Sección de Matemáticas
Universidad de La Laguna

Christian Manuel Bartolomé Moreno

Métodos de Krylov para sistemas lineales

Krylov methods for linear systems

Trabajo de Fin de Grado
Grado en Matemáticas
La Laguna, Julio de 2019

Dirigido por
Domingo Hernández Abreu

Agradecimientos

A mi familia, por los valores que siempre me han infundado y el incondicional apoyo brindado en cada una de las etapas de mi vida. En especial, a la memoria de mi padre.

A mis amigos, tanto los nuevos como los de siempre, cuyo compañerismo unido a los inolvidables momentos juntos que estos años han deparado, han hecho mucho más llevadero el camino recorrido hasta la finalización del grado.

A Domingo Hernández Abreu, al que admiro por la manera impecable y gratificante que ha tenido de acompañarme a lo largo de estos meses, facilitando y amenizando la confección de este trabajo.

En definitiva, a todas las personas que han colaborado de manera firme en mi desarrollo, tanto personal como académico, a lo largo de estos años de carrera, haciéndome crecer hasta ser mejor matemático pero, sobre todo, mejor persona.

Autor del TFG
La Laguna, 8 de julio de 2019

Resumen · Abstract

Resumen

Los subespacios de Krylov, que deben su nombre al matemático ruso Alekséi Krylov, son hoy en día la base sobre la que se fundamentan los métodos iterativos modernos a la hora de calcular vectores y valores propios o para resolver sistemas de ecuaciones lineales $Ax = b$, empleando una menor cantidad de memoria y tiempo de proceso que el resto de métodos convencionales. Sin ir más lejos, los métodos numéricos de Gauss-Seidel y Jacobi, que se enseñan a lo largo del Grado, presentan un rendimiento bastante pobre cuando se les aplica a problemas diferenciales en 2 y 3 dimensiones espaciales.

Así, estos algoritmos basados en estos subespacios, llamados métodos de subespacios de Krylov,han registrado buenos resultados dentro del álgebra lineal numérica. Entre ellos, los más conocidos son el método del Gradiente Conjugado (GC) junto a su variante para ecuaciones normales (GCNR) y el Método del Residual Mínimo Generalizado (GMRES), siendo su eficiencia y tasa de convergencia los aspectos clave en torno a los cuales gira el estudio acometido en este trabajo. Todo ello quedará reflejado gráficamente a lo largo del último capítulo, en el que se detallará el comportamiento de estos métodos aplicados a Ecuaciones en Derivadas Parciales lineales elípticas y comparando cómo varían los resultados al aplicar precondicionamiento a la matriz de los sistemas resultantes de su discretización espacial mediante diferencias Diferencias Finitas.

Palabras clave: *Métodos de Proyección – Subespacios de Krylov – Gradiente Conjugado – Gradiente Conjugado para Ecuaciones Normales – Método del Residual Mínimo Generalizado.*

Abstract

Krylov subspaces, which owe their name to the Russian mathematician Alekséi Krylov are today the basis on which modern iterative methods are based when calculating vectors and eigenvalues or solving systems of linear equations $Ax = b$, using a smaller amount of memory and CPU time than the rest of conventional methods. Without going any further, the numerical methods of Gauss-Seidel and Jacobi, which are taught throughout the degree, show a rather poor performance when applied to differential problems in two and three spatial dimensions. Thus, algorithms based on these subspaces, called Krylov subspace methods, have great acceptance in numerical linear algebra. Among them, the best known are the Conjugate Gradient (GC) method with its variant for normal equations (GCNR) and the Generalized Minimum Residual Method (GMRES), being their efficiency and convergence rate the key aspects of the study undertaken in this work. All this will be reflected graphically throughout the last chapter, which will detail the behavior of these methods applied to linear elliptic Partial Differential Equations and comparing how the results vary when applying preconditioning to the matrix resulting after spatial discretization by means of Finite Differences.

Keywords: *Projection Methods – Krylov Subspaces – Conjugate Gradient – Conjugate Gradient for normal equations – Generalized Minimum Residual Method.*

Contenido

1

Métodos de proyección para la resolución de sistemas lineales

1.1. Introducción

El estudio de los sistemas lineales $Ax = b$ ha sido siempre una cuestión que ha suscitado gran interés entre la comunidad matemática por aparecer asociados a numerosos fenómenos que se presentan comúnmente en las aplicaciones. Sin embargo, estos sistemas pueden llegar a contar con dimensión tan grande que su resolución deja de ser en cualquier caso sencilla.
Enmarcado dentro del álgebra lineal numérica, este trabajo presenta los llamados subespacios de Krylov y se enfoca en el estudio de la eficacia y tasa de convergencia de una serie de métodos numéricos, conocidos como métodos de proyección, entre los que destacan aquellos de tipo Krylov, por ser lo suficientemente potentes como para poder estimar la solución de sistemas lineales cuyas dimensiones son enormes.
De esta forma, siguiendo [8, Cap. 5], el primer capítulo de este trabajo introduce unas primeras nociones acerca de los métodos de proyección y sus propiedades, relacionándolos con algunos métodos conocidos como Jacobi y Gauss-Seidel y tratando algunos métodos de proyección unidimensional. Lo anterior funciona como un preámbulo del Capítulo 2, en el que se emplean [7, Cap. 11], [8, Cap. 6-10] y [9, Parte VI] para exponer el método del Gradiente Conjugado (GC) junto a su versión para ecuaciones normales (GCNR), y el Método del Residual Mínimo Generalizado (GMRES). Una vez presentados los métodos y sus algoritmos, la atención recaerá en cómo acelerar su convergencia, para lo cual se empleará el precondicionamiento de matrices y se hará uso de las descomposiciones LU y de Cholesky, así como de sus variantes incompletas y modificadas. Finalmente, el tercer capítulo recoge una aplicación de estos métodos en EDPs lineales elípticas, ilustrando y comparando cómo varían su eficacia y convergencia según las propiedades de simetría de la matriz A y el precondicionador empleado.

1.2. Métodos de proyección: definición y propiedades

Sea el sistema lineal $Ax = b$, con $A \in \mathbb{R}^{N\times N}$ regular y $b \in \mathbb{R}^N$. Si denotamos la solución exacta del problema como $x^* = A^{-1}b$ y $x^{(0)} \in \mathbb{R}^N$ una aproximación inicial a x^*, definimos entonces el vector residual inicial como:

$$r^{(0)} := A(x - x^{(0)}) = b - Ax^{(0)}.$$

Definición 1.1. *Sean $\mathcal{K}$ y $\mathcal{L}$ subespacios vectoriales de $\mathbb{R}^N$. El método de proyección sobre $\mathcal{K}$ ortogonal a $\mathcal{L}$ consiste en:*

$$\boxed{\text{Hallar } \delta \in \mathcal{K} \text{ tal que } \langle r^{(0)} - A\delta, w\rangle_2 = 0, \forall w \in \mathcal{L} \text{ y tomar } x^{(1)} := x^{(0)} + \delta,} \quad (1.1)$$

siendo $r^{(0)} := b - Ax^{(0)}$ *el residual de la aproximación* $x^{(0)}$ *y* $\langle ., . \rangle_2$ *el producto euclídeo de* $\mathbb{R}^N$.

Nótese que el residual de $x^{(1)}$ es precisamente $r^{(1)} := b - Ax^{(1)} = b - Ax^{(0)} - A\delta = r^{(0)} - A\delta$. Los siguientes ejemplos muestran cómo los métodos clásicos de Jacobi y Gauss-Seidel pueden interpretarse como métodos de proyección.

Ejemplo 1.2. (**Método de Jacobi**). El método de Jacobi sigue el siguiente esquema iterativo:

$$a_{ii} \cdot x_i^{(1)} = b_i - \sum_{\substack{j=1 \\ j \neq i}}^{N} a_{ij} \cdot x_j^{(0)}, \quad 1 \leq i \leq N, \quad \text{si } a_{ii} \neq 0, \forall i \in 1, ..., N.$$

Ahora bien, este algoritmo se puede interpretar como método de proyección prosiguiendo como sigue:
Para cada $i \in \{1, ..., N\}$ definimos $\mathcal{K}_i = \mathcal{L}_i = span\{e_i\}$, siendo e_i el i-ésimo vector canónico de $\mathbb{R}^N$ y buscamos $\delta^{(i)} = \delta_i \cdot e_i \in K_i$, con $\delta_i \in \mathbb{R}$ cumpliendo que

$$\langle r^{(0)} - A\delta^{(i)}, w \rangle_2, \forall w \in \mathcal{L}_i,$$

esto es,

$$0 = \langle r^{(0)} - A\delta^{(i)}, e_i \rangle_2 = b_i - \sum_{j=1}^{N} a_{ij} x_j^{(0)} - \delta_i a_{ii}.$$

Luego, despejando, se obtiene $\delta_i = \dfrac{1}{a_{ii}} \left[b_i - \displaystyle\sum_{j=1}^{N} a_{ij} x_j^{(0)} \right]$. Tomamos entonces $x^{(1)} = x^{(0)} + \delta^{(1)} + ... + \delta^{(N)}$. Se tiene así:

$$x_i^{(1)} = x_i^{(0)} + \delta_i = \frac{1}{a_{ii}} \left[b_i - \sum_{\substack{j=1 \\ j \neq i}}^{N} a_{ij} \cdot x_j^{(0)} \right], 1 \leq i \leq N,$$

que coincide con la nueva aproximación del método de Jacobi.

Ejemplo 1.3. (**Método de Gauss-Seidel**). Recordamos que el método de Gauss-Seidel sigue el siguiente esquema iterativo:

$$a_{ii} x_i^{(1)} = b_i - \sum_{j<i} a_{ij} x_j^{(1)} - \sum_{j>i} a_{ij} x_j^{(0)}, \quad 1 \leq j \leq N. \tag{1.2}$$

Consideremos $x^{(0,0)} := x^{(0)}, r^{(0,0)} := r^{(0)} = b - Ax^{(0)}$. Para cada $i \in \{1, ..., N\}$, sean $\mathcal{K}_i = \mathcal{L}_i = span\{e_i\}$ y busquemos $\delta^{(i)} = \delta_i e_i \in \mathcal{K}_i$, con $\delta_i \in \mathbb{R}$ tal que:

$$\langle r^{(0,i-1)} - A\delta^{(i)}, w \rangle_2 = 0, \forall w \in \mathcal{L}_i, \text{ con } r^{(0,i-1)} = b - Ax^{(0,i-1)}, \tag{1.3}$$

y tomemos

$$x^{(0,i)} := x^{(0,i-1)} + \delta^{(i)}. \tag{1.4}$$

Finalmente, tomamos

$$x^{(1)} := x^{(0,N)}. \tag{1.5}$$

Análogamente al caso previo, tendremos por (1.3) que:

$$\delta_i = \frac{1}{a_{ii}} \left[b_i - \sum_{j=1}^{N} a_{ij} x_j^{(0,i-1)} \right], 1 \leq i \leq N.$$

Además, de (1.4):

$$x_j^{(0,i)} = x_j^{(0,i-1)}, \quad \text{si } j \neq i. \tag{1.6}$$

$$x_i^{(0,i)} = x_i^{(0,i-1)} + \delta_i = \frac{1}{a_{ii}} \left[b_i - \sum_{j \neq i} a_{ij} x_j^{(0,i-1)} \right]. \tag{1.7}$$

Por otra parte, de (1.5) y (1.6):

$$x_i^{(1)} = x_i^{(0,N)} = x_i^{(0,i)}. \tag{1.8}$$

Luego, de (1.6), (1.7) y (1.8):

$$\begin{aligned} x_i^{(1)} = x_i^{(0,i)} &= \frac{1}{a_{ii}} \left[b_i - \sum_{j<i} a_{ij} x_j^{(0,i-1)} - \sum_{j>i} a_{ij} x_j^{(0,i-1)} \right] = \\ &= \frac{1}{a_{ii}} \left[b_i - \sum_{j<i} a_{ij} x_j^{(0,j)} - \sum_{j>i} a_{ij} x_j^{(0,0)} \right] = \\ &= \frac{1}{a_{ii}} \left[b_i - \sum_{j<i} a_{ij} x_j^{(1)} - \sum_{j>i} a_{ij} x_j^{(0)} \right], \quad 1 \leq i \leq N, \end{aligned}$$

llegando finalmente a obtener así el método de Gauss-Seidel.

En lo que sigue asumimos que $\mathcal{K}$ y $\mathcal{L}$ tienen igual dimensión $m \leq N$, y están respectivamente generados por sendos sistemas linealmente independientes $\{v_1, \ldots, v_m\}, \{w_1, \ldots, w_m\}$ de vectores en $\mathbb{R}^N$, esto es, $\mathcal{K} = span\{v_1, ..., v_m\}$ y $\mathcal{L} = span\{w_1, ..., w_m\}$, e identifiquemos $V = [v_1|v_2|...|v_m]$, $W = [w_1|...|w_m] \in \mathbb{R}^{N \times M}$.

Lema 1.4. *Sea A regular. $W^T AV$ es singular si y solo si existe $v \in A\mathcal{K} \setminus \{0\}$ tal que $v \perp \mathcal{L}$, siendo $A\mathcal{K} = \{AVy / y \in \mathbb{R}^N\}$.*

Demostración. $\Leftarrow$) Sabemos que existe $y \in \mathbb{R}^m$, $y \neq 0$, tal que $v = AVy$ y $w_j^T AVy = 0, \quad 1 \leq j \leq m$. Luego, $W^T AVy = 0$ y $W^T AV$ es singular.
$\Rightarrow$) Ahora, tenemos que existe $y \in \mathbb{R}^m, y \neq 0$, tal que $W^T AVy = 0$. Además, $AVy \neq 0$, pues $\{v_1, v_2, ..., v_m\}$ son linealmente independientes y A es regular. Sea $v = AVy$. Entonces, $W^T v = 0$ y $w_j^T \cdot v = 0, \quad 1 \leq j \leq m$. □

Observación 1.5. El lema previo nos induce a afirmar que el método de proyección sobre $\mathcal{K}$ ortogonal a $\mathcal{L}$ tiene solución única si y solo si $A\mathcal{K} \cap \mathcal{L}^\perp = \{0\}$, donde $\mathcal{L}^\perp = \{v \in \mathbb{R}^N / \langle v, w \rangle_2 = 0, \forall w \in \mathcal{L}\}$ denota el complemento ortogonal de $\mathcal{L}$.

Proposición 1.6. *Sea A regular y $\mathcal{K}$ tal que $A\mathcal{K} \subseteq \mathcal{K}$, con $A\mathcal{K} \cap \mathcal{L}^\perp = \{0\}$. Si $r^{(0)} \in \mathcal{K}$, entonces, $x^{(1)} = x^*$.*

Demostración. Como $A\mathcal{K} \subseteq \mathcal{K}$ y $dim(A\mathcal{K}) = dim(\mathcal{K})$ por ser A regular, sigue que $A\mathcal{K} = \mathcal{K}$. Luego, $r^{(0)} \in A\mathcal{K}$, y $r^{(0)} - A\delta \in A\mathcal{K} \cap \mathcal{L}^\perp = \{0\}$. En definitiva, $r^{(0)} - A\delta = 0$, siendo $r^{(0)} - A\delta = b - Ax^{(1)}$. Como A es regular, $x^{(1)} = x^*$. □

Observación 1.7. Nótese que, si $\mathcal{K} = span\{v_1, ..., v_m\}$, entonces $A\mathcal{K} = span\{Av_1, ..., Av_m\}$, siendo $Av_1, ..., Av_m$ linealmente independientes, siempre que A sea regular.

Observación 1.8. Dado $x \in \mathbb{R}^N$, consideramos el problema de hallar $\delta \in \mathcal{K}$ tal que $Ax - A\delta \in \mathcal{L}^\perp$, siendo $A\mathcal{K} \cap \mathcal{L}^\perp = \{0\}$, y definimos la aplicación:

$$\begin{aligned} Q_{\mathcal{K}}^{\mathcal{L}}\colon\ \mathbb{R}^N &\longrightarrow \mathbb{R}^N \\ x &\longmapsto Q_{\mathcal{K}}^{\mathcal{L}}(x) := \delta. \end{aligned}$$

Esta aplicación está bien definida y es lineal, ya que, si ponemos $\delta = Vy \in \mathcal{K}$, se tiene que

$$W^T[Ax - AVy] = 0 \Leftrightarrow y = (W^TAV)^{-1} \cdot W^TAx,$$

de forma que $\delta = V(W^TAV)^{-1} \cdot W^TAx$, siendo V y W matrices asociadas a bases cualesquiera de $\mathcal{K}$ y $\mathcal{L}$, respectivamente. Luego:

$$Q_{\mathcal{K}}^{\mathcal{L}}(x) = V(W^TAV)^{-1} \cdot W^TAx$$

El operador $Q_{\mathcal{K}}^{\mathcal{L}}$ verifica las siguiente propiedades:

(I) Si $x \in \mathcal{K}, x = Vy$, para algún $y \in \mathbb{R}^m$, y entonces $Q_{\mathcal{K}}^{\mathcal{L}}(x) = V(W^TAV)^{-1} \cdot W^TA(Vy) = Vy = x$. En particular, $(Q_{\mathcal{K}}^{\mathcal{L}})^2 = Q_{\mathcal{K}}^{\mathcal{L}}$.

(II) $\forall x \in \mathbb{R}^N$, $Ax - AQ_{\mathcal{K}}^{\mathcal{L}}(x) \in \mathcal{L}^\perp$. En efecto,

$$W^T[Ax - AV(W^TAV)^{-1}W^TAx] = W^TAx - W^TAx = 0.$$

Finalmente, observamos que el método de proyección sobre $\mathcal{K}$ ortogonal a $\mathcal{L}$ consistente en hallar $x^{(1)} = x^{(0)} + \delta$, con $\delta \in \mathcal{K}$ y $r^{(0)} - A\delta = A(x^* - x^{(0)}) - A\delta \in \mathcal{L}^\perp$, coincide con:

$$x^{(1)} = x^{(0)} + Q_{\mathcal{K}}^{\mathcal{L}}(x^* - x^{(0)}) = (I - Q_{\mathcal{K}}^{\mathcal{L}})(x^{(0)}) + Q_{\mathcal{K}}^{\mathcal{L}}(x^*).$$

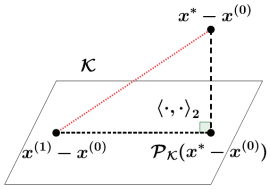
En particular, $Q_{\mathcal{K}}^{\mathcal{L}}(x^* - x^{(1)}) = 0$, pues$(Q_{\mathcal{K}}^{\mathcal{L}})^2 = Q_{\mathcal{K}}^{\mathcal{L}}$.

Observación 1.9. Observamos que el error $x^{(1)} - x^*$ verifica

$$\left\|x^{(1)} - x^*\right\|_2 = \left\|\delta - (x^* - x^{(0)})\right\|_2 \geq dist_2(x^* - x^{(0)}, \mathcal{K}),$$

donde $dist_2(x, \mathcal{K}) := \text{ínf}\{z \in \mathcal{K}, \|x - z\|_2\}$.

El siguiente teorema nos da una cota superior general para el error $\left\|x^{(1)} - x^*\right\|_2$ en términos de $dist_2(x^* - x^{(0)}, \mathcal{K})$. Para ello, introduciremos primero el operador de proyección ortogonal sobre $\mathcal{K}$:



$$\mathcal{P}_{\mathcal{K}}\colon\ \begin{aligned} \mathbb{R}^N &\longrightarrow \mathbb{R}^N \\ x &\longmapsto \mathcal{P}_{\mathcal{K}}(x), \end{aligned}\ \text{cumpliendo que}\ \begin{cases} \mathcal{P}_{\mathcal{K}}(x) \in \mathcal{K} \\ x - \mathcal{P}_{\mathcal{K}}(x) \in \mathcal{K}^\perp. \end{cases}$$

Es sencillo comprobar que, si V es la matriz asociada a una base de $\mathcal{K}$, entonces $\mathcal{P}_{\mathcal{K}}(x) = V(V^TV)^{-1}V^Tx$ donde $V(V^TV)^{-1}V^T$ es independiente de la base elegida. En particular,

- $\mathcal{P}_{\mathcal{K}}$ es lineal.
- $\mathcal{P}_{\mathcal{K}}(x) = Vy$, con $y = \left(V^TV\right)^{-1}V^Tx$. Luego, $\mathcal{P}_{\mathcal{K}} \in \mathcal{K}$.
- $\mathcal{P}_{\mathcal{K}}^2 = \mathcal{P}_{\mathcal{K}}$, puesto que $\mathcal{P}_{\mathcal{K}}^2(x) = \left[V(V^TV)^{-1}V^T\right]\left[V(V^TV)^{-1}V^T\right]x = \mathcal{P}_{\mathcal{K}}(x)$.
- $V^T\left[x - \mathcal{P}_{\mathcal{K}}(x)\right] = V^Tx - V^TV\left(V^TV\right)^{-1}V^Tx = V^Tx - V^Tx = 0$.
- Se cumple que $\forall v \in \mathcal{K}$:

$$\begin{aligned} \|x - v\|_2^2 &= \|x - \mathcal{P}_{\mathcal{K}}(x)\|_2^2 + \|\mathcal{P}_{\mathcal{K}}(x) - v\|_2^2 + 2\langle x - \mathcal{P}_{\mathcal{K}}(x), \mathcal{P}_{\mathcal{K}}(x) - v\rangle_2 = \\ &= \|x - \mathcal{P}_{\mathcal{K}}(x)\|_2^2 + \|\mathcal{P}_{\mathcal{K}}(x) - v\|_2^2 \end{aligned}$$

  puesto que $(x - \mathcal{P}_{\mathcal{K}}) \in \mathcal{K}^\perp$ y $(\mathcal{P}_{\mathcal{K}}(x) - v) \in \mathcal{K}$. Luego, $dist_2(x, \mathcal{K}) = \|x - \mathcal{P}_{\mathcal{K}}(x)\|_2$.

Teorema 1.10. $\left\|x^{(1)} - x^*\right\|_2 \leq \left(1 + \left\|Q_{\mathcal{K}}^{\mathcal{L}}(I - \mathcal{P}_{\mathcal{K}})\right\|_2\right) \cdot dist_2\left(x^* - x^{(0)}, \mathcal{K}\right)$.

Demostración. Pongamos $z^{(1)} = x^{(1)} - x^{(0)}$ y $z^* = x^* - x^{(0)}$. Entonces, por definición del método de proyección, se tiene que $z^{(1)} = Q_{\mathcal{K}}^{\mathcal{L}}(z^*)$. Luego: $x^{(1)} - x^* = z^{(1)} - z^* = (z^{(1)} - \mathcal{P}_{\mathcal{K}}(z^*)) + (\mathcal{P}_{\mathcal{K}}(z^*) - z^*)$, siendo $\|\mathcal{P}_{\mathcal{K}}(z^*) - z^*\|_2 = dist_2(z^*, \mathcal{K})$. Por otra parte:

$$z^{(1)} - \mathcal{P}_{\mathcal{K}}(z^*) = Q_{\mathcal{K}}^{\mathcal{L}}(z^*) - Q_{\mathcal{K}}^{\mathcal{L}}\mathcal{P}_{\mathcal{K}}(z^*) = Q_{\mathcal{K}}^{\mathcal{L}}(I - \mathcal{P}_{\mathcal{K}})(z^*) = Q_{\mathcal{K}}^{\mathcal{L}}(I - \mathcal{P}_{\mathcal{K}}) \cdot (I - \mathcal{P}_{\mathcal{K}})(z^*),$$

teniendo en cuenta que $\mathcal{P}_{\mathcal{K}}^2 = \mathcal{P}_{\mathcal{K}}$. Así, se llega a que

$$\left\|z^{(1)} - \mathcal{P}_{\mathcal{K}}(z^*)\right\|_2 \leq \left\|Q_{\mathcal{K}}^{\mathcal{L}}(I - \mathcal{P}_{\mathcal{K}})\right\|_2 \cdot \|z^* - \mathcal{P}_{\mathcal{K}}(z^*)\|_2 ,$$

lo cual concluye la prueba. □

Observación 1.11. En forma matricial, el teorema anterior se escribe como:

$$\left\|x^{(1)} - x^*\right\|_2 \leq \left\{1 + \left\|V\left(W^T A V\right)^{-1} W^T A \left(I - V\left(V^T V\right)^{-1} V^T\right)\right\|_2\right\} \cdot dist_2\left(x^* - x^{(0)}, \mathcal{K}\right)$$

independientemente de las bases que se tomen para $\mathcal{K}$ y $\mathcal{L}$, respectivamente.

1.2.1. Métodos de proyección del error.

Consideramos aquí el caso particular de métodos de proyección (1.1) con A simétrica y definida positiva y $\mathcal{L} = \mathcal{K}$. Previamente, enunciamos el célebre Teorema Espectral para matrices simétricas ([2, Cap. 6]).

Teorema 1.12. *Sea $A \in \mathbb{R}^{N\times N}$ una matriz simétrica. Entonces:*

(I) *Cada autovalor λ de A es real y admite un autovector real u, esto es, para cada $\lambda \in \sigma[A]$, $\lambda \in \mathbb{R}$ y existe $u \in \mathbb{R}^N \backslash \{0\}$ tal que $Au = \lambda u$.*

(II) *Los autovectores correspondientes a autovalores distintos son ortogonales respecto de $\langle\cdot,\cdot\rangle_2$, es decir, si λ_1, λ_2 son autovalores de A tal que $\lambda_1 \neq \lambda_2$ y u_1, u_2 son autovectores correspondientes, entonces $u_1^T u_2 = 0$.*

(III) *Existe una matriz diagonal $D \in \mathbb{R}^{N\times N}$ y una matriz ortogonal $U \in \mathbb{R}^{N\times N}$, esto es, $U^T U = UU^T = I$, tal que $A = UDU^T$. Los elementos diagonales de D son los autovalores de A y las columnas de U son sus autovectores correspondientes.*

Definición 1.13. *Sea $A \in \mathbb{R}^{N\times N}$ simétrica y definida positiva, esto es, $v^T Av > 0, \forall v \in \mathbb{R}^N, v \neq 0$. Se define el producto energía en $\mathbb{R}^N$ como sigue:*

$$\langle u, v\rangle_A = \langle Au, v\rangle_2 = v^T Au, \quad \forall u, v \in \mathbb{R}^N,$$

con norma inducida (norma energía) $\|u\|_A = \sqrt{u^T Au}, \forall u \in \mathbb{R}^N$.

Proposición 1.14. *Sea A simétrica y definida positiva, $\mathcal{L} = \mathcal{K}$, $x^{(1)} = x^{(0)} + \delta$, con $\delta \in \mathcal{K}$, y $r^{(0)} = b - Ax^{(0)}$. Sea x^* la solución de $Ax = b$ y definamos $E(x) := \|x^* - x\|_A^2$, $x \in \mathbb{R}^N$. Entonces,*

$$\langle r^{(0)} - A\delta, v\rangle_2 = 0, \forall v \in \mathcal{K} \Leftrightarrow x^{(1)} \text{ minimiza } E(x) \text{ sobre } x^{(0)} + \mathcal{K}.$$

Demostración. $\Rightarrow$) Sea $\widetilde{x} = x^{(0)} + \widetilde{\delta}$, con $\widetilde{\delta} \in \mathcal{K}$. Escribimos:

$$\|x_* - \widetilde{x}\|_A^2 = \left\|x^* - x^{(1)}\right\|_A^2 + 2\langle A(x^* - x^{(1)}), x^{(1)} - \widetilde{x}\rangle_2 + \left\|x^{(1)} - \widetilde{x}\right\|_A^2 .$$

Como $A(x^* - x^{(1)}) = b - Ax^{(1)} = r^{(0)} - A\delta$ y $x^{(1)} - \widetilde{x} = \delta - \widetilde{\delta} \in \mathcal{K}$, concluimos

$$E(\widetilde{x}) = E(x^{(1)}) + \left\|x^{(1)} - \widetilde{x}\right\|_A^2 \geq E(x^{(1)}).$$

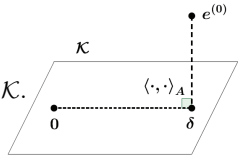
$\Leftarrow$) Si existiera $v \in \mathcal{K}$ tal que $\lambda := \langle r^{(0)} - A\delta, v\rangle_2 \neq 0$, entonces, tomando $\widetilde{x} = x^{(1)} + \dfrac{\lambda}{\|v\|_A^2} v \in x^{(0)} + \mathcal{K}$ se tiene análogamente que:

$$\|x^* - \widetilde{x}\|_A^2 = \left\|x^* - x^{(1)}\right\|_A^2 + 2\langle r^{(0)} - A\delta, \left(\frac{-\lambda}{\|v\|_A^2}\right) v\rangle_2 + \frac{\lambda^2}{\|v\|_A^2} =$$
$$= \left\|x^* - x^{(1)}\right\|_A^2 - \frac{\lambda^2}{\|v\|_A^2} < \left\|x^* - x^{(1)}\right\|_A^2 .$$

Por tanto, $E(\widetilde{x}) < E(x^{(1)})$. □

Denotemos $e^{(0)} = x^* - x^{(0)}$ y $e^{(1)} = x^* - x^{(1)} = e^{(0)} - \delta$. Entonces, se tiene que

$Ae^{(1)} = A(e^{(0)} - \delta) = r^{(0)} - A\delta$, siendo $\langle e^{(0)} - \delta, v\rangle_A = 0, \quad \forall v \in \mathcal{K}$.



Como $\delta \in \mathcal{K}$, δ es la proyección ortogonal respecto del producto $\langle \cdot, \cdot \rangle_A$ del error $e^{(0)}$ sobre $\mathcal{K}$. Así, estos métodos se denominan **métodos de proyección del error**.

Corolario 1.15. *En las mismas hipótesis de la Proposición 1.14, se tiene que*

$$\left\|x^* - x^{(1)}\right\|_A \leq \left\|x^* - x^{(0)}\right\|_A$$

Demostración. $\left\|e^{(0)}\right\|_A^2 = \left\|(e^{(0)} - \delta) + \delta\right\|_A^2 = \left\|e^{(0)} - \delta\right\|_A^2 + \|\delta\|_A^2 + 2\langle e^{(0)} - \delta, \delta\rangle_A$. Como $\langle e^{(0)} - \delta, \delta\rangle_A = 0$, se concluye que

$$\left\|e^{(1)}\right\|_A^2 = \left\|e^{(0)}\right\|_A^2 - \|\delta\|_A^2 \leq \left\|e^{(0)}\right\|_A^2 .$$

□

1.2.2. Métodos de proyección del residual.

Otra clase interesante de métodos de proyección surge cuando A es regular arbitraria y $\mathcal{L} = A\mathcal{K}$.

Definición 1.16. *Sea $A \in \mathbb{R}^{N\times N}$ arbitraria y $\mathcal{L} = A\mathcal{K}$. Definamos*

$$R(x) := \|b - Ax\|_2^2 = \|A(x^* - x)\|_2^2 , \quad x \in \mathbb{R}^N ,$$

la norma euclídea al cuadrado del vector residual asociado al vector x.

Proposición 1.17. *Sea* $A \in \mathbb{R}^{N\times N}$, $\mathcal{L} = A\mathcal{K}, x^{(1)} = x^{(0)} + \delta$, *con* $\delta \in \mathcal{K}$, *y* $r^{(0)} = b - Ax^{(0)}$. *Entonces,*

$$\langle r^{(0)} - A\delta, w\rangle_2 = 0, \forall w \in \mathcal{L} \Leftrightarrow x^{(1)} \text{ minimiza } R(x) \text{ sobre } x^{(0)} + \mathcal{K}.$$

Demostración. $\Rightarrow$) Sea $\widetilde{x} = x^{(0)} + \widetilde{\delta}$, con $\widetilde{\delta} \in \mathcal{K}$. Escribimos:

$$\|b - A\widetilde{x}\|_2^2 = \|A(x^* - \widetilde{x})\|_2^2 = \left\|A(x^* - x^{(1)})\right\|_2^2 + 2\langle A(x^* - x^{(1)}), A(x^{(1)} - \widetilde{x})\rangle + \left\|A(x^{(1)} - \widetilde{x})\right\|_2^2.$$

Como $A(x^* - x^{(1)}) = b - Ax^{(1)} = r^{(0)} - A\delta$ y $A(x^{(1)} - \widetilde{x}) = A(\delta - \widetilde{\delta}) \in \mathcal{L}$, se puede concluir que $R(\widetilde{x}) = R(x^{(1)}) + \left\|A(x^{(1)} - \widetilde{x})\right\|_2^2 \geq R(x^{(1)})$.

$\Leftarrow$) Si existiera $w = Av \in \mathcal{L}, w \neq 0, v \in \mathcal{K}$ tal que $\lambda := \langle r^{(0)} - A\delta, w\rangle_2 \neq 0$, entonces tomando $\widetilde{x} = x^{(1)} + \dfrac{\lambda}{\|w\|_2^2} v \in x^{(0)} + \mathcal{K}$, se obtiene que

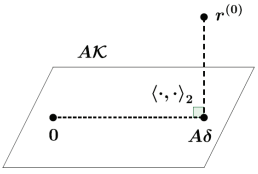
$$R(\widetilde{x}) = \|A(x^* - \widetilde{x})\|_2^2 = R(x^{(1)}) + 2\langle r^{(0)} - A\delta, \frac{-\lambda}{\|w\|_2^2} Av\rangle_2 + \frac{\lambda^2}{\|w\|_2^2} =$$

$$= R(x^{(1)}) - \frac{\lambda^2}{\|w\|_2^2} < R(x^{(1)}).$$

□

Dados $r^{(0)} = b - Ax^{(0)}$ y $r^{(1)} = b - Ax^{(1)} = r^{(0)} - A\delta$, tenemos que

$$\langle r^{(1)}, w\rangle_2 = 0, \quad \forall w \in A\mathcal{K}.$$



Como $A\delta \in A\mathcal{K}$, esto significa que $A\delta$ es la proyección ortogonal respecto de $\langle\cdot,\cdot\rangle_2$ del residual $r^{(0)}$ sobre $A\mathcal{K}$. Estos métodos se denominan **métodos de proyección del residual**.

Corolario 1.18. *En las mismas hipótesis de la proposición 1.17, se tiene que*

$$\left\|b - Ax^{(1)}\right\|_2 \leq \left\|b - Ax^{(0)}\right\|_2.$$

Demostración.

$$\left\|r^{(0)}\right\|_2^2 = \left\|(r^{(0)} - A\delta) + A\delta\right\|_2^2 = \left\|r^{(0)} - A\delta\right\|_2^2 + \|A\delta\|_2^2 + 2\langle r^{(0)} - A\delta, A\delta\rangle.$$

Ahora bien, como $r^{(0)} - A\delta = r^{(1)}$, por ortogonalidad, obtenemos:

$$\left\|r^{(1)}\right\|_2^2 = \left\|r^{(0)}\right\|_2^2 - \|A\delta\|_2^2 \leq \left\|r^{(0)}\right\|_2^2.$$

□

1.2.3. Métodos de proyección unidimensional

Sean $\mathcal{K} = span\{v\}$ y $\mathcal{L} = span\{w\}$ cumpliendo $A\mathcal{K} \cap \mathcal{L}^{\perp} = \{0\}$, esto es, $\langle Av, w\rangle_2 \neq 0$. Entonces, el método de proyección sobre $\mathcal{K}$ ortogonal a $\mathcal{L}$ (1.1) viene dado por:

$$x^{(1)} = x^{(0)} + \delta, \delta = \alpha \cdot v, \text{ con } \alpha \in \mathbb{R} \text{ tal que } \langle r^{(0)} - A\delta, w\rangle_2 = 0.$$

Luego, $\alpha = \dfrac{\langle r^{(0)}, w\rangle_2}{\langle Av, w\rangle_2}$ y, con todo ello:

$$\boxed{x^{(1)} = x^{(0)} + \frac{\langle r^{(0)}, w\rangle_2}{\langle Av, w\rangle_2} \cdot v} \text{ con } r^{(0)} = b - Ax^{(0)}. \tag{1.9}$$

Método del descenso más rápido

Consideramos A simétrica y definida positiva. Tomando $v = w = r^{(0)}$, se tiene que $\langle Ar^{(0)}, r^{(0)}\rangle_2 \neq 0$, siempre que $r^{(0)} \neq 0$, y el método (1.9) queda como:

$$\boxed{x^{(1)} = x^{(0)} + \frac{\left\|r^{(0)}\right\|_2^2}{\langle Ar^{(0)}, r^{(0)}\rangle_2} \cdot r^{(0)}} \text{ con } r^{(0)} = b - Ax^{(0)}. \tag{1.10}$$

Observamos que

$$r^{(1)} = b - Ax^{(1)} = b - Ax^{(0)} - \alpha Ar^{(0)} = r^{(0)} - \alpha Ar^{(0)}, \text{ con } \alpha = \frac{\left\|r^{(0)}\right\|_2^2}{\langle Ar^{(0)}, r^{(0)}\rangle_2}.$$

El siguiente algoritmo permite implementar el método del descenso más rápido con una única multiplicación matriz-vector por iteración. Dado $r^{(0)} = b - Ax^{(0)}$, calcular:

- $\boxed{d^{(0)} = Ar^{(0)},}$ $\boxed{\alpha = \dfrac{\langle r^{(0)}, r^{(0)}\rangle_2}{\langle d^{(0)}, r^{(0)}\rangle_2},}$
- $\boxed{x^{(1)} = x^{(0)} + \alpha r^{(0)},}$ $\boxed{r^{(1)} = r^{(0)} - \alpha \cdot d^{(0)}.}$ (1.11)

Observación 1.19. $\langle r^{(1)}, r^{(0)}\rangle_2 = 0$. En efecto, $\langle r^{(1)}, r^{(0)}\rangle_2 = \langle r^{(0)}, r^{(0)}\rangle_2 - \alpha \cdot \langle d^{(0)}, r^{(0)}\rangle_2 = 0$, por cómo ha sido definido α.

Observación 1.20. Por la proposición 1.17, $x^{(1)}$ minimiza $E(x) = \|x - x^*\|_A^2$ sobre $x^{(0)} + \mathcal{K}$, siendo $\|u\|_A = \sqrt{\langle Au, u\rangle_2}$ (con A simétrica y definida positiva), y $\mathcal{K} = span\{r^{(0)}\}$. Ahora bien, $E(x) = (x - x^*)^T A(x - x^*)$ y $\nabla E(x) = 2A(x - x^*)$. Como $-\nabla E(x^{(0)}) = 2 \cdot r^{(0)}$, se tiene que $\mathcal{K} = span\left\{-\nabla E(x^{(0)})\right\}$.
Luego, $x^{(1)}$ minimiza $E(x)$ sobre todos los vectores de la forma $x^{(0)} + \lambda(-\nabla E(x^{(0)})), \lambda \in \mathbb{R}$, siendo $-\nabla E(x^{(0)})$ localmente la dirección de decrecimiento más rápido para $E(x)$ en $x^{(0)}$, lo cual explica el nombre del método.

El siguiente teorema establecerá la convergencia del método del descenso más rápido para cualquier valor inicial $x^{(0)}$.

Teorema 1.21. *Sea A simétrica y definida positiva. Entonces, la solución de avance del método del descenso más rápido verifica:*

$$\left\|x^{(1)} - x^*\right\|_A \leq \frac{\kappa_2(A) - 1}{\kappa_2(A) + 1} \cdot \left\|x^{(0)} - x^*\right\|_A, \tag{1.12}$$

siendo $\kappa_2(A) = \|A\|_2 \cdot \left\|A^{-1}\right\|_2 = \dfrac{\lambda_{max}(A)}{\lambda_{min}(A)} := Cond_2(A)$.

Observación 1.22. Si A es simétrica y definida positiva, entonces todos sus autovalores son positivos. Además, $\|A\|_2 = \sqrt{\rho(A^*A)} = \rho(A) = \lambda_{max}(A)$, $\left\|A^{-1}\right\|_2 = \lambda_{max}(A^{-1}) = \dfrac{1}{\lambda_{min}(A)}$ y $\kappa_2(A) = \|A\|_2 \cdot \left\|A^{-1}\right\|_2 = \dfrac{\lambda_{max}(A)}{\lambda_{min}(A)} := Cond_2(A)$.

Demostración (Teorema 1.21). Sean $e^{(i)} = x^* - x^{(i)}, i = 0, 1$. Entonces, $A \cdot e^{(i)} = b - Ax^{(i)} = r^{(i)}, i = 0, 1$. Luego:

$$\begin{aligned}\left\|e^{(1)}\right\|_A^2 &= \langle e^{(1)}, e^{(0)} - \alpha r^{(0)}\rangle_A = \langle e^{(1)}, e^{(0)}\rangle_A - \alpha\langle e^{(1)}, r^{(0)}\rangle_A = \\ &= \langle e^{(1)}, r^{(0)}\rangle_2 - \alpha\langle r^{(1)}, r^{(0)}\rangle_2 = \langle e^{(1)}, r^{(0)}\rangle_2.\end{aligned}$$

Por tanto, $\left\|e^{(1)}\right\|_A^2 = \langle e^{(0)} - \alpha r^{(0)}, r^{(0)}\rangle_2 = \langle A^{-1} \cdot r^{(0)}, r^{(0)}\rangle_2 - \alpha\langle r^{(0)}, r^{(0)}\rangle_2$. Observamos que $\langle A^{-1}r^{(0)}, r^{(0)}\rangle_2 = \langle e^{(0)}, e^{(0)}\rangle_A = \left\|e^{(0)}\right\|_A^2$ y, si $e^{(0)} = 0$, entonces el enunciado es obvio (dado que $x^{(0)} = x^*$, $r^{(0)} = 0$ y $x^{(1)} = x^{(0)} = x^*$). En otro caso, si $e^{(0)} \neq 0$, se tiene que:

$$\begin{aligned}\left\|e^{(1)}\right\|_A^2 &= \left\|e^{(0)}\right\|_A^2 \cdot \left[1 - \alpha\frac{\langle r^{(0)}, r^{(0)}\rangle_2}{\langle A^{-1}r^{(0)}, r^{(0)}\rangle_2}\right] = \left\|e^{(0)}\right\|_A^2 \cdot \left[1 - \frac{\langle r^{(0)}, r^{(0)}\rangle_2^2}{\langle A^{-1}r^{(0)}, r^{(0)}\rangle_2\langle Ar^{(0)}, r^{(0)}\rangle_2}\right] \\ &\leq \left\|e^{(0)}\right\|_A^2 \cdot \left[1 - \frac{4 \cdot \lambda_{max}(A)\lambda_{min}(A)}{(\lambda_{max}(A) + \lambda_{min}(A))^2}\right] = \left\|e^{(0)}\right\|_A^2 \cdot \left[\frac{\lambda_{max}(A) - \lambda_{min}(A)}{\lambda_{max}(A) + \lambda_{min}(A)}\right]^2\end{aligned}$$

donde hemos empleado la desigualdad de Kantorovich del lema 1.23.

□

Lema 1.23 (Desigualdad de Kantorovich). *Sea B una matriz simétrica y definida positiva. Entonces*

$$\frac{\langle Bx, x\rangle_2\langle B^{-1}x, x\rangle_2}{\|x\|_2^4} \leq \frac{(\lambda_{max} + \lambda_{min})^2}{4 \cdot \lambda_{max}(B)\lambda_{min}(B)}, \quad \forall x \neq 0.$$

Demostración. Basta probarlo para todo $x \neq 0$ con $\|x\|_2 = 1$. Como B es simétrica, la podemos escribir como $B = Q^T DQ$, con $D = Diag(\lambda_i)_{i=1}^N$ y $Q \in \mathbb{R}^{N\times N}$ ortogonal, es decir, $Q^TQ = QQ^T = I_{N\times N}$. Luego,

$$\langle Bx, x\rangle_2\langle B^{-1}x, x\rangle_2 = \langle DQx, Qx\rangle_2 \cdot \langle D^{-1}Qx, Qx\rangle_2.$$

Poniendo $u := Qx$, con $\|u\|_2 = 1$, sigue que

$$\begin{aligned}\langle Bx, x\rangle_2\langle B^{-1}x, x\rangle_2 &= \left(\sum_{i=1}^N \lambda_i u_i^2\right) \cdot \left(\sum_{i=1}^N \lambda_i^{-1} u_i^2\right) \\ &= \left(\sum_{i=1}^N (c\lambda_i) u_i^2\right) \cdot \left(\sum_{i=1}^N (c\lambda_i)^{-1} u_i^2\right), \quad \forall c > 0 \text{ fijo.}\end{aligned}$$

Usando que $a \cdot b \leq \left(\frac{a+b}{2}\right)^2, \forall a, b \geq 0$, sigue que:

$$\langle Bx, x\rangle_2 \langle B^{-1}x, x\rangle_2 \leq \frac{1}{4}\left[\sum_{i=1}^{N}((c\lambda_i) + (c\lambda_i)^{-1})u_i^2\right]^2.$$

Como $f(\lambda) := c \cdot \lambda + (c\lambda)^{-1}$ es convexa si $\lambda > 0$, se tiene que:

$$f(\lambda) \leq max\{f(\lambda_1), f(\lambda_N)\}, \text{ cuando } \lambda \in [\lambda_1, \lambda_N].$$

Tomando $c = \frac{1}{\sqrt{\lambda_1\lambda_N}}$, se obtiene $f(\lambda_1) = f(\lambda_N) = \frac{\lambda_1 + \lambda_N}{\sqrt{\lambda_1\lambda_N}}$. Luego, con $\lambda_1 = \lambda_{min}(B)$ y $\lambda_N = \lambda_{max}(B)$:

$$\langle Bx, x\rangle_2 \langle B^{-1}x, x\rangle_2 \leq \frac{1}{4} \cdot \left(\frac{\lambda_1 + \lambda_N}{\sqrt{\lambda_1\lambda_N}}\right)^2 \cdot \left(\sum_{i=1}^{N} u_i^2\right)^2 = \frac{(\lambda_{max}(B) + \lambda_{min}(B))^2}{4 \cdot \lambda_{max}(B)\lambda_{min}(B)}$$

□

Método del Residual Mínimo

Consideremos $A \in \mathbb{R}^{N\times N}$ regular arbitraria y los vectores $v = r^{(0)}$, $w = Ar^{(0)}$, con subespacios asociados $\mathcal{K} = span\left\{r^{(0)}\right\}$ y $\mathcal{L} = A\mathcal{K} = span\left\{Ar^{(0)}\right\}$. Como $\langle Av, w\rangle_2 = \left\|Ar^{(0)}\right\|_2^2 \neq 0$ siempre que $r^{(0)} \neq 0$, el método de proyección (1.9) resultante será:

$$\boxed{x^{(1)} = x^{(0)} + \frac{\langle r^{(0)}, Ar^{(0)}\rangle_2}{\|Ar^{(0)}\|_2^2} r^{(0)}} \text{ con } r^{(0)} = b - Ax^{(0)}. \tag{1.13}$$

Nuevamente, $r^{(1)} = b - Ax^{(1)} = r^{(0)} - \alpha Ar^{(0)}$, siendo ahora $\alpha := \frac{\langle r^{(0)}, Ar^{(0)}\rangle_2}{\|Ar^{(0)}\|_2^2}$.
Llegamos entonces al siguiente algoritmo, que únicamente requiere una multiplicación matriz-vector por iteración. Dado$r^{(0)} = b - Ax^{(0)}$, calcular:

- $\boxed{d^{(0)} = Ar^{(0)},}$ $\boxed{\alpha = \frac{\langle r^{(0)}, d^{(0)}\rangle_2}{\langle d^{(0)}, d^{(0)}\rangle_2},}$
- $\boxed{x^{(1)} = x^{(0)} + \alpha r^{(0)},}$ $\boxed{r^{(1)} = r^{(0)} - \alpha d^{(0)}.}$

Observación 1.24. $\langle r^{(1)}, Ar^{(0)}\rangle_2 = \langle r^{(0)}, Ar^{(0)}\rangle_2 - \alpha\langle Ar^{(0)}, Ar^{(0)}\rangle_2 = 0$ por definicion de α.

Observación 1.25. Por la proposición 1.17, $x^{(1)}$ minimiza $R(x) = \|b - Ax\|_2^2$ sobre $x^{(0)} + \mathcal{K}$ con $\mathcal{K} = span\left\{r^{(0)}\right\}$. Esta propiedad motiva el nombre del método.

Teorema 1.26. *Sea A regular, $r^{(0)} \neq 0$ y $\theta \in \left[0, \frac{\pi}{2}\right]$ el ángulo entre los vectores $r^{(0)}$ y $Ar^{(0)}$, siendo* $\cos(\theta) := \frac{\langle r^{(0)}, Ar^{(0)}\rangle_2}{\|r^{(0)}\|_2 \|Ar^{(0)}\|_2}$. *Entonces,* $\left\|r^{(1)}\right\|_2 \leq \sin(\theta)\left\|r^{(0)}\right\|_2$.

Demostración.

$$\left\|r^{(1)}\right\|_2^2 = \langle r^{(0)} - \alpha A r^{(0)}, r^{(0)} - \alpha A r^{(0)}\rangle_2 = \left\|r^{(0)}\right\|_2^2 - 2\alpha\langle r^{(0)}, Ar^{(0)}\rangle_2 + \alpha^2 \left\|Ar^{(0)}\right\|_2^2 =$$
$$= \left\|r^{(0)}\right\|_2^2 - \alpha^2 \left\|Ar^{(0)}\right\|_2^2 = \left\|r^{(0)}\right\|_2^2 \cdot \left[1 - \frac{\langle r^{(0)}, Ar^{(0)}\rangle_2^2}{\|r^{(0)}\|^2 \|Ar^{(0)}\|^2}\right] =$$
$$= \left\|r^{(0)}\right\|_2^2 \cdot \left(1 - \cos^2(\theta)\right) = \sin^2(\theta) \cdot \left\|r^{(0)}\right\|_2^2.$$

□

Observación 1.27. En el teorema previo, $0 \leq sen(\theta) \leq 1$, pero no se excluye la posibilidad de que $sen(\theta) = 1$. Se puede garantizar la convergencia cuando A es definida positiva (no necesariamente simétrica).

Teorema 1.28. *Sea A definida positiva, $\mu = \lambda_{min}[\frac{1}{2}(A + A^T)]$ y $\sigma = \|A\|_2$. Entonces,*

$$\left\|r^{(1)}\right\|_2 \leq \left(1 - \left(\frac{\mu}{\sigma}\right)^2\right)^{\frac{1}{2}} \cdot \left\|r^{(0)}\right\|_2^2.$$

Demostración. Asumimos, sin pérdida de generalidad, que $r^{(0)} \neq 0$, puesto que si $r^{(0)} = 0$, entonces $r^{(1)} = 0$. Siguiendo los pasos de la demostración anterior, podemos escribir:

$$\left\|r^{(1)}\right\|_2^2 = \left\|r^{(0)}\right\|_2^2 \cdot \left[1 - \left(\frac{\langle r^{(0)}, Ar^{(0)}\rangle_2}{\langle r^{(0)}, r^{(0)}\rangle_2}\right)^2 \cdot \left(\frac{\left\|r^{(0)}\right\|_2}{\|Ar^{(0)}\|_2}\right)^2\right]$$

En particular, $0 \leq 1 - \left(\frac{\langle r^{(0)}, Ar^{(0)}\rangle_2}{\langle r^{(0)}, r^{(0)}\rangle_2}\right)^2 \cdot \left(\frac{\left\|r^{(0)}\right\|_2}{\|Ar^{(0)}\|_2}\right)^2 \leq 1$. Además,

1. Para $x \neq 0$, $\quad \frac{\left\|Ar^{(0)}\right\|_2}{\|r^{(0)}\|_2} \leq \sup_{x \in \mathbb{R}^N} \frac{\|Ax\|_2}{\|x\|_2} = \|A\|_2$.
2. $$\frac{\langle r^{(0)}, Ar^{(0)}\rangle_2}{\langle r^{(0)}, r^{(0)}\rangle_2} = \frac{\langle r^{(0)}, \frac{1}{2}\left(A + A^T\right) r^{(0)}\rangle_2}{\langle r^{(0)}, r^{(0)}\rangle_2} \geq \min_{x \in \mathbb{R}^N} \frac{\langle x, \frac{1}{2}(A + A^T)x\rangle_2}{\langle x, x\rangle_2} = \lambda_{min}(\frac{1}{2}(A + A^T)) > 0,$$

   puesto que $\frac{1}{2}(A + A^T)$ es simétrica y definida positiva. Luego:

$$\left\|r^{(1)}\right\|_2^2 \leq \left\|r^{(0)}\right\|_2^2 \left[1 - \left(\lambda_{min}\left(\frac{1}{2}(A + A^T)\right)^2\right) \left(\|A\|_2^{-1}\right)^2\right] = \left\|r^{(0)}\right\|_2^2 \left(1 - \frac{\mu^2}{\sigma^2}\right).$$

□

Observación 1.29. En la demostración previa hemos usado que, al ser $\frac{1}{2}(A + A^T)$ simétrica, admite la descomposición espectral $Q^T \cdot D \cdot Q$, con $D = diag(\lambda_i)$ y Q ortogonal ($Q^T = Q^{-1}$). Luego, para u unitario, se tiene:

$$\langle u, Q^T DQu\rangle_2 = \langle Qu, DQu\rangle_2 = \langle \widetilde{u}, D\widetilde{u}\rangle_2 = \sum_{i=1}^{N} \lambda_i \widetilde{u_i}^2 \geq \lambda_1 = \lambda_{min}\left(\frac{1}{2}\left(A + A^T\right)\right).$$

Corolario 1.30. *Si A es simétrica y definida positiva, entonces*

$$\left\|r^{(1)}\right\|_2 \leq \left(\frac{\kappa_2(A)^2-1}{\kappa_2(A)^2}\right)^{\frac{1}{2}} \cdot \left\|r^{(0)}\right\|_2.$$

Demostración. Como A es simétrica y definida positiva, se tiene que $\lambda_{min}\left(\frac{1}{2}\left(A+A^T\right)\right) = \lambda_{min}(A)$ y $\|A\|_2 = \lambda_{max}(A)$. El teorema previo nos daría

$$1-\left(\frac{\mu}{\sigma}\right)^2 = 1-\left(\frac{\lambda_{min}(A)}{\lambda_{max}(A)}\right)^2 = 1-\frac{1}{\kappa_2(A)^2} = \frac{\kappa_2(A)^2-1}{\kappa_2(A)^2}.$$

□

Método del descenso más rápido para las ecuaciones normales

Sea A una matriz regular arbitraria y los vectores $v = A^T r^{(0)}$ y $w = Av = AA^T r^{(0)}$, con subespacios $\mathcal{K} = span\left\{A^T r^{(0)}\right\}$ y $\mathcal{L} = A\mathcal{K}$. Nuevamente:

$$\langle Av, w\rangle_2 = \left\|AA^T r^{(0)}\right\|_2^2 \neq 0, \text{ si } r^{(0)} \neq 0.$$

Se obtiene entonces el siguiente método de proyección (1.9):

$$x^{(1)} = x^{(0)} + \frac{\langle A^T r^{(0)}, A^T r^{(0)}\rangle_2}{\|AA^T r^{(0)}\|_2^2} \cdot A^T r^{(0)} \text{ con } r^{(0)} = b - Ax^{(0)}.$$

En este caso, $r^{(1)} = r^{(0)} - \alpha \cdot AA^T r^{(0)}$, siendo $\alpha = \dfrac{\left\|A^T r^{(0)}\right\|_2^2}{\|AA^T r^{(0)}\|_2^2}$. Se obtiene entonces el siguiente algoritmo, que requiere dos productos matriz-vector por iteración. Dados $r^{(0)} = b - Ax^{(0)}$ y $\widehat{r}^{(0)} = A^T r^{(0)}$, calcular:

- $\boxed{d^{(0)} = A \cdot \widehat{r}^{(0)},}$ $\boxed{\alpha = \dfrac{\left\|\widehat{r}^{(0)}\right\|_2^2}{\|d^{(0)}\|_2^2},}$
- $\boxed{x^{(1)} = x^{(0)} + \alpha \cdot \widehat{r}^{(0)},}$ $\boxed{r^{(1)} = r^{(0)} - \alpha \cdot d^{(0)},}$ $\boxed{\widehat{r}^{(1)} = A^T r^{(1)}.}$ (1.14)

Observación 1.31.

$$\langle \widehat{r}^{(1)}, \widehat{r}^{(0)}\rangle_2 = \langle A^T r^{(1)}, A^T r^{(0)}\rangle_2 = \langle r^{(1)}, AA^T r^{(0)}\rangle_2 =$$
$$= \langle r^{(0)}, AA^T r^{(0)}\rangle_2 - \alpha \cdot \left\|AA^T r^{(0)}\right\|_2^2 = 0, \text{ por la definición de } \alpha.$$

Observación 1.32. Por la proposición 1.17, $x^{(1)}$ minimiza $R(x) = \|b - Ax\|_2^2$ sobre el espacio afín $x^{(0)} + span\{\widehat{r}^{(0)}\}$. Ahora bien: $R(x) = (b-Ax)^T \cdot (b-Ax)$ y $\nabla R(x) = -2A^T(b-Ax)$. Como $-\nabla R(x^{(0)}) = 2 \cdot \widehat{r}^{(0)}$, sigue que $x^{(1)}$ minimiza $R(x)$ sobre todos los vectores de la forma $x^{(0)} + \lambda\left(-\nabla R\left(x^{(0)}\right)\right)$, con $\lambda \in \mathbb{R}$. Así pues, este método coincide con el método del descenso más rápido aplicado al sistema de ecuaciones normales $A^T Ax = A^T b$. Esto se puede comprobar directamente, aplicando el algoritmo (1.11), cambiando b por $A^T b$ y A por $A^T A$.

Observación 1.33. Puesto que A^TA es simétrica y definida positiva, el teorema 1.21 da como corolario un teorema de convergencia para el algoritmo (1.14).

Teorema 1.34. *Sea A regular. Entonces la solución de avance del método del descenso más rápido para las ecuaciones normales verifica:*

$$\left\|b - Ax^{(1)}\right\|_2 \leq \frac{\kappa_2(A)^2 - 1}{\kappa_2(A)^2 + 1} \cdot \left\|b - Ax^{(0)}\right\|_2 .$$

Demostración. Aplicando el teorema 1.21, basta observar que:

$$\left\|x^{(i)} - x^*\right\|^2_{A^TA} = \langle A(x^{(i)} - x^*), A(x^{(i)} - x^*)\rangle_2 = \left\|b - Ax^{(i)}\right\|^2_2, \quad i = 0, 1,$$

y tener en cuenta la propiedad de que $\kappa_2(A^TA) = \left\|A^TA\right\|_2 \cdot \left\|(A^TA)^{-1}\right\|_2 = \kappa_2(A)^2$. □

2

Métodos de proyección basados en subespacios de Krylov

2.1. Subespacios de Krylov: definición y propiedades

Sea el sistema lineal $Ax = b$, con $A \in \mathbb{R}^{N\times N}$ regular y $b \in \mathbb{R}^N$. Denotemos $x^* = A^{-1}b$ y sea $x^{(0)} \in \mathbb{R}^N$ una aproximación dada a x^* con residual $r^{(0)} := b - Ax^{(0)}$. Asumimos que $r^{(0)} \neq 0$. Entonces, resolver $Ax = b$ equivale a resolver $Az = r^{(0)}$, con $z := x - x^{(0)}$. Así, definimos también $z^* := x^* - x^{(0)}$, de forma que $z^* = A^{-1}r^{(0)}$. Sea ahora $p \in \Pi_\gamma$, $1 \leq \gamma \leq N$, el polinomio mónico de menor grado γ tal que $p(A)\cdot r^{(0)} = 0$, que será de la forma $p(x) = \alpha_0 + \alpha_1 x + ... + \alpha_{\gamma-1}x^{\gamma-1} + x^\gamma$, con $\alpha_0 \neq 0$. Se debe observar que $\gamma \leq N$, pues toda matriz es anulada por su polinomio característico (Teorema de Caley-Hamilton). Por tanto,

$$A^{-1}r^{(0)} = \frac{-1}{\alpha_0}\left[\alpha_1 I + \alpha_2 A + ... + \alpha_{\gamma-1}A^{\gamma-2} + A^{\gamma-1}\right]\cdot r^{(0)},$$

y entonces $z^* \in span\left\{r^{(0)}, Ar^{(0)}, ..., A^{\gamma-1}r^{(0)}\right\}$ o bien $x^* \in x^{(0)} + span\left\{r^{(0)}, Ar^{(0)}, ..., A^{\gamma-1}r^{(0)}\right\}$, con $r^{(0)} = b - Ax^{(0)}$. Así, llegamos a que la solución x^* del sistema lineal $Ax = b$ se encuentra en el espacio afín $x^{(0)} + span\left\{r^{(0)}, Ar^{(0)}, ..., A^{\gamma-1}r^{(0)}\right\}$, cuyo espacio vectorial asociado será de especial interés a lo largo de este capítulo.

Definición 2.1. *Sea $A \in \mathbb{R}^{N\times N}$ una matriz y $r^{(0)} \in \mathbb{R}^N$ un vector. Definimos la sucesión de subespacios de Krylov asociados a A y $r^{(0)}$ como:*

$$\mathcal{K}_n(A, r^{(0)}) = span\left\{r^{(0)}, Ar^{(0)}, ..., A^{n-1}r^{(0)}\right\} \subset \mathbb{R}^N, \quad n = 0, 1, ... \tag{2.1}$$

Nótese que $\{0\} := \mathcal{K}_0(A, r^{(0)}) \subset \mathcal{K}_1(A, r^{(0)}) \subset \mathcal{K}_2(A, r^{(0)}) \subset ...$

Veamos ahora algunas propiedades que albergan los subespacios de Krylov.

Lema 2.2. *Sea $A \in \mathbb{R}^{N\times N}$ una matriz regular, $r^{(0)} \in \mathbb{R}^N$ un vector y $n \geq 1$ natural. Las siguientes afirmaciones son equivalentes:*

(a) $r^{(0)}, Ar^{(0)}, ..., A^n r^{(0)}$ son vectores linealmente dependientes.
(b) $\mathcal{K}_n(A, r^{(0)}) = \mathcal{K}_{n+1}(A, r^{(0)})$.
(c) $A\mathcal{K}_n(A, r^{(0)}) \subseteq \mathcal{K}_n(A, r^{(0)})$.
(d) $z^* := A^{-1}r^{(0)} \in \mathcal{K}_n(A, r^{(0)})$.

Demostración. (a)⇒(b): Por hipótesis, existen $\gamma_0, \gamma_1, ..., \gamma_{n-1} \in \mathbb{R}$ tales que $A^n r^{(0)} = \sum_{j=0}^{n-1} \gamma_j A^j r^{(0)}$. Por tanto, se tiene que $A^n r^{(0)} \in \mathcal{K}_n(A, r^{(0)})$ y $\mathcal{K}_{n+1}(A, r^{(0)}) \subseteq \mathcal{K}_n(A, r^{(0)})$, por lo que tales subespacios son iguales.
(b)⇒(c): Por definición de subespacios de Krylov y la hipótesis dada, obtenemos

$A\mathcal{K}_n(A, r^{(0)}) \subseteq \mathcal{K}_{n+1}(A, r^{(0)}) = \mathcal{K}_n(A, r^{(0)})$.
(c)⇒(d): Sea la aplicación lineal $\mathcal{L} : \mathcal{K}_n(A, r^{(0)}) \longrightarrow \mathcal{K}_n(A, r^{(0)})$ definida por $\mathcal{L}(v) = Av$, para todo $v \in \mathcal{K}_n(A, r^{(0)})$. Por hipótesis, $\mathcal{L}$ está bien definida y, además, es inyectiva por ser A una matriz regular. Por tanto, $dim(Im(\mathcal{L})) = dim(\mathcal{K}_n(A, r^{(0)}))$ y $\mathcal{L}$ es una aplicación biyectiva. Luego, existe un vector $v \in \mathcal{K}_n(A, r^{(0)})$ tal que $Av = r^{(0)}$. Asimismo, como A es una matriz regular, $v = z^*$.
(d)⇒(a): Sea $z^* \in \mathcal{K}_n(A, r^{(0)})$ la solución del sistema de ecuaciones $Ax = r^{(0)}$. Entonces, $r^{(0)} \in span\{Ar^{(0)}, ..., A^n r^{(0)}\}$ y $\{r^{(0)}, Ar^{(0)}, ..., A^n r^{(0)}\}$ son linealmente dependientes. □

Estas propiedades nos permiten describir el comportamiento de la sucesión de subespacios de Krylov, llegando a conocer la dimensión n^* del subespacio en el que se encuentra la solución única de $Ax = b$.

Corolario 2.3. *Sea $A \in \mathbb{R}^{N\times N}$ una matriz regular y $r^{(0)} \in \mathbb{R}^N$ un vector no nulo. Existe un único $n^* \in \mathbb{N}$, $1 \leq n^* \leq N$, de modo que*

$$\{0\} = \mathcal{K}_0(A, r^{(0)}) \subsetneq \mathcal{K}_1(A, r^{(0)}) \subsetneq ... \subsetneq \mathcal{K}_{n^*-1}(A, r^{(0)}) \subsetneq \mathcal{K}_{n^*}(A, r^{(0)}) = \mathcal{K}_{n^*+1}(A, r^{(0)}).$$

Además, $z^* = A^{-1}r^{(0)} \in \mathcal{K}_{n^*}(Ar^{(0)}) \setminus \mathcal{K}_{n^*-1}(A, r^{(0)})$.

Demostración. Es inmediata por el lema 2.2. □

Observación 2.4. Sea $v \in \mathbb{R}^N$ un autovector no nulo de la matriz $A \in \mathbb{R}^{N\times N}$ y $r^{(0)} = kv$, donde $k \in \mathbb{R}\backslash\{0\}$. Entonces, $\mathcal{K}_1(A, r^{(0)}) = \mathcal{K}_2(A, r^{(0)})$. En efecto, $\mathcal{K}_1(A, r^{(0)}) = span\{r^{(0)}\} = span\{r^{(0)}, Ar^{(0)}\} = \mathcal{K}_2(A, r^{(0)})$. Esta propiedad se puede generalizar para dar una interpretación algebraica del número de iteraciones $n^* \leq N$ necesarias para la estabilización de la sucesión de subespacios de Krylov.

Definición 2.5. *Sea $A \in \mathbb{R}^{N\times N}$ una matriz regular y $v \in \mathbb{R}^N$ un vector. Se llama polinomio mínimo de v respecto de A al polinomio mónico $p \in \Pi_\gamma$ de menor grado $\gamma \geq 0$ tal que:*

$$p(A) \cdot r^{(0)} = 0. \tag{2.2}$$

Denominamos a γ como grado de v respecto de A.

Observación 2.6. Por el Teorema de Cayley-Hamilton, es claro que $\gamma \leq N$. Observar además que el polinomio (2.2) es único, pues de existir otro $q \in \Pi_\gamma$, $q \neq p$, polinomio mónico tal que $q(A)r^{(0)} = 0$, entonces $p - q \in \Pi_{\gamma-1}$ y $(p - q)(A)r^{(0)} = 0$. Luego, si $\alpha \neq 0$ es el coeficiente director de $p - q$, el polinomio $r = \dfrac{p-q}{\alpha}$ es mónico, verifica que $r(A)r^{(0)} = 0$ y su grado es menor que γ, lo cual es un absurdo.

Teorema 2.7. *Sea $A \in \mathbb{R}^{N\times N}$ una matriz regular y $r^{(0)} \in \mathbb{R}^N$ un vector no nulo. Sea $\gamma \geq 1$ el grado de $r^{(0)}$ respecto de A. Entonces, $\mathcal{K}_0(A, r^{(0)}) \subsetneq \mathcal{K}_1(A, r^{(0)}) \subsetneq \cdots \subsetneq \mathcal{K}_\gamma(A, r^{(0)}) = \mathcal{K}_{\gamma+1}(A, r^{(0)})$. En particular, $z^* = A^{-1}r^{(0)} \in \mathcal{K}_\gamma(A, r^{(0)})$.*

En otras palabras, $\gamma = n^*$ en (2.3).

Demostración. Observamos que $\gamma \geq 1$ es el menor número natural tal que $A^\gamma r^{(0)}$ es combinación lineal de $r^{(0)}, \ldots, A^{\gamma-1}r^{(0)}$. Equivalentemente, por el lema 2.2, γ es el menor número natural tal que $\mathcal{K}_\gamma(A, r^{(0)}) = \mathcal{K}_{\gamma+1}(A, r^{(0)})$, esto es, $\mathcal{K}_0(A, r^{(0)}) \subsetneq \mathcal{K}_1(A, r^{(0)}) \subsetneq \cdots \subsetneq \mathcal{K}_\gamma(A, r^{(0)}) = \mathcal{K}_{\gamma+1}(A, r^{(0)}) = \ldots$. El mismo lema permite asegurar que $z^* \in \mathcal{K}_\gamma(A, r^{(0)})\backslash\mathcal{K}_{\gamma-1}(A, r^{(0)})$.

Observación 2.8. A lo largo de este capítulo, nos centraremos en la resolución de sistemas lineales $Ax = b$ mediante los métodos de proyección del Gradiente Conjugado(GC), Gradiente Conjugado para ecuaciones normales(GCNR) y del residual mínimo generalizado(GMRES), que son métodos de proyección basados en los subespacios de Krylov introducidos en la definición 2.1. Ahora bien, de cara al algoritmo GMRES que se detallará en la sección 2.3, nos será de gran ayuda conocer un algoritmo que genere una base ortogonal para $\mathcal{K}_n(A, r^{(0)})$: **el proceso de Arnoldi**.

Dado un vector $v_1 \in \mathbb{R}^N$ con $\|v_1\|_2 = 1$, el proceso de Arnoldi consiste en generar una sucesión de vectores $v_1, v_2, ...$, con $\|v_i\|_2 = 1$, $\forall i = 1, 2, ...$, ortogonales con respecto al producto escalar $\langle \cdot, \cdot \rangle_2$ mediante una variante del proceso de Gram-Schmidt de la siguiente forma:
Definimos $v_1 = \dfrac{r^{(0)}}{\|r^{(0)}\|_2}$ y consideramos el sistema $\{v_1, Av_1\}$, sobre el que aplicamos el proceso de Gram-Schmidt, obteniendo

$$\widehat{v_2} = Av_1 - \langle Av_1, v_1 \rangle_2 v_1.$$

Si $\widehat{v_2} \neq 0$, definimos $v_2 := \dfrac{\widehat{v_2}}{\|v_2\|_2}$ y obtenemos el sistema ortonormal $\{v_1, v_2\}$. Seguidamente, aplicamos el proceso de Gram-Schmidt a $\{v_1, v_2, Av_2\}$, obteniendo

$$\widehat{v_3} = Av_2 - \langle Av_2, v_1 \rangle_2 v_1 - \langle Av_2, v_2 \rangle_2 v_2.$$

Si $\widehat{v_3} \neq 0$, podemos definir $v_3 := \dfrac{\widehat{v_3}}{\|\widehat{v_3}\|_2}$, resultando un nuevo sistema ortogonal $\{v_1, v_2, v_3\}$. Repitiendo recursivamente este proceso, podemos definirlo en el siguiente algoritmo:

Algoritmo del proceso de Arnoldi
Dado $r^{(0)} \in \mathbb{R}^N, r^{(0)} \neq 0$, definimos $\widehat{v}_1 := r^{(0)}$. Para $n \geq 1$, mientras $\widehat{v}_n \neq 0$, calcular:

- $v_n = \dfrac{\widehat{v}_n}{\|\widehat{v}_n\|_2}$.
- $\widehat{v}_{n+1} = Av_n - \displaystyle\sum_{j=1}^{n} \langle Av_n, v_j \rangle_2 v_j$.

Observación 2.9. Nótese que el cálculo de $\widehat{v}_{n+1}$ requiere tener almacenados los vectores $v_1, v_2, ..., v_n$, así como una multiplicación Matriz-Vector.

Observación 2.10. El proceso de Arnoldi parará al llegar a la primera iteración tal que $Av_n \in span\{v_1, v_2, ..., v_n\}$, es decir, cuando $\widehat{v}_{n+1} = 0$ en el método Gram-Schmidt.

Observación 2.11. Si A es una matriz simétrica, entonces, para $k \leq n-2$, $\langle Av_n, v_k \rangle_2 = \langle v_n, Av_k \rangle_2 = 0$, ya que los vectores v_n y v_j son ortogonales para $j = 0, ..., n-1$ y Av_k es combinación lineal de $v_1, v_2, ..., v_k, v_{k+1}$. Luego, el proceso de Arnoldi se convierte en una recursión a tres términos:

$$\widehat{v}_{n+1} := Av_n - \langle Av_n, v_n \rangle_2 v_n - \langle Av_n, v_{n-1} \rangle_2 v_{n-1}, \qquad n = 1, 2, ... \tag{2.3}$$

Este caso especial de aplicar el proceso de Arnoldi a matrices simétricas se conoce como el proceso de Lanczos.

Veamos ahora un lema que nos permite afirmar que el proceso de Arnoldi genera una base ortonormal de los subespacios de Krylov $\mathcal{K}_n(A, r^{(0)})$.

Lema 2.12. *Los vectores $v_1, v_2, ..., v_{n^*}$ generados por el proceso de Arnoldi forman un sistema ortonormal respecto del producto euclídeo $\langle ., . \rangle_2$ en $\mathbb{R}^N$ y*

$$span\{v_1, ..., v_n\} = span\{v_1, ..., v_{n-1}, Av_{n-1}\} = \mathcal{K}_n(A, r^{(0)}),\ para\ 1 \leq n \leq n^*.$$

Demostración. La propia construcción de los vectores $v_1, v_2, ..., v_{n^*}$ conlleva inmediantamente la ortonormalidad de los mismos, ya que es evidente que $\|v_j\|_2 = 1$ para $1 \leq j \leq n^*$ y, asumiendo que $\langle v_j, v_k \rangle_2 = \delta_{jk}$ para $1 \leq j < k \leq n$, se tiene que, dado k tal que $1 \leq k \leq n$

$$\langle v_{n+1}, v_k \rangle_2 = \frac{1}{\|\hat{v}_{n+1}\|_2} \left(\langle Av_n, v_k \rangle_2 - \sum_{j=1}^{n} \langle Av_n, v_j \rangle_2 \langle v_j, v_k \rangle_2 \right) \overset{j=k}{=} 0$$

Aplicando inducción, comprobemos que

$$span\{v_1, ..., v_n\} = span\{v_1, ..., v_{n-1}, Av_{n-1}\} = \mathcal{K}_n(A, r^{(0)}), \quad \text{para } 1 \leq n \leq n^*$$

Como $v_1 = \dfrac{r^{(0)}}{\|r^{(0)}\|_2}$, la propiedad resulta trivial para $n = 1$. Ahora, supongamos que se cumple para un n con $1 \leq n \leq n^* - 1$, y veamos que se verifica para $n + 1$. De esta forma, escribimos

$$v_{n+1} = \frac{1}{\|\hat{v}_{n+1}\|_2} \left(Av_n - \sum_{j=1}^{n} \langle Av_n, v_j \rangle_2 v_j \right) \in span\left\{v_1, ..., v_n, Av_n\right\}.$$
$$\Rightarrow span\left\{v_1, ..., v_{n+1}\right\} \subseteq span\left\{v_1, ..., v_n, Av_n\right\} \subseteq \mathcal{K}_{n+1}.$$

Ahora bien, como $dim\left(\{v_1, ..., v_{n+1}\}\right) = n + 1$, $dim\left(span\left\{v_1, ..., v_n, Av_n\right\}\right) \leq n + 1$ y $dim(\mathcal{K}_{n+1}) \leq n + 1$, por un argumento de dimensión, los tres subespacios son iguales. □

Observación 2.13. Este lema nos permite afirmar que el número de iteraciones n^* necesarias para que el proceso de Arnoldi se detenga coincide con el menor valor n^* tal que $\mathcal{K}_{n^*}(A, r^{(0)}) = \mathcal{K}_{n^*+1}(A, r^{(0)})$ (véase Corolario 2.3).

2.2. El método del Gradiente Conjugado

Sea el sistema lineal $Ax = b$, $b \in \mathbb{R}^N$, $A \in \mathbb{R}^{N \times N}$ simétrica y definida positiva. Sea además $x^{(0)} \in \mathbb{R}^N$ una aproximación inicial fija de la solución $x^* = A^{-1}b$, con residual $r^{(0)} = b - Ax^{(0)} \neq 0$.

Definición 2.14. *Para cada $n \geq 1$, el método del Gradiente Conjugado produce una aproximación $x^{(n)} \in \mathbb{R}^N$ a través del método de proyección ortogonal (1.1) sobre $\mathcal{K} = \mathcal{K}_n(A, r^{(0)})$, siendo $\mathcal{K}_n(A, r^{(0)}) = span\left\{r^{(0)}, Ar^{(0)}, ..., A^{n-1}r^{(0)}\right\}$ el enésimo subespacio de Krylov generado por A y $r^{(0)}$.*

Así pues,

$$x^{(n)} = x^{(0)} + \delta^{(n)},\ \text{con } \delta^{(n)} \in \mathcal{K}_n(A, r^{(0)}) \text{ tal que} \tag{2.4}$$
$$r^{(n)} \in \mathcal{K}_n(A, r^{(0)})^{\perp},\ \text{siendo } r^{(n)} = b - Ax^{(n)} = r^{(0)} - A\delta^{(n)}.$$

Teniendo en cuenta el Lema 2.2, el corolario 2.3 y la proposición 1.14, observamos:

(I) $x^{(n)} = x^*$, $\forall n \geq n^*$, pues $A\mathcal{K}_n \subseteq \mathcal{K}_n$, $\forall n \geq n^*$.
(II) $x^{(n)} \neq x^*$, $\forall n < n^*$, pues $z^* = x^* - x^{(0)} \notin \mathcal{K}_n(A, r^{(0)})$ si $n < n^*$.

Es decir, el método del Gradiente Conjugado converge a la solución exacta del sistema lineal $Ax = b$ en exactamente n^* iteraciones.

Por otra parte, según la proposición 1.17 del capítulo 1, $x^{(n)}$ minimiza $E(x) = \|x^* - x\|_A^2$ sobre el subespacio $x^{(0)} + \mathcal{K}_n(A, r^{(0)})$. Puesto que $\mathcal{K}_{n-1}(A, r^{(0)}) \subsetneq \mathcal{K}_n(A, r^{(0)})$ para todo n, $1 \leq n \leq n^*$, sigue que $\left\|x^* - x^{(n)}\right\|_A \leq \left\|x^* - x^{(n-1)}\right\|_A$, $1 \leq n \leq n^*$.

Observación 2.15. Observamos que, fijada una base $\{v_1, ..., v_n\}$ de $\mathcal{K}_n(A, r^{(0)})$, $1 \leq n \leq n^*$, con matriz asociada $V \in \mathbb{R}^{N \times n}$, entonces la condición $r^{(0)} \in \mathcal{K}_n(A, r^{(0)})^{\perp}$ se traduce en que $V^T \left(r^{(0)} - A\delta^{(n)}\right) = 0$, siendo $\delta^{(n)} := V \cdot y^{(n)} \in \mathcal{K}_n(A, r^{(0)})$ y $y^{(n)} \in \mathbb{R}^n$.

Como $V^T AV$ es regular, del lema 1.4 sigue que:

$$x^{(n)} = x^{(0)} + \delta^{(n)} = x^{(0)} + V \left(V^T AV\right)^{-1} \cdot V^T r^{(0)} \tag{2.5}$$

y, además:

$$r^{(n)} = r^{(0)} - A\delta^{(n)} = \left[I - AV \left(V^T AV\right)^{-1} \cdot V^T\right] \cdot r^{(0)}. \tag{2.6}$$

Así, la elección de la base $\{v_1, v_2, ..., v_n\}$ juega un papel importante en la implementación eficiente del método GC.

A continuación, consideramos tres bases alternativas para $\mathcal{K}_n(A, r^{(0)})$ que dan lugar a tres implementaciones distintas del método GC.

1°)Consideremos directamente la base $\{r^{(0)}, Ar^{(0)}, ..., A^{n-1}r^{(0)}\}$ $(1 \leq n \leq n^*)$. Entonces $x^{(n)} = x^{(0)} + \sum_{j=0}^{n-1} \alpha_j \cdot A^j r^{(0)}$, $\alpha_0, ..., \alpha_{n-1} \in \mathbb{R}$ Además, $r^{(n)} = b - Ax^{(n)} = r^{(0)} - A(x^{(n)} - x^{(0)}) = r^{(0)} - \sum_{j=0}^{n-1} \alpha_j \cdot A^{j+1} r^{(0)}$. Es decir:

- $x^{(n)} = x^{(0)} + q_n(A)r^{(0)}$, con $q_n(t) := \sum_{j=0}^{n-1} \alpha_j \cdot t^j \in \Pi_{n-1}$.
- $r^{(n)} = p_n(A)r^{(0)}$, con $p_n(t) := 1 - t \cdot q_n(t)$.

Veamos cómo determinar los coeficientes $\alpha_0, \alpha_1, ..., \alpha_{n-1}$. Como

$$\left\|z^* - z^{(n)}\right\|_A = \min_{z \in \mathcal{K}_n(A, r^{(0)})} \|z^* - z\|_A, \text{ siendo } z^* = x^* - x^{(0)}, z^{(n)} = x^{(n)} - x^{(0)}, \tag{2.7}$$

$z^{(n)}$ es la mejor aproximación de z^* en $\mathcal{K}_n(A, r^{(0)})$ respecto de la norma energía $\|.\|_A$. Planteando el sistema de ecuaciones normales del problema de Mínimos Cuadrados, con $z^{(n)} = \sum_{j=0}^{n-1} \alpha_j \cdot A^j r^{(0)}$ respecto de la base $\{r^{(0)}, ..., A^{n-1}r^{(0)}\}$ sigue que $G \cdot c = F$, con $c = (\alpha_0, ..., \alpha_{n-1})^T, F = \left(\langle z^*, A^{j-1}r^{(0)}\rangle_A\right)_{j=1}^n$ y $G = (\langle A^{i-1}r^{(0)}, A^{j-1}r^{(0)}\rangle_A)_{i,j=1}^n$. Como $\langle A^{i-1}r^{(0)}, A^{j-1}r^{(0)}\rangle_A = {r^{(0)}}^T \cdot A^{i+j-1}r^{(0)}$, $1 \leq i, j \leq n$, y $\langle z^*, A^{j-1}r^{(0)}\rangle_A = \left(r^{(0)}\right)^T \cdot A^{j-1} \cdot Az^* = {r^{(0)}}^T A^{j-1}r^{(0)}$, $1 \leq j \leq n$, $\alpha_0, ..., \alpha_{n-1}$ se pueden obtener resolviendo el

siguiente sistema lineal de dimensión n:

$$\begin{bmatrix} {r^{(0)}}^T A r^{(0)} & {r^{(0)}}^T A^2 r^{(0)} & \cdots & {r^{(0)}}^T A^n r^{(0)} \\ {r^{(0)}}^T A^2 r^{(0)} & {r^{(0)}}^T A^3 r^{(0)} & \cdots & {r^{(0)}}^T A^{n+1} r^{(0)} \\ \vdots & \vdots & \ddots & \vdots \\ {r^{(0)}}^T A^n r^{(0)} & {r^{(0)}}^T A^{n+1} r^{(0)} & \cdots & {r^{(0)}}^T A^{2n-1} r^{(0)} \end{bmatrix} \begin{bmatrix} \alpha_0 \\ \alpha_1 \\ \vdots \\ \alpha_{n-1} \end{bmatrix} = \begin{bmatrix} {r^{(0)}}^T r^{(0)} \\ {r^{(0)}}^T A r^{(0)} \\ \vdots \\ {r^{(0)}}^T A^{n-1} r^{(0)} \end{bmatrix}. \quad (2.8)$$

2º) Si se considera la base $\{v_1, ..., v_n\}$ de $\mathcal{K}_n(A, r^{(0)})$ generada por el proceso de Arnoldi (Lanczos, 2.3), con matriz asociada $V = [v_1|v_2|\dots|v_n]$, entonces:

- $V_n \cdot e_1 = v_1 = \dfrac{r^{(0)}}{\|r^{(0)}\|_2}$ y $V_n^T r^{(0)} = \left\|r^{(0)}\right\|_2 \cdot e_1 \in \mathbb{R}^n$
- Si se premultiplica (2.3) por v_j^T, $1 \leq j \leq n$, se obtiene que $V_n^T A V_n = T_n$, para cierta matriz regular, tridiagonal y simétrica.

Por tanto, usando (2.5)-(2.6)

- $x^{(n)} = x^{(0)} + V_n T_n^{-1} \left(\left\|r^{(0)}\right\|_2 e_1\right)$,
- $r^{(n)} = r^{(0)} - A V_n T_n^{-1} \left(\left\|r^{(0)}\right\|_2 e_1\right), \quad 1 \leq n \leq n^*$.

Estas fórmulas se pueden implementar considerando descomposición LU para la matriz T_n en cada iteración, para dar lugar a la implementación del método del Gradiente Conjugado basada en el algoritmo de Lanczos (para más detalles, ver [8, Pág, 187-192]). No obstante, nos centraremos a continuación en un algoritmo más eficiente que surge al considerar una base de $\mathcal{K}_n(A, r^{(0)})$ formada por vectores A-conjugados, esto es, vectores ortogonales respecto de $\langle ., . \rangle_A$.

3º) Observemos que, si dispusiéramos de una base $\{d^{(0)}, d^{(1)}, ..., d^{(n-1)}\}$ de $\mathcal{K}_n(A, r^{(0)})$ ortogonal respecto del $\langle ., . \rangle_A$, entonces, el problema de mínimos cuadrados con solución $z^{(n)} = \sum_{j=0}^{n-1} \alpha_j \cdot d^{(j)}$ se reduciría a resolver el sistema lineal de ecuaciones normales (2.8) con matriz diagonal regular $G = Diag\left(\left\|d^{(i)}\right\|_A^2\right)_{i=0}^{n-1}$ y $F = \left(\langle z^*, d^{(i)} \rangle_A\right)_{i=0}^{n-1}$. En particular,

$$\alpha_i = \frac{\langle z^*, d^{(i)} \rangle_A}{\|d^{(i)}\|_A^2} = \frac{\langle r^{(0)}, d^{(i)} \rangle_2}{\langle A d^{(i)}, d^{(i)} \rangle_2}, \; i = 0, ..., n-1, \quad (2.9)$$

y $x^{(n)} = x^{(0)} + \sum_{i=0}^{n-1} \dfrac{\langle r^{(0)}, d^{(i)} \rangle_2}{\langle A d^{(i)}, d^{(i)} \rangle_2} \cdot d^{(i)}$. Además, $x^{(n+1)} = x^{(n)} + \alpha_n \cdot d^{(n)}$ y $r^{(n+1)} = b - A x^{(n+1)} = b - A x^{(n)} - \alpha_n A d^{(n)} = r^{(n)} - \alpha_n A d^{(n)}$.
Respecto al numerador del coeficiente α_i (2.9), podemos observar que, dado $r^{(i)} = r^{(0)} - \sum_{j=0}^{i-1} \alpha_j \cdot A d^{(j)}$, y teniendo en cuenta la A-ortogonalidad de la base $\{d^{(0)}, ..., d^{(n-1)}\}$,

$$\langle r^{(0)}, d^{(i)} \rangle_2 = \langle r^{(i)}, d^{(i)} \rangle_2 + \sum_{j=0}^{i-1} \alpha_j \underbrace{\langle A d^{(j)}, d^{(i)} \rangle_2}_{\langle d^{(j)}, d^{(i)} \rangle_A} = \langle r^{(i)}, d^{(i)} \rangle_2,$$

Por tanto,

$$\alpha_i = \frac{\langle r^{(i)}, d^{(i)}\rangle_2}{\langle Ad^{(i)}, d^{(i)}\rangle_2}. \tag{2.10}$$

El siguiente teorema indica cómo construir una base ortogonal respecto de $\langle\cdot,\cdot\rangle_A$ para $\mathcal{K}_n(A, r^{(0)})$.

Teorema 2.16. *Sea $A \in \mathbb{R}^{N\times N}$ una matriz simétrica y definida positiva, $r^{(0)}\in\mathbb{R}^N$ un vector no nulo y $x^{(n)}\in \mathbb{R}^N$ la n-ésima iteración del método del Gradiente Conjugado, es decir, el único vector solución de*

$$x^{(n)} \in x^{(0)} + \mathcal{K}_n(A, r^{(0)}), \quad r^{(n)} = b - Ax^{(n)} \in \mathcal{K}_n(A, r^{(0)})^\perp,\ 0 \le n \le n*,$$

siendo $n^* \ge 1$ el menor número natural tal que $r^{(n^*)} = 0$. Sean $d^{(0)} := r^{(0)}$, $d^{(n)} := r^{(n)} + \beta_{n-1}d^{(n-1)}$ con $\beta_{n-1} := \dfrac{-\langle r^{(n)}, d^{(n-1)}\rangle_A}{\|d^{(n-1)}\|_A^2}$. Entonces:

(I) $d^{(0)}, \ldots, d^{(n-1)}$ son ortogonales respecto de $\langle\cdot,\cdot\rangle_A$.
(II) $\mathcal{K}_n(A, r^{(0)}) = span\{d^{(0)}, \ldots, d^{(n-1)}\} = span\{r^{(0)}, \ldots, r^{(n)}\}$ para $1 \le n \le n^*$.

En particular, $\langle r^{(n)}, r^{(j)}\rangle_2 = 0$, para $0 \le j \le n-1$ y $1 \le n \le n^* - 1$.

Demostración. En primer lugar, observamos que $r^{(n)} \neq 0$ para $0 \le n \le n^* - 1$. Luego, por el lema 2.2 y corolario 2.3, se sigue que:

$$dim(\mathcal{K}_n) = n \text{ para } 0 \le n \le n^* \text{ y } \mathcal{K}_0 \subsetneq \mathcal{K}_1 \subsetneq \cdots \mathcal{K}_{n^*} = \mathcal{K}_{n*+1}$$

Ahora, demostraremos los enunciados (I) y (II) por inducción sobre $n = 1, 2, \ldots, n^*$. Si $n = 1$, dado que $d^{(0)} = r^{(0)}$, es obvio que $\mathcal{K}_1(A, r^{(0)}) = span\{d^{(0)}\} = span\{r^{(0)}\}$. Supongamos cierto para $1 \le n \le n^*-1$ que los vectores $d^{(0)}, \ldots, d^{(n-1)}$ son ortogonales respecto de $\langle\cdot,\cdot\rangle_A$ y que $\mathcal{K}_n(A, r^{(0)}) = span\{d^{(0)}, \ldots, d^{(n-1)}\} = span\{r^{(0)}, \ldots, r^{(n-1)}\}$. Veamos que se cumplen las propiedades para $n + 1 \le n^*$:
Como $r^{(n)} \neq 0$ y $r^{(n)} \in \mathcal{K}_n(A, r^{(0)})^\perp$, entonces $d^{(0)}, \cdots, d^{(n-1)}, r^{(n)}$ son linealmente independientes. Por ello, el proceso de ortogonalización de Gram-Schmidt respecto de $\langle\cdot,\cdot\rangle_A$ permite afirmar que el vector

$$v_n := r^{(n)} - \sum_{j=0}^{n-1} \frac{\langle r^{(n)}, d^{(j)}\rangle_A}{\langle d^{(j)}, d^{(j)}\rangle_A} \cdot d^{(j)} \tag{2.11}$$

es ortogonal a $d^{(0)}, \cdots, d^{(n-1)}$ respecto de $\langle .,.\rangle_A$. Ahora bien, si $0 \le j \le n-2$, entonces $Ad^{(j)} \in A\mathcal{K}_{n-1}(A, r^{(0)}) \subset \mathcal{K}_n(A, r^{(0)})$. Como $r^{(n)} \in \mathcal{K}_n(A, r^{(0)})^\perp$, entonces

$$\langle r^{(n)}, d^{(j)}\rangle_A = \langle r^{(n)}, Ad^{(j)}\rangle_2 = 0 \text{ para } 0 \le j \le n-2. \tag{2.12}$$

Luego, sustituyendo (2.12) en (2.11), tenemos que

$$v_n = r^{(n)} - \frac{\langle r^{(n)}, d^{(n-1)}\rangle_A}{\langle d^{(n-1)}, d^{(n-1)}\rangle_A} \cdot d^{(n-1)} = r^{(n)} + \beta_{n-1}d^{(n-1)} = d^{(n)}.$$

Por tanto, $d^{(0)}, \ldots, d^{(n)}$ son ortogonales respecto de $\langle\cdot,\cdot\rangle_A$. Además, $d^{(n)} \neq 0$ ya que, si $d^{(n)} = 0$, entonces $r^{(n)} = -\beta_{n-1}d^{(n-1)} \in \mathcal{K}_n(A, r^{(0)})$ y, como $r^{(n)} \in \mathcal{K}_n(A, r^{(0)})^\perp$, se tendría que $r^{(n)} = 0$, lo cual es un absurdo.
Por hipótesis de inducción, como $r^{(n)} = d^{(n)} - \beta_{n-1}d^{(n-1)}$, se obtiene:

$$span\{r^{(0)}, \cdots, r^{(n-1)}, r^{(n)}\} = span\{d^{(0)}, \cdots, d^{(n-1)}, r^{(n)}\} = span\{d^{(0)}, d^{(1)} \cdots, d^{(n)}\}.$$

Además, como $A\mathcal{K}_n \subset \mathcal{K}_{n+1}$, y $r^{(0)}, \ldots, r^{(n-1)} \in \mathcal{K}_n \subset \mathcal{K}_{n+1}$, entonces, $r^{(n)} = r^{(0)} - A(x^{(n)} - x^{(0)}) \in \mathcal{K}_{n+1}$.
Por tanto, $span\{r^{(0)}, \ldots, r^{(n-1)}, r^{(n)}\} \subseteq \mathcal{K}_{n+1}$ siendo $dim(\mathcal{K}_{n+1}) = n+1$. Considerando que $n+1 \leq n^*$, tenemos que:

$$dim(span\{r^{(0)}, \ldots, r^{(n-1)}, r^{(n)}\}) = dim(span\{d^{(0)}, \cdots, d^{(n-1)}, d^{(n)}\}) = n+1$$

ya que $d^{(0)}, \ldots, d^{(n-1)}, d^{(n)}$ son linealmente independientes. En definitiva,

$$\mathcal{K}_{n+1} = span\{r^{(0)}, \ldots, r^{(n-1)}, r^{(n)}\} = span\{d^{(0)}, \ldots, d^{n-1}, d^{(n)}\}$$

□

Observación 2.17. Nótese que $r^{(n)} \in \mathcal{K}_n^{\perp}$ y $\mathcal{K}_n = span\{r^{(0)}, \ldots, r^{(n-1)}\}$ implica que

$$\langle r^{(n)}, r^{(j)} \rangle_2 = 0 \text{ para } 0 \leq j \leq n-1.$$

Teorema 2.18. *Usando la notación del teorema 2.16, se cumple que:*

$$\alpha_n = \frac{\left\| r^{(n)} \right\|_2^2}{\langle Ad^{(n)}, d^{(n)} \rangle_2},\ n = 0, 1, \cdots, n^* - 1 \text{ y } \beta_{n-1} = \frac{\left\| r^{(n)} \right\|_2^2}{\|r^{(n-1)}\|_2^2},\ n = 1, ..., n^* - 1.$$

Demostración. En primer lugar, notamos que $\langle r^{(n)}, d^{(n)} \rangle_2 = \langle r^{(n)}, r^{(n)} - \beta_{n-1} d^{(n-1)} \rangle_2 = \left\| r^{(n)} \right\|_2^2$, pues $d^{(n-1)} \in \mathcal{K}_n(A, r^{(0)})$ y $r^{(n)} \in \mathcal{K}_n(A, r^{(0)})^{\perp}$. Considerando (2.10), esto demuestra la expresión para α_n en la tesis.
Por otra parte, $\langle r^{(n)}, r^{(n)} \rangle_2 = \langle r^{(n)}, r^{(n-1)} - \alpha_{n-1} Ad^{(n-1)} \rangle_2 = -\alpha_{n-1} \langle r^{(n)}, Ad^{(n-1)} \rangle_2$, en virtud de la observación 2.4. Usando la expresión recién demostrada para α_{n+1} y el teorema 2.16:

$$\left\| r^{(n)} \right\|_2^2 = \frac{-\left\| r^{(n-1)} \right\|_2^2}{\langle Ad^{(n-1)}, d^{(n-1)} \rangle_2} \cdot \langle r^{(n)}, Ad^{(n-1)} \rangle_2 = \left\| r^{(n-1)} \right\|_2^2 \beta_{n-1}.$$

□

Compaginando los resultados de los teoremas previos, obtenemos el siguiente **Algoritmo para el método del Gradiente Conjugado:**
Sea $x^{(0)} \in \mathbb{R}^N$, con residual $r^{(0)} = b - Ax^{(0)}$,

1. Definimos $\boxed{d^{(0)} := r^{(0)}}$
2. Dados $\boxed{x^{(n)}, r^{(n)} \text{ y } d^{(n)}, n \geq 0,}$ mientras $r^{(n)} \neq 0$, calcular:
   - $\boxed{x^{(n+1)} = x^{(n)} + \alpha_n d^{(n)},}$ con $\alpha_n = \dfrac{\left\| r^{(n)} \right\|_2^2}{\langle Ad^{(n)}, d^{(n)} \rangle_2}$
   - $\boxed{r^{(n+1)} = r^{(n)} - \alpha_n Ad^{(n)}.}$
   - $\boxed{d^{(n+1)} = r^{(n+1)} + \beta_n d^{(n)},}$ con $\beta_n = \dfrac{\left\| r^{(n+1)} \right\|_2^2}{\|r^{(n)}\|_2^2}.$

Observación 2.19. Sea $A \in \mathbb{R}^{N \times N}$ una matriz simétrica, definida positiva y $r^{(0)} \in \mathbb{R}^N$ un vector no nulo. Observamos que, para todo n natural, $r^{(n)} = r^{(0)} - Ax^{(n)}$ coincide con el vector opuesto al gradiente de la función energía $\mathcal{E}(x) := \frac{1}{2}\langle Ax, x \rangle_2 - \langle x, r^{(0)} \rangle_2$ evaluado en $x^{(n)}$. Esto es, $r^{(n)} = -\nabla \mathcal{E}(x^{(n)})$.

Nótese que esta última observación 2.19, junto con la 2.17, son las que dan nombre al método del Gradiente Conjugado.

Según el Teorema 2.7, el método del Gradiente Conjugado podrá ser interpretado como un método directo, puesto que siempre se obtendrá la solución exacta $x^{(n^*)} = x^*$ para $Ax = b$ después de un número finito de pasos si se trabaja en aritmética infinita. No obstante, dicho número de iteraciones puede ser considerablemente grande. Por esta razón, será de interés calcular estimaciones para el error.

Lema 2.20. *Sea $A \in \mathbb{R}^{N\times N}$ una matriz simétrica y definida positiva, con autovalores $\{\lambda_j\}_{j=1}^N$ y autovectores correspondientes ortonormales $\{v_j\}_{j=1}^N \subset \mathbb{R}^N$. Si $x = \sum_{j=1}^N c_j v_j \in \mathbb{R}^N$, $c_j \in \mathbb{R}$, entonces, para todo polinomio p:*

$$\|p(A)x\|_2 = \left(\sum_{j=1}^N c_j^2 p(\lambda_j)^2\right)^{1/2}, \quad \|p(A)x\|_A \left(\sum_{j=1}^N c_j^2 \lambda_j p(\lambda_j)^2\right)^{1/2}. \tag{2.13}$$

En particular, se cumple que

$$\sqrt{m}\,\|x\|_2 \le \|x\|_A \le \sqrt{M}\,\|x\|_2, \qquad x \in \mathbb{R}^N \tag{2.14}$$

donde $m := \min_{j=1,\dots,N} \lambda_j$ y $M := \max_{j=1,\dots,N} \lambda_j$.

Demostración. Puesto que $A^\gamma x = \sum_{j=1}^N c_j \lambda_j^\gamma v_j$, $\gamma = 0, 1, \dots$, y dado que $\{v_j\}_{j=1}^N \subset \mathbb{R}^N$ es un sistema ortogonal respecto de $\langle\cdot,\cdot\rangle_2$, se tiene:

$$\|p(A)x\|_2 = \langle \sum_{k=1}^N c_k p(\lambda_k) v_k, \sum_{j=1}^N c_j p(\lambda_j) v_j\rangle_2^{1/2} = \left(\sum_{j=1}^N c_j^2 p(\lambda_j)^2\right)^{1/2},$$

teniendo en cuenta que, si λ_j, $j = 1, \dots, N$, son los autovalores de A, entonces $p(\lambda_j)$ son los autovalores de $p(A)$ y, además, tienen los mismos autovectores asociados. De manera análoga,

$$\|p(A)x\|_A = \langle \sum_{k=1}^N c_k \lambda_k p(\lambda_k) v_k, \sum_{j=1}^N c_j p(\lambda_j) v_j\rangle_2^{1/2} = \left(\sum_{j=1}^N c_j^2 \lambda_j p(\lambda_j)^2\right)^{1/2}$$

Finalmente, con $p(t) = 1$ en (2.13) sigue que $\|x\|_2 = \left(\sum_{j=1}^N c_j^2\right)^{1/2}$ y $\|x\|_A = \left(\sum_{j=1}^N c_j^2 \lambda_j\right)^{1/2}$. De aquí se obtiene (2.14) dado que $\min_{j=1,\dots,N} \lambda_j \le \lambda_j \le \max_{j=1,\dots,N} \lambda_j$, para $j = 1, \dots, N$. □

El siguiente teorema da una primera estimación de convergencia del método del Gradiente Conjugado.

Teorema 2.21. *Sea $A \in \mathbb{R}^{N\times N}$ matriz simétrica y definida positiva. Si $\{x^{(n)}\}_{n=0}^{n^*}$ son las iteraciones del método del Gradiente Conjugado, entonces*

$$\left\|x^* - x^{(n)}\right\|_A \le \left(\inf_{\substack{p\in\prod_n \\ p(0)=1}} \max_{\lambda\in\sigma[A]} |p(\lambda)|\right) \cdot \left\|x^* - x^{(0)}\right\|_A, \quad n = 0, 1, \dots, n^*.$$

Demostración. Para cada polinomio $p \in \prod_n$ con $p(0) = 1$, definimos $q(t) := \dfrac{1-p(t)}{t} \in \prod_{n-1}$, y el vector $x = x^{(0)} + q(A)r^{(0)} \in x^{(0)} + \mathcal{K}_n(A, r^{(0)})$. Entonces, se tiene que

$$x^* = x^{(0)} + [x^* - x^{(0)}] = x^{(0)} + A^{-1}r^{(0)},$$

puesto que $A(x^* - x^{(0)}) = Ax^* - Ax^{(0)} = b - Ax^{(0)} = r^{(0)}$, y se llega así a que

$$x - x^* = q(A)r^{(0)} - A^{-1}r^{(0)} = [I - p(A) - I]\, A^{-1}r^{(0)} = -p(A)\left(x^* - x^{(0)}\right).$$

Por el lema 2.20, con $x^* = x^{(0)} + \sum_{j=1}^{N} c_j v_j \in \mathbb{R}^N$, obtenemos

$$\left\|x^{(n)} - x^*\right\|_A \leq \|x - x^*\|_A = \left\|p(A)(x^* - x^{(0)})\right\|_A =$$

$$= \left(\sum_{j=1}^{N} c_j^2 \lambda_j p\left(\lambda_j\right)^2\right)^{1/2} \leq \sup_{\lambda \in \sigma[A]} |p(\lambda)| \overbrace{\left(\sum_{j=1}^{N} c_j^2 \lambda_j\right)^{1/2}}^{\|x^* - x^{(0)}\|_A}.$$

Luego, la anterior desigualdad también se cumple para el ínfimo de todos los polinomios $p \in \prod_n$ con $p(0) = 1$. □

Este último teorema 2.21 y el siguiente Lema permitirán dar una cota de error más práctica.

Lema 2.22. *Sean $T_n(t) = cos(n \cdot arccost)$, para $t \in [-1, 1]$ y $n \geq 0$, los polinomios de Chebyshev de primera especie. Entonces,*

$$T_n(t) = \frac{1}{2}\left[\left(t + \sqrt{t^2-1}\right)^n + \left(t - \sqrt{t^2-1}\right)^n\right], \text{ para } |t| \geq 1$$

y

$$T_n\left(\frac{k+1}{k-1}\right) \geq \frac{1}{2}\left(\frac{\sqrt{k}+1}{\sqrt{k}-1}\right)^n \quad , \text{ para } k \in \mathbb{R}, k > 1. \tag{2.15}$$

Demostración. Los polinomios de Chebyshev satisfacen la recurrencia a tres términos $T_0(t) = 1$, $T_1(t) = t$, $T_{n+1}(t) = 2tT_n(t) - T_{n-1}(t)$ para $n \geq 1$ en base a lo cual se deduce por inducción que $T_n(t)$ es un polinomio de grado exacto n.
Ahora bien, con $\theta = arccos(t)$, $T_n(t) = cos(n\theta)$ y para $t \in [-1, 1]$,

$$T_n(t) = \frac{1}{2}\left[(cos\theta + isen\theta)^n + (cos\theta - isen\theta)^n\right] = \frac{1}{2}\left[\left(t + i\sqrt{1-t^2}\right)^n + \left(t - i\sqrt{1-t^2}\right)^n\right].$$

Puesto que para $|t| \geq 1$,

$$\frac{1}{2}\left[\left(t + i\sqrt{1-t^2}\right)^n + \left(t - i\sqrt{1-t^2}\right)^n\right] = \frac{1}{2}\left[\left(t + \sqrt{t^2-1}\right)^n + \left(t - \sqrt{t^2-1}\right)^n\right]$$

$$= \sum_{\substack{k=0 \\ kpar}}^{n} \binom{n}{k} t^{n-k}\left(\sqrt{t^2-1}\right)^k$$

es un polinomio de grado n en t que coincide con $T_n(t)$ en $[-1, 1]$, debe tenerse que

$$T_n(t) = \frac{1}{2}\left[\left(t + \sqrt{t^2-1}\right)^n + \left(t - \sqrt{t^2-1}\right)^n\right], \; |t| \geq 1.$$

Finalmente, dado $k > 1$ y $t := \dfrac{k+1}{k-1}$, se tiene que $t \pm \sqrt{t^2-1} = \dfrac{\sqrt{k} \pm 1}{\sqrt{k} \mp 1}$ y

$$T_n\left(\frac{k+1}{k-1}\right) = \frac{1}{2}\left[\left(\frac{\sqrt{k}+1}{\sqrt{k}-1}\right)^n + \underbrace{\left(\frac{\sqrt{k}-1}{\sqrt{k}+1}\right)^n}_{\geq 0}\right] \geq \frac{1}{2}\left(\frac{\sqrt{k}+1}{\sqrt{k}-1}\right)^n.$$

□

Teorema 2.23. *Sea $A \in \mathbb{R}^{N\times N}$ una matriz simétrica y definida positiva. Entonces, las iteraciones $x^{(n)}$, $n \geq 0$, del método del Gradiente Conjugado para $Ax = b$, con $x^{(0)}$ arbitrario verifican*

$$\left\|x^{(n)} - x^*\right\|_A \leq 2\gamma^n \left\|x^{(0)} - x^*\right\|_A \quad y \quad \left\|x^{(n)} - x^*\right\|_2 \leq 2\sqrt{\kappa_A}\gamma^n \left\|x^{(0)} - x^*\right\|_2,\ n \geq 1,$$

siendo $x^* = A^{-1}b$, $\kappa_A = cond_2(A)$ y $\gamma = \dfrac{\sqrt{\kappa_A}-1}{\sqrt{\kappa_A}+1}$.

Observación 2.24. Observar que si A es una matriz simétrica y definida positiva, entonces su norma euclídea será su radio espectral, por lo que $cond_2(A) := \|A\|_2 \left\|A_2^{-1}\right\| = \rho(A)\rho(A^{-1}) = \dfrac{\lambda_{max}(A)}{\lambda_{min}(A)} \geq 1$, siendo $\lambda_{max}(A)$ y $\lambda_{min}(A)$ el mayor y menor autovalor de A, respectivamente.

Demostración. Por la observación 1.22, sabemos que $\kappa_A \geq 1$.

(I) En primer lugar, si $\kappa_A = 1$, entonces $\lambda = \lambda_{max}(A) = \lambda_{min}(A)$ para cada $\lambda \in \sigma[A]$. Luego, como A es simétrica y definida positiva, por el teorema 1.12, tenemos que $A = \lambda I$ con $\lambda > 0$. Por tanto, $\mathcal{K}_2(A, r^{(0)}) = \mathcal{K}_1(A, r^{(0)})$ y $x_1 = \ldots = x_n = x^* = \lambda^{-1}b$ para todo $n \geq 1$, y el enunciado es evidente.

(II) Ahora, supongamos que $\kappa_A > 1$, de modo que $0 < \gamma < 1$. Denotamos $M := \lambda_{max}(A)$ y $m := \lambda_{min}(A)$ con lo cual $\sigma[A] \subset [m, M]$. Definimos:

$$p(t) := \frac{1}{T_n\left(\frac{M+m}{M-m}\right)} T_n\left(\frac{M+m-2t}{M-m}\right).$$

Es claro que $p \in \prod_n$ con $p(0) = 1$ y si $t \in [m, M]$ entonces, $-1 \leq \dfrac{M+m-2t}{M-m} \leq 1$. Por tanto, puesto que $k_A = \dfrac{M}{m}$, usando (2.15) se tiene que:

$$\max_{m\leq t\leq M} |p(t)| = \left|T_n\left(\frac{M+m}{M-m}\right)\right|^{-1} = \left|T_n\left(\frac{k_A+1}{k_A-1}\right)\right|^{-1} \leq 2\gamma^n.$$

Aplicando el teorema 2.21, sigue que $\left\|x^{(n)} - x^*\right\|_A \leq 2\gamma^n \left\|x^{(0)} - x^*\right\|_A$. Finalmente, de (2.14) se obtiene

$$\left\|x^{(n)} - x^*\right\|_2 \leq \frac{1}{\sqrt{m}} \left\|x^{(n)} - x^*\right\|_A \leq \frac{2\gamma^n}{\sqrt{m}} \left\|x^{(0)} - x^*\right\|_A \leq 2\sqrt{k_A}\gamma^n \left\|x^{(0)} - x^*\right\|_2.$$

□

Observación 2.25. Este Teorema refleja que la convergencia del método GC puede verse ralentizada si $\kappa_A \gg 1$.

2.2.1. Método del Gradiente Conjugado para ecuaciones normales

Sea un sistema lineal de ecuaciones $Ax = b$ donde $A \in \mathbb{R}^{N\times N}$ es regular pero no necesariamente simétrica o definida positiva. Sin embargo, observamos que A^TA es una matriz regular, pues $det(A^TA) = det(A)^2 > 0$, simétrica y también definida positiva, puesto que $x^TA^TAx = (Ax)^T(Ax) = \|Ax\|_2^2 > 0$ para todo $x \in \mathbb{R}^N\backslash\{0\}$. El método del Gradiente Conjugado para las ecuaciones normales (GCNR) consiste en aplicar el método GC al sistema de ecuaciones normales $A^TAx = A^Tb$. Así pues, el método GCNR considera la proyección ortogonal sobre los subespacios de Krylov $\mathcal{K}_n(A^TA, A^Tr^{(0)})$.

Algoritmo para el método del Gradiente Conjugado para ecuaciones normales:

Considerar el sistema lineal $Ax = b$, donde $A \in \mathbb{R}^{N\times N}$ es una matriz regular y $b \in \mathbb{R}^N$. Denotamos $r^{(j)} := b - Ax^{(j)}$ y $\widehat{r}^{(j)} = A^Tr^{(j)}$, $j \geq 0$.

1. Dado $x^{(0)} \in \mathbb{R}^N$, definimos $\boxed{r^{(0)} := b - Ax^{(0)}}$, $\boxed{\widehat{r}^{(0)} := A^Tr^{(0)}}$ y $\boxed{\widehat{d}_0 := \widehat{r}^{(0)}}$.
2. Dados $\boxed{x^{(n)}, r^{(n)}, \widehat{r}^{(n)} \text{ y } \widehat{d}^{(n)}}$ con $n \geq 0$, mientras $r^{(n)} \neq 0$, calcular:
   - $\boxed{x^{(n+1)} = x^{(n)} + \alpha_n\widehat{d}^{(n)}}$, con $\alpha_n = \dfrac{\left\|\widehat{r}^{(n)}\right\|_2^2}{\langle A^TA\widehat{d}^{(n)}, \widehat{d}^{(n)}\rangle_2} = \dfrac{\left\|\widehat{r}^{(n)}\right\|_2^2}{\left\|A\widehat{d}^{(n)}\right\|_2^2}$
   - $\boxed{r^{(n+1)} = r^{(n)} - \alpha_nA\widehat{d}^{(n)}}$
   - $\boxed{\widehat{r}^{(n+1)} = \widehat{r}^{(n)} - \alpha_nA^TA\widehat{d}^{(n)} = A^Tr^{(n+1)}}$
   - $\boxed{\widehat{d}^{(n+1)} = \widehat{r}^{(n+1)} + \beta_n\widehat{d}^{(n)}}$, con $\beta_n = \dfrac{\left\|\widehat{r}^{(n+1)}\right\|_2^2}{\|\widehat{r}^{(n)}\|_2^2}$.

Observación 2.26. A diferencia del método GC para el caso simétrico definido positivo, el método GCNR requiere otro producto matriz-vector $A^Tr^{(j)}$ para cada $j \geq 0$. Sin embargo, se evita calcular la operación A^TA, que requiere un gran costo computacional.

Observación 2.27. Como consecuencia directa de la proposición 1.14, cambiando A por A^TA, se da la siguiente propiedad minimal para las iteraciones del método GCNR:

$$\left\|b - Ax^{(n)}\right\|_2 = \min_{x\in x^{(0)}+\mathcal{K}_n(A^TA, A^Tr^{(0)})} \|b - Ax\|_2.$$

En efecto, para todo $v \in \mathbb{R}^N$, $\|x^* - v\|_{A^TA}^2 = \langle A(x^* - v), A(x^* - v)\rangle_2 = \|b - Av\|_2^2$. Esta propiedad justifica la presencia de la letra R en la notación del método GCNR, ya que en esta variante se minimizan los residuales en norma euclídea. Asimismo, la letra N hace referencia a las ecuaciones normales.

Teorema 2.28. *Sea $A \in \mathbb{R}^{N\times N}$ una matriz regular. Entonces, las iteraciones $x^{(n)}$, $n \geq 0$, del método del Gradiente Conjugado para Ecuaciones Normales, con $x^{(0)}$ arbitrario, satisfacen*

$$\left\|b - Ax^{(n)}\right\|_2 \leq 2\gamma^n \left\|b - Ax^{(0)}\right\|_2 \quad y \quad \left\|x^{(n)} - x^*\right\|_2 \leq 2k_A\gamma^n \left\|x^{(0)} - x^*\right\|_2, n \geq 1,$$

siendo $\gamma = \dfrac{\kappa_A - 1}{\kappa_A + 1}$, $\kappa_A = cond_2(A)$.

Demostración. Es consecuencia directa del teorema 2.23. Nótese que $\kappa_{A^TA} = cond_2(A^TA) = cond_2(A)^2 = (\kappa_A)^2$. □

Observación 2.29. Por su propia definición, la iteración $x^{(n)}$ del método GCNR se obtiene como la proyección ortogonal de $x^{(0)}$ sobre $\mathcal{K} = \mathcal{K}_n(A^TA, A^Tr^{(0)})$ (con $\mathcal{L} = \mathcal{K}$):

$$x^{(n)} = x^{(0)} + \widehat{\delta}^{(n)}, \text{ con } \widehat{\delta}^{(n)} \in \mathcal{K}_n(A^TA, A^Tr^{(0)})$$

y

$$\widehat{r}^{(n)} = A^Tb - A^TAx^{(n)} = \widehat{r}^{(0)} - A^TA\widehat{\delta}^{(n)} \in \mathcal{K}^\perp,$$

de tal modo que $x^{(n)}$ minimiza 1.14

$$E_{A^TA}(x) = \|x - x^*\|^2_{A^TA} = \|Ax - b\|^2_2 = R_A(x) \text{ sobre } x^{(0)} + \mathcal{K}_n(A^TA, A^Tr^{(0)}).$$

Alternativamente, esta última propiedad nos dice que $x^{(n)}$ también se puede interpretar como la proyección de $x^{(0)}$ sobre $\mathcal{K} = \mathcal{K}_n(A^TA, A^Tr^{(0)})$ ortogonal a $\mathcal{L} := A\mathcal{K}$. En efecto, por la proposición 1.17,

$$x^{(n)} \text{ minimiza } R_A(x) = \|b - Ax\|^2_2 \text{ sobre } x^{(0)} + \mathcal{K} \Leftrightarrow r^{(0)} - A\widehat{\delta}^{(n)} \in \mathcal{L}^\perp.$$

Esto equivale a decir que

$$\begin{aligned} &\langle r^{(0)} - A\widehat{\delta}^{(n)}, Av\rangle_2 = 0, \forall v \in \mathcal{K} = \mathcal{K}_n(A^TA, A^Tr^{(0)}) \\ &\Leftrightarrow \langle A^Tr^{(0)} - A^TA\widehat{\delta}^{(n)}, v\rangle_2 = 0, \forall v \in \mathcal{K} = \mathcal{K}_n(A^TA, A^Tr^{(0)}) \Leftrightarrow \widehat{r}^{(n)} \in \mathcal{K}^\perp. \end{aligned}$$

2.3. El método del Residual Mínimo Generalizado (GMRES)

Sea el sistema lineal $Ax = b$, $b \in \mathbb{R}^N$, $A \in \mathbb{R}^{N\times N}$ regular y $x^{(0)}$ aproximación inicial fija a $x^* = A^{-1}b$, con residual $r^{(0)} = b - Ax^{(0)} \neq 0$.

Definición 2.30. *Para cada $n \geq 1$, el método del Residual Mínimo Generalizado (GMRES) produce una aproximación $x^{(n)} \in \mathbb{R}^N$ a través del método de proyección sobre $\mathcal{K} = \mathcal{K}_n(A, r^{(0)})$ ortogonal a $\mathcal{L} = A\mathcal{K}$, siendo $\mathcal{K}_n(A, r^{(0)})$ el enésimo subespacio de Krylov generado por A y $r^{(0)}$.*

Así:

$$\begin{gathered} x^{(n)} = x^{(0)} + \delta^{(n)}, \text{ con } \delta^{(n)} \in \mathcal{K}_n(A, r^{(0)}) \text{ tal que } r^{(n)} \in \mathcal{L}^\perp, \\ \text{siendo } r^{(n)} = b - Ax^{(n)} = r^{(0)} - A\delta^{(n)}. \end{gathered} \tag{2.16}$$

Observación 2.31. Al igual que el método GC para matrices simétricas y definidas positivas, el método GMRES converge a la solución exacta x^* en exactamente n^* iteraciones, pues ambos métodos están basados en los subespacios de Krylov.

Observación 2.32. Según la proposición 1.14, $x^{(n)}$ minimiza $R(x) = \|b - Ax\|^2_2$ sobre $x^{(0)} + \mathcal{K}_n(A, r^{(0)})$. Así, puesto que $\mathcal{K}_{n-1}(A, r^{(0)}) \subsetneq \mathcal{K}_n(A, r^{(0)})$, si $1 \leq n \leq n^*$, se tiene que $\left\|b - Ax^{(n)}\right\|_2 \leq \left\|b - Ax^{(n-1)}\right\|_2$, $1 \leq n \leq n^*$.

Esta última desigualdad puede darse con igualdad incluso para $1 \leq n \leq N - 1$, como muestra el siguiente ejemplo.

Ejemplo 2.33. Sean e_i el i-ésimo vector canónico de $\mathbb{R}^N$, $1 \leq i \leq N$, y

$$A = \begin{bmatrix} e_N^T \\ e_1^T \\ e_2^T \\ \vdots \\ e_{N-1}^T \end{bmatrix} = \begin{bmatrix} 0 & 0 & \cdots & 1 \\ 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \vdots & \vdots & \vdots & \ddots \\ 0 & \cdots & 1 & 0 \end{bmatrix} \in \mathbb{R}^{N\times N}, b = e_1 = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \\ 0 \end{bmatrix} \in \mathbb{R}^N, x^* = e_N = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix} \in \mathbb{R}^N$$

de modo que x^* es la única solución del sistema lineal $Ax = b$. Tomando $x^{(0)} = 0$, el método GMRES converge en, exactamente, N iteraciones, pues $r^{(0)} = b = e_1$, $Ar^{(0)} = e_2$, $A^2r^{(0)} = e_3, \ldots, A^{N-1}r^{(0)} = e_N$, $A^N r^{(0)} = e_1$. Luego, $\mathcal{K}_j(A, r^{(0)}) = span\{e_1, \ldots, e_j\}$ para $1 \leq j \leq N$, y $\mathcal{K}_{N+1}(A, r^{(0)}) = \mathcal{K}_N(A, b)$. Además, si $j = 1, \ldots, N-1$, entonces

$$\min_{x\in\mathcal{K}_j(A,r^{(0)})} \|b - Ax\|_2 = \min_{\lambda_1,\ldots,\lambda_j\in\mathbb{R}} \left\| e_1 - \sum_{k=2}^{j+1} \lambda_{k-1} e_k \right\|_2 = \min_{\lambda_1,\cdots,\lambda_j\in\mathbb{R}} \sqrt{1 + \lambda_1^2 + \ldots + \lambda_j^2} = 1$$

que se obtiene con $\lambda_1, \ldots, \lambda_j = 0$, esto es, $x^{(j)} = 0$ si $1 \leq j \leq N-1$ y $x^{(N)} = x^*$.

Observación 2.34. Fijemos una base $\{v_1, \ldots, v_n\}$ de $\mathcal{K} = \mathcal{K}_n(A, r^{(0)})$, $1 \leq n \leq n^*$, con matriz asociada $V \in \mathbb{R}^{N\times n}$, y tomemos $\{Av_1, \ldots, Av_n\}$ como base de $\mathcal{L} = A\mathcal{K}$ (por ser A regular). Entonces, poniendo $\delta^{(n)} = Vy^{(n)}$, para cierto $y^{(n)} \in \mathbb{R}^n$, se tiene que:

$$\begin{aligned} r^{(0)} - AVy^{(n)} = r^{(n)} \in \mathcal{L}^\perp &\Leftrightarrow (Av_j)^T \left(r^{(0)} - AVy^{(n)} \right) = 0,\ 1 \leq j \leq n \\ &\Leftrightarrow V^T A^T \left(r^{(0)} - AVy^{(n)} \right) = 0. \end{aligned}$$

Como $V^T A^T AV$ es regular (ver lema 1.4), sigue que $y^{(n)} = \left(V^T A^T AV\right)^{-1} \left(V^T A^T\right) r^{(0)}$ y

$$x^{(n)} = x^{(0)} + V \left(V^T A^T AV\right)^{-1} \left(V^T AT\right) r^{(0)}$$

con residual

$$r^{(n)} = \left[I - (AV) \left(V^T A^T AV\right)^{-1} \left(V^T A^T\right) \right] r^{(0)}.$$

Observación 2.35. En el resto de esta sección, consideramos que $\{v_1, \ldots, v_n\}$ es la base ortogonal respecto de $\langle \cdot, \cdot \rangle_2$ para $\mathcal{K}_n(A, r^{(0)})$ generada por el proceso de Arnoldi (ver lema 2.12), y denotaremos $h_{kn} := \langle Av_n, v_k \rangle_2$ para $n \geq 1$ y $k = 1, \ldots, n$. Asimismo, si $\widehat{v}_{n+1} \neq 0$, definimos

$$h_{n+1,n} := \|\widehat{v}_{n+1}\|_2 \text{ para } n \geq 1.$$

Por tanto, obtenemos una matriz rectangular de Hessenberg superior:

$$H_n = \begin{bmatrix} h_{11} & h_{12} & \cdots & h_{1n} \\ h_{21} & h_{22} & \cdots & h_{2n} \\ \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & h_{nn} \\ 0 & 0 & \cdots & h_{n+1,n} \end{bmatrix} \in \mathbb{R}^{(n+1)\times n}, \tag{2.17}$$

donde se tiene que $\widehat{H}_n := I_{n,n+1} \cdot H_n$ representa la parte cuadrada de la matriz, siendo $I_{n,n+1} = [I_n|0] \subset \mathbb{R}^{n\times(n+1)}$.

Además, según la observación 2.11, si A es una matriz simétrica, entonces $(H_n)_{i,j} = 0$, si $|i-j| \geq 2$. En particular, $\widehat{H}_n = T_n$ es una matriz cuadrada tridiagonal.

Teorema 2.36. *Sea una matriz $A \in \mathbb{R}^{N\times N}$ y un vector $r^{(0)} \in \mathbb{R}^N$ no nulo. Asumiendo la notación 2.10 del proceso de Arnoldi, se tiene que:*

$$\boxed{AV_n = V_{n+1}H_n,} \quad 1 \leq n \leq n^*-1, \quad \boxed{AV_{n^*} = V_{n^*}H_{n^*},} \; y \; \boxed{V_n^T AV_n = \widehat{H}_n,} \; 1 \leq n \leq n^*, \tag{2.18}$$

donde $V_j = [v_1|\cdots|v_j] \in \mathbb{R}^{N\times j}$, $1 \leq j \leq n^*$ y

$$H_{n^*} = \begin{bmatrix} h_{11} & h_{12} & \cdots & h_{1n^*} \\ h_{21} & h_{22} & \cdots & h_{2n^*} \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & h_{n^*,n^*-1} & h_{n^*n^*} \end{bmatrix} \in \mathbb{R}^{n^*\times n^*}.$$

Nótese que, en la última iteración, la matriz de Hessenberg es cuadrada, $H_{n^*} = \widehat{H}_{n^*}$.

Demostración. Por el proceso de Arnoldi, para cada $n = 1, 2, \ldots, n^*$, sabemos que $h_{n+1,n}v_{n+1} = \widehat{v}_{n+1} = Av_n - \sum_{k=1}^{n} h_{kn}v_k$, habiendo definido $h_{n^*+1,n^*} := 0$ y $v_{n^*+1} := 0$. Luego, nos queda $Av_n = \sum_{k=1}^{n+1} h_{kn}v_k$ para $1 \leq n \leq n^*$, lo cual demuestra la primera parte del resultado.
Por otro lado,

$$V_n^T AV_n = V_n^T V_{n+1}H_n = V_n^T[V_n|v_{n+1}]H_n = [I_n|0]H_n = \widehat{H}_n, \text{ para } 1 \leq n \leq n^*-1.$$

Para $n = n^*$, $V_{n^*}^T AV_{n^*} = V_{n^*}^T V_{n^*}H_{n^*} = H_{n^*}$. □

Veamos ahora cómo expresar las iteraciones $x^{(1)}, \ldots, x^{(n^*)}$ del método GMRES a través de la resolución del problema de mínimos cuadrados

$$\min_{x\in x^{(0)}+\mathcal{K}_n(A,r^{(0)})} \|b - Ax\|_2. \tag{2.19}$$

Teorema 2.37. *Sean $x^{(1)}, x^{(2)}, \ldots, x^{(n^*)} \in \mathbb{R}^N$ las iteraciones del método GMRES aplicado al sistema lineal $Ax = b$ con $A \in \mathbb{R}^{N\times N}$ matriz regular y $x^{(0)} \in \mathbb{R}^N$ tal que $r^{(0)} = b - Ax^{(0)} \neq 0$. Para cada $n = 1, \ldots, n^*$, el problema de mínimos cuadrados*

$$\min_{z\in\mathbb{R}^n} \|H_n z - c_n\|_2 \text{ donde } c_n := \left\|r^{(0)}\right\|_2 e_1, \text{ con } e_1 = (1, 0, \ldots, 0)^T \in \mathbb{R}^{\min\{n+1,n^*\}},$$

admite solución única $z_n \in \mathbb{R}^n$. Además, $x^{(n)} = x^{(0)} + V_n z_n$ y $\left\|Ax^{(n)} - b\right\|_2 = \|H_n z_n - c_n\|_2$ para $1 \leq n \leq n^*$, donde V_n y H_n están dadas por (2.18) y (2.17), respectivamente.

En lo que sigue, denotamos $\widehat{n} := \min\{n+1, n^*\}$.

Demostración. Se sabe que $x \in x^{(0)} + \mathcal{K}_n(A, r^{(0)}) \Leftrightarrow x = x^{(0)} + V_n \cdot z$, con $z \in \mathbb{R}^n$, y esta expresión es única. Entonces, aplicando el teorema 2.36 y sabiendo que $V_{\widehat{n}} \cdot e_1 = v_1$:

$$Ax - b = Ax^{(0)} + AV_n z - b = V_{\widehat{n}} H_n z - r^{(0)} = V_{\widehat{n}} H_n z - \left\| r^{(0)} \right\|_2 v_1 = V_{\widehat{n}} (H_n z - c_n).$$

Como $V_{\widehat{n}}^T V_{\widehat{n}} = I_{\widehat{n}}$, $\|Ax - b\|_2 = \|H_n z - c_n\|_2$. Luego, (2.19) equivale a resolver

$$\min_{z \in \mathbb{R}^n} \|H_n z - c_n\|_2. \tag{2.20}$$

Como $x^{(n)}$ es el único vector que minimiza $\|Ax - b\|_2$ en $x^{(0)} + \mathcal{K}_n(A, r^{(0)})$, y como $x^{(n)}$ se expresa de forma única como $x^{(n)} = x^{(0)} + V_n z_n$, para cierto $z_n \in \mathbb{R}^n$, entonces z_n es el único vector que minimiza $\|H_n z - C_n\|_2$. □

Observación 2.38. La matriz de Hessenberg $H_n \in \mathbb{R}^{\widehat{n} \times n}$, donde $\widehat{n} = \text{mín}\{n+1, n^*\}$ para $1 \leq n \leq n^*$, se puede factorizar como $H_n = Q_n R_n$ siendo $Q_n \in \mathbb{R}^{\widehat{n} \times \widehat{n}}$ una matriz ortogonal, es decir, $Q_n^T = Q_n^{-1}$, y $R_n \in \mathbb{R}^{\widehat{n} \times n}$ una matriz triangular superior, esto es, $(R_n)_{i,j} = 0$ si $i > j$. Además, $\widehat{R}_n := ((R_n)_{i,j})_{i,j=1}^n \in \mathbb{R}^{n \times n}$ es una matriz regular. Probaremos esta propiedad en el lema 2.43 y la utilizaremos a continuación para dar una expresión cerrada para las iteraciones del método GMRES y las correspondientes normas de los residuales.

Teorema 2.39. *Para cada $n \in \{1, \ldots, n^*\}$, sea $z_n \in \mathbb{R}^n$ la única solución del problema de mínimos cuadrados (2.20) con $H_n = Q_n R_n$, donde Q_n y R_n son las matrices definidas en la observación 2.38. Entonces,*

(i) $z_n = \widehat{R}_n^{-1} I_{n \times \widehat{n}} Q_n^T c_n$, donde $I_{n \times \widehat{n}}$ es la identidad rectangular de dimensión $n \times \widehat{n}$.
(ii) $\|H_n z_n - c_n\|_2 = \rho_n$, siendo

$$\rho_n = \begin{cases} |e_{n+1}^T Q_n^T c_n|, \text{ con } e_{n+1} := (0, \ldots, 0, 1)^T \in \mathbb{R}^{n+1}, \text{ si } 1 \leq n \leq n^* - 1, \\ 0, \qquad\qquad\qquad\qquad\qquad \text{si} \qquad n = n^*. \end{cases}$$

Demostración. Para $z \in \mathbb{R}^n$ arbitrario, $H_n z - c_n = Q_n \left(R_n z - Q_n^T c_n \right)$. Por lo tanto, como $\|Hy\|_2^2 = \left(y^T H^T\right)(Hy) = y^T R^T \left(Q^T Q\right) Ry = \|Ry\|_2^2$, se tiene que

$$\|H_n z - c_n\|_2 = \left\| R_n z - Q_n^T c_n \right\|_2.$$

Distinguimos dos casos:

1. Si $n = n^*$, entonces $R_n = \widehat{R}_n$ es una matriz regular y $\|H_n z - c_n\|_2$ toma calor mínimo igual a cero con $z = z_n := \widehat{R}_n^{-1} Q_n^T c_n$.
2. Si $n < n^*$, $R_n z = \begin{bmatrix} \widehat{R}_n z \\ 0 \end{bmatrix}$ y $Q_n^T c_n = \begin{bmatrix} I_{n \times (n+1)} Q_n^T c_n \\ e_{n+1}^T Q_n^T c_n \end{bmatrix}$. Luego,

$$\left\| R_n z - Q_n^T c_n \right\|_2^2 = |e_{n+1}^T Q_n^T c_n|^2 + \left\| \widehat{R}_n z - I_{n \times (n+1)} Q_n^T c_n \right\|_2^2,$$

y esta expresión se minimiza tomando $z = z_n := \widehat{R}_n^{-1} I_{n \times \widehat{n}} Q_n^T c_n$, con valor mínimo $\left\| R_n z - Q_n^T c_n \right\|_2 = |e_{n+1}^T Q_n^T c_n|$. □

El siguiente corolario expresa explícitamente las iteraciones y residuales del método GMRES.

Corolario 2.40. *Las iteraciones* $x^{(1)}, \cdots, x^{(n^*)}$ *del método GMRES se pueden expresar como* $x^{(n)} = x^{(0)} + \left(V_n \widehat{R}_n^{-1} I_{n\times\widehat{n}} Q_n^T e_1\right) \left\|r^{(0)}\right\|_2$, *con* $e_1 = (1, 0, \cdots, 0)^T \in \mathbb{R}^{\widehat{n}}$, *y residual:*

$$b - Ax^{(n)} = \begin{cases} \left(V_{n+1} Q_n e_{n+1}\right)\left(e_{n+1}^T Q_n^T e_1\right)\left\|r^{(0)}\right\|_2 & si \;\; 1 \le n \le n^* - 1, \\ 0, & si \qquad n = n^*. \end{cases}$$

Demostración. Como $x^{(n)} = x^{(0)} + V_n z_n$ y $AV_n = V_{n+1} H_n$, se tiene que

$$b - Ax^{(n)} = r^{(0)} - AV_n z_n = \left\|r^{(0)}\right\|_2 v_1 - AV_n z_n = V_{n+1}\left(c_n - H_n z_n\right).$$

Como $V_{n+1}\left(c_n - H_n z_n\right) = V_{n+1} Q_n \left(Q_n^T c_n - R_n z_n\right)$, usando el procedimiento de la demostración del teorema 2.39, para $1 \le n \le n^* - 1$, se tiene que

$$V_{n+1} Q_n \left(Q_n^T c_n - R_n z_n\right) = \left(V_{n+1} Q_n e_{n+1}\right)\left(e_{n+1}^T Q_n^T e_1\right)\left\|r^{(0)}\right\|_2.$$

□

Observación 2.41. En el lema 2.43 veremos que las matrices superiores de Hessenberg H_n, para $1 \le n \le n^*$, se pueden factorizar como el producto de una matriz ortogonal y una matriz triangular superior. Presentamos la prueba de este resultado por completitud en la exposición, aunque se base meramente en un caso particular de la descomposición QR para matrices rectangulares en general. En la prueba consideramos matrices de rotación de Givens, que cuentan con la propiedad de ser ortogonales:

$$G(i, i+1, \theta) := \left[\begin{array}{c|cc|c} I_{i-1} & & & \\ \hline & a & b & \\ & -b & a & \\ \hline & & & I_{\widehat{n}-1-i} \end{array}\right] \in \mathbb{R}^{\widehat{n}\times\widehat{n}} \text{ con } a = cos(\theta), b = sin(\theta), 1 \le i \le \widehat{n} - 1.$$

Observar que dado $\begin{pmatrix} c \\ s \end{pmatrix} \neq \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ se obtiene $\begin{pmatrix} a & b \\ -b & a \end{pmatrix}\begin{pmatrix} c \\ s \end{pmatrix} = \begin{pmatrix} r \\ 0 \end{pmatrix}$ con $r := \sqrt{c^2 + s^2}$ tomando $a = \dfrac{c}{r}$ y $b = \dfrac{s}{r}$. Si $\begin{pmatrix} c \\ s \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$, basta tomar $a = 1, b = 0$, ya que ha de cumplirse $a^2 + b^2 = 1$. Emplearemos este procedimiento para anular las componentes de H_n ubicadas bajo su diagonal.

Veamos ahora una propiedad del rango de matrices que nos será de utilidad en el lema 2.43.

Lema 2.42. *Sean* $M \in \mathbb{R}^{p\times q}$ *y* $N \in \mathbb{R}^{q\times r}$ *con* $rango(M) = q$, $(q \le p)$. *Entonces,* $rango(MN) = rango(N)$.

Demostración. Consideramos las aplicaciones lineales asociadas a M, N y MN respectivamente; estas son $\mathcal{L}_M : \mathbb{R}^q \to \mathbb{R}^p, \mathcal{L}_N : \mathbb{R}^r \to \mathbb{R}^q, \mathcal{L}_{MN} : \mathbb{R}^r \to \mathbb{R}^p$.
Como, $dim(Im(\mathcal{L}_M)) = rango(M) = q$, se tiene que $dim(ker(\mathcal{L}_M)) = 0$ y $ker(\mathcal{L}_M) = \{0\}$. El enunciado quedará demostrado por el primer Teorema de Isomorfía si probamos que $ker(\mathcal{L}_N) = ker(\mathcal{L}_{MN})$.
Resulta evidente que $ker(\mathcal{L}_N) \subseteq ker(\mathcal{L}_{MN})$, pues si $Nv = 0$, entonces $MNv = 0$. Por otro lado, sea $v \in ker(\mathcal{L}_{MN})$, $MNv = 0$. Entonces, $Nv \in ker(\mathcal{L}_M) = \{0\}$ y, por tanto, $v \in ker(\mathcal{L}_N)$. □

Lema 2.43. *La matriz H_n (2.17)-(2.18) con $1 \leq n \leq n^*$, admite la descomposición $H_n = Q_n R_n$ donde Q_n y R_n son las matrices definidas en la observación 2.38.*

Demostración. Denotamos la matriz de Hessenberg como $H_n^{(0)} := H_n$. Para cada $i \in \{1, \ldots, \widehat{n} - 1\}$, definimos matrices de rotación de Givens $G_i = G(i, i+1, \theta_i) \in \mathbb{R}^{\widehat{n} \times \widehat{n}}$ tales que

$$H_n^{(i)} := G_i H_n^{(i-1)} \text{ con } \left(H_n^{(i)}\right)_{i+1,i} = 0.$$

Es directo comprobar que $R_n := H_n^{(n)} = G_n G_{n-1} \cdots G_1 H_n$ verifica que $(R_n)_{i,j} = 0$ para $i > j$. Además, puesto que $G_i^T = G_i^{-1}$, $1 \leq i \leq \widehat{n} - 1$, definiendo la matriz $Q_n := G_1^T \cdots G_{n-1}^T G_n^T$ obtenemos $Q_n Q_n^T = Q_n^T Q_n = I$ y $H_n = Q_n R_n$.
Finalmente, veamos que la matriz $\widehat{R}_n := ((R_n)_{i,j})_{i,j=1}^n$ es regular. Por (2.18), sabemos que $AV_n = V_{\widehat{n}} H_n = V_{\widehat{n}} Q_n R_n$ y $R_n \in \mathbb{R}^{\widehat{n} \times n}$. Aplicando el lema 2.42 y, teniendo en cuenta que $rango(Q_n) = \widehat{n} = rango(V_{\widehat{n}})$, entonces, $rango(R_n) = rango(Q_n R_n) = rango(V_{\widehat{n}} Q_n R_n)$. Además, como A es una matriz inversible, sigue que $rango(V_{\widehat{n}} Q_n R_n) = rango(AV_n) = rango(V_n) = n$. Luego, $rango(\widehat{R}_n) = rango(R_n) = n$ y $\widehat{R}_n$ es una matriz regular. □

Así todo, el desarrollo del método GMRES queda recogido en el siguiente algoritmo:

Algoritmo para el método del GMRES:

Considerar el sistema lineal $Ax = b$, donde $A \in \mathbb{R}^{N \times N}$ es una matriz regular y $r^{(0)} = b - Ax^{(0)} \in \mathbb{R}^N$ es un vector no nulo.

1. Definir $\boxed{v_1 := \dfrac{r^{(0)}}{\|r^{(0)}\|_2}} \in \mathbb{R}^N$, y $\boxed{c := \left\|r^{(0)}\right\|_2 e_1} \in \mathbb{R}^N$. Para $n \geq 1$:
2. Calcular $h_{in} := \langle Av_n, v_i \rangle_2$, $1 \leq i \leq n$, $\widehat{v}_{n+1} := Av_n - \sum_{i=1}^{n} h_{in} v_i$ y $h_{n+1,n} := \|\widehat{v}_{n+1}\|_2$.
3. Si $n = 1$, ir al paso 4. De lo contrario, aplicar la rotación de Givens $G_j = \begin{bmatrix} a_j & b_j \\ -b_j & a_j \end{bmatrix}$ a $(h_{jn}, h_{j+1,n})^T$ para $1 \leq j \leq n-1$, y redefinir
$$\begin{bmatrix} h_{jn} \\ h_{j+1,n} \end{bmatrix} := G_j \begin{bmatrix} h_{jn} \\ h_{j+1,n} \end{bmatrix} \text{ para } 1 \leq j \leq n-1.$$
4. Construir $G_n = \begin{bmatrix} a_n & b_n \\ -b_n & a_n \end{bmatrix}$ tal que $(0,1) G_n \begin{bmatrix} h_{nn} \\ h_{n+1,n} \end{bmatrix} = 0$ y redefinir
$$\begin{bmatrix} h_{nn} \\ 0 \end{bmatrix} := G_n \begin{bmatrix} h_{nn} \\ h_{n+1,n} \end{bmatrix}.$$
5. Redefinir c de modo que $\begin{bmatrix} c_n \\ c_{n+1} \end{bmatrix} := G_n \begin{bmatrix} c_n \\ c_{n+1} \end{bmatrix}$.
   - Si $|c_{n+1}| \leq TOL$, entonces resolver $(h_{ij})_{i,j=1}^n z = (c_i)_{i=1}^n$ y tomar
   $$\boxed{x^{(n)} = x^{(0)} + \sum_{j=1}^{n} z_j v_j}.$$
   - En otro caso, calcular $v_{n+1} := \dfrac{\widehat{v}_{n+1}}{\|\widehat{v}_{n+1}\|_2}$ y volver al paso 2 con $n = n + 1$.

Finalizamos esta sección con los siguientes resultados, que dan una estimación de convergencia del método GMRES, en la línea del teorema 2.21.

Teorema 2.44. *Sea $A \in \mathbb{R}^{N\times N}$ regular. Si $x^{(n)}$ denota la enésima iteración del método GMRES, entonces*

$$\left\|b - Ax^{(n)}\right\|_2 \leq \left(\inf_{\substack{p\in\prod_n \\ p(0)=1}} \|p(A)\|_2\right) \cdot \left\|r^{(0)}\right\|_2$$

Demostración. Dado cualquier polinomio $p \in \prod_n$ con $p(0) = 1$, sea $q(t) := \dfrac{1-p(t)}{t} \in \prod_{n-1}$. Luego, definiendo $x = x^{(0)} + q(A)r^{(0)} \in x^{(0)} + \mathcal{K}_n(A, r^{(0)})$, tenemos que

$$b - Ax = r^{(0)} - Aq(A)r^{(0)} = \left[I - Aq(A)\right] r^{(0)} = p(A)r^{(0)}.$$

La propiedad minimal del método GMRES (2.19) implica que

$$\left\|b - Ax^{(n)}\right\|_2 \leq \|b - Ax\|_2 = \left\|p(A)r^{(0)}\right\|_2 \leq \|p(A)\|_2 \left\|r^{(0)}\right\|_2.$$

Entonces, la desigualdad anterior se cumple para el ínfimo. □

Corolario 2.45. *Sea $A \in \mathbb{R}^{N\times N}$ una matriz diagonalizable, es decir, existe $T \in \mathbb{R}^{N\times N}$ una matriz tal que $T^{-1}AT = D$ donde $D := diag\,(\lambda_1, \cdots, \lambda_N)$. Entonces, si $x^{(n)}$ denota la enésima iteración del método GMRES:*

$$\left\|b - Ax^{(n)}\right\|_2 \leq cond_2(T) \left(\inf_{\substack{p\in\prod_n \\ p(0)=1}} \max_{k=1,\cdots,N} |p(\lambda_k)|\right) \cdot \left\|r^{(0)}\right\|_2$$

Demostración. La prueba sigue del teorema 2.44, teniendo en cuenta que, como $A = T^{-1}DT$,

$$\|p(A)\|_2 \leq \|T\|_2 \|p(D)\|_2 \left\|T^{-1}\right\|_2 = cond_2(T) \max_{k=1,\cdots,N} |p(\lambda_k)|.$$

□

Corolario 2.46. *Sea $A \in \mathbb{R}^{N\times N}$ una matriz regular diagonalizable con exactamente m autovalores distintos, entonces $x^{(m)} = x^*$, esto es, el método GMRES converge en, a lo sumo, m iteraciones.*

Demostración. Definimos el polinomio $p(t) = \displaystyle\prod_{i=1}^{m} \frac{\lambda_i - t}{\lambda_i} \in \prod_m$ con $p(0) = 1$. Entonces, $\max\limits_{k=1,\cdots,N} |p(\lambda_k)| = 0$. Finalmente, por el teorema 2.44, tenemos $\left\|b - Ax^{(m)}\right\|_2 = 0$. □

2.4. Precondicionamiento y factorizaciones incompletas

Las secciones anteriores han deparado estimaciones de la convergencia de los métodos GC, GCNR y GMRES. Así, basándonos en el teorema 2.23, la observación 2.27 y el teorema 2.44, respectivamente, ha quedado patente que la tasa de convergencia de estos métodos se puede ver afectada por el número de condición, respecto de la norma euclídea, de cierta matriz relacionada con la matriz de coeficientes del sistema

$Ax = b$.
Así todo, se llega a la conclusión de que los métodos GC, GCNR y GMRES pueden ver ralentizada su convergencia en caso de que este número de condición sea lo suficientemente grande. De hecho, el próximo capítulo servirá para ilustrar que, en problemas prácticos, esta lenta convergencia se presenta de manera habitual.
Una alternativa para atenuar los problemas de condicionamiento consiste en *precondicionar* el sistema lineal $Ax = b$, esto es, considerar una matriz regular M, denominada precondicionador, para convertir el sistema $Ax = b$ en el sistema equivalente $M^{-1}Ax = M^{-1}b$ de modo que, en cierto sentido que no precisaremos aquí, M sea próxima a A y la resolución de sistemas $My = c$ sea poco costosa.
Esta alternativa se conoce como precondicionamiento a la izquierda (para otras variantes de precondicionamiento, véase [9, Cap. 9]).

1. El Método del Gradiente Conjugado Precondicionado (PGC): consiste en aplicar el método del gradiente Conjugado al sistema $M^{-1}Ax = M^{-1}b$ respecto del producto interior $\langle u, v\rangle_M = v^T Mu$, para $u, v \in \mathbb{R}^N$, siendo M una matriz simétrica y definida positiva. Considerando el método de proyección ortogonal sobre los subespacios de Krylov $\mathcal{K}_n(M^{-1}A, M^{-1}r^{(0)})$, se obtiene el siguiente algoritmo:

   ALGORITMO: Método del Gradiente Conjugado Precondicionado (PGC)
   Considerar el sistema lineal $Ax = b$ donde $A \in \mathbb{R}^{N\times N}$ es una matriz simétrica y definida positiva, y $b \in \mathbb{R}^N$ un vector no nulo. Dada M una matriz simétrica y definida positiva, denotamos $r^{(j)} = b - Ax^{(j)}$, $\widetilde{r}^{(j)} = M^{-1}r^{(n)}$ para $j \geq 0$.
   (1) Dado $x^{(0)}$, definimos $\boxed{r^{(0)} := b - Ax^{(0)}}$, $\boxed{\widetilde{r}^{(0)} := M^{-1}r^{(0)}}$ y $\boxed{\widetilde{d}^{(0)} := \widetilde{r}^{(0)}}$.
   (2) Dados $x^{(n)}, r^{(n)}, \widetilde{r}^{(n)}$ y $\widetilde{d}^{(n)}$, mientras $r^{(n)} \neq 0$, calcular
   - $x^{(n+1)} = x^{(n)} + \alpha_n \widetilde{d}^{(n)}$, con $\alpha_n = \dfrac{\langle \widetilde{r}^{(n)}, \widetilde{r}^{(n)}\rangle_M}{\langle M^{-1}A\widetilde{d}^{(n)}, \widetilde{d}^{(n)}\rangle_M} = \dfrac{\langle r^{(n)}, \widetilde{r}^{(n)}\rangle_2}{\langle A\widetilde{d}^{(n)}, \widetilde{d}^{(n)}\rangle_2}$
   - $r^{(n+1)} = r^{(n)} - \alpha_n A\widetilde{d}^{(n)}$
   - $\widetilde{r}^{(n+1)} = \widetilde{r}^{(n)} - \alpha_n \left(M^{-1}A\right)\widetilde{d}^{(n)} = M^{-1}\left[r^{(n)} - \alpha_n A\widetilde{d}^{(n)}\right] = M^{-1}r^{(n+1)}$
   - $\widetilde{d}^{(n+1)} = \widetilde{r}^{(n+1)} + \beta_n \widetilde{d}^{(n)}$, con $\beta_n = \dfrac{\langle \widetilde{r}^{(n+1)}, \widetilde{r}^{(n+1)}\rangle_M}{\langle \widetilde{r}^{(n)}, \widetilde{r}^{(n)}\rangle_M} = \dfrac{\langle r^{(n+1)}, \widetilde{r}^{(n+1)}\rangle_2}{\langle r^{(n)}, \widetilde{r}^{(n)}\rangle_2}$.

   Observación 2.47. Obsérvese que cada iteración del método PGC requiere hallar el producto matriz-vector $A\widetilde{d}^{(n)}$ y resolver el sistema lineal $M\widetilde{r}^{n+1} = r^{(n)} - \alpha_n A\widetilde{d}^{(n)}$.

2. El Método GCNR Precondicionado (PGCNR): análogamente consiste en aplicar el método GC al sistema precondicionado $M^{-1}A^TAx = M^{-1}A^Tb$ respecto del producto interior $\langle\cdot,\cdot\rangle_M$ siendo M una matriz simétrica y definida positiva.
   Considerando ahora los subespacios de Krylov $\mathcal{K}_n\left(M^{-1}A^TA, M^{-1}A^Tr^{(0)}\right)$ y el método de proyección ortogonal se obtiene el siguiente algoritmo:

   ALGORITMO: Método GCNR Precondicionado (PGCNR)
   Considerar el sistema lineal $Ax = b$ donde $A \in \mathbb{R}^{N\times N}$ es una matriz regular, y $b \in \mathbb{R}^N$ un vector no nulo. Dada M una matriz simétrica y definida positiva, denotamos $r^{(j)} = b - Ax^{(j)}$, $\widehat{r}^{(j)} = A^Tr^{(j)}$ y $\widetilde{r}^{(j)} = M^{-1}\widehat{r}^{(j)}$ para $j \geq 0$.
   (1) Dado $x^{(0)} \in \mathbb{R}^N$, definimos $\boxed{r^{(0)} := b - Ax^{(0)}}$, $\boxed{\widehat{r}^{(0)} := A^Tr^{(0)}}$, $\boxed{\widetilde{r}^{(0)} := M^{-1}\widehat{r}^{(0)}}$ y $\boxed{\widetilde{d}^{(0)} := \widetilde{r}^{(0)}}$.
   (2) Dados $x^{(n)}, r^{(n)}, \widehat{r}^{(n)}, \widetilde{r}^{(n)}$ y $\widetilde{d}^{(n)}$, mientras $r^{(n)} \neq 0$, calcular

- $x^{(n+1)} = x^{(n)} + \alpha_n \widetilde{d}^{(n)}$, con $\alpha_n = \dfrac{\langle \widetilde{r}^{(n)}, \widetilde{r}^{(n)} \rangle_M}{\langle M^{-1}A^T A\widetilde{d}^{(n)}, \widetilde{d}^{(n)} \rangle_M} = \dfrac{\langle \widehat{r}^{(n)}, \widetilde{r}^{(n)} \rangle_2}{\left\| A\widetilde{d}^{(n)} \right\|_2^2}$
- $r^{(n+1)} = r^{(n)} - \alpha_n A\widetilde{d}^{(n)}$
- $\widehat{r}^{(n+1)} = A^T r^{(n+1)}$
- $$\widetilde{r}^{(n+1)} = \widetilde{r}^{(n)} + \alpha_n \left(M^{-1}A^T A \right) \widetilde{d}^{(n)} = M^{-1} \left[\widehat{r}^{(n)} - \alpha_n A^T A\widetilde{d}^{(n)} \right] = \\ = M^{-1}A^T \left[r^{(n)} - \alpha_n A\widetilde{d}^{(n)} \right] = M^{-1}\widehat{r}^{(n+1)}$$
- $\widetilde{d}^{(n+1)} = \widetilde{r}^{(n+1)} + \beta_n \widetilde{d}^{(n)}$, con $\beta_n = \dfrac{\langle \widetilde{r}^{(n+1)}, \widetilde{r}^{(n+1)} \rangle_M}{\langle \widetilde{r}^{(n)}, \widetilde{r}^{(n)} \rangle_M} = \dfrac{\langle \widehat{r}^{(n+1)}, \widetilde{r}^{(n+1)} \rangle_2}{\langle \widehat{r}^{(n)}, \widetilde{r}^{(n)} \rangle_2}$.

Observación 2.48. A nivel computacional, cada iteración de PGCNR requiere dos productos matriz-vector $A\widetilde{d}^{(n)}$, $\widehat{r}^{(n+1)} = A^T r^{(n+1)}$ y resolver $M\widetilde{r}^{(n+1)} = \widehat{r}^{(n+1)}$.

3. El Método GMRES Precondicionado (PGMRES): consiste en aplicar directamente el método GMRES al sistema precondicionado $M^{-1}Ax = M^{-1}b$, siendo M una matriz regular. A partir de los subespacios de Krylov $\mathcal{K}_n \left(M^{-1}A, M^{-1}r^{(0)} \right)$, se trata de hallar el vector $x^{(n)} \in x^{(0)} + \mathcal{K}_n \left(M^{-1}A, M^{-1}r^{(0)} \right)$ tal que

$$\left\| M^{-1}b - M^{-1}Ax^{(n)} \right\|_2 = \min_{x \in x^{(0)} + \mathcal{K}_n \left(M^{-1}A, M^{-1}r^{(0)} \right)} \left\| M^{-1}b - M^{-1}Ax \right\|_2.$$

Se obtiene el siguiente algoritmo:

ALGORITMO: Método GMRES Precondicionado (PGMRES)
Considerar el sistema lineal $Ax = b$ donde $A \in \mathbb{R}^{N \times N}$ es una matriz regular, $b \in \mathbb{R}^N$ un vector no nulo y M una matriz regular. Dados $x^{(0)} \in \mathbb{R}^N$ y $r^{(0)} = b - Ax^{(0)}$,

(1) Definimos $\boxed{\widetilde{r}^{(0)} := M^{-1}r^{(0)}}$, $\boxed{v_1 = \dfrac{\widetilde{r}^{(0)}}{\left\| \widetilde{r}^{(0)} \right\|_2}}$ y $\boxed{c = \left\| \widetilde{r}^{(0)} \right\|_2 e_1} \in \mathbb{R}^N$.

(2) Sean $\widetilde{v}_n := M^{-1}Av_n$ y $h_{in} := \langle \widetilde{v}_n, v_i \rangle_2$ para $1 \leq i \leq n$. Calcular $\widehat{v}^{(n+1)} = \widetilde{v}^{(n)} - \sum_{i=1}^{n} h_{in}v_i$ y $h_{n+1,n} = \left\| \widehat{v}^{(n+1)} \right\|_2$.

(3) Continuar con los pasos 3-5 del Algoritmo del Método GMRES.

Observación 2.49. Nótese que se requiere resolver el sistema $M\widetilde{v}^{(n)} = Av_n$ en cada iteración de PGMRES.

Observación 2.50. La elección del precondicionador M puede influir bastante en la eficiencia del correspondiente método precondicionado y, por lo general, M ha de elegirse adaptado al sistema lineal que se desea resolver. No obstante, en este trabajo consideraremos dos tipos de precondicionadores generales, ambos basados en las denominadas *factorizaciones matriciales incompletas*. Una situación ideal sería elegir $M = L \cdot U$, donde L y U forman una descomposición LU de la matriz A, con L triangular inferior y U triangular superior, o bien $M = L \cdot L^T$, la descomposición de Cholesky de A, siempre que A sea simétrica y definida positiva. Sin embargo, tales descomposiciones pueden ser extremadamente costosas o incluso inviables computacionalmente si la dimensión N de la matriz A es suficientemente grande. Una primera estrategia para relajar la

exigencia computacional de las descomposiciones LU y de Cholesky consiste en fijar un subconjunto de índices

$$\mathbb{S} \subseteq \{1, \ldots, N\} \times \{1, \ldots, N\}$$

y seguir los pasos usuales de los algoritmos para computar una descomposición LU o de Cholesky pero sólo en aquellos índices $(i, j) \in \mathbb{S}$. Esto da lugar a descomposiciones LU y de Cholesky incompletas. En particular, una usual opción a considerar es tomar

$$\mathbb{S} = \{(i, j) \in \{1, \ldots, N\}^2 / a_{i,j} \neq 0\},$$

esto es, realizar los pasos de la descomposición LU o de Cholesky pero respetando las componentes nulas de la matriz A original. Esta elección da lugar a las factorizaciones incompletas LU y de Cholesky con "nivel 0 de relleno"denotadas como ILU(0) e IC(0), respectivamente. Observar que si A es una matriz sparse, esto es, con una elevada proporción de elementos nulos, el cálculo de factorizaciones LU (o de Cholesky) hace que los factores L y U tengan un mayor número de componentes no nulas, lo cual conlleva un mayor gasto computacional en la resolución de sistemas $L \cdot U \cdot y = c$ por sustitución hacia atrás y hacia adelante. Por ello, la elección de un precondicionador $M = L \cdot U$ (o $M = L \cdot L^T$) a través de una factorización incompleta suele ser ventajosa si la dimensión N de A es elevada, a pesar de que $A - LU \neq 0$ (resp. $A - LL^T \neq 0$).

De [1] tomamos los siguientes pseudocódigos para las factorizaciones LU y de Cholesky incompletas.

- ALGORITMO ILU(0) (por el Método de Doolittle, esto es, L con unos en la diagonal)
  - Para $r = 1 : N - 1$
  - Para $i = r + 1 : N$
  - Si $(i, r) \in \mathbb{S}$, entonces $a_{i,r} = a_{i,r}/a_{r,r}$
  - Para $j = r + 1 : N$
  - Si $(i, j) \in \mathbb{S}$ y $(r, j) \in \mathbb{S}$, entonces

$$a_{i,j} = a_{i,j} - \frac{a_{i,r} \cdot a_{r,j}}{a_{r,r}} \tag{2.21}$$

- ALGORITMO IC(0) (para hallar L, factor triangular inferior)
  - Para $r = 1 : N$, $a_{r,r} = \sqrt{a_{r,r}}$
  - Para $i = r + 1 : N$
  - Si $(i, r) \in \mathbb{S}$, entonces $a_{i,r} = a_{i,r}/a_{r,r}$
  - Para $j = r + 1 : N$
  - Para $i = j : N$
  - Si $(i, j) \in \mathbb{S}$, $(i, r) \in \mathbb{S}$ y $(j, r) \in \mathbb{S}$, entonces

$$a_{i,j} = a_{i,j} - a_{i,r} \cdot a_{j,r} \tag{2.22}$$

Naturalmente, cuando $\mathbb{S} = \{1, \ldots, N\}^2$ se obtienen los algoritmos usuales para las descomposiciones LU (por el método de Doolittle) y la de Cholesky, respectivamente.

Observación 2.51 (Factorizaciones incompletas modificadas). En la factorización $ILU(0)$, si $(i, j) \notin \mathbb{S}$, entonces el valor $a_{i,r} \cdot a_{r,j}/a_{r,r}$ no se sustrae de $a_{i,j}$, como correspondería a una factorización LU completa (ver (2.21)). Del mismo modo, en la descomposición de Cholesky incompleta $IC(0)$, el valor $a_{i,r} \cdot a_{j,r}$ no se sustrae de $a_{i,j}$ si $(i, j) \notin \mathbb{S}$ (ver (2.22)).

Una opción alternativa que suele ser ventajosa en la práctica consiste en no despreciar completamente el valor $a_{i,r} \cdot a_{r,j}/a_{r,r}$ (respectivamente $a_{i,r} \cdot a_{j,r}$) cuando $(i,j) \notin \mathbb{S}$ y sustraerlo de la componente diagonal de la fila correspondiente, esto es, $a_{i,i}$. Esto da lugar a las descomposiciones incompletas modificadas LU y de Cholesky, $MILU(0)$ y $MIC(0)$, respectivamente.

- ALGORITMO MILU(0) Análogo a $ILU(0)$ pero modificando la condición en (2.21) por:
  - Si $(r,j) \in \mathbb{S}$, entonces
  - Si $(i,j) \in \mathbb{S}$, entonces

$$a_{i,j} = a_{i,j} - \frac{a_{i,r} \cdot a_{r,j}}{a_{r,r}}; \qquad \text{en caso contrario,} \qquad a_{i,i} = a_{i,i} - \frac{a_{i,r} \cdot a_{r,j}}{a_{r,r}}. \tag{2.23}$$

- ALGORITMO MIC(0) Análogo a $IC(0)$ pero modificando la condición en (2.22) por:
  - Si $(i,r) \in \mathbb{S}$ y $(j,r) \in \mathbb{S}$, entonces
  - Si $(i,j) \in \mathbb{S}$

$$a_{i,j} = a_{i,j} - a_{i,r} \cdot a_{j,r} \qquad \text{en caso contrario,} \qquad a_{i,i} = a_{i,i} - a_{i,r} \cdot a_{j,r} \tag{2.24}$$

Entrar en detalles acerca de la existencia, estabilidad y propiedades de las factorizaciones incompletas, y sus múltiples variantes, se escapa de los objetivos del presente trabajo. Para ello, remitimos al lector a [6], Cap. 7. No obstante, comentamos que las variantes modificadas $MILU(0)$ y $MIC(0)$ suelen resultar ventajosas como precondicionadores sobre $ILU(0)$ e $IC(0)$, respectivamente, pues los factores L y U resultantes de la factorización modificada garantizan que

$$A \cdot \mathbb{1} = L \cdot U \cdot \mathbb{1} \text{ (resp. } A \cdot \mathbb{1} = L \cdot L^T \cdot \mathbb{1} \text{),}$$

esto es, A y la descomposición incompleta modificada coinciden al actuar sobre el vector prueba $\mathbb{1} = (1, \cdots, 1)^T \in \mathbb{R}^N$ (véase [4, Pág. 170], y [6, Cap. 11]).

Finalizamos esta sección con un ejemplo explicativo de cómo se obtienen las descomposiciones $ILU(0)$, $IC(0)$, $MILU(0)$ y $MIC(0)$ para una matriz resultante de la discretización espacial del operador Laplaciano $-\Delta = -\dfrac{\partial^2}{\partial x^2} - \dfrac{\partial^2}{\partial y^2}$ en el cuadrado $[0,1]^2$ mediante diferencias centrales de orden 2 en una malla formada tomando cuatro puntos en cada dirección espacial (véase Capítulo 3). Así, la matriz simétrica y definida positiva con la que ilustraremos las factorizaciones es:

$$A = \begin{bmatrix} 4 & -1 & -1 & 0 \\ -1 & 4 & 0 & -1 \\ -1 & 0 & 4 & -1 \\ 0 & -1 & -1 & 4 \end{bmatrix}.$$

Ejemplo 1: Descomposiciones LU y variantes incompletas ILU(0) y MILU(0)

(1°) Recordemos cómo se obtiene la descomposición LU de A (por el método de Doolittle):

$$\begin{bmatrix} 4 & -1 & -1 & \boxed{0} \\ -1 & 4 & \boxed{0} & -1 \\ -1 & \boxed{0} & 4 & -1 \\ \boxed{0} & -1 & -1 & 4 \end{bmatrix} \xrightarrow[F_3 + \boxed{\frac{1}{4}} F_1]{F_2 + \boxed{\frac{1}{4}} F_1} \begin{bmatrix} 4 & -1 & -1 & 0 \\ 0 & \frac{15}{4} & -\frac{1}{4} & -1 \\ 0 & -\frac{1}{4} & \frac{15}{4} & -1 \\ 0 & -1 & -1 & 4 \end{bmatrix} \xrightarrow[F_4 + \boxed{\frac{4}{15}} F_2]{F_3 + \boxed{\frac{1}{15}} F_2} \begin{bmatrix} 4 & -1 & -1 & 0 \\ 0 & \frac{15}{4} & -\frac{1}{4} & -1 \\ 0 & 0 & \frac{56}{15} & -\frac{16}{15} \\ 0 & 0 & -\frac{16}{15} & \frac{56}{15} \end{bmatrix} \xrightarrow{F_4 + \boxed{\frac{2}{7}} F_3}$$

$$\begin{bmatrix} 4 & -1 & -1 & \boxed{0} \\ 0 & \frac{15}{4} & \boxed{-\frac{1}{4}} & -1 \\ 0 & 0 & \frac{56}{15} & -\frac{16}{15} \\ 0 & 0 & 0 & \frac{24}{7} \end{bmatrix} = U, \quad \text{con } L = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -\frac{1}{4} & 1 & 0 & -1 \\ -\frac{1}{4} & -\frac{1}{15} & 1 & 0 \\ 0 & -\frac{4}{15} & -\frac{2}{7} & 1 \end{bmatrix}$$

Observar que la matriz L se obtiene a partir de los $\boxed{\text{factores de eliminación}}$ con signo opuesto y unos en la diagonal. Además, observamos que $a_{i,j} = 0$, para $(i,j) \in \mathbb{S}^c := \{(i,j)/j \neq 5-1, 1 \leq i \leq 4\}$, pero esta propiedad no se mantiene para las matrices L y U.

(2º) La descomposición incompleta $ILU(0)$ se realiza de modo análogo, pero de tal modo que se fuerza que $u_{i,j} = l_{i,j} = 0$, para $(i,j) \in \mathbb{S}^c$. Esto es:

$$\begin{bmatrix} 4 & -1 & -1 & 0 \\ -1 & 4 & 0 & -1 \\ -1 & 0 & 4 & -1 \\ 0 & -1 & -1 & 4 \end{bmatrix} \xrightarrow[F_3+\boxed{\frac{1}{4}}F_1]{F_2+\boxed{\frac{1}{4}}F_1} \begin{bmatrix} 4 & -1 & -1 & 0 \\ 0 & \frac{15}{4} & \boxed{0} & -1 \\ 0 & \boxed{0} & \frac{15}{4} & -1 \\ 0 & -1 & -1 & 4 \end{bmatrix} \xrightarrow{F_4+\boxed{\frac{1}{15}}F_2} \begin{bmatrix} 4 & -1 & -1 & 0 \\ 0 & \frac{15}{4} & 0 & -1 \\ 0 & 0 & \frac{15}{4} & -1 \\ 0 & 0 & -1 & \frac{56}{15} \end{bmatrix} \xrightarrow{F_4+\boxed{\frac{2}{7}}F_3}$$

$$\begin{bmatrix} 4 & -1 & -1 & \boxed{0} \\ 0 & \frac{15}{4} & \boxed{0} & -1 \\ 0 & 0 & \frac{15}{4} & -1 \\ 0 & 0 & 0 & \frac{52}{15} \end{bmatrix} = U^{(I)}, \qquad \text{con } L^{(I)} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -\frac{1}{4} & 1 & 0 & 0 \\ -\frac{1}{4} & 0 & 1 & 0 \\ 0 & -\frac{4}{15} & -\frac{4}{15} & 1 \end{bmatrix}.$$

En este caso, nos encontramos con que $A - L^{(I)} \cdot U^{(I)} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & -\frac{1}{4} & 0 \\ 0 & -\frac{1}{4} & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \neq 0,$

pero $L^{(I)}$ y $U^{(I)}$ verifican que $u_{i,j} = l_{i,j} = 0$ para el mismo conjunto de índices $(i,j) \in \mathbb{S}^c$ que A.

(3º) La descomposición incompleta modificada $MILU(0)$ por filas (resp. por columnas) se lleva a cabo de modo análogo a $ILU(0)$, pero añadiendo al elemento diagonal de la correspondiente fila (resp. columna) el valor de las componentes descartadas en cada paso de $ILU(0)$ al forzar que $u_{i,j} = 0$ para el mismo conjunto de índices $\mathbb{S}^c$ que A. Así, tendremos:

$$\begin{bmatrix} 4 & -1 & -1 & 0 \\ -1 & 4 & 0 & -1 \\ -1 & 0 & 4 & -1 \\ 0 & -1 & -1 & 4 \end{bmatrix} \xrightarrow[F_3+\boxed{\frac{1}{4}}F_1]{F_2+\boxed{\frac{1}{4}}F_1} \begin{bmatrix} 4 & -1 & -1 & 0 \\ 0 & (\frac{15}{4}-\frac{1}{4}) & \boxed{0} & -1 \\ 0 & \boxed{0} & (\frac{15}{4}-\frac{1}{4}) & -1 \\ 0 & -1 & -1 & 4 \end{bmatrix} \xrightarrow{F_4+\boxed{\frac{2}{7}}F_2} \begin{bmatrix} 4 & -1 & -1 & 0 \\ 0 & \frac{7}{2} & 0 & -1 \\ 0 & 0 & \frac{7}{2} & -1 \\ 0 & 0 & -1 & \frac{26}{7} \end{bmatrix} \xrightarrow{F_4+\boxed{\frac{2}{7}}F_3}$$

$$\begin{bmatrix} 4 & -1 & -1 & \boxed{0} \\ 0 & \frac{7}{2} & \boxed{0} & -1 \\ 0 & 0 & \frac{7}{2} & -1 \\ 0 & 0 & 0 & \frac{24}{7} \end{bmatrix} = U^{(Im)}, \quad \text{con } L^{(Im)} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -\frac{1}{4} & 1 & 0 & 0 \\ -\frac{1}{4} & \boxed{0} & 1 & 0 \\ \boxed{0} & -\frac{2}{7} & -\frac{2}{7} & 1 \end{bmatrix}.$$

En este caso, se tiene que

$$A - L^{(Im)} \cdot U^{(Im)} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & \frac{1}{4} & -\frac{1}{4} & 0 \\ 0 & -\frac{1}{4} & \frac{1}{4} & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \neq 0.$$

En particular, con la descomposición $MILU(0)$ se garantiza que

$$A \cdot \mathbb{1} = L^{(Im)} \cdot U^{(Im)} \cdot \mathbb{1}, \text{ siendo } \mathbb{1} = (1,1,1,1)^T,$$

esto es, que la suma de las filas de A coincide con la de las filas de $L^{(Im)}$ y $U^{(Im)}$.

Ejemplo 2: Descomposiciones de Cholesky y sus variantes incompletas IC(0) y MIC(0)

(1º) Recordemos cómo se obtiene la descomposición de Cholesky $A = L \cdot L^T$, con L triangular inferior:

$$\begin{bmatrix} 4 & -1 & -1 & \boxed{0} \\ -1 & 4 & \boxed{0} & -1 \\ -1 & \boxed{0} & 4 & -1 \\ \boxed{0} & -1 & -1 & 4 \end{bmatrix} \xrightarrow{F_1 \boxed{\frac{1}{\sqrt{4}}}} \begin{bmatrix} 2 & -\frac{1}{2} & -\frac{1}{2} & 0 \\ -1 & 4 & 0 & -1 \\ -1 & 0 & 4 & -1 \\ 0 & -1 & -1 & 4 \end{bmatrix} \xrightarrow[F_3+\boxed{\frac{1}{2}}F_1]{F_2+\boxed{\frac{1}{2}}F_1} \begin{bmatrix} 2 & -\frac{1}{2} & -\frac{1}{2} & 0 \\ -1 & \frac{15}{4} & \boxed{-\frac{1}{4}} & -1 \\ -1 & \boxed{-\frac{1}{4}} & \frac{15}{4} & -1 \\ 0 & -1 & -1 & 4 \end{bmatrix} \xrightarrow{F_2 \cdot \boxed{\frac{4}{\sqrt{15}}}}$$

$$\begin{bmatrix} 2 & -\frac{1}{2} & -\frac{1}{2} & 0 \\ 0 & \frac{\sqrt{15}}{2} & -\frac{1}{2\sqrt{15}} & -\frac{2}{\sqrt{15}} \\ 0 & -\frac{1}{4} & \frac{15}{4} & -1 \\ 0 & -1 & -1 & 4 \end{bmatrix} \xrightarrow[F_4+\boxed{\frac{2}{\sqrt{15}}}F_2]{F_3+\boxed{\frac{1}{2\sqrt{15}}}F_2} \begin{bmatrix} 2 & -\frac{1}{2} & -\frac{1}{2} & 0 \\ 0 & \frac{\sqrt{15}}{2} & -\frac{1}{2\sqrt{15}} & -\frac{2}{\sqrt{15}} \\ 0 & 0 & \frac{56}{15} & -\frac{16}{15} \\ 0 & 0 & -\frac{16}{15} & \frac{56}{15} \end{bmatrix} \xrightarrow{F_3 \cdot \boxed{\sqrt{\frac{15}{56}}}}$$

$$\begin{bmatrix} 2 & -\frac{1}{2} & -\frac{1}{2} & 0 \\ 0 & \frac{\sqrt{15}}{2} & -\frac{1}{2\sqrt{15}} & -\frac{2}{\sqrt{15}} \\ 0 & 0 & 2\sqrt{\frac{14}{15}} & -4\sqrt{\frac{2}{105}} \\ 0 & 0 & -\frac{16}{15} & \frac{56}{15} \end{bmatrix} \xrightarrow{F_4+\boxed{4\sqrt{\frac{2}{105}}}F_3} \begin{bmatrix} 2 & -\frac{1}{2} & -\frac{1}{2} & 0 \\ 0 & \frac{\sqrt{15}}{2} & -\frac{1}{2\sqrt{15}} & -\frac{2}{\sqrt{15}} \\ 0 & 0 & 2\sqrt{\frac{14}{15}} & -4\sqrt{\frac{2}{105}} \\ 0 & 0 & 0 & \frac{24}{7} \end{bmatrix} \xrightarrow{F_4+\boxed{\sqrt{\frac{7}{24}}}}$$

$$\begin{bmatrix} 2 & -\frac{1}{2} & -\frac{1}{2} & \boxed{0} \\ 0 & \frac{\sqrt{15}}{2} & \boxed{-\frac{1}{2\sqrt{15}}} & -\frac{2}{\sqrt{15}} \\ 0 & 0 & 2\frac{14}{15} & -4\sqrt{\frac{2}{105}} \\ 0 & 0 & 0 & \sqrt{\frac{24}{7}} \end{bmatrix} = L^T.$$

Nuevamente, las matrices L (y L^T) no verifican $l_{i,j} = 0$, para $(i,j) \in \mathbb{S}^c$. Forzar esta propiedad da lugar a la descomposición de Cholesky incompleta $IC(0)$, que tratamos a continuación:

(2º) Descomposición IC(0):

$$\begin{bmatrix} 4 & -1 & -1 & 0 \\ -1 & 4 & 0 & -1 \\ -1 & 0 & 4 & -1 \\ 0 & -1 & -1 & 4 \end{bmatrix} \xrightarrow{F_1 \cdot \boxed{\frac{1}{\sqrt{4}}}} \begin{bmatrix} 2 & -\frac{1}{2} & -\frac{1}{2} & 0 \\ -1 & 4 & 0 & -1 \\ -1 & 0 & 4 & -1 \\ 0 & -1 & -1 & 4 \end{bmatrix} \xrightarrow[F_3+\boxed{\frac{1}{2}}F_1]{F_2+\boxed{\frac{1}{2}}F_1} \begin{bmatrix} 2 & -\frac{1}{2} & -\frac{1}{2} & 0 \\ 0 & \frac{15}{4} & \boxed{0} & -1 \\ 0 & \boxed{0} & \frac{15}{4} & -1 \\ 0 & -1 & -1 & 4 \end{bmatrix} \xrightarrow{F_2 \cdot \boxed{\sqrt{\frac{4}{15}}}}$$

$$\begin{bmatrix} 2 & -\frac{1}{2} & -\frac{1}{2} & 0 \\ 0 & \frac{\sqrt{15}}{2} & 0 & -\frac{2}{\sqrt{15}} \\ 0 & 0 & \frac{15}{4} & -1 \\ 0 & -1 & -1 & 4 \end{bmatrix} \xrightarrow{F_4+\boxed{\frac{2}{\sqrt{15}}}F_2} \begin{bmatrix} 2 & -\frac{1}{2} & -\frac{1}{2} & 0 \\ 0 & \frac{\sqrt{15}}{2} & 0 & -\frac{2}{\sqrt{15}} \\ 0 & 0 & \frac{15}{4} & -1 \\ 0 & 0 & -1 & \frac{56}{15} \end{bmatrix} \xrightarrow{F_3 \cdot \boxed{\sqrt{\frac{4}{15}}}} \begin{bmatrix} 2 & -\frac{1}{2} & -\frac{1}{2} & 0 \\ 0 & \frac{\sqrt{15}}{2} & 0 & -\frac{2}{\sqrt{15}} \\ 0 & 0 & \frac{\sqrt{15}}{2} & -\frac{2}{\sqrt{15}} \\ 0 & 0 & -1 & \frac{56}{15} \end{bmatrix}$$

$$\overset{F_4+\boxed{\frac{2}{\sqrt{15}}}F_3}{\longrightarrow}\begin{bmatrix}2 & -\frac{1}{2} & -\frac{1}{2} & 0\\ 0 & \frac{\sqrt{15}}{2} & 0 & -\frac{2}{\sqrt{15}}\\ 0 & 0 & \frac{\sqrt{15}}{2} & -\frac{2}{\sqrt{15}}\\ 0 & 0 & 0 & \frac{52}{15}\end{bmatrix}\overset{F_4\cdot\boxed{\sqrt{\frac{15}{52}}}}{\longrightarrow}\begin{bmatrix}2 & -\frac{1}{2} & -\frac{1}{2} & \boxed{0}\\ 0 & \frac{\sqrt{15}}{2} & \boxed{0} & -\frac{2}{\sqrt{15}}\\ 0 & 0 & \frac{\sqrt{15}}{2} & -\frac{2}{\sqrt{15}}\\ 0 & 0 & 0 & 2\sqrt{\frac{13}{15}}\end{bmatrix} = L^{(I)^T}.$$

Nuevamente,

$$A - L^{(I)}\cdot L^{(I)^T} = \begin{bmatrix}0 & 0 & 0 & 0\\ 0 & 0 & -\frac{1}{4} & 0\\ 0 & -\frac{1}{4} & 0 & 0\\ 0 & 0 & 0 & 0\end{bmatrix} \neq 0, \text{ pero } l_{i,j} = 0,\ (i,j) \in \mathbb{S}^c.$$

(3º) La descomposición incompleta modificada de Cholesky por filas ($MIC(0)$) se produce de modo análogo a $MILU(0)$, añadiendo al elemento diagonal de la correspondiente fila el valor de las componentes descartadas en cada paso de $IC(0)$.

$$\begin{bmatrix}4 & -1 & -1 & 0\\ -1 & 4 & 0 & -1\\ -1 & 0 & 4 & -1\\ 0 & -1 & -1 & 4\end{bmatrix}\overset{F_1+\boxed{\frac{1}{\sqrt{4}}}}{\longrightarrow}\begin{bmatrix}2 & -\frac{1}{2} & -\frac{1}{2} & 0\\ -1 & 4 & 0 & -1\\ -1 & 0 & 4 & -1\\ 0 & -1 & -1 & 4\end{bmatrix}\overset{\substack{F_2+\boxed{\frac{1}{2}}F_1\\ F_3+\boxed{\frac{1}{2}}F_1}}{\longrightarrow}\begin{bmatrix}2 & -\frac{1}{2} & -\frac{1}{2} & 0\\ 0 & \frac{15}{4} & -\frac{1}{4} & \boxed{0}\ -1\\ 0 & \boxed{0} & \frac{15}{4} & -\frac{1}{4}\ -1\\ 0 & -1 & -1 & 4\end{bmatrix}\overset{F_2\cdot\boxed{\sqrt{\frac{4}{14}}}}{\longrightarrow}$$

$$\begin{bmatrix}2 & -\frac{1}{2} & -\frac{1}{2} & 0\\ 0 & \sqrt{\frac{7}{2}} & 0 & -\sqrt{\frac{2}{7}}\\ 0 & 0 & \frac{7}{2} & -1\\ 0 & -1 & -1 & 4\end{bmatrix}\overset{F_4+\boxed{\sqrt{\frac{2}{7}}}F_2}{\longrightarrow}\begin{bmatrix}2 & -\frac{1}{2} & -\frac{1}{2} & 0\\ 0 & \sqrt{\frac{7}{2}} & 0 & -\sqrt{\frac{2}{7}}\\ 0 & 0 & \frac{7}{2} & -1\\ 0 & 0 & -1 & \frac{26}{7}\end{bmatrix}\overset{F_3\cdot\boxed{\sqrt{\frac{2}{7}}}}{\longrightarrow}\begin{bmatrix}2 & -\frac{1}{2} & -\frac{1}{2} & 0\\ 0 & \sqrt{\frac{7}{2}} & 0 & -\sqrt{\frac{2}{7}}\\ 0 & 0 & \sqrt{\frac{7}{2}} & -\sqrt{\frac{2}{7}}\\ 0 & 0 & -1 & \frac{26}{7}\end{bmatrix}$$

$$\overset{F_4+\boxed{\sqrt{\frac{2}{7}}}F_3}{\longrightarrow}\begin{bmatrix}2 & -\frac{1}{2} & -\frac{1}{2} & 0\\ 0 & \sqrt{\frac{7}{2}} & 0 & -\sqrt{\frac{2}{7}}\\ 0 & 0 & \sqrt{\frac{7}{2}} & -\sqrt{\frac{2}{7}}\\ 0 & 0 & 0 & \frac{24}{7}\end{bmatrix}\overset{F_4\cdot\boxed{\sqrt{\frac{7}{24}}}}{\longrightarrow}\begin{bmatrix}2 & -\frac{1}{2} & -\frac{1}{2} & \boxed{0}\\ 0 & \sqrt{\frac{7}{2}} & \boxed{0} & -\sqrt{\frac{2}{7}}\\ 0 & 0 & \sqrt{\frac{7}{2}} & -\sqrt{\frac{2}{7}}\\ 0 & 0 & 0 & 2\sqrt{\frac{6}{7}}\end{bmatrix} = L^{(Im)^T}.$$

De nuevo, se observa que

$$A - L^{(Im)}\cdot L^{(Im)^T} = \begin{bmatrix}0 & 0 & 0 & 0\\ 0 & \frac{1}{4} & -\frac{1}{4} & 0\\ 0 & -\frac{1}{4} & \frac{1}{4} & 0\\ 0 & 0 & 0 & 0\end{bmatrix} \neq 0, \text{ y } A\cdot\mathbb{1} = L^{(Im)}\cdot L^{(Im)^T}\cdot\mathbb{1}.$$

3

Ilustración numérica

3.1. Aplicación a la discretización de EDPs lineales elípticas

Nuestro objetivo en este capítulo será el de mostrar el comportamiento de los métodos numéricos presentados en los capítulos anteriores cuando se aplican a los grandes sistemas lineales que resultan de la discretización espacial mediante diferencias finitas de EDPs lineales elípticas con coeficientes constantes en un dominio rectangular n-dimensional Ω, esto es:

$$-d\sum_{i=1}^{n} u_{x_i,x_i} + a\sum_{i=1}^{n} u_{x_i} + r\cdot u = f, \text{ en } \Omega = (0,1)^n, \tag{3.1}$$

con condiciones de frontera en $\Gamma = \partial\Omega$ de tipo Dirichlet. En (3.1), a, d y r son constantes, $u = u(x_1,\ldots,x_n)$, $f = f(x_1,\ldots,x_n)$, mientras que $u_{x_i} = \dfrac{\partial u}{\partial x_i}$, $u_{x_i,x_i} = \dfrac{\partial^2 u}{\partial^2 x_i}$, $1 \le i \le n$.

Los códigos de los algoritmos empleados para llevar a cabo esta experimentación numérica se han tomado adaptados de [3], estando también disponibles en https://it.mathworks.com/matlabcentral/fileexchange/2158-templates-for-the-solution-of-linear-systems.

En los ejemplos que mostraremos a continuación, consideraremos las dimensiones espaciales $n = 2, 3$. Para la discretización de las derivadas parciales primeras u_{x_i} y segundas u_{x_i,x_i} consideraremos diferencias finitas centrales de segundo orden en una malla de amplitud $h > 0$ en cada dirección espacial.

En primer lugar, en el caso unidimensional $n = 1$, si $u = u(x)$ es suficientemente regular, las aproximaciones en diferencias centrales para $u'(x)$ y $u''(x)$ con paso $h > 0$ cumplen:

$$u'(x) = \frac{u(x+h) - u(x-h)}{2h} - \frac{1}{6}u'''(\xi_1)\cdot h^2, \quad \xi_1 \in (x-h, x+h)\,, \tag{3.2}$$

$$u''(x) = \frac{u(x+h) - 2u(x) + u(x-h)}{h^2} - \frac{1}{12}u^{(4)}(\xi_2)\cdot h^2, \quad \xi_2 \in (x-h, x+h)\,, \tag{3.3}$$

de tal modo que los cocientes en diferencias resultan exactos si u es un polinomio de grado menor o igual que 2.

Así, la discretización de (3.1), con condiciones de frontera $u(0) = \alpha$ y $u(1) = \beta$, en una malla $\Omega_h^{(1)}$ de $N+2$ puntos $\{i\cdot h\}_{i=0}^{N+1}$, con $h = \dfrac{1}{N+1}$, produce el esquema en diferencias siguiente:

$$d\left(-U_{i+1}+2U_i-U_{i-1}\right)+a\cdot\frac{h}{2}\left(U_{i+1}-U_{i-1}\right)+r\cdot h^2\cdot U_i=h^2 f_i \quad 1\le i\le N, \quad (3.4)$$

con $U_0=\alpha$, $U_{N+1}=\beta$, $f_i=f\left(i\cdot h\right)$ y $U_i\simeq u(i\cdot h)$, $1\le i\le N$.
En forma matricial, (3.4) se escribe como:

$$A^{(1)}\cdot U=b \quad (3.5)$$

donde $U=(U_1,\ldots,U_N)^T$, $b=\left(f_1+\left(d+\frac{h}{2}\right)\alpha, f_2,\ldots,f_{N-1},f_N+\left(d-a\frac{h}{2}\right)\beta\right)^T\in$ $\mathbb{R}^N$ y $A^{(1)}=d\cdot T_d^{(1)}+a\cdot\frac{h}{2}\cdot T_a^{(1)}+rh^2 I_N$, siendo I_N la matriz identidad de orden N y

$$T_d^{(1)}=Tridiag(-1,2,-1)\in\mathbb{R}^{N\times N},\quad T_a^{(1)}=Tridiag(1,0,-1)\in\mathbb{R}^{N\times N}. \quad (3.6)$$

En el caso bidimensional $n=2$, tendremos el problema de valores en la frontera

$$\begin{cases} -d(u_{xx}+u_{yy})+a(u_x+u_y)+r\cdot u=f, & (x,y)\in\Omega=(0,1)^2,\\ u=g, & (x,y)\in\Gamma=\partial\Omega, \end{cases} \quad (3.7)$$

y al considerar la malla $\Omega_h^{(2)}=\{(ih,jh)\}_{i,j=0}^{N+1}$, con $h=\dfrac{1}{N+1}$, y diferencias centrales de orden dos:

$$u_x(x,y)\simeq\frac{u(x+h,y)-u(x-h,y)}{2h},\quad u_y(x,y)\simeq\frac{u(x,y+h)-u(x,y-h)}{2h},$$

$$u_{xx}(x,y)\simeq\frac{u(x+h,y)-2u(x,y)+u(x-h,y)}{h^2},u_{yy}(x,y)\simeq\frac{u(x,y+h)-2u(x,y)+u(x,y-h)}{h^2},$$

se obtiene el esquema en diferencias

$$\begin{aligned} &d\left[(-U_{i+1,j}+2U_{i,j}-U_{i-1,j})+(-U_{i,j+1}+2U_{i,j}-U_{i,j-1})\right]+\\ &+a\frac{h}{2}\left[(U_{i+1,j}-U_{i-1,j})+(U_{i,j+1}-U_{i,j-1})\right]+r\cdot h^2\cdot U_{i,j}=h^2\cdot f_{i,j},\quad 1\le i,j\le N, \end{aligned}$$

que corresponde a un sistema lineal de dimensión N^2

$$A^{(2)}\cdot U=b, \quad (3.8)$$

donde $b\in\mathbb{R}^{N^2}$ contiene la discretización del dato $f(x,y)$ y las condiciones de frontera $g(x,y)$, $U=(U_{11},U_{21},\ldots,U_{N1},U_{12},U_{22},\ldots,U_{N2},\ldots,U_{1N},U_{2N},\ldots,U_{NN})^T\in\mathbb{R}^{N^2}$, $U_{i,j}\simeq u(ih,jk)$, y $A^{(2)}=d\cdot T_d^{(2)}+a\cdot\dfrac{h}{2}\cdot T_a^{(2)}+r\cdot h^2\cdot I_{N^2}$, siendo $T_d^{(2)}$ la matriz de dimensión $N^2\times N^2$ dada en (3.10) y $T_a^{(2)}$ como en (3.10) pero con elementos diagonales 0 y elementos en la parte triangular inferior estricta iguales a $+1$. Por bloques, las matrices $T_d^{(2)}$ y $T_a^{(2)}$ se escriben como:

$$T_d^{(2)}=\left[\begin{array}{c|c|c|c|c|c} \widehat{T}_d^{(1)} & -I & 0 & \cdots & 0 & 0\\ \hline -I & \widehat{T}_d^{(1)} & -I & \cdots & 0 & 0\\ \hline & \ddots & \ddots & \ddots & & \\ \hline & & \ddots & \ddots & \ddots & \\ \hline & & & \ddots & \ddots & -I\\ \hline 0 & 0 & & & -I & \widehat{T}_d^{(1)} \end{array}\right],T_a^{(2)}=\left[\begin{array}{c|c|c|c|c|c} T_a^{(1)} & -I & 0 & \cdots & 0 & 0\\ \hline I & T_a^{(1)} & -I & \cdots & 0 & 0\\ \hline & \ddots & \ddots & \ddots & & \\ \hline & & \ddots & \ddots & \ddots & \\ \hline & & & \ddots & \ddots & -I\\ \hline 0 & 0 & & & I & T_a^{(1)} \end{array}\right], \quad (3.9)$$

con $I = I_N$ y $\widehat{T}_d^{(1)} = Tridiag(-1, 2n, -1)$, para $n = 2$.

$$T_d^{(2)} = \left[\begin{array}{cccc|cccc|cccc|cccc} 4 & -1 & & & -1 & & & & & & & & & & & \\ -1 & \ddots & \ddots & & & \ddots & & & & & & & & & & \\ & \ddots & \ddots & -1 & & & \ddots & & & & & & & & & \\ & & -1 & 4 & & & & -1 & & & & & & & & \\ \hline -1 & & & & 4 & -1 & & & \ddots & & & & & & & \\ & \ddots & & & -1 & \ddots & \ddots & & & \ddots & & & & & & \\ & & \ddots & & & \ddots & \ddots & -1 & & & \ddots & & & & & \\ & & & -1 & & & -1 & 4 & & & & \ddots & & & & \\ \hline & & & & \ddots & & & & \ddots & & & & -1 & & & \\ & & & & & \ddots & & & & \ddots & & & & \ddots & & \\ & & & & & & \ddots & & & & \ddots & & & & \ddots & \\ & & & & & & & \ddots & & & & \ddots & & & & -1 \\ \hline & & & & & & & & -1 & & & & 4 & -1 & & \\ & & & & & & & & & \ddots & & & -1 & \ddots & \ddots & \\ & & & & & & & & & & \ddots & & & \ddots & \ddots & -1 \\ & & & & & & & & & & & -1 & & & -1 & 4 \end{array}\right] \tag{3.10}$$

De forma más compacta, usando el producto de Kronecker de matrices, con $A = (a_{ij}) \in \mathbb{R}^{n\times m}, C \in \mathbb{R}^{p\times q}$,

$$A \otimes C = (a_{ij}C) = \left[\begin{array}{c|c|c|c} a_{11}C & a_{12}C & \cdots & a_{1m}C \\ \hline a_{21}C & a_{22}C & \cdots & a_{2m}C \\ \hline \vdots & \vdots & & \vdots \\ \hline a_{n1}C & a_{n2}C & \cdots & a_{nm}C \end{array}\right] \in \mathbb{R}^{np\times mq},$$

la matriz $A^{(2)}$ en (3.8)-(3.9) se escribe como

$$\begin{aligned} A^{(2)} &= d\left(I_N \otimes T_d^{(1)} + T_d^{(1)} \otimes I_N\right) + \frac{a}{2}\left(I_N \otimes T_a^{(1)} + T_a^{(1)} \otimes I_N\right) + r \cdot h^2 \cdot I_{N^2} = \\ &= I_N \otimes \left[dT_d^{(1)} + a\frac{h}{2}T_a^{(1)} + \frac{r}{2}h^2 I_N\right] + \left[dT_d^{(1)} + a\frac{h}{2}T_a^{(1)} + \frac{r}{2}h^2 I_N\right] \otimes I_N, \end{aligned}$$

donde $T_d^{(1)}$ y $T_a^{(1)}$ fueron definidas en (3.6).

De modo similar, para la ecuación tridimensional (3.1) con condiciones de frontera de tipo Dirichlet:

$$\begin{cases} -d\left(u_{xx} + u_{yy} + u_{zz}\right) + a\left(u_x + u_y + u_z\right) + r \cdot u = f, & (x, y, z) \in \Omega = (0, 1)^3, \\ u = g, & (x, y, z) \in \Gamma = \partial\Omega, \end{cases} \tag{3.11}$$

al usar diferencias centrales de orden 2 en la malla $\Omega_h^{(3)} = \{(ih, jh, kh)\}_{i,j,k=0}^{N+1}$, con $h = \dfrac{1}{N+1}$, se obtiene el esquema en diferencias

$$d\cdot[(-U_{i+1,j,k}+2U_{i,j,k}-U_{i-1,j,k})+(-U_{i,j+1,k}+2U_{i,j,k}-U_{i,j-1,k})+$$
$$(-U_{i,j,k+1}+2U_{i,j,k}-U_{i,j,k-1})]++a\frac{h}{2}[(U_{i+1,j,k}-U_{i-1,j,k})+$$
$$(U_{i,j+1,k}-U_{i,j-1,k})+(U_{i,j,k+1}-U_{i,j,k-1})]+r\cdot h^2\cdot U_{i,j,k}=f_{i,j,k},\quad 1\le i,j,k\le N,$$

que produce un sistema lineal de dimensión N^3 con incógnitas $U_{i,j,k}\simeq u\,(ih,jh,kh)$, $1\le i,j,k\le N$. De hecho, en forma matricial, este sistema se escribe como

$$A^{(3)}\cdot U=b, \tag{3.12}$$

donde nuevamente b contiene la discretización del dato $f(x,y,z)$ en los puntos interiores de la malla, así como de las condiciones de frontera $g(x,y,z)$, $U=(U_{i,j,k})_{i,j,k=1}^{N}\in\mathbb{R}^{N^3}$ ordenado de acuerdo al orden lexicográfico y $A^{(3)}\in\mathbb{R}^{N^3\times N^3}$ dada por $A^{(3)}=dT_d^{(3)}+a\frac{h}{2}T_a^{(3)}+r\cdot h\cdot I_{N^3}$, siendo

$$T_d^{(3)}=\left[\begin{array}{c|c|c|c|c|c}
T_d^{(2)} & -I & 0 & \cdots & 0 & 0\\ \hline
-I & T_d^{(2)} & -I & \cdots & 0 & 0\\ \hline
 & \ddots & \ddots & \ddots & & \\ \hline
 & & \ddots & \ddots & \ddots & \\ \hline
 & & & \ddots & \ddots & -I\\ \hline
0 & 0 & & & -I & T_d^{(2)}
\end{array}\right],\ T_a^{(3)}=\left[\begin{array}{c|c|c|c|c|c}
T_a^{(2)} & -I & 0 & \cdots & 0 & 0\\ \hline
I & T_a^{(2)} & -I & \cdots & 0 & 0\\ \hline
 & \ddots & \ddots & \ddots & & \\ \hline
 & & \ddots & \ddots & \ddots & \\ \hline
 & & & \ddots & \ddots & -I\\ \hline
0 & 0 & & & I & T_a^{(2)}
\end{array}\right],\ I=I_{N^2},$$

con $T_d^{(2)}$ y $T_a^{(2)}$ dadas como en (3.9), con $n=3$ para $\widehat{T}_d^{(1)}$.
De forma más compacta, usando el producto de Kronecker de matrices:

$$\begin{aligned}
A^{(3)} &= d\left[I_N\otimes I_N\otimes T_d^{(1)}+I_N\otimes T_d^{(1)}\otimes I_N+T_d^{(1)}\otimes I_N\otimes I_N\right]\\
&+\tfrac{ah}{2}\left[I_N\otimes I_N\otimes T_a^{(1)}+I_N\otimes T_a^{(1)}\otimes I_N+T_a^{(1)}\otimes I_N\otimes I_N\right]\\
&=I_N\otimes I_N\otimes\left[dT_d^{(1)}+\tfrac{ah}{2}T_a^{(1)}+\tfrac{rh^2}{3}I_N\right]+I_N\otimes\left[dT_d^{(1)}+\tfrac{ah}{2}T_a^{(1)}+\tfrac{rh^2}{3}I_N\right]\otimes I_N\\
&+\left[dT_d^{(1)}+\tfrac{ah}{2}T_a^{(1)}+\tfrac{rh^2}{3}I_N\right]\otimes I_N\otimes I_N.
\end{aligned} \tag{3.13}$$

Observación 3.1. A continuación, presentamos tres ejemplos en dimensiones $n=2$ y $n=3$ en los que la solución de la EDP (3.1) es un polinomio de grado menor o igual que dos en cada variable espacial, de tal modo que la discretización espacial mediante diferencias centrales de orden dos sea exacta y la solución exacta del sistema lineal $A^{(n)}\cdot U=b$ (véase (3.8) y (3.12)) es precisamente $U=u_{|\Omega_h^{(n)}}$, es decir, la restricción de la solución exacta de la EDP a la malla espacial $\Omega_h^{(n)}$. El vector U se usará como solución de referencia para medir el error de las iteraciones provistas por los métodos numéricos de los capítulos anteriores. Respecto al dato f de la EDP, éste se obtiene a partir de la solución exacta u directamente de (3.1). Además, el vector b, en los sistemas lineales (3.8) y (3.12), que contiene la discretización del dato f y las condiciones de frontera, se ha obtenido directamente en *Matlab* a partir del vector U y la matriz $A^{(n)}$ como $b=A^{(n)}U$. En la resolución numérica de $A^{(n)}\cdot U=b$, hemos considerado como valor inicial $\widehat{U}^{(0)}=0$ para todos los métodos numéricos. Además, los tiempos de CPU (en segundos) que figuran en las siguientes gráficas corresponden a un procesador *Intel Core i5* a 2.8 GHz y 8 Gb de memoria RAM.

3.1.1. Ejemplo 1.

La ecuación de Poisson con condiciones de frontera de tipo Dirichlet homogéneas en $[0,1]^n$, $n = 2, 3$, con solución exacta $u(x_1, \cdots, x_n) = 4^n \cdot \prod_{i=1}^{n} x_i(1 - x_i)$.

Este ejemplo se corresponde con la EDP (3.1) con $d = 1$ y $a = r = 0$. Para las dimensiones $n = 2$ y $n = 3$, la discretización de (3.1) mediante diferencias centrales de orden 2 produce los sistemas lineales (3.8) y (3.12) $A^{(n)} \cdot U = b$, de dimensión N^n, con $A^{(n)}$ simétrica y definida positiva. Tomamos, en cada caso, $N = 100$.

- En dimensión $n = 2$, mostramos en la Figura 3.1 el comportamiento del error $\left\|U - \widehat{U}\right\|_\infty$ en relación al número de iteraciones y al tiempo de cómputo, respectivamente. $\widehat{U}$ denota la aproximación provista por cada uno de los métodos numéricos en cada iteración, con un total de 1000 iteraciones. Hemos considerado la comparación de los métodos de Gauss-Seidel (GS, 1.2), del Descenso Más Rápido (DR, 1.10), del Residual Mínimo (MR, 1.13), del Gradiente Conjugado GC (2.4) y GMRES (2.16).

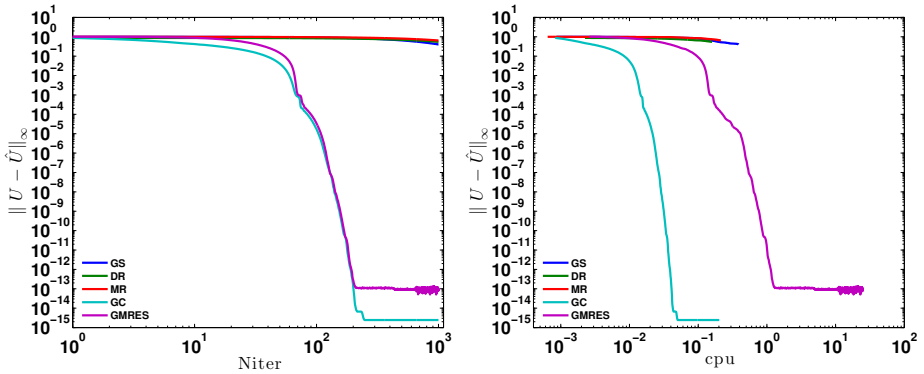


Figura 3.1. Error frente a número de iteraciones (izquierda) y frente al tiempo de CPU en segundos (derecha) de los métodos de Gauss-Seidel(GS), del descenso más rápido (DR), del residual mínimo (MR), del Gradiente Conjugado (GC) y GMRES aplicados a la ecuación de Poisson 2D (3.1) con $d = 1$, $a = r = 0$ y solución exacta con condiciones de frontera de tipo Dirichlet homogéneas $u(x, y) = 4^2 x(1 - x)y(1 - y)$ en la malla $\Omega_h^{(2)}$ con $N = 100$.

Observamos en la Figura 3.1 (izquierda) que la convergencia de GS, DR y MR en este problema de alta dimensión es bastante lenta, mientras que GC y GMRES proveen de una solución numérica bastante precisa en 110 iteraciones. Además, GC resulta claramente más eficiente que GMRES, alcanzando la solución exacta en apenas $5 \cdot 10^{-2}$ segundos (ver Figura 3.1, derecha).

- En dimensión $n = 3$, mostramos en la Figura 3.2 el comportamiento del error $\left\|U - \widehat{U}\right\|_\infty$ para los métodos GC y GMRES, y sus variantes con reinicialización GMRES(m) tras m iteraciones, con $m = 10$, 50 y 100, dando un total de 300 iteraciones. En este caso, la matriz $A^{(3)}$ es simétrica, definida positiva y de dimensión $N^3 = 10^6$. Observamos nuevamente en la Figura 3.2 que el decaimiento del error con las iteraciones es similar para GC y GMRES (ver Fig 3.2, izquierda), siendo el primero más eficiente (ver Fig 3.2, derecha). En este caso, GC da una

aproximación con error absoluto de magnitud 10^{-14} en un tiempo aproximado de 10 segundos. Asimismo, observamos que en esta EDP con condiciones de frontera homogéneas, el método GMRES parece ligeramente más eficiente que sus variantes con reinicialización.

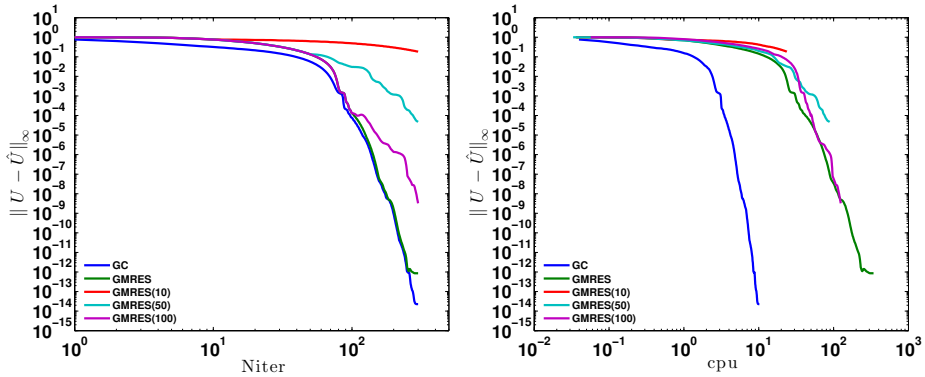


Figura 3.2. Error frente a número de iteraciones (izquierda) y frente al tiempo de CPU en segundos (derecha) de los métodos GC, GMRES y GMRES(m), $m = 10$, 50 y 100, aplicados a la ecuación de Poisson 3D (3.1) con $d = 1$, $a = r = 0$ y solución exacta con condiciones de frontera de tipo Dirichlet homogéneas $u(x, y, z) = 4^3x(1 - x)y(1 - y)z(1 - z)$ en la malla $\Omega_h^{(3)}$ con N=100.

3.1.2. Ejemplo 2

La ecuación de Poisson con condiciones de frontera de tipo Dirichlet no homogéneas en $[0, 1]^3$ con solución exacta $u(x, y, z) = x^2 + y^2 + z^2$. Nuevamente, este ejemplo sigue la EDP modelo (3.1) con $d = 1$ y $a = r = 0$ y, con $N = 100$, la discretización mediante diferencias centrales de orden dos produce el sistema lineal (3.12) $A^{(3)} \cdot U = b$, con matriz $A^{(3)}$ simétrica y definida positiva de dimensión $N^3 = 10^6$ dada por (3.13). En la Figura 3.3 presentamos el comportamiento del error $\left\|U - \widehat{U}\right\|_\infty$ en función del número de iteraciones y del tiempo de CPU, con un máximo de 300 iteraciones, para los métodos GC, GMRES y GMRES(m), con $m = 10$, 50 y 100. Observamos que la imposición de condiciones de frontera no homogéneas exige a los métodos numéricos un mayor número de iteraciones para disminuir el error. En particular, observamos en la Figura (3.3) (derecha) que GC y GMRES necesitan alrededor de 10 y 300 segundos, respectivamente, para proveer un error de magnitud 10^{-6}. Asimismo, observamos que las variantes con reinicialización GMRES(m) resultan ligeramente más eficientes que GMRES porque para una misma magnitud de error requieren menor tiempo de cómputo.

A efectos de mejorar la eficiencia y acelerar la convergencia de los métodos numéricos, en este mismo ejemplo hemos considerado las variantes de GC, GMRES y GMRES(10) con precondicionamiento (ver Sección 2.4). Puesto que la matriz $A^{(3)}$ es simétrica, hemos considerado dos opciones para el precondicionador $M = L \cdot L^T$: en primer lugar, el provisto por la descomposición de Cholesky incompleta IC(0); y, en segundo lugar, el provisto por la descomposición de Cholesky incompleta modificada MIC(0).

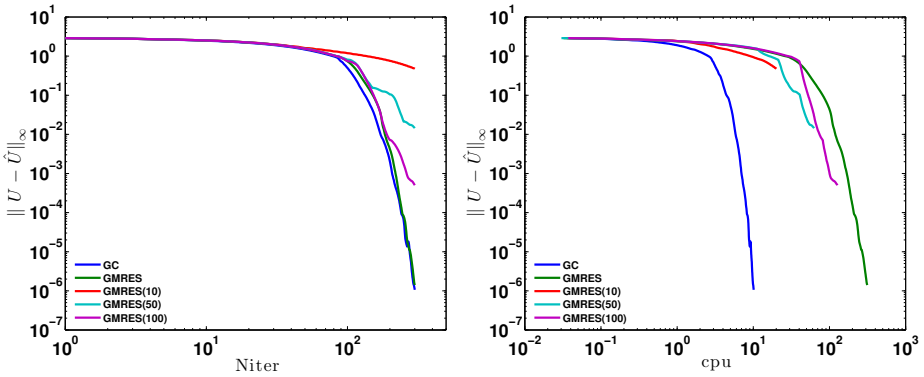


Figura 3.3. Error frente a número de iteraciones (izquierda) y frente al tiempo de CPU en segundos (derecha) de los métodos GC, GMRES y GMRES(m), $m = 10$, 50 y 100, aplicados a la ecuación de Poisson 3D (3.1) con d=1, a=r=0 y solución exacta con condiciones de frontera de tipo Dirichlet no homogéneas $u(x, y, z) = x^2 + y^2 + z^2$ en la malla $\Omega_h^{(3)}$ con $N = 100$.

Observamos en la Figura 3.4 cómo ambas opciones de precondicionamiento aceleran la convergencia en comparación con los correspondientes métodos originales sin precondicionar. En este caso, observamos en la Figura 3.4 (derecha) que el método del Gradiente Conjugado precondicionado con MIC(0), esto es, PGC-MIC(0), resulta la opción más eficiente, prácticamente convergiendo a la solución exacta en 7 segundos y 100 iteraciones. A su vez, observamos que, para cada método, el precondicionador MIC(0) resulta más eficiente que el precondicionador IC(0).

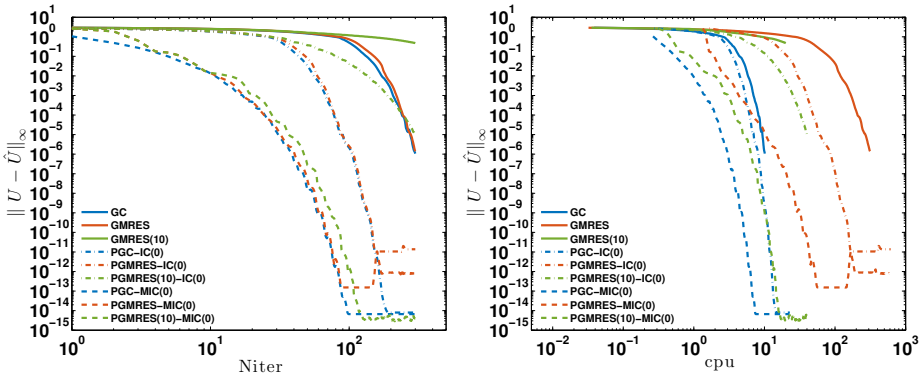


Figura 3.4. Error frente a número de iteraciones (izquierda) y frente al tiempo de CPU en segundos (derecha) de los métodos GC, GMRES, GMRES(10) y sus variantes con precondicionador Cholesky incompleto IC(0) y Cholesky incompleto modificado MIC(0), aplicados a la ecuación de Poisson 3D (3.1) con d=1, a=r=0 y solución exacta con condiciones de frontera de tipo Dirichlet no homogéneas $u(x, y, z) = x^2 + y^2 + z^2$ en la malla $\Omega_h^{(3)}$ con $N = 100$.

3.1.3. Ejemplo 3

La ecuación de difusión-convección-reacción (3.1), con $d = 1$, $a = 100$ y $r = -300$ (convección-reacción dominante) con condiciones de frontera de tipo Dirichlet no homogéneas en $[0, 1]^3$ y solución exacta $u(x, y, z) = x^2 + y^2 + z^2$.
Con $N = 100$ en la malla $\Omega_h^{(3)}$, la discretización mediante diferencias centrales de orden 2 produce el sistema lineal (3.12) $A^{(3)} \cdot U = b$, donde ahora la matriz $A^{(3)}$ dada en (3.13) tiene dimensión $N^3 = 10^6$ pero es no simétrica. En tal caso, el método del Gradiente Conjugado GC no tiene porqué estar bien definido ni porqué converger (de hecho, en este problema concreto, diverge). Siguiendo el mismo criterio que en el Ejemplo (3.1.2), procedemos de modo análogo pero sustituyendo el método GC por su variante para ecuaciones normales GCNR. En la Figura 3.5, presentamos el comportamiento del error en términos del número de iteraciones y de tiempo de CPU para los métodos GCNR, GMRES y sus variantes con reinicialización GMRES(m), $m = 10$, 50 y 100. Observamos que GMRES necesita más de 200 iteraciones y 600 segundos para una aproximación con error absoluto de 10^{-5}. Los restantes métodos necesitan más de 300 iteraciones para que se observe un decaimiento del error. En este ejemplo, se aprecia claramente la necesidad de encontrar un buen precondicionador que permita acelerar la convergencia de los métodos numéricos.

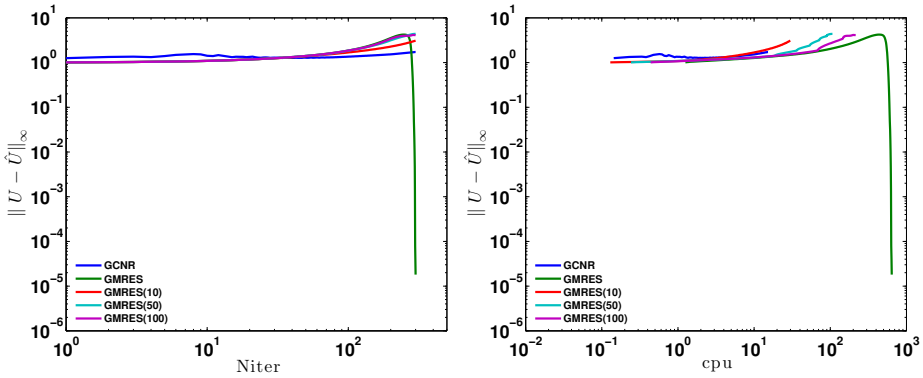


Figura 3.5. Error frente a número de iteraciones (izquierda) y frente al tiempo de CPU en segundos (derecha) de los métodos GCNR, GMRES y GMRES(m), $m = 10$, 50 y 100, aplicados a la ecuación (3.1) con $d = 1$, $a = 100$, $r = -300$ y solución exacta con condiciones de frontera de tipo Dirichlet no homogéneas $u(x, y, z) = x^2 + y^2 + z^2$ en la malla $\Omega_h^{(3)}$ con $N = 100$.

Finalmente, para mejorar el rendimiento de los métodos sobre este sistema lineal, hemos considerado sus variantes precondicionadas, considerando como precondicionadores los provistos por la descomposición LU incompleta ILU(0) y su versión modificada MILU(0) (ver Sección 2.4): $M = L \cdot U$. En el caso del método GCNR, el correspondiente precondicionador se ha elegido como $M = (L \cdot U)^T (L \cdot U)$, donde L y U forman una descomposición incompleta de $A^{(3)}$. Observamos en la Figura 3.6 cómo ambas opciones de precondicionamiento mejoran el rendimiento de los métodos originales, resultando el precondicionador MILU(0) más eficiente que ILU(0). En particular, sobre este sistema lineal GMRES(10) precondicionado con MILU(0) resulta la opción más eficiente, prácticamente convergiendo a la solución exacta en tan solo 11 iteraciones y 3 segundos. Observamos, además, que para cada opción de precondicionamiento, GCNR resulta menos eficiente debido a los productos adicionales requeridos por la matriz $A^{(3)^T}$.

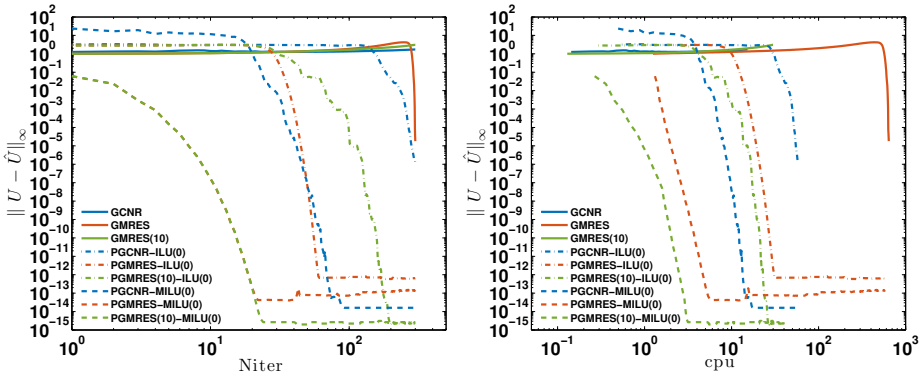


Figura 3.6. Error frente a número de iteraciones (izquierda) y frente al tiempo de CPU en segundos (derecha) de los métodos GCNR, GMRES, GMRES(10) y sus variantes con precondicionador LU incompleto ILU(0) y LU incompleto modificado MILU(0), aplicados a la ecuación (3.1) con $d = 1$, $a = 100$, $r = -300$ y solución exacta con condiciones de frontera de tipo Dirichlet no homogéneas $u(x, y, z) = x^2 + y^2 + z^2$ en la malla $\Omega_h^{(3)}$ con $N = 100$.

3.2. Conclusiones

Los métodos de proyección que se han presentado en este estudio parecen ser tan eficaces como sensibles a la naturaleza del sistema lineal a resolver. Su comportamiento queda recogido en los ejemplos pertenecientes al último capítulo de este trabajo, donde se aplican a Ecuaciones en Derivadas Parciales lineales con coeficientes constantes en n dimensiones espaciales, discretizadas mediante Diferencias Finitas.

Se intuye que el método del Gradiente Conjugado (GC) resulta el más eficiente cuando las condiciones de frontera son homogéneas y la matriz A es simétrica y definida positiva, tanto en dimensión $n = 2$ como $n = 3$, obteniendo errores del orden 10^{-15} necesitando entre $5 \cdot 10^{-2}$ y 10 segundos. Esta mejora del GC se extiende también si consideramos condiciones no homogéneas, pero se observa claramente cómo el número de iteraciones y de tiempo de cómputo requerido crece considerablemente, motivo por el cual se introduce como precondicionador la descomposición de Cholesky incompleta IC(0) y su variante modificada MIC(0), obteniendo esta última los mejores resultados cuando se le aplica al Gradiente Conjugado.

Sin embargo, estos efectos cambian cuando en el tercer ejemplo se emplea una matriz A que no es simétrica, lo cual provoca que el GC no tenga porqué estar bien definido ni converger, de hecho diverge. En este caso, se emplea su variante para Ecuaciones Normales (GCNR) y se observa que únicamente el GMRES presenta un notable decaimiento del error con respecto al resto, aunque necesite para ello más de 200 iteraciones y 600 segundos, por lo que, de nuevo, parece necesario encontrar un buen precondicionador. Así, se dota a los métodos de los precondicionadores dados por la descomposición LU incompleta ILU(0) y su versión modificada MILU(0), los cuales mejoran considerablemente los resultados iniciales y es aquí donde parece destacar el método GMRES con reinicialización cada 10 iteraciones precondicionado con MILU(0).

En cualquier caso, el objetivo inicial de este estudio queda ilustrado en estas gráficas, donde se refleja el comportamiento que tienen los métodos de proyección basados en subespacios de Krylov cuando se enfrentan a grandes sistemas lineales, mostrando qué condiciones son las que más favorecen a cada uno.

Bibliografía

[1] Axelsson, O. *Iterative solution methods.* Cambridge University Press (1996).

[2] Barnett, S. *Matrices. Methods and applications.* Oxford University Press (1996).

[3] Barrett, R. et al. *Templates for the solution of linear systems: building blocks for iterative systems.* SIAM (1994).

[4] Chan, T.F., y Van der Vorst, H.A. *Approximate and incomplete factorizations.* Parallel Numerical Algorithms, ICASE/LaRC Interdiscip. Ser. Sci. Engrg. 4, D. E. Keyes, A. Sameh, and V. Venkatakrishnan, eds., Kluwer Academic, Dordecht, The Netherlands, p. 167-202 (1997).

[5] Greenbaum, A. *Iterative methods for solving linear systems.* SIAM (1987).

[6] Hackbusch, W. *Iterative Solution of Large Sparse Systems of Equations.* Springer (2016)

[7] Plato, R. *Concise numerical mathematics.* AMS (2003).

[8] Saad, Y. *Iterative methods for sparse linear systems.* SIAM (2003).

[9] Trefethen, L.N., Bau, D. *Numerical Linear Algebra.* SIAM (1997).

Krylov Methods for linear systems

Sección de Matemáticas
Universidad de La Laguna

Christian Manuel Bartolomé Moreno
Facultad de Ciencias · Sección de Matemáticas
Universidad de La Laguna
alu0100964240@ull.edu.es

Abstract

Krylov subspaces, which owe their name to the Russian mathematician Alekséi Krylov are today the basis on which modern iterative methods are based when calculating vectors and eigenvalues or solving systems of linear equations $Ax = b$, using a smaller amount of memory and CPU time than the rest of conventional methods. Without going any further, the numerical methods of Gauss-Seidel and Jacobi, which are taught throughout the degree, show a rather poor performance when applied to differential problems in two and three spatial dimensions. Thus, algorithms based on these subspaces, called Krylov subspace methods, have great acceptance in numerical linear algebra. Among them, the best known are the Conjugate Gradient (GC) method with its variant for normal equations (GCNR) and the Generalized Minimum Residual Method (GMRES), being their efficiency and convergence rate the key aspects of the study undertaken in this work. All this will be reflected graphically throughout the last chapter, which will detail the behavior of these methods applied to linear elliptic Partial Differential Equations and comparing how the results vary when applying preconditioning to the matrix resulting after spatial discretization by means of Finite Differences.

1. Projection methods: definition and properties

If we denote the exact solution of the problem $x^* = A^{-1}b$ and $x^{(0)} \in \mathbb{R}^N$ an initial approximation to x^*, we then define the initial residual vector as:

$$r^{(0)} := A(x - x^{(0)}) = b - Ax^{(0)}.$$

Let $\mathcal{K}$ y $\mathcal{L}$ be vectorial subspaces of$\mathbb{R}^{\mathbb{N}}$. The projection method on $\mathcal{K}$ orthogonal to $\mathcal{L}$ consists of:

$$\text{find } \delta \in \mathcal{K} : \langle r^{(0)} - A\delta, w\rangle_2 = 0, \forall w \in \mathcal{L} \quad \text{and take } x^{(1)} := x^{(0)} + \delta, \tag{1}$$

being $r^{(0)} := b - Ax^{(0)}$ the residual of the approximation $x^{(0)}$ and $\langle .,. \rangle_2$ the Euclidean product of $\mathbb{R}^{\mathbb{N}}$.

Proposition 1 *Let A symmetric and positive definite, $\mathcal{L} = \mathcal{K}$,$x^{(1)} = x^{(0)} + \delta$, with $\delta \in \mathcal{K}$, and $r^{(0)} = b - Ax^{(0)}$. Let x^* be the solution of $Ax = b$ and let $E(x) := \|x^* - x\|_A^2$, $x \in \mathbb{R}^{\mathbb{N}}$. Then,*

$$\langle r^{(0)} - A\delta, v\rangle_2 = 0, \forall v \in \mathcal{K} \Leftrightarrow x^{(1)} \text{ minimizes } E(x) \text{ on } x^{(0)} + \mathcal{K}. \tag{2}$$

Moreover

$$\|x^* - x^{(1)}\|_A \leq \|x^* - x^{(0)}\|_A.$$

Another interesting class of projection methods arises when A is arbitrary and $\mathcal{L} = A\mathcal{K}$.

Proposition 2 *Let $A \in \mathbb{R}^{\mathbb{N}\times\mathbb{N}}$, $\mathcal{L} = A\mathcal{K}$,$x^{(1)} = x^{(0)} + \delta$, with $\delta \in \mathcal{K}$, and $r^{(0)} = b - Ax^{(0)}$. So,*

$$\langle r^{(0)} - A\delta, w\rangle_2 = 0, \forall w \in \mathcal{L} \Leftrightarrow x^{(1)} \text{ minimizes R(x) on } x^{(0)} + \mathcal{K}. \tag{3}$$

Moreover

$$\|b - Ax^{(1)}\|_2 \leq \|b - Ax^{(0)}\|_2.$$

Some examples are the Deepest Descent Method, the Minimum Residual Method and the Deepest Descent Method for Normal Equations.

2. Projection methods based on Krylov subspaces

Definition 3 *Let $A \in \mathbb{R}^{\mathbb{N}\times\mathbb{N}}$ be a matrix and $r^{(0)} \in \mathbb{R}^{\mathbb{N}}$ a vector. We define the Krylov Subspaces sequence associated to A and $r^{(0)}$, with $n = 0, 1, ...$ as:*

$$\mathcal{K}_n\left(A, r^{(0)}\right) = span\left\{r^{(0)}, Ar^{(0)}, ..., A^{n-1}r^{(0)}\right\} \subset \mathbb{R}^{\mathbb{N}}. \tag{4}$$

Note that $\{0\} := \mathcal{K}_0(A, r^{(0)}) \subset \mathcal{K}_1(A, r^{(0)}) \subset \mathcal{K}_2(A, r^{(0)}) \subset ...$

Definition 4 *For every $n \geq 1$, the Conjugate Gradient method produces an approximation $x^{(n)} \in \mathbb{R}^{\mathbb{N}}$ using the orthogonal projection method (1) over $\mathcal{K} = \mathcal{K}_n\left(A, r^{(0)}\right)$, being $\mathcal{K}_n\left(A, r^{(0)}\right) = span\left\{r^{(0)}, Ar^{(0)}, ..., A^{n-1}r^{(0)}\right\}$ the nth Krylov subspace generated by A y $r^{(0)}$.*

Then,

$$x^{(n)} = x^{(0)} + \delta^{(n)}, \text{ with } \delta^{(n)} \in \mathcal{K}_n\left(A, r^{(0)}\right) \text{ such that } r^{(n)} \in \mathcal{K}_n\left(A, r^{(0)}\right)^{\perp}, \tag{5}$$

being $r^{(n)} = b - Ax^{(n)} = r^{(0)} - A\delta^{(n)}$.

The Conjugate Gradient version for Normal Equations consists on applying the CG method to the normal equation systems $A^TAx = A^Tb$. In this way, the GCNR method considers the orthogonal projection about Krylov subspaces $\mathcal{K}_n(A^TA, A^Tr^{(0)})$.

Definition 5 *For every $n \geq 1$, the Generalized Minimal Residual Method produces an approximation $x^{(n)} \in \mathbb{R}^{\mathbb{N}}$ using the projection method about $\mathcal{K} = \mathcal{K}_n\left(A, r^{(0)}\right)$ ortogonal t $\mathcal{L} = A\mathcal{K}$, being $\mathcal{K}_n\left(A, r^{(0)}\right)$ the nth Krylov subspace generated by A y $r^{(0)}$.*

Then:

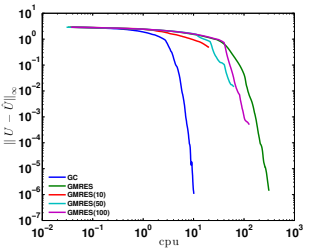
$$x^{(n)} = x^{(0)} + \delta^{(n)}, \text{ with } \delta^{(n)} \in \mathcal{K}_n\left(A, r^{(0)}\right) \text{ such that } r^{(n)} \in \mathcal{L}^{\perp}, \tag{6}$$

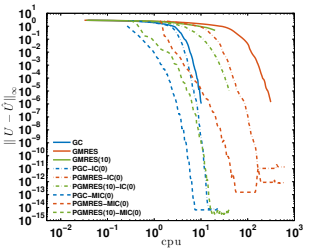
being $r^{(n)} = b - Ax^{(n)} = r^{(0)} - A\delta^{(n)}$.

As it is concluded that the GC, GCNR and GMRES methods can see their convergence slowed down in case the condition number, with respect to the Euclidean norm, of a certain matrix related to the system coefficient matrix $Ax = b$ is large enough, the alternative to precondition the linear system $Ax = b$ is introduced. This consists in considering a regular matrix M, called preconditioner, to convert the system $Ax = b$ into the equivalent system $M^{-1}Ax = M^{-1}b$ so that, in certain sense that we will not need here, M is close to A and the systems resolution $My = c$ is inexpensive.

3. Numerical illustration

For example, using the Poisson equation with non-homogeneous Dirichlet boundary conditions in $[0, 1]^3$ with exact solution $u(x, y, z) = x^2 + y^2 + z^2$, we can observe the following difference in performance when we use the methods naturally and with preconditioning.





References

[1] Plato, R. *Concise numerical mathematics.* AMS (2003).

[2] Saad, Y. *Iterative methods for sparse linear systems.* SIAM, Springer (2003).