Métodos Iterativos para Sistemas Lineales

Alan Reyes-Figueroa

Métodos Numéricos II (Aula 10) 08.agosto.2023

Refinamiento Iterativo

El **refinamiento iterativo** es una técnica para mejorar la precisión en la solución de un sistema lineal $A\mathbf{x} = \mathbf{b}$, hallada por un método directo.
Suponga que el sistema lineal $A\mathbf{x} = \mathbf{b}$ ha sido resuelto mediante factoración LU (con pivoteo parcial o completo), y sea $\mathbf{x}^{(0)} \in \mathbb{R}^n$ la solución calculada. Habiendo fijado una tolerancia de error, *tol*, el el refinamiento iterativo se realiza de la siguiente manera:

Algoritmo: (Refinamiento iterativo, FPR):
`While end condition doesn't hold.`
    `Compute the residual` $\mathbf{r}^{(k)} = \mathbf{b} - A\mathbf{x}^{(k)}$,
    `Solve the system` $A\mathbf{z} = \mathbf{r}^{(k)}$,
    `Update the solution` $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \mathbf{z}$,

Los términos $\mathbf{r}^{(k)}$ se denominan residuos. En ausencia de errores de redondeo, el proceso se detendría en el primer paso, dando como resultado la solución exacta.
A este procedimiento lo llamamos refinamiento iterativo de precisión simple o fija (FRP), aunque existen mejoras como el refinamiento iterativo de precisión mixta (MPR).

Métodos Iterativos

Una técnica iterativa para resolver el sistema lineal $A\mathbf{x} = \mathbf{b}$, con $A \in \mathbb{R}^{n\times n}$, $\mathbf{x}, \mathbf{b} \in \mathbb{R}^n$, comienza con una aproximación $\mathbf{x}^{(0)}$ de $\mathbf{x}$, y genera una secuencia de aproximaciones $\{\mathbf{x}^{(0)}, \mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots\} \subset \mathbb{R}^n$ tal que $\mathbf{x}^{(k)} \to \mathbf{x}$.

Estas técnicas transforman el sistema $A\mathbf{x} = \mathbf{b}$ en un sistema equivalente $\mathbf{x} = T\mathbf{x} + \mathbf{c}$, $T \in \mathbb{R}^{n\times n}$, $\mathbf{c} \in \mathbb{R}^n$. Usualmente se obtiene este sistema equivalente buscando alguna forma de “despejar” la $\mathbf{x}$ en el sistema original. Luego, se aproxima la solución mediante la secuencia

$$\mathbf{x}^{(k+1)} = T\mathbf{x}^{(k)} + \mathbf{c}, \qquad k = 0, 1, 2 \ldots \tag{1}$$

Para estudiar la convergencia de estas técnicas de iteración, necesitamos analizar la fórmula (1), con $\mathbf{x}^{(0)}$ arbitrario.
Recordemos que una matriz $T \in \mathbb{R}^{n\times n}$ es **convergente** si $T^k \to \mathbf{0}$.

Lema *Sea $T \in \mathbb{R}^{n\times n}$. Si el radio espectral $\rho(T) < 1$, entonces $(I - T)^{-1}$ existe y*

$$(I - T)^{-1} = I + T + T^2 + T^3 + \ldots = \sum_{k\geq 0} T^k.$$

Prueba: Observe que para cualquier autovalor λ de T, $1-\lambda$ es autovalor de $I-T$, pues $T\mathbf{x}=\lambda\mathbf{x} \;\Rightarrow\; (I-T)\mathbf{x}=\mathbf{x}-T\mathbf{x}=\mathbf{x}-\lambda\mathbf{x}=(1-\lambda)\mathbf{x}$.
Como $|\lambda|\leq\rho(T)<1$, entonces $|1-\lambda|>\big||1|-|\lambda|\big|=1-|\lambda|>0$. Así, ningún autovalor de $I-T$ es cero, y $I-T$ es no singular, y la ianversa $(I-T)^{-1}$ existe.
Por otro lado, $\rho(T)<1$ implica que T es convergente y $\lim T^k=\mathbf{0}$.

Sea $S_m=I+T+T^2+\ldots+T^m$. Entonces

$$(I-T)S_m=(I-T)(I+T+T^2+\ldots+T^m)=I-T^{m+1},$$

y como T es convergente, entonces

$$(I-T)\lim_{m\to\infty}S_m=\lim_{m\to\infty}(I-T)S_m=\lim_{m\to\infty}(I-T^{m+1})=\lim_{m\to\infty}I-\underbrace{\lim_{m\to\infty}T^{m+1}}_{=0}=I.$$

Portanto, $(I-T)^{-1}=\lim_{m\to\infty}S_m=\sum_{k\geq 0}T^k$. $\square$

Teorema *Para cualquier* $\mathbf{x}^{(0)} \in \mathbb{R}^n$, *la secuencia* $\{\mathbf{x}^{(k)}\}_{k\geq 0}$ *definida por*

$$\mathbf{x}^{(k+1)} = T\mathbf{x}^{(k)} + \mathbf{c}, \qquad k = 0, 1, 2, \ldots, \ y \ \mathbf{c} \neq \mathbf{0}, \tag{2}$$

converge a la solución única de $\mathbf{x} = T\mathbf{x} + \mathbf{c}$ si, y sólo si, $\rho(T) < 1$.

Prueba: De (2), para todo $k \in \mathbb{N}$ obtenemos

$$\begin{aligned}
\mathbf{x}^{(k)} &= T\mathbf{x}^{(k-1)} + \mathbf{c} \ = \ T(T\mathbf{x}^{(k-2)} + \mathbf{c}) + \mathbf{c} = T^2\mathbf{x}^{(k-2)} + (T+I)\mathbf{c} \\
&= T^2(T\mathbf{x}^{(k-3)} + \mathbf{c}) + \mathbf{c} = T^3\mathbf{x}^{(k-3)} + (T^2 + T + I)\mathbf{c} \\
&= \ldots \\
&= T^k\mathbf{x}^{(0)} + (T^{k-1} + \ldots + T^2 + T + I)\mathbf{c}.
\end{aligned}$$

($\Leftarrow$) Si $\rho(T) < 1$, entonces T es convergente y $\lim_k T^k = \mathbf{0}$. Luego,

$$\mathbf{x}^* = \lim_{k\to\infty} \mathbf{x}^{(k)} = \lim_{k\to\infty} T^k\mathbf{x}^{(0)} + \lim_{k\to\infty} \Big(\sum_{j=0}^{k-1} T^j\Big)\mathbf{c} = \big(\lim_{k\to\infty} T^k\big)\mathbf{x}^{(k)} + (I-T)^{-1}\mathbf{c} = (I-T)^{-1}\mathbf{c}.$$

Así, $(I - T)\mathbf{x}^* = \mathbf{c} \ \Rightarrow \mathbf{x}^* = T\mathbf{x}^* + \mathbf{c}$, y $\mathbf{x}^{(k)}$ converge a la única solución de (2).

($\Rightarrow$) Sea $\mathbf{x}^*$ la solución de $\mathbf{x} = T\mathbf{x} + \mathbf{c}$, y suponga que $\mathbf{x}^{(k)} \to \mathbf{x}^*$. Para todo $k \geq 0$ vale

$$\mathbf{x}^* - \mathbf{x}^{(k)} = T(\mathbf{x}^* - \mathbf{x}^{(k-1)}) = T^2(\mathbf{x}^* - \mathbf{x}^{(k.2)}) = \ldots = T^k(\mathbf{x}^* - \mathbf{x}^{(0)}). \tag{3}$$

Sea $\mathbf{z} \in \mathbb{R}^n$ arbitrario. Haciendo $\mathbf{x}^{(0)} = \mathbf{x}^* - \mathbf{z}$, resulta

$$\lim_{k\to\infty} T^k\mathbf{z} = \lim_{k\to\infty} T^k(\mathbf{x}^* - \mathbf{x}^{(0)}) = \lim_{k\to\infty} (\mathbf{x}^* - \mathbf{x}^{(k)}) = \mathbf{x}^* - \lim_{k\to\infty} \mathbf{x}^{(k)} = \mathbf{0}.$$

Esto equivale a tener $\rho(T) < 1$. $\square$

Observaciones: Si $||T|| < 1$ para alguna norma matricial, entonces la ecuación (3) nos produce fórmulas para medir el error del k-ésimo iterado $\mathbf{x}^{(k)}$ con respecto a la solución exacta $\mathbf{x}^*$:

$$||\mathbf{x}^{(k)} - \mathbf{x}^*|| \leq ||T||^k \, ||\mathbf{x}^{(0)} - \mathbf{x}^*||, \tag{4}$$

$$||\mathbf{x}^{(k)} - \mathbf{x}^*|| \leq \frac{||T||^k}{1 - ||T||} \, ||\mathbf{x}^{(1)} - \mathbf{x}^{(0)}||. \tag{5}$$

Métodos Iterativos

Una técnica general para diseñar métodos iterativos lineales se basa en una descomposición aditiva de A, en la forma $A = P - N$, con P no singular. P se llama **matriz de preacondicionamiento** o **precondicionador**.

Dado $\mathbf{x}^{(0)} \in \mathbb{R}^n$, se puede calcular la secuencia $\{x^{(k)}\}$ para $k \geq 1$, resolviendo

$$P\mathbf{x}^{(k+1)} = N\mathbf{x}^{(k)} + \mathbf{b}, \qquad k \geq 0. \tag{6}$$

La matriz de iteración en (6) es $T = P^{-1}N$, mientras que $\mathbf{c} = P^{-1}\mathbf{b}$. Alternativamente, (6) puede escribirse en la forma

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + P^{-1}\mathbf{r}^{(k)}, \qquad k \geq 0. \tag{7}$$

donde $\mathbf{r}^{(k)} = \mathbf{b} - A\mathbf{x}^{(k)}$ es el vector residual en el paso k.

- P debe ser no singular,
- Además, P debería ser fácilmente invertible, para mantener bajo el costo computacional general.
- Si $P = A$ y $N = \mathbf{0}$, (6) convergería en una iteración, pero en el mismo costo de un método directo).

Mencionamos dos resultados que aseguran la convergencia de la iteración (6).

Propiedad

Sea $A = P - N$, con A y P simétricas y positivas definidas. Si la matriz $2P - A = N + P$ es positiva definida, entonces el método iterativo definido en (6) converge para cualquier elección del dato inicial $\mathbf{x}^{(0)} \in \mathbb{R}^n$, y

$$\rho(T) = ||T||_A = ||T||_P < 1.$$

Además, la convergencia de la iteración es monótona con respecto a las normas $||\cdot||_P$ y $||\cdot||_A$ (es decir, $||\mathbf{e}^{(k+1)}||_P < ||\mathbf{e}^{(k)}||_P$ y $||\mathbf{e}^{(k+1)}||_A < ||\mathbf{e}^{(k)}||_A$, para $k = 0, 1, 2, \ldots$

Propiedad

Sea $A = P - N$, con A simétrica y positiva definida. Si la matriz $P + P^T - A$ es positiva definida, entonces P es invertible, el método iterativo en (6) es monótonamente convergente con respecto a la norma $||\cdot|_A$, para cualquier $\mathbf{x}^{(0)} \in \mathbb{R}^n$, y

$$\rho(T) \leq ||T||_A < 1.$$

Método de Jacobi

Considere el sistema lineal $A\mathbf{x} = \mathbf{b}$, con $A \in \mathbb{R}^{nn}$, $\mathbf{x}, \mathbf{b} \in \mathbb{R}^n$. El **método de** Jacobi es un método iterativo que parte de descomponer la matriz A como una suma $A = -L + D - U$,

$$A = -\underbrace{\begin{pmatrix} & & & \\ -a_{21} & & & \\ \vdots & \ddots & & \\ -a_{n1} & \cdots & -a_{n,n-1} & \end{pmatrix}}_{L} + \underbrace{\begin{pmatrix} a_{11} & & & \\ & a_{22} & & \\ & & \ddots & \\ & & & a_{nn} \end{pmatrix}}_{D} - \underbrace{\begin{pmatrix} & -a_{12} & \cdots & -a_{1n} \\ & & \ddots & \vdots \\ & & & -a_{n-1,n} \\ & & & \end{pmatrix}}_{U}.$$

Entonces, el sistema $A\mathbf{x} = \mathbf{b}$ puese escribirse como

$$(-L + D - U)\mathbf{x} = -L\mathbf{x} + D\mathbf{x} - U\mathbf{x} = \mathbf{b} \qquad \Rightarrow \qquad D\mathbf{x} = (L + U)\mathbf{x} + \mathbf{b}.$$

Si D^{-1} existe, entonces

$$\mathbf{x} = D^{-1}(L + U)\mathbf{x} + D^{-1}\mathbf{b}.$$

y hemos escrito el sistema en la forma $\mathbf{x} = T_j\mathbf{x} + \mathbf{c}_j$, donde $T_j = D^{-1}(L + U)$ y $\mathbf{c}_j = D^{-1}\mathbf{b}$.

Método de Jacobi

Algoritmo: (Método de JACOBI).
Inputs: $A \in \mathbb{R}^{n\times n}$ no-singular y diagonal dominante, $\mathbf{b} \in \mathbb{R}^n$, $\mathbf{x}^{(0)} \in \mathbb{R}^n$.
Outputs: $\mathbf{x} \in \mathbb{R}^n$, solución del sistema $A\mathbf{x} = \mathbf{b}$.
`Initialize` $\mathbf{x} = \mathbf{x}^{(0)}$.
`while end conditions doesn't hold:`
    `for` $i = 1$ `to` n:
        $x_i = (b_i - \sum_{j\neq i} a_{ij}x_j)/a_{ii}$,
    `If (`$||\mathbf{x} - \mathbf{x}^{(0)}|| \leq tol$`) or (`$iter >= MaxIter$`):  end condition = True`.

Algoritmo: (Método de JACOBI en forma vectorial).
Inputs: $A \in \mathbb{R}^{n\times n}$ no-singular y diagonal dominante, $\mathbf{b} \in \mathbb{R}^n$, $\mathbf{x}^{(0)} \in \mathbb{R}^n$.
Outputs: $\mathbf{x} \in \mathbb{R}^n$, solución del sistema $A\mathbf{x} = \mathbf{b}$.
`Initialize` $\mathbf{x} = \mathbf{x}^{(0)}$, $T = D^{-1}(L+U)$, $\mathbf{c} = D^{-1}\mathbf{b}$.
`while end conditions doesn't hold:`
    $\mathbf{x} = T\mathbf{x} + \mathbf{c}$,
    `If (`$||\mathbf{x} - \mathbf{x}^{(0)}|| \leq tol$`) or (`$iter >= MaxIter$`):  end condition = True`.

Método de Seidel

Consideramos de nuevo el sistema $A\mathbf{x} = \mathbf{b}$, con $A \in \mathbb{R}^{nn}$, $\mathbf{x}, \mathbf{b} \in \mathbb{R}^n$. El **método de Seidel** es un método iterativo que también descompone la matriz A como la suma $A = -L + D - U$,

$$A = -\underbrace{\begin{pmatrix} & & \\ -a_{21} & & \\ \vdots & \ddots & \\ -a_{n1} & \cdots & -a_{n,n-1} \end{pmatrix}}_{L} + \underbrace{\begin{pmatrix} a_{11} & & & \\ & a_{22} & & \\ & & \ddots & \\ & & & a_{nn} \end{pmatrix}}_{D} - \underbrace{\begin{pmatrix} & -a_{12} & \cdots & -a_{1n} \\ & & \ddots & \vdots \\ & & & -a_{n-1,n} \\ & & & \end{pmatrix}}_{U}.$$

Sin embargo, ahora el sistema $A\mathbf{x} = \mathbf{b}$ se reescribe como

$$(-L + D - U)\mathbf{x} = -L\mathbf{x} + D\mathbf{x} - U\mathbf{x} = \mathbf{b} \qquad \Rightarrow \qquad (D - L)\mathbf{x} = U\mathbf{x} + \mathbf{b}.$$

Si $(D - L)^{-1}$ existe, entonces

$$\mathbf{x} = (D - L)^{-1}U\mathbf{x} + (D - L)^{-1}\mathbf{b}.$$

y hemos escrito el sistema en la forma $\mathbf{x} = T_s\mathbf{x} + \mathbf{c}_s$, donde $T_s = (D - L)^{-1}U$ y $\mathbf{c}_s = (D - L)^{-1}\mathbf{b}$.

Método de Seidel

Algoritmo: (Método de SEIDEL).
Inputs: $A \in \mathbb{R}^{n\times n}$ no-singular y diagonal dominante, $\mathbf{b} \in \mathbb{R}^n$, $\mathbf{x}^{(0)} \in \mathbb{R}^n$.
Outputs: $\mathbf{x} \in \mathbb{R}^n$, solución del sistema $A\mathbf{x} = \mathbf{b}$.
`Initialize` $\mathbf{x} = \mathbf{x}^{(0)}$.
`while end conditions doesn't hold:`
 $\mathbf{z} = \mathbf{x}$,
 `for` $i = 1$ `to` n:
  $x_i = (b_i - \sum_{j=1}^{i-1} a_{ij}x_j - \sum_{j=i+1}^{n} a_{ij}z_j)/a_{ii}$,
 `If` $(||\mathbf{x} - \mathbf{x}^{(0)}|| \leq tol)$ `or` $(iter >= MaxIter)$`:  end condition = True.`

Algoritmo: (Método de SEIDEL en forma vectorial).
Inputs: $A \in \mathbb{R}^{n\times n}$ no-singular y diagonal dominante, $\mathbf{b} \in \mathbb{R}^n$, $\mathbf{x}^{(0)} \in \mathbb{R}^n$.
Outputs: $\mathbf{x} \in \mathbb{R}^n$, solución del sistema $A\mathbf{x} = \mathbf{b}$.
`Initialize` $\mathbf{x} = \mathbf{x}^{(0)}$, $T = (D-L)^{-1}U$, $\mathbf{c} = (D-L)^{-1}\mathbf{b}$.
`while end conditions doesn't hold:`
 $\mathbf{x} = T\mathbf{x} + \mathbf{c}$,

Para que funcione el método de JACOBI es necesario:

1. $a_{ii} > 0 \ \forall i$, (para que D^{-1} exista).
2. $\rho(D^{-1}(L+U)) < 1$, (condición de convergencia).

Definición

*Sean $A \in \mathbb{R}^{n\times n}$. El **radio espectral** de A es el mayor módulo de sus autovalores. Esto es*

$$\rho(A) = \max_{1\leq i\leq n} |\lambda_i|, \qquad \text{donde } \lambda_i \text{ es autovalor de } A.$$

Lema

$\rho(A) \leq ||A||$, para cualquier norma inducida $||\cdot||$.

Prueba: Sea λ autovalor de A. Entonces existe $\mathbf{x} \neq 0$, con $||\mathbf{x}|| = 1$, tal que $A\mathbf{x} = \lambda\mathbf{x}$. Luego

$$|\lambda| = ||\lambda\mathbf{x}|| = ||A\mathbf{x}|| \leq ||A||\,||\mathbf{x}|| = ||A|| \qquad \Rightarrow \qquad \rho(A) = \max|\lambda| \leq ||A||. \ \square$$

Teorema *Sea $A \in \mathbb{R}^{n\times n}$, una matriz estrictamente diagonal dominante. Para cualquier elección inicial $\mathbf{x}^{(0)} \in \mathbb{R}^n$, el método iterativo de* Jacobi*,*

$$\mathbf{x}^{(k+1)} = T_j\mathbf{x}^{(k)} + \mathbf{c}_j, \qquad \text{con } T_j = D^{-1}(L+U),\ \mathbf{c}_j = D^{-1}\mathbf{b},$$

converge a la solución del sistema $A\mathbf{x} = \mathbf{b}$.

Prueba: Basta mostrar las condiciones (1) y (2) anteriores. Como A es diagonal dominante, entonces $|a_{ii}| > \sum_{j\neq i} |a_{ij}|$. En particular $|a_{ii}| > 0$ $a_{ii} \neq 0$, para todo $1 \leq i \leq n$, de modo que D es invertible.

Por otro lado, las entradas t_{ij} de la matriz $D^{-1}(L+U)$ son de la forma $t_{ij} = \begin{cases} -\frac{a_{ij}}{a_{ii}}, & i \neq j; \\ 0, & i = j. \end{cases}$ Tomando la norma $||\cdot||_\infty$ para $T_j = D^{-1}(L+U)$, obtenemos

$$\rho(T_j) \leq ||T_j||_\infty = \max_{1\leq i\leq n} \sum_{j=1}^{n} |t_{ij}| = \max_{1\leq i\leq n} \sum_{j\neq i} \frac{|-a_{ij}|}{|a_{ii}|} = \max_{1\leq i\leq n} \underbrace{\frac{1}{|a_{ii}|} \sum_{j\neq i} |a_{ij}|}_{<1} < 1. \qquad \square$$

Teorema *Sea $A \in \mathbb{R}^{n\times n}$, una matriz estrictamente diagonal dominante. Para cualquier elección inicial $\mathbf{x}^{(0)} \in \mathbb{R}^n$, el método iterativo de* SEIDEL,

$$\mathbf{x}^{(k+1)} = T_s\mathbf{x}^{(k)} + \mathbf{c}_s, \qquad \text{con } T_s = (D-L)^{-1}U,\ \mathbf{c}_s = (D-L)^{-1}\mathbf{b},$$

converge a la solución del sistema $A\mathbf{x} = \mathbf{b}$.

Prueba: Basta mostrar las condiciones (1) y (2). $\square$

Referencias: Para este teorema (y los subsiguientes sin demostración), consultar

- Axelsson O. (1994). *Iterative Solution Methods*. Cambridge University Press.
- Young D. (1971). *Iterative Solution of Large Linear Systems*. Academic Press.
- Varga R. (1962). *Matrix Iterative Analysis*. Prentice-Hall.

Métodos de Relajación

Sea $\mathbf{r}_i^{(k)} = (r_{i1}^{(k)}, r_{i2}^{(k)}, \ldots, r_{in}^{(k)})$ el vector residual para el método de SEIDEL correspondiente a la solución aproximada $\mathbf{x}^{(k)}$

$$\mathbf{x}^{(k)} = (x_1^{(k)}, \ldots, x_{i-1}^{(k)}, x_i^{(k-1)}, \ldots, x_n^{(k-1)}).$$

El m-ésimo componente de $\mathbf{r}_i^{(k)}$ es

$$r_{mi}^{(k)} = b_m - \sum_{j=1}^{i-1} a_{mj} x_j^{(k)} - \sum_{j=i}^{n} a_{mj} x_j^{(k-1)}, \qquad m = 1, 2, \ldots, n.$$

En particular, el i-ésimo componente de $\mathbf{r}_i^{(k)}$ es

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

así que

$$a_{ii} x_i^{(k-1)} + r_{ii}^{(k)} = b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(k)} - \sum_{j=i+1}^{n} a_{ij} x_j^{(k-1)} = a_{ii} x_i^{(k)}. \tag{8}$$

o equivalentemente

$$x_i^{(k)} = x_i^{(k-1)} + \frac{r_{ii}^{(k)}}{a_{ii}}. \tag{9}$$

El método de SEIDEL está diseñado para que el residuo $r_{i,i+1}^{(k)}$ sea 0. Sin embargo, reducir a cero una coordenada del vector residuo no siempre es el camino más eficiente para reducir la norma del vector $\mathbf{r}_i^{(k)}$. Es posible modificar la ecuación (9) de modo que

$$x_i^{(k)} = x_i^{(k-1)} + \omega\,\frac{r_{ii}^{(k)}}{a_{ii}}, \tag{10}$$

para ciertos valores de ω.

Los métodos que usan (10) se llaman **métodos de relajación**. Para valores $0 < \omega < 1$ se llaman *de subrelajación*, mientras que valores $\omega > 1$ se llaman *de sobre-relajación*, y se utilizan para acelerar la convergencia de los métodos iterativos. Los métodos de **sobre relajación sucesiva** (*Succesive Over Relaxation*), se abrevian como **SOR**, ó, **OR**.

Métodos de Sobre-relajación

Llamaremos JOR a la implementación de sobre-relajación en el caso de la iteración de JACOBI. En este caso, tenemos

$$x_i^{(k+1)} = \frac{\omega}{a_{ii}}\Big(b_i - \sum_{j\neq i} a_{ij}x_j^{(k)}\Big) + (1-\omega)x_i^{(k)}, \qquad i = 1, 2, \ldots, n.$$

La matriz de iteración en este caso es $T_{j,\omega} = \omega D^{-1}(L+U) + (1-\omega)I$. Así, en forma matricial el método se escribe

$$\mathbf{x}^{(k+1)} = \big[\omega D^{-1}(L+U) + (1-\omega I)\big]\mathbf{x}^{(k)} + \omega D^{-1}\mathbf{b}.$$

Equivalentemente,

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \omega D^{-1}\mathbf{r}^{(k)}.$$

En el caso del método de SEIDEL, la iteración para el métos SOR resulta

$$x_i^{(k+1)} = \frac{\omega}{a_{ii}}\Big(b_i - \sum_{j=1}^{i-1} a_{ij}x_j^{(k+1)} + \sum_{j=1}^{i-1} a_{ij}x_j^{(k)}\Big) + (1-\omega)x_i^{(k)}, \qquad i = 1, 2, \ldots, n.$$

La matriz de iteración en este caso es $T_{s,\omega} = (I - \omega D^{-1}L)^{-1}\big[\omega D^{-1}U + (1-\omega)I\big]$. Así, en forma matricial el método se escribe

$$(I - \omega D^{-1}L)\mathbf{x}^{(k+1)} = \big[\omega D^{-1}U + (1-\omega)I\big]\mathbf{x}^{(k)} + \omega D^{-1}\mathbf{b}.$$

Equivalentemente,

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \left(\tfrac{1}{\omega}D - L\right)^{-1}\mathbf{r}^{(k)}.$$

Resultados sobre convergencia: Resumimos algunos resultados de convergencia para los métodos de JACOBI , SEIDEL, JOR y SOR.

Teorema

Si A y $2D - A$ son simétricas y positiva definidas, entonces el método de Jacobi es convergente y $\rho(T_j) = ||T_j||_A = ||T_J||_D$.

En el caso del método JOR, la suposición en $2D - A$ se puede eliminar, produciendo el siguiente resultado.

Teorema

Si A es simétrica y positiva definida, entonces el método JOR es convergente si $0 < \omega < \frac{2}{\rho(D^{-1}A)}$.

Con respecto al método de Gauss-Seidel, se cumple el siguiente resultado.

Teorema (Teorema de Kahan)

Para cualquier $\omega \in \mathbb{R}$, *se tiene que* $\rho(T_{s,\omega}) \geq |\omega - 1|$; *por lo tanto, el método SOR converge sólo si,* $0 < \omega < 2$.

<u>Prueba</u>: Si $\{\lambda_i\}$ son los autovalores de la matriz de iteración SOR, entonces

$$\left|\prod_{i=1}^{n} \lambda_i\right| = \left|\det\left[\omega D^{-1}L + (1-\omega)I\right]\right| = |1-\omega|^n.$$

Por lo tanto, debe existir al menos un autovalor λ_i tal que $|\lambda_i| \geq |1-\omega|$ y por lo tanto, para que se mantenga la convergencia, se debe tener $|1-\omega| < 1$, es decir $0 < \omega < 2$. $\square$

Suponiendo que A es simétrica y definida positiva, la condición $0 < \omega < 2$, además de ser necesario, también se vuelve suficiente para la convergencia.

Teorema (Teorema de Ostrowki-Reich)

Si A es simétrica y positiva definida, entonces el método SOR es convergente si $0 < \omega < 2$. Además, su convergencia es monótona con respecto a $||\cdot||_A$.

Finalmente, si A es estrictamente diagonale dominante, el método SOR converge si $0 < \omega \leq 1$. Los resultados anteriores muestran que el método SOR es más o menos rápidamente convergente, dependiendo de la elección del parámetro de relajación ω.

La pregunta de cómo determinar el valor óptimo ω_{opt} para el que la tasa de convergencia es la más alta posible se puede dar una respuesta satisfactoria sólo en casos especiales.

Teorema

Si la matriz A es tridiagonal T_j tiene autovalores reales, entonces el método SOR converge para cualquier elección de $\mathbf{x}^{(0)}$ si $\rho(T_j) < 1$ y $0 < \omega < 2$. Además,

$$\omega_{opt} = \frac{2}{1 + \sqrt{1 - \rho(T_j)^2}},$$

y el factor de convergencia asintótica correspondiente es

$$\rho(T_{s,\omega_{opt}}) = \frac{1 - \sqrt{1 - \rho(T_j)^2}}{1 + \sqrt{1 - \rho(T_j)^2}}.$$