Métodos iterativos para sistemas lineales de ecuaciones

Laboratori de Càlcul Numèric (LaCàN)

23 de septiembre de 2010

Se desea resolver el sistema lineal de ecuaciones

$$\mathbf{A}\mathbf{x} = \mathbf{b},$$

con $\mathbf{A} \in \mathbb{R}^{n\times n}$ regular, $\mathbf{x} \in \mathbb{R}^n$ y $\mathbf{b} \in \mathbb{R}^n$. Se define el residuo

$$\mathbf{r}(\mathbf{x}) := \mathbf{A}\mathbf{x} - \mathbf{b}$$

y se denota con un asterisco, $\mathbf{x}^*$, a la solución del sistema,

$$\mathbf{A}\mathbf{x}^* = \mathbf{b} \quad \Leftrightarrow \quad \mathbf{x}^* = \mathbf{A}^{-1}\mathbf{b} \quad \Leftrightarrow \quad \mathbf{r}(\mathbf{x}^*) = \mathbf{0}.$$

1. Métodos iterativos estacionarios

Un método iterativo estacionario es un esquema iterativo que genera una sucesión de vectores $\{\mathbf{x}^0, \mathbf{x}^1, \ldots, \mathbf{x}^k\}$ a partir de una expresión de la forma

$$\mathbf{x}^{k+1} = \mathbf{G}\mathbf{x}^k + \mathbf{h}, \tag{1}$$

donde $\mathbf{G}$ y $\mathbf{h}$ son una matriz y un vector constantes que no dependen de la iteración k. La sucesión de vectores generados a partir de la fórmula anterior debería converger a la única solución del sistema de ecuaciones planteado. Es decir,

$$\lim_{k\to\infty} \mathbf{x}^k = \mathbf{x}^*.$$

Para ello, se estudia la consistencia y la convergencia en las siguientes secciones.

En vez de construir una matriz $\mathbf{G}$ y un vector $\mathbf{h}$ es muy habitual la siguiente estructura para los métodos iterativos estacionarios: dado $\mathbf{x}^0$, hallar $\mathbf{x}^{k+1}$ a partir de $\mathbf{x}^k$ realizando los siguientes pasos

$$\mathbf{r}^k = \mathbf{A}\mathbf{x}^k - \mathbf{b} \qquad \text{(cálculo del residuo)} \tag{2a}$$

$$\mathbf{C}\,\Delta\mathbf{x}^k = -\mathbf{r}^k \qquad \text{(resolución de un sistema)} \tag{2b}$$

$$\mathbf{x}^{k+1} = \mathbf{x}^k + \Delta\mathbf{x}^k \qquad \text{(actualización de la aproximación)} \tag{2c}$$

donde se denota el residuo de la iteración k como $\mathbf{r}^k = \mathbf{r}(\mathbf{x}^k)$. La primera instrucción (2a) permite evaluar el módulo del residuo y así comprobar si se ha llegado a convergencia. La segunda instrucción (2b) implica resolver un sistema de ecuaciones. Aquí existe un compromiso entre tomar $\mathbf{C} = \mathbf{A}$ (en una iteración converge pero con coste muy elevado, a veces no factible, equivalente al de un método directo) o tomar $\mathbf{C} = \mathbf{I}$ (coste por iteración despreciable, pero requiere en principio muchas iteraciones y la convergencia no está asegurada). A la práctica, para evitar tener que escribir cada vez las ecuaciones descritas en (2), se simboliza el proceso en

$$\mathbf{x}^{k+1} = \mathbf{x}^k - \mathbf{C}^{-1}\left(\mathbf{A}\mathbf{x}^k - \mathbf{b}\right). \tag{3}$$

Esta forma simbólica *no implica la inversión de la matriz* $\mathbf{C}$, sólo es una manera de reescribir el esquema de forma compacta para dejar claro que se trata de un método de iteración funcional, $\mathbf{x}^{k+1} = \boldsymbol{\psi}(\mathbf{x}^k)$. De hecho, representa un caso particular de (1) donde

$$\begin{cases} \mathbf{G} = \mathbf{I} - \mathbf{C}^{-1}\mathbf{A} \\ \mathbf{h} = \mathbf{C}^{-1}\mathbf{b}. \end{cases} \tag{4}$$

1.1. Consistencia y convergencia

Definición (Consistencia para sistemas lineales de ecuaciones): Un método iterativo es consistente si y sólo si $\mathbf{x}^* = \mathbf{A}^{-1}\mathbf{b}$ es el único punto fijo del esquema iterativo.

Para esquemas iterativos estacionarios, exigir consistencia equivale a asegurar que

1. $\mathbf{x}^*$ es punto fijo:

$$\mathbf{x}^* = \mathbf{G}\mathbf{x}^* + \mathbf{h} \quad \Leftrightarrow \quad \mathbf{h} = (\mathbf{I} - \mathbf{G})\mathbf{A}^{-1}\mathbf{b}.$$

2. $\mathbf{x}^*$ es el único punto fijo:

   Consideremos $\mathbf{y} \in \mathbb{R}^n$ tal que $\mathbf{y} = \mathbf{G}\mathbf{y} + \mathbf{h}$, entonces,

   $$(\mathbf{x}^* - \mathbf{y}) = \mathbf{G}(\mathbf{x}^* - \mathbf{y}) \quad \Leftrightarrow \quad (\mathbf{I} - \mathbf{G})(\mathbf{x}^* - \mathbf{y}) = \mathbf{0}.$$

   Para que este sistema tenga sólo la solución trivial, es decir $\mathbf{y} = \mathbf{x}^*$, hay que exigir que $(\mathbf{I} - \mathbf{G})$ sea regular.

Por lo tanto, el esquema iterativo $\mathbf{x}^{k+1} = \mathbf{G}\mathbf{x}^k + \mathbf{h}$ es consistente si y sólo si $\mathbf{h} = (\mathbf{I} - \mathbf{G})\mathbf{A}^{-1}\mathbf{b}$ y $(\mathbf{I} - \mathbf{G})$ es regular. En el caso particular de los esquemas que se escriben como (3), es trivial verificar (se deja como ejercicio) que $\mathbf{x}^*$ es el único punto fijo. Por consiguiente, todos los esquemas que pueden escribirse según (3) son consistentes.

Definición (Convergencia): un método iterativo es convergente si el error en la iteración k, $\mathbf{e}^k := \mathbf{x}^* - \mathbf{x}^k$, verifica $\lim_{k\to\infty} \mathbf{e}^k = \mathbf{0}$ para cualquier $\mathbf{e}^0$.

Para estudiar las condiciones de convergencia de los métodos iterativos estacionarios se analiza la evolución del error, $\mathbf{e}^k$, iteración a iteración. Se supone la consistencia del esquema, es decir se asume que (único punto fijo)

$$\mathbf{x}^* = \mathbf{G}\mathbf{x}^* + \mathbf{h} \tag{5}$$

Restando las ecuaciones (5) y (1), se obtiene la ecuación que relaciona el error en dos iteraciones consecutivas,

$$\mathbf{e}^{k+1} = \mathbf{G}\,\mathbf{e}^k. \tag{6}$$

Aplicando esta igualdad recursivamente se obtiene una expresión que relaciona el error con el error en la aproximación inicial

$$\mathbf{e}^k = [\mathbf{G}]^k\,\mathbf{e}^0,$$

donde $[\mathbf{G}]^k$ denota la potencia k de la matriz $\mathbf{G}$. De manera que, el esquema iterativo será convergente si

$$\lim_{k\to\infty} [\mathbf{G}]^k = \mathbf{0}.$$

El teorema 2 demuestra que esta condición es equivalente a decir que el esquema será convergente si y sólo si el radio espectral de la matriz $\mathbf{G}$ es menor que 1,

$$\rho(\mathbf{G}) := \max_i |\lambda_i| \leq 1.$$

Sin embargo, la demostración de este resultado requiere de un resultado previo que se enuncia a continuación.

Teorema 1. *Para cada matriz* $\mathbf{A}$ *de orden* n *y cada real* $\varepsilon > 0$ *arbitrario existe una norma matricial asociada a una norma vectorial*[1]*,* $\|\cdot\|$*, tal que*

$$\rho(\mathbf{A}) \leq \|\mathbf{A}\| \leq \rho(\mathbf{A}) + \varepsilon.$$

La demostración se puede encontrar en [1] (teorema 1.3) o en [2] (teoremas 6.9.1 y 6.9.2).

Teorema 2. *Un esquema iterativo definido por* (1) *es convergente si y sólo si* $\rho(\mathbf{G}) < 1$*.*

Demostración. Para que un método sea convergente se debe cumplir que $\lim_{k\to\infty} \mathbf{e}^k = \mathbf{0}$ para cualquier $\mathbf{e}^0$. En particular, considerando $\mathbf{e}^0$ un vector propio de la matriz $\mathbf{G}$ con valor propio λ, $\mathbf{e}^k = \lambda^k \mathbf{e}^0$ y, por consiguiente, $\lim_{k\to\infty} \mathbf{e}^k = \mathbf{0}$ si $|\lambda| < 1$. Por lo tanto,

$$\lim_{k\to\infty} \mathbf{e}^k = \mathbf{0} \; \forall \; \mathbf{e}^0 \quad \Rightarrow \quad \rho(\mathbf{G}) < 1.$$

Para ver la implicación inversa, es suficiente comprobar que $\rho(\mathbf{G}) < 1 \Rightarrow \lim_{k\to\infty} \|[\mathbf{G}]^k\| = \mathbf{0}$. Para ello se considera el teorema 1. Si $\rho(\mathbf{G}) < 1$, por el teorema anterior considerando $\varepsilon > 0$ tal que $\rho(\mathbf{G}) + \varepsilon < 1$, existe una norma matricial tal que $\|\mathbf{G}\| \leq \rho(\mathbf{G}) + \varepsilon = \theta < 1$. Utilizando la desigualdad de Schwarz[2], $\|[\mathbf{G}]^k\| \leq \|\mathbf{G}\|^k \leq \theta^k$ y, por lo tanto, se concluye que

$$\rho(\mathbf{G}) < 1 \quad \Rightarrow \quad \lim_{k\to\infty} \|[\mathbf{G}]^k\| = \mathbf{0} \quad \Rightarrow \quad \lim_{k\to\infty} \mathbf{e}^k = \mathbf{0} \; \forall \; \mathbf{e}^0.$$

□

Notar que para cualquier norma matricial la desigualdad $\rho(\mathbf{G}) \leq \|\mathbf{G}\|$ es siempre cierta. En efecto, cualquier vector propio $\mathbf{u}_i$ de $\mathbf{G}$, de valor propio λ_i, verifica

$$|\lambda_i| \|\mathbf{u}_i\| = \|\lambda_i \mathbf{u}_i\| = \|\mathbf{G}\mathbf{u}_i\| \leq \|\mathbf{G}\| \|\mathbf{u}_i\|.$$

Por lo tanto, $|\lambda_i| \leq \|\mathbf{G}\|$ para cualquier valor propio λ_i. Es decir,

$$\rho(\mathbf{G}) \leq \|\mathbf{G}\|,$$

para cualquier norma matricial y cualquier matriz $\mathbf{G}$. Esto da lugar al siguiente corolario del teorema 2.

[1] Dada una norma vectorial $\|\cdot\|$ la norma matricial asociada es $\|\mathbf{A}\| = \max_{\|x\|=1} \|\mathbf{A}\mathbf{x}\|$. El hecho de tener una norma matricial asociada a una norma vectorial permite utilizar la desigualdad $\|\mathbf{A}\mathbf{x}\| \leq \|\mathbf{A}\| \|\mathbf{x}\|$.

[2] La desigualdad de Schwarz asegura que cualquier norma matricial cumple $\|\mathbf{A} \cdot \mathbf{B}\| \leq \|\mathbf{A}\| \|\mathbf{B}\|$ para matrices cualesquiera $\mathbf{A}$ y $\mathbf{B}$.

Corolario 1. *Una condición suficiente de convergencia para el esquema* (1) *es* $\|\mathbf{G}\| < 1$.

En particular este resultado es cierto para las normas matriciales

$$\|\mathbf{G}\|_\infty = \max_i \sum_{j=1}^{n} |g_{ij}| \quad \text{y} \quad \|\mathbf{G}\|_1 = \max_j \sum_{i=1}^{n} |g_{ij}| \tag{7}$$

puesto que se trata de normas matriciales asociadas a normas vectoriales. A diferencia del cálculo del radio espectral, $\rho(\mathbf{G})$, la evaluación de estas normas es sencilla y, por lo tanto, tienen un gran interés a la hora de establecer condiciones suficientes de convergencia.

1.2. Velocidad de convergencia

En el diseño de un método iterativo, además de asegurar la consistencia y convergencia del método, es importante el concepto de velocidad de convergencia.

Supongamos que se dispone de una aproximación de la solución, $\mathbf{x}^k$, con q cifras significativas, $\|\mathbf{e}^k\| \leq \frac{1}{2}10^{-q}\|\mathbf{x}^*\|$, y se desea saber cuántas iteraciones más se tienen que hacer para obtener m cifras correctas más. Interesa, por lo tanto, saber cuántas iteraciones ν hay que hacer para que, $\|\mathbf{e}^{k+\nu}\| \leq \frac{1}{2}10^{-(q+m)}\|\mathbf{x}^*\|$, es decir

$$\|\mathbf{e}^{k+\nu}\| \leq 10^{-m}\|\mathbf{e}^k\|.$$

Esto nos dará información sobre la velocidad con que la sucesión de aproximaciones $\mathbf{x}^k$ se acerca a la solución. Aplicando recursivamente la igualdad (6) se obtiene la relación $\mathbf{e}^{k+\nu} = [\mathbf{G}]^{(\nu)}\mathbf{e}^k$, que tomando normas es

$$\|\mathbf{e}^{k+\nu}\| \leq \|[\mathbf{G}]^\nu\|\|\mathbf{e}^k\|.$$

De manera que para conseguir las m cifras significativas es suficiente exigir que $\|[\mathbf{G}]^\nu\| \leq 10^{-m}$ o, equivalentemente

$$-\log_{10}\|[\mathbf{G}]^\nu\| \geq m.$$

Así, para conseguir m cifras significativas es suficiente hacer ν iteraciones con

$$\nu \geq \frac{m}{R_\nu},$$

donde se define

$$R_\nu := -\frac{1}{\nu}\log_{10}\|[\mathbf{G}]^\nu\| = -\log_{10}\left(\|[\mathbf{G}]^\nu\|^{\frac{1}{\nu}}\right).$$

El parámetro R_ν se puede interpretar como una velocidad de convergencia: cuanto mayor es R_ν menos iteraciones hay que hacer para incrementar la precisión en la aproximación. A la práctica se trabaja con la velocidad asintótica de convergencia que corresponde al límite cuando $\nu \to \infty$. Se puede comprobar que $\lim_{k\to\infty} \|[\mathbf{G}]^\nu\|^{\frac{1}{\nu}} = \rho(\mathbf{G})$ y, por lo tanto, la **velocidad asintótica de convergencia** del método es

$$R_\infty = \log_{10}\left[\frac{1}{\rho(\mathbf{G})}\right].$$

Para las matrices usuales en problemas de ingeniería (en la resolución con diferencias finitas o elementos finitos) el límite con $\nu \to \infty$ converge rápidamente y es una buena aproximación considerar por ejemplo $R_\infty \simeq R_{10}$. El radio espectral hace el papel del factor asintótico de convergencia del método. Para $\rho(\mathbf{G}) \simeq 0$, la velocidad de convergencia R_∞ es muy grande y el método tiene una convergencia muy rápida. Para $\rho(\mathbf{G}) = 1 - \varepsilon$, con $\varepsilon \simeq 0$ positivo, la velocidad de convergencia es positiva pero pequeña y el método converge per lo hace lentamente. En cambio, si $\rho(\mathbf{G}) > 1$ la velocidad de convergencia es negativa y, tal como se comenta en la sección 1.1, la convergencia no está asegurada.

1.3. Descripción de los métodos

Vistas las condiciones por las que un método estacionario es consistente y convergente, estas se pueden particularizar para los esquemas según (4), es decir,

$$\mathbf{x}^{k+1} = \mathbf{x}^k - \mathbf{C}^{-1}(\mathbf{A}\mathbf{x}^k - \mathbf{b}).$$

Estos esquemas son consistentes por definición, y serán convergentes si y sólo si $\rho(\mathbf{G}) = \rho(\mathbf{I} - \mathbf{C}^{-1}\mathbf{A}) < 1$. A continuación se presentan algunos métodos clásicos que corresponden a diferentes elecciones de la matriz $\mathbf{C}$. Para ello en lo sucesivo se considera la descomposición aditiva de la matriz $\mathbf{A}$ como

$$\mathbf{A} = \mathbf{L}_A + \mathbf{D}_A + \mathbf{U}_A$$

donde $\mathbf{L}_A$ es la parte triangular inferior de la matriz sin incluir la diagonal, $\mathbf{D}_A$ es la diagonal de $\mathbf{A}$, y $\mathbf{U}_A$ es la parte triangular superior sin incluir la diagonal (ver figura 1).

1.3.1. Método de Richardson: $\mathbf{C} = \mathbf{I}$

El método de Richardson corresponde a considerar la opción más barata computacionalmente, es decir $\mathbf{C} = \mathbf{I}$, obteniendo el esquema

$$\mathbf{x}^{k+1} = \mathbf{x}^k - \left[\mathbf{A}\mathbf{x}^k - \mathbf{b}\right],$$

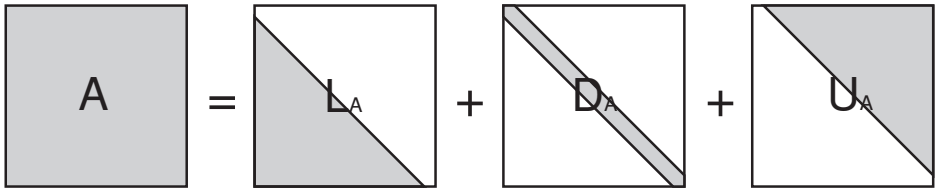

Figura 1: Descomposición aditiva de la matriz

o, componente a componente,

$$\boxed{x_i^{k+1} = x_i^k - \sum_{j=1}^{n} a_{ij}\, x_j^k + b_i\ , \quad i = 1...n, \quad k = 0, 1, ...}$$

El inconveniente del método es que la condición para convergencia es muy restrictiva: el método es convergente sólo si $\rho(\mathbf{I}-\mathbf{A}) < 1$. Es decir, para tener una convergencia rápida es necesario que todos los valores propios de $\mathbf{A}$ sean cercanos a la unidad[3], que evidentemente es una condición que no se suele verificar.

1.3.2. Método de Jacobi: $\mathbf{C} = \mathbf{D}_A$

El método de Jacobi se escribe como

$$\boxed{x_i^{k+1} = \frac{1}{a_{ii}} \left[b_i - \sum_{j=1}^{i-1} a_{ij} x_j^k - \sum_{j=i+1}^{n} a_{ij} x_j^k \right], \quad i = 1...n, \quad k = 0, 1, ...}$$

La expresión del algoritmo se puede interpretar como la aproximación que se obtiene al despejar la incógnita x_i de la ecuación i-ésima del sistema. Sin embargo, este método se puede clasificar dentro de los métodos según (4) considerando $\mathbf{C}$ como la diagonal de $\mathbf{A}$, $\mathbf{C} = \mathbf{D}_A$. En efecto el esquema para $\mathbf{C} = \mathbf{D}_A$,

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

se puede reescribir como

$$\mathbf{D}_A\mathbf{x}^{k+1} = \mathbf{D}_A\mathbf{x}^k + \mathbf{b} - \mathbf{A}\mathbf{x}^k.$$

Utilizando la descomposición aditiva de $\mathbf{A}$ y simplificando se llega a la expresión

$$\mathbf{D}_A\mathbf{x}^{k+1} = \mathbf{b} - \mathbf{L}_A\mathbf{x}^k - \mathbf{U}_A\mathbf{x}^k,$$

que corresponde a la forma matricial del algoritmo.

[3]Si A tiene valores propios $\{\lambda_i\}_{i=1,...,n}$, entonces $\mathbf{A} - p\mathbf{I}$ tiene valores propios $\{\lambda_i - p\}_{i=1,...,n}$ con los mismos vectores propios (traslación).

Observación 1. *El método de Jacobi no se puede utilizar si la matriz tiene algún coeficiente nulo en la diagonal. Es decir, para poder usar el método es necesario que* $a_{ii} \neq 0 \ \forall \ i$.

Observación 2. *Una condición suficiente para convergencia del método de Jacobi es que* $\mathbf{A}$ *sea diagonalmente dominante*[4]*. En efecto, la condición suficiente de convergencia del corolario 1 para el método de Jacobi,* $\mathbf{G} = \mathbf{I} - \mathbf{C}^{-1}\mathbf{A}$*, con las normas matriciales* (7) *se reescribe como,*

$$\|\mathbf{G}\|_\infty = \max_i \sum_{j \neq i} |\frac{a_{ij}}{a_{ii}}| < 1 \quad o \quad \|\mathbf{G}\|_1 = \max_j \sum_{i \neq j} |\frac{a_{ij}}{a_{ii}}| < 1,$$

respectivamente. La primera es trivial de verificar si la matriz es diagonalmente dominante por filas y la segunda si lo es por columnas.

Observación 3. *La velocidad asintótica de convergencia se puede estimar recordando que* $\rho(\mathbf{G}) \leq \|\mathbf{G}\|$ *y, por consiguiente*

$$\mathbf{R}_\infty = \log_{10}\left(\frac{1}{\rho(\mathbf{G})}\right) \geq \log_{10}\left(\frac{1}{\min(\|\mathbf{G}\|_1, \|\mathbf{G}\|_\infty)}\right).$$

Observación 4. *Para la implementación del método es necesario dimensionar dos vectores para almacenar la aproximación actual,* $\mathbf{x}^k$*, y la aproximación nueva,* $\mathbf{x}^{k+1}$*, en cada iteración.*

1.3.3. Método de Gauss-Seidel: $\mathbf{C} = \mathbf{D}_A + \mathbf{L}_A$

En el método de Jacobi cada componente x_i^{k+1} se calcula utilizando las componentes de la aproximación anterior, $\mathbf{x}^k$. Sin embargo, a la hora de calcular la componente x_i^{k+1} ya se han calculado previamente las componentes anteriores, $x_1^{k+1},\ x_2^{k+1}, \ldots,\ x_{i-1}^{k+1}$. En el método de Gauss-Seidel se aprovechan estas $i-1$ componentes para el cálculo de x_i^{k+1}. El esquema resultante es

$$\boxed{x_i^{k+1} = \frac{1}{a_{ii}}\left[b_i - \sum_{j=1}^{i-1} a_{ij}\ x_j^{\mathbf{k+1}} - \sum_{j=i+1}^{n} a_{ij}\ x_j^k\right],\ i = 1...n,\ k = 0, 1, ...} \quad (8)$$

donde se ha resaltado el superíndice $k+1$ para destacar la diferencia con el método de Jacobi. En cada paso se utiliza la información más actual de que se dispone. Este método se clasifica en los métodos según (4) considerando $\mathbf{C} = \mathbf{D}_A + \mathbf{L}_A$. En efecto, la particularización de (4) para $\mathbf{C} = \mathbf{D}_A + \mathbf{L}_A$ es

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

[4]Una matriz $\mathbf{A}$ es diagonalmente dominante si verifica $|a_{ii}| > \sum_{j \neq i} |a_{ij}| \ \forall \ i$

Premultiplicando esta ecuación por la matriz $(\mathbf{D}_A + \mathbf{L}_A)$ se obtiene

$$(\mathbf{D}_A + \mathbf{L}_A)\mathbf{x}^{k+1} = (\mathbf{D}_A + \mathbf{L}_A)\mathbf{x}^k + (\mathbf{b} - \mathbf{A}\mathbf{x}^k),$$

que, teniendo en cuenta la descomposición aditiva de A, es

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

Esta ecuación finalmente se escribe como

$$\mathbf{D}_A\mathbf{x}^{k+1} = \mathbf{b} - \mathbf{L}_A\mathbf{x}^{k+1} - \mathbf{U}_A\mathbf{x}^k,$$

que es la versión matricial del algoritmo.

Observación 5. *El método de Gauss-Seidel se puede implementar dimensionando un sólo vector para almacenar las componentes que interesan de* $\mathbf{x}^k$ *y* $\mathbf{x}^{k+1}$*. El mismo vector se va actualizando en cada iteración.*

Observación 6. *Se puede comprobar que el método es convergente para matrices diagonalmente dominantes, y para matrices simétricas y definidas positivas* [5] *(ver [1] y [2]).*

Observación 7. *Generalmente, si ambos convergen, la velocidad de convergencia del método de Gauss-Seidel es mayor que la del método de Jacobi. De hecho, para matrices diagonalmente dominantes* $\|\mathbf{G}_{GS}\|_\infty \leq \|\mathbf{G}_J\|_\infty$*, aunque eso no siempre implica que* $\rho(\mathbf{G}_{GS}) \leq \rho(\mathbf{G}_J)$*.*

Observación 8. *A diferencia del método de Jacobi, en el método de Gauss-Seidel las aproximaciones* $\mathbf{x}^k$ *dependen de la numeración de las incógnitas. Para reducir esta dependencia de la numeración es habitual utilizar una versión simetrizada del método en la que se alterna una iteración con aproximación estándar* (8) *y una iteración donde el cálculo de las componentes se hace en sentido inverso,*

$$x_i^{k+1} = \frac{1}{a_{ii}}\left[b_i - \sum_{j=1}^{i-1} a_{ij}x_j^k - \sum_{j=i+1}^{n} a_{ij}x_j^{\mathbf{k+1}}\right], \quad i = n, \ldots, 1,$$

es decir,

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

[5] Una matriz $\mathbf{A}$ es definida positiva si $\mathbf{x}^T\mathbf{A}\mathbf{x} > 0 \; \forall \; \mathbf{x} \neq \mathbf{0}$. Otras condiciones equivalentes para la definición positiva de $\mathbf{A}$ son (1) todos los valores propios de $\mathbf{A}$ son positivos o (2) todos los menores principales de $\mathbf{A}$ tienen determinante positivo.

Observación 9. *Considerando una descomposición por bloques de la matriz* $\mathbf{A}$*, del vector* $\mathbf{b}$ *y de la solución* $\mathbf{x}$*,*

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} & \dots & \mathbf{A}_{1s} \\ \mathbf{A}_{21} & \mathbf{A}_{22} & \dots & \mathbf{A}_{2s} \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{A}_{s1} & \mathbf{A}_{s42} & \dots & \mathbf{A}_{ss} \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \\ \vdots \\ \mathbf{b}_s \end{pmatrix}, \quad \mathbf{x} = \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_s \end{pmatrix},$$

se puede plantear el llamado método de Gauss-Seidel por bloques,

$$\mathbf{A}_{ii}\mathbf{x}_i^{k+1} = \mathbf{b}_i - \sum_{j=1}^{i-1} \mathbf{A}_{ij}\, \mathbf{x}_j^{\mathbf{k+1}} - \sum_{j=i+1}^{s} \mathbf{A}_{ij}\, \mathbf{x}_j^{k},\ i = 1...s,\ k = 0, 1, \dots,$$

cuya implementación sólo requiere resolver sistemas con las matrices $\mathbf{A}_{ii}$ *(de dimensión pequeña).*

1.3.4. Método de sobrerelajaciones sucesivas (SOR)

Existen infinidad de métodos para *acelerar* la convergencia, *extrapolar* o *relajar* los distintos métodos iterativos. Todos ellos se basan en introducir como mínimo un parámetro ω en la definición de $\mathbf{G}$, de manera que se puede considerar un valor óptimo del parámetro

$$\omega_{opt} = \arg\min_{\omega} \rho(\mathbf{G}(\omega)),$$

y así acelerar la convergencia. De estos métodos el más conocido es el método de sobrerelajaciones sucesivas (SOR). La forma general se puede escribir como,

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

Se deja como ejercicio para el lector comprobar que este método corresponde a considerar $\mathbf{C} = \frac{1}{\omega}\left(\omega\mathbf{L}_A + \mathbf{D}_A\right)$. El parámetro ω se puede elegir de manera que el radio espectral sea lo más pequeño posible y así obtener una convergencia rápida. Este método es popular sobretodo en el contexto de las diferencias finitas, en el que existen varios resultados que permiten determinar el parámetro óptimo de forma analítica.

2. Métodos de los gradientes conjugados

El método de los gradientes conjugados es un método iterativo no estacionario. Es decir, es un método cuya definición depende de la iteración k.

De hecho, es uno de los llamados métodos de Krylov [3]. Estos métodos son métodos que en cada iteración k calculan una aproximación $\mathbf{x}^k$ que minimiza una cierta medida del error en el espacio $\mathbf{x}^0 + \mathcal{K}_k$, donde $\mathcal{K}_k \subset \mathbb{R}^n$ es el llamado k-esimo espacio de Krylov. Si no se llega a convergencia antes de la iteración n, el n-ésimo espacio de Krylov es $\mathcal{K}_n = \mathbb{R}^n$ y la minimización sobre $\mathcal{K}_n$ proporciona la solución $\mathbf{x}^*$. Los métodos de Krylov se pueden considerar métodos directos, ya que encuentran la solución $\mathbf{x}^*$ del sistema lineal como mucho en n iteraciones. Sin embargo, para sistemas lineales con un gran número de ecuaciones, se utilizan como métodos iterativos.

El más popular de los métodos de Krylov es el método de los gradientes conjugados. Aunque se puede hacer una deducción del método de los gradientes conjugados como un método de Krylov, aquí se va a hacer una deducción más sencilla motivada por el método del máximo descenso. El método de los gradientes conjugados se basa en la minimización de una cierta función ϕ. En la siguiente sección se comprueba la equivalencia del problema $\mathbf{Ax} = \mathbf{b}$ con la minimización de la función ϕ. A continuación, se presenta el método del máximo descenso. Finalmente, el método de los gradientes conjugados se plantea como una mejora de este método.

Observación 10. *Para poder hacer la analogía con el problema de minimización, el método de los gradientes conjugados requiere que la matriz sea simétrica y definida positiva. Sin embargo, existen métodos de Krylov que no tienen estas restricciones. El método GMRES (Generalized Minimum Residual) sólo necesita que la matriz sea regular. A cambio, tiene un coste computacional y unos requerimientos de memoria considerablemente mayores que el método de los gradientes conjugados.*

2.1. Problema de minimización equivalente

Notación: Para simplificar las expresiones en lo sucesivo se denota el producto escalar de vectores como

$$\langle \mathbf{u}, \mathbf{v} \rangle = \mathbf{u}^T \mathbf{v}.$$

Proposición 1. *Sea* $\mathbf{A}$ *es simétrica y definida positiva (SDP). Entonces,* $\mathbf{x}^*$ *es la solución de* $\mathbf{Ax} = \mathbf{b}$ *si y sólo si es el mínimo del funcional cuadrático*

$$\phi(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T\mathbf{Ax} - \mathbf{x}^T\mathbf{b} = \frac{1}{2}\langle \mathbf{x}, \mathbf{Ax} \rangle - \langle \mathbf{x}, \mathbf{b} \rangle.$$

Es decir, $\mathbf{x}^* = \mathbf{A}^{-1}\mathbf{b} \quad \Leftrightarrow \quad \mathbf{x}^* = \arg\min_{\mathbf{x}\in\mathbb{R}^n} \phi(\mathbf{x})$.

Demostración. La demostración comprueba las dos implicaciones:

- $\mathbf{x}^* = \mathbf{A}^{-1}\mathbf{b} \quad \Rightarrow \quad \mathbf{x}^* = \arg\min_{\mathbf{x}\in\mathbb{R}^n} \phi(\mathbf{x})$:

  Supongamos que $\mathbf{x}^*$ es solución: $\mathbf{A}\mathbf{x}^* = \mathbf{b}$. Cualquier $\mathbf{x} \in \mathbb{R}$ puede expresarse como $\mathbf{x} = \mathbf{z} + \mathbf{x}^*$ y, por consiguiente,

  $$\phi(\mathbf{x}) = \phi(\mathbf{z}+\mathbf{x}^*) = \frac{1}{2}\left[\langle \mathbf{z}, \mathbf{A}\mathbf{z}\rangle + 2\langle \mathbf{z}, \mathbf{A}\mathbf{x}\rangle + \langle \mathbf{x}^*, \mathbf{A}\mathbf{x}^*\rangle\right] - \langle \mathbf{z}, \mathbf{b}\rangle - \langle \mathbf{x}^*, \mathbf{b}\rangle,$$

  donde se ha tenido en cuenta la simetría de la matriz $\mathbf{A}$. Empleando ahora la hipótesis de que $\mathbf{x}^*$ es solución del sistema se obtiene

  $$\phi(\mathbf{x}) = \phi(\mathbf{x}^*) + \frac{1}{2}\langle \mathbf{z}, \mathbf{A}\mathbf{z}\rangle + \langle \mathbf{z}, \mathbf{A}\mathbf{x}^* - \mathbf{b}\rangle = \phi(\mathbf{x}^*) + \frac{1}{2}\langle \mathbf{z}, \mathbf{A}\mathbf{z}\rangle.$$

  Puesto que $\mathbf{A}$ es una matriz definida positiva, $\langle \mathbf{z}, \mathbf{A}\mathbf{z}\rangle \geq 0$, se deduce que

  $$\phi(\mathbf{x}) \leq \phi(\mathbf{x}^*) \qquad \forall\ \mathbf{x}.$$

  Además, sólo habrá una igualdad si $\mathbf{z} = \mathbf{0}$, es decir, $\mathbf{x} = \mathbf{x}^*$.

- $\mathbf{x}^* = \mathbf{A}^{-1}\mathbf{b} \quad \Leftarrow \quad \mathbf{x}^* = \arg\min_{\mathbf{x}\in\mathbb{R}^n} \phi(\mathbf{x})$:

  Si $\mathbf{x}^*$ es un mínimo de la función ϕ, debe verificar la condición necesaria de extremo relativo $\nabla\phi(\mathbf{x}) = \mathbf{0}$. Dado que $\nabla\phi(\mathbf{x}) = \mathbf{r}(\mathbf{x}) = \mathbf{A}\mathbf{x} - \mathbf{b}$, la condición de extremo es

  $$\mathbf{A}\mathbf{x} - \mathbf{b} = \mathbf{0}.$$

□

2.2. Motivación: método del máximo descenso, del gradiente o de máxima pendiente

El método del máximo descenso se basa en el hecho de que el gradiente de una función ϕ en un punto $\mathbf{x}^k$ es la dirección en la que esta función ϕ crece más rápidamente. Por esto, dado $\mathbf{x}^k$, la nueva aproximación $\mathbf{x}^{k+1}$ se busca avanzando en la dirección de $-\nabla\phi = -\mathbf{r}^k$,

$$\boxed{\mathbf{x}^{k+1} = \mathbf{x}^k + \alpha_k \mathbf{r}^k}.$$

El parámetro α_k se elige de forma que se minimice ϕ en esta dirección. Se trata de un problema de minimización unidimensional de $\phi(\mathbf{x}^{k+1}) = \phi(\mathbf{x}^k + \alpha_k\mathbf{r}^k)$ en la dirección de $\mathbf{r}^k$, es decir,

$$\alpha_k = \arg\min_{\alpha\in\mathbb{R}} \phi(\mathbf{x}^k + \alpha\mathbf{r}^k). \tag{9}$$

Imponiendo la condición necesaria de extremo, $\frac{d\phi}{d\alpha}|_{\alpha=\alpha_k} = 0$[6], se obtiene la ecuación para α_k,

$$0 = \langle \mathbf{r}^k, \nabla\phi(\mathbf{x}^k + \alpha_k \mathbf{r}^k)\rangle = \langle \mathbf{r}^k, \mathbf{A}(\mathbf{x}^k + \alpha_k \mathbf{r}^k) - \mathbf{b}\rangle = \langle \mathbf{r}^k, \mathbf{r}^k + \alpha_k \mathbf{A}\mathbf{r}^k\rangle,$$

de donde se despeja la expresión

$$\boxed{\alpha_k = -\frac{\langle \mathbf{r}^k, \mathbf{r}^k\rangle}{\langle \mathbf{r}^k, \mathbf{A}\mathbf{r}^k\rangle}}.$$

Observación 11. *Si no se ha llegado a convergencia, debido a la definición positiva de la matriz* $\mathbf{A}$ *el parámetro* α_k *es siempre negativo. Por lo tanto, siempre se avanza en sentido contrario al gradiente.*

Observación 12. *Teniendo en cuenta que* $\nabla\phi(\mathbf{x}^k + \alpha_k \mathbf{r}^k) = \nabla\phi(\mathbf{x}^{k+1}) = \mathbf{r}^{k+1}$*, la ecuación* (9) *está exigiendo que*

$$\langle \mathbf{r}^k, \mathbf{r}^{k+1}\rangle = 0.$$

Es decir, la dirección de avance en cada iteración es perpendicular a la dirección de avance anterior. Esto provoca que generalmente se repitan direcciones de avance. Por ejemplo, en 2D el método tiene sólo 2 direcciones de avance, $\mathbf{r}^0$ *y su perpendicular, que se alternan en cada iteración. Esto provoca un fenómeno de "zig-zag"que hace que la convergencia del método pueda ser extraordinariamente lenta (ver ejemplo en la sección 2.3.3).*

Observación 13. *Se deja como ejercicio verificar que considerando* $\alpha_k = -1$ *constante se obtiene el método de Richardson. Así, el método de máximo descenso se puede interpretar como una aceleración o extrapolación del método de Richardson.*

2.3. Método de los gradientes conjugados

Con la misma idea que en el método del máximo descenso, en cada iteración del método de los gradientes conjugados se busca una aproximación de la forma

$$\boxed{\mathbf{x}^{k+1} = \mathbf{x}^k + \alpha_k \mathbf{p}^k}$$

que minimice ϕ. De nuevo el parámetro escalar α_k se escoge para minimizar ϕ en la dirección de avance, en este caso $\mathbf{p}^k$. Tal como se vio en el apartado anterior, la minimización unidireccional proporciona la expresión para α_k,

$$\boxed{\alpha_k = -\frac{\langle \mathbf{p}^k, \mathbf{r}^k\rangle}{\langle \mathbf{p}^k, \mathbf{A}\mathbf{p}^k\rangle}}.$$

[6]La derivada de una función ϕ en una dirección $\boldsymbol{\xi}$ se puede calcular aplicando la regla de la cadena como $\frac{d\phi(\mathbf{x}+\alpha\boldsymbol{\xi})}{d\alpha} = [\nabla\phi(\mathbf{x} + \alpha\boldsymbol{\xi})]^T\boldsymbol{\xi}$.

La gran y única diferencia con el método del máximo descenso es la definición de las direcciones de avance. En el método de los gradientes conjugados las direcciones $\mathbf{p}^k$ se escogen de manera que sean $\mathbf{A}$-conjugadas, es decir, $(\mathbf{p}^i)^T\mathbf{A}\mathbf{p}^j = \langle \mathbf{p}^i, \mathbf{A}\mathbf{p}^j\rangle = \delta_{ij}$. De esta forma se evitan las repeticiones que tanto relentizan el método del máximo descenso.

En $\mathbb{R}^n$ existen n direcciones $\mathbf{A}$-conjugadas distintas, $\{\mathbf{p}^0, \mathbf{p}^1, \ldots, \mathbf{p}^{k-1}\}$. Para obtenerlas se podría plantear una ortogonalización de Gram-Schmidt con el producto escalar definido por $\mathbf{A}$,

$$\mathbf{p}^k = \begin{cases} \mathbf{r}^0 & k = 0 \\ \mathbf{r}^k + \displaystyle\sum_{s=0}^{k-1} \varepsilon_{ks}\mathbf{p}^s & k > 0 \end{cases},$$

donde los escalares ε_{ks} se determinan al imponer $\langle \mathbf{p}^k, \mathbf{p}^s\rangle = 0$ para $s = 0, \ldots, k-1$. Esta opción tiene un coste computacional importante que, por suerte, se puede evitar, ya que para matrices simétricas y definidas es suficiente exigir que

$$\boxed{\mathbf{p}^k = \begin{cases} \mathbf{r}^0 & k = 0 \\ \mathbf{r}^k + \beta_k\mathbf{p}^{k-1} & k > 0 \end{cases}}, \tag{10}$$

con $\langle \mathbf{p}^k, \mathbf{A}\mathbf{p}^{k-1}\rangle = 0$ para obtener, como se comenta más adelante, $\langle \mathbf{p}^k, \mathbf{A}\mathbf{p}^s\rangle = 0$ para $s = 0, \ldots, k-1$.

El escalar β_k se determina al exigir que $\langle \mathbf{p}^k, \mathbf{A}\mathbf{p}^{k-1}\rangle = 0$,

$$\mathbf{0} = \langle \mathbf{r}^k + \beta_k\mathbf{p}^{k-1}, \mathbf{A}\mathbf{p}^{k-1}\rangle = \langle \mathbf{r}^k, \mathbf{A}\mathbf{p}^{k-1}\rangle + \beta_k\langle \mathbf{p}^{k-1}, \mathbf{A}\mathbf{p}^{k-1}\rangle,$$

es decir,

$$\boxed{\beta_k = -\frac{\langle \mathbf{r}^k, \mathbf{A}\mathbf{p}^{k-1}\rangle}{\langle \mathbf{p}^{k-1}, \mathbf{A}\mathbf{p}^{k-1}\rangle}}. \tag{11}$$

Así, el método de los gradientes conjugados, mediante las ecuaciones (10) y (11), proporciona una forma sencilla y eficiente de calcular las direcciones $\mathbf{A}$-conjugadas de avance. A continuación se presenta el algoritmo resultante.

2.3.1. Algoritmo de gradientes conjugados (versión 1):

$\mathbf{x}^0$: aproximación inicial

$k = 0$, $\mathbf{r}^0 = \mathbf{A}\mathbf{x}^0 - b$, $\mathbf{p}^0 = \mathbf{r}^0$

do while (no convergencia)

$\mathbf{q}^k = \mathbf{A}\mathbf{p}^k$

$\alpha_k = -\frac{\langle \mathbf{p}^k, \mathbf{r}^k \rangle}{\langle \mathbf{p}^k, \mathbf{q}^k \rangle}$

$\mathbf{x}^{k+1} = \mathbf{x}^k + \alpha_k \mathbf{p}^k$

$\mathbf{r}^{k+1} = \mathbf{A}\mathbf{x}^{k+1} - \mathbf{b}$

$\beta_{k+1} = -\frac{\langle \mathbf{r}^{k+1}, \mathbf{q}^k \rangle}{\langle \mathbf{p}^k, \mathbf{q}^k \rangle}$

$\mathbf{p}^{k+1} = \mathbf{r}^{k+1} + \beta_{k+1} \mathbf{p}^k$

$k = k + 1$

enddo

El coste computacional por iteración de este algoritmo es el de 2 productos de matriz por vector, 2 cálculos de vector más escalar por vector, 3 productos escalares vector por vector y 1 suma de vectores. Obviamente, el coste computacional más importante es debido al producto matriz por vector.

Observación 14. *El coste computacional del método se puede reducir suprimiendo un cálculo de matriz por vector. En efecto, multiplicando por* $\mathbf{A}$ *y restando* $\mathbf{b}$ *a los dos lados de la ecuación* $\mathbf{x}^{k+1} = \mathbf{x}^k + \alpha_k \mathbf{p}^k$ *se obtiene la siguiente expresión recursiva para el cálculo del residuo*

$$\mathbf{r}^{k+1} = \mathbf{r}^k + \alpha_k \mathbf{q}^k, \tag{12}$$

con coste computacional considerablemente menor que $\mathbf{r}^{k+1} = \mathbf{A}\mathbf{x}^{k+1} - \mathbf{b}$. *Sin embargo, para matrices mal condicionadas en la práctica es recomendable calcular el residuo,* $\mathbf{r}^{k+1} = \mathbf{A}\mathbf{x}^{k+1} - \mathbf{b}$, *cada cierto número de iteraciones para corregir errores de redondeo acumulados por la fórmula recursiva.*

2.3.2. Propiedades del método de los gradientes conjugados

Proposición 2. *Las iteraciones del algoritmo de los gradientes conjugados verifican, para todo* $k \leq n$, *las siguientes propiedades:*

(a) $\langle \mathbf{r}^k, \mathbf{p}^i \rangle = 0$ *para* $i < k$,

(b) $\langle \mathbf{r}^i, \mathbf{r}^i \rangle = \langle \mathbf{r}^i, \mathbf{p}^i \rangle$ *para* $i \leq k$,

(c) $\langle \mathbf{r}^k, \mathbf{r}^i \rangle = 0$ *para* $i < k$ *(residuos ortogonales),*

(d) $\langle \mathbf{p}^k, \mathbf{A}\mathbf{p}^i \rangle = 0$ *para* $i < k$ *(direcciones A-conjugadas).*

Demostración. La demostración se hace por inducción sobre k.

Comprobamos primero que cada una de las propiedades es cierta para $k = 1$. Dado que $\mathbf{p}^0 = \mathbf{r}^0$ por definición, las propiedades (a) y (c) son equivalentes para $k = 1$. Así, teniendo en cuenta la ecuación (12), es decir $\mathbf{r}^1 = \mathbf{r}^0 + \alpha_0 \mathbf{A}\mathbf{p}^0$, se llega a

$$\langle \mathbf{r}^1, \mathbf{p}^0 \rangle = \langle \mathbf{r}^0 + \alpha_0 \mathbf{A}\mathbf{p}^0, \mathbf{p}^0 \rangle = \langle \mathbf{r}^0, \mathbf{p}^0 \rangle + \alpha_0 \langle \mathbf{A}\mathbf{p}^0, \mathbf{p}^0 \rangle = 0, \tag{13}$$

por definición de α_0, lo cual demuestra (a) y (c). La propiedad (b) se comprueba teniendo en cuenta (13) y la definición $\mathbf{p}^1 = \mathbf{r}^1 + \beta_1 \mathbf{p}^0$,

$$\langle \mathbf{r}^1, \mathbf{p}^1 \rangle = \langle \mathbf{r}^1, \mathbf{r}^1 \rangle + \beta_1 \underbrace{\langle \mathbf{r}^1, \mathbf{p}^0 \rangle}_{0} = \langle \mathbf{r}^1, \mathbf{r}^1 \rangle.$$

El caso $k = 1$ se completa con la comprobación de la propiedad (d), usando la definición de $\mathbf{p}^1$ y del parámetro β_1,

$$\langle \mathbf{p}^1, \mathbf{A}\mathbf{p}^0 \rangle = \langle \mathbf{r}^1 + \beta_1 \mathbf{p}^0, \mathbf{A}\mathbf{p}^0 \rangle = \langle \mathbf{r}^1, \mathbf{A}\mathbf{p}^0 \rangle + \beta_1 \langle \mathbf{p}^0, \mathbf{A}\mathbf{p}^0 \rangle = 0.$$

A continuación, suponiendo que las cuatro propiedades son ciertas para k (hipótesis de inducción) se comprueba que también son ciertas para $k + 1$:

(a) La primera propiedad se comprueba utilizando la ecuación (12), es decir $\mathbf{r}^{k+1} = \mathbf{r}^k + \alpha_k \mathbf{A}\mathbf{p}^k$, de donde se deduce

$$\langle \mathbf{r}^{k+1}, \mathbf{p}^i \rangle = \langle \mathbf{r}^k, \mathbf{p}^i \rangle + \alpha_k \langle \mathbf{p}^k, \mathbf{A}\mathbf{p}^i \rangle.$$

Así, para $i = k$ tenemos $\langle \mathbf{r}^{k+1}, \mathbf{p}^k \rangle = 0$ por definición de α_k. Para $k < i$, por hipótesis de inducción con las propiedades (a) y (d), se verifica que $\langle \mathbf{r}^k, \mathbf{p}^i \rangle = 0$ y $\langle \mathbf{p}^k, \mathbf{A}\mathbf{p}^i \rangle = 0$, de lo que se concluye que $\langle \mathbf{r}^{k+1}, \mathbf{p}^i \rangle = 0$.

(b) Para demostrar la segunda propiedad se utiliza la definición de $\mathbf{p}^{k+1}$ y la propiedad (a) para $k + 1$ e $i = k$ (acabada de demostrar),

$$\langle \mathbf{r}^{k+1}, \mathbf{p}^{k+1} \rangle = \langle \mathbf{r}^{k+1}, \mathbf{r}^{k+1} \rangle + \beta_{k+1} \underbrace{\langle \mathbf{r}^{k+1}, \mathbf{p}^k \rangle}_{0} = \langle \mathbf{r}^{k+1}, \mathbf{r}^{k+1} \rangle.$$

(c) La tercera propiedad se comprueba utilizando la defínción de $\mathbf{p}^i$, que se reescribe como

$$\mathbf{r}^i = \mathbf{p}^i - \beta_i \mathbf{p}^{i-1}$$

y la propiedad (a) ya demostrada,

$$\langle \mathbf{r}^{k+1}, \mathbf{r}^i \rangle = \underbrace{\langle \mathbf{r}^{k+1}, \mathbf{p}^i \rangle}_{0} - \beta_i \underbrace{\langle \mathbf{r}^{k+1}, \mathbf{p}^{i-1} \rangle}_{0} = 0.$$

(d) Para comprobar la última propiedad se utiliza la ecuación (12) para $k = i$, que se reescribe como

$$\mathbf{A}\mathbf{p}^i = \frac{1}{\alpha_i}(\mathbf{r}^{i+1} - \mathbf{r}^i),$$

y la definición de $\mathbf{p}^{k+1} = \mathbf{r}^{k+1} + \beta_{k+1}\mathbf{p}^k$,

$$\begin{aligned} \langle \mathbf{p}^{k+1}, \mathbf{A}\mathbf{p}^i \rangle &= \langle \mathbf{r}^{k+1}, \mathbf{A}\mathbf{p}^i \rangle + \beta_{k+1} \langle \mathbf{p}^k, \mathbf{A}\mathbf{p}^i \rangle \\ &= \frac{1}{\alpha_i} \langle \mathbf{r}^{k+1}, \mathbf{r}^{i+1} - \mathbf{r}^i \rangle + \beta_{k+1} \langle \mathbf{p}^k, \mathbf{A}\mathbf{p}^i \rangle. \end{aligned}$$

Para $i < k$ el primer término de abajo es cero por la propiedad (c) y el segundo término es nulo por hipótesis de inducción con la propiedad (d). Por lo tanto, $\langle \mathbf{p}^{k+1}, \mathbf{A}\mathbf{p}^i \rangle = 0$ para $i < k$. Para i=k, utilizando la definición de $\mathbf{p}^{k+1}$ y β_{k+1},

$$\begin{aligned} \langle \mathbf{p}^{k+1}, \mathbf{A}\mathbf{p}^k \rangle &= \langle \mathbf{r}^{k+1}, \mathbf{A}\mathbf{p}^k \rangle + \beta_{k+1} \langle \mathbf{p}^k, \mathbf{A}\mathbf{p}^k \rangle \\ &= \langle \mathbf{r}^{k+1}, \mathbf{A}\mathbf{p}^k \rangle - \frac{\langle \mathbf{r}^{k+1}, \mathbf{A}\mathbf{p}^k \rangle}{\langle \mathbf{p}^k, \mathbf{A}\mathbf{p}^k \rangle} \langle \mathbf{p}^k, \mathbf{A}\mathbf{p}^k \rangle = 0, \end{aligned}$$

tal como se quiere demostrar.

□

Corolario 2. *El método de los gradientes conjugados converge como máximo en n iteraciones.*

Demostración. La propiedad (a) de la proposición 2 particularizada para $k = n$ asegura que el residuo tras n iteraciones es ortogonal a las n direcciones de avance utilizadas,

$$\langle \mathbf{r}^n, \mathbf{p}^j \rangle = 0 \text{ para } j = 0, \ldots, n-1.$$

Dado que las direcciones son $\mathbf{A}$ conjugadas y, por lo tanto, linealmente independientes (si el residuo no se anula antes de la iteración n), esta condición es equivalente a decir que $\mathbf{r}^n$ es ortogonal a una base de $\mathbb{R}^n$, de lo que se concluye que $\mathbf{r}^n = \mathbf{0}$. □

De hecho, el método de GC se desarrolló como un método directo, pero se dejó de utilizar por su alto coste computacional (si se realizan las n iteraciones). Ahora se ha recuperado como método iterativo y se suele utilizar con matrices de dimensión grande pero casi vacías sin llegar generalmente a hacer las n iteraciones.

Proposición 3. *(Convergencia) El error en las iteraciones del método de los gradientes conjugados está acotado por la siguiente expresión*

$$\boxed{\|\mathbf{x}^i - \mathbf{x}^*\|_A \leq 2\left[\frac{\sqrt{k_2(\mathbf{A})}-1}{\sqrt{k_2(\mathbf{A})}+1}\right]^i \|\mathbf{x}^0 - \mathbf{x}^*\|_A}$$

donde $\|\mathbf{x}\|_A = \langle \mathbf{x}^T, \mathbf{A}\mathbf{x}\rangle^{1/2}$ *es la norma inducida por* $\mathbf{A}$ *(norma de la energía) y* $k_2(\mathbf{A})$ *es el número de condición de la matriz.*

La demostración de este resultado se basa en la interpretación del método como un método de Krylov. Se basa en la propiedad de que el método de los gradientes conjugados minimiza el error en la aproximación, con la norma inducida por $\mathbf{A}$, dentro de un subespacio de $\mathbb{R}^n$ cuya dimensión se incrementa en cada iteración.

Dado que la matriz es simétrica y definida positiva, el número de condición de $\mathbf{A}$ es el cociente entre sus valores propios mayor y menor en valor absoluto,

$$k_2(\mathbf{A}) = \frac{\lambda_{max}}{\lambda_{min}}.$$

Observación 15. *El término* $\frac{\sqrt{k_2(\mathbf{A})}-1}{\sqrt{k_2(\mathbf{A})}+1}$ *en la fórmula de la proposición 3 es siempre menor que* 1*. Esto asegura la convergencia monótona en la norma inducida por* $\mathbf{A}$*, es decir,*

$$\|\mathbf{x}^{i+1} - \mathbf{x}^*\|_A \leq \|\mathbf{x}^i - \mathbf{x}^*\|_A,$$

(ver representación gráfica en la figura 2). En general esto no implica la convergencia en norma $\|\cdot\|_2$[7] *o* $\|\cdot\|_\infty$*. Sin embargo, para el método de los gradientes conjugados se puede demostrar que la convergencia es monótona también en estas normas.*

Observación 16. *Para matrices bien condicionadas* $k_2(\mathbf{A}) = 1 + \varepsilon$*, con* $\varepsilon > 0$ *pequeño, y la ecuación de la proposición 3 se aproxima como*

$$\|\mathbf{x}^i - \mathbf{x}^*\|_A \leq 2\left(\frac{\varepsilon}{2}\right)^i \|\mathbf{x}^0 - \mathbf{x}^*\|_A,$$

[7] $\|\mathbf{x}\|_2 = \sqrt{\langle \mathbf{x}, \mathbf{x}\rangle}$

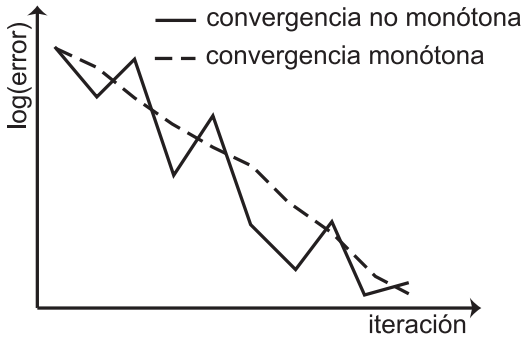

Figura 2: Representación del error para un método con convergencia monótona y un método con convergencia no monótona.

de donde se deduce que el error decrece rápidamente. Si la matriz es mal condicionada, $k_2(\mathbf{A}) \simeq \infty$*, entonces* $\dfrac{\sqrt{k_2(\mathbf{A})}-1}{\sqrt{k_2(\mathbf{A})}+1} = 1-\varepsilon$ *con* $\varepsilon > 0$ *pequeño. Entonces, la fórmula de la proposición 3 se escribe como*

$$\|\mathbf{x}^i - \mathbf{x}^*\|_A \leq 2\,(1-\varepsilon)^i\,\|\mathbf{x}^0 - \mathbf{x}^*\|_A,$$

de donde se deduce que la sucesión convergerá lentamente a la solución. En general, la velocidad de convergencia depende del número de condición de la matriz (ver figura 3). Cuanto menor sea el número de condición de la matriz más rápida será la convergencia del método de los gradientes conjugados. Para mejorar la convergencia se pueden usar precondicionadores, tal como se comenta en la sección 2.3.5.

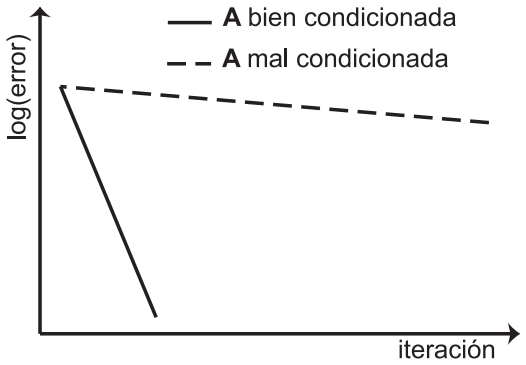

Figura 3: Ejemplo de convergencia del método de los gradientes conjugados para una matriz mal condicionada y una matriz con número de condición cercano a 1

2.3.3. Ejemplo de dimensión 2

Se plantea la resolución del sistema de ecuaciones de dimensión $n = 2$ $\mathbf{Ax} = \mathbf{b}$ con

$$\mathbf{A} = \begin{pmatrix} 3 & 2 \\ 2 & 6 \end{pmatrix}, \qquad \mathbf{b} = \begin{pmatrix} 2 \\ -8 \end{pmatrix}.$$

La figura 4 muestra la representación gráfica de la función $\phi(\mathbf{x}) = \frac{1}{2}\langle \mathbf{x}, \mathbf{Ax} \rangle -$

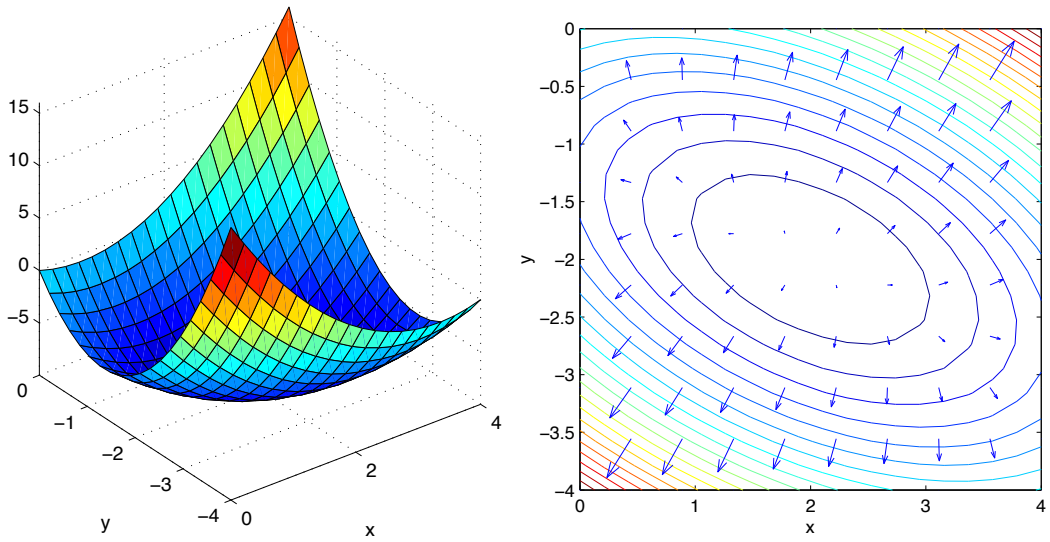

Figura 4: Función $\phi(\mathbf{x})$ (izquierda), curvas de nivel y gradientes (derecha)

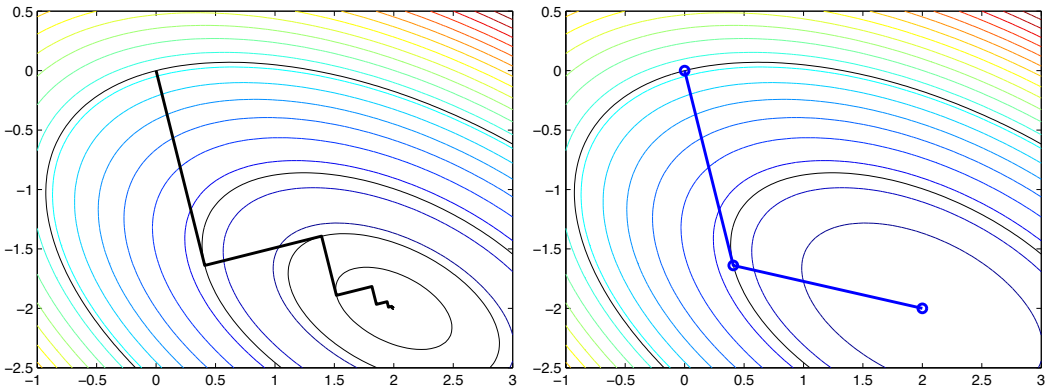

Figura 5: Curvas de nivel de $\phi(\mathbf{x})$ e iteraciones del método del máximo descenso (izquierda) y del método de los gradientes conjugados (derecha).

$\langle \mathbf{x}, \mathbf{b} \rangle$ para este ejemplo, así como sus curvas de nivel y el gradiente, $\nabla\phi(\mathbf{x}) =$

$\mathbf{Ax} - \mathbf{b}$. En la figura 5 se muestran las iteraciones del método del máximo descenso y del método de los gradientes conjugados con la misma aproximación inicial, $\mathbf{x}^0 = \mathbf{0}$. En el método del máximo descenso (izquierda) la dirección de avance es perpendicular a las curvas de nivel, además se observa como la dirección de avance en cada iteración es perpendicular a la dirección de avance anterior. Esto provoca un fenómeno de "zig-zag"que hace que la convergencia del método sea realmente lenta: son necesarias 25 iteraciones para conseguir un error relativo en el residuo menor que $0.5\ 10^{-6}$, es decir, $\|\mathbf{r}^k\| \leq 0.5\ 10^{-6}\|\mathbf{b}\|$. El método de los gradientes conjugados (derecha), como cabe esperar, converge en $n = 2$ iteraciones.

2.3.4. Algoritmo de gradientes conjugados (versión 2)

Utilizando las propiedades enunciadas en la proposición 2, se pueden simplificar algunas de las expresiones que aparecen en el algoritmo realizando los cálculos de forma más eficiente.

Utilizando la definición $\mathbf{p}^k = \mathbf{r}^k + \beta_k \mathbf{p}^{k-1}$ y una particularización de la propiedad (a) de la proposición 2, $\langle \mathbf{r}^k, \mathbf{p}^{k-1} \rangle = 0$, se obtiene una nueva expresión para α_k:

$$\alpha_k = -\frac{\langle \mathbf{p}^k, \mathbf{r}^k \rangle}{\langle \mathbf{p}^k, \mathbf{q}^k \rangle} = -\frac{\langle \mathbf{r}^k + \beta_k \mathbf{p}^{k-1}, \mathbf{r}^k \rangle}{\langle \mathbf{p}^k, \mathbf{q}^k \rangle} = -\frac{\langle \mathbf{r}^k, \mathbf{r}^k \rangle}{\langle \mathbf{p}^k, \mathbf{q}^k \rangle} \tag{14}$$

El cálculo de β_{k+1} se puede optimizar utilizando la fórmula recursiva (12), reescrita como $\mathbf{q}^k = \left(\mathbf{r}^{k+1} - \mathbf{r}^k\right)/\alpha_k$, la propiedad (c) de ortogonalidad de los residuos, $\langle \mathbf{r}^{k+1}, \mathbf{r}^k \rangle = 0$, y la expresión de α_k dada por la ecuación (14),

$$\begin{aligned} \beta_{k+1} &= -\frac{\langle \mathbf{r}^{k+1}, \mathbf{q}^k \rangle}{\langle \mathbf{p}^k, \mathbf{q}^k \rangle} = -\frac{\langle \mathbf{r}^{k+1}, \frac{\mathbf{r}^{k+1} - \mathbf{r}^k}{\alpha_k} \rangle}{\langle \mathbf{p}^k, \mathbf{q}^k \rangle} = -\frac{1}{\alpha_k} \frac{\langle \mathbf{r}^{k+1}, \mathbf{r}^{k+1} \rangle}{\langle \mathbf{p}^k, \mathbf{q}^k \rangle} \\ &= \frac{\langle \mathbf{r}^{k+1}, \mathbf{r}^{k+1} \rangle}{\langle \mathbf{r}^k, \mathbf{r}^k \rangle}. \end{aligned} \tag{15}$$

La nueva versión del algoritmo, más eficiente, se obtiene utilizando las expresiones dadas por (14) y (15) y la fórmula recursiva introducida en (12). El coste computacional de esta nueva versión es de 1 producto matriz por vector, 3 operaciones vector más escalar por vector y 2 productos escalares vector por vector.

$\mathbf{x}^0$: aproximación inicial

$k = 0$, $\mathbf{r}^0 = \mathbf{A}\mathbf{x}^0 - b$, $\mathbf{p}^0 = \mathbf{r}^0$, $\rho_0 = \langle \mathbf{r}^0, \mathbf{r}^0 \rangle$

do while (no convergencia)

$\mathbf{q}^k = \mathbf{A}\mathbf{p}^k$

$\alpha_k = -\frac{\rho_k}{\langle \mathbf{p}^k, \mathbf{q}^k \rangle}$

$\mathbf{x}^{k+1} = \mathbf{x}^k + \alpha_k \mathbf{p}^k$

$\mathbf{r}^{k+1} = \mathbf{r}^k + \alpha_k \mathbf{q}^k$

$\rho_{k+1} = \langle \mathbf{r}^{k+1}, \mathbf{r}^{k+1} \rangle$

$\beta_{k+1} = \frac{\rho_{k+1}}{\rho_k}$

$\mathbf{p}^{k+1} = \mathbf{r}^{k+1} + \beta_{k+1} \mathbf{p}^k$

$k = k + 1$

enddo

2.3.5. Método de gradientes conjugados precondicionado

Como ya se ha comentado, la convergencia del método de gradientes conjugados depende del número de condición de la matriz (ver proposición 3). La convergencia del método se puede mejorar resolviendo un sistema equivalente pero con una matriz con número de condición menor, es decir, precondicionado.

Consideremos una matriz $\mathbf{P}$ (precondicionador) *fácilmente invertible*[8] y con $k_2(\mathbf{P}^{-1}\mathbf{A}) \simeq 1$ (típicamente se trata de una matriz $\mathbf{P} \simeq \mathbf{A}$). El sistema de ecuaciones

$$\left(\mathbf{P}^{-1}\mathbf{A}\right)\mathbf{x} = \mathbf{P}^{-1}\mathbf{b},$$

sistema precondicionado, es equivalente al sistema $\mathbf{A}\mathbf{x} = \mathbf{b}$. Es decir, tiene la misma solución. Sin embargo, al reducir el numero de condición la convergencia del método sea más rápida para el sistema precondicionado, obteniendo así la solución con la precisión deseada en menos iteraciones. Esta transformación del sistema se denomina precondicionamiento por la derecha y es muy utilizada para mejorar la convergencia de métodos iterativos en general. Sin embargo, no es la utilizada en el caso del método de gradientes conjugados.

[8]Se dice que una matriz es fácilmente invertible cuando es sencillo resolver sistemas con esa matriz. En la práctica nunca resulta rentable calcular la inversa de una matriz.

La razón es que, aunque $\mathbf{P}$ y $\mathbf{A}$ sean matrices simétricas y definidas positivas, el producto $\mathbf{P}^{-1}\mathbf{A}$ en general no es una matriz simétrica y, por lo tanto, no es posible aplicar el método de los gradientes conjugados.

En el método de los gradientes conjugados el precondicionamiento del sistema se hace de manera que la matriz siga siendo simétrica y definida positiva. Para ello se considera un precondicionador $\mathbf{P}$ simétrico y definido positivo, de manera que es posible considerar su matriz raíz cuadrada[9] (aunque, como se comenta más adelante, a la práctica no es necesario calcularla), $\mathbf{P}^{\frac{1}{2}}$, y se puede transformar el sistema de ecuaciones $\mathbf{Ax} = \mathbf{b}$ en

$$\mathbf{P}^{-\frac{1}{2}}\mathbf{A}\mathbf{P}^{-\frac{1}{2}}\underbrace{\mathbf{P}^{\frac{1}{2}}\mathbf{x}}_{\tilde{\mathbf{x}}} = \mathbf{P}^{-\frac{1}{2}}\mathbf{b}.$$

Así, definiendo $\tilde{\mathbf{A}} = \mathbf{P}^{-\frac{1}{2}}\mathbf{A}\mathbf{P}^{-\frac{1}{2}}$, $\tilde{\mathbf{x}} = \mathbf{P}^{\frac{1}{2}}\mathbf{x}$ y $\tilde{\mathbf{b}} = \mathbf{P}^{-\frac{1}{2}}\mathbf{b}$, se puede obtener la solución resolviendo con gradientes conjugados el sistema equivalente

$$\tilde{\mathbf{A}}\tilde{\mathbf{x}} = \tilde{\mathbf{b}},$$

y calculando $\mathbf{x} = \mathbf{P}^{-\frac{1}{2}}\tilde{\mathbf{x}}$. Se deja como ejercicio para el lector comprobar que la matriz del sistema equivalente $\tilde{\mathbf{A}}$ es una matriz simétrica y definida positiva y, por lo tanto, no hay problemas para resolver el sistema mediante gradientes conjugados. Además, si $\mathbf{P} \simeq \mathbf{A}$ es de esperar que el número de condición de $\tilde{\mathbf{A}}$ sea pequeño y que, por lo tanto, la velocidad se acelere. La mejora en la velocidad de convergencia dependerá de la elección del precondicionador $\mathbf{P}$.

A la práctica el algoritmo se simplifica **evitando el cálculo de la matriz raíz cuadrada** $\mathbf{P}^{\frac{1}{2}}$. Si al aplicar el método de gradientes conjugados al sistema equivalente se utilizan las variables auxiliares

$$\tilde{\mathbf{x}}^k = \mathbf{P}^{\frac{1}{2}}\mathbf{x}^k, \quad \tilde{\mathbf{r}}^k = \tilde{\mathbf{A}}\tilde{\mathbf{x}}^k - \tilde{\mathbf{b}} = \mathbf{P}^{-\frac{1}{2}}\mathbf{r}^k, \quad \tilde{\mathbf{p}}^k = \mathbf{P}^{\frac{1}{2}}\mathbf{p}^k,$$

se puede deshacer el cambio para expresar el algoritmo en las variables sin tilde. Simplificando algunas expresiones el algoritmo final se puede escribir cómo

[9]Dada una matriz $\mathbf{P}$ definida positiva, la matriz raíz cuadrada $\mathbf{P}^{\frac{1}{2}}$ es una matriz que cumple $\mathbf{P} = \mathbf{P}^{\frac{1}{2}}\mathbf{P}^{\frac{1}{2}}$. Si $\mathbf{P}$ diagonaliza como $\mathbf{P} = \mathbf{SDS^{-1}}$ la matriz raíz cuadrada se calcula como $\mathbf{P}^{\frac{1}{2}} = \mathbf{S}\mathbf{D}^{\frac{1}{2}}\mathbf{S^{-1}}$, donde $\mathbf{D}^{\frac{1}{2}}$ se obtiene calculando la raíz cuadrada de cada coeficiente de la matriz diagonal.

$\mathbf{x}^0$: aproximación inicial

$k = 0$, $\mathbf{r}^0 = \mathbf{A}\mathbf{x}^0 - b$, $\mathbf{q}^0 = \mathbf{P}^{-1}\mathbf{r}^0$,

$\mathbf{p}^0 = \mathbf{q}^0$, $\rho_0 = (\mathbf{r}^0, \mathbf{q}^0)$

do while (no convergencia)

$\alpha_k = -\frac{\rho_k}{(\mathbf{p}^k, \mathbf{A}\mathbf{p}^k)}$

$\mathbf{x}^{k+1} = \mathbf{x}^k + \alpha_k \mathbf{p}^k$

$\mathbf{r}^{k+1} = \mathbf{r}^k + \alpha_k \mathbf{A}\mathbf{p}^k$

$\mathbf{q}^{k+1} = \mathbf{P}^{-1}\mathbf{r}^{k+1}$

$\rho_{k+1} = (\mathbf{r}^{k+1}, \mathbf{q}^{k+1})$

$\beta_{k+1} = \frac{\rho_{k+1}}{\rho_k}$

$\mathbf{p}^{k+1} = \mathbf{q}^{k+1} + \beta_{k+1}\mathbf{p}^k$

$k = k + 1$

enddo

donde se puede observar que no aparece la matriz $\mathbf{P}^{\frac{1}{2}}$ y sólo es necesario resolver un sistema con el precondicionador $\mathbf{P}$ en cada iteración.

Existen varias posibilidades para la elección del precondicionador $\mathbf{P}$, teniendo en cuenta que el único requisito que debe cumplir es que sea simétrico y definido positivo. Algunos de los más populares son $\mathbf{P} = \mathbf{D}_A$, precondicionador diagonal, u otros precondicionadores inspirados en los método iterativos estacionarios, y las factorizaciones incompletas, $\mathbf{P} = \mathbf{L}\mathbf{L}^T$. Las factorizaciones incompletas son similares a las factorizaciones estándar de la matriz, por ejemplo $\mathbf{L}\mathbf{L}^T \simeq \mathbf{A}$, pero sin almacenar algunos de los coeficientes. De esta manera se obtiene un precondicionador *fácilmente invertible*, con $\mathbf{P} \simeq \mathbf{A}$, pero sin un excesivo llenado de la factorización y manteniendo bajo el tiempo de CPU.

Referencias

[1] E. Isaacson and H.B. Keller, H.B. *Analysis of numerical methods*, Wiley, 1966.

[2] J. Stoer and R. Bulirsch. *Introduction to numerical analysis*, Springer-Verlag, 1980.

[3] R. Barret, M. Berry, T. Chan, J. Demmel, J. Donato, J. Dongarra, V. Eijkhout, R. Pozo, C. Romine and H. van der Vorst. *Templates for the Solution of Linear Systems: Building Blocks for Iterative Methods*, SIAM, Philadelphia, PA, 1993.