmatrix} A & 0 \\ 0 & Q \end{pmatrix} \quad \mathcal{L} = \begin{pmatrix} 0 & 0 \\ B^T & 0 \end{pmatrix} \quad \mathcal{U} = \begin{pmatrix} 0 & -B \\ 0 & Q \end{pmatrix}$$

y Q es una matriz simétrica y no singular.

Tenemos entonces

$$(\mathcal{D} - \mathcal{L} - \mathcal{U}) \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} b \\ -q \end{pmatrix}$$

es decir,

$$\left(\begin{pmatrix} A & 0 \\ 0 & Q \end{pmatrix} - \begin{pmatrix} 0 & 0 \\ B^T & 0 \end{pmatrix} - \begin{pmatrix} 0 & -B \\ 0 & Q \end{pmatrix} \right) \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} b \\ -q \end{pmatrix}$$

Definimos la matriz de iteración

$$\begin{aligned} \mathcal{M}_\omega &= (\mathcal{D} - \omega\mathcal{L})^{-1}[(1-\omega)\mathcal{D} + \omega\mathcal{U}] \\ &= \begin{pmatrix} A & 0 \\ -\omega B^T & Q \end{pmatrix}^{-1} \begin{pmatrix} (1-\omega)A & -\omega B \\ 0 & Q \end{pmatrix} \end{aligned} \tag{2.3}$$

de tal forma que podemos describir cada iteración como

$$\begin{pmatrix} x^{(k+1)} \\ y^{(k+1)} \end{pmatrix} = \mathcal{M}_\omega \begin{pmatrix} x^{(k)} \\ y^{(k)} \end{pmatrix} + \omega(\mathcal{D} - \omega\mathcal{L})^{-1} \begin{pmatrix} b \\ -q \end{pmatrix}$$

De esta forma, el método SOR para sistemas aumentados en su expresión componente a componente se escribe como

$$\begin{cases} x^{(k+1)} = (1-\omega)x^{(k)} + \omega A^{-1}(b - By^{(k)}) \\ y^{(k+1)} = y^{(k)} + \omega Q^{-1}(B^T x^{(k+1)} - q) \end{cases} \tag{2.4}$$

En la práctica, Q es una aproximación a la matriz $B^T A^{-1} B$, que es el complemento de Schur del bloque diagonal A en el sistema aumentado. La situación más simple corresponde a tomar $Q = I$, la matriz identidad $n \times n$. El resto del capítulo se centra en caracterizar en términos de los parámetros de convergencia de este método iterativo y la determinación, cuando es posible, de los parámetros óptimos.

2.3. Convergencia del método SOR para sistemas aumentados

Estudiamos ahora la convergencia del método SOR para sistemas aumentados (2.1), para lo que necesitaremos un par de lemas previos.

2.1 Lema. *Supongamos que μ es un autovalor de $Q^{-1}B^TA^{-1}B$. Si λ satisface*

$$(\lambda - 1)(1 - \omega - \lambda) = \lambda\omega^2\mu, \tag{2.5}$$

entonces λ es un autovalor de $\mathcal{M}_\omega$. Recíprocamente, si λ es un autovalor de $\mathcal{M}_\omega$ tal que $\lambda \neq 1$, $\lambda \neq 1-\omega$ y μ satisface (2.5), entonces μ es un autovalor no nulo de $Q^{-1}B^TA^{-1}B$.

Demostración. Sea λ un autovalor de $\mathcal{M}_\omega$, $\lambda \neq 1$, $\lambda \neq 1-\omega$, con autovector $\begin{pmatrix} x \\ y \end{pmatrix}$. Entonces,

$$\mathcal{M}_\omega \begin{pmatrix} x \\ y \end{pmatrix} = \lambda \begin{pmatrix} x \\ y \end{pmatrix}$$

$$\begin{pmatrix} A & 0 \\ -\omega B^T & Q \end{pmatrix}^{-1} \begin{pmatrix} (1-\omega)A & -\omega B \\ 0 & Q \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \lambda \begin{pmatrix} x \\ y \end{pmatrix}$$

$$\begin{pmatrix} (1-\omega)A & -\omega B \\ 0 & Q \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \lambda \begin{pmatrix} A & 0 \\ -\omega B^T & Q \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}$$

$$\begin{cases} (1-\omega)Ax - \omega By = \lambda Ax \\ Qy = -\lambda\omega B^T x + \lambda Qy \end{cases}$$

$$\begin{cases} (1-\omega-\lambda)Ax = \omega By \\ (\lambda - 1)Qy = \lambda\omega B^T x \end{cases}$$

Como A es simétrica y definida positiva entonces es invertible. Luego de la primera igualdad se deduce,

$$(1 - \omega - \lambda)x = \omega A^{-1}By.$$

Como habíamos supuesto que $\lambda \neq 1 - \omega$,

$$x = \frac{\omega}{1 - \omega - \lambda} A^{-1}By.$$

Como Q es no singular, de la segunda ecuación se tiene

$$(\lambda - 1)y = \lambda\omega Q^{-1}B^T x.$$

Sustituyendo ahora lo obtenido para x

$$(\lambda-1)y = \lambda\omega Q^{-1}B^T\left(\frac{\omega}{1-\omega-\lambda}A^{-1}By\right)$$

$$(\lambda-1)(1-\omega-\lambda)y = \lambda\omega^2 Q^{-1}B^TA^{-1}By.$$

Entonces, como λ satisface (2.5), se tiene

$$\lambda\omega^2\mu y = \lambda\omega^2 Q^{-1}B^TA^{-1}By.$$

Simplificando $\lambda\omega^2$,

$$\mu y = Q^{-1}B^TA^{-1}By$$

es decir, μ es autovalor de $Q^{-1}B^TA^{-1}B$ con autovector asociado y.

Recíprocamente, supongamos que μ es autovalor de $Q^{-1}B^TA^{-1}B$ con autovector asociado y,

$$\mu y = Q^{-1}B^TA^{-1}By.$$

Multipliquemos ambos lados por $\lambda\omega^2$,

$$\lambda\omega^2\mu y = \lambda\omega^2 Q^{-1}B^TA^{-1}By.$$

Como λ satisface (2.5),

$$(\lambda-1)(1-\omega-\lambda)y = \lambda\omega^2 Q^{-1}B^TA^{-1}By$$

$$(\lambda-1)y = \frac{\lambda\omega^2}{1-\omega-\lambda}Q^{-1}B^TA^{-1}By.$$

Llamando $x = \frac{\omega}{1-\omega-\lambda}A^{-1}By$, entonces

$$(\lambda-1)y = \lambda\omega Q^{-1}B^Tx$$

$$(\lambda-1)Qy = \lambda\omega B^Tx$$

$$Qy = -\lambda\omega B^Tx + \lambda Qy. \tag{2.6}$$

Como hemos tomado $x = \frac{\omega}{1-\omega-\lambda}A^{-1}By$, entonces

$$(1-\omega-\lambda)x = \omega A^{-1}By$$

$$(1-\omega-\lambda)Ax = \omega By$$

$$(1-\omega)Ax - \omega By = \lambda Ax. \tag{2.7}$$

Entonces, de (2.6) y (2.7) se deduce que λ es un autovalor de $\mathcal{M}_\omega$ con autovector asociado $\begin{pmatrix} x \\ y \end{pmatrix}$.

■

2.2 Lema. *Consideremos la ecuación cuadrática $x^2 - bx + c = 0$, donde b y c son números reales. Ambas raices de la ecuación tienen módulo estrictamente menor que 1 si y solo si $|c| < 1$ y $|b| < 1 + c$.*

Demostración. Denotemos por λ_1 y λ_2 a las raíces de la ecuación cuadrática considerada $x^2 - bx + c = 0$.

Supongamos que ambas raíces tienen módulo menor que 1, esto es, $|\lambda_1| < 1$ y $|\lambda_2| < 1$. Sabemos que $c = \lambda_1\lambda_2$, luego se tiene trivialmente que $|c| = |\lambda_1\lambda_2| < 1$. Además sabemos que $b = \lambda_1 + \lambda_2$, por tanto,

$$1+c-|b| = \begin{cases} 1 + \lambda_1\lambda_2 - (\lambda_1 + \lambda_2) & = (1 - \lambda_1)(1 - \lambda_2) > 0 \quad \text{si } \lambda_1 + \lambda_2 \geq 0 \\ 1 + \lambda_1\lambda_2 + \lambda_1 + \lambda_2 & = (1 + \lambda_1)(1 + \lambda_2) > 0 \quad \text{si } \lambda_1 + \lambda_2 < 0 \end{cases}$$

En ambos casos, $1 + c - |b| > 0$, luego $|b| < 1 + c$.

Recíprocamente, supongamos $|c| < 1$ y $|b| < 1+c$. Entonces, $1+c-|b| > 0$, luego o bien

$$(1 - \lambda_1)(1 - \lambda_2) > 0 \qquad (\lambda_1 + \lambda_2 \geq 0) \tag{2.8}$$

o bien

$$(1 + \lambda_1)(1 + \lambda_2) > 0 \qquad (\lambda_1 + \lambda_2 < 0) \tag{2.9}$$

Si $(1 - \lambda_1)(1 - \lambda_2) > 0$, entonces o bien $\lambda_1 < 1$ y $\lambda_2 < 1$ o bien $\lambda_1 > 1$ y $\lambda_2 > 1$. Como ocurre que $|c| < 1$ y sabemos que $c = \lambda_1\lambda_2$, entonces no puede ser $\lambda_1 > 1, \lambda_2 > 1$. Luego debe ser $\lambda_1 < 1, \lambda_2 < 1$. Ahora, si ocurriera que $\lambda_1 \leq -1$ o $\lambda_2 \leq -1$, entonces como $\lambda_1 + \lambda_2 \geq 0$, debería ser $\lambda_1 \geq 1$ o $\lambda_2 \geq 1$, lo cual es imposible en este caso. Por tanto debe ocurrir $|\lambda_1| < 1$ y $|\lambda_2| < 1$

Si $(1 + \lambda_1)(1 + \lambda_2) > 0$, entonces o bien $\lambda_1 < -1$ y $\lambda_2 < -1$ o bien $\lambda_1 > -1$ y $\lambda_2 > -1$. Como ocurre que $|c| < 1$ y sabemos que $c = \lambda_1\lambda_2$, entonces no puede ser $\lambda_1 < -1, \lambda_2 < -1$. Luego debe ser $\lambda_1 > -1, \lambda_2 > -1$. Ahora, si ocurriera que $\lambda_1 \geq 1$ o $\lambda_2 \geq 1$, entonces como $\lambda_1 + \lambda_2 < 0$, debería ser $\lambda_1 \leq -1$ o $\lambda_2 \leq -1$, lo cual es imposible en este caso. Por tanto debe ocurrir $|\lambda_1| < 1$ y $|\lambda_2| < 1$

■

2.1 Teorema. *Supongamos que B tiene rango máximo y A es simétrica y definida positiva. Asumimos también que todos los autovalores, μ, de $Q^{-1}B^TA^{-1}B$ son reales. Entonces, si $\rho > 0$, el método SOR para sistemas aumentados* (2.1) *converge para todo ω tal que*

$$0 < \omega < \frac{4}{\sqrt{4\rho + 1} + 1},$$

donde ρ es el radio espectral de $Q^{-1}B^TA^{-1}B$.

Demostración. Del primer lema se sigue que

$$\lambda^2 + (\omega^2\mu + \omega - 2)\lambda + 1 - \omega = 0.$$

Entonces, por el segundo lema, $|\lambda| < 1$ si y solo si

$$|1 - \omega| < 1$$

y

$$|\omega^2\mu + \omega - 2| < 1 + 1 - \omega = 2 - \omega.$$

De $|1-\omega| < 1$, se tiene que $-1 < 1-\omega < 1$. Luego, directamente obtenemos que debe ser

$$0 < \omega < 2.$$

De la segunda condición , $|\omega^2\mu + \omega - 2| < 2 - \omega$, se tiene

$$-2 + \omega < \omega^2\mu + \omega - 2 < 2 - \omega \quad \Longrightarrow \quad 0 < \omega^2\mu < 4 - 2\omega.$$

Como ρ es el radio espectral de $Q^{-1}B^TA^{-1}B$, $\rho \geq |\mu| = \mu$, luego

$$0 < \omega^2\mu \leq \omega^2\rho < 4 - 2\omega.$$

Entonces,

$$\begin{aligned}
&\omega^2\rho < 4 - 2\omega \implies 4\omega^2\rho < 16 - 8\omega \implies 4\rho < \frac{16}{\omega^2} - \frac{8}{\omega} \\
&\implies 4\rho + 1 < \frac{16}{\omega^2} - \frac{8}{\omega} + 1 = \left(\frac{4}{\omega} - 1\right)^2 \implies \sqrt{4\rho + 1} < \frac{4}{\omega} - 1 \\
&\implies \sqrt{4\rho + 1} + 1 < \frac{4}{\omega} \implies \omega < \frac{4}{\sqrt{4\rho + 1} + 1} < 2.
\end{aligned}$$

■

2.4. Optimización del parámetro del método SOR para sistemas aumentados

Estudiemos ahora las posibles elecciones del parámetro ω que permitirán optimizar la convergencia del método SOR. Para simplificar la notación denotemos $\rho = \rho(Q^{-1}B^TA^{-1}B)$ y si μ es un autovalor no nulo de la matriz $Q^{-1}B^TA^{-1}B$, llamemos $0 < \mu_0 = \min_{\mu \neq 0} \mu$.

2.2 Teorema. *Sea $\mu_0 > \frac{1}{4}$, entonces*

$$\rho(\mathcal{M}_\omega) = \begin{cases} \sqrt{1-\omega}, & 0 < \omega \leq \frac{2\sqrt{\rho}-1}{\rho} \\ \dfrac{\left[|2 - \omega - \omega^2\rho| + \omega\sqrt{(\omega\rho+1)^2 - 4\rho}\right]}{2}, & \frac{2\sqrt{\rho}-1}{\rho} \leq \omega < \frac{4}{\sqrt{4\rho+1}+1} \end{cases} \tag{2.10}$$

Además, el valor óptimo del parámetro ω_b y $\rho(\mathcal{M}_{\omega_b})$ vienen dados por

$$\omega_b = \frac{2\sqrt{\rho}-1}{\rho} \leq 1$$

y

$$\rho(\mathcal{M}_{\omega_b}) = \frac{|\sqrt{\rho}-1|}{\sqrt{\rho}}.$$

Demostración. Como hemos visto en el primer lema de la sección anterior, λ satisface

$$(\lambda-1)(1-\omega-\lambda) = \lambda\omega^2\mu$$

es decir,

$$\lambda^2 + (\omega^2\mu+\omega-2)\lambda - \omega + 1 = 0.$$

Entonces,

$$\begin{aligned}\lambda =& \frac{-(\omega^2\mu+\omega-2) \pm \sqrt{(\omega^2\mu+\omega-2)^2+4(\omega-1)}}{2} \\ &= 0{,}5[2-\omega-\omega^2\mu \pm \omega\sqrt{(\omega\mu+1)^2-4\mu}].\end{aligned} \tag{2.11}$$

Luego, λ es complejo si $\omega \leq \frac{2\sqrt{\mu}-1}{\mu}$ y es real si $\omega \geq \frac{2\sqrt{\mu}-1}{\mu}$. Por tanto,

$$|\lambda| = \begin{cases} \sqrt{1-\omega} & \text{si } 0 < \omega \leq \frac{2\sqrt{\mu}-1}{\mu} \\ 0{,}5\left[|2-\omega-\omega^2\mu| + \omega\sqrt{(\omega\mu+1)^2-4\mu}\right] & \text{si } \frac{2\sqrt{\mu}-1}{\mu} \leq \omega < \frac{4}{\sqrt{4\mu+1}+1} \end{cases}$$

Entonces, como $\frac{2\sqrt{\mu}-1}{\mu}$ es una función monótona decreciente y $\sqrt{1-\omega} \geq 1-\omega$ para todo $\omega \leq 1$, si $\rho_0 \geq \frac{1}{4}$, se obtiene (2.10).

Para obtener el parámetro óptimo trabajemos con la expresión

$$(\lambda-1)(1-\omega-\lambda) = \lambda\omega^2\mu$$

hasta poder escribirla como $g_\omega(\lambda) = f_\omega(\lambda)$ para ciertas funciones de las que podamos hacer un análisis más apropiado.

$$\begin{aligned}&(\lambda-1)(1-\omega-\lambda) = \lambda\omega^2\mu \implies \lambda - \lambda\omega - \lambda^2 - 1 + \omega + \lambda = \lambda\omega^2\mu \\ &\lambda^2 - 2\lambda + 1 + \lambda\omega - \omega = -\lambda\omega^2\mu \implies (\lambda-1)^2 + (\lambda-1)\omega = -\lambda\omega^2\mu \\ &(\lambda-1)^2 + \left(\frac{\omega}{2}\right)^2 + (\lambda-1)\omega = \left(\frac{\omega}{2}\right)^2 - \lambda\omega^2\mu \implies \\ &\left(\lambda - 1 + \frac{\omega}{2}\right)^2 = \left(\frac{\omega}{2}\right)^2(1-4\lambda\mu) \implies \left(\frac{\lambda-1+\frac{\omega}{2}}{\frac{\omega}{2}}\right)^2 = 1-4\lambda\mu\end{aligned}$$

Denotamos así

$$g_\omega(\lambda) = \left(\frac{\lambda-1+\frac{\omega}{2}}{\frac{\omega}{2}}\right)^2$$

y

$$f_\omega(\lambda) = 1 - 4\lambda\mu.$$

Observemos que $g_\omega(1) = 1$ y $f_\omega(0) = 1$, esto es,$g_\omega(\lambda)$ pasa por el punto $(1, 1)$ y $f_\omega(\lambda)$ pasa por $(0, 1)$. Entonces, la recta $f_\omega$cruza la curva parabólica g_ω. El valor óptimo para el parámetro ω se obtendrá cuando f_ω sea tangente a g_ω. Esto sucede cuando

$$\omega_b = \frac{2\sqrt{\rho} - 1}{\rho}$$

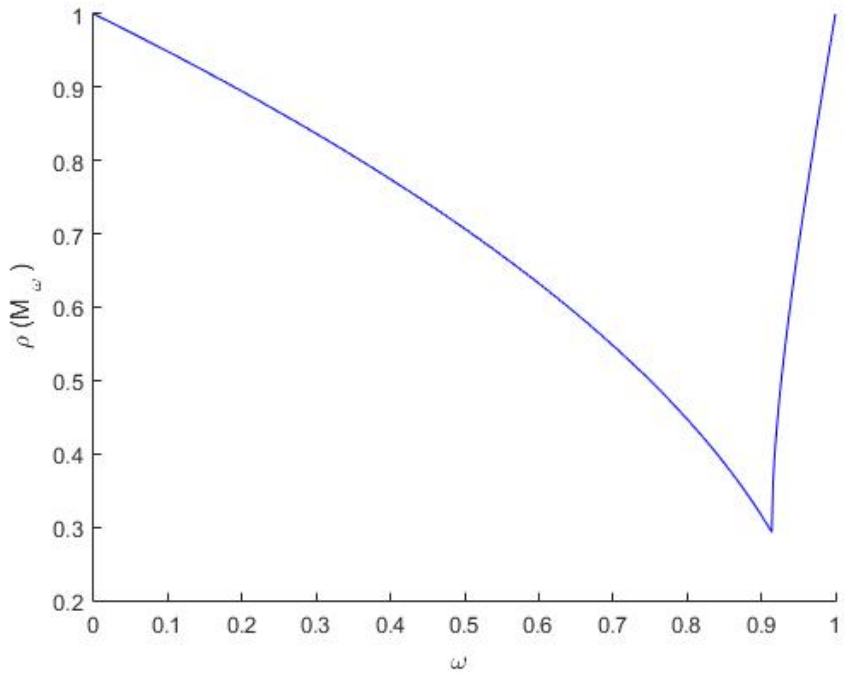
ya que $g'_\omega(\lambda) = 8(\lambda - 1 + \omega/2)/\omega^2$ y $f'_\omega(\lambda) = -4\mu$.

Además,

$$\rho(\mathcal{M}_{\omega_b}) = \sqrt{1 - \omega_b} = \frac{|\sqrt{\rho} - 1|}{\sqrt{\rho}}.$$

■

Figura 2.1: Curva del radio espectral del método SOR aumentado para $\rho = 2$



2.3 Teorema. *Supongamos que* $\rho_0 \leq \frac{1}{4}$. *Entonces,*

$$\rho(\mathcal{M}_\omega) = \begin{cases} \dfrac{\left[|2-\omega-\omega^2\mu_0| + \omega\sqrt{(\omega\mu_0+1)^2-4\mu_0}\right]}{2}, & 0 < \omega \leq \omega_b \\ \dfrac{\left[|2-\omega-\omega^2\rho| + \omega\sqrt{(\omega\rho+1)^2-4\rho}\right]}{2}, & \omega_b < \omega < \frac{4}{\sqrt{4\rho+1}+1} \end{cases} \tag{2.12}$$

donde $\omega_b < 2$ *es la raíz positiva de la ecuación*

$$|2-\omega-\omega^2\mu_0| + \omega\sqrt{(\omega\mu_0+1)^2-4\mu_0} = |2-\omega-\omega^2\rho| + \omega\sqrt{(\omega\rho+1)^2-4\rho}.$$

Demostración. De (2.11) deducimos que para todo $\mu_0 < \mu \leq \frac{1}{4}$, ocurre $|\lambda(\mu)| \leq |\lambda(\mu_0)|$. Tenemos

$$|\lambda(\mu_0)| = 0{,}5\left[|2-\omega-\omega^2\mu_0| + \omega\sqrt{(\omega\mu_0+1)^2-4\mu_0}\right].$$

Además, $|\lambda(\mu_0)| \geq \sqrt{1-\omega}$ para $\mu > \frac{1}{4}$.

Para $\omega > \frac{(2\sqrt{\rho}-1)}{\rho}$,se tiene $|\lambda(\mu_0)| \leq |\lambda(\rho)|$, donde

$$|\lambda(\rho)| = 0{,}5\left[|2-\omega-\omega^2\rho| + \omega\sqrt{(\omega\rho+1)^2-4\rho}\right].$$

Combinando las dos igualdades que hemos obtenido, llegamos al resultado que queríamos probar.

■

2.5. El método SOR generalizado para sistemas aumentados

Con el objetivo de mejorar la velocidad de convergencia del método SOR para sistemas aumentados, se desarrolla, mediante la introducción de un nuevo parámetro, el método generalizado SOR (GSOR).

Consideramos, para el sistema aumentado (2.2), la misma escisión que en el método SOR, es decir,

$$\begin{pmatrix} A & B \\ -B^T & 0 \end{pmatrix} = \mathcal{D} - \mathcal{L} - \mathcal{U}$$

donde

$$\mathcal{D} = \begin{pmatrix} A & 0 \\ 0 & Q \end{pmatrix} \quad \mathcal{L} = \begin{pmatrix} 0 & 0 \\ B^T & 0 \end{pmatrix} \quad \mathcal{U} = \begin{pmatrix} 0 & -B \\ 0 & Q \end{pmatrix}$$

y $Q \in \mathbb{R}^{n\times n}$ es una matriz simétrica y no singular.

Sean ahora ω y τ dos reales no nulos. Llamemos

$$\Omega = \begin{pmatrix} \omega I_m & 0 \\ 0 & \tau I_n \end{pmatrix}$$

donde $I_m \in \mathbb{R}^{m\times m}$ e $I_n \in \mathbb{R}^{n\times n}$ son, respectivamente, la matriz identidad $m \times m$ y $n \times n$.

Generalizamos, para el sistema aumentado (2.2), el método SOR como sigue

$$\begin{pmatrix} x^{(k+1)} \\ y^{(k+1)} \end{pmatrix} = (\mathcal{D}-\Omega\mathcal{L})^{-1}[(I-\Omega)\mathcal{D}+\Omega\mathcal{U}] \begin{pmatrix} x^{(k)} \\ y^{(k)} \end{pmatrix} + (\mathcal{D}-\Omega\mathcal{L})^{-1}\Omega \begin{pmatrix} b \\ -q \end{pmatrix}.$$

Equivalentemente, si definimos las matrices

$$\begin{aligned} \mathcal{H}(\omega,\tau) &= (\mathcal{D}-\Omega\mathcal{L})^{-1}[(I-\Omega)\mathcal{D}+\Omega\mathcal{U}] \\ &= \begin{pmatrix} (1-\omega)I & -\omega A^{-1}B \\ (1-\omega)\tau Q^{-1}B^T I & -\omega\tau Q^{-1}B^T A^{-1}B \end{pmatrix} \end{aligned}$$

y

$$\mathcal{M}(\omega,\tau) = \Omega^{-1}(\mathcal{D}-\Omega\mathcal{L}) = \begin{pmatrix} \frac{1}{\omega}A & 0 \\ -B^T & \frac{1}{\tau}Q \end{pmatrix}$$

podemos escribir la versión matricial del método GSOR para sistemas ampliados

$$\begin{pmatrix} x^{(k+1)} \\ y^{(k+1)} \end{pmatrix} = \mathcal{H}(\omega,\tau) \begin{pmatrix} x^{(k)} \\ y^{(k)} \end{pmatrix} + \mathcal{M}(\omega,\tau)^{-1} \begin{pmatrix} b \\ -q \end{pmatrix}.$$

De esta forma, el método generalizado SOR en su versión componente a componente se escribe como

$$\begin{cases} x^{(k+1)} = (1-\omega)x^{(k)} + \omega A^{-1}(b - By^{(k)}) \\ y^{(k+1)} = y^{(k)} + \tau Q^{-1}(B^T x^{(k+1)} - q) \end{cases} \tag{2.13}$$

donde Q es una matriz que aproxima al complemento de Schur, $B^T A^{-1} B$, y sirve como preacondicionamiento.

Observemos que de forma trivial cuando $\tau = \omega$, el método generalizado SOR se reduce al método SOR expuesto en el apartado anterior.

Denotemos ahora

$$\mathcal{N}(\omega,\tau) = \mathcal{M}(\omega,\tau) - \mathcal{A} = \begin{pmatrix} (\frac{1}{\omega}-1)A & -B \\ 0 & \frac{1}{\tau}Q \end{pmatrix}$$

Entonces, podemos considerar la siguiente escisión para la matriz $\mathcal{A}$, matriz de coeficientes del sistema aumentado (2.2), $\mathcal{A} = \mathcal{M}(\omega,\tau) - \mathcal{N}(\omega,\tau)$. Entonces, es fácil ver que

$$\mathcal{H}(\omega,\tau) = \mathcal{M}(\omega,\tau)^{-1}\mathcal{N}(\omega,\tau)$$

es la matriz de iteración del método SOR generalizado.

2.6. Teoremas de convergencia para el método SOR generalizado

Hemos visto que la matriz $\mathcal{H}(\omega,\tau)$ es la matriz de iteración del método GSOR. Por tanto, el método será convergente si y solo si el radio espectral de la matriz $\mathcal{H}(\omega,\tau)$ es estrictamente menor que 1, esto es, $\rho(\mathcal{H}(\omega,\tau)) < 1$.

2.4 Teorema. *Sean $A \in \mathbb{R}^{m\times m}$ y $Q \in \mathbb{R}^{n\times n}$ matrices simétricas y definidas positivas. Sea también $B \in \mathbb{R}^{m\times n}$ de rango máximo. Denotemos $\mathcal{J} = Q^{-1}B^TA^{-1}B$ y $\mu_{\text{mín}}$, $\mu_{\text{máx}}$ al menor y mayor autovalor de $\mathcal{J}$, respectivamente. Entonces, el método GSOR converge si ω verifica que $0 < \omega < 2$ y τ satisfacela condición:*

$$0 < \tau < \frac{2(2-\omega)}{\omega\mu_{\text{máx}}}.$$

Demostración. Por las condiciones establecidas en el enunciado del teorema, se tiene que todos los autovalores, μ, de la matriz $\mathcal{J} = Q^{-1}B^TA^{-1}B$ son reales y no nulos.

Sea λ un autovalor no nulo de la matriz de iteración $\mathcal{H}(\omega,\tau)$, con autovector asociado $(x^T, y^T)^T \in \mathbb{R}^{m+n}$. Recordemos que

$$\mathcal{H}(\omega,\tau) = \mathcal{M}(\omega,\tau)^{-1}\mathcal{N}(\omega,\tau) = \begin{pmatrix} \frac{1}{\omega}A & 0 \\ -B^T & \frac{1}{\tau}Q \end{pmatrix}^{-1} \begin{pmatrix} (\frac{1}{\omega}-1)A & -B \\ 0 & \frac{1}{\tau}Q \end{pmatrix}$$

Entonces, la relación $\mathcal{H}(\omega,\tau)\begin{pmatrix} x \\ y \end{pmatrix} = \lambda\begin{pmatrix} x \\ y \end{pmatrix}$ se puede escribir como:

$$\begin{pmatrix} (\frac{1}{\omega}-1)A & -B \\ 0 & \frac{1}{\tau}Q \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} = \lambda\begin{pmatrix} \frac{1}{\omega}A & 0 \\ -B^T & \frac{1}{\tau}Q \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix}$$

Equivalentemente,

$$\begin{pmatrix} (1-\omega)A & -\omega B \\ 0 & Q \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} = \lambda\begin{pmatrix} A & 0 \\ -\tau B^T & Q \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix}$$

Entonces tenemos

$$\begin{cases} (1-\omega)Ax - \omega By = \lambda Ax \\ Qy = -\lambda\tau B^Tx + \lambda Qy \end{cases}$$

luego

$$\begin{cases} (1-\omega-\lambda)Ax = \omega By \\ \lambda\tau B^Tx = (\lambda-1)Qy \end{cases} \tag{2.14}$$

De la primera ecuación de (2.14) obtenemos, $(1-\omega-\lambda)x = \omega A^{-1}By$. Luego cuando $\lambda \neq 1-\omega$, se tiene que

$$\lambda\tau(1-\omega-\lambda)B^Tx = \lambda\tau\omega B^TA^{-1}By.$$

De aquí y de la segunda ecuación de (2.14), se tiene que

$$\begin{aligned}&(\lambda-1)(1-\omega-\lambda)Qy=\lambda\tau\omega B^TA^{-1}By \implies\\&(\lambda-1)(1-\omega-\lambda)y=\lambda\tau\omega Q^{-1}B^TA^{-1}By \implies\\&(\lambda-1)(1-\omega-\lambda)y=\lambda\tau\omega\mathcal{J}y\end{aligned}$$

Si $\lambda=1-\omega\neq 0$, entonces de la primera ecuación de (2.14) tenemos $0=By$ y de la segunda ecuación de (2.14), $-\omega Qy=\lambda\tau B^Tx$. Por tanto, $y=0$ y $x\in\ker(B^T)$, donde $\ker(B^T)$ es el núcleo de B^T. Luego, $\lambda=1-\omega$ es un autovalor de $\mathcal{H}(\omega,\tau)$ cuyo autovector correspondiente es $(x^T,0)^T$, donde $x\in\ker(B^T)$.

Entonces, los autovalores λ (excepto $\lambda=1-\omega$)de la matriz $\mathcal{H}(\omega,\tau)$ y los autovalores μ de la matriz $\mathcal{J}$ satisfacen la relación siguiente

$$(1-\omega-\lambda)(\lambda-1)=\lambda\tau\omega\mu$$

es decir, λ satisface la ecuación cuadrática

$$\lambda^2+(\tau\omega\mu+\omega-2)\lambda+1-\omega=0 \tag{2.15}$$

Entonces, de acuerdo con el lema de Young, sabemos que tanto $\lambda=1-\omega$ como las dos raíces de la ecuación cuadrática anterior, satisfacen $|\lambda|<1$ si y solo si $|1-\omega|<1$ y $|\tau\omega\mu+\omega-2|<2-\omega$. De la primera ecuación se obtiene directamente que $0<\omega<2$ y de la segunda,

$$\begin{aligned}&-2<\tau\omega\mu+\omega-2<2-\omega \implies 0<\tau\omega\mu<4-2\omega \implies\\&0<\tau<\frac{2(2-\omega)}{\omega\mu} \implies 0<\tau<\frac{2(2-\omega)}{\omega\mu_{\text{máx}}}.\end{aligned}$$

■

De forma análoga se puede probar que si $Q\in\mathbb{C}^{n\times n}$ es simétrica y definida negativa, entonces el método GSOR converge si ω satisface $0<\omega<2$ y τ satisface

$$\frac{2(2-\omega)}{\omega\mu_{\text{mín}}}<\tau<0.$$

2.1 Corolario. *Sean $A\in\mathbb{R}^{m\times m}$ una matriz simétrica y positiva definida,$B\in\mathbb{R}^{m\times n}$ de rango máximo y $Q\in\mathbb{R}^{n\times n}$ no singular y simétrica. Sea también $\mathcal{J}=Q^{-1}B^TA^{-1}B$. Si μ es un autovalor de la matriz $\mathcal{J}$, entonces el λ determinado por la ecuación cuadrática* (2.15) *es un autovalor de la matriz $\mathcal{H}(\omega,\tau)$. Recíprocamente, si λ es un autovalor de la matriz $\mathcal{H}(\omega,\tau)$, entonces el μ determinado por* (2.15) *es un autovalor de la matriz $\mathcal{J}$. Por tanto, los autovalores no nulos de la matriz $\mathcal{H}(\omega,\tau)$ vienen dados por $\lambda=1-\omega$ o*

$$\lambda=\frac{1}{2}\left(2-\omega-\tau\omega\mu\pm\sqrt{(2-\omega-\tau\omega\mu)^2-4(1-\omega)}\right).$$

2.2 Corolario. *Sean $A \in \mathbb{R}^{m\times m}$ una matriz simétrica y positiva definida,$B \in \mathbb{R}^{m\times n}$ de rango máximo y $Q \in \mathbb{R}^{n\times n}$ no singular y simétrica. Sea también $\mathcal{J} = Q^{-1}B^TA^{-1}B$. Si μ es un autovalor de la matriz $\mathcal{J}$, entonces el λ determinado por la ecuación cuadrática* (2.15) *con $\tau = \omega$ es un autovalor de la matriz $\mathcal{L}(\omega)$.Recíprocamente, si λ es un autovalor de la matriz $\mathcal{L}(\omega)$, entonces el μ determinado por* (2.15) *con $\tau = \omega$ es un autovalor de la matriz $\mathcal{J}$. Por tanto, los autovalores no nulos de la matriz $\mathcal{L}(\omega)$ vienen dados por $\lambda = 1 - \omega$ o*

$$\lambda = \frac{1}{2}\left(2 - \omega - \omega^2\mu \pm \sqrt{(2 - \omega - \omega^2\mu)^2 - 4(1-\omega)}\right).$$

2.7. Optimización de los parámetros en el método SOR generalizado

Determinamos en esta sección los valores óptimos de los parámetros del método GSOR y el correspondiente factor óptimo de convergencia. Para aligerar la notación llamemos:

$\mu_k (k = 1, 2, \ldots, n)$ son los autovalores de la matriz $\mathcal{J} = Q^{-1}B^TA^{-1}B$
$\mu_{\text{mín}} = \min\limits_{1\le k\le n} \mu_k$
$\mu_{\text{máx}} = \max\limits_{1\le k\le n} \mu_k$

Asumimos, sin pérdida de generalidad, que estos autovalores están ordenados, es decir

$$0 < \mu_{\text{mín}} = \mu_1 \le \mu_2 \le \ldots \le \mu_{n-1} \le \mu_n = \mu_{\text{máx}}$$

Hemos probado en la sección anterior que los parámetros del método deben satisfacer

$$0 < \omega < 2 \quad \text{y} \quad 0 < \tau < \frac{2(2-\omega)}{\omega\mu_{\text{máx}}}$$

Denotemos también

$$\begin{cases} f_1(\omega,\tau,\mu) = & \frac{1}{2}\left(2 - \omega - \tau\omega\mu + \sqrt{(2-\omega-\tau\omega\mu)^2 - 4(1-\omega)}\right) \\ & \text{para} \quad \frac{4\tau\mu}{(1+\tau\mu)^2} \le \omega < \frac{2}{1+\tau\mu} \quad \text{y} \quad \tau\mu < 1 \\ f_2(\omega,\tau,\mu) = & \frac{1}{2}\left(\tau\omega\mu + \omega - 2 + \sqrt{(\tau\omega\mu+\omega-2)^2 - 4(1-\omega)}\right) \\ & \text{para} \quad \omega \ge \frac{4\tau\mu}{(1+\tau\mu)^2} \quad \text{y} \quad \tau\mu > 1 \\ & \text{o} \quad \omega > \frac{2}{1+\tau\mu} \quad \text{y } \tau\mu < 1 \\ f_3(\omega,\tau,\mu) = & g(\omega) = \sqrt{1-\omega} \\ & \text{para } \omega \le \frac{4\tau\mu}{(1+\tau\mu)^2} \end{cases}$$

Estas funciones se obtienen de calcular el valor absoluto de los autovalores no nulos de la matriz de iteración $\mathcal{H}(\omega,\tau)$ del método GSOR, donde hemos distinguidos tres casos según sea positivo o negativo el discriminante de la ecuación cuadrática y según el signo del término $2-\omega-\tau\omega\mu$.

Se puede ver que $f_j(\omega,\tau,\mu) \geq \sqrt{1-\omega} \geq 1-\omega, \quad j=1,2.$

Estudiemos ahora la monotonía de estas funciones.

$$\left\{\begin{array}{lll} \frac{\partial f_1(\omega,\tau,\mu)}{\partial\mu} & = & -\frac{\tau\omega}{2}\left(1+\frac{2-\omega-\tau\omega\mu}{\sqrt{(2-\omega-\tau\omega\mu)^2-4(1-\omega)}}\right) \\ \frac{\partial f_2(\omega,\tau,\mu)}{\partial\mu} & = & \frac{\tau\omega}{2}\left(1+\frac{\tau\omega\mu+\omega-2}{\sqrt{(\tau\omega\mu+\omega-2)^2-4(1-\omega)}}\right) \\ \frac{\partial f_3(\omega,\tau,\mu)}{\partial\mu} & = & \frac{dg(\omega)}{d\mu}=0 \end{array}\right.$$

Luego,

$$\left\{\begin{array}{lll} \frac{\partial f_1(\omega,\tau,\mu)}{\partial\mu}<0 & \text{para} & \frac{4\tau\mu}{(1+\tau\mu)^2}<\omega<\frac{2}{1+\tau\mu} \quad \text{y} \quad \tau\mu<1 \\ \frac{\partial f_2(\omega,\tau,\mu)}{\partial\mu}>0 & \text{para} & \omega>\frac{4\tau\mu}{(1+\tau\mu)^2} \quad \text{y} \quad \tau\mu>1 \\ & \text{o} & \omega>\frac{2}{1+\tau\mu} \quad \text{y} \quad \tau\mu<1 \end{array}\right.$$

Es decir, la función $f_1(\omega,\tau,\mu)$ decrece con respecto a μ para $\frac{4\tau\mu}{(1+\tau\mu)^2} < \omega < \frac{2}{1+\tau\mu}$ y $\tau\mu<1$.

La función $f_2(\omega,\tau,\mu)$ crece con respecto a μ para $\omega>\frac{4\tau\mu}{(1+\tau\mu)^2}$ y $\tau\mu>1$ o $\omega>\frac{2}{1+\tau\mu}$ y $\tau\mu<1$.

Por otro lado,

$$\left\{\begin{array}{lll} \frac{\partial f_1(\omega,\tau,\mu)}{\partial\omega} & = & -\frac{1}{2}\left(1+\tau\mu+\frac{(2-\omega-\tau\omega\mu)(1+\tau\mu)-2}{\sqrt{(2-\omega-\tau\omega\mu)^2-4(1-\omega)}}\right) \\ \frac{\partial f_2(\omega,\tau,\mu)}{\partial\omega} & = & \frac{1}{2}\left(1+\tau\mu+\frac{(\tau\omega\mu+\omega-2)(1+\tau\mu)+2}{\sqrt{(\tau\omega\mu+\omega-2)^2-4(1-\omega)}}\right) \\ \frac{\partial f_3(\omega,\tau,\mu)}{\partial\omega} & = & \frac{dg(\omega)}{d\mu}=-\frac{1}{2\sqrt{1-\omega}} \end{array}\right.$$

Luego,

$$\left\{\begin{array}{lll} \frac{\partial f_1(\omega,\tau,\mu)}{\partial\omega}>0 & \text{para} & \frac{4\tau\mu}{(1+\tau\mu)^2}<\omega<\frac{2}{1+\tau\mu} \quad \text{y} \quad \tau\mu<1 \\ \frac{\partial f_2(\omega,\tau,\mu)}{\partial\omega}>0 & \text{para} & \omega>\frac{4\tau\mu}{(1+\tau\mu)^2} \quad \text{y} \quad \tau\mu>1 \\ & \text{o} & \omega>\frac{2}{1+\tau\mu} \quad \text{y} \quad \tau\mu<1 \\ \frac{\partial f_3(\omega,\tau,\mu)}{\partial\omega}<0 & \text{para } \omega\leq\frac{4\tau\mu}{(1+\tau\mu)^2} & \end{array}\right.$$

Es decir, la función $f_1(\omega,\tau,\mu)$ crece con respecto a ω para $\frac{4\tau\mu}{(1+\tau\mu)^2} < \omega < \frac{2}{1+\tau\mu}$ y $\tau\mu<1$.

La función $f_2(\omega,\tau,\mu)$ crece con respecto a ω para $\omega>\frac{4\tau\mu}{(1+\tau\mu)^2}$ y $\tau\mu>1$ o $\omega>\frac{2}{1+\tau\mu}$ y $\tau\mu<1$.

La función $f_3(\omega,\tau,\mu)$ decrece con respecto a ω para $\omega\leq\frac{4\tau\mu}{(1+\tau\mu)^2}$.

Además, para cualesquiera $\tilde{\mu}$ y $\tilde{\tau}$ reales positivos para los que las funciones estén bien definidas,

$$\begin{cases} f_1(\omega,\tilde{\tau},\tilde{\mu}) = f_3(\omega,\tilde{\tau},\tilde{\mu}) & \text{para} \quad \omega = \frac{4\tilde{\tau}\tilde{\mu}}{(1+\tilde{\tau}\tilde{\mu})^2} \\ f_2(\omega,\tilde{\tau},\tilde{\mu}) = f_3(\omega,\tilde{\tau},\tilde{\mu}) & \text{para} \quad \omega = \frac{4\tilde{\tau}\tilde{\mu}}{(1+\tilde{\tau}\tilde{\mu})^2} \\ f_1(\omega,\tilde{\tau},\tilde{\mu}) = f_2(\omega,\tilde{\tau},\tilde{\mu}) & \text{para} \quad \omega = \frac{2}{1+\tilde{\tau}\tilde{\mu}} \end{cases}$$

y para dos autovalores de $\mathcal{J}$, $\mu_\alpha, \mu_\beta \in \{\ \mu_k \mid 1 \leq k \leq n\}$ para los que f_1 y f_2 estén bien definidas, tenemos

$$f_1(\omega,\tilde{\tau},\mu_\alpha) = f_2(\omega,\tilde{\tau},\mu_\beta) \quad \text{si} \quad \omega = \frac{4}{\tilde{\tau}(\mu_\alpha+\mu_\beta)+2}$$

Por otra parte, definimos las siguientes funciones

$$\begin{aligned} \omega_-(\tau) &= \tfrac{4\tau\mu_{\text{mín}}}{(1+\tau\mu_{\text{mín}})^2} \\ \omega_+(\tau) &= \tfrac{4\tau\mu_{\text{máx}}}{(1+\tau\mu_{\text{máx}})^2} \\ \omega_0(\tau) &= \tfrac{4}{\tau(\mu_{\text{mín}}+\mu_{\text{máx}})+2} \end{aligned}$$

Usando el primer corolario de la sección anterior, podemos expresar el valor absoluto de los autovalores λ de la matriz $\mathcal{H}(\omega,\tau)$ como:

Cuando $\mu\tau < 1$ o bien $|\lambda| = |1-\omega|$ o bien

$$|\lambda| = \begin{cases} f_1(\omega,\tau,\mu) & \text{para } \frac{4\tau\mu}{(1+\tau\mu)^2} < \omega < \frac{2}{1+\tau\mu} \\ f_2(\omega,\tau,\mu) & \text{para } \omega \geq \frac{2}{1+\tau\mu} \\ g(\omega) & \text{para } \omega \leq \frac{4\tau\mu}{(1+\tau\mu)^2} \end{cases}$$

y cuando $\mu\tau \geq 1$ entonces o bien $|\lambda| = |1-\omega|$ o bien

$$|\lambda| = \begin{cases} f_2(\omega,\tau,\mu) & \text{para } \omega > \frac{4\tau\mu}{(1+\tau\mu)^2} \\ g(\omega) & \text{para } \omega \leq \frac{4\tau\mu}{(1+\tau\mu)^2} \end{cases}$$

2.5 Teorema. *Consideremos el método GSOR. Sea $A \in \mathbb{R}^{m\times m}$ y $Q \in \mathbb{R}^{n\times n}$ simétrica y definida positiva, y sea $B \in \mathbb{R}^{m\times n}$ de rango máximo. Denotemos $\mu_{\text{mín}}$, $\mu_{\text{máx}}$ al menor y mayor autovalor de $\mathcal{J} = Q^{-1}B^TA^{-1}B$, respectivamente. Entonces,*

(i) cuando $\tau \leq \frac{1}{\sqrt{\mu_{\text{mín}}\mu_{\text{máx}}}}$, se tiene que

$$\rho(\mathcal{H}(\omega,\tau)) = \begin{cases} \sqrt{1-\omega} \\ \textit{para } 0 < \omega < \omega_-(\tau). \\ \frac{1}{2}\left(2-\omega-\tau\omega\mu_{\text{mín}} + \sqrt{(2-\omega-\tau\omega\mu_{\text{mín}})^2 - 4(1-\omega)}\right) \\ \textit{para } \omega_-(\tau) \leq \omega \leq \omega_0(\tau). \\ \frac{1}{2}\left(\tau\omega\mu_{\text{máx}}+\omega-2 + \sqrt{(\tau\omega\mu_{\text{máx}}+\omega-2)^2 - 4(1-\omega)}\right) \\ \textit{para } \omega_0(\tau) < \omega < 2. \end{cases} \tag{2.16}$$

(ii) cuando $\frac{1}{\sqrt{\mu_{\text{mín}}\mu_{\text{máx}}}} < \tau < \frac{2(2-\omega)}{\omega\mu_{\text{máx}}}$

$$\rho(\mathcal{H}(\omega,\tau)) = \begin{cases} \sqrt{1-\omega} \\ \textit{para } 0 < \omega < \omega_+(\tau). \\ \frac{1}{2}\left(\tau\omega\mu_{\text{máx}} + \omega - 2 + \sqrt{(\tau\omega\mu_{\text{máx}} + \omega - 2)^2 - 4(1-\omega)}\right) \\ \textit{para } \omega_+(\tau) \leq \omega < 2. \end{cases} \tag{2.17}$$

Además, los parámetros óptimos ω_{opt} *y* τ_{opt} *vienen dados por*

$$\begin{aligned} \omega_{opt} &= \frac{4\sqrt{\mu_{\text{mín}}\mu_{\text{máx}}}}{(\sqrt{\mu_{\text{mín}}}+\sqrt{\mu_{\text{máx}}})^2} \\ \tau_{opt} &= \frac{1}{\sqrt{\mu_{\text{mín}}\mu_{\text{máx}}}} \end{aligned} \tag{2.18}$$

y por tanto,el factor de convergencia óptimo para el método GSOR es

$$\rho(\mathcal{H}(\omega_{opt},\tau_{opt})) = \frac{\sqrt{\mu_{\text{máx}}} - \sqrt{\mu_{\text{mín}}}}{\sqrt{\mu_{\text{máx}}} + \sqrt{\mu_{\text{mín}}}}.$$

Demostración. Para simplificar la prueba vamos a distinguir tres casos en función del parámetro τ.

Caso (a): $\tau \leq \frac{1}{\mu_{\text{máx}}}$.

Caso (b): $\tau \geq \frac{1}{\mu_{\text{mín}}}$.

Caso (c): $\frac{1}{\mu_{\text{máx}}} < \tau < \frac{1}{\mu_{\text{mín}}}$

Caso (a): $\tau \leq \frac{1}{\mu_{\text{máx}}}$. Como $\tau\mu_{\text{máx}} \leq 1$, entonces para todo μ autovalor de $\mathcal{J}$ se tiene $\tau\mu < 1$. Además,

$$\omega_-(\tau) \leq \omega_+(\tau) < 1 < \omega_0(\tau) < 2.$$

Recordemos que para λ , autovalor de $\mathcal{H}(\omega,\tau)$ o bien $|\lambda| = |1-\omega|$ o bien

$$|\lambda| = \begin{cases} f_1(\omega,\tau,\mu) & \text{para } \frac{4\tau\mu}{(1+\tau\mu)^2} < \omega < \frac{2}{1+\tau\mu} \\ f_2(\omega,\tau,\mu) & \text{para } \omega \geq \frac{2}{1+\tau\mu} \\ g(\omega) & \text{para } \omega \leq \frac{4\tau\mu}{(1+\tau\mu)^2} \end{cases}$$

Consideremos fijos $\omega, \tau > 0$, entonces por la monotonía estudiada para las funciones $f_i(\omega,\tau,\mu), i = 1,2,3$ con respecto a μ se tiene que

$$\begin{cases} \text{máx}_\mu \left\{ f_1(\omega,\tau,\mu) | \frac{4\tau\mu}{(1+\tau\mu)^2} < \omega < \frac{2}{1+\tau\mu} \right\} & = f_1(\omega,\tau,\mu_{\text{mín}}) \\ \text{máx}_\mu \left\{ f_2(\omega,\tau,\mu) | \omega \geq \frac{2}{1+\tau\mu} \right\} & = f_2(\omega,\tau,\mu_{\text{máx}}) \\ \text{máx}_\mu \left\{ f_3(\omega,\tau,\mu) | \omega \leq \frac{4\tau\mu}{(1+\tau\mu)^2} \right\} & = g(\omega) \end{cases}$$

Además, sabemos que el punto de intersección de $f_1(\omega,\tau,\mu_{\text{mín}})$ con $f_2(\omega,\tau,\mu_{\text{máx}})$ es $\omega_0(\tau)$. El punto de intersección de $f_1(\omega,\tau,\mu_{\text{mín}})$ y de $g(\omega)$ es $\omega_-(\tau)$.

Entonces, por la monotonía estudiada de las funciones $f_i(\omega,\tau,\mu), i = 1,2,3$ con respecto a ω se tiene que

$$\rho(\mathcal{H}(\omega,\tau)) = \begin{cases} g(\omega) & \text{para } 0 < \omega < \omega_-(\tau) \\ f_1(\omega,\tau,\mu_{\text{mín}}) & \text{para } \omega_-(\tau) \leq \omega \leq \omega_0(\tau) \\ f_2(\omega,\tau,\mu_{\text{máx}}) & \text{para } \omega_0(\tau) < \omega < 2. \end{cases}$$

Por tanto, para cualquier τ fijo

$$\text{argmin}_\omega \rho(\mathcal{H}(\omega,\tau)) = \omega_-(\tau)$$

Como $\rho(\mathcal{H}(\omega_-(\tau),\tau)) = \sqrt{1-\omega_-(\tau)}$, el valor de τ para el cual se minimiza $\rho(\mathcal{H}(\omega_-(\tau),\tau))$ es aquel que maximiza $\omega_-(\tau)$.

En conclusión, para este caso, los parámetros óptimos son

$$\tau^{(1)}_{\text{opt}} = \frac{1}{\mu_{\text{máx}}} \quad \text{y } \omega^{(1)}_{\text{opt}} = \frac{4\mu_{\text{mín}}\mu_{\text{máx}}}{(\mu_{\text{mín}}+\mu_{\text{máx}})^2}$$

y el factor de convergencia óptimo asociado es

$$\rho(\mathcal{H}(\omega^{(1)}_{\text{opt}}, \tau^{(1)}_{\text{opt}})) = \frac{\mu_{\text{máx}} - \mu_{\text{mín}}}{\mu_{\text{máx}} + \mu_{\text{mín}}}.$$

Caso (b): $\tau \geq \frac{1}{\mu_{\text{mín}}}$.

Como $\tau\mu_{\text{mín}} \geq 1$, entonces para todo μ autovalor de $\mathcal{J}$ se tiene $\tau\mu > 1$. Recordemos que para λ , autovalor de $\mathcal{H}(\omega,\tau)$ o bien $|\lambda| = |1-\omega|$ o bien

$$|\lambda| = \begin{cases} f_2(\omega,\tau,\mu) & \text{para } \omega > \frac{4\tau\mu}{(1+\tau\mu)^2} \\ g(\omega) & \text{para } \omega \leq \frac{4\tau\mu}{(1+\tau\mu)^2} \end{cases}$$

Consideremos fijos $\omega, \tau > 0$, entonces por la monotonía estudiada para las funciones $f_i(\omega,\tau,\mu), i = 2,3$ con respecto a μ se tiene que

$$\begin{cases} \text{máx}_\mu \left\{ f_2(\omega,\tau,\mu) | \omega > \frac{4\tau\mu}{(1+\tau\mu)^2} \right\} & = f_2(\omega,\tau,\mu_{\text{máx}}) \\ \text{máx}_\mu \left\{ f_3(\omega,\tau,\mu) | \omega \leq \frac{4\tau\mu}{(1+\tau\mu)^2} \right\} & = g(\omega) \end{cases}$$

Además, sabemos que el punto de intersección de $f_2(\omega,\tau,\mu_{\text{máx}})$ y de $g(\omega)$ es $\omega_+(\tau)$.

Entonces, por la monotonía estudiada de las funciones $f_i(\omega,\tau,\mu), i = 2,3$ con respecto a ω se tiene que

$$\rho(\mathcal{H}(\omega,\tau)) = \begin{cases} g(\omega) & \text{para } 0 < \omega < \omega_+(\tau) \\ f_2(\omega,\tau,\mu_{\text{máx}}) & \text{para } \omega_+(\tau) < \omega < 2. \end{cases}$$

Por tanto, para cualquier τ fijo

$$\text{argmin}_\omega \rho(\mathcal{H}(\omega,\tau)) = \omega_+(\tau)$$

Como $\rho(\mathcal{H}(\omega_+(\tau),\tau)) = \sqrt{1-\omega_+(\tau)}$, el valor de τ tal que $\rho(\mathcal{H}(\omega_+(\tau),\tau))$ es mínimo es aquel que maximiza $\omega_+(\tau)$.

En conclusión, para este caso, los parámetros óptimos son

$$\tau_{\text{opt}}^{(2)} = \frac{1}{\mu_{\text{mín}}} \quad \text{y } \omega_{\text{opt}}^{(2)} = \frac{4\mu_{\text{mín}}\mu_{\text{máx}}}{(\mu_{\text{mín}}+\mu_{\text{máx}})^2}$$

y el factor de convergencia óptimo asociado es

$$\rho(\mathcal{H}(\omega_{\text{opt}}^{(2)}, \tau_{\text{opt}}^{(2)})) = \frac{\mu_{\text{máx}}-\mu_{\text{mín}}}{\mu_{\text{máx}}+\mu_{\text{mín}}}.$$

Caso (c): $\frac{1}{\mu_{\text{máx}}} < \tau < \frac{1}{\mu_{\text{mín}}}$.

Llamemos $\widehat{\tau} = \frac{1}{\sqrt{\mu_{\text{mín}}\mu_{\text{máx}}}}$. Para este valor ocurre

$$\omega_-(\widehat{\tau}) = \omega_+(\widehat{\tau}) = \omega_0(\widehat{\tau}) = \frac{4\sqrt{\mu_{\text{mín}}\mu_{\text{máx}}}}{(\sqrt{\mu_{\text{mín}}}+\sqrt{\mu_{\text{máx}}})^2}.$$

Distinguimos de nuevo dos casos en función del valor de τ, según sea $\tau \in \left(\frac{1}{\mu_{\text{máx}}}, \widehat{\tau}\right]$ o $\tau \in \left[\widehat{\tau}, \frac{1}{\mu_{\text{mín}}}\right)$.

Para el caso en que $\tau \in \left(\frac{1}{\mu_{\text{máx}}}, \widehat{\tau}\right]$, se tiene:

$$\omega_-(\tau) \leq \omega_+(\tau) < 1 < \omega_0(\tau).$$

Sea α un entero positivo para el que

$$\mu_\alpha = \text{máx}\left\{\mu | \mu < \frac{1}{\tau} \text{ y } \frac{1}{\mu_{\text{máx}}} < \tau \leq \widehat{\tau}\right\}.$$

Consideremos fijos ω, τ, entonces por la monotonía estudiada para las funciones $f_i(\omega, \tau\mu), i = 1,2,3$ con respecto a μ se tiene que

$$\begin{cases} \text{máx}_{\mu_{\text{mín}} \leq \mu \leq \mu_\alpha} \left\{ f_1(\omega,\tau,\mu) | \frac{4\tau\mu}{(1+\tau\mu)^2} < \omega < \frac{2}{1+\tau\mu} \right\} & = f_1(\omega,\tau,\mu_{\text{mín}}) \\ \text{máx}_{\mu_{\text{mín}} \leq \mu \leq \mu_\alpha} \left\{ f_2(\omega,\tau,\mu) | \omega \geq \frac{2}{1+\tau\mu} \right\} & = f_2(\omega,\tau,\mu_\alpha) \\ \text{máx}_{\mu_{\text{mín}} \leq \mu \leq \mu_\alpha} \left\{ f_3(\omega,\tau,\mu) | \omega \leq \frac{4\tau\mu}{(1+\tau\mu)^2} \right\} & = g(\omega) \end{cases}$$

Además, sabemos que el punto de intersección de $f_1(\omega,\tau,\mu_{\text{mín}})$ y de $f_2(\omega,\tau,\mu_\alpha)$ es $\widehat{\omega}_0(\tau) = \frac{4}{\tau(\mu_{\text{mín}}+\mu_\alpha)+2}$ y la intersección de $f_1(\omega,\tau,\mu_{\text{mín}})$ y de $g(\omega)$ es $\omega_-(\tau)$.

Por otro lado, por la monotonía de $f_i(\omega,\tau,\mu), i = 1,2,3$ con respecto a ω se tiene que

$$\text{máx}\{|\lambda| \,|\, \mu_{\text{mín}} \leq \mu \leq \mu_\alpha\} = \begin{cases} g(\omega) & \text{para } 0 < \omega < \omega_-(\tau) \\ f_1(\omega,\tau,\mu_{\text{mín}}) & \text{para } \omega_-(\tau) \leq \omega \leq \widehat{\omega}_0(\tau) \\ f_2(\omega,\tau,\mu_\alpha) & \text{para } \widehat{\omega}_0(\tau) < \omega < 2. \end{cases}$$

De forma similar, sea β un entero positivo tal que

$$\mu_\beta = \text{mín}\left\{\mu | \mu > \frac{1}{\tau} \text{ y } \frac{1}{\mu_{\text{máx}}} < \tau \leq \widehat{\tau}\right\}.$$

Consideremos fijos ω, τ, entonces por la monotonía estudiada para las funciones $f_i(\omega, \tau, \mu), i = 2, 3$ con respecto a μ se tiene que

$$\begin{cases} \text{máx}_{\mu_\beta \leq \mu \leq \mu_{\text{máx}}} \left\{ f_2(\omega, \tau, \mu) | \omega > \frac{4\tau\mu}{(1+\tau\mu)^2} \right\} & = f_2(\omega, \tau, \mu_{\text{máx}}) \\ \text{máx}_{\mu_\beta \leq \mu \leq \mu_{\text{máx}}} \left\{ f_3(\omega, \tau, \mu) | \omega \leq \frac{4\tau\mu}{(1+\tau\mu)^2} \right\} & = g(\omega) \end{cases}$$

Además, sabemos que el punto de intersección de $f_2(\omega, \tau, \mu_{\text{máx}})$ y de $g(\omega)$ es $\omega_+(\tau)$.

Por otro lado, por la monotonía de $f_i(\omega, \tau, \mu), i = 2, 3$ con respecto a ω se tiene que

$$\text{máx}\left\{|\lambda| \,|\, \mu_\beta \leq \mu \leq \mu_{\text{máx}}\right\} = \begin{cases} g(\omega) & \text{para } 0 < \omega \leq \omega_+(\tau) \\ f_2(\omega, \tau, \mu_{\text{máx}}) & \text{para } \omega_+(\tau) < \omega < 2. \end{cases}$$

Como $\omega_-(\tau) \leq \omega_+(\tau) < 1 \leq \omega_0(\tau) \leq \widehat{\omega}_0(\tau)$, por la monotonía de $f_2(\omega, \tau, \mu)$ respecto a ω, se tiene que

$$f_2(\omega, \tau, \mu_{\text{máx}}) \geq f_2(\omega, \tau, \mu_\alpha).$$

Además, como $f_1(\omega, \tau, \mu_{\text{mín}})$ y $f_2(\omega, \tau, \mu_{\text{máx}})$ intersecan en $\omega_0(\tau)$, tenemos

$$\rho(\mathcal{H}(\omega, \tau)) = \begin{cases} g(\omega) & \text{para } 0 < \omega < \omega_-(\tau) \\ f_1(\omega, \tau, \mu_{\text{mín}}) & \text{para } \omega_-(\tau) \leq \omega \leq \omega_0(\tau) \\ f_2(\omega, \tau, \mu_{\text{máx}}) & \text{para } \omega_0(\tau) < \omega < 2. \end{cases}$$

Por tanto, para cualquier $\tau \in \left(\frac{1}{\mu_{\text{máx}}}, \widehat{\tau}\right]$ fijo

$$\text{argmin}_\omega \rho(\mathcal{H}(\omega, \tau)) = \omega_-(\tau)$$

Como $\rho(\mathcal{H}(\omega_-(\tau), \tau)) = \sqrt{1 - \omega_-(\tau)}$, el valor de τ tal que $\rho(\mathcal{H}(\omega_+(\tau), \tau))$ es mínimo es aquel que maximiza $\omega_-(\tau)$.

En conclusión, para este caso, los parámetros óptimos son

$$\tau_{\text{opt}}^{(3)} = \widehat{\tau} = \frac{1}{\sqrt{\mu_{\text{mín}}\mu_{\text{máx}}}} \quad \text{y } \omega_{\text{opt}}^{(3)} = \omega_-(\widehat{\tau}) = \frac{4\sqrt{\mu_{\text{mín}}\mu_{\text{máx}}}}{(\sqrt{\mu_{\text{mín}}} + \sqrt{\mu_{\text{máx}}})^2}$$

y el factor de convergencia óptimo asociado es

$$\rho(\mathcal{H}(\omega_{\text{opt}}^{(3)}, \tau_{\text{opt}}^{(3)})) = \frac{\sqrt{\mu_{\text{máx}}} - \sqrt{\mu_{\text{mín}}}}{\sqrt{\mu_{\text{máx}}} + \sqrt{\mu_{\text{mín}}}}.$$

Para el caso en que $\tau \in \left[\widehat{\tau}, \frac{1}{\mu_{\text{mín}}}\right)$, se tiene:

$$\omega_0(\tau) \leq \omega_+(\tau) \leq \omega_-(\tau).$$

Sea α un entero positivo para el que

$$\mu_\alpha = \text{máx}\left\{\mu | \mu < \frac{1}{\tau} \text{ y } \widehat{\tau} \leq \tau < \frac{1}{\mu_{\text{mín}}}\right\}.$$

Consideremos fijos ω, τ, entonces por la monotonía estudiada para las funciones $f_i(\omega, \tau, \mu), i = 1, 2, 3$ con respecto a μ se tiene que

$$\begin{cases} \text{máx}_{\mu_{\text{mín}} \leq \mu \leq \mu_\alpha} \left\{ f_1(\omega, \tau, \mu) | \frac{4\tau\mu}{(1+\tau\mu)^2} < \omega < \frac{2}{1+\tau\mu} \right\} & = f_1(\omega, \tau, \mu_{\text{mín}}) \\ \text{máx}_{\mu_{\text{mín}} \leq \mu \leq \mu_\alpha} \left\{ f_2(\omega, \tau, \mu) | \omega \geq \frac{2}{1+\tau\mu} \right\} & = f_2(\omega, \tau, \mu_\alpha) \\ \text{máx}_{\mu_{\text{mín}} \leq \mu \leq \mu_\alpha} \left\{ f_3(\omega, \tau, \mu) | \omega \leq \frac{4\tau\mu}{(1+\tau\mu)^2} \right\} & = g(\omega) \end{cases}$$

Además, sabemos que el punto de intersección de $f_1(\omega, \tau, \mu_{\text{mín}})$ y de $f_2(\omega, \tau, \mu_\alpha)$ es $\widehat{\omega}_0(\tau) = \frac{4}{\tau(\mu_{\text{mín}}+\mu_\alpha)+2}$ y la intersección de $f_1(\omega, \tau, \mu_{\text{mín}})$ y de $g(\omega)$ es $\omega_-(\tau)$ y de las funciones $f_2(\omega, \tau, \mu_\alpha)$ y de $g(\omega)$ es $\widehat{\omega}_+(\tau) = \frac{4\tau\mu_\alpha}{(1+\tau\mu_\alpha)^2}$.

Por otro lado, por la monotonía de $f_i(\omega, \tau, \mu), i = 1, 2, 3$ con respecto a ω se tiene que

$$\text{máx}\left\{|\lambda| \,|\, \mu_{\text{mín}} \leq \mu \leq \mu_\alpha\right\} = \begin{cases} g(\omega) & \text{para } 0 < \omega < \omega_-(\tau) \\ f_1(\omega, \tau, \mu_{\text{mín}}) & \text{para } \omega_-(\tau) \leq \omega \leq \widehat{\omega}_0(\tau) \\ f_2(\omega, \tau, \mu_\alpha) & \text{para } \widehat{\omega}_0(\tau) < \omega < 2. \end{cases}$$

De forma similar, sea β un entero positivo tal que

$$\mu_\beta = \text{mín}\left\{\mu | \mu > \frac{1}{\tau} \text{ y } \widehat{\tau} \leq \tau < \frac{1}{\mu_{\text{mín}}}\right\}.$$

Consideremos fijos ω, τ, entonces por la monotonía estudiada para las funciones $f_i(\omega, \tau, \mu), i = 2, 3$ con respecto a μ se tiene que

$$\begin{cases} \text{máx}_{\mu_\beta \leq \mu \leq \mu_{\text{máx}}} \left\{ f_2(\omega, \tau, \mu) | \omega > \frac{4\tau\mu}{(1+\tau\mu)^2} \right\} & = f_2(\omega, \tau, \mu_{\text{máx}}) \\ \text{máx}_{\mu_\beta \leq \mu \leq \mu_{\text{máx}}} \left\{ f_3(\omega, \tau, \mu) | \omega \leq \frac{4\tau\mu}{(1+\tau\mu)^2} \right\} & = g(\omega) \end{cases}$$

Además, sabemos que el punto de intersección de la intersección de $f_2(\omega, \tau, \mu_{\text{máx}})$ y de $g(\omega)$ es $\omega_+(\tau)$.

Por otro lado, por la monotonía de $f_i(\omega, \tau, \mu), i = 2, 3$ con respecto a ω se tiene que

$$\text{máx}\,\{|\lambda|\,|\,\mu_\beta \le \mu \le \mu_{\text{máx}}\} = \begin{cases} g(\omega) & \text{para } 0 < \omega < \omega_+(\tau) \\ f_2(\omega,\tau,\mu_{\text{máx}}) & \text{para } \omega_+(\tau) < \omega < 2. \end{cases}$$

Como $\omega_0(\tau) \le \omega_+(\tau) \le \omega_-(\tau) < 1 \le \widehat{\omega}_0(\tau)$, por la monotonía de $f_2(\omega,\tau,\mu)$ respecto a ω, se tiene que

$$f_2(\omega,\tau,\mu_{\text{máx}}) \ge f_2(\omega,\tau,\mu_\alpha).$$

Además, como $f_1(\omega,\tau,\mu_{\text{mín}})$ y $f_2(\omega,\tau,\mu_{\text{máx}})$ intersecan en $\omega_0(\tau)$, tenemos

$$\rho(\mathcal{H}(\omega,\tau)) = \begin{cases} g(\omega) & \text{para } 0 < \omega < \omega_+(\tau) \\ f_2(\omega,\tau,\mu_{\text{máx}}) & \text{para } \omega_+(\tau) < \omega < 2. \end{cases}$$

Por tanto, para cualquier $\tau \in \left[\widehat{\tau}, \frac{1}{\mu_{\text{mín}}}\right)$ fijo

$$\text{argmin}_\omega \rho(\mathcal{H}(\omega,\tau)) = \omega_+(\tau)$$

Como $\rho(\mathcal{H}(\omega_+(\tau),\tau)) = \sqrt{1-\omega_+(\tau)}$, el valor de τ tal que $\rho(\mathcal{H}(\omega_+(\tau),\tau))$ es mínimo es aquel que maximiza $\omega_+(\tau)$.

En conclusión, para este caso, los parámetros óptimos son

$$\tau_{\text{opt}}^{(4)} = \widehat{\tau} = \frac{1}{\sqrt{\mu_{\text{mín}}\mu_{\text{máx}}}} \quad \text{y } \omega_{\text{opt}}^{(4)} = \omega_+(\widehat{\tau}) = \frac{4\mu_{\text{mín}}\mu_{\text{máx}}}{(\sqrt{\mu_{\text{mín}}} + \sqrt{\mu_{\text{máx}}})^2}$$

y el factor de convergencia óptimo asociado es

$$\rho(\mathcal{H}(\omega_{\text{opt}}^{(4)}, \tau_{\text{opt}}^{(4)})) = \frac{\sqrt{\mu_{\text{máx}}} - \sqrt{\mu_{\text{mín}}}}{\sqrt{\mu_{\text{máx}}} + \sqrt{\mu_{\text{mín}}}}.$$

De los tres casos (a), (b), (c) podemos deducir que los parámetros óptimos son

$$\tau_{\text{opt}} = \frac{1}{\sqrt{\mu_{\text{mín}}\mu_{\text{máx}}}} \quad \text{y } \omega_{\text{opt}} = \frac{4\mu_{\text{mín}}\mu_{\text{máx}}}{(\sqrt{\mu_{\text{mín}}} + \sqrt{\mu_{\text{máx}}})^2}$$

y el factor de convergencia óptimo asociado para el método GSOR es

$$\rho(\mathcal{H}(\omega_{\text{opt}}, \tau_{\text{opt}})) = \frac{\sqrt{\mu_{\text{máx}}} - \sqrt{\mu_{\text{mín}}}}{\sqrt{\mu_{\text{máx}}} + \sqrt{\mu_{\text{mín}}}}.$$

■

Capítulo 3

Métodos HSS para sistemas generales y sistemas aumentados

3.1. Introducción

Consideremos el sistema

$$\begin{pmatrix} A & B^H \\ B & -C \end{pmatrix} \begin{pmatrix} u \\ p \end{pmatrix} = \begin{pmatrix} f \\ g \end{pmatrix}.$$

donde $A \in \mathbb{C}^{n\times n}$, $B \in \mathbb{C}^{m\times n}$, $C \in \mathbb{C}^{m\times m}$, $f \in \mathbb{C}^n$, $g \in \mathbb{C}^m$ y $m \leq n$. Asumimos que A, B y C son matrices grandes y dispersas.

Un caso particular del anterior con especial interés es el correspondiente a problemas de punto de silla, para el cual la matriz A es simétrica definidda positiva, $C = 0$, B tiene rango máximo, m y $\ker(A) \cap \ker(B) = \{0\}$. En este caso el problema tiene solución única.

Podemos considerar también problemas generalizados de punto de silla. En este caso, la matriz A tiene una parte simétrica y semidefinida positiva, $H = \frac{1}{2}(A + A^H)$; el rango de B es m; $\ker(H) \cap \ker(B) = \{0\}$; C es simétrica y semidefinida positiva.

De forma equivalente al sistema que estamos considerando, podemos resolver el sistema

$$\begin{pmatrix} A & B^H \\ -B & C \end{pmatrix} \begin{pmatrix} u \\ p \end{pmatrix} = \begin{pmatrix} f \\ -g \end{pmatrix} \tag{3.1}$$

o $\mathcal{A}x = b$, donde llamamos $\mathcal{A}$ a la matriz del sistema (3.1), $x = \begin{pmatrix} u \\ p \end{pmatrix}$ y $b = \begin{pmatrix} f \\ -g \end{pmatrix}$.

El desarrollo de este último capítulo está fundamentado en los artículos [1] y [4] .

3.2. El método HSS para sistemas aumentados

Consideremos el sistema $\mathcal{A}x = b$ que proviene de un problema de punto de silla, donde

$$\mathcal{A} = \begin{pmatrix} A & B^H \\ -B & 0 \end{pmatrix}, \quad x = \begin{pmatrix} u \\ p \end{pmatrix}, \quad b = \begin{pmatrix} f \\ -g \end{pmatrix}.$$

donde $A \in \mathbb{C}^{n\times n}$ es una matriz hermítica y semidefinida positiva, $B \in \mathcal{C}^{m\times n}$ tiene rango máximo ($m \leq n$), $f \in \mathcal{C}^n$ y $g \in \mathcal{C}^m$.

Denotemos ahora, $\mathcal{H} = \frac{1}{2}(\mathcal{A}+\mathcal{A}^H)$, a la parte hermítica de $\mathcal{A}$. Entonces,

$$\mathcal{H} = \begin{pmatrix} A & 0 \\ 0 & 0 \end{pmatrix}.$$

Denotemos también, $\mathcal{S} = \frac{1}{2}(\mathcal{A} - \mathcal{A}^H)$, a la parte anti-hermítica de $\mathcal{A}$. Entonces,

$$\mathcal{S} = \begin{pmatrix} 0 & B^H \\ -B & 0 \end{pmatrix}.$$

Consideremos las dos siguientes escisiones de la matriz $\mathcal{A}$, donde $\alpha > 0$ es un parámetro e $\mathcal{I}$ denota la matriz indentidad $(n+m)\times(n+m)$:

$$\mathcal{A} = (\mathcal{H} + \alpha\mathcal{I}) - (\alpha\mathcal{I} - \mathcal{S})$$

$$\mathcal{A} = (\mathcal{S} + \alpha\mathcal{I}) - (\alpha\mathcal{I} - \mathcal{H})$$

Observemos que tanto

$$\mathcal{H} + \alpha\mathcal{I} = \begin{pmatrix} A + \alpha I_n & 0 \\ 0 & \alpha I_m \end{pmatrix}$$

como

$$\mathcal{S} + \alpha\mathcal{I} = \begin{pmatrix} \alpha I_n & B^H \\ -B & \alpha I_m \end{pmatrix}$$

son matrices regulares. La iteración hermítica/anti-hermítica (HSS) se obtiene alternando las dos escisiones anteriores en la definición de los iterantes. Para un vector inicial $x^{(0)} = (u^{(0)}, p^{(0)})$, la secuencia de vectores $\{x^{(m)}\}$ para el método HSS viene dada por:

$$\begin{cases} (\mathcal{H} + \alpha\mathcal{I})x^{(m+1/2)} = (\alpha\mathcal{I} - \mathcal{S})x^{(m)} + b, \\ (\mathcal{S} + \alpha\mathcal{I})x^{(m+1)} = (\alpha\mathcal{I} - \mathcal{H})x^{(m+1/2)} + b. \end{cases} \tag{3.2}$$

La primera ecuación se traduce en

$$\begin{pmatrix} A + \alpha I_n & 0 \\ 0 & \alpha I_m \end{pmatrix} \begin{pmatrix} u^{(m+1/2)} \\ p^{(m+1/2)} \end{pmatrix} = \begin{pmatrix} \alpha I_n & -B^H \\ B & \alpha I_m \end{pmatrix} \begin{pmatrix} u^{(m)} \\ p^{(m)} \end{pmatrix} + \begin{pmatrix} f \\ -g \end{pmatrix}$$

Entonces,

$$\begin{cases} (A+\alpha I_n)u^{(m+1/2)} = \alpha u^{(m)} - B^H p^{(m)} + f \\ p^{(m+1/2)} = p^{(m)} + \frac{1}{\alpha}(Bu^{(m)} - g) \end{cases} \tag{3.3}$$

Como la matriz $(A+\alpha I_n)$ es no singular, el sistema tiene solución única que podemos determinar. Además, la matriz es hermítica y definida positiva, por lo que para determinar la solución se puede emplear cualquiera de los métodos para resolver sistemas con matriz hermítica y definida positiva, como pueden ser la factorización de Cholesky o el método del gradiente conjugado.

Para resolver la segunda parte de (3.2), nos encontramos con el siguiente sistema

$$\begin{pmatrix} \alpha I_n & B^H \\ -B & \alpha I_m \end{pmatrix} \begin{pmatrix} u^{(m+1)} \\ p^{(m+1)} \end{pmatrix} = \begin{pmatrix} \alpha I_n - A & 0 \\ 0 & \alpha I_m \end{pmatrix} \begin{pmatrix} u^{(m+1/2)} \\ p^{(m+1/2)} \end{pmatrix} + \begin{pmatrix} f \\ -g \end{pmatrix}$$

es decir,

$$\begin{cases} \alpha u^{(m+1)} + B^H p^{(m+1)} = (\alpha I_n - A)u^{(m+1/2} + f \\ -Bu^{(m+1)} + \alpha p^{(m+1)} = \alpha p^{(m+1/2)} - g \end{cases} \tag{3.4}$$

Para resolver este sistema, denotemos

$$f^{(m)} := (\alpha I_n - A)u^{(m+1/2)} + f$$

y

$$g^{(m)} := \alpha p^{(m+1/2)} - g.$$

Entonces podemos reducir el sistema anterior eliminando $u^{(m+1)}$ obteniendo

$$(BB^H + \alpha^2 I_m)p^{(m+1)} = Bf^{(m)} + \alpha g^{(m)}$$

y después, despejando de la primera ecuación tenemos que

$$u^{(m+1)} = \frac{1}{\alpha}(f^{(m)} - B^H p^{(m+1)}).$$

Sin embargo, observemos que el método HSS se puede usar como preacondicionamiento, sin necesidad de resolver los sistemas (3.3) y (3.4) de forma exacta. Se pueden resolver de forma inexacta (IHSS) para hacer este método más competitivo.

3.3. Convergencia del método HSS

Para analizar la convergencia del método HSS (3.2) podemos eliminar el vector intermedio $x^{(m+1/2)}$ de (3.2), para ello utlizamos el siguiente lema que describe un criterio de convergencia para métodos con iteración en dos pasos.

3.1 Lema. *Sea $\mathcal{A} \in \mathbb{C}^{n\times n}$. Consideremos dos escisiones para $\mathcal{A}$, $\mathcal{A} = M_i - N_i$ para $i = 1, 2$, donde las matrices M_1 y M_2 son no singulares. Sea entonces $x^{(0)} \in \mathbb{C}^n$ un vector inicial dado. Si la secuencia de iterantes $\{x^{(m)}\}$ viene descrita por:*

$$\begin{cases} M_1 x^{(m+1/2)} = N_1 x^{(m)} + b, \\ M_2 x^{(m+1)} = N_2 x^{(m+1/2)} + b, \end{cases}$$

para $m = 0, 1, 2, \ldots$. Entonces, se puede escribir

$$x^{(m+1)} = M_2^{-1} N_2 M_1^{-1} N_1 x^{(m)} + M_2^{-1}(I + N_2 M_1^{-1})b, \quad m = 0, 1, 2, \ldots$$

Además, si el radio espectral de la matriz de iteración $M_2^{-1} N_2 M_1^{-1} N_1$ es $\rho(M_2^{-1} N_2 M_1^{-1} N_1) < 1$, entonces la secuencia de iterantes $\{x^{(m)}\}$ converge a la solución única $x^* \in \mathbb{C}^n$ del sistema de ecuaciones lineales anterior para todo vector inicial $x^{(0)} \in \mathbb{C}^n$.

Tomemos entonces en el lema anterior, $M_1 = \alpha\mathcal{I} + \mathcal{H}$, $N_1 = \alpha\mathcal{I} - \mathcal{S}$, $M_2 = \alpha\mathcal{I} + \mathcal{S}$ y $N_2 = \alpha\mathcal{I} - \mathcal{H}$.

Así, podemos escribir el método como

$$x^{(m+1)} = \mathcal{T}_\alpha x^{(m)} + c, \tag{3.5}$$

donde la matriz de iteración del método es

$$\mathcal{T}_\alpha := (\mathcal{S} + \alpha\mathcal{I})^{-1}(\alpha\mathcal{I} - \mathcal{H})(\mathcal{H} + \alpha\mathcal{I})^{-1}(\alpha\mathcal{I} - \mathcal{S}) \tag{3.6}$$

y

$$c := (\mathcal{S} + \alpha\mathcal{I})^{-1}[\mathcal{I} + (\alpha\mathcal{I} - \mathcal{H})(\mathcal{H} + \alpha\mathcal{I})^{-1}]b.$$

Por otro lado, este método HSS también puede escribirse en "forma corregida":

$$x^{(m+1)} = x^{(m)} + \mathcal{M}_\alpha^{-1} r^{(m)},$$

donde la matriz de iteración es $\mathcal{M}_\alpha^{-1}$ para

$$\mathcal{M}_\alpha := \frac{1}{2\alpha}(\mathcal{H} + \alpha\mathcal{I})(\mathcal{S} + \alpha\mathcal{I}),$$

y

$$r^{(m)} = b - \mathcal{A}x^{(m)}.$$

Observemos que $\mathcal{T}_\alpha = \mathcal{I} - \mathcal{M}_\alpha^{-1}\mathcal{A}$.

3.1 Teorema. *Sea $\mathcal{A} \in \mathbb{C}^{n\times n}$ una matriz definida positiva y α una constante positiva. Consideremos el método HSS para esta matriz, donde la matriz de iteración del método viene dada por*

$$\mathcal{T}_\alpha = (\mathcal{S} + \alpha\mathcal{I})^{-1}(\alpha\mathcal{I} - \mathcal{H})(\mathcal{H} + \alpha\mathcal{I})^{-1}(\alpha\mathcal{I} - \mathcal{S}),$$

para $\mathcal{S} = \frac{1}{2}(\mathcal{A} - \mathcal{A}^{\mathcal{H}})$ *y* $\mathcal{H} = \frac{1}{2}(\mathcal{A} + \mathcal{A}^{\mathcal{H}})$*. Entonces, para el radio espectral de esta matriz se tiene que*

$$\rho(\mathcal{T}_\alpha) \leq \sigma(\alpha) < 1, \quad \forall \alpha > 0,$$

es decir, el método HSS converge a la solución única $x^* \in \mathbb{C}^n$ *del sistema de ecuaciones lineales* $\mathcal{A}x = b$*, para*

$$\sigma(\alpha) = \max_{\lambda_i \in \lambda(\mathcal{H})} \left| \frac{\alpha - \lambda_i}{\alpha + \lambda_i} \right|$$

donde $\lambda(\mathcal{H})$ *es el espectro de la matriz* $\mathcal{H}$*.*

Demostración. Observemos que las matrices $(\alpha\mathcal{I} + \mathcal{S})^{-1}(\alpha\mathcal{I} - \mathcal{H})(\alpha\mathcal{I} + \mathcal{H})^{-1}(\alpha\mathcal{I} - \mathcal{S})$ y $(\alpha\mathcal{I} - \mathcal{H})(\alpha\mathcal{I} + \mathcal{H})^{-1}(\alpha\mathcal{I} - \mathcal{S})(\alpha\mathcal{I} + \mathcal{S})^{-1}$ son matrices semejantes. Por tanto, como el espectro de una matriz es invariante por semejanza, se tiene que

$$\begin{aligned}
\rho(\mathcal{T}_\alpha) &= \rho((\alpha\mathcal{I} + \mathcal{S})^{-1}(\alpha\mathcal{I} - \mathcal{H})(\alpha\mathcal{I} + \mathcal{H})^{-1}(\alpha\mathcal{I} - \mathcal{S})) \\
&= \rho((\alpha\mathcal{I} - \mathcal{H})(\alpha\mathcal{I} + \mathcal{H})^{-1}(\alpha\mathcal{I} - \mathcal{S})(\alpha\mathcal{I} + \mathcal{S})^{-1}) \\
&\leq \|(\alpha\mathcal{I} - \mathcal{H})(\alpha\mathcal{I} + \mathcal{H})^{-1}(\alpha\mathcal{I} - \mathcal{S})(\alpha\mathcal{I} + \mathcal{S})^{-1}\|_2 \\
&\leq \|(\alpha\mathcal{I} - \mathcal{H})(\alpha\mathcal{I} + \mathcal{H})^{-1}\|_2 \|(\alpha\mathcal{I} - \mathcal{S})(\alpha\mathcal{I} + \mathcal{S})^{-1}\|_2.
\end{aligned}$$

Veamos que $\|(\alpha\mathcal{I} - \mathcal{S})(\alpha\mathcal{I} + \mathcal{S})^{-1}\|_2 = 1$. Denotemos para ello $Q(\alpha) = (\alpha\mathcal{I} - \mathcal{S})(\alpha\mathcal{I} + \mathcal{S})^{-1}$ y como sabemos que $\mathcal{S}$ es anti-hermítica, es decir $\mathcal{S}^H = -\mathcal{S}$, tenemos que

$$\begin{aligned}
Q(\alpha)^H Q(\alpha) &= \{(\alpha\mathcal{I} - \mathcal{S})(\alpha\mathcal{I} + \mathcal{S})^{-1}\}^H (\alpha\mathcal{I} - \mathcal{S})(\alpha\mathcal{I} + \mathcal{S})^{-1} \\
&= \{(\alpha\mathcal{I} + \mathcal{S})^{-1}\}^H (\alpha\mathcal{I} - \mathcal{S})^H (\alpha\mathcal{I} - \mathcal{S})(\alpha\mathcal{I} + \mathcal{S})^{-1} \\
&= \{(\alpha\mathcal{I} + \mathcal{S})^H\}^{-1} \{(\alpha\mathcal{I})^H - (\mathcal{S})^H\}(\alpha\mathcal{I} - \mathcal{S})(\alpha\mathcal{I} + \mathcal{S})^{-1} \\
&= (\alpha\mathcal{I} - \mathcal{S})^{-1}(\alpha\mathcal{I} + \mathcal{S})^{-1}(\alpha\mathcal{I} - \mathcal{S})(\alpha\mathcal{I} + \mathcal{S})^{-1} \\
&= (\alpha\mathcal{I} - \mathcal{S})^{-1}(\alpha\mathcal{I} - \mathcal{S})^{-1}(\alpha\mathcal{I} + \mathcal{S})(\alpha\mathcal{I} + \mathcal{S})^{-1} \\
&= I
\end{aligned}$$

Por tanto, $\|Q(\alpha)\|_2 = 1$. De aquí se deduce que

$$\rho(\mathcal{T}_\alpha) \leq \|(\alpha\mathcal{I} - \mathcal{H})(\alpha\mathcal{I} + \mathcal{H})^{-1}\|_2 = \max_{\lambda_i \in \lambda\mathcal{H}} \left| \frac{\alpha - \lambda_i}{\alpha + \lambda_i} \right|.$$

Como $\mathcal{A}$ es definida positiva, también lo es $\mathcal{H}$ y, por tanto $\lambda_i > 0$ para todo $i = 1, 2, \ldots, n$. Además, $\alpha > 0$, luego se deduce que $\rho(\mathcal{T}_\alpha) \leq \sigma(\alpha) < 1$.

■

Como acabamos de ver, para el caso en que la matriz $\mathcal{A}$ es definida positiva, el método HSS es convergente para cualquier valor de α. Sin embargo, para el caso de problemas generales de punto de silla, la matriz $\mathcal{H}$ es solo semidefinida positiva y es, en general, singular. Por tanto, no es aplicable a este caso el análisis de convergencia anterior.

3.2 Teorema. *Para el problema* (3.1) *supongamos que la matriz A es una matriz real y positiva, C es una matriz simétrica y semidefinida positiva y B tiene rango máximo. Entonces, para la iteración* (3.5) *se tiene que* $\rho(\mathcal{T}_\alpha) < 1$ *para todo* $\alpha > 0$.

Demostración. La matriz de iteración del método, $\mathcal{T}_\alpha := (\mathcal{S} + \alpha\mathcal{I})^{-1}(\alpha\mathcal{I} - \mathcal{H})(\mathcal{H} + \alpha\mathcal{I})^{-1}(\alpha\mathcal{I} - \mathcal{S})$ es semejante a $\hat{\mathcal{T}}_\alpha := (\alpha\mathcal{I} - \mathcal{H})(\mathcal{H} + \alpha\mathcal{I})^{-1}(\alpha\mathcal{I} - \mathcal{S})(\mathcal{S} + \alpha\mathcal{I})^{-1} = \mathcal{R}\mathcal{U}$, para

$$\mathcal{R} = (\alpha\mathcal{I} - \mathcal{H})(\mathcal{H} + \alpha\mathcal{I})^{-1}$$

y

$$\mathcal{U} = (\alpha\mathcal{I} - \mathcal{S})(\mathcal{S} + \alpha\mathcal{I})^{-1}.$$

Observemos que la matriz $\mathcal{R}$ es simétrica por serlo $\mathcal{H}$ y la matriz $\mathcal{U}$ es ortogonal, es decir, su inversa coincide con su traspuesta. La matriz $\mathcal{R}$ es ortogonalmete semejante a la matriz diagonal $(n+m)\times(n+m)$, $\mathcal{D}$, es decir, existe una matriz ortogonal P tal que $\mathcal{R} = P^T\mathcal{D}P$, donde la matriz $\mathcal{D}$ viene dada por

$$\begin{aligned}\mathcal{D} &= \begin{pmatrix} \frac{\alpha-\mu_1}{\alpha+\mu_1} & & & & & & & & \\ & \frac{\alpha-\mu_2}{\alpha+\mu_2} & & & & & & & \\ & & \ddots & & & & & & \\ & & & \frac{\alpha-\mu_n}{\alpha+\mu_n} & & & & & \\ & & & & \frac{\alpha-\nu_1}{\alpha+\nu_1} & & & & \\ & & & & & \frac{\alpha-\nu_2}{\alpha+\nu_2} & & & \\ & & & & & & \ddots & & \\ & & & & & & & \frac{\alpha-\nu_m}{\alpha+\nu_m} & \end{pmatrix} \\ &= \begin{pmatrix} D_1 & 0 \\ 0 & D_2 \end{pmatrix}\end{aligned}$$

para μ_i , $i = 1, \dots n$ autovalores de H y ν_i , $i = 1, \dots n$ son los autovalores no negativos de C. D_1 es la matriz diagonal $n \times n$ cuyas entradas son los autovalores de H y D_2 es la matriz diagonal $m \times m$ cuyas entradas son los autovalores no negativos de C.

Observemos que $\left|\frac{\alpha-\mu_i}{\alpha+\mu_i}\right| < 1$ para $i = 1, \dots, n$ y $\left|\frac{\alpha-\nu_i}{\alpha+\nu_i}\right| < 1$ para $i = 1, \dots, m$. Como para $\mathcal{R}$ tenemos $P^T\mathcal{R}P = \mathcal{D}$, entonces $\mathcal{R}\mathcal{U}$ es ortogonalmente semejante a

$$P^T\mathcal{R}\mathcal{U}P = (P^T\mathcal{R}P)(P^T\mathcal{U}P) = \mathcal{D}\mathcal{Q}$$

para $\mathcal{Q} = P^T\mathcal{U}P$. Además, por ser producto de matrices ortogonales es una matriz ortogonal. Entonces, la matriz $\mathcal{T}_\alpha$ es semejante a $\mathcal{D}\mathcal{Q}$, luego

$$\rho(\mathcal{T}_\alpha) = \rho(\mathcal{D}\mathcal{Q}) = \rho(\mathcal{Q}\mathcal{D}).$$

Veamos ahora que $\rho(\mathcal{QD}) < 1$ para todo $\alpha > 0$. Consideremos la partición de $\mathcal{Q}$,

$$\mathcal{Q} = \begin{pmatrix} Q_{1,1} & Q_{1,2} \\ Q_{2,1} & Q_{2,2} \end{pmatrix}.$$

Entonces,

$$\mathcal{QD} = \begin{pmatrix} Q_{1,1}D_1 & Q_{1,2}D_2 \\ Q_{2,1}D_1 & Q_{2,2}D_2 \end{pmatrix}.$$

Sea $\lambda \in \mathbb{C}$ un autovalor de $\mathcal{QD}$ y sea $x \in \mathbb{C}^{n+m}$ un autovector asociado. Podemos suponer, sin pérdida de generalidad que $\|x\|_2 = 1$.

Si $\lambda = 0$, entonces es obvio que $|\lambda| < 1$, luego no hay nada que probar.

Supongamos entonces que $\lambda \neq 0$. Como λ es autovalor, tenemos $\mathcal{QD}x = \lambda x$, luego por ser $\mathcal{Q}$ ortogonal se tiene que $\mathcal{D}x = \lambda \mathcal{Q}^t x$. Luego

$$\|\mathcal{D}x\|_2 = |\lambda| \|\mathcal{Q}^T x\|_2 = |\lambda|.$$

Luego,

$$|\lambda|^2 = \|\mathcal{D}x\|_2^2 = \sum_{i=1}^{n} \left(\frac{\alpha - \mu_i}{\alpha + \mu_i}\right)^2 x_i \tilde{x}_i + \sum_{i=n+1}^{n+m} \left(\frac{\alpha - \nu_i}{\alpha + \nu_i}\right)^2 x_i \tilde{x}_i \leq \|x\|_2^2 = 1. \tag{3.7}$$

Por tanto, $|\lambda| \leq 1$. Veamos que, de hecho, esta desigualdad es estricta. Para ello vamos a probar que existe, al menos, un i $(1 \leq i \leq n)$ tal que $x_i \neq 0$. Si fuera $x_i = 0$ para todo $i = 1, \ldots, n$, es decir, si fuera $x = \begin{pmatrix} 0 \\ \hat{x} \end{pmatrix}$, entonces $\mathcal{QD}x = \lambda x$ sería

$$\mathcal{QD}x = \begin{pmatrix} Q_{1,1}D_1 & Q_{1,2}D_2 \\ Q_{2,1}D_1 & Q_{2,2}D_2 \end{pmatrix} \begin{pmatrix} 0 \\ \hat{x} \end{pmatrix} = \begin{pmatrix} Q_{1,2}D_2\hat{x} \\ Q_{2,2}D_2\hat{x} \end{pmatrix} = \begin{pmatrix} 0 \\ \lambda\hat{x} \end{pmatrix}.$$

De forma que sería $Q_{1,2}D_2\hat{x} = 0$.

Veamos ahora que $Q_{1,2}$ tiene rango máximo. Recordemos que $\mathcal{Q} = P^T\mathcal{U}P$, con $P = \begin{pmatrix} P_{1,1} & 0 \\ 0 & P_{2,2} \end{pmatrix}$. Donde $P_{1,1} \in \mathbb{R}^n$ es la matriz ortogonal que diagonaliza $(\alpha I_n - H)(\alpha I_n + H)^{-1}$ y $P_{2,2} \in \mathbb{R}^m$ es la matriz ortogonal que diagonaliza $(\alpha I_m - C)(\alpha I_m + C)^{-1}$. Recordemos además que la matriz $\mathcal{U}$ viene dada por

$$\begin{aligned} \mathcal{U} = (\alpha\mathcal{I} - \mathcal{S})(\alpha\mathcal{I} + \mathcal{S})^{-1} &= \begin{pmatrix} \alpha I_n - S & -B^T \\ B & \alpha I_m \end{pmatrix} \begin{pmatrix} \alpha I_n + S & B^T \\ -B & \alpha I_m \end{pmatrix}^{-1} \\ &= \begin{pmatrix} U_{1,1} & U_{1,2} \\ U_{2,1} & U_{2,2} \end{pmatrix}. \end{aligned}$$

Operando,

$$U_{1,2} = -\left[(\alpha I_n - S)(\alpha I_n + S)^{-1} + I_n\right] B^T \left[\alpha I_m + B(\alpha I_n + S)^{-1} B^T\right]^{-1}.$$

Observemos que $(\alpha I_n - S)(\alpha I_n + S)^{-1}$ no puede tener a -1 como autovalor, por tanto, $(\alpha I_n - S)(\alpha I_n + S)^{-1} + I_n$ es no singular. Además, como $B(\alpha I_n + S)^{-1}B^T$ es real y positiva, la matriz $\alpha I_n + B(\alpha I_n + S)^{-1}B^T$ es no singular. También,

$$\mathcal{Q} = P^T \mathcal{U} P = \begin{pmatrix} P_{1,1}^T U_{1,1} P_{1,1} & P_{1,1}^T U_{1,2} P_{2,2} \\ P_{2,2}^T U_{2,1} P_{1,1} & P_{2,2}^T U_{2,2} P_{2,2} \end{pmatrix}.$$

Luego,

$$\begin{aligned} Q_{1,2} &= P_{1,1}^T U_{1,2} P_{2,2} \\ &= -P_{1,1}^T \left[(\alpha I_n - S)(\alpha I_n + S)^{-1} + I_n \right] B^T \left[\alpha I_m + B(\alpha I_n + S)^{-1} B^T \right]^{-1} P_{2,2}. \end{aligned}$$

Esto implica que $Q_{1,2}$ tiene rango máximo puesto que tanto $P_{1,1}^T$ como $P_{2,2}$ son ortogonales y B^T tiene rango máximo.

Recordemos que teníamos que $Q_{1,2} D_2 \hat{x} = 0$ y como acabamos de probar que $Q_{1,2}$ tiene rango máximo, entonces ha de ser $D_2 \hat{x} = 0$. Pero por otro lado teníamos que $Q_{2,2} D_2 \hat{x} = \lambda \hat{x}$, luego tendríamos que $\lambda \hat{x} = 0$ y como habíamos asumido que $\lambda \neq 0$, entonces debería ser $\hat{x} = 0$, y por tanto, $x = 0$. Sin embargo, esto contradice el hecho de que $\|x\|_2 = 1$, por tanto, debe existir al menos un i $(1 \leq i \leq n)$ tal que $x_i \neq 0$. Entonces de (3.7) y como $\left| \frac{\alpha - \mu_i}{\alpha + \mu_i} \right| < 1$ para $1 \leq i \leq n$, se deduce que $|\lambda| < 1$. Por tanto, el método es convergente.

■

3.4. Optimización del parámetro para el método HSS

Usando el primer teorema de la sección anterior y conociendo el máximo autovalor y el mínimo autovalor de la matriz $\mathcal{H}$ podemos obtener el valor de α que optimiza $\sigma(\alpha)$.

3.1 Proposición. *Sea $A \in \mathbb{C}^{n \times n}$ una matriz definida positiva y sean $\gamma_{\text{mín}}$ y $\gamma_{\text{máx}}$ los autovalores mínimos y máximos, respectivamente, de la matriz $\mathcal{H} = \frac{1}{2}(\mathcal{A} + \mathcal{A}^H)$ y sea α una constante positiva. Entonces,*

$$\alpha^* = \arg\min_{\alpha} \left\{ \max_{\gamma_{\text{mín}} \leq \lambda \leq \gamma_{\text{máx}}} \left| \frac{\alpha - \lambda}{\alpha + \lambda} \right| \right\} = \sqrt{\gamma_{\text{mín}} \gamma_{\text{máx}}}$$

y

$$\sigma(\alpha^*) = \frac{\sqrt{\gamma_{\text{máx}}} - \sqrt{\gamma_{\text{mín}}}}{\sqrt{\gamma_{\text{máx}}} + \sqrt{\gamma_{\text{mín}}}} = \frac{\sqrt{\kappa(\mathcal{H})} - 1}{\sqrt{\kappa(\mathcal{H})} + 1}$$

donde $\kappa(\mathcal{H})$ es el número de condición espectral de $\mathcal{H}$.

Demostración. En este caso,

$$\sigma(\alpha) = \max_{\gamma_{\text{mín}} \leq \lambda \leq \gamma_{\text{máx}}} \left| \frac{\alpha - \lambda}{\alpha + \lambda} \right| = \max \left\{ \left| \frac{\alpha - \gamma_{\text{mín}}}{\alpha + \gamma_{\text{mín}}} \right|, \left| \frac{\alpha - \gamma_{\text{máx}}}{\alpha + \gamma_{\text{máx}}} \right| \right\}.$$

Para obtener una aproximación al factor α que optimice $\rho(\mathcal{T}_\alpha)$, podemos minimizar la cota superior que tenemos para $\rho(\mathcal{T}_\alpha)$, es decir, minimizamos $\sigma(\alpha)$. Sea α^* ese valor que minimiza $\sigma(\alpha)$. Entonces, α^* debe satisfacer, $\alpha^* - \gamma_{\text{mín}} > 0$, $\alpha^* - \gamma_{\text{máx}} < 0$ y

$$\frac{\alpha^* - \gamma_{\text{mín}}}{\alpha^* + \gamma_{\text{mín}}} = \frac{\gamma_{\text{máx}} - \alpha^*}{\gamma_{\text{máx}} + \alpha^*}.$$

Operando,

$$(\alpha^* - \gamma_{\text{mín}})(\gamma_{\text{máx}} + \alpha^*) = (\gamma_{\text{máx}} - \alpha^*)(\alpha^* + \gamma_{\text{mín}})$$

$$\alpha^*\gamma_{\text{máx}} + (\alpha^*)^2 - \gamma_{\text{mín}}\gamma_{\text{máx}} - \alpha^*\gamma_{\text{mín}} = \alpha^*\gamma_{\text{máx}} + \gamma_{\text{máx}}\gamma_{\text{mín}} - (\alpha^*)^2 - \alpha^*\gamma_{\text{mín}}$$

$$2(\alpha^*)^2 = 2\gamma_{\text{máx}}\gamma_{\text{mín}}$$

$$\alpha^* = \sqrt{\gamma_{\text{máx}}\gamma_{\text{mín}}}$$

Entonces,

$$\begin{aligned}\sigma(\alpha) &= \frac{\sqrt{\gamma_{\text{máx}}\gamma_{\text{mín}}} - \gamma_{\text{mín}}}{\sqrt{\gamma_{\text{máx}}\gamma_{\text{mín}}} + \gamma_{\text{mín}}} = \frac{\sqrt{\gamma_{\text{máx}}\gamma_{\text{mín}}} - \sqrt{\gamma_{\text{mín}}^2}}{\sqrt{\gamma_{\text{máx}}\gamma_{\text{mín}}} + \sqrt{\gamma_{\text{mín}}^2}} = \frac{\sqrt{\gamma_{\text{mín}}}}{\sqrt{\gamma_{\text{mín}}}} \frac{(\sqrt{\gamma_{\text{máx}}} - \sqrt{\gamma_{\text{mín}}})}{(\sqrt{\gamma_{\text{máx}}} + \sqrt{\gamma_{\text{mín}}})} \\ &= \frac{\sqrt{\gamma_{\text{máx}}} - \sqrt{\gamma_{\text{mín}}}}{\sqrt{\gamma_{\text{máx}}} + \sqrt{\gamma_{\text{mín}}}}.\end{aligned}$$

■

Nótese que este corolario nos da el valor óptimo α^* que minimiza la cota superior del radio espectral de la matriz de iteración, pero no minimiza el radio espectral en sí.

Bibliografía

[1] Z. Z. Bai, G. H. Golub, M. K. Ng: *Hermitian and Skew-Hermitian splitting methods for non-Hermitian positive definite linear systems*. Siam J. Matrix Appl. Vol 24, No. 3, pp. 603-626 (2003) (Cited on p. 41)

[2] Z.Z. Bai, B. N. Parlett, Z.Q. Wang: *On generalized successive overrelaxation methods for augmented linear systems* . Nummer. Math. 102; 1-38 (2005) (Cited on p. 19)

[3] M. Benzi, M. J. Gander, G. H. Golub: *Optimization of the Hermitian and Skew-Hermitian splitting iteration for saddle-point problems*. BIT Numerical Mathematics 43: 881-900 (2003) (Not cited)

[4] M. Benzi, G. H. Golub: *A preconditioner for generalized saddle point problems*. Siam J. Matrixx Anal. Appl. Vol. 26, No. 1, pp. 20-41 (2004) (Cited on p. 41)

[5] G. H. Golub, X. Wu, J.Y. Yuan:*SOR-like methods for augmented systems*. BIT 41, 71-85 (2001) (Cited on p. 19)

[6] R. S. Varga: *Matrix Iterative Analysis*. Prentice-Hall, Englewood Cliffs, N.J.(1962) (Cited on pp. 1, 6, 8, 9, 10)

[7] D. M. Young, *Iterative Solution of Large Linear Systems*, Academic Press, New York, USA, 1971. (Cited on p. 1)