Análisis Numérico Matricial
Universidad de Murcia
curso 2016-2017

Unidad 3: Métodos Iterativos para Resolver Ecuaciones Lineales

Antonio José Pallarés Ruiz

Índice general

Capítulo 3

Métodos iterativos de resolución de sistemas de ecuaciones

Interrogantes centrales del capítulo

- Análizar técnicas iterativas de resolución de sistemas de ecuaciones lineales.
- Aprender los métodos de resolución :
  - Método de Jacobi.
  - Método de Gauss-Seidel.
  - Método de relajación.
  - Método del gradiente conjugado

Destrezas a adquirir en el capítulo

- Resolver sistemas de ecuaciones lineales utilizando los métodos iterativos.
- Implementar los métodos en el ordenador.
- Compararlos entre si y con los métodos directos del capítulo anterior.

En esta unidad se estudian distintos métodos iterativos de resolución de sistemas de ecuaciones lineales. Los métodos iterativos no suelen utilizarse para resolver problemas lineales de dimensión pequeña ya que, para obtener una precisión razonable, requieren más operaciones que los métodos directos. Sin embargo, en el caso de sistemas grandes con muchos ceros en sus coeficientes (matrices banda o estrictamente diagonal dominantes que aparecen en problemas de ecuaciones diferenciales con condiciones frontera) hay métodos iterativos muy eficientes.

Proponemos tres métodos iterativos concretos a partir de una misma idea general de iteraciones de punto fijo, y completamos la unidad con el método del gradiente conjugado que nació como un método directo y que se ha convertido en el método más popular utilizado como método da aproximación iterativa a las soluciones de sistemas grandes de ecuaciones con matrices de coeficientes simétricas, definidas positivas y con muchos ceros (“sparse”).

Desarrollo de los contenidos fundamentales

- Métodos iterativos, convergencia.
- Método de Jacobi.
- Método de Gauss-Seidel.
- Método de Relajación.
- Método del gradiente conjugado.

3.1. Métodos iterativos. Criterios de Convergencia

Idea general:

Un método iterativo para resolver un sistema lineal

$$Ax = b$$

consiste en transformar el sistema de ecuaciones en una ecuación de punto fijo

$$x = Tx + c;$$

donde T es una aplicación lineal. Si el radio espectral $\rho(T) < 1$, la aplicación $Tx+c$ es contractiva y la solución del sistema $\vec{x}$ (el punto fijo) se obtiene como el límite de una sucesión de iteradas funcionales $\vec{x}_k = T\vec{x}_{k-1} + c$, comenzando en una aproximación inicial $\vec{x}_0$ a la solución.

Ejemplo 3.1.1 Consideremos el sistema $Ax = b$ dado por las ecuaciones:

$$\begin{cases} 2x_1 - 2x_2 & = 1 \\ 2x_1 + 3x_2 + x_3 & = 5 \\ -x_1 - 2x_3 & = 7 \end{cases}$$

que tiene como única solución $x = (\frac{20}{9}, \frac{31}{18}, -\frac{83}{18}) \approx (2.22222, 1.72222, -4.61111)$.

Despejando x_i en la ecuación i se tiene la ecuación equivalente

$$\begin{cases} x_1 = x_2 + \frac{1}{2} \\ x_2 = -\frac{2}{3}x_1 - \frac{1}{3}x_3 + \frac{5}{3} \\ x_3 = -\frac{1}{2}x_1 - \frac{7}{2} \end{cases}$$

Ésta es una ecuación de punto fijo $x = Tx + c$, con

$$T = \begin{pmatrix} 0 & 1 & 0 \\ -\frac{2}{3} & 0 & -\frac{1}{3} \\ -\frac{1}{2} & 0 & 0 \end{pmatrix}$$

$\rho(T) = 0.84657...$, El límite de la sucesión de iteradas $\vec{x}_k = T\vec{x}_{k-1}+c$ obtenido en 195 iteraciones comenzando en $\vec{x}_0 = (1, 1, 1)^t$, con un error relativo en la imagen menor que 10^{-14}, es la solución del sistema lineal $(2.22222, 1.72222, -4.61111)$

Definición 3.1.2 *Dado un sistema lineal $Ax = b$, vamos a llamar "método iterativo de resolución del sistema" a cualquier par (T, c) formado por una matriz T y un vector c tales que la solución de $Ax = b$ es el único punto fijo de la transformación afín $\Phi(x) = Tx + c$, es decir*

$$Ax = b \quad \Leftrightarrow \quad x = Tx + c.$$

Se dice que el método iterativo es convergente cuando la sucesión de iteradas $x_k = \Phi(x_{k-1}) = Tx_{k-1} + c$ converge hacia el punto fijo para cualquier elección del vector x_0.

A la hora de implementar los métodos iterativos interesa tener presente que podemos utilizar como condición de parada el tamaño de los "vectores residuales" en cada etapa:

$$r_k = Ax_k - b.$$

3.1.1. Criterios de Convergencia

En el siguiente teorema recogemos los resultados estudiados en el capítulo 1 que dan condiciones necesarias y suficientes para que un método iterativo sea convergente:

Teorema 3.1.3 *Sea T una matriz cuadrada de dimensión n. Entonces son equivalentes:*

(I) Existe una norma matricial (subordinada) tal que $\|T\| < 1$.

(II) El radio espectral $\rho(T) < 1$.

(III) $\lim_{k\to\infty} T^k v = 0$, para todo vector v.

(IV) Las sucesiones de iteradas $\vec{x}_k = T\vec{x}_{k-1} + c$ convergen comenzando en cualquier vector $\vec{x}_0$.

DEMOSTRACIÓN:

La equivalencia entre (I) y (II) es el Teorema 1.3.2.

La equivalencia entre (II) y (III) es el Teorema 1.3.4.

La implicación (I) $\Rightarrow$ (IV), la da el Teorema del punto fijo de Banach porque si $\Phi(x) = Tx+c$,

$$\|\Phi(x) - \Phi(y)\| = \|Tx - Ty\| = \|T(x - y)\| \leq \|T\|\|x - y\|,$$

y si $\|T\| < 1$, Φ será contractiva.

Para la implicación (IV) $\Rightarrow$ (III), consideremos v un vector arbitrario, y x el vector solución de $x = Tx + c$. Si tomamos $\vec{x}_0 = x - v$, la sucesión $\vec{x}_k = T\vec{x}_{k-1} + c$ converge hacia x y por lo tanto:

$$x - \vec{x}_k = Tx + c - (T\vec{x}_{k-1} + c) = T(x - \vec{x}_{k-1}) = T^2(x - \vec{x}_{k-2})... = T^k(x - \vec{x}_0) = T^k(v) \to 0.$$

□

Ejemplo 3.1.4 Volviendo al ejemplo 3.1.1, el polinomio característico de la matriz T es

$$p_T(\lambda) = 6\lambda^3 + 4\lambda + 1.$$

El radio espectral es $\rho(T) = 0.84657...$[1] por lo tanto la iteración $\vec{x}_k = T\vec{x}_{k-1} + c$, comenzando en cualquier vector $\vec{x_0}$, converge hacia la solución del sistema de punto fijo $x = Tx + c$; que también es la solución de la ecuación $Ax = b$.

[1] Aunque no es fácil calcular raíces de polinomios de tercer grado, si es fácil comprobar que los ceros de este polinomio característico (valores propios de T) tienen módulo menor o igual que $\sqrt[3]{\frac{5}{6}} < 1$

Ejemplo 3.1.5 (Método iterativo de Richardson.Método del gradiente) Para resolver el sistema $Ax = b$ se considera el método iterativo

$$x_{k+1} = x_k + \alpha(b - Ax_k)$$

que corresponde a la ecuación de punto fijo $x = T_\alpha x + c$ con $T_\alpha = Id - \alpha A$ y $c = \alpha b$.

Los valores propios de T_α son de la forma $(1 - \alpha\lambda_i)$, con $\lambda_i \in \sigma(A)$. El método de Richardson es convergente si, y sólo si, $|1 - \alpha\lambda_i| < 1$ para cada λ_i, e.d. $-1 < 1 - \alpha\lambda_i < 1$ que si el espectro de A está en $\mathbb{R}$, equivale a $0 < \alpha\lambda_i < 2$. Cuando todos los valores propios de A son positivos $0 < \lambda_i \leq \rho(A)$, esta última condición equivale a que $\alpha \in (0, 2/\rho(A))$. Obsérvese que si A tiene valores propios reales de distinto signo el método de Richardon no converge para ningún valor de $\alpha \in \mathbb{R}$ porque en ese caso $\rho(T_\alpha) \geq 1 - \alpha\lambda > 1$ (para algún valor propio real λ).

La denominación de método del gradiente se debe a que $Ax - b$ es la dirección del gradiente de la función $g(x) = x^*Ax - 2x^*b$ que si A es una matriz simétrica definida positiva tiene un mínimo absoluto en la solución de la ecuación $Ax = b$ (véanse el Teorema 3.5.1 y la Proposición 3.5.2).

Ejercicio 3.1 Comprobad que en el ejemplo 3.1.5, si todos los valores propios de A son reales y positivos (p.e. si A es simétrica definida positiva), $0 < \lambda_1 \leq \lambda_2 \leq ... \leq \lambda_n = \rho(A)$, el valor de α que minimiza el radio espectral de T_α es

$$\alpha_{opt} = \frac{2}{\lambda_1 + \lambda_n} \text{ y } \rho(T_{\alpha_{opt}}) = \frac{\lambda_n - \lambda_1}{\lambda_n + \lambda_1} < 1.$$

3.1.2. Construcción de Métodos iterativos

Idea general:

Supongamos que tenemos un sistema lineal

$$Ax = b$$

y que la matriz A se puede expresar como diferencia de dos matrices

$$A = M - N,$$

donde M es una matriz "fácil" de invertir (por ejemplo si es diagonal o triangular). Entonces:

$$Ax = b \longleftrightarrow Mx - Nx = b \longleftrightarrow Mx = Nx + b,$$

$$Ax = b \longleftrightarrow x = M^{-1}Nx + M^{-1}b$$

Con la notación de la Definición 3.1.2 el método iterativo correspondiente viene dado por la matriz $T = M^{-1}N$ y el vector $c = M^{-1}b$.

Ejemplo 3.1.6 (Método iterativo de Richardson) En el ejemplo 3.1.5, la matriz $M = \alpha^{-1}Id$ que es fácil de invertir y $N = \alpha^{-1}Id - A$.

Dada una matriz cuadrada de dimensión n, $A = (a_{ij})$, tal que $a_{ii} \neq 0$ para todo $1 \leq i \leq n$. Entonces la matriz diagonal $D = diagonal(a_{ii})$ es muy fácil de invertir. Escribiendo $M = D$ y $N = D - A$ se obtiene el método iterativo de Jacobi.

Los métodos de Gauss-Seidel y de relajación son variaciones del método de Jacobi. Para descibirlos vamos a utilizar la siguiente notación para describir $A = L + D + U$ con

$$L = \begin{pmatrix} 0 & 0 & & & 0 \\ a_{21} & 0 & & & 0 \\ . & . & . & & . \\ . & . & & . & . \\ . & . & & 0 & . \\ a_{n1} & . & . & a_{n(n-1)} & 0 \end{pmatrix} \quad U = \begin{pmatrix} 0 & a_{12} & . & . & a_{1n} \\ & 0 & a_{23} & . & a_{2n} \\ & & . & & . \\ & & & . & . \\ & & & 0 & a_{(n-1)n} \\ & & & & 0 \end{pmatrix}.$$

3.2. Método de Jacobi

Tal y como hemos mencionado, el método de Jacobi para buscar la solución de un sistema lineal $Ax = b$ si no hay ceros en la diagonal D de A consiste en construir la sucesión de iteradas

$$x_{k+1} = D^{-1}(-(L+U))x_k + D^{-1}b = T_J x_k + D^{-1}b,$$

donde la matriz del método de jacobi es $T_J = D^{-1}(-(L+U))$ y el vector $c = D^{-1}b$.

Para realizar los cálculos de forma eficiente y para utilizarlos en las condiciones de parada comenzamos analizando los vectores residuales y observando cómo pueden utilizarse para construir cada iteración del método:

- $r_k = Ax_k - b = Dx_k + (L+U)x_k - b$
- $D^{-1}r_k = x_k - (D^{-1}(-(L+U))x_k + D^{-1}b) = x_k - x_{k+1}$
- $x_{k+1} = x_k - D^{-1}r_k$

Cada etapa del cálculo de los vectores $r_k = (r_i^k)_i$ y $x_k = (x_i^k)_i$ se realiza coordenada a coordenada, comezando en $i = 1$ y consecutivamente $i = 2, 3, ..., n$

`Para` $i = 1$`, mientras que` $i \leq n$`, haciendo en cada paso` $i = i + 1$`:`

$$r_i^k = a_{ii}x_i^k + \sum_{j=1;j\neq i} a_{ij}x_j^k - b_i \tag{3.1}$$
$$x_i^{k+1} = x_i^k - \frac{1}{a_{ii}}r_i^k$$

En la página siguiente el algoritmo 3.1 implementa el método de Jacobi. Prestad atención a la construcción del vector residual y como se utiliza el cuadrado de su norma euclídea como condición de parada.

Ejemplo 3.2.1 En el ejemplo 3.1.1 se ha considerado la iteración de Jacobi que, como hemos señalado, converge hacia la solución del sistema lineal $Ax = b$.

Ejercicio 3.2 Comprueba que el espectro de $T_J = D^{-1}(-(L+U))$, es el conjunto de ceros del polinomio $q_J(\lambda) = \det(L + U + \lambda D)$.

Algoritmo 3.1 **Método de Jacobi para resolución de sistemas lineales**

Datos de entrada: $A[n][n]$ (Matriz de coeficientes del sistema.);
$b[n]$ (vector término independiente.);
n (dimensión de A y b);
ε (precisión para la condición de parada);
$nmax$ (número máximo de iteraciones);
Variables:$xa[n]$;// (x_k) vector para aproximar la solución del sistema.
$xb[n]$;// (x_{k+1}) vector para las nuevas aproximaciones de la solución del sistema.
$e[n]$;// un vector auxiliar para almacenar el vector residual y el vector de corrección $x_k - x_{k+1}$.
$eadmisible = 0$;// precisión admisible (se usan errores relativos.
$norma = 0$; // registro para el cuadrado de la norma del vector residual.
Fujo del programa:
// Condiciones iniciales y evaluacion de la diagonal
`for(j=1;j<=n;j++){`
$xa(j) = 1$; $eadmisible = eadmisible + b(j)^2$;
`if(`$A_{j,j} == 0$`){`
ERROR; Jacobi no es aplicable;
`}`
`}`
$eadmisible = \varepsilon^2 * eadmisible$ // $(\varepsilon * \|b\|)^2$.
// Vamos a hacer las etapas $k = 1, 2, ..., nmax$.
`for(k=1;k<=nmax;k++){`
$norma = 0$; // 1. cálculo de la corrección.
`for(i=1;i<=n;i++){`
$e(i) = -b(i)$;
`for(j=1;j<=n;j++){`
$e(i) = e(i) + A_{i,j} * xa(j)$;
`}`
$norma = norma + e(i)^2$;
$e(i) = e(i)/A_{i,i}$;
$xb(i) = xa(i) - e(i)$;
`}`
`if(`$norma < eadmisible$`){`
Parada, la solucion es xa
`}`
$xa = xb$;
`}`
Parada, no hay convergencia en nmax iteraciones;
Datos de salida: Solución x del sistema o mensajes de error si la diagonal de A tiene algún cero o la iteración no converge.

Ejercicio 3.3 Considera el sistema de ecuaciones

$$\begin{cases} 10x_1 - x_2 + 2x_3 = 6 \\ -x_1 + 11x_2 - x_3 + 3x_4 = 25 \\ 2x_1 - x_2 + 10x_3 - x_4 = -11 \\ 3x_2 - x_3 + 8x_4 = 15 \end{cases}$$

que tiene matriz de coeficientes con diagonal estrictamente dominante y solución en $x = (1, 2, -1, 1)^t$.

(I) Demuestra que el método de Jacobi para este sistema es convergente.

(II) Evalúa las primeras iteraciones del método comenzando en el origen $x^0 = (0, 0, 0, 0)^t$.

3.3. Método de Gauss-Seidel

Los vectores del algoritmo de Jacobi se van construyendo coordenada a coordenada de manera que cuando se va a calcular x_i^{k+1} ya se conocen los valores de x_j^{k+1} para $j < i$. La idea en la modificación propuesta en el método de Gauss-Seidel consiste en utilizar para el cálculo de x_i^{k+1} en 3.1 las coordenadas conocidas x_j^{k+1} para $j < i$ junto con x_j^k para $i \leq j$, en lugar de utilizar sólo las coordenadas de x_k, en la hipótesis de que si el método va a converger, las coordenadas de x_{k+1} son una mejor aproximación a las de la solución que las de x_k.

`Para` $i = 1$`, mientras que` $i \leq n$`, haciendo en cada paso` $i = i + 1$`:`

$$\tilde{r}_i^k = a_{ii}x_i^k + \sum_{j=1}^{i-1} a_{ij}x_j^{k+1} + \sum_{j=i+1}^{n} a_{ij}x_j^k - b_i \tag{3.2}$$

$$x_i^{k+1} = x_i^k - \frac{1}{a_{ii}}\tilde{r}_i^k$$

En términos matriciales, el método de Gauss-Seidel consiste en considerar la descomposición $A = (L + D) + U = M - N$, $M = L + D$ y $N = -U$. Para comprobarlo basta con observar las coordenadas de la expresión: $(L + D)x_{k+1} = -Ux_k + b$. La matriz del método iterativo de Gauss-Seidel es $T_G = (L + D)^{-1}(-U)$.

En la página siguiente está el algoritmo 3.2 correspondiente a el método de Gauss-Seidel. Observad que para la condición de parada hemos utilizado como vector residual el vector $\tilde{r}_k$, donde cada una de las coordenadas $\tilde{r}_i^k$ se corresponde con la coordenada i de los vectores $A\tilde{x}_{ik} - b$ con $\tilde{x}_{ik} = (x_1^{k+1}, ..., x_{i-1}^{k+1}, x_i^k, ...x_n^k)^t$ $(i = 1, ..., n)$. Si la sucesión x_k converge hacia la solución del sistema lineal, también lo hace $\tilde{x}_{ik}$ y $\tilde{r}_k \to 0$.

Ejercicio 3.4 Comprueba que el espectro de $T_G = (L + D)^{-1}(-U)$, es el conjunto de ceros del polinomio $q_G(\lambda) = \det(U + \lambda L + \lambda D)$.

Ejercicio 3.5 Continuando con el sistema de ecuaciones del ejercicio 3.3

(I) Demuestra que el método de Gauss-Seidel para este sistema es convergente.

(II) Evalúa las primeras iteraciones del método comenzando en el origen $x^0 = (0, 0, 0, 0)^t$.

Algoritmo 3.2 **Método de Gauss-Seidel para resolución de sistemas lineales**

Datos de entrada: $A[n][n]$ (Matriz de coeficientes del sistema.);
$b[n]$ (vector término independiente.);
n (dimensión de A y b);
ε (precisión para la condición de parada);
$nmax$ (número máximo de iteraciones);
Variables:$xa[n]$;// vector para aproximar la solución del sistema.
$xb[n]$;// vector para las nuevas aproximaciones de la solución del sistema.
$\tilde{e}[n]$;// un vector residual modificado.
$eadmisible = 0$;// precisión admisible.
$norma = 0$; // registro para el cuadrado de la norma del vector residual modificado.
Fujo del programa:
// Condiciones iniciales y evaluacion de la diagonal
`for(j=1;j<=n;j++){`
$xa(j) = 1$; $eadmisible = eadmisible + b(j)^2$;
`if(`$A_{j,j} == 0$`){` ERROR; Gauss-Seidel no es aplicable;`}`
`}`
$eadmisible = \varepsilon^2 * eadmisible$ // $(\varepsilon * \|b\|)^2$.
// Vamos a hacer las etapas $k = 1, 2, ..., nmax$.
`for(k=1;k<=nmax;k++){`
$norma = 0$; // 1. cálculo del residuo.
`for(i=1;i<=n;i++){`
$\tilde{e}(i) = -b(i)$;
`for(j=i;j<=n;j++){`
$\tilde{e}(i) = \tilde{e}(i) + A_{i,j} * xa(j)$; // se usan coordenadas de x_k.
`}`

`for(j=1;j<i;j++){`
$\tilde{e}(i) = \tilde{e}(i) + A_{i,j} * xb(j)$// se usan coordenadas de x_{k+1}.
`}`
$norma = norma + \tilde{e}(i)^2$;
$\tilde{e}(i) = \tilde{e}(i)/A_{i,i}$;
$xb(i) = xa(i) - \tilde{e}(i)$;
`}`
`if(`$norma < eadmisible$`){` Parada, la solucion es xa`}`
$xa = xb$
`}`
Parada, no hay convergencia en nmax iteraciones;
Datos de salida: Solución x del sistema o mensajes de error si la diagonal de A tiene algún cero o la iteración no converge.

ç

Ejercicio 3.6 Considera las matrices:

$$A_1 = \begin{pmatrix} 2 & -1 & 1 \\ 2 & 2 & 2 \\ -1 & -1 & 2 \end{pmatrix} \qquad A_2 = \begin{pmatrix} 1 & 2 & -2 \\ 1 & 1 & 1 \\ 2 & 2 & 1 \end{pmatrix}$$

Los sistemas de ecuaciones $A_1x = \begin{pmatrix} -1 \\ 4 \\ -5 \end{pmatrix}$ y $A_2x = \begin{pmatrix} 7 \\ 2 \\ 5 \end{pmatrix}$ tienen la misma solución $x = (1, 2, -1)^t$.

(I) Tomando como vector de inicio $x_0 = (0, 0, 0)^t$ realiza las tres primeras iteraciones de de los métodos de Jacobi y Gauss-Seidel para los dos sistemas

(II) Calcula los radios espectrales de las matrices de los métodos de Jacobi y Gauss-Seidel para los dos sistemas.

(III) Estudia la convergencia de los métodos de Jacobi y Gauss-Seidel para los dos sistemas.

Ejercicio 3.7 Se considera una matriz $A = \left(\begin{smallmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{smallmatrix}\right)$ verificando $a_{11} \neq 0$ y $a_{22} \neq 0$.

Demostrar que los métodos de Jacobi y Gauss-Seidel para A convergen o divergen simultáneamente.

Encontrar una condición sobre los elementos de A que sea necesaria y suficiente para la convergencia de los dos métodos.

En el caso de que los dos métodos converjan, ¿cuál lo hace más rápidamente?

3.3.1. Convergencia de los Métodos de Jacobi y Gauss-Seidel

No existen resultados generales que digan cuál de los dos métodos será más eficaz en un sistema lineal arbitrario. Sin embargo, en algunos casos especiales se conoce la respuesta. En esta sección describimos algunos de estos resultados.

Teorema 3.3.1 (P. Stein and R. L. Rosenberg, 1948) *Si $A = (a_{i,j})$ es una matriz cuadrada tal que $a_{i,i} > 0$ para cada i y si $a_{i,j} \leq 0$ para cada $i \neq j$, entonces se cumple una de las siguientes afirmaciones excluyentes relativas a las matrices de los métodos de Jacobi (T_J) y Gauss-Seidel (T_G):*

(I) $0 \leq \rho(T_G) \leq \rho(T_J) < 1$

(II) $1 < \rho(T_J) \leq \rho(T_G)$

(III) $\rho(T_J) = \rho(T_G) = 1$

(IV) $\rho(T_J) = \rho(T_G) = 0$

Una prueba completa está en la sección 3 del capítulo 8 libro de Hämmerlin y Hoffman [7]. Observad que la hipótesis de este resultado equivale a que los elementos de T_J sean todos positivos. El teorema dice que en este caso los dos métodos convergen o divergen simultáneamente y que en el caso de converger Gauss-Seidel lo hace más rápidamente que Jacobi, mientras que si divergen Gauss-Seidel sigue dando la nota divergiendo también más rápidamente que Jacobi.

Para las matrices especiales del capítulo anterior se tienen buenos criterios de convergencia:

Teorema 3.3.2 *Sea A una matriz diagonal estrictamente dominante:*

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

para todo $i = 1, ..., n$. Entonces el método de Jacobi y el método de Gauss-Seidel para resolver el sistema $Ax = b$ son convergentes, y el método de Gauss-Seidel converge al menos a la misma velocidad que el de Jacobi. De forma más concreta

$$\|T_G\|_\infty \leq \|T_J\|_\infty < 1.$$

DEMOSTRACIÓN: [Prueba completa en la sección 3 del capítulo 8 libro de Hämmerlin y Hoffman [7].]
Ideas que intervienen

- *(1) Primero se prueba que* $\|T_J\|_\infty = \max\{\sum_{j\neq i} \frac{|a_{i,j}|}{|a_{i,i}|} : i = 1, ..., n\} < 1$.
- *(2) Si* $z = T_G y$, *se prueba por inducción en* $i = 1, 2, ..., n$ *que*

$$|z_i| \leq \left(\sum_{j\neq i} \frac{|a_{i,j}|}{|a_{i,i}|} \right) \|y\|_\infty \leq \|T_J\|_\infty \|y\|_\infty$$

  y de la definición de la norma subordinada y (1) $\|T_G\|_\infty \leq \|T_J\|_\infty < 1$.

□

Ejercicio 3.8 (ver [7, pág. 351-352]) Una matriz $n \times n$ $A = (A_{i,j})$ de números reales o complejos se dice que es descomponible, si existen dos conjuntos disjuntos no vacíos N_1 y N_2 tales que $N_1 \bigcup N_2 = \{1, 2, ..., n\}$ y $a_{i,j} = 0$ cuando $i \in N_1$ y $j \in N_2$. En estos casos las soluciones de los sistemas lineales $Ax = b$ se pueden encontrar resolviendo dos sistemas lineales más pequeños (de ahí el término descomponible). Comprobar que para sistemas lineales cuya matriz de coeficientes A no es descomponible (por ejemplo si es tridiagonal con las diagonales inferiores y superiores sin ceros) el método de Jacobi es convergente si A tiene diagonal débilmente dominante, en el sentido de que

$$\sum_{j=1, j\neq i}^{n} |a_{i,j}| \leq |a_{i,i}| \qquad \text{si } 1 \leq i \leq n, \text{ y}$$

$$\sum_{j=1, j\neq i_0}^{n} |a_{i_0,j}| < |a_{i_0,i_0}| \quad \text{para algún } 1 \leq i_0 \leq n.$$

Teorema 3.3.3 *Sea A una matriz tridiagonal. Entonces los radios espectrales de las matrices de los métodos de Jacobi y Gauss-Seidel cumplen:*

$$\rho(T_G) = \rho(T_J)^2.$$

Así los métodos de Jacobi y de Gauss-Seidel para resolver el sistema $Ax = b$ *convergen simultáneamente y cuando lo hacen, el método de Gauss-Seidel converge más rápidamente que el de Jacobi.*

La demostración de este teorema la podéis seguir en el libro de Ciarlet [5, Th. 5.3-4].
DEMOSTRACIÓN:
Ideas que intervienen

- *Considera los polinomios* q_J *y* q_G *de los ejercicios 3.2 y 3.4.*
- *Utilizando el Ejercicio 8 de la relación 2 de ejercicios del curso, comprueba que*

$$q_G(\lambda^2) = \lambda^n q_J(\lambda)$$

  de donde resulta que

$$\lambda \in \sigma(T_J) \setminus \{0\} \Leftrightarrow \lambda^2 \in \sigma(T_G) \setminus \{0\}.$$

□

Ejercicio 3.9 Recuerda el ejercicio 2.6 en el que se considera el sistema lineal

$$\begin{pmatrix} 2 & -1 & & \\ -1 & 2 & -1 & \\ & -1 & 2 & -1 \\ & & -1 & 2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} \begin{pmatrix} 19 \\ 19 \\ -3 \\ -12 \end{pmatrix}$$

Comprueba, utilizando la norma subordinada a la del supremo y el ejercicio 3.8, que el radio espectral de la matriz de Jacobi es estrictamente menor que 1. Consecuentemente tanto Jacobi como Gauss-Seidel son métodos iterativos convergentes siendo Gauss-Seidel más rápido que Jacobi. Realiza algunas iteraciones de Gauss-Seidel para aproximar la solución.

En la siguiente sección veremos que los dos métodos son convergentes cuando la matriz es tridiagonal definida positiva (Teorema 3.4.2). Este es el caso de este tipo de matrices.

Por otra parte, de forma análoga a como se trabaja con los polinomios de Tchevichev, se pueden encontrar los valores y vectores propios de la matriz anterior en cualquier dimensión n (n=4 en este caso) y de las matrices de Jacobi correspondientes, pudiendo evaluar $\rho(T_J) = \cos\frac{\pi}{n}$. *Véase [2, sec 5.3.3 y 8.4]*

3.4. Método de Relajación

En la construcción de la sucesión de Gauss-Seidel hemos ido definiendo x_{k+1} coordenada a coordenada conjuntamente con las coordenadas del vector residual $\tilde{r}_k$:

`Para` $i = 1$, `mientras que` $i \leq n$, `haciendo en cada paso` $i = i + 1$`:`

$$\begin{aligned} \tilde{r}_i^k &= a_{ii}x_i^k + \sum_{j=1}^{i-1} a_{ij}x_j^{k+1} + \sum_{j=i+1}^{n} a_{ij}x_j^k - b_i \\ x_i^{k+1} &= x_i^k - \frac{1}{a_{ii}}\tilde{r}_i^k \end{aligned} \tag{3.2}$$

Los métodos de relajación consisten en considerar un peso $\omega > 0$ para corregir las coordenadas de x_k poniendo en la ecuación 3.2

$$x_i^{k+1} = x_i^k - \frac{\omega}{a_{ii}}\tilde{r}_i^k$$

Observad que ahora

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

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

Con esta modificación, en términos matriciales, el método de relajación consiste en considerar

$$(D + \omega L)x_{k+1} = ((1-\omega)D - \omega U)x_k + \omega b.$$

Así, la matriz de la iteración del método de relajación es

$$T_{R(\omega)} = (D+\omega L)^{-1}((1-\omega)D - \omega U) = (1/\omega D + L)^{-1}(\frac{1-\omega}{\omega}D - U) = M^{-1}N.$$

con $A = M - N$, $(M = 1/\omega D + L$ y $N = \frac{1-\omega}{\omega}D - U)$.

Algoritmo 3.3 **Método de Relajación para resolución de sistemas lineales**

Datos de entrada: $A[n][n]$ (Matriz de coeficientes del sistema.);
$b[n]$ (vector término independiente.);
n (dimensión de A y b);
ω (parámetro de construcción);
ε (precisión para la condición de parada);
$nmax$ (número máximo de iteraciones);
Variables:$xa[n]$;// vector para aproximar la solución del sistema.
$xb[n]$;// vector para las nuevas aproximaciones de la solución del sistema.
$\tilde{e}[n]$;// un vector de corrección para xa.
$eadmisible = 0$;// precisión admisible.
$norma = 0$; // registro para el cuadrado de la norma de la corrección.
Fujo del programa:
// Condiciones iniciales y evaluacion de la diagonal
`for(j=1;j<=n;j++){`
$xa(j) = 1$; $eadmisible = eadmisible + b(j)^2$;
`if(`$A_{i,i} == 0$`){` ERROR; Relajación no es aplicable;`}`
`}`
$eadmisible = \varepsilon^2 * eadmisible$ // $(\varepsilon * \|b\|)^2$.
// Vamos a hacer las etapas $k = 1, 2, ..., nmax$.
`for(k=1;k<=nmax;k++){`
$norma = 0$; // 1. cálculo de la corrección.
`for(i=1;i<=n;i++){`
$\tilde{e}(i) = -b(i)$;
`for(j=i;j<=n;j++){`
$\tilde{e}(i) = \tilde{e}(i) + A_{i,j} * xa(j)$; `}`
`for(j=1;j<i;j++){`
$\tilde{e}(i) = \tilde{e}(i) + A_{i,j} * xb(j)$; `}`
$norma = norma + \tilde{e}(i)^2$;
$\tilde{e}(i) = \tilde{e}(i) * \omega / A_{i,i}$;
$xb(i) = xa(i) - \tilde{e}(i)$;
`}`
`if(`$norma < eadmisible$`){` Parada, la solucion es xa`}`
$xa = xb$
`}`
Parada, no hay convergencia en nmax iteraciones;
Datos de salida: Solución x del sistema o mensajes de error si la diagonal de A tiene algún cero o la iteración no converge.

Ejemplo 3.4.1 En el caso concreto del ejemplo 3.1.1 el método de Gauss-Seidel no rebaja el número de iteraciones utilizadas por el de Jacobi para aproximar la solución de $Ax = b$ con la precisión de 10^{-14}, sin embargo el método de relajación con $w = 0.85$ si que las rebaja significativamente pues alcanza la solución en sólo 34 iteraciones.

En relación con la convergencia de los métodos de relajación tenemos los siguientes resultados cuyas demostraciones podéis seguir en la sección 5.3 del libro de Ciarlet [5].

Teorema 3.4.2 *Para cualquier matriz A, el radio espectral de la matriz del método de relajación $T_{R(\omega)}$ cumple siempre que $\rho(R(\omega)) > |\omega - 1|$. Por lo tanto el método de relajación sólo puede ser convergente si $0 < \omega < 2$.*

Demostración:

Ideas que intervienen

- *Si* $A = L + D + U \in \mathcal{M}_n(\mathbb{C})$
  $T_{R(\omega)} = (D + \omega L)^{-1}((1 - \omega)D - \omega U) = (1/\omega D + L)^{-1}(\frac{1-\omega}{\omega}D - U) = M^{-1}N.$
  con $A = M - N$, $(M = 1/\omega D + L$ *y* $N = \frac{1-\omega}{\omega}D - U)$.
- $\det(T_{R(\omega)}) = \det((1 - \omega)D - \omega U)/\det(D + \omega L) = (1 - \omega)^n.$
- $\rho(T_{R(\omega)})^n \geq \prod_{i=1}^{n} |\lambda_i(T_{R(\omega)})| = |\det(T_{R(\omega)})| = |1 - \omega|^n.$

□

Teorema 3.4.3 *Si A es una matriz simétrica (o Hermitiana) definida positiva, el método de relajación converge para $0 < \omega < 2$.*

Demostración:

Ideas que intervienen

- *Si* $A = L + D + U$ *es definida positiva, entonces* D *es definida positiva y* $D + \omega L$ *es no singular.*
- $T_{R(\omega)} = (D + \omega L)^{-1}((1 - \omega)D - \omega U) = (1/\omega D + L)^{-1}(\frac{1-\omega}{\omega}D - U) = M^{-1}N.$
  con $A = M - N$, $(M = 1/\omega D + L$ *y* $N = \frac{1-\omega}{\omega}D - U)$.
- $M^* + N = \frac{2-\omega}{\omega}D$ *es simétrica y definida positiva.*

  Lema 3.4.4 *Si $A = M - N$ es simétrica (o Hermitiana) y definida positiva, y si $M^* + N$ es definida positiva entonces $\rho(M^{-1}N) < 1$.*

  Para la prueba del lema basta considerar el producto escalar $< x, y >_A := y^*Ax$, la norma euclídea que define $\|x\|_A^2 = x^*Ax$ y observar que la norma subordinada $\|M^{-1}N\|_A < 1$.

□

Teorema 3.4.5 *Si A es una matriz simétrica definida positiva, tridiagonal, los métodos de Jacobi, Gauss-Seidel y Relajación para $0 < \omega < 2$ son convergentes.*

El mínimo de los radios espectrales de las matrices de los métodos de Relajación se alcanza en

$$\omega_0 = \frac{2}{1+\sqrt{1-\rho(T_J)^2}},$$

de manera que

$$\rho(T_{R(\omega_0)}) = \min_{0<\omega<2}\{\rho(T_{R(\omega)})\} < \rho(T_G) = \rho(T_J)^2 < \rho(T_J).$$

Además $\rho(T_{R(\omega_0)}) = \dfrac{\rho(T_J)^2}{(1+\sqrt{1-\rho(T_J)^2})^2} = \omega_0 - 1.$

DEMOSTRACIÓN: (La prueba la podéis encontrar en [5] o en [2, chapter 8])
Ideas que intervienen

- *El teorema 3.4.3 prueba que el método de relajación converge solo para $0 < w < 2$. Además por el teorema 3.3.3 sabemos que $1 > \rho(T_{R(1)}) = \rho(T_G) = \rho(T_J)^2$ y que los tres métodos convergen.*

  Además el teorema 3.4.3 nos asegura que $\rho(T_{R(\omega)}) < 1$ si, y sólo si, $0 < \omega < 2$.

- *Tenemos que comparar el radio espectral de $T_{R(\omega)}$ y el de T_J.*

  Si $A = L+D+U \in \mathcal{M}_n(\mathbb{C})$, con las mismas ideas de la prueba del Teorema 3.3.3 aplicadas a la matriz $B(\lambda^2) = \frac{\omega+1}{\omega}D + U + \lambda^2(1/\omega D + L) = -(1/\omega D + L)(T_{R(\omega)} - \lambda^2 Id)$. con $\mu = \frac{1}{\lambda}$ se comprueba que

$$p_{T_{R(\omega)}}(\lambda^2) = \det(T_{R(\omega)} - \lambda^2 Id) = c\lambda^n q_J\left(\frac{\lambda^2+\omega-1}{\lambda\omega}\right).$$

- *Escribiendo $\alpha = \frac{\lambda_\alpha^2+\omega-1}{\lambda_\alpha\omega}$ se tiene que*

$$\alpha \in \sigma(T_J) \Longleftrightarrow \mu = \lambda_\alpha^2 \in \sigma(T_{R(\omega)})$$

  y despejando λ_α se tiene

$$\lambda_\alpha^{\pm} = \frac{1}{2}(\alpha\omega \pm \sqrt{\alpha^2\omega^2 - 4(\omega-1)})$$

$$\mu_\alpha^{\pm} = \lambda_\alpha^{\pm 2} = \frac{1}{2}(\alpha^2\omega^2 - 2\omega + 2) \pm \frac{\alpha\omega}{2}\sqrt{\alpha^2\omega^2 - 4(\omega-1)}$$

  En el teorema 3.4.3 se probó que $\alpha \in \sigma(T_J) \Longleftrightarrow \alpha^2 \in \sigma(T_G)$, de donde se deduce que $\alpha \in \sigma(T_J) \Longleftrightarrow -\alpha \in \sigma(T_J)$. Así $|\lambda_\alpha^+| = |\lambda_{-\alpha}^-|$ y a la hora de buscar el radio espectral de $T_{R(\omega)}$ bastará considerar los valores $|\mu_\alpha^+|$

$$\rho(T_{R(\omega)}) = \max\{|\mu_\alpha^+| = \left|(\alpha^2\omega^2 - 2\omega + 2) + \alpha\omega\sqrt{\alpha^2\omega^2 - 4(\omega-1)}\right|/2 : \alpha \in \sigma(T_J)\}$$

- *Observad que si A es PD, $A = L + D + U$, entonces todos los elementos de la diagonal D son positivos ($a_ii = \vec{e_i}^t A\vec{e_i} > 0$), y si $\alpha \in \in \sigma(T_J)$ con $v \neq 0$ y $Jv = -D^{-1}(L+U)v = \alpha v$, entonces $Av = Dv + (L+U)v = Dv - \alpha Dv = (1-\alpha)Dv$, y $(1-\alpha)v^tDv = v^tAv > 0$. Teniendo en cuenta que $v^tDv > 0$ resulta que $\alpha \in \mathbb{R}$ y además $\alpha \in (-1, +1)$.*

- *Para medir el módulo $|\mu_\alpha^+|$ hay que distinguir cuando $p(w) = \alpha^2\omega^2 - 4(\omega-1) < 0$. Calculando las dos raíces de $p(w)$ esto sucede para*

$$w_\alpha^+ = \frac{2}{1+\sqrt{1-\alpha^2}} < \omega < w_\alpha^- = \frac{2}{1-\sqrt{1-\alpha^2}}$$

Como $0 < w < 2 < w_\alpha^-$, el módulo de $|\mu_\alpha^+|$ toma el valor:

$$|\mu_\alpha^+| = \begin{cases} |\omega - 1| & \text{si } 1 < w_\alpha^+ < \omega < 2 \quad (\mu_\alpha^+ \in \mathbb{C}) \\ \left|(\alpha^2\omega^2 - 2\omega + 2) + \alpha\omega\sqrt{\alpha^2\omega^2 - 4(\omega-1)}\right|/2 & \text{si } 0 < \omega \le w_\alpha^+ \quad (\mu_\alpha^+ \in \mathbb{R}) \end{cases}$$

- *Considerad la función de 2 variables:*

$$f(\alpha, \omega) = \mu_\alpha^+ = {\lambda_\alpha}^2.$$

(I) *Si $0 < \omega < \omega_\alpha^+$, teniendo en cuenta que $\sqrt{\alpha^2\omega^2 - 4(\omega-1)} = 2\lambda_\alpha^+ - \alpha\omega$ se cumple que*

$$\begin{aligned} \frac{\partial f(\alpha,\omega)}{\partial \alpha} &= 2\lambda_\alpha^+ \frac{\partial \lambda_\alpha^+}{\partial \alpha} \\ &= \lambda_\alpha^+ \left(\omega + \frac{\alpha\omega^2}{\sqrt{\alpha^2\omega^2 - 4(\omega-1)}} \right) \\ &= \lambda_\alpha^+ \left(\frac{\omega(2\lambda_\alpha^+ - \alpha\omega) + \alpha\omega^2}{\sqrt{\alpha^2\omega^2 - 4(\omega-1)}} \right) \\ &= \lambda_\alpha^+ \left(\frac{2\lambda_\alpha^+\omega}{\sqrt{\alpha^2\omega^2 - 4(\omega-1)}} \right) > 0 \end{aligned}$$

La función $f(\alpha, \omega)$ es creciente en la variable α si

$$0 < \alpha \le \rho(T_J) \text{ y } 0 < w \le w_{\rho(T_J)}^+ =: \omega_0.$$

Observad que también se cumple que $w < w_\alpha^+$ y por lo tanto

$$\rho(T_{R(\omega)}) = \max\{f(\alpha,\omega) : \alpha \in \sigma(T_J), \alpha > 0\} = f(\rho(T_J), \omega).$$

(II) *También si $0 < \omega < \omega_\alpha^+$, derivando con respecto a ω*

$$\begin{aligned} \frac{\partial f(\alpha,\omega)}{\partial \omega} &= 2\lambda_\alpha^+ \frac{\partial \lambda_\alpha^+}{\partial \omega} \\ &= \lambda_\alpha^+ \left(\alpha + \frac{2\alpha^2\omega - 4}{2\sqrt{\alpha^2\omega^2 - 4(\omega-1)}} \right) \\ &= \lambda_\alpha^+ \left(\frac{\alpha(2\lambda_\alpha^+ - \alpha\omega) + \alpha^2\omega - 2}{\sqrt{\alpha^2\omega^2 - 4(\omega-1)}} \right) \\ &= \lambda_\alpha^+ \left(\frac{2(\alpha\lambda_\alpha^+ - 1)}{\sqrt{\alpha^2\omega^2 - 4(\omega-1)}} \right) < 0 \end{aligned}$$

ya que $0 < \alpha < \rho(T_J) < 1$ y $\lambda_\alpha^2 \le \rho(T_{R(w)}) < 1$, por lo que $(\alpha\lambda_\alpha^+ - 1) < 0$.

La función $f(\alpha, \omega)$ es decreciente en la variable ω si

$$0 < \alpha \leq \rho(T_J) \text{ y } 0 < \omega \leq \omega^+_{\rho(T_J)} =: \omega_0.$$

Observad que también se cumple que $w < w^+_\alpha$ y por lo tanto

$$\rho(T_{R(\omega)}) = f(\rho(T_J), \omega) \geq f(\rho(T_J), \omega^+_{\rho(T_J)}) = \rho(T_{R(\omega_0)}).$$

- *Es fácil comprobar que*

$$\rho(T_{R(\omega_0)}) = f(\rho(T_J), \omega^+_{\rho(T_J)}) = \omega^+_{\rho(T_J)} - 1 = \omega_0 - 1$$

- *Por último, si $1 < \omega_0 < \omega < 2$*

$$\rho(T_{R(\omega)}) \geq |\omega - 1| = \omega - 1 > \omega_0 - 1 = (T_{R(\omega_0)})$$

.

□

El parámetro ω_0 nos da el método iterativo óptimo, con convergencia más rápida, de entre los estudiados hasta este momento.

Ejercicio 3.10 Considera el sistema de ecuaciones

$$\begin{cases} 10x_1 - x_2 = 9 \\ -x_1 + 10x_2 - 2x_3 = 7 \\ -2x_2 + 10x_3 = 6 \end{cases}$$

que tiene matriz de coeficientes con tridiagonal .

(I) Demuestra que los métodos de Jacobi y Gauss-Seidel para este sistema son convergentes.

(II) Evalúa las primeras iteraciones de estos métodos comenzando en el origen $x^0 = (0, 0, 0,)^t$.

(III) Comprueba si su matriz de coeficientes es definida positiva y en el caso afirmativo, encuentra la elección óptima de w_0 para el método de relajación y efectúa las primeras iteraciones de ese método empezando en el origen.

Ejercicio 3.11 Prueba que para $0 < w \leq 1$ y $\lambda \in \mathbb{C}$ con $|\lambda| \geq 1$ se cumple: $\left|\frac{1-w-\lambda}{\lambda w}\right| \geq 1$.

Demuestra que si A es una matriz con diagonal estrictamente dominante, el método de relajación para A es convergente si $0 < w \leq 1$.

Indicación: Para la primera desigualadad utiliza la desigualdad triangular $|\lambda + w - 1| \geq |\lambda| - (1 - w)$ y para la demostración de la convergencia del método de relajación en el caso de matrices con diagonal estrictamente dominante el que $\rho(T_J) < 1$ y parte de la prueba del Teorema 3.4.5

3.5. Método del gradiente conjugado

El método del gradiente conjugado (Hestenes y Stiefel 1952) fue creado originalmente como un método directo para resolver sistemas lineales de n ecuaciones cuya matriz de coefiecientes es definida positiva. Como método directo es, en general, menos eficiente que el método de Gauss porque los dos necesitan n pasos y los cálculos en el del gradiente son más costosos. Sin embargo, cuando la matriz de coeficientes es simétrica definida positiva con muchos ceros y la matriz esta preparada ("precondicionada") para que los cálculos sean eficientes, se tienen muy buenas aproximaciones en $\sqrt{n}$ iteraciones. Por eso para valores de n grandes el método del gradiente conjugado se ha convertido en el más popular de los métodos utilizados.

La primera idea a tener en cuenta en este método es la de identificar la solución del sistema de ecuaciones

$$Ax = b$$

como el valor donde alcanza el mínimo absoluto una forma cuadrática. En otras palabras, buscar la solución de un problema lineal como la solución de un problema de optimización (no lineal).

Teorema 3.5.1 *Si A es una matriz simétrica definida positiva, son equivalentes*

(I) $xs \in \mathbb{R}^n$ es la solución del sistema lineal $Ax = b$, y

(II) $xs \in \mathbb{R}^n$ es el valor que minimiza la función $g(x) = x^*Ax - 2x^*b$, $g : \mathbb{R}^n \to \mathbb{R}$.

DEMOSTRACIÓN: Véase [4, Th 7.31 pag. 466].

Si x y $v \neq 0$ son dos vectores fijos y $t \in \mathbb{R}$

$$\begin{aligned} g(x+tv) &= (x+tv)^*A(x+tv) - 2(x+tv)^*b \\ &= x^*Ax + tv^*Ax + tx^*Av + t^2v^*Av - 2x^*b - t2v^*b \\ &= g(x) - 2tv^*(b-Ax) + t^2v^*Av \end{aligned}$$

la parabola $h(t) = g(x+tv)$ tiene su mínimo ($v^*Av > 0$) en $t_{x,v}$ con $h'(t_{x,v}) = 0$

$$h'(t_{x,v}) = 2(-v^*(b-Ax) + t_{x,v}v^*Av) = 0 \Longleftrightarrow t_{x,v} = \frac{v^*(b-Ax)}{v^*Av}, \text{ y}$$

$$h(t_{x,v}) = g(x) - \frac{(v^*(b-Ax))^2}{v^*Av}$$

Así, para cada $v \neq 0$, $g(x+t_{x,v}) < g(x)$ salvo si $v^*(b-Ax) = 0$, en cuyo caso $g(x+t_{x,v}) = g(x)$.

Ahora la prueba del teorema es fácil:

Si $Ax = b$, $v^*(b-Ax) = 0$ para cualquier dirección $v \neq 0$, y $g(x) = g(x+t_{x,v}v)$ es el mínimo de g en cualquier dirección.

Recíprocamente, si g tiene un mínimo en x, para cualquier dirección $v \neq 0$ $g(x) = g(x+t_{x,v}v)$, y en consecuencia $v^*(b-Ax) = 0$, en particular para $v = b-Ax$, $\|b-Ax\|^2 = (b-Ax)^*(b-Ax) = 0$ y $Ax = b$. □

Recordad, del cálculo con funciones de varias variables, que dado un vector $x \in \mathbb{R}^n$ la dirección de máximo crecimiento de g viene dada por el vector gradiente

$$\nabla g(x) = \left(\frac{\partial g(x)}{\partial x_1}, \frac{\partial g(x)}{\partial x_2}, ..., \frac{\partial g(x)}{\partial x_n} \right),$$

y por lo tanto la dirección de máximo decrecimiento de g es $-\nabla g(x)$. Así, dada una aproximación xi al mínimo de g, se puede buscar una mejor aproximación en la dirección de máximo decrecimiento $-\nabla g(xi)$, $xf = xi - \alpha \nabla g(xi)$, buscando un menor valor de $g(xf)$. Esta es la idea del método del descenso rápido que estudiaremos para resolver ecuaciones no lineales 5.4.

En el caso de la función $g(x) = x^*Ax - 2x^*b$ del teorema anterior, el cálculo del gradiente es muy sencillo:

Proposición 3.5.2

$$\nabla g(x) = \left(\frac{\partial g(x)}{\partial x_1}, \frac{\partial g(x)}{\partial x_2}, ..., \frac{\partial g(x)}{\partial x_n} \right) = 2(Ax - b) = -2r(x),$$

donde $r(x) = b - Ax$ es el vector residual en x que usamos frecuentemente para medir la aproximación de x a la solución del sistema lineal.

DEMOSTRACIÓN: Véase [4, pag. 467]. □

Hay que señalar que el método del descenso rápido aplicado a la busqueda del mínimo de g no es muy eficiente porque aunque puede ser seguro, no es nada rápido y requiere de muchos cálculos.

El método del gradiente conjugado es una alternativa al método del descenso rápido que consiste en buscar una base ortogonal para un determinado producto escalar en $\mathbb{R}^n$ que proporciona un método directo para localizar la solución del sistema lineal.

Para elaborar esta receta nuestro primer ingrediente va a ser el producto escalar asociado a la matriz simétrica definida positiva A:

Definición 3.5.3 *Si A es una matriz simétrica definida positiva de dimensión n, el producto definido por*

$$< x, y >_A = y^*Ax$$

es un producto escalar en $\mathbb{R}^n$ y $\|x\|_A = \sqrt{x^*Ax}$ define la norma euclídea asociada.

Ejercicio 3.12 Comprobad las propiedades del producto escalar para el producto definido en la definición anterior, 3.5.3

El segundo ingrediente es el siguiente teorema que proporciona un metodo directo de resolución del sistema de ecuaciones $Ax = b$ en un proceso de n pasos que dependen de un sistema ortogonal de vectores para el producto $<,>_A$ (sistema A-ortogonal).

Teorema 3.5.4 *Sean A una matriz simétrica definida positiva de dimensión n, $\{v_1, ..., v_n\}$ un sistema A-ortogonal de vectores en $\mathbb{R}^n$, y sea $x_0 \in \mathbb{R}^n$ un vector arbitrario. Sean*

(I) $r_{k-1} = b - Ax_{k-1} = -0.5\nabla g(x_{k-1})$ *el residual en* x_{k-1},

(II) $t_k = \frac{v_k^* r_{k-1}}{v_k^* A v_k} = \frac{v_k^*(b - Ax_{k-1})}{v_k^* A v_k}$ *y* $x_k = x_{k-1} + t_k v_k$

para $k = 1, 2, ..., n$. Entonces $Ax_n = b$, e.d. x_n es la solución del sistema lineal $Ax = b$.

DEMOSTRACIÓN:
Ideas que intervienen

- *Se trata de comprobar que por la construcción de las iteradas x_k, el residual $Ax_n - b$ es perpendicular a todos los vectores del sistema A-ortogonal.*

  De la definición resulta que

$$r_k = b - Ax_k = r_{k-1} - t_k Av_k = b - Ax_0 - t_1 Av_1 - ... - t_k Av_k.$$

  (I) *para cada $j \leq k$ como los vectores v_j son A-ortogonales, se tiene*

$$v_j^* r_k = v_j^* r_0 - t_j \, v_j^* Av_j;$$

  (II) *para cada $j > m$ se cumple $v_j^* r_m = v_j^* r_0$;*

  (III) *para cada j de la definición de los t_j se tiene que $t_j \, v_j^* Av_j = v_j^* r_{j-1} = v_j^* r_0$*

  reuniendo (I) y (III), se tiene que

$$v_j^* r_k = 0 \text{ para todo } j \leq k.$$

  En particular $r_n \in \mathbb{R}^n$ es un vector ortogonal (para el producto escalar usual) a todos los n vectores del sistema A-ortogonal, que forman una base de $\mathbb{R}^n$, en consecuencia, $Ax_n - b = r_n = 0$.

□

Obsérvese que de la prueba anterior se tiene que los residuales r_k son ortogonales a los vectores $v_1, ..., v_k$.

El tercer ingrediente en el método del gradiente conjugado es el ir construyendo los vectores del sistema A-ortogonal, v_k, de manera que los vectores residuales r_k sean mutuamente ortogonales

(I) Partimos de una aproximación inicial x_0 a la solución del sistema lineal y consideramos la dirección de máximo descenso como primer vector del sistema A-ortogonal para buscar una mejor aproximación:
$v_1 = r_0 = b - Ax_0$, $t_1 = \frac{r_0^* r_0}{r_0^* Ar_0} = \frac{\|r_0\|_2^2}{r_0^* Ar_0}$ y $x_1 = x_0 + t_1 r_0$, ($s_1 = 0$).

Observad que $r_1 = b - Ax_1 = r_0 - \frac{r_0^* r_0}{r_0^* Ar_0} Ar_0$, de manera que

$$< r_0, r_1 >= r_0^* r_0 - \frac{r_0^* r_0}{r_0^* Ar_0} r_0^* Ar_0 = 0.$$

Supongamos construidas las aproximaciones de la solución $x_0, x_1, ..., x_{k-1}$ $(k \geq 2)$ y los vectores A-ortogonales $v_1, ..., v_{k-1}$ de manera que
$< v_i, v_j >_A = v_j^* Av_i = 0$ y $< r_i, r_j >= r_j^* r_i = 0$ para $i < j \leq k - 1$.

(II) Si $Ax_{k-1} == b \Leftrightarrow r_{k-1} = b - Ax_{k-1} == 0$ (Condición de parada: $\|r_{k-1}\|_2$ es suficientemente pequeña), hemos terminado. x_{k-1} es la solución del sistema.

En otro caso

(III) ($k \geq 2$) Buscamos construir v_k de la forma $v_k = r_{k-1} + s_{k-1}v_{k-1}$, para lo que elegimos s_{k-1} de modo que $< v_k, v_{k-1} >_A = v_{k-1}^* A v_k = 0$.

$$< v_k, v_{k-1} >_A = v_{k-1}^* A v_k = v_{k-1}^* A r_{k-1} + s_{k-1} v_{k-1}^* A v_{k-1} = 0$$

$$s_{k-1} = -\frac{v_{k-1}^* A r_{k-1}}{v_{k-1}^* A v_{k-1}}$$

$$v_k = r_{k-1} + s_{k-1} v_{k-1}.$$

También se puede demostrar que con esta elección $< v_k, v_i >_A = 0$ para $i = 1, ..., k-2$ (Esta prueba no es en absoluto trivial, véase por ejemplo [10, Sec 8, pag. 30-31]).

Una vez determinado el vector v_k , como $v_{k-1}^* r_{k-1} = 0$, al evaluar t_k se tiene

$$\begin{aligned} t_k &= \frac{v_k^* r_{k-1}}{v_k^* A v_k} \\ &= \frac{r_{k-1}^* r_{k-1}}{v_k^* A v_k} + \frac{s_{k-1} v_{k-1}^* r_{k-1}}{v_k^* A v_k} = \frac{r_{k-1}^* r_{k-1}}{v_k^* A v_k} \\ &= \frac{\|r_{k-1}\|_2^2}{v_k^* A v_k} \end{aligned}$$

$$x_k = x_{k-1} + t_k v_k, \text{ y}$$

$$r_k = b - A x_k = r_{k-1} - t_k A v_k.$$

Observad que

$$\|r_k\|_2^2 = r_k^* r_k = r_k^* r_{k-1} - t_k r_k^* A v_k$$

de donde se tiene, dado que A es simétrica,

$$v_k^* A r_k = r_k^* A v_k = \frac{-\|r_k\|_2^2}{t_k}.$$

Por otra parte de la construcción de t_k

$$v_k^* A v_k = \frac{\|r_{k-1}\|_2^2}{t_k}$$

De este modo al calcular s_k también se puede escribir

$$s_k = -\frac{v_k^* A r_k}{v_k^* A v_k} = \frac{\|r_k\|_2^2}{\|r_{k-1}\|^2}$$

(IV) El teorema 3.5.4 garantiza que alcanzaremos la solución del sistema $Ax = b$ en no más de n iteraciones.

Reuinendo todas las ideas se tiene el algoritmo del método del gradiente conjugado que en resumen consiste en

(A) Comenzamos con

$x = x_0$

$r = b - A x_0$;

$v = r$;

$\gamma = \|r_0\|^2$

(B) (Para $i = 1, ..., n$ y mientras que $\gamma > errorAdmisible$)

$y = Av$

$t = \frac{\gamma}{v^* y}$

$x = x + tv$

$r = r - ty$

$s = \frac{\|r\|_2^2}{\gamma}$

$\gamma = \|r\|_2^2$

$v = r + sv$

(fin del bucle)

(C) Si $\gamma > errorAdmisible$ mandar mensaje de Error por inestabilidad en cálculos.

Precondicionamiento

La velocidad de convergencia del método del gradiente conjugado para resolver un sistema de ecuaciones $Ax = b$ depende del tamaño del número de condición $cond_2(A)$. En concreto, considerando la función creciente en $[1, +\infty)$ definida por $\varphi(t) = \left(\frac{t-1}{t+1}\right)$, se tiene que la siguiente estimación de la distancia de la iterada k-ésima en el método del gradiente conjugado, x_k, a la solución x del sistema lineal:

$$\|x_k - x\|_2 \leq 2\sqrt{cond_2(A)}\varphi(cond_2(A))^k \|x_0 - x\|_2,$$

(véase [2, Prop. 9.5.1]).

Esta estimación nos debe hacer pensar que si pasamos del sistema $Ax = b$ a otro equivalente con menor número de condición, conseguiremos acelerar la convergencia del método.

Definición 3.5.5 (Precondicionamiento) *Sea $Ax = b$ el sistema lineal a resolver, con A una matriz simétrica definida positiva. Se llama precondicionamiento de A, a una matriz fácil de invertir C tal que $\widetilde{A} = C^{-1}A(C^{-1})^t$ que sigue siendo simétrica definida positiva cumpla*

$$cond_2(\widetilde{A}) < cond_2(A).$$

Al sistema equivalente $(C^{-1}A(C^{-1})^t)\tilde{x} = C^{-1}Ax = C^{-1}b$, con $\tilde{x} = C^t x$, se le llama sistema precondicionado.

Normalmente se presenta el método del gradiente conjugado incluyendo la acción de una matriz de "precondicionamiento" porque si la matriz A está mal condicionada el método del gradiente es muy inestable. Así en lugar de aplicar el método a la ecuación $Ax = b$ se considera una matriz C de manera que la matriz simétrica $\widetilde{A} = C^{-1}A(C^{-1})^t$ esté mejor condicionada y se aplica el método del gradiente conjugado al sistema $\widetilde{A}\widetilde{x} = C^{-1}b$.

Por ejemplo, si A tiene diagonal dominante los valores de los elementos de la diagonal, aproximan a los valores propios de A, así el número de condición de $cond_2(A)$ estaría cerca del cociente entre el mayor valor absoluto y el menor valor absoluto de los elementos de la diagonal. En estos casos si se considera como matriz de precondicionamiento, $C = D^{1/2}$ los valores de la diagonal de $\widetilde{A}$ estarían más equilibrados y $\widetilde{A}$ estaría mejor condicionada.

Obsérvese que a la hora de implementar el método al sistema $\widetilde{A}\widetilde{x} = C^{-1}b$ se va a poder hacer de forma "implicita" en el siguiente sentido (véase [4, Alg. 7.5, pag 473]). Las asignaciones recuadradas son las necesarias para la elaboración del algoritmo.

- Sea $P = CC^t$ (facilitará los cálculos)
- si $\widetilde{x_k} = C^t x_k$ y $\widetilde{v_k} = C^t v_k$,
  $\widetilde{v_k}^t \widetilde{A} \widetilde{v_k} = v_k^t C (C^{-1} A (C^t)^{-1}) C^t v_k = v_k^t A v_k$.
- $\widetilde{r_k} = C^{-1} b - C^{-1} A (C^{-1})^t \widetilde{x_k} = C^{-1} r_k$.
  $\boxed{z_k = P^{-1} r_k} \iff$ resolver $\boxed{P z_k = r_k}$ (fácil si C es triangular porque equivale a dos sistemas triangulares resolubles por los métodos ascendente y descendente).
  $\widetilde{r_k}^t \widetilde{r_k} = (C^{-1} r_k)^t C^{-1} r_k = r_k^t (C^{-1})^t C^{-1} r_k = r_k^t (P^{-1} r_k) = r_k^t z_k$.
- $\boxed{\widetilde{t_k} = \frac{r_{k-1}^t z_{k-1}}{v_k^t A v_k}}$
- $\widetilde{x_k} = \widetilde{x_{k-1}} + \widetilde{t_k} \widetilde{v_k}$, de donde se tiene
  $\boxed{x_k = x_{k-1} + \widetilde{t_k} v_k}$
- $\widetilde{r_k} = \widetilde{r_{k-1}} - \widetilde{t_k} \widetilde{A} \widetilde{v_k}$, de donde se tiene
  $C^-1 r_k = C^{-1} r_{k-1} - \widetilde{t_k} C^{-1} A v_k$, y
  $\boxed{r_k = r_{k-1} - \widetilde{t_k} A v_k.}$
  $\boxed{z_k = P^{-1} r_k}$
- $\boxed{\widetilde{s_k} = \frac{{r_k}^t z_k}{r_{k-1}^t z_{k-1}}}$
- Por último $C^t v_k = \widetilde{v_k} = \widetilde{r_k} + \widetilde{s_k} C^t v_{k-1}$ y así
  $v_k = (C^{-1})^t \widetilde{r_k} + \widetilde{s_k} v_{k-1} = (C^{-1})^t C^{-1} r_k + \widetilde{s_k} v_{k-1} = P^{-1} r_k + \widetilde{s_k} v_{k-1} = z_k + \widetilde{s_k} v_{k-1}$
  $\boxed{v_k = z_k + \widetilde{s_k} v_{k-1}}$

Ejercicio 3.13

(I) Describir el algoritmo del gradiente conjugado con precondicionamiento.

(II) Implementar el algoritmo como un método de vuestro proyecto de prácticas y comprobad su funcionamientos mimetizando los ejemplos 2 y 3 de la sección 7.5 de [4].

(III) Resolver el ejercicio 11 de la sección 7.5 de [4].

3.6. Actividades complementarias del capítulo

Los documentos se pueden descargar de la Zona de Recursos en el AULA VIRTUAL.

- Hoja de problemas nº3
- Prácticas 5 y 6.

Bibliografía

[1] A. Delshams A. Aubanell, A. Benseny, *Útiles básicos de cálculo numérico*, Ed. Labor - Univ. Aut. de Barcelona, Barcelona, 1993.

[2] G. Allaire and S.M. Kaber, *Numerical linear algebra*, TAM 55, Springer,, New York, etc., 2008.

[3] C. Brézinski, *Introducton á la pratique du calcul numérique*, Dunod Université, Paris, 1988.

[4] R.L. Burden and J.D. Faires, *Análisis numérico, 7ª edición*, Thomson-Learning, Mexico, 2002.

[5] P.G. Ciarlet, *Introduction à l'analyse numérique matricielle et à l'optimisation*, Masson, Paris, 1990.

[6] P.M. Cohn, *Algebra, vol. 1 y 2*, Ed. Hohn Wiley and Sons, London, 1974.

[7] G. Hammerlin and K.H. Hoffmann, *Numerical mathematics*, Springer-Verlag, New York, 1991.

[8] J.Stoer and R. Burlisch, *Introduction to numerical analysis*, Springer Verlag, New York, 1980.

[9] D. Kincaid and W. Cheney, *Análisis numérico*, Ed. Addison-Wesley Iberoamericana, Reading, USA, 1994.

[10] J. R. Shewchuk, *An introduction to the conjugate gradient method without the agonizing pain*, Tech. report, Carnegie Mellon University, USA,https://www.cs.cmu.edu/ quake-papers/painless-conjugate-gradient.pdf, 1994.