Boletín de Matemáticas
Vol. XI (1977), págs. 1-22

METODOS ITERATIVOS PARA LA SOLUCION DE UN SISTEMA DE ECUACIONES LINEALES

por

Julio C. Díaz

Sumario.

En este artículo se presenta una introducción a los Métodos Iterativos para la solución de Sistemas de Ecuaciones Lineales. Se da la descripción de los métodos de Gauss-Seidel, Jacobi y S.O.R. Se prueba la convergencia de los métodos asumiendo ciertas propiedades de los sistemas. Se comparan los distintos métodos en su eficiencia y en su utilización de memoria.

1. Introducción.

Dado el sistema de ecuaciones lineales representado

$$AX = b \,, \qquad (1.1)$$

donde A es una matriz $n \times n$ y b un vector de n componentes, el problema de encontrar un vector X que satisfaga (1.1) lo atacaremos por medio de métodos iterativos.

Cuando A es no singular, los métodos descritos aquí pueden adaptarse para dar algorítmos para calcular aproximadamente A^{-1}, la inversa de A. Sin embargo, queremos hacer énfasis en que en la mayor parte de las aplicaciones el cálculo de A^{-1} es innecesario y consume tiempo exageradamente. Si se desea únicamente la solución de (1.1) los métodos expuestos aquí serán más rápidos, y por lo tanto más baratos, que primero calcular A^{-1} y luego formar $A^{-1}b$.

El método más conocido y estudiado para la solución de (1.1) es el método sistemático de eliminación de Gauss, ver [1], [2]. La eliminación gaussiana consiste en factorizar la matriz $A = LU$ en una matriz L triangular inferior y una matriz U triangular superior. Sin embargo, cuando los sistemas lineales son de un orden muy grande, las matrices son en general dispersas (cuando muy pocos de los elementos no son cero). Por ejemplo, la solución de un problema de Ecuaciones Diferenciales puede ser aproximada por medio de Ecuaciones de Diferen-

cias Finitas sobre una red de digamos 5.000 puntos. La matriz correspondiente es de orden 5.000×5.000; pero sólamente hay 25.000 de ellos diferentes a cero. Con este tipo de matriz la eliminación de Gauss tiene una desventaja impresionante pues cambia muchos de los elementos cero en elementos que no son cero. En el ejemplo de arriba, el usuario de eliminación gaussiana tiene que contentarse con reservar más de 100.000 elementos. En contraste con la eliminación gaussiana los métodos iterativos, como veremos, hacen uso solamente de la matriz original, y entonces se conservan las propiedades de la matriz a través del proceso.

La clase de métodos que nos interesa aquí se obtiene separando o dividiendo la matriz A en dos partes como sigue:

$$A = M - N \qquad (1.2)$$

donde M y N son matrices de dimensión $n \times n$.

Definimos una sucesión de vectores $X^{(m)}$, por medio de las ecuaciones

$$MX^{(m+1)} = NX^{(m)}, \quad m = 0,1,2,\ldots, \qquad (1.3)$$

donde $X^{(o)}$ debe especificarse inicialmente. Usualmente se requiere que M tenga una inversa fácilmente calculable; claro está, si la sucesión de

vectores así definida converge, el límite será una solución de (1.1).

Los ejemplos más comunes de separación son:

(i) El método de Jacobi o método de desplazamiento simultáneos,

(ii) El método de Gauss-Seidel o método de desplazamientos sucesivos,

(iii) El método de sobre relajación sucesiva ó S.O.R. (Succesive Over Relaxation),

los cuales se pueden describir como sigue:

En el caso del método de Jacobi, D denota la matriz diagonal cuya diagonal es la de A y se toma

$$M = D \,, \qquad (1.4)$$

$$N = D - A \,,$$

luego, en este caso, cada paso de la iteración en (1.3) solo necesita la solución de un sistema diagonal de ecuaciones. Obviamente una condición necesaria para que esto funcione es que $a_{ii} \neq 0$.

En el método de Gauss-Seidel, se escribe $A = D + E + F$ donde E y F son matrices triangulares estrictamente inferior y superior respectivamente y D es la diagonal mencionada antes.

Entonces se toma

$$M = D + E \ , \qquad (1.5)$$

$$N = - F \ ,$$

en este caso un sistema de matrices triangulares es todo lo que hay que resolver en cada iteración.

En el tercer método introducimos un parámetro numérico w , parámetro de relajación como sigue:

$$M = \frac{1}{w} (D + wE) \ , \qquad (1.6)$$

$$N = \frac{1}{w} ((1 - w) D - wF) \ ,$$

ya que M sigue siendo una matriz triangular inferior, el sistema que hay que resolver en cada iteración es un sistema triangular.

Observamos que para el valor $w = 1$ se obtiene el método de Gauss-Seidel.

2. Convergencia.

Antes de proseguir debemos determinar condiciones bajo las cuales las sucesiones $\mathbf{X}^{(m)}$, dadas por las separaciones de A descritas en 1., convergen. En general, de (1.3) podemos observar que

$$X^{(m+1)} = M^{-1} NX^{(m)} + M^{-1}d , \quad m \geq 0 \qquad (2.1)$$

Definamos

$$e^{(m)} = X^{(m)} - X , \quad m \geq 0 , \qquad (2.2)$$

donde X es una solución de (1.1) y $e^{(m)}$ es el vector de error asociado con $X^{(m)}$. Dado que

$$MX = NX + d ,$$

tenemos:

$$X = M^{-1} NX + M^{-1}d.$$

Si definimos

$$B = M^{-1}N ,$$

obtenemos

$$e^{(m)} = Be^{(m-1)} = \ldots = B^{(m)} e^{(o)}$$

Luego la sucesión de vectores $X^{(m)}$ satisface

$$\lim_{n\to\infty} X^{(m)} = X \ ,$$

si y sólo si ,

$$\lim_{m\to o} e^{(m)} = 0 \ ,$$

donde los límites son tomados en cada componente. Luego lo que buscamos son condiciones que aseguren que

$$\lim_{m\to\infty} B^m \, e^{(o)} = 0 \ ,$$

para todos los vectores $e^{(o)}$, lo cual es equivalente a determinar cuando

$$\lim_{m\to\infty} B^m = 0$$

donde 0 es la matriz nula.

Para esto definimos el radio espectral $\rho(A)$ de una matriz A como el máximo de los valores absolutos de los valores propios de A, o sea

$$\rho(A) = \max\{|\lambda| \ : \exists X \neq 0, \ AX = \lambda X\} = \max_{\lambda\in\sigma(A)} |\lambda|$$

Con esta definición podemos enunciar el siguiente teorema.

Teorema 1. Sea A una matriz cuadrada cualquiera, entonces $\lim_{r\to\infty} A^r = 0$, si y sólo si, $\rho(A) < 1$.

Demostración. ($\Longrightarrow$) : Asumamos que $|\lambda| \geq 1$ para todo valor propio λ de A. Entonces $A^r \neq 0$ puesto que si $AX = \lambda X$ entonces $A^rX = \lambda^rX$ ($r = 1,2,\ldots$). El cual no tiende a cero cuando $r\to\infty$.

($\Longleftarrow$) : Si A tiene valores propios distintos entonces existe una matriz no singular P tal que

$$A = P^{-1} \Lambda \, Pm \quad \text{donde} \quad \Lambda = \text{diag}\,(X_1,\ldots, \lambda_n)$$

Pero $A^r = P^{-1}\Lambda^r P \to 0$ cuando $r \to \infty$.

El caso en que A no tiene valores propios distintos, puede encontrarse en Varga [3].

3. Teoremas de Gersgorin.

Los siguientes dos teoremas de gran simplicidad, que casi nunca se incluyen en cursos de Algebra Lineal, son muy útiles para estimar el radio espectral de una matriz.

Teorema 2. (Teorema del círculo de Gersgorin). Sea $A = (a_{ij})$ una matriz compleja de $n\times n$ y sea

$$\Lambda_j = \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}| \ , \quad 1 \leqslant i \leqslant n$$

Entonces, todos los valores propios λ de A están en la unión de los discos.

$$D_i = \{z \ | |z - a_{ij}| \leqslant \Lambda_j\} \ , \quad 1 \leqslant i \leqslant n$$

Demostración. Sea $\lambda \in \sigma(A)$ y sea X un vector propio de A correspondiente a λ. Normalizamos X de forma tal que $\max_{1\leqslant i\leqslant n} |X_i| = 1$. Luego

$$(\lambda - a_{ij})\, X_i = \sum_{\substack{j=1 \\ j\neq i}}^{n} a_{ij}\, X_j \ , \quad 1 \leqslant i \leqslant n \ ,$$

en particular si $|X_r| = 1$, entonces

$$|\lambda - a_{rr}| \leqslant \sum_{\substack{j=1 \\ j\neq r}}^{n} |a_{rj}||X_j| \leqslant \sum_{\substack{j=1 \\ j\neq r}}^{n} |a_{rj}| = \Lambda_r \ ,$$

luego $\lambda \in D_r$. Ya que λ era arbitrario, se concluye que todos los valores propios de A están en la unión de los discos.

Antes de presentar el segundo teorema de Gersgorin

debemos definir el concepto de irreducibilidad. Decimos que una matriz A es reducible si existe una matriz de permutación P tal que

$$PAP^T = \begin{bmatrix} B & C \\ 0 & D \end{bmatrix},$$

donde B y C son matrices cuadradas. Si tal matriz P no existe, se dice que la matriz es irreducible. Una matriz de permutación es una matriz que consta de ceros y unos, con exactamente un elemento no nulo en cada fila o columna.

Teorema 3. (Segundo Teorema de Gersgorin). Supongamos que $A = (a_{ij})$ es una matriz irreducible y que un valor propio λ de A es un punto de frontera de la unión de los discos $\bigcup_{i=1}^{n} D_i$. Entonces,

$$|\lambda - a_{ij}| = \Lambda_i , \quad 1 \leqslant i \leqslant n$$

La demostración la puede encontrar el lector en Varga [3].

Una matriz se dice de diagonal estrictamente dominante si

$$|a_{ij}| > \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}| , \quad i = 1,2,\ldots,n \qquad (3.1)$$

Ahora podemos establecer las siguientes consecuencias directas del teorema del círculo de Gersgorin y de esta definición. Primero una proposición que nos ayudará a establecer una propiedad suficiente para que los métodos de Jacobi y Gauss-Seidel converjan.

Proposición. Si A es una matriz compleja de diagonal estrictamente dominante, entonces A es no singular y los elementos diagonales de A no son cero.

Demostración. Si los elementos de la matriz A satisfacen (3.1), entonces por el teorema del círculo de Gersgorin los valores propios son distintos de cero.

Y segundo, un teorema que establece una propiedad suficiente para que el método de Jacobi converja.

Teorema 4. Si A es de diagonal estrictamente dominante, el método de Jacobi converge.

Demostración. De las ecuaciones (1.3), (1.4) y (2.2) tenemos

$$e^{(k+1)} = - D^{-1}(E + F)\, e^{(k)} .$$

Sea $G = - D^{-1}(E + F)$, $G = (g_{ij})$. Entonces

$$g_{ii} = 0 \ , \quad g_{ij} = - \frac{a_{ij}}{a_{ii}} \ , \quad \text{para} \quad i \neq j \ .$$

Pero $\sum_{\substack{j=1 \\ j \neq i}}^{n} \frac{a_{ij}}{a_{ii}} < 1$, puesto que A es de diagonal estrictamente dominante; luego

$$\Lambda_i = \sum_{\substack{j=1 \\ j \neq j}}^{n} |g_{ij}| < 1 \ ,$$

entonces $\rho(G) < 1$ por el teorema del círculo de Gersgorin. Finalmente con la ayuda de la proposición establecemos el siguiente teorema.

Teorema 5. Si A es de diagonal estrictamente dominante, el método de Gauss-Seidel converge.

Demostración. De (1.5) vemos que en este caso debemos probar que $\rho(L_1) < 1$ donde

$$L_1 = - (D + E)^{-1} F \ ,$$

Observamos que $(D + E)^{-1}$ existe puesto que $a_{ii} \neq 0$ para cada i por la Proposición. Tomemos $X \neq 0$ un vector propio de L_1 correspondiente a un valor propio $\lambda \in \sigma(L_1)$. Esto es $L_1 X = \lambda X$. Luego,

$$[\lambda(D + E) + F] X = 0 .$$

Asumamos que $|\lambda| \geq 1$, entonces

$$(D + E + \lambda^{-1}F) X = 0 ,$$

ya que A es de diagonal estrictamente dominante y $|\lambda| \geq 1$ no es difícil mostrar que $D + E + \lambda^{-1}F$ también es de diagonal estrictamente dominante; luego, por la proposición, $D + E + \lambda^{-1}F$ es no singular. Entonces $X \equiv 0$ lo cual es una contradicción. Entonces $|\lambda| < 1$, $\lambda \in \sigma(L_1)$.

4. Comparación de los métodos de Jacobi y de Gauss Seidel.

Mirando a las ecuaciones (1.4) observamos que el método de Jacobi, componente por componente, consiste en la iteración

$$x_i^{(k+1)} = - \sum_{\substack{j=1 \\ j \neq i}}^{n} \frac{a_{ij}}{a_{ii}} x_j^{(k)} + \frac{d_j}{a_{ii}} , \qquad (4.1)$$

$$i = 1,2,\ldots,n, \quad k \geq 0.$$

Observamos que en el método de Jacobi uno debe usar todas las componentes del vector $X^{(k)}$ mien-

tras calcula las componentes del vector $X^{(k+1)}$. Intuitivamente, parecería más atractivo usar los últimos estimativos de las componentes en todas los cálculos sucesivos, esto da origen al siguiente método.

$$a_{ii}X_i^{(k+1)} = -\sum_{j=1}^{i-1} a_{ij}X_j^{(k+1)} - \sum_{j-i+1}^{n} a_{ij}X_j^{(k)} + d_i, \quad i=1,2,..n, \quad k \geq 0 \tag{4.2}$$

Pero el método descrito en (4.2) no es otro que una descripción componente por componente del método de Gauss-Seidel dado por (1.5). Luego el método de Gauss-Seidel tiene la ventaja sobre el método de Jacobi de que no requiere que simultáneamente se guarde en memoria las dos aproximaciones $X_i^{(k)}$ y $X_i^{(k+1)}$ durante el curso del cálculo.

Lo anterior nos da un criterio de comparación desde el punto de vista de economía de memoria; ahora queremos comparar la efectividad de los métodos iterativos con respecto a convergencia y para esto introducimos el concepto de rata de convergencia $R(H)$ para un método cuya matriz de iteración es H dado por

$$R(H) = -\log \rho(H)$$

La utilidad de la rata de convergencia proviene de que se puede demostrar que el número de iteraciones necesarias para reducir un error inicial por un factor prescrito es inversamente proporcional a la rata de convergencia; ver [3]. Claramente a menor radio espectral mayor rata de convergencia. Luego, si H depende de un parámetro w, éste se escogerá de tal forma que $\rho(A(w))$ se minimice.

En los casos de importancia práctica, se puede mostrar que el método de Gauss-Seidel converge más rápidamente que el método de Jacobi. Por ejemplo, cuando una matriz es de diagonal estrictamente dominante. También se puede demostrar que en ciertos casos esos métodos son ambos convergentes o ambos divergentes.

5. S.O.R.

El tercero de los métodos descritos en 1 está relacionado al método de Gauss-Seidel. Si definimos $\tilde{x}^{(k)}$ promedio de

$$a_{ii}\tilde{x}_i^{(k+1)} = -\sum_{j=1}^{i-1} a_{ij}x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij}x_j^{(k)} + d_i, \tag{5.1}$$

$$i = 1,2,\ldots,n \ , \quad k \geq 0,$$

y luego aplicamos un factor de relajación w en

la siguiente forma:

$$X_i^{(k+1)} = X_i^{(k)} + w\{\tilde{X}_i^{(k+1)} - X_i^{(k)}\} = \tag{5.2}$$

$$= (1 - w)X_i^{(k)} + w\tilde{X}_i^{(k+1)}$$

combinando (5.1) y (5.2) en una sola ecuación, obtenemos:

$$(D + wF)X^{(k+1)} = \{(1 - w)D - wF\}\, X^{(k)} + wd, \tag{5.3}$$

$$k \geq 0,$$

que es el método de sobre relajación sucesiva (S.O.R) el cual requiere guardar en memoria un solo vector de aproximación en el curso del Cálculo.

La ecuación (5.3) se puede reescribir como

$$X^{(k+1)} = (I - wL)^{-1}\{(1 - w)\, I + wU\}X^{(k)} +$$

$$+ w(I - uL)^{-1}\, D^{-1}d\ ,$$

donde $L = D^{-1}E$ y $U = -D^{-1}F$. La matriz de iteración del método S.O.R L_w está dada por

$$L_w = (1 - wL)^{-1}\{(1 - w)I + wU\}. \tag{5.4}$$

Notemos que $w = 1$ da el método de Gauss-Seidel, como era de esperarse a partir de (5.2).

El factor de relajación w se debe escoger de tal forma que acelere la convergencia, luego se ha de escoger en tal forma que minimice $\rho(L_w)$ como función de w. El siguiente teorema nos da una idea de dónde buscar ese mínimo.

Teorema 6. (Kahan). Si L_w está definida como en (5,4), entonces

$$\rho(L_w) \geq |w - 1|$$

con igualdad solamente si todos los valores propios de L_w tienen módulo $|w - 1|$.

Demostración. Sea $\phi(\lambda) = \det(\lambda I - L_w)$, el polinomio característico de L_w; ya que L es estrictamente triangular inferior, $(I - wL)^{-1}$ existe y $\det(I - wL) = 1$. Luego

$$\phi(\lambda) = \det(I - wL)\det(\lambda I - L_w) =$$

$$= \det\left[(I - wL)(I - L_w)\right]$$

$$= \det\left[(I - wL)(I - (I - wL)^{-1}\{(1 - w)I + wU\})\right]$$

$$= \det\left[\lambda(I - wL) - (I - w)I - wU\right]$$

$$= \det\left[\lambda(I - wL) - (I - w)\, I - wU\right]$$

$$= \det\left[(w + \lambda - I)\, I - w(\lambda L + U)\right] ,$$

Pero

$$\phi(\lambda) = \prod_{i=1}^{n} (\lambda - \lambda_i) ,$$

luego

$$\phi(0) = (-1)^n \prod_{i=1}^{n} \lambda_i = \det((w - 1)\, I - wU) = (w - 1)^n$$

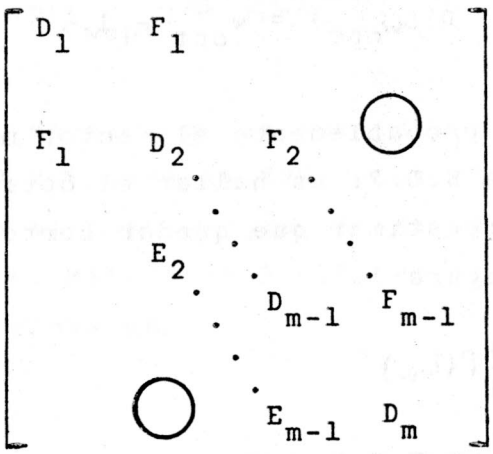
de donde

$$|\rho(L_w)|^n = \max_{\lambda_i \in \sigma(L_w)} |\lambda_i|^n \geq |w - 1|^n ,$$

con igualdad solamente si todos los valores propios de L_w tienen módulo $|w - 1|$. Luego debemos buscar el valor óptimo en el intervalo $(0,2)$.

6. Determinación de parámetro óptimo w_{opt}.

Dada una matriz A de orden n, y un entero $m \geq 2$, decimos que A es tridiagonal de m bloqués si se puede escribir como



donde los D_i's son matrices cuadradas. Si las matrices D_i son diagonales entonces decimos que A es **diagonalmente tridiagonal de m bloques**. Por último se dice que si existe una matriz de permutación P tal que PAP^T es diagonalmente tridiagonal de m bloques, entonces A tiene la propiedad (A).

Ahora podemos enunciar un teorema que nos da el valor óptimo de w en un caso particular.

Teorema 7. Si A tiene propiedad (A), $A = D + E + F$, $G = L + U$ la matriz del método de Jacobi y L_w como antes, entonces

$$w_{opt} = \frac{2}{1 + \sqrt{1 - \rho(G)}} \qquad (7.1)$$

$$\rho(L_{w_{opt}}) = w_{opt} - 1 \ . \qquad (7.2)$$

En práctica, probablemente el factor más importante del uso de S.O.R. es hallar el óptimo parámetro. Es mejor sobreestimar que quedar corto como lo ilustra la figura 1.

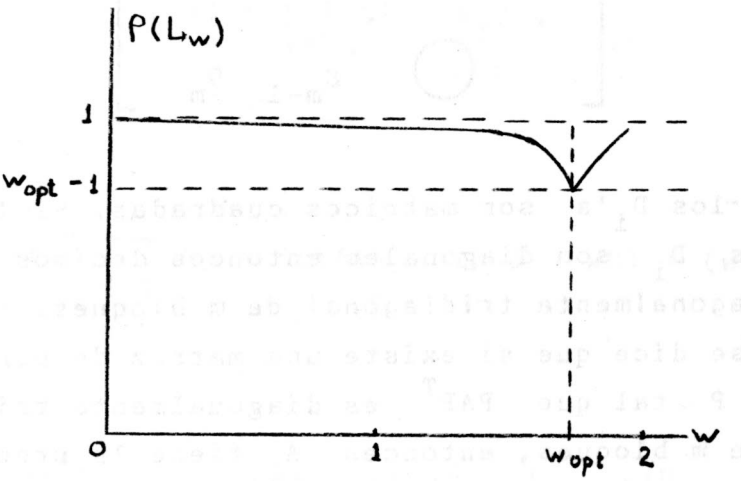


Figura 1

7. Comparación de los métodos.

Si una matriz cuadrada A satisface la propiedad (A) entonces

$$R(L_1) = 2R(G),$$

y

$$R(L_{wopt}) \approx 2\sqrt{R(L_1)} ,$$

por ejemplo, en el caso de que $R(L_1) = 10^{-4}$, $R(L_{wopt}) \sim 0.2$, esto es 200 veces más rápido. Por lo tanto, en el caso de que A satisfaga la propiedad (A), el método de S.O.R. es el más rápido de todos en converger.

Referencias.

[1]. Forsythe, G., Moler C. Computer Solucion of Linear Algebraic Systems. Prentice Hall, Englewood Cliffs, N.J. (1967).

[2]. Stewart, G.W. Introduction to Matrix Computation. Academic Press, Inc. New York, N.Y. (1973).

[3]. Varga, S.R. Matrix Iterative Analysis. Prentice Hall, Englewood Cliffs, N.Y. (1962).

Esta presentación ha sido sustancialmente influenciada en forma y contenido por "Iterative Methods for the Solution of Systems of Linear Equations". Notas de clase escritas por el Dr. G. Gairweather para los estudiantes de la Universidad de Kentucky.

Department of Mathematics
University of Kentucky
Lexington, Ky 40506