Universidad de La Laguna

SECCIÓN DE MATEMÁTICAS | FACULTAD DE CIENCIAS

Ana Gómez Pérez

Métodos iterativos para sistemas de ecuaciones lineales

Iterative methods for systems of linear equations

Trabajo Fin de Grado
Grado en Matemáticas
La Laguna, Julio de 2018

Dirigido por

Domingo Hernández Abreu

Agradecimientos

A Domingo Hernández Abreu por su excelente orientación
durante todo el proceso de realización del proyecto.
A mi familia y amigos por su apoyo incondicional.

Resumen · Abstract

Resumen

En la presente memoria se estudian diversos métodos iterativos para resolver sistemas lineales de ecuaciones. Inicialmente se incluye una breve introducción sobre los métodos directos. A continuación se pasa a analizar los métodos iterativos clásicos de Jacobi y Gauss-Seidel así como la convergencia, particularizando en el análisis de las matrices estrictamente diagonal dominantes e irreduciblemente diagonal dominantes. Asimismo, se considera el método de relajación, estudiando su convergencia en función del parámetro de relajación, en concreto para matrices consistentemente ordenadas. También se introduce otra clase de métodos iterativos basados en subespacios de Krylov: el método del Gradiente Conjugado para matrices simétricas y definidas positivas, y el método del residual mínimo generalizado para cualquier matriz arbitraria. Finalmente se presentan algunos ejemplos comparando los diferentes métodos según su tiempo CPU, error y número de iteraciones para sistemas lineales resultantes de la discretización espacial de la ecuación en derivadas parciales con coeficientes constantes de tipo difusión-convección-reacción.

Palabras clave: *Métodos iterativos-Jacobi-Gauss-Seidel-Subespacios de Krylov-Gradiente Conjugado-Residual mínimo generalizado-Precondicionamiento-Discretización espacial de EDPs.*

Abstract

In this project several iterative methods are studied to solve linear systems of equations. First, a brief introduction about the direct methods are included. Next, the classic iterative methods of Jacobi and Gauss-Seidel are considered. A convergence analysis focusing on strictly diagonally dominant and irreducibly diagonally dominant matrices is presented. Likewise, the relaxation method is considered by studying its convergence as a function of the relaxation parameter, specifically, for consistently ordered matrices. Also, other classes of iterative methods based on Krylov subspaces are introduced: the Conjugate Gradient method, for symmetric and positive definite matrices, and the generalized minimal residual method for any nonsingular matrix. Finally, some examples are illustrated comparing the different methods according to their CPU time, error and number of iterations for linear systems arising from the spatial discretization of diffusion-convection-reaction partial differential equations with constant coefficients.

Keywords: *Iterative methods-Jacobi-Gauss-Seidel-Krylov subspaces-Gradient Conjugate-Generalized minimal residual-Preconditioning-Spatial discretization of PDEs.*

Contenido

1

Métodos iterativos básicos

1.1. Introducción

A lo largo de este trabajo estudiaremos diferentes métodos para resolver grandes sistemas lineales de ecuaciones

$$Ax = b \qquad \text{con A} \in \mathbb{R}^{\mathrm{N}\times\mathrm{N}} \text{ regular, b} \in \mathbb{R}^{\mathrm{N}}, \tag{1.1}$$

o bien, en forma desarrollada

$$\sum_{j=1}^{N} a_{kj}x_j = b_k, \qquad k = 1, \ldots, N,$$

con $x_* = A^{-1}b \in \mathbb{R}$ solución única. Mediante métodos directos se obtiene la solución exacta en un número finito de pasos. El método directo más clasico es el método de la Eliminación Gaussiana, el cual consiste en descomponer la matriz A en el producto LU donde L es una matriz triangular inferior y U es matriz triangular superior, de modo que el sistema $Ax = b$ resulta equivalente a dos sistemas triangulares $Ly = b$ y $Ux = y$ que a su vez se resuelven por sustitución hacia delante y hacia atrás. En [3, Cap.3-4] pueden encontrarse diversos resultados que permiten garantizar la existencia de una descomposición LU bajo condiciones adecuadas. En particular, de [3, p.144], si $det(A(1:k, 1:k)) \neq 0$ para $k = 1, \ldots, N$, entonces existe una única descomposición $A = LU$, con $l_{ii} = 1$, $1 \leq i \leq N$.

Por otra parte, si A es una matriz simétrica y definida positiva, es decir, $A^\top = A$ y $w^\top Aw > 0$ para todo $w \in \mathbb{R}^N \backslash \{0\}$, entonces A se puede factorizar de forma única según la descomposición de Cholesky como $A = LL^\top$ donde L es una matriz triangular inferior con elementos diagonales positivos [3, p.163]. Por tanto, determinar la solución del sistema $Ax = b$ equivale a resolver dos sistemos triangulares $Ly = b$ y $L^\top x = y$.

Tanto la descomposición LU como la de Cholesky requieren $\mathcal{O}(N^3)$ operaciones.

Otro tipo diferente de descomposición para una matrix rectangular A se denomina descomposición QR, que consiste en factorizar $A \in \mathbb{R}^{M\times N}$ en un producto de una matriz ortogonal $Q \in \mathbb{R}^{M\times M}$ y una matriz triangular $R \in \mathbb{R}^{M\times N}$, (ver [3, p.245]). Recordemos que una matriz $Q \in \mathbb{R}^{M\times M}$ se dice ortogonal si $QQ^\top = Q^\top Q = I$. Además, la matriz rectangular $R \in \mathbb{R}^{M\times N}$ se dice triangular superior si $R_{ij} = 0$, para todo $i > j$.

En particular, esta descomposición QR se utiliza para hallar las soluciones de problemas de mínimos cuadrados. Además, el sistema de ecuaciones normales $A^\top Ax = A^\top b$ puede simplificarse en dos sistemas más sencillos $y = Q^\top b$ y $Rx = y$, ya que $0 = A^\top(b - Ax) = R^\top Q^\top(b - Ax) = R^\top(Q^\top b - Rx)$.

Finalmente, el calculo de las matrices Q y R puede realizarse en $\mathcal{O}(MN^2)$ operaciones por muchos algoritmos entre los que destacaremos: el método de Householder que podemos encontrar en [3, p.248], una modificación del método de ortonormalización de Gram-Schmidt y el método de rotaciones de Givens que veremos en el Capítulo 2.

A continuación introducimos un ejemplo donde se refleja el alto coste computacional que pueden conllevar los métodos directos.

Ejemplo 1.1. El problema de Dirichlet para la ecuación de Poisson con condiciones de frontera homogéneas en el cuadrado unidad,

$$\begin{cases} -\dfrac{\partial^2 u}{\partial x^2} - \dfrac{\partial^2 u}{\partial y^2} = f & \text{en} \quad \Omega := (0,1)^2, \\ u = 0 & \text{en} \quad \Gamma := \partial[0,1]^2, \end{cases} \tag{1.2}$$

donde $f : [0,1]^2 \longrightarrow \mathbb{R}$ es una función continua y la función $u : [0,1]^2 \longrightarrow \mathbb{R}$ debe determinarse. Asumimos que el problema (1.2) tiene una única solución u que es continua y dos veces continuamente diferenciable en Ω.

Consideramos una malla de puntos equidistantes que abarque $[0,1]^2$

$$x_k = kh, \quad y_j = jh, \qquad k, j = 0, 1, ..., M+1, \quad h = \frac{1}{M+1} \tag{1.3}$$

Con respecto a la malla (1.3), la ecuación de Poisson $-\frac{\partial^2 u}{\partial x^2} - \frac{\partial^2 u}{\partial y^2} = f$ se considera solamente en el interior, con puntos (x_k, y_j), $1 \leq k, j \leq M$, mientras que las derivadas parciales se aproximan por diferencias centrales de segundo orden,

$$\begin{cases} -\dfrac{\partial^2 u}{\partial x^2}(x_k, y_j) = \dfrac{-u(x_{k-1}, y_j) + 2u(x_k, y_j) - u(x_{k+1}, y_j)}{h^2} + \mathcal{O}(h^2), \\ -\dfrac{\partial^2 u}{\partial y^2}(x_k, y_j) = \dfrac{-u(x_k, y_{j-1}) + 2u(x_k, y_j) - u(x_k, y_{j+1})}{h^2} + \mathcal{O}(h^2), \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad k, j = 1, ..., M, \end{cases} \tag{1.4}$$

siempre que $u \in C^4([0,1]^2)$. Truncando los términos $\mathcal{O}(h^2)$ en (1.4), se obtiene el sistema lineal de $N = M^2$ ecuaciones lineales

$$-U_{k-1,j} - U_{k,j-1} + 4U_{k,j} - U_{k,j+1} - U_{k+1,j} = h^2 f_{k,j}, \qquad 1 \leq j, k \leq M \tag{1.5}$$

para las aproximaciones, $U_{k,j} \approx u(x_k, y_j)$ con $U_{k,0} = U_{0,j} = 0$ y $f_{k,j} = f(x_k, y_j)$ para $k, j = 1, 2, ..., M$, donde la matriz asociada se puede escribir como

$$A = \begin{bmatrix} A^{(1)} & -I & & \\ -I & A^{(1)} & \ddots & \\ & \ddots & \ddots & -I \\ & & -I & A^{(1)} \end{bmatrix}$$

con $A^{(1)} = Tridiag(-1, 4, -1)$. En la Sección 3.1 se dará una expresión aún más compacta para la matriz A usando el denominado producto de Kronecker de matrices.

En concreto, las ecuaciones (1.5) se escriben en forma matricial de la siguiente manera,

$$\underbrace{\left[\begin{array}{cc|ccc|c} 4 & -1 & -1 & & & \\ -1 & \ddots & \ddots & \ddots & & \\ & \ddots & \ddots & \ddots & & \\ & & -1 & 4 & -1 & \\ \hline -1 & & 4 & -1 & \ddots & \\ & \ddots & -1 & \ddots & \ddots & \\ & & \ddots & \ddots & -1 & \\ & & & -1 & 4 & \\ \hline & & \ddots & \ddots & & -1 \\ & & & \ddots & \ddots & \\ & & & & \ddots & -1 \\ \hline & & & -1 & 4 & -1 \\ & & & & -1 & 4 \end{array}\right]}_{:=A} \left(\begin{array}{c} U_{1,1} \\ \vdots \\ U_{M,1} \\ U_{1,2} \\ \vdots \\ U_{M,2} \\ \vdots \\ U_{1,M} \\ \vdots \\ U_{M,M} \end{array}\right) = h^2 \left(\begin{array}{c} f_{1,1} \\ \vdots \\ f_{M,1} \\ f_{1,2} \\ \vdots \\ f_{M,2} \\ \vdots \\ f_{1,M} \\ \vdots \\ f_{M,M} \end{array}\right)$$

La matriz $A \in \mathbb{R}^{N\times N}$ es una matriz sparse (dispersa) y servirá de referencia para las clases de matrices que introduciremos en el desarrollo de este capítulo.

Observación 1.2. En el lado izquierdo de la ecuación (1.5) aparecen aproximaciones para $u(x, y)$ en el punto (x_k, y_j) y cuatro puntos vecinos $(x_{k\pm1}, y_{j\pm1})$ para todo (i, j). Por ello el esquema es conocido como fórmula de 5 puntos. Por otra parte, el orden dado en la cuadrícula de puntos para obtener la formulación matricial se conoce como orden lexicográfico (por filas).

En las siguientes secciones, introduciremos los métodos iterativos de Jacobi, Gauss-Seidel y sobre relajación sucesiva. En general, estos métodos consideran un vector inicial arbitrario $x^{(0)}$ y calculan sucesivamente vectores $x^{(1)}, x^{(2)}, x^{(3)}, ...$ de acuerdo al método aplicado. Cualquiera de estos métodos involucrarán:

- Una o dos operaciones matriciales, cada una de las cuales requiere N^2 multiplicaciones.
- Varias operaciones pequeñas, por ejemplo, calcular la suma de vectores que requiere un total de $\mathcal{O}(N)$ operaciones aritméticas.

En total, la ejecución de una sola iteración requiere $\mathcal{O}(N^2)$ operaciones aritméticas. Si el rendimiento del método iterativo es suficientemente bueno después de un número considerable de $n \ll N$ iteraciones, entonces la cantidad total será considerablemente menor que $\mathcal{O}(N^3)$ operaciones ariméticas, que es el costo asociado a métodos directos como la eliminación gaussiana, la descomposición de Cholesky o la factorización QR.

En el Capítulo 2 presentaremos los subespacios de Krylov en los que se basarán los métodos iterativos que estudiaremos: método de Gradiente Conjugado para matrices simétricas y definidas positivas y su variante GCNR para el caso general tratado a través del sistema de ecuaciones normales $A^TAx = A^Tb$. Seguidamente desarrollaremos el método del residual

mínimo generalizado para cualquier matriz regular. Por último se presentarán sus variantes precondicionadas para una mayor eficiencia de los métodos.

Finalmente, en la Sección 3 se realizarán diferentes ejemplos en Matlab comparando los tiempos de computación, los errores y el número de iteraciones de los métodos tanto directos como iterativos.

Observación 1.3. Durante las siguientes secciones haremos uso de la siguiente notación. Sea $A = (a_{ij}) \in \mathbb{R}^{N\times N}$ y $v \in \mathbb{R}^N$, definimos las siguientes normas.

La p-norma de la matriz A es

$$\|A\|_p := \max_{\|v\|_p=1} \|Av\|_p, \quad 1 \le p \le \infty \tag{1.6}$$

donde la p-norma del vector v es $\|v\|_p := \left(\sum_{j=1}^{n} |v_j|^p\right)^{1/p}$, $1 \le p < \infty$, y $\|v\|_\infty = \max_{1\le j\le n} |v_j|$.

Recordemos $\|A\|_1 := \max_{1\le j\le N} \sum_{i=1}^{N} |a_{ij}|$ y $\|A\|_\infty := \max_{1\le i\le N} \sum_{j=1}^{N} |a_{ij}|$.

Por otra parte, sean $\lambda_1, ..., \lambda_N$ autovalores de A, denominamos al radio espectral de A como

$$\rho(A) := \max_{i=1,...,N} |\lambda_i|.$$

Con esto, $\|A\|_2 = \sqrt{\rho(A^*A)}$ siendo A^* la matriz traspuesta conjugada de A. Además, si A es simétrica, $\|A\|_2 = \rho(A)$.

Finalmente, definimos el número de condición de A regular con respecto a la norma matricial $\|\cdot\|$ como el número real $cond(A) := \|A\|\|A^{-1}\|$. En particular, denotamos $cond_p(A)$ al número de condición de la matriz A con respecto a la norma (1.6). Además,

$$cond_2(A) = \sqrt{\frac{\lambda_{\text{máx}}(A^*A)}{\lambda_{\text{mín}}(A^*A)}}.$$

Por otra parte, es interesante tener en cuenta el número de condición espectral

$$cond_{sp}(A) := \frac{\text{máx}\{|\lambda_i(A)/i = 1, \ldots, N\}}{\text{mín}\{|\lambda_i(A)/i = 1, \ldots, N\}}$$

Claramente, $cond_2(A) = cond_{sp}(A)$ si A es simétrica.

1.2. Métodos iterativos de punto fijo y algunos tipos especiales de matrices

Una clase de método iterativo que produce una aproximación a la solución x_* de la ecuación (1.1) es la obtenida reformulando $Ax = b$ como una ecuación de punto fijo de la forma:

$$x = \mathcal{H}x + z \tag{1.7}$$

donde $\mathcal{H} \in \mathbb{R}^{N\times N}$ es una matriz de iteración adecuada y $z \in \mathbb{R}^N$ un vector fijo. Luego, lo único que necesitamos exigir es que la solución $x_* \in \mathbb{R}^N$ sea el único punto fijo de (1.7). La iteración del método del punto fijo que corresponde a (1.7) viene dada por

$$x^{(n+1)} = \mathcal{H}x^{(n)} + z, \quad n = 0, 1, ... \tag{1.8}$$

donde $x^{(0)} \in \mathbb{R}^N$ es un vector inicial arbitrario.

Definición 1.4. *El método de punto fijo (1.8) para determinar la solución $x_* \in \mathbb{R}^N$ de (1.1) es convergente si, para cualquier valor inicial $x^{(0)} \in \mathbb{R}^N$, se cumple*

$$\lim_{n\to\infty} \|x^{(n)} - x_*\| = 0$$

donde $\|\cdot\| : \mathbb{R}^N \to \mathbb{R}$ es una norma vectorial cualquiera en $\mathbb{R}^N$. En caso de que el método no converja, entonces se dice divergente.

Observación 1.5. En lo que sigue denotamos por $\sigma[M]$ al conjunto de autovalores, o espectro, de la matriz M y $\rho(M) = \max_{\lambda\in\sigma(M)} |\lambda|$ su radio espectral.

Dada una norma vectorial $\|\cdot\|$ en $\mathbb{R}^N$ consideramos la norma matricial inducida definida como $\|M\| := \sup_{x\neq 0} \frac{\|Mx\|}{\|x\|} = \sup_{\|x\|=1} \|Mx\|$ con $M \in \mathbb{R}^{N\times N}$. Observar que $\|M\| \geq \rho(M)$, pues tomando $\lambda \in \sigma[M]$ con $|\lambda| = \rho(M)$ y $v \neq 0$ autovector de M asociado a λ, entonces $\|M\| \geq \frac{\|Mv\|}{\|v\|} = \frac{|\lambda|\|v\|}{\|v\|} = \rho(M)$.

Para abordar la convergencia de métodos iterativos de punto fijo consideramos el siguiente Teorema (de Von Neumann) cuya prueba puede encontrarse en [5].

Teorema 1.6. *([5, p.256]) Para una matriz $A \in \mathbb{R}^{N\times N}$, los siguientes enunciados son equivalentes:*

1. *$\sigma[A] \subset \{\lambda \in \mathbb{C} : |\lambda| < 1\}$.*
2. *Existe una norma vectorial $\|\cdot\| : \mathbb{R}^N \to \mathbb{R}$ tal que $\|A\| < 1$.*
3. *La serie $\sum_{\nu=0}^{\infty} A^\nu$ es convergente.*
4. *$A^\nu \to 0$ para $\nu \to \infty$.*

Si se cumple al menos una de las anteriones condiciones, entonces $(I - A)^{-1} = \sum_{\nu=0}^{\infty} A^\nu$.

Teorema 1.7. *El método iterativo (1.8) es convergente si y sólo si $\rho(\mathcal{H}) < 1$.*

Demostración. Puesto que $x_* = \mathcal{H}x_* + z$, se tiene que $x^{(n+1)} - x_* = \mathcal{H}(x^{(n)} - x_*)$, $n \geq 0$. Luego,

$$x^{(n)} - x_* = \mathcal{H}^n(x^{(0)} - x_*) \qquad , n \geq 0. \tag{1.9}$$

Por lo tanto, la convergencia es equivalente a que se verifique la condición $\mathcal{H}^n \to 0$ para $n \to \infty$. Esto es a su vez, por el Teorema 1.6, equivalente a $\rho(\mathcal{H}) < 1$. □

A efectos de mejorar la tasa de convergencia del método (1.8) interesa que $\rho(\mathcal{H})$ sea lo menor posible. Esta afirmación está justificada por los dos resultados siguientes.

Lema 1.8. *Para toda matriz $\mathcal{H} \in \mathbb{R}^{N\times N}$ y para cada $\varepsilon > 0$ existe una norma vectorial $\|\cdot\| : \mathbb{R}^N \to \mathbb{R}$ tal que $\|\mathcal{H}\| \leq \rho(\mathcal{H}) + \varepsilon$.*

Demostración. Definiendo $a := \frac{1}{\rho(\mathcal{H})+\varepsilon}$ obtenemos $\rho(a\mathcal{H}) = a\rho(\mathcal{H}) < 1$ y, por el Teorema 1.6, existe una norma vectorial $\|\cdot\| : \mathbb{R}^N \to \mathbb{R}$ tal que $\|a\mathcal{H}\| < 1$. Luego, $\|\mathcal{H}\| \leq \rho(\mathcal{H}) + \varepsilon$. □

Como una consecuencia directa del Lema 1.8 y la Observación 1.5 obtenemos el siguiente

Corolario 1.9. *Para toda matriz $\mathcal{H} \in \mathbb{R}^{N\times N}$,*

$$\rho(\mathcal{H}) = \text{ínf}\{\|\mathcal{H}\| \,/\, \|\cdot\| \textit{norma matricial inducida por una norma vectorial real}\}.$$

□

A continuacion, estudiaremos un teorema que nos proporciona una cota del error.

Teorema 1.10. *Sea $\mathcal{H} \in \mathbb{R}^{N\times N}$ una matriz arbitraria. Para cada $\varepsilon > 0$, existe una norma vectorial $\|\cdot\| : \mathbb{R}^N \to \mathbb{R}$ tal que el método iterativo (1.8) cumple*

$$\|x^{(n)} - x_*\| \leq (\rho(\mathcal{H}) + \varepsilon)^n \|x^{(0)} - x_*\|, \qquad n \geq 0.$$

Demostración. Aplicando (1.9) y el Lema 1.9, tenemos que
$\|x^{(n)} - x_*\| \leq \|\mathcal{H}\|^n \|x^{(0)} - x_*\| \leq (\rho(\mathcal{H}) + \varepsilon)^n \|x^{(0)} - x_*\|, \qquad n = 0, 1, ...$ □

Antes de introducir los métodos iterativos clásicos, consideramos algunas clases de matrices especiales.

Definición 1.11. *Una matriz $B = (b_{kj}) \in \mathbb{R}^{N\times N}$ se denomina reducible si existen conjuntos $\mathcal{K}, \mathcal{J} \subset \{1, ..., N\}$ tales que*

$$\begin{gathered} \mathcal{K} \neq \emptyset, \quad \mathcal{J} \neq \emptyset, \quad \mathcal{K} \cap \mathcal{J} = \emptyset, \quad \mathcal{K} \cup \mathcal{J} = \{1, ..., N\}, \\ b_{kj} = 0, \quad \forall k \in \mathcal{K}, \quad j \in \mathcal{J}. \end{gathered} \tag{1.10}$$

En otro caso, la matriz se dice irreducible.

Ejemplo 1.12. La matriz $\begin{bmatrix} 1 & 2 & 0 \\ -1 & 1 & 0 \\ 3 & 0 & 1 \end{bmatrix}$ es reducible considerando $\mathcal{K} = \{1, 2\}$ y $\mathcal{J} = \{3\}$.

Observación 1.13. Sea $Ax = b$ un sistema de ecuaciones con A regular. Si la matriz $A = (a_{kj}) \in \mathbb{R}^{N\times N}$ es reducible, entonces el sistema lineal se puede descomponer en dos sistemas más pequeños con dimensiones $|\mathcal{K}|$ y $|\mathcal{J}|$. En efecto de la Definición 1.11

(i) Primero, determinamos las incógnitas x_k, para $k \in \mathcal{K}$:

$$\sum_{j\in\mathcal{K}} a_{kj} x_j = b_k, \qquad k \in \mathcal{K}.$$

(ii) Luego, determinamos las incógnitas x_k, para $k \in \mathcal{J}$:

$$\sum_{j\in\mathcal{J}} a_{kj} x_j = b_k - \sum_{j\in\mathcal{K}} a_{kj} x_j, \qquad k \in \mathcal{J}.$$

El siguiente resultado describe cuándo una matriz tridiagonal es irreducible. Observemos que $A = (a_{ij}) \in \mathbb{R}^{N\times N}$ es tridiagonal si $a_{ij} = 0$ para todo $i, j \in \{1, ..., N\}$ con $|i - j| \geq 2$.

Proposición 1.14. *Una matriz tridiagonal es irreducible si y sólo si cada uno de sus elementos no diagonales son distintos de cero.*

Demostración. Sea $B = (b_{kj}) \in \mathbb{R}^{N\times N}$ una matriz tridiagonal.
" $\Longrightarrow$ " Supongamos que B es irreducible y veamos que los elementos no diagonales son distintos de cero.

Para cualquier índice $k_* \in \{1, ..., N-1\}$, definimos los conjuntos $\mathcal{K} = \{1, ..., k_*\}$ y $\mathcal{J} = \{k_*+1, ..., N\}$. Para cualquier $k \in \mathcal{K}$ y $j \in \mathcal{J}$ con $|k-j| \geq 2$ se tiene que $b_{kj} = 0$ y debido a que la matriz B es irreducible, entonces $b_{k_*,k_*+1} \neq 0$. Análogamente, podemos concluir que $b_{k_*+1,k_*} \neq 0$ intercambiando los papeles de $\mathcal{K}$ y $\mathcal{J}$.
" $\Longleftarrow$ " Sean ahora dos conjuntos cualesquiera $\mathcal{K}, \mathcal{J} \subset \{1, ..., N\}$ verificando (1.10). Entonces existen $k \in \mathcal{K}, j \in \mathcal{J}$ adyacentes, es decir, $j = k + 1$ o $j = k - 1$. Por hipótesis, para estos índices $b_{kj} \neq 0$. Por tanto, B es irreducible. □

A continuación veamos algunas propiedades de las matrices irreducibles.

Lema 1.15. *Sea $B = (b_{kj}) \in \mathbb{R}^{N\times N}$ una matriz irreducible.*

1. *Para cada matriz diagonal $D \in \mathbb{R}^{N\times N}$, la matriz $B + D$ es irreducible.*
2. *Si $c_{kj} \in \mathbb{R}$, con $c_{kj} \neq 0$, entonces la matriz $(c_{kj}b_{kj}) \in \mathbb{R}^{N\times N}$ es irreducible.*

Demostración. Si una matriz es irreducible, entonces esta propiedad, por definición, no cambia si se añaden elementos arbitrarios a la diagonal. También se mantiene la propiedad si se multiplican los elementos no diagonales por cantidades no nulas arbitrarias. □

Definición 1.16. *Una matriz $B = (b_{kj}) \in \mathbb{R}^{N\times N}$ se dice irreduciblemente diagonal dominante por filas si B es irreducible y, además, satisface que*

$$\begin{cases} \sum\limits_{\substack{j=1\\ j\neq k}}^{N} |b_{kj}| \leq |b_{kk}|, \quad k = 1, ..., N. \\ \sum\limits_{\substack{j=1\\ j\neq k}}^{N} |b_{kj}| < |b_{kk}|, \quad \textit{para al menos un} \quad k \in \{1, ..., N\}. \end{cases} \tag{1.11}$$

Definición 1.17. *Una matriz $B = (b_{kj}) \in \mathbb{R}^{N\times N}$ se dice irreduciblemente diagonal dominante por columnas si B es irreducible y, además, satisface que*

$$\begin{cases} \sum\limits_{\substack{k=1\\ j\neq k}}^{N} |b_{kj}| \leq |b_{jj}|, \quad j = 1, ..., N. \\ \sum\limits_{\substack{k=1\\ j\neq k}}^{N} |b_{kj}| < |b_{jj}|, \quad \textit{para al menos un} \quad j \in \{1, ..., N\}. \end{cases}$$

Ejemplo 1.18. La matriz correspondiente al Ejemplo 1.1 es una matriz irreduciblemente diagonal dominante por filas y columnas.

Teorema 1.19. *Si una matriz $B = (b_{kj}) \in \mathbb{R}^{N\times N}$ es irreduciblemente diagonal dominante, entonces es regular.*

Demostración. Presentamos la prueba para el caso de dominancia diagonal por filas. Para el caso por columnas es análoga. Por reducción al absurdo, supongamos que B no es una matriz regular. Entonces existe un vector $x \in \mathbb{R}^N$, $x \neq 0$, tal que $Bx = 0$. Definimos los conjuntos de índices $\mathcal{K} := \{k \ / \ |x_k| = \|x\|_\infty\}$, $\mathcal{J} := \{k \ / \ |x_k| < \|x\|_\infty\}$. Es claro que $\mathcal{K} \cup \mathcal{J} = \{1, ..., N\}$, $\mathcal{K} \cap \mathcal{J} = \emptyset$ y $\mathcal{K} \neq \emptyset$. Entonces, también $\mathcal{J} \neq \emptyset$, pues de lo contrario $|x_k| = \|x\|_\infty$ se cumpliría para todos los índices k, y entonces, dado que $Bx = 0$,

$$|b_{kk}||x_k| \leq \sum_{\substack{j=1\\ j\neq k}}^{N} |b_{kj}||x_j|, \qquad k = 1, 2, ..., N$$

produciría una contradicción con la segunda condición de (1.11). Ahora, como B es irreducible, existen $k_* \in \mathcal{K}$, $j_* \in \mathcal{J}$, con $b_{k_*j_*} \neq 0$. Luego,

$$|b_{k_*k_*}| \leq \sum_{\substack{j=1\\ j\neq k_*}}^{N} |b_{k_*j}| \frac{|x_j|}{|x_{k_*}|},$$

siendo $|b_{k_*j_*}| \neq 0$, $\dfrac{|x_{j_*}|}{|x_{k_*}|} < 1$ y $|x_j| \leq |x_{k_*}| = \|x\|_\infty$ para $j \in \{1, ..., N\}$. Luego,

$$\sum_{\substack{j=1\\ j\neq k_*}}^{N} |b_{k_*j}| \frac{|x_j|}{|x_{k_*}|} < \sum_{\substack{j=1\\ j\neq k_*}}^{N} |b_{k_*j}|,$$

lo que contradice la primera condición de (1.11). □

1.3. Método de Jacobi

Dada una matriz $A \in \mathbb{R}^{N\times N}$ regular y un vector $b \in \mathbb{R}^N$, el sistema lineal (1.1) es equivalente a decir que:

$$a_{kk}x_k = b_k - \sum_{\substack{j=1\\ j\neq k}}^{N} a_{kj}x_j, \qquad k = 1, ..., N. \tag{1.12}$$

Podemos escribir (1.12) en forma matricial como $Dx = b-(L+R)x$ usando la descomposición:

$$\underbrace{\begin{bmatrix} a_{11} & \dots & \dots & a_{1N} \\ \vdots & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ a_{N1} & \dots & \dots & a_{NN} \end{bmatrix}}_{:=A} = \underbrace{\begin{bmatrix} a_{11} & & & 0 \\ & \ddots & & \\ & & \ddots & \\ 0 & & & a_{NN} \end{bmatrix}}_{:=D} + \underbrace{\begin{bmatrix} 0 & & & 0 \\ a_{21} & & & \vdots \\ \vdots & \ddots & & \vdots \\ a_{N1} & \dots & a_{NN-1} & 0 \end{bmatrix}}_{:=L} + \underbrace{\begin{bmatrix} 0 & a_{12} & \dots & a_{1N} \\ \vdots & & \ddots & \vdots \\ \vdots & & & a_{N-1N} \\ 0 & & & 0 \end{bmatrix}}_{:=R} \tag{1.13}$$

Definición 1.20. *Sea una matriz $A = (a_{kj}) \in \mathbb{R}^{N\times N}$ regular tal que $a_{kk} \neq 0$ para $k = 1, ..., N$. Dado un vector inicial $x^{(0)} \in \mathbb{R}^N$, el método de Jacobi para el sistema $Ax = b$ es:*

$$x_k^{(n+1)} = \frac{1}{a_{kk}} \left(b_k - \sum_{\substack{j=1\\ j\neq k}}^{N} a_{kj}x_j^{(n)} \right), \qquad k = 1, ..., N, \quad n \geq 0. \tag{1.14}$$

Observación 1.21. En forma matricial, el método de Jacobi queda de la forma:

$$x^{(n+1)} = D^{-1}b - D^{-1}(L + R)x^{(n)} = \mathcal{H}_J x^{(n)} + z$$

donde $z = D^{-1}b$ y la matriz de iteración es

$$\mathcal{H}_J = -D^{-1}(L + R) = -\begin{bmatrix} 0 & \frac{a_{12}}{a_{11}} & \dots & \frac{a_{1N}}{a_{11}} \\ \frac{a_{21}}{a_{22}} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \frac{a_{N-1,N}}{a_{N-1,N-1}} \\ \frac{a_{N1}}{a_{NN}} & \dots & \frac{a_{N,N-1}}{a_{NN}} & 0 \end{bmatrix}. \tag{1.15}$$

El siguiente teorema da una condición suficiente para la convergencia del método de Jacobi.

Teorema 1.22. *El método de Jacobi es convergente si se cumple alguna de las siguientes condiciones:*

1. *A es estrictamente diagonal dominante por filas:* $\sum_{\substack{j=1\\ j\neq k}}^{N} |a_{kj}| < |a_{kk}|$, $k = 1, 2, ..., N$.
2. *A es estrictamente diagonal dominante por columnas:* $\sum_{\substack{k=1\\ k\neq j}}^{N} |a_{kj}| < |a_{jj}|$, $j = 1, 2, ..., N$.

Demostración. Cada una de las condiciones implica directamente que el método de Jacobi es factible, esto es que $a_{kk} \neq 0$ para $k = 1, ..., N$.

i) La primera condición equivale a decir que $\|\mathcal{H}_\mathcal{J}\|_\infty < 1$, lo cual implica $\rho(\mathcal{H}_J) < 1$.

ii) La segunda condición equivale a $\| - (L+R)D^{-1}\|_1 < 1$. Además, las matrices $\mathcal{H}_J$ y $-(L+R)D^{-1}$ son semejantes ya que $D\mathcal{H}_J D^{-1} = -(L+R)D^{-1}$. Luego, $\sigma[\mathcal{H}_J] = \sigma\left[-(L+R)D^{-1}\right]$ y por lo tanto,

$$\rho(\mathcal{H}_J) = \rho\left(-(L+R)D^{-1}\right) \leq \| - (L+R)D^{-1}\|_1 < 1. \qquad \square$$

Muchos sistemas lineales que provienen de la discretización espacial de EDPs involucran matrices que no satisfacen las suposiciones relativamente fuertes del Teorema 1.22. El siguiente teorema también implica la convergencia del método de Jacobi bajo condiciones menos restrictivas.

Teorema 1.23. *Sea $A \in \mathbb{R}^{N\times N}$ una matriz irreduciblemente diagonal dominante. Entonces el método de Jacobi es convergente.*

Demostración. Presentamos la prueba para el caso de dominancia diagonal por filas. Para el caso por columnas la prueba es análoga. Primero veamos que el método está bien definido. Por reducción al absurdo, supongamos que existe algún elemento nulo en la diagonal de la matriz A. Por la Definición 1.16, entonces todas las entradas de la misma fila serían iguales a cero, pero esto contradice que A sea irreducible.

Ahora, para ver la convergencia del método, tenemos que ver que $\rho(\mathcal{H}_J) < 1$, o lo que es lo mismo, que para todo $\lambda \in \mathbb{C}$ con $|\lambda| \geq 1$, la matriz $\mathcal{H}_J - \lambda I$ es regular. Por el Teorema 1.19, es suficiente ver que $\mathcal{H}_J - \lambda I$ es una matriz irreduciblemente diagonal dominante. Pongamos $\mathcal{H}_J - \lambda I := (b_{kj})$ y veamos que cumple las condiciones de la Definición 1.16:

(i) Como A es una matriz irreducible, de (1.13) y el Lema 1.15, concluimos que $A+D = L+R$, $\mathcal{H}_J = -D^{-1}(L+R)$ y $\mathcal{H}_J - \lambda I$ son irreducibles.

(ii) Además, se tiene que:

$$\sum_{\substack{j=1\\ j\neq k}}^{N} |a_{kj}| \leq |a_{kk}| \Rightarrow \sum_{\substack{j=1\\ j\neq k}}^{N} |b_{kj}| = \sum_{\substack{j=1\\ j\neq k}}^{N} \frac{|a_{kj}|}{|a_{kk}|} \leq 1 \leq |\lambda| = |b_{kk}|,\ k = 1, ..., N.$$

Por otra parte, existe $k \in \{1, ..., N\}$ tal que $\sum_{\substack{j=1\\ j\neq k}}^{N} |a_{kj}| < |a_{kk}|$. Luego, para este k

$$\sum_{\substack{j=1\\ j\neq k}}^{N} |b_{kj}| = \sum_{\substack{j=1\\ j\neq k}}^{N} \frac{|a_{kj}|}{|a_{kk}|} < 1 \leq |\lambda| = |b_{kk}|.$$

Por tanto, la matriz $\mathcal{H}_J - \lambda I$ es una matriz irreduciblemente diagonal dominante y, por consiguiente, es regular. $\square$

1.4. Método de Gauss-Seidel

Para el método de Jacobi (1.14) se consideran los valores $x_1^{(n)}, ..., x_{k-1}^{(n)}$ para calcular $x_k^{(n+1)}$. Sin embargo, podemos considerar intuitivamente que las aproximaciones $x_1^{(n+1)}, ..., x_{k-1}^{(n+1)}$ son mejores que $x_1^{(n)}, ..., x_{k-1}^{(n)}$ para calcular $x_k^{(n+1)}$, por lo que se puede definir el siguiente método de iteración de punto fijo en el que se van sustituyendo las antiguas aproximaciones $x_1^{(n)}, ..., x_{k-1}^{(n)}$ por las nuevas $x_1^{(n+1)}, ..., x_{k-1}^{(n+1)}$ conforme se van calculando.

Definición 1.24. *Sea $A = (a_{kj}) \in \mathbb{R}^{N\times N}$ una matriz regular donde $a_{kk} \neq 0$ para $k = 1, ..., N$. Dado un vector inicial $x^{(0)} \in \mathbb{R}^N$, el método de Gauss-Seidel para la solución del sistema lineal de ecuaciones $Ax = b$ viene dado por:*

$$x_k^{(n+1)} = \frac{1}{a_{kk}} \left(b_k - \sum_{j=1}^{k-1} a_{kj} x_j^{(n+1)} - \sum_{j=k+1}^{N} a_{kj} x_j^{(n)} \right), \quad k = 1, ..., N, \ n \geq 0.$$

En forma matricial, el método de Gauss-Seidel es

$$x^{(n+1)} = (D+L)^{-1} b - (D+L)^{-1} R x^{(n)} = \mathcal{H}_{GS} x^{(n)} + z \qquad \text{para } n = 0, 1, ...$$

donde $z = (D+L)^{-1} b$ *y la matriz de iteración es* $\mathcal{H}_{GS} = -(D+L)^{-1} R$.

A efectos de obtener resultados sobre la convergencia del método de Gauss-Seidel introducimos previamente los siguientes dos lemas.

Definición 1.25. *El módulo de una matriz* $A = (a_{kj}) \in \mathbb{R}^{N \times N}$ *es* $|A| := (|a_{kj}|) \in \mathbb{R}^{N \times N}$.

Definición 1.26. *Sean matrices* $A, B \in \mathbb{R}^{N \times M}$. *Escribimos* $A \leq B$ *si* $a_{ij} \leq b_{ij}$ *para todo* i, j. *En particular,* $A \geq 0$ *si* $a_{ij} \geq 0$ *para todo* i, j.

Lema 1.27. *Sean matrices* $A, B \in \mathbb{R}^{N \times N}$. *Entonces:*

(i) $|A| \geq 0$.

(ii) $|A + B| \leq |A| + |B|$.

(iii) $|AB| \leq |A||B|$.

(iv) $\|A\|_\infty = \||A|\|_\infty = \||A|e\|_\infty$ con $e = (1, ..., 1)^{\mathrm{T}} \in \mathbb{R}^{\mathrm{N}}$.

(v) $|A| \leq |B| \Rightarrow \|A\|_\infty \leq \|B\|_\infty$.

Demostración. Es trivial. □

Lema 1.28. *Sean matrices* $S, T \in \mathbb{R}^{N \times N}$ *tales que* $|S| \leq T$ *y* $\rho(T) < 1$. *Entonces*

$$I - S \text{ es regular y } |(I-S)^{-1}| \leq (I-T)^{-1}.$$

Demostración. Por el Lema 1.27, obtenemos que $\left\| \sum_{j=k_0}^{k_1} S^j \right\|_\infty \leq \left\| \sum_{j=k_0}^{k_1} |S|^j \right\|_\infty \leq \left\| \sum_{j=k_0}^{k_1} T^j \right\|_\infty$. Puesto que $\rho(T) < 1$, tenemos que $\left\| \sum_{j=k_0}^{k_1} T^j \right\|_\infty \longrightarrow 0$ para $k_0 \leq k_1$, $k_0, k_1 \to \infty$. Luego, $\sum_{j=0}^{\infty} S^j$ es convergente. Por lo tanto, por el Teorema 1.6, $I - S$ es regular y $(I-S)^{-1} = \sum_{j=0}^{\infty} S^j$. Además, $|(I-S)^{-1}| \leq \sum_{j=0}^{\infty} |S|^j \leq \sum_{j=0}^{\infty} T^j = (I-T)^{-1}$. □

Una condición preliminar que es suficiente para la convergencia del método de Gauss-Seidel viene dada por el siguiente teorema.

Teorema 1.29. *Para toda matriz estrictamente diagonal dominante por filas* $A \in \mathbb{R}^{N \times N}$ *se cumple que* $\|\mathcal{H}_{GS}\|_\infty \leq \|\mathcal{H}_J\|_\infty < 1$.

Demostración. Una matriz $A \in \mathbb{R}^{N \times N}$ es estrictamente diagonal dominante por filas si y sólo si $\|\mathcal{H}_J\|_\infty < 1$. Por lo tanto, nos queda ver que $\|\mathcal{H}_{GS}\|_\infty \leq \|\mathcal{H}_J\|_\infty$. Para ello veremos que

$$|\mathcal{H}_{GS}| e \leq \|\mathcal{H}_J\|_\infty e \quad \text{con } e = (1, ..., 1)^T.$$

Por el Lema 1.27, sabemos que $\|\mathcal{H}_{GS}\|_\infty = \||\mathcal{H}_{GS}|e\|_\infty$. Además, por ser $D^{-1}L$ y $D^{-1}R$ matrices estrictamente triangulares inferior y superior, respectivamente, tenemos que

$$|\mathcal{H}_J| = |D^{-1}L| + |D^{-1}R|. \tag{1.16}$$

Por lo tanto,

$$\begin{aligned}|\mathcal{H}_{GS}| &= |-(D+L)^{-1}R| \\ &= |-(I+D^{-1}L)^{-1}D^{-1}R)| \\ &\leq |(I+D^{-1}L)^{-1}||D^{-1}R)| \\ &\leq (I-|D^{-1}L|)^{-1}|D^{-1}R)|\end{aligned}$$

donde la última desigualdad sigue del Lema 1.28, con $S = -D^{-1}L$ y $T = |-D^{-1}L|$. Ahora de (1.16) obtenemos que:

$$\begin{aligned}|\mathcal{H}_{GS}| &\leq (I-|D^{-1}L|)^{-1}\left[(|\mathcal{H}_J|-I)+\left(I-|D^{-1}L|\right)\right] \\ &= I+(I-|D^{-1}L|)^{-1}(|\mathcal{H}_J|-I).\end{aligned}$$

Luego,

$$|\mathcal{H}_{GS}|\, e \leq e + (I-|D^{-1}L|)^{-1}(|\mathcal{H}_J|e - e).$$

Como $(I-|D^{-1}L|)^{-1} = \sum_{j=0}^{N-1}|D^{-1}L|^j \geq I$ y $|\mathcal{H}_J|e \leq \|\mathcal{H}_J\|_\infty e$, con $\|\mathcal{H}_J\| < 1$, sigue que

$$|\mathcal{H}_{GS}|\, e \leq e + \left[\sum_{j=0}^{N-1}|D^{-1}L|^j\right](\|\mathcal{H}_J\|_\infty - 1)\, e \leq \|\mathcal{H}_J\|_\infty e. \qquad \square$$

De forma análoga, enunciaremos un resultado similar para matrices estrictamente diagonal dominantes por columnas.

Teorema 1.30. *Sea $A \in \mathbb{R}^{N\times N}$ estrictamente diagonal dominante por columnas. Entonces, $\|(D+L)\mathcal{H}_{GS}(D+L)^{-1}\|_1 \leq \|D\mathcal{H}_J D^{-1}\|_1 < 1$. En particular, el método de Gauss-Seidel converge.*

Demostración. Es análoga a la del Teorema 1.29. Primero observamos que A es estrictamente diagonal dominante por columnas si y sólo si $\|(R+L)D^{-1}\|_1 < 1$, o bien, $\|D\mathcal{H}_J D^{-1}\|_1 < 1$. Sean $\widehat{\mathcal{H}}_J := D\mathcal{H}_J D^{-1} = -(LD^{-1}+RD^{-1})$ y $\widehat{\mathcal{H}}_{GS} := (D+L)\mathcal{H}_{GS}(D+L)^{-1} = -R(D+L)^{-1} = -(RD^{-1})(I+LD^{-1})^{-1}$. Luego, por Lema 1.28

$$\begin{aligned}|\widehat{\mathcal{H}}_{GS}| &\leq |RD^{-1}||I+LD^{-1}|^{-1} \leq |RD^{-1}|(I-|LD^{-1}|)^{-1} \\ &= \left[\left(|\widehat{\mathcal{H}}_J|-I\right)+\left(I-|LD^{-1}|\right)\right]\left(I-|LD^{-1}|\right)^{-1} \\ &= I+\left(|\widehat{\mathcal{H}}_J|-I\right)\left(I-|LD^{-1}|\right)^{-1}.\end{aligned}$$

Premultiplicando por $e^\top$:

$$e^\top|\widehat{\mathcal{H}}_{GS}| \leq e^\top + e^\top(|\widehat{\mathcal{H}}_J|-I)\left(I-|LD^{-1}|\right)^{-1}.$$

Ahora $\left(I-|LD^{-1}|\right)^{-1} = \sum_{j=0}^{N-1}|LD^{-1}| \geq I$, mientras que $e^\top(|\widehat{\mathcal{H}}_J|-I) \leq e^\top\|\widehat{\mathcal{H}}_J\|_1 - e^\top$. Luego: $e^\top|\widehat{\mathcal{H}}_{GS}| \leq e^\top + e^\top(\|\widehat{\mathcal{H}}_J\|_1 - 1) = e^\top\|\widehat{\mathcal{H}}_J\|_1$, y entonces, $\|\widehat{\mathcal{H}}_{GS}\|_1 \leq \|\widehat{\mathcal{H}}_J\|_1 < 1$. $\square$

Observación 1.31. De los Teoremas 1.29 o 1.30 no se deduce que $\rho(\mathcal{H}_{GS}) \leq \rho(\mathcal{H}_J)$. De hecho, presentamos el siguiente contraejemplo de matriz estrictamente diagonal dominante, tanto por filas como por columnas, en el que $\rho(\mathcal{H}_J) = 0$ y $0 < \rho(\mathcal{H}_{GS}) < 1$, para todo $a \in \left(\frac{-1}{2}, \frac{1}{2}\right)$, con $a \neq 0$.

Ejemplo 1.32. Sea $A = \begin{bmatrix} 1 & a & -a \\ \frac{a}{2} & 1 & a \\ -\frac{a}{2} & -a & 1 \end{bmatrix}$ con $a \in \left(-\frac{1}{2}, \frac{1}{2}\right)$ de modo que A es estrictamente diagonal dominante por filas y por columnas. Además, si $|a| = \frac{1}{2}$, entonces A es irreduciblemente diagonal dominante. Para esta matriz se tiene que las matrices de iteración del método de Jacobi y de Gauss-Seidel, respectivamente, son

$$\mathcal{H}_J = \begin{bmatrix} 0 & -a & a \\ -\frac{a}{2} & 0 & -a \\ \frac{a}{2} & a & 0 \end{bmatrix}, \qquad \mathcal{H}_{GS} = \begin{bmatrix} 0 & -a & a \\ 0 & \frac{a^2}{2} & -a-\frac{a^2}{2} \\ 0 & -a(\frac{a}{2}-\frac{a^2}{2}) & a^2+(\frac{a}{2}-\frac{a^2}{2}) \end{bmatrix}.$$

Por tanto, calculando sus correspondientes radios espectrales, tenemos que

$$0 = \rho(\mathcal{H}_J) < \rho(\mathcal{H}_{GS}) = \frac{1}{4}\,\text{máx}\left\{0, \left|a(-a^2-\sqrt{a}\sqrt{8+a^3}\right|, \left|a(-a^2+\sqrt{a}\sqrt{8+a^3}\right|\right\} < 1,$$

para todo $a \in \left[-\frac{1}{2}, \frac{1}{2}\right], a \neq 0$.

Otra condición suficiente menos restrictiva para la convergencia del método de Gauss-Seidel se da en el siguiente

Teorema 1.33. *Sea $A \in \mathbb{R}^{N\times N}$ matriz irreduciblemente diagonal dominante por filas o columnas. Entonces el método de Gauss-Seidel es convergente.*

Demostración. Consideramos el caso de dominancia diagonal por filas. El caso por columnas es análogo.

Sea $A = (a_{kj}) \in \mathbb{R}^{N\times N}$ una matriz irreduciblemente diagonal dominante por filas. Se garantiza que el método de Gauss-Seidel está bien definido tal como demostramos en el Teorema 1.23.

Para ver la convergencia del método de Gauss-Seidel necesitamos demostrar que $\rho(\mathcal{H}_{GS}) < 1$, o lo que es lo mismo, que para todo $\lambda \in \mathbb{C}$ con $|\lambda| \geq 1$, la matriz $\mathcal{H}_{GS} - \lambda I$ es regular. Observemos que $\mathcal{H}_{GS} - \lambda I = -(D+L)^{-1}(\lambda D + \lambda L + R)$.

Sea $\lambda \in \mathbb{C}$ con $|\lambda| \geq 1$ y $\lambda \neq 0$, veamos que $M = (\lambda D + \lambda L + R)$ es regular. Por el Teorema 1.19, es suficiente ver que M es irreduciblemente diagonal dominante por filas.

1. Por el Lema 1.15, concluimos que M es irreducible.
2. Ahora, veamos que M satisface las condiciones (1.11):

$$\sum_{j=1}^{k-1} |\lambda a_{kj}| + \sum_{j=k+1}^{N} |a_{kj}| \leq |\lambda| \sum_{\substack{j=1 \\ j\neq k}}^{N} |a_{kj}| \leq |\lambda||a_{kk}|, \quad k = 1, ..., N. \tag{1.17}$$

Además, existe al menos un k para el que se cumple la desigualdad estricta en (1.17). □

Observación 1.34. Sin las condiciones adecuadas para la matriz A, cualquiera de los métodos, Jacobi o Gauss-Seidel, puede converger mientras que el otro no.

Ejemplo 1.35. (a) Sea $A = \begin{bmatrix} 1 & -2 & 2 \\ -1 & 1 & -1 \\ -2 & -2 & 1 \end{bmatrix}$ con la descomposición de (1.13). Calculando las matrices de iteración de los métodos de Jacobi y Gauss-Seidel:

$$\mathcal{H}_J = -D^{-1}(L+R) = \begin{bmatrix} 0 & 2 & -2 \\ 1 & 0 & 1 \\ 2 & 2 & 0 \end{bmatrix} \text{ y } \mathcal{H}_{GS} = -(D+L)^{-1}R = \begin{bmatrix} 0 & 2 & -2 \\ 0 & 2 & -1 \\ 0 & 8 & -6 \end{bmatrix}$$

de modo que $\rho(\mathcal{H}_J) = 0$ y $\rho(\mathcal{H}_{GS}) = 2(1+\sqrt{2})$. Luego, el método de Jacobi es convergente, mientras que el método de Gauss-Seidel no.

(b) A continuación, veamos un ejemplo en el cual el método de Gauss-Seidel converge pero el método de Jacobi no. Dada la siguiente matriz $A = \frac{1}{2}\begin{bmatrix} 2 & 1 & 1 \\ -2 & 2 & -2 \\ -1 & 1 & 2 \end{bmatrix}$ entonces las matrices de iteración son: $\mathcal{H}_J = \frac{1}{2}\begin{bmatrix} 0 & -1 & -1 \\ 2 & 0 & 2 \\ 1 & -1 & 0 \end{bmatrix}$ y $\mathcal{H}_{GS} = \frac{1}{2}\begin{bmatrix} 0 & -1 & -1 \\ 0 & -1 & 1 \\ 0 & 0 & 1 \end{bmatrix}$ de modo que $\rho(\mathcal{H}_J) = \frac{1}{2}\sqrt{5}$ y $\rho(\mathcal{H}_{GS}) = \frac{1}{2}$, respectivamente. Por tanto, el método de Gauss-Seidel converge, mientras que el método de Jacobi no.

1.5. Método de relajación

A continuación, introducimos el método de relajación asociado al método de Gauss-Seidel. Dado $w \in \mathbb{R}$ y $x^{(n)}$ aproximación a x_* solución de $Ax = b$, el método de relajación computa una siguiente aproximación mediante

$$x^{(n+1)} = w\widehat{x}^{(n+1)} + (1-w)x^{(n)},\ n \geq 0 \tag{1.18}$$

siendo $\widehat{x}^{(n+1)}$ la solución de avance de Gauss-Seidel partiendo $x^{(n)}$. Observar que para la solución de avance de Gauss-Seidel se cumple que

$$\widehat{x}^{(n+1)} = (D+L)^{-1}b - (D+L)^{-1}Rx^{(n)} = x^{(n)} + (D+L)^{-1}\left[b - Ax^{(n)}\right].$$

El método de relajación (1.18) coincide entonces con

$$x^{(n+1)} := x^{(n)} + w(D+L)^{-1}\left[b - Ax^{(n)}\right].$$

Estas dos últimas expresiones permiten interpretar ambos métodos en términos del vector residual $Ax^{(n)} - b$.

Observación 1.36. Para el método de Jacobi, se puede definir el método de relajación análogamente como $x^{(n+1)} := (1-w)x^{(n)} + w\widehat{x}^{(n+1)}$ con $\widehat{x}^{(n+1)} = x^{(n)} + D^{-1}(b - Ax^{(n)})$. No obstante, nuestro análisis se centrará en el método de relajación asociado a Gauss-Seidel. Se puede encontrar un análisis de convergencia para el método de relajación de Jacobi en [4, p.360-362].

Definición 1.37. *Sea $A = (a_{kj}) \in \mathbb{R}^{N\times N}$ con $a_{kk} \neq 0$ para $k = 1, ..., N$. El método de relajación (SOR) asociado al método de Gauss-Seidel para la solución de $Ax = b$ es:*

$$a_{kk}x_k^{(n+1)} = w\left(b_k - \sum_{j=1}^{k-1} a_{kj}x_j^{(n+1)} - \sum_{j=k+1}^{N} a_{kj}x_j^{(n)}\right) + (1-w)a_{kk}x_k^{(n)},\ 1 \leq k \leq N,\ n \geq 0.$$

En forma matricial, $x^{(n+1)} = (D+wL)^{-1}\left[wb + ((1-w)D - wR)\,x^{(n)}\right]$, $n \geq 0$, con matriz de iteración:

$$\mathcal{H}(w) = (D+wL)^{-1}\left((1-w)D - wR\right). \tag{1.19}$$

Podemos observar que para $w = 0$, $\mathcal{H}(0) = I$; mientras que para $w = 1$ el método de relajación coincide con el método de Gauss-Seidel, esto es, $\mathcal{H}(1) = \mathcal{H}_{GS}$.

Observación 1.38. Para matrices arbitrarias $A \in \mathbb{R}^{N\times N}$ no es posible, en general, dar una expresión explícita para $\rho(\mathcal{H}(w))$. Por tanto, en general, no es posible determinar exactamente un valor óptimo para el párametro de relajación.

Teorema 1.39 (Kahan). *Sea $A = (a_{kj}) \in \mathbb{R}^{N\times N}$ con $a_{kk} \neq 0$ para $1 \leq k \leq N$, y $\mathcal{H}(w)$ la matriz de iteración (1.19) del método de relajación. Entonces,*

$$\rho(\mathcal{H}(w)) \geq |w-1|, \qquad w \in \mathbb{R}.$$

Demostración. Sean $\lambda_1, ..., \lambda_N \in \mathbb{C}$ los autovalores de $\mathcal{H}(w)$, contando con sus multiplicidades. De (1.19), sigue que

$$\prod_{k=1}^{N} \lambda_k = det(I - wD^{-1}L)^{-1}det((1-w)I - wD^{-1}R) = (1-w)^N,$$

donde la última igualdad es consecuencia de que L y R son estrictamente triangulares. Por lo tanto, existe al menos un λ_k, $1 \leq k \leq N$, tal que $|\lambda_k| \geq |1-w|$. □

Corolario 1.40. *Si el método de relajación es convergente entonces* $0 < w < 2$.

Demostración. Para $w \notin (0,2)$ se cumple, por el Teorema 1.39, que $\rho(\mathcal{H}(w)) \geq 1$. □

Una condición suficiente para la convergencia del método de relajación viene dada por el siguiente teorema.

Teorema 1.41 (Ostrowski, Reich). *Sea* $A \in \mathbb{R}^{N\times N}$ *una matriz simétrica y definida (positiva o negativa). El método de relajación es convergente para cada* $0 < w < 2$, *es decir,* $\rho(\mathcal{H}(w)) < 1$ *para* $0 < w < 2$.

Demostración. Suponemos sin pérdida de generalidad que $A \in \mathbb{R}^{N\times N}$ es una matriz definida positiva. Entonces A es regular, sus autovalores son positivos y $a_{kk} = e_k^T A e_k > 0$, para todo $1 \leq k \leq N$, siendo e_k el k-ésimo vector canónico de $\mathbb{R}^N$. Por lo tanto, el método está bien definido. Para demostrar la convergencia del método en primer lugar observamos que

$$\mathcal{H}(w) = (D + wL)^{-1}\left[(D + wL) - wA\right] = I - w(D + wL)^{-1}A.$$

Como la matriz A es regular, podemos escribir

$$\mathcal{H}(w) = I - 2\left(2A^{-1}\left(\frac{1}{w}D + L\right)\right)^{-1} = I - 2(Q + I)^{-1} = (Q - I)(Q + I)^{-1},$$

de modo que $\mathcal{H}(w)$ es una función racional de $Q := 2A^{-1}\left(\frac{1}{w}D + L\right) - I$. A continuación veremos que

$$\sigma\left[Q\right] \subset \{\lambda \in \mathbb{C} \,/\, Re\lambda > 0\}. \tag{1.20}$$

Habiendo demostrado (1.20), como $\sigma\left[\mathcal{H}(w)\right] = \left\{\dfrac{\lambda - 1}{\lambda + 1} \,/\, \lambda \in \sigma[Q]\right\}$, se tendría que

$$\frac{\lambda - 1}{\lambda + 1} = \frac{|\lambda|^2 - 2Re\lambda + 1}{|\lambda|^2 + 2Re\lambda + 1} < 1, \text{ para } Re\lambda > 0,$$

y $\rho(\mathcal{H}(w)) < 1$, lo cual demostraría la convergencia del método de relajación.

Nos quedaría demostrar (1.20). Sea $\lambda \in \mathbb{C}$ autovalor de Q y $x \in \mathbb{C}^N$, $x \neq 0$, tal que $Qx = \lambda x$. Considerando la definición de Q y premultiplicando por x^*A, siendo x^* el vector traspuesto conjugado de x,

$$\lambda(x^*Ax) = x^*AQx = 2x^*(\frac{1}{w}D + L)x - x^*Ax = \frac{2}{w}(x^*Dx) + 2(x^*Lx) - x^*Ax.$$

Como A y D son matrices definidas positivas, entonces $x^*Ax > 0$ y $x^*Dx > 0$. Luego,

$$Re\lambda(x^*Ax) = \frac{2}{w}(x^*Dx) + 2Re(x^*Lx) - x^*Ax = \frac{2}{w}(x^*Dx) + x^*(L + L^*)x - x^*Ax$$

siendo L^* la matriz traspuesta conjugada de L. Además, A es real y simétrica, por lo que $L^* = L^T = R$. En definitiva,

$$Re\lambda(x^*Ax) = \frac{2}{w}(x^*Dx) + x^*(L + R)x - x^*(D + L + R)x = (\frac{2}{w} - 1)(x^*Dx).$$

Como $0 < w < 2$, se deduce entonces que $Re\lambda > 0$. □

Ahora consideraremos una clase amplia de matrices para la cual el radio espectral de la matriz de iteración $\mathcal{H}(w)$ puede expresarse en función del parámetro w y así poder elegirlo de modo óptimo.

Definición 1.42. *Sea* $A = (a_{kj}) \in \mathbb{R}^{N\times N}$ *una matriz con la descomposición (1.13) y* $a_{kk} \neq 0$ *para todo* k. *Se dice que* A *es una matriz consistentemente ordenada si los autovalores de la matriz*

$$\mathcal{J}(\alpha) := \alpha D^{-1}L + \alpha^{-1}D^{-1}R, \qquad \alpha \in \mathbb{C}, \alpha \neq 0, \tag{1.21}$$

son independientes de α, *es decir, si se cumple que* $\sigma[\mathcal{J}(\alpha)] = \sigma[\mathcal{J}(1)]$ *para todo* $\alpha \in \mathbb{C}, \alpha \neq 0$.

Observación 1.43. Para una interpretación del concepto de matriz consistentemente ordenada en términos de grafos remitimos al lector en [7, Sec. 4.2.5].

Lema 1.44. *Sea $A \in \mathbb{R}^{N\times N}$ consistentemente ordenada y $\mathcal{H}_J$ la matriz de iteración (1.15) del método de Jacobi. Entonces, $\lambda \in \sigma(\mathcal{H}_J)$ si y sólo si $-\lambda \in \sigma(\mathcal{H}_J)$.*

Demostración. Por (1.21), $\mathcal{J}(1) = -\mathcal{H}_J$ y $\mathcal{J}(-1) = \mathcal{H}_J$. La prueba sigue del hecho de que $\mathcal{J}(1)$ y $\mathcal{J}(-1)$ tienen los mismos autovalores. □

El siguiente teorema establece una relación entre los autovalores de $\mathcal{H}_J$ y $\mathcal{H}(w)$.

Teorema 1.45. *Sea $A \in \mathbb{R}^{N\times N}$ matriz consistentemente ordenada y $w \in \mathbb{R}, w \neq 0$.*

$$\lambda \in \sigma\left[\mathcal{H}(w)\right] \Longleftrightarrow \frac{\lambda + w - 1}{w} = \sqrt{\lambda}\mu \qquad \text{con } \mu \in \sigma[\mathcal{H}_J].$$

Demostración. Sea $\lambda \in \mathbb{C}$. Tenemos que

$$\lambda I - \mathcal{H}(w) = w(I + wD^{-1}L)^{-1}\left[\frac{\lambda + w - 1}{w} I + D^{-1}(\lambda L + R)\right].$$

Luego, $\lambda \in \sigma\left[\mathcal{H}(w)\right]$ si y sólo si $\dfrac{\lambda + w - 1}{w} \in \sigma\left[-D^{-1}(\lambda L + R)\right]$.

Observamos que si $\lambda \neq 0$, entonces $D^{-1}(\lambda L + R) = \sqrt{\lambda}\left[D^{-1}(\sqrt{\lambda}L + \dfrac{1}{\sqrt{\lambda}}R)\right]$ y, puesto que A es consistentemente ordenada, sigue que

$$\sigma\left[-D^{-1}(\lambda L + R)\right] = \sqrt{\lambda}\sigma\left[D^{-1}(\sqrt{\lambda}L + \frac{1}{\sqrt{\lambda}}R)\right] = \sqrt{\lambda}\sigma\left[D^{-1}(L + R)\right] = \sqrt{\lambda}\sigma\left[\mathcal{H}_J\right].$$

Por otra parte, si $\lambda = 0$, $\sigma\left[-D^{-1}(\lambda L + R)\right] = \sigma\left[-D^{-1}R\right] = 0 = \sqrt{\lambda}\sigma\left[\mathcal{H}_J\right]$. □

Observación 1.46. En el caso $w = 1$ obtenemos que $\lambda \in \sigma\left[\mathcal{H}_{GS}\right]$ si y sólo si $\lambda \in \sigma\left[\sqrt{\lambda}\mathcal{H}_J\right]$. Además, en tal caso

$$\lambda I - \mathcal{H}_{GS} = (I + D^{-1}L)^{-1}\left[\lambda I + D^{-1}(\lambda I + R)\right].$$

Más aún, sean $p_{GS}(x) = det(xI - \mathcal{H}_{GS})$ y $p_J(x) = det(xI - \mathcal{H}_J)$ los polinomios característicos de las matrices de iteración de Gauss-Seidel y Jacobi, respectivamente. Poniendo $\lambda = \mu^2$,

$$\mu^2 I - \mathcal{H}_{GS} = \mu(I + D^{-1}L)^{-1}\left[\mu I + D^{-1}(\mu L + \mu^{-1}R)\right], \quad \text{para todo } \mu \in \mathbb{R},\ \mu \neq 0.$$

Tomando determinantes y teniendo en cuenta que A es consistentemente ordenada, entonces $p_{GS}(\mu^2) = \mu^N K p_J(\mu)$, siendo K una constante, o bien $p_{GS}(\lambda) = \sqrt[N]{\lambda}K p_J(\sqrt{\lambda})$. Observamos que

(i) si N es par, por el Lema 1.44, $p_J(x)$ sólo contiene potencias pares de x y

$$p_{GS}(\lambda) = K\sqrt[N]{\lambda}p_J(\sqrt{\lambda}) \text{ con } p_J(x) = a_0 + a_2x^2 + ... + a_Nx^N$$

de modo que $\lambda = 0$ es un autovalor de $\mathcal{H}_{GS}$ con multiplicidad al menos $\frac{N}{2}$.

(ii) Si N es impar, $p_J(x)$ sólo contiene potencias impares de x y

$$p_{GS}(\lambda) = K\sqrt[N]{\lambda}p_J(\sqrt{\lambda}) \text{ con } p_J(x) = a_1x + a_3x^3 + ... + a_Nx^N$$

de modo que $\lambda = 0$ es autovalor de $\mathcal{H}_{GS}$ con multiplicidad al menos $\frac{N+1}{2}$.

Corolario 1.47. *Si $A \in \mathbb{R}^{N\times N}$ es una matriz consistentemente ordenada entonces $\rho(\mathcal{H}_{GS}) = \rho(\mathcal{H}_J)^2$.* □

De lo anterior, podemos deducir que para una matriz consistentemente ordenada, tanto el método de Jacobi como el de Gauss-Seidel son, o bien ambos convergentes o bien ambos divergentes. En caso de convergencia, se puede esperar que el método de Gauss-Seidel converja el doble de rápido que el método de Jacobi.

A continuación, damos una forma explícita para $\rho(\mathcal{H}(w)))$ bajo condiciones adecuadas para la matriz A. En este caso se puede determinar explícitamente un valor óptimo para el parámetro de relajación.

Teorema 1.48. *Sea $A \in \mathbb{R}^{N\times N}$ una matriz consistentemente ordenada tal que $\sigma[\mathcal{H}_J] \subset (-1,1)$. Entonces,*

$$\rho(\mathcal{H}(w)) = \begin{cases} \frac{1}{4}\left(w\rho_J + \sqrt{w^2\rho_J^2 - 4(w-1)}\right)^2, & 0 < w \leq w_*, \\ (w-1), & w_* \leq w < 2. \end{cases} \tag{1.22}$$

siendo $w_* := \dfrac{2}{1+\sqrt{1-\rho_J^2}}$ *y* $\rho_J := \rho(\mathcal{H}_J)$.

Observación 1.49. Observar que

(i) dado que $0 \leq \rho_J < 1$, se tiene que $1 \leq w_* < 2$.

(ii) $w^2\rho_J^2 - 4(w-1) \geq 0$, con $w \in (0,2)$, si y sólo si, $w \leq \dfrac{2}{1+\sqrt{1-\rho_J^2}} = w_*$.

(iii) $\rho(\mathcal{H}(w))$, dada por (1.22), es una función continua para $w \in (0,2)$ y derivable en $(0,2)\backslash\{w_*\}$ pues $\frac{\partial}{\partial w}\rho(\mathcal{H}(w))|_{w_*^-} = -\infty$ y $\frac{\partial}{\partial w}\rho(\mathcal{H}(w))|_{w_*^+} = 1$.

La dependencia del radio espectral $\rho(\mathcal{H}(w))$ con respecto al párametro w se ilustra en la Figura 1.1.

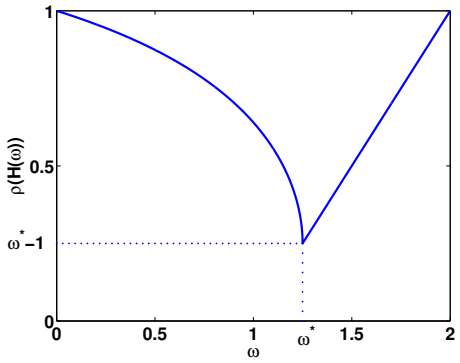


Figura 1.1. Radio espectral de la matriz $\mathcal{H}(\omega)$ para $\rho_J = 0'8$, con valor óptimo del parámetro de relajación $\omega^* = 1'25$.

Demostración (del Teorema 1.48). Sea $\lambda \in \sigma[\mathcal{H}(w)]$. Por el Teorema 1.45, esto equivale a poner $(\lambda + w - 1)^2 = \lambda w^2\mu^2$, para algún $\mu \in \sigma[\mathcal{H}_J] \subset (-1,1)$. O bien,

$$\lambda^2 + \lambda\left[2(w-1) - w^2\mu^2\right] + (w-1)^2 = 0.$$

Entonces, $\lambda = -(w-1) + \dfrac{w^2\mu^2}{2} \pm w|\mu|\sqrt{\dfrac{w^2\mu^2}{4} - (w-1)}$.

Por la Observación 1.49 apartado (ii),

$$\frac{w^2\mu^2}{4} - (w-1) \geq 0 \Longleftrightarrow w \leq \frac{2}{1+\sqrt{1-\mu^2}}.$$

Por tanto, en primer lugar, si $\dfrac{2}{1+\sqrt{1-\mu^2}} < w (< 2)$, entonces $\lambda \in \mathbb{C}\backslash\mathbb{R}$, $w > 1$ y

$$|\lambda|^2 = \left[-(w-1) + \frac{w^2\mu^2}{2}\right]^2 + w^2\mu^2\left(\frac{w^2\mu^2}{4} - (w-1)\right)(-1) = (w-1)^2.$$

Luego, $|\lambda| = w - 1$.

En segundo lugar, si $0 < w \leq \dfrac{2}{1+\sqrt{1-\mu^2}}$, entonces $\lambda \in \mathbb{R}$ y se puede expresar como $\lambda = \frac{1}{4}\left(w|\mu| \pm \sqrt{w^2\mu^2 - 4(w-1)}\right)^2 \geq 0$. Como la determinación positiva provee un autovalor mayor y como la función $\lambda(\mu) = \frac{1}{4}\left(w|\mu| + \sqrt{w^2\mu^2 - 4(w-1)}\right)^2$ es creciente para $\mu \geq 0$ (con $\mu^2 \geq \frac{4(w-1)}{w^2}$) se deduce, tomando $\mu = \rho_J$, donde $\rho_J \in \sigma\left[\mathcal{H}_J\right]$, que

$$\rho(\mathcal{H}(w)) = \begin{cases} \frac{1}{4}(w\rho_J + \sqrt{w^2\rho_J^2 - 4(w-1)}, & 0 < w \leq w^*, \\ (w-1), & w^* \leq w < 2. \end{cases}$$

□

Finalizamos este capítulo con dos ejemplos de matrices consistentemente ordenadas de interés en la discretización espacial de ecuaciones en derivadas parciales.

Ejemplo 1.50. Sea A una matriz tridiagonal por bloques:

$$A = \begin{bmatrix} D_1 & C_1 & & \\ B_1 & \ddots & \ddots & \\ & \ddots & \ddots & C_{M-1} \\ & & B_{M-1} & D_M \end{bmatrix} \in \mathbb{R}^{N\times N}$$

donde $D_j \in \mathbb{R}^{N_j\times N_j}, j = 1, ..., M$, es diagonal y regular, y $\sum_{j=1}^{M} N_j = N$. Además, $B_j \in \mathbb{R}^{N_{j+1}\times N_j}$ y $C_j \in \mathbb{R}^{N_j\times N_{j+1}}$ para $j = 1, ..., M-1$. Veamos que A es consistentemente ordenada. Ponemos A en la forma (1.13),

$$A = D + L + R = \begin{bmatrix} D_1 & & & \\ & D_2 & & \\ & & \ddots & \\ & & & D_M \end{bmatrix} + \begin{bmatrix} 0 & & & \\ B_1 & 0 & & \\ & \ddots & \ddots & \\ & & B_{M-1} & 0 \end{bmatrix} + \begin{bmatrix} 0 & C_1 & & \\ & 0 & \ddots & \\ & & \ddots & C_{M-1} \\ & & & 0 \end{bmatrix}.$$

Tenemos que ver que se cumple que $\sigma\left[\mathcal{J}(1)\right] = \sigma\left[\mathcal{J}(\alpha)\right]$ para todo $\alpha \neq 0$ (ver 1.21). Es fácil ver que

$$\mathcal{J}(\alpha) = \begin{bmatrix} 0 & \frac{1}{\alpha}D_1^{-1}C_1 & & & \\ \alpha D_2^{-1}B_1 & 0 & \frac{1}{\alpha}D_2^{-1}C_2 & & \\ & \alpha D_3^{-1}B_2 & 0 & \ddots & \\ & & \ddots & \ddots & \frac{1}{\alpha}D_{M-1}^{-1}C_{M-1} \\ & & & \alpha D_M^{-1}B_{M-1} & 0 \end{bmatrix}$$

Tomando

$$S_\alpha = \begin{bmatrix} \alpha^0 I & & & \\ & \alpha^1 I & & \\ & & \ddots & \\ & & & \alpha^{M-1} I \end{bmatrix}$$

sigue que

$$S_\alpha^{-1}\mathcal{J}(\alpha)S_\alpha = \begin{bmatrix} 0 & D_1^{-1}C_1 & & & \\ D_2^{-1}B_1 & 0 & D_2^{-1}C_2 & & \\ & D_3^{-1}B_2 & 0 & \ddots & \\ & & \ddots & \ddots & D_{M-1}^{-1}C_{M-1} \\ & & & D_M^{-1}B_{M-1} & 0 \end{bmatrix} = \mathcal{J}(1).$$

Por lo tanto, $\mathcal{J}(1)$ y $\mathcal{J}(\alpha)$ son semejantes para todo $\alpha \neq 0$ y por lo tanto, tienen los mismos autovalores.

Ejemplo 1.51. Sea A la matriz tridiagonal por bloques

$$A = \left[\begin{array}{c|c|c|c} B & b_1 D & & \\ \hline a_1 D & \ddots & \ddots & \\ \hline 0 & \ddots & \ddots & b_{M-1}D \\ \hline & & a_{M-1}D & B \end{array}\right] \in \mathbb{R}^{N\times N}$$

donde $D = diag(b_{11}, ..., b_{KK})$ es una matriz diagonal siendo $b_{kk} \neq 0$ los elementos diagonales de una matriz $B \in \mathbb{R}^{K\times K}$ consistentemente ordenada. Con la descomposición $B = D + L + R$, por (1.21) existe una matriz de paso S_α tal que $\mathcal{J}(\alpha) = S_\alpha \mathcal{J}(1) S_\alpha^{-1}$. Veamos que A también es consistentemente ordenada.

Sean $A = \widehat{D} + \widehat{L} + \widehat{R}$ y $\widehat{\mathcal{J}}(\alpha) = \alpha\widehat{D}^{-1}\widehat{L} + \frac{1}{\alpha}\widehat{D}^{-1}\widehat{R}$, con

$$\widehat{D} = \begin{bmatrix} D & & & \\ & D & & \\ & & \ddots & \\ & & & D \end{bmatrix}, \widehat{L} = \left[\begin{array}{c|c|c|c|c} L & & & & \\ \hline a_1 D & L & & & \\ \hline & a_2 D & \ddots & & \\ \hline & & \ddots & \ddots & \\ \hline & & & a_{M-1}D & L \end{array}\right] \text{ y } \widehat{R} = \left[\begin{array}{c|c|c|c|c} R & b_1 D & & & \\ \hline & R & b_2 D & & \\ \hline & & \ddots & \ddots & \\ \hline & & & \ddots & b_{M-1}D \\ \hline & & & & R \end{array}\right].$$

Tomando $\widehat{S}_\alpha := \begin{bmatrix} \alpha^0 S_\alpha & & & \\ & \alpha^1 S_\alpha & & \\ & & \ddots & \\ & & & \alpha^{M-1} S_\alpha \end{bmatrix}$, sigue usando $S_\alpha \mathcal{J}(1) S_\alpha^{-1} = \mathcal{J}(\alpha)$ que

$$\widehat{S}_\alpha \widehat{\mathcal{J}}(1) S_\alpha^{-1} = \begin{bmatrix} \mathcal{J}(\alpha) & \frac{1}{\alpha}b_1 I & & & \\ \alpha a_1 I & \mathcal{J}(\alpha) & \frac{1}{\alpha}b_2 I & & \\ & \alpha a_2 I & \ddots & \ddots & \\ & & \ddots & \ddots & \frac{1}{\alpha}b_{M-1}I \\ & & & \alpha a_{M-1} I & \mathcal{J}(\alpha) \end{bmatrix} = \widehat{\mathcal{J}}(\alpha).$$

Por lo tanto, $\widehat{\mathcal{J}}(1)$ y $\widehat{\mathcal{J}}(\alpha)$ son semejantes para todo $\alpha \neq 0$ y A es consistentemente ordenada.

Inductivamente este ejercicio nos permite ver que, en particular, la discretización espacial de la ecuacion de Laplace (1.1) en cualquier dimensión mediante diferencias centrales de segundo orden produce sistemas lineales con matrices consistentemente ordenadas.

2

Métodos basados en subespacios de Krylov

2.1. Subespacios de Krylov

Antes de presentar nuevos métodos para resolver el sistema lineal de ecuaciones $Ax = b$, introduciremos los denominados subespacios de Krylov y sus diferentes propiedades ya que posteriormente se utilizarán en los diferentes métodos que estudiaremos en este capítulo.

Definición 2.1. *Sea $A \in \mathbb{R}^{N\times N}$ una matriz y $b \in \mathbb{R}^N$ un vector. Definimos la sucesión de subespacios de Krylov como*

$$\mathcal{K}_n(A,b) = span\left\{b, Ab, \ldots, A^{n-1}b\right\} \subset \mathbb{R}^N, \qquad n = 0, 1, \ldots \tag{2.1}$$

Obviamente se cumple que $\{0\} = \mathcal{K}_0(A,b) \subset \mathcal{K}_1(A,b) \subset \ldots$

En el siguiente lema veremos algunas propiedades esenciales de los subespacios de Krylov.

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

(a) $b, Ab, \ldots, A^n b$ son vectores linealmente dependientes.

(b) Existe un subespacio vectorial $\mathcal{M} \subseteq \mathbb{R}^N$, con $dim\,(\mathcal{M}) \leq n$, tal que $b \in \mathcal{M}$ y $A \cdot \mathcal{M} \subseteq \mathcal{M}$; es decir, $Av \in \mathcal{M}$ para todo $v \in \mathcal{M}$.

(c) $\mathcal{K}_n(A,b) = \mathcal{K}_{n+1}(A,b)$.

(d) $A\mathcal{K}_n(A,b) \subseteq \mathcal{K}_n(A,b)$.

(e) $x_* := A^{-1}b \in \mathcal{K}_n(A,b)$.

Demostración. $(a) \Rightarrow (b)$: Definimos el subespacio $\mathcal{M} := \mathcal{K}_n(A,b) = span\left\{b, Ab, \ldots, A^{n-1}b\right\}$. Es claro que $dim\,(\mathcal{M}) \leq n$ y $b \in \mathcal{M}$. Por hipótesis, existen $\gamma_0, \gamma_1, \ldots, \gamma_{n-1} \in \mathbb{R}$ tales que $A^n b = \sum_{j=0}^{n-1} \gamma_j A^j b$. Por tanto, $A^n b \in \mathcal{K}_n(A,b)$. Luego, si $v \in \mathcal{M}$, $v = \sum_{j=0}^{n-1} \alpha_j A^j b$, y $Av = \sum_{j=0}^{n-1} \alpha_j A^{j+1} b = \sum_{j=1}^{n} \alpha_{j-1} A^j b \in \mathcal{K}_n(A,b) = \mathcal{M}$. Esto implica que $A \cdot \mathcal{M} \subseteq \mathcal{M}$.

$(b) \Rightarrow (c)$: Por la definición de los subespacios de Krylov, $\mathcal{K}_n(A,b) \subseteq \mathcal{K}_{n+1}(A,b)$. Además, por hipótesis, sabemos que $\mathcal{K}_{n+1}(A,b) \subseteq \mathcal{M}$. Luego, $dim\,(\mathcal{K}_{n+1}(A,b)) \leq n$. Por tanto, el vector $A^n b$ es combinación lineal de $\left\{b, Ab, \ldots, A^{n-1}b\right\}$. En definitiva, $\mathcal{K}_{n+1}(A,b) \subseteq \mathcal{K}_n(A,b)$ y $\mathcal{K}_{n+1}(A,b) = \mathcal{K}_n(A,b)$.

$(c) \Rightarrow (d)$: Por definición de subespacios de Krylov y la hipótesis, obtenemos $A\mathcal{K}_n(A,b) \subseteq \mathcal{K}_{n+1}(A,b) = \mathcal{K}_n(A,b)$.

$(d) \Rightarrow (e)$: Sea la aplicación lineal $\mathcal{L} : \mathcal{K}_n(A,b) \longrightarrow \mathcal{K}_n(A,b)$ definida por $\mathcal{L}(v) = Av$ para todo $v \in \mathcal{K}_n(A,b)$. Por hipótesis, $\mathcal{L}$ está bien definida. Además, $\mathcal{L}$ es una aplicación lineal e inyectiva, ya que A es una matriz regular. Por tanto, $dim\,(Im\,(\mathcal{L})) = dim\,(\mathcal{K}_n)$ y $\mathcal{L}$ es una aplicación biyectiva. Luego, existe un vector $v \in \mathcal{K}_n(A,b)$ tal que $Av = b$. Asimismo, como A es una matriz regular, $v = x_*$.

$(e) \Rightarrow (a)$: Sea $x_* \in \mathcal{K}_n(A,b)$ la solución del sistema de ecuaciones $Ax = b$. Entonces $b \in span\,\{Ab, \ldots, A^n b\}$ y $\{b, Ab, \ldots, A^n b\}$ son linealmente dependientes. □

Como consecuencia del Lema 2.2, podemos obtener el comportamiento detallado de la sucesión de subespacios de Krylov y ubicar en qué subespacio se encuentra la solución única del sistema lineal $Ax = b$.

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

$$\{0\} = \mathcal{K}_0(A,b) \subsetneq \mathcal{K}_1(A,b) \subsetneq \ldots \subsetneq \mathcal{K}_{n_*-1}(A,b) \subsetneq \mathcal{K}_{n_*}(A,b) = \mathcal{K}_{n_*+1}(A,b). \tag{2.2}$$

Además, $x_* = A^{-1}b \in \mathcal{K}_{n_*}(A,b) \setminus \mathcal{K}_{n_*-1}(A,b)$.

Demostración. Por definición de los subespacios de Krylov,

$$dim\left(\mathcal{K}_n(A,b)\right) \leq dim\left(\mathcal{K}_{n+1}(A,b)\right) \leq \text{mín}\left\{n+1, N\right\}.$$

Entonces, existe un número natural $n \geq 1$ tal que $\mathcal{K}_n(A,b) = \mathcal{K}_{n+1}(A,b)$. Consideramos $n_* \in \mathbb{N}$ con $n_* \geq 1$, el menor número natural verificando dicha propiedad. Luego, se tiene (2.2). Por el Lema 2.2, apartados (c) y (e), se deduce que $x_* \in \mathcal{K}_{n_*}(A,b)$. Además, como $\mathcal{K}_{n_*-1}(A,b) \neq \mathcal{K}_{n_*}(A,b)$, entonces x_* no pertenece a $\mathcal{K}_{n_*-1}$. □

Observación 2.4. En esta sección resolveremos el sistema lineal de ecuaciones (1.1) mediante los métodos del Gradiente Conjugado (GC), del Gradiente Conjugado para ecuaciones normales (GCNR) y del residual mínimo generalizado (GMRES), que están basados en los subespacios de Krylov definidos en la Definición 2.1. Para aplicar estos métodos sobre el sistema $Ax = b$ consideraremos el valor inical $x_0 = 0$. Para un valor inicial x_0 arbitrario podemos tener en cuenta lo siguiente,

$$Ax = b \Leftrightarrow A\widetilde{x} = \widetilde{b}$$

donde $\widetilde{x} := x - x_0$ y $\widetilde{b} := b - Ax_0$. Además, para $x_n := \widetilde{x}_n + x_0$ se tiene que $Ax_n - b = A\widetilde{x}_n - \widetilde{b}$.

2.2. Método del Gradiente Conjugado

2.2.1. Método del Gradiente Conjugado para matrices simétricas y definidas positivas

De forma genérica, consideramos una sucesión de subespacios vectoriales de la forma:

$$\{0\} \subsetneq \mathcal{D}_1 \subseteq \mathcal{D}_2 \subseteq \ldots \subseteq \mathbb{R}^N. \tag{2.3}$$

Definición 2.5. *Dado los subespacios (2.3), el método del residual ortogonal consiste en hallar*

$$\left.\begin{array}{r} x_n \in \mathcal{D}_n \\ Ax_n - b \in \mathcal{D}_n^{\perp} \end{array}\right\} \qquad n = 1, 2, \ldots \tag{2.4}$$

Nota 2.1 *El vector $Ax - b$ se denomina residual para cada $x \in \mathbb{R}^N$.*

Observación 2.6. Denotaremos el conjunto ortogonal de un conjunto $\mathcal{M} \subset \mathbb{R}^N$ arbitrario como $\mathcal{M}^{\perp} := \left\{y \in \mathbb{R}^N : \langle y, x\rangle_2 = y^{\top}x = 0 \text{ para cada } x \in \mathcal{M}\right\}$ donde $\langle\cdot,\cdot\rangle_2$ denota el producto euclídeo de $\mathbb{R}^N$.

Observación 2.7. En particular, el método del residual ortogonal de la Definición 2.5 se considerará para una matriz $A \in \mathbb{R}^{N\times N}$ simétrica y definida positiva, es decir, $A^{\top} = A$ y $x^{\top}Ax > 0$ para todo $x \in \mathbb{R}^N \backslash \{0\}$. Para estudiar la existencia y unicidad del vector x_n en el método del residual ortogonal, introducimos

$$\langle x, y\rangle_A = x^{\top}Ay \quad \text{y} \quad \|x\|_A = \sqrt{x^{\top}Ax}, \qquad x \in \mathbb{R}^N$$

donde $\langle\cdot,\cdot\rangle_A$ define un producto escalar en $\mathbb{R}^N$ y $\|\cdot\|_A$ es su norma vectorial asociada. A continuación, demostraremos la existencia y unicidad de solución x_n para (2.4) y establecemos una propiedad minimal para dicha solución.

Teorema 2.8. *Sea $A \in \mathbb{R}^{N\times N}$ una matriz simétrica y definida positiva. Para cada $n \geq 1$, el método residual ortogonal (2.4) admite solución única x_n y*

$$\|x_n - x_*\|_A = \min_{x\in\mathcal{D}_n} \|x - x_*\|_A, \qquad n = 1, 2, \ldots, \tag{2.5}$$

siendo $x_* = A^{-1}b$; esto es, x_n es la mejor aproximación de x_* en $\mathcal{D}_n$ respecto de $\langle\cdot,\cdot\rangle_A$.

Demostración. Consideramos una base arbitraria $d_0, d_1, \ldots, d_{m-1}$ de $\mathcal{D}_n$ siendo $dim(\mathcal{D}_n) = m$ y consideremos x_n de la forma $x_n = \sum_{j=0}^{m-1} \alpha_j d_j$, con $\alpha_0, \ldots, \alpha_{m-1} \in \mathbb{R}$. Imponiendo (2.4)

$$Ax_n - b \in \mathcal{D}_n^{\top} \Leftrightarrow \sum_{j=0}^{m-1} \alpha_j \langle Ad_j, d_k\rangle_2 = \langle b, d_k\rangle_2, \qquad k = 0, \ldots, m-1 \tag{2.6}$$

donde $G := (\langle Ad_j, d_k\rangle_2)_{j,k=0}^{m-1}$ es la matriz de Gram de $\{d_0, \ldots, d_{m-1}\}$ respecto de $\langle\cdot,\cdot\rangle_A$. Como G es definida positiva e inversible, entonces existe solución única $\alpha_0, \ldots, \alpha_{m-1} \in \mathbb{R}$ para (2.6). Finalmente, probamos la propiedad minimal (2.5). Para ello, consideramos $x \in \mathcal{D}_n$ vector arbitrario:

$$\begin{aligned}\|x - x_*\|_A^2 &= \|\,(x_n - x_*) + (x - x_n)\,\|_A^2 \\ &= \|x_n - x_*\|_A^2 + 2\langle A(x_n - x_*), x - x_n\rangle_2 + \|x - x_n\|_A^2 \\ &= \|x_n - x_*\|_A^2 + 2\langle (Ax_n - b) - (Ax_* - b), x - x_n\rangle_2 + \|x - x_n\|_A^2.\end{aligned}$$

Por (2.4), sabemos que $Ax_n - b \in \mathcal{D}_n^{\top}$. Como $x - x_n \in \mathcal{D}_n$ y $Ax_* - b = 0$ sigue que
$\langle (Ax_n - b) - (Ax_* - b), x - x_n\rangle_2 = 0$ y $\|x - x_*\|_A^2 \geq \|x_n - x_*\|_A^2$. □

El sistema lineal (2.6) se reduce a un sistema diagonal en el caso de que los vectores $d_0, \ldots, d_{m-1}$ formen una base ortogonal de $\mathcal{D}_n$.

Teorema 2.9. *Sean $A \in \mathbb{R}^{N\times N}$ una matriz simétrica y definida positiva, $d_0, d_1, \ldots, d_{N-1} \in \mathbb{R}^N \setminus \{0\}$ vectores ortogonales respecto de $\langle\cdot,\cdot\rangle_A$ y $\mathcal{D}_n = span\,\{d_0, ..., d_{n-1}\}$, con $1 \leq n \leq N$. Entonces, para cada $1 \leq n \leq N$, la solución única de (2.4) viene dada por*

$$x_n = \sum_{j=0}^{n-1} \alpha_j d_j \ \text{ con } \alpha_j = -\frac{\langle r_j, d_j\rangle_2}{\langle Ad_j, d_j\rangle_2}, \qquad 0 \leq j \leq n-1, \tag{2.7}$$

donde $r_j := Ax_j - b$ para $j \geq 1$ y $r_0 := -b$.

Demostración. Por el Teorema 2.8 con $m = n$, dado que los vectores $d_0, \ldots, d_{n-1}$ son ortogonales respecto de $\langle\cdot,\cdot\rangle_A$, obtenemos

$$x_n = \sum_{j=0}^{n-1} \alpha_j d_j \text{ con } \alpha_j = \frac{\langle b, d_j\rangle_2}{\langle Ad_j, d_j\rangle_2}, \qquad 0 \leq j \leq n-1.$$

Finalmente observamos para $0 \leq j \leq n-1$ que

$$\langle r_j, d_j\rangle_2 = \langle Ax_j - b, d_j\rangle_2 = \sum_{l=0}^{j-1} \alpha_l \langle Ad_j, d_l\rangle_2 - \langle b, d_j\rangle_2 = -\langle b, d_j\rangle_2 .$$ □

Observación 2.10. (i) La representación (2.7) implica que

$$x_{n+1} = x_n + \alpha_n d_n, \quad r_{n+1} = r_n + \alpha_n Ad_n, \text{ para } n = 0, 1, \ldots \tag{2.8}$$

Además, el residual r_{n+1} no requiere calcular un producto adicional Ad_n, ya que dicho producto se requiere para la determinación de α_n.

(ii) Dados x_n y d_n, según (2.5) el coeficiente α_n es óptimo en el siguiente sentido

$$\|x_{n+1} - x_*\|_A = \|x_n + \alpha_n d_n - x_*\|_A = \min_{t\in\mathbb{R}} \|x_n + t d_n - x_*\|_A.$$

El método del residual ortogonal particularizado en los subespacios de Krylov de la Sección 2.1 da lugar al método del Gradiente Conjugado.

Definición 2.11. *Dada una matriz $A \in \mathbb{R}^{N\times N}$ simétrica y definida positiva, el método del gradiente conjugado (GC) es el método (2.4) siendo $\mathcal{D}_n = \mathcal{K}_n(A,b)$ dados por (2.1) para $n = 0, 1, \ldots$*

Teorema 2.12. *Sea $A \in \mathbb{R}^{N\times N}$ una matriz simétrica y definida positiva, $b \in \mathbb{R}^N$ un vector no nulo y $x_n \in \mathbb{R}^N$ el único vector solución de*

$$x_n \in \mathcal{K}_n(A,b), \quad r_n = Ax_n - b \in \mathcal{K}_n(A,b)^\perp, \qquad 0 \le n \le n_*,$$

siendo $n_* \ge 1$ el menor número natural tal que $r_{n_*} = 0$. Sean $d_0 := b$,

$$d_n := -r_n + \beta_{n-1} d_{n-1} \text{ con } \beta_{n-1} := \frac{\langle r_n, d_{n-1}\rangle_A}{\|d_{n-1}\|_A^2} \text{ para } 1 \le n \le n_* - 1. \tag{2.9}$$

Entonces:

(i) $d_0, \ldots, d_{n_*-1}$ son ortogonales respecto de $\langle\cdot,\cdot\rangle_A$.

(ii) $\mathcal{K}_n(A,b) = span\{d_0, \ldots, d_{n-1}\} = span\{r_0, \ldots, r_{n-1}\}$ 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 sigue que

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

En efecto, observar que si $x_* \in \mathcal{K}_n$, como $r_* = Ax_* - b = 0 \in \mathcal{K}_n^\top$, seguiría que $x_* = x_n$ y $r_n = 0$. Luego, si $r_n \neq 0$, entonces x_* no pertenece a $\mathcal{K}_n$.

Ahora, demostraremos los enunciados (i) y (ii) por inducción sobre $n = 1, 2, \ldots, n_*$. Si $n = 1$, dado que $d_0 = b$, $x_0 = 0$ y $r_0 = -b$, es obvio que $\mathcal{K}_1(A,b) = span\{d_0\} = span\{r_0\}$. 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 $\mathcal{K}_n(A,b) = span\{d_0, \ldots, d_{n-1}\} = span\{r_0, \ldots, r_{n-1}\}$. Veamos que se cumplen las propiedades para $n+1 \le n_*$:

Como $r_n \neq 0$ y $r_n \in \mathcal{K}_n(A,b)^\perp$, entonces $d_0, d_1, \ldots, d_{n-1}, -r_n$ son linealmente independientes. Por tanto, el proceso de ortogonalización de Gram-Schmidt respecto de $\langle\cdot,\cdot\rangle_A$ permite afirmar que el vector

$$v_n := -r_n - \sum_{j=0}^{n-1} \frac{\langle -r_n, d_j\rangle_A}{\langle d_j, d_j\rangle_A} d_j \tag{2.10}$$

es ortogonal a $d_0, d_1, \ldots, d_{n-1}$ respecto de $\langle\cdot,\cdot\rangle_A$. Ahora bien, si $0 \le j \le n-2$, entonces $Ad_j \in A\mathcal{K}_{n-1}(A,b) \subset \mathcal{K}_n(A,b)$. Como $r_n \in \mathcal{K}_n(A,b)^\top$, entonces

$$\langle r_n, d_j\rangle_A = \langle r_n, Ad_j\rangle_2 = 0 \text{ para } 0 \le j \le n-2. \tag{2.11}$$

Luego, sustituyendo (2.11) en (2.10), tenemos que

$$v_n = -r_n - \frac{\langle -r_n, d_{n-1}\rangle_A}{\langle d_{n-1}, d_{n-1}\rangle_A} d_{n-1} = -r_n + \beta_{n-1} d_{n-1} = d_n.$$

Por tanto, $d_0, d_1, \ldots, d_n$ son ortogonales respecto de $\langle\cdot,\cdot\rangle_A$. Además, $d_n \neq 0$ ya que si $d_n = 0$, entonces $r_n = \beta_{n-1} d_{n-1} \in \mathcal{K}_n$ y como $r_n \in \mathcal{K}_n^\perp$, entonces $r_n = 0$, lo cual es un absurdo. Por hipótesis de inducción, como $d_n = -r_n + \beta_{n-1} d_{n-1}$, obtenemos

$$span\{d_0,\dots,d_{n-1},d_n\} = span\{r_0,\dots,r_{n-1},d_n\} = span\{r_0,\dots,r_{n-1},r_n\}.$$

Además, como $A\mathcal{K}_n \subset \mathcal{K}_{n+1}$, y $r_0,\dots,r_{n-1} \in \mathcal{K}_n \subset \mathcal{K}_{n+1}$, entonces $r_n = Ax_n - b \in \mathcal{K}_{n+1}$. Por tanto, $span\{r_0,\dots,r_{n-1},r_n\} \subseteq \mathcal{K}_{n+1}$ siendo $dim(\mathcal{K}_{n+1}) = n+1$. Considerando que $n+1 \le n_*$, tenemos que:

$$dim(span\{r_0,\dots,r_{n-1},r_n\}) = dim(span\{d_0,\dots,d_{n-1},d_n\}) = n+1$$

ya que $d_0,\dots,d_{n-1},d_n$ son linealmente independientes. En definitiva,

$$\mathcal{K}_{n+1} = span\{r_0,\dots,r_{n-1},r_n\} = span\{d_0,\dots,d_{n-1},d_n\}. \qquad \square$$

Observación 2.13. Observar que $r_n \in \mathcal{K}_n^{\perp}$ y $\mathcal{K}_n = span\{r_0,\dots,r_{n-1}\}$ implica que

$$\langle r_n, r_j\rangle_2 = 0 \text{ para } 0 \le j \le n-1.$$

Teorema 2.14. *Usando la notación del Teorema 2.12 , se cumple que*

$$\alpha_n = \frac{\|r_n\|_2^2}{\langle Ad_n, d_n\rangle_2}, \quad n = 0,1,\dots,n_*-1 \quad y \quad \beta_{n-1} = \frac{\|r_n\|_2^2}{\|r_{n-1}\|_2^2}, \quad n = 1,\dots,n_*-1.$$

Demostración. La expresión de α_n sigue de (2.7) y (2.9) teniendo en cuenta que

$$-\langle r_n, d_n\rangle_2 = \langle -r_n, -r_n + \beta_{n-1}d_{n-1}\rangle_2 = \|r_n\|_2^2 + \beta_{n-1}\underbrace{\langle -r_n, \overbrace{d_{n-1}}^{\in\mathcal{K}_n(A,b)}\rangle_2}_{=0} = \|r_n\|_2^2.$$

Ahora, demostraremos la representación de β_{n-1}. Por (2.8),

$$\|r_n\|_2^2 = \langle r_n, r_{n-1} + \alpha_{n-1}Ad_{n-1}\rangle_2 = \langle r_n, r_{n-1}\rangle_2 + \alpha_{n-1}\langle r_n, Ad_{n-1}\rangle_2.$$

Como $r_{n-1} \in \mathcal{K}_n(A,b)$, entonces $\langle r_n, r_{n-1}\rangle_2 = 0$ y

$$\|r_n\|_2^2 = \frac{\|r_{n-1}\|_2^2}{\langle Ad_{n-1}, d_{n-1}\rangle_2}\langle r_n, Ad_{n-1}\rangle_2 = \|r_{n-1}\|_2^2\frac{\langle r_n, d_{n-1}\rangle_A}{\|d_{n-1}\|_A^2} = \|r_{n-1}\|_2^2\beta_{n-1}. \qquad \square$$

Compaginando los resultados de los teorema previos, obtenemos el siguiente algoritmo para el método del Gradiente Conjugado:

Algoritmo 1 Método del Gradiente Conjugado (GC)

Considerar el sistema lineal $Ax = b$ con $A \in \mathbb{R}^{N\times N}$ matriz simétrica y definida positiva y $b \in \mathbb{R}^N$ un vector no nulo. Sea $r_j = Ax_j - b$ el residual de la iteración j.

1: Definimos $x_0 := 0$, $r_0 := -b$ y $d_0 := -r_0 = b$.
2: Dados x_n, r_n y d_n, $n \ge 0$, mientras $r_n \neq 0$ calcular

$x_{n+1} = x_n + \alpha_n d_n$, con $\alpha_n = \frac{\|r_n\|_2^2}{\langle Ad_n, d_n\rangle_2}$

$r_{n+1} = r_n + \alpha_n Ad_n$

$d_{n+1} = -r_{n+1} + \beta_n d_n$, con $\beta_n = \frac{\|r_{n+1}\|_2^2}{\|r_n\|_2^2}$.

Observación 2.15. Observamos que el Algoritmo 1, requiere computar el producto matriz-vector Ad_n en cada iteración.

Observación 2.16. Otra forma de interpretar el método del Gradiente Conjugado es mediante las ecuaciones normales del problema de mínimos cuadrados (2.5). Sea $A \in \mathbb{R}^{N\times N}$ simétrica y definida positiva y $b \in \mathbb{R}^N$. Veamos que:

(i) Para $n = 1, \ldots, n_*$ se cumple que $x_n = q_n(A)b$ y $r_n = -p_n(A)b$ para cierto $q_n \in \prod_{n-1}$ y $p_n(t) = 1 - tq_n(t)$.
En efecto, sea $x_n \in \mathcal{K}_n(A,b) = span\left\{b, Ab, \ldots, A^{n-1}b\right\}$ con $Ax_n - b = r_n \in \mathcal{K}_n(A,b)^{\perp}$. Entonces $x_n = \sum_{j=0}^{n-1} c_j A^j b = q_n(A)b$ con $q_n(t) = \sum_{j=0}^{n-1} c_j t^j \in \prod_{n-1}$. Además, $r_n = Ax_n - b = -\left[I - A\sum_{j=0}^{n-1} c_j A^j\right] b = -p_n(A)b$ con $p_n(t) = 1 - tq_n(t)$.

(ii) Se cumple que $q_n(t) = \sum_{j=0}^{n-1} c_j t^j$ donde el vector $(c_0, \ldots, c_{n-1})^{\top} \in \mathbb{R}^n$ es la solución del sistema lineal:

$$\begin{bmatrix} b^{\top}Ab & b^{\top}A^2b & \ldots & b^{\top}A^nb \\ b^{\top}A^2b & b^{\top}A^3b & \ldots & b^{\top}A^{n+1}b \\ \vdots & \vdots & \ddots & \vdots \\ b^{\top}A^nb & b^{\top}A^{n+1}b & \ldots & b^{\top}A^{2n-1}b \end{bmatrix} \begin{bmatrix} c_0 \\ c_1 \\ \vdots \\ c_{n-1} \end{bmatrix} = \begin{bmatrix} b^{\top}b \\ b^{\top}Ab \\ \vdots \\ b^{\top}A^{n-1}b \end{bmatrix}. \tag{2.12}$$

En efecto, sabemos que x_n resuelve el problema de mínimos cuadrados $\|x_n - x_*\|_A = \min_{x \in \mathcal{K}_n(A,b)} \|x - x_*\|_A$ donde $\mathcal{K}_n(A,b)$ es un subespacio vectorial de dimensión finita de $\left(\mathbb{R}^N, \langle\cdot,\cdot\rangle_A\right)$. Luego, por el estudio de mejor aproximación en espacios pre-Hilbert, existe un único $x_n \in \mathcal{K}_n(A,b)$ tal que $\|x_n - x_*\|_A = \min_{x \in \mathcal{K}_n(A,b)} \|x - x_*\|_A$.

Para $n \leq n_*$, $\left\{b, Ab, \ldots, A^{n-1}b\right\}$ es un sistema linealmente independiente y es una base de $\mathcal{K}_n(A,b)$. Luego, los coeficientes $c_0, \ldots, c_{n-1} \in \mathbb{R}^N$ de $x_n = \sum_{j=0}^{n-1} c_j A^j b$ se obtienen resolviendo las ecuaciones normales del problema de mínimos cuadrados

$$Gc = F$$

donde $c = (c_0, \ldots, c_{n-1})^{\top}$, $F = \left(\langle x_*, A^{j-1}b\rangle_A\right)_{j=1}^n = \left(b^{\top}A^{j-1}b\right)_{j=1}^n$ y con matriz de Gram $G = \left(\langle A^{i-1}b, A^{j-1}b\rangle_A\right)_{i,j=1}^n = \left(b^{\top}A^{i+j-1}b\right)_{i,j=1}^n$. Este sistema lineal coincide con (2.12).

Observación 2.17. Sea $A \in \mathbb{R}^{N\times N}$ una matriz simétrica, definida positiva y $b \in \mathbb{R}^N$ un vector no nulo. Observamos que para todo n número natural, $r_n = Ax_n - b$ coincide con el gradiente del funcional energía $\mathcal{J}(x) := \frac{1}{2}\langle Ax, x\rangle_2 - \langle x, b\rangle_2$ evaluado en x_n. Esto es, $\nabla\mathcal{J}(x_n) = r_n$.

Observación 2.18. Sea $v \in \mathbb{R}^N$ un autovector no nulo de la matriz $A \in \mathbb{R}^{N\times N}$ y $b = kv$ donde $k \in \mathbb{R}\setminus\{0\}$. Entonces, el método del Gradiente Conjugado para $Ax = b$ converge en una iteración, es decir, $\mathcal{K}_1(A,b) = \mathcal{K}_2(A,b)$. En efecto, $\mathcal{K}_1(A,b) = span\{b\} = span\{b, Ab\} = \mathcal{K}_2(A,b)$. Esta propiedad se puede generalizar para dar una interpretación algebraica del número de iteraciones $n_* \leq N$ necesarias para la convergencia del método del Gradiente Conjugado.

Definición 2.19. *Sea $A \in \mathbb{R}^{N\times N}$ una matriz regular y $b \in \mathbb{R}^N$ un vector. Se llama polinomio mínimo de b respecto de A al polinomio mónico $p \in \prod_\nu$ de menor grado $\nu \geq 0$ tal que*

$$p(A)b = 0. \tag{2.13}$$

ν se denomina grado de b respecto de A.

Observación 2.20. Por el Teorema de Cayley-Hamilton, es claro que $\nu \leq N$.

Observación 2.21. Observar que el polinomio (2.13) es único pues si existe otro $q \in \prod_\nu$, $q \neq p$, polinomio mónico tal que $q(A)b = 0$, entonces $p - q \in \prod_{\nu-1}$ y $(p-q)(A)b = 0$. Luego, si α es el coeficiente director de $p - q$, el polinomio $r = \frac{p-q}{\alpha}$ es mónico, verifica que $r(A)b = 0$ y su grado es menor que ν, lo cual es un absurdo.

Teorema 2.22. *Sea $A \in \mathbb{R}^{N\times N}$ una matriz regular y $b \in \mathbb{R}^N$ un vector no nulo. Sea $\nu \geq 1$ el grado de b respecto de A. Entonces, $\mathcal{K}_0(A,b) \subsetneq \mathcal{K}_1(A,b) \subsetneq \ldots \subsetneq \mathcal{K}_\nu(A,b) = \mathcal{K}_{\nu+1}(A,b)$. En particular, el método del Gradiente Conjugado converge exactamente en ν iteraciones, esto es, $x_* = A^{-1}b \in \mathcal{K}_\nu(A,b)\backslash\mathcal{K}_{\nu-1}(A,b)$. En otras palabras, $\nu = n_*$ en (2.2).*

Demostración. Consideramos $\nu \geq 1$ es el menor número natural tal que $A^\nu b$ es combinación lineal de $b, \ldots, A^{\nu-1}b$. Equivalentemente, por el Lema 2.2, ν es el menor número natural tal que $\mathcal{K}_\nu(A,b) = \mathcal{K}_{\nu+1}(A,b)$, esto es, $\mathcal{K}_0(A,b) \subsetneq \mathcal{K}_1(A,b) \subsetneq \ldots \subsetneq \mathcal{K}_\nu(A,b) = \mathcal{K}_{\nu+1}(A,b) = \ldots$ El mismo lema permite asegurar que $x_* \in \mathcal{K}_\nu(A,b)\backslash\mathcal{K}_{\nu-1}(A,b)$. Como $Ax_* - b = 0 \in \mathcal{K}_\nu(A,b)^\perp$, por el Teorema 2.8 sigue que $x_* = x_\nu$. Además, como x_* no pertenece a $\mathcal{K}_{\nu-1}(A,b)$, se tiene que $x_* \neq x_n$ para $0 \leq n \leq \nu - 1$. En definitiva, el método del Gradiente Conjugado converge exactamente en ν iteraciones. □

Según el Teorema 2.22, el método del Gradiente Conjugado puede interpretarse 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, usaremos los siguientes resultados para calcular estimaciones para el error del método del Gradiente Conjugado basándonos en la propiedad de optimalidad dada en (2.5). Previamente enunciamos el célebre teorema espectral para matrices simétricas.

Teorema 2.23. *([1, p.151]) 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 todo $\lambda \in \sigma[A]$, $\lambda \in \mathbb{R}$ 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 $u_1^\top u_2 = 0$.

(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, $U^\top U = UU^\top = I$, tal que $A = UDU^\top$. Los elementos diagonales de D son los autovalores de A y las columnas de U son sus autovectores correspondientes.

Lema 2.24. *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 $x = \sum_{j=1}^N c_j v_j \in \mathbb{R}^N$, entonces para todo polinomio p:*

$$\|p(A)x\|_2 = \left(\sum_{j=1}^N c_j^2 p(\lambda_j)^2\right)^{1/2}, \qquad \|p(A)x\|_A = \left(\sum_{j=1}^N c_j^2 \lambda_j p(\lambda_j)^2\right)^{1/2}. \tag{2.14}$$

En particular, se cumple que

$$\sqrt{m}\|x\|_2 \leq \|x\|_A \leq \sqrt{M}\|x\|_2, \qquad x \in \mathbb{R}^N \tag{2.15}$$

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

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

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

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 (2.14) 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 (2.15) dado que $\min_{j=i,\ldots,N} \lambda_j \leq \lambda_j \leq \max_{j=i,\ldots,N} \lambda_j$ para $j = 1, \ldots, N$. □

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

Teorema 2.25. *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*

$$\|x_n - x_*\|_A \leq \left(\inf_{\substack{p\in\prod_n \\ p(0)=1}} \sup_{\lambda\in\sigma[A]} |p(\lambda)| \right) \|x_* - x_0\|_A, \qquad n = 0, 1, \ldots, n_*.$$

Demostración. Para cada polinomio $p \in \prod_n$ con $p(0) = 1$ definimos el polinomio $q(t) := (1 - p(t))/t \in \prod_{n-1}$ y el vector $x := q(A)b \in \mathcal{K}_n(A, b)$. Entonces,

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

donde en la última igualdad se ha usado que $A^{-1}p(A) = p(A)A^{-1}$. Por (2.5) y el Lema 2.24 con $x_* = \sum_{j=1}^{N} c_j v_j \in \mathbb{R}^N$, obtenemos

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

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

El Teorema 2.25 y el siguiente lema permitirán dar una cota de error más práctica.

Lema 2.26. *Sean $T_n(t) = \cos(n \arccos t)$, para $t \in [-1, 1]$ y $n \geq 0$, los polinomios de Chebyshev de primera especie. Entonces,*

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

$$T_n\left(\frac{k+1}{k-1} \right) \geq \frac{1}{2} \left(\frac{\sqrt{k}+1}{\sqrt{k}-1} \right)^n, \qquad \text{para } k \in \mathbb{R}, k \geq 1. \tag{2.16}$$

Demostración. Es bien conocido que 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 $t \in [-1, 1]$

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

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

$$\begin{aligned}\tfrac{1}{2}\left[\left(t + i\sqrt{1-t^2}\right)^n + \left(t - i\sqrt{1-t^2}\right)^n\right] &= \tfrac{1}{2}\left[\left(t + \sqrt{t^2-1}\right)^n + \left(t - \sqrt{t^2-1}\right)^n\right] \\ &= \sum_{\substack{k=0 \\ k \text{ par}}}^{n} \binom{n}{k} t^{n-k}\left(\sqrt{t^2-1}\right)^k\end{aligned}$$

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

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

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

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

Teorema 2.27. *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 $x_0 = 0$, verifican*

$$\|x_n - x_*\|_A \leq 2\gamma^n \|x_0 - x_*\|_A, \quad \|x_n - x_*\|_2 \leq 2\sqrt{k_A}\gamma^n \|x_0 - x_*\|_2, \qquad n \geq 1,$$

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

Observación 2.28. Observar que si A es una matriz simétrica y definida positiva, entonces $cond_2(A) := \|A\|_2\|A^{-1}\|_2 = \rho(A)\rho(A^{-1}) = \frac{\lambda_{\max}(A)}{\lambda_{\min}(A)} \geq 1$, siendo $\lambda_{\max}(A)$ y $\lambda_{\min}(A)$ el mayor y menor autovalor de A, respectivamente.

Demostración (Teorema 2.27). Por la observación anterior sabemos que $k_A \geq 1$.

1. En primer lugar, si $k_A = 1$ entonces $\lambda = \lambda_{\max}(A) = \lambda_{\min}(A)$ para todo $\lambda \in \sigma[A]$. Luego, como A es simétrica y definida positiva, por el Teorema 2.23 tenemos que $A = \lambda I$ con $\lambda > 0$. Por tanto, $\mathcal{K}_2(A,b) = \mathcal{K}_1(A,b)$ y $x_1 = \ldots = x_n = x_* = \lambda^{-1}b$ para todo $n \geq 1$. Por consiguiente, el enunciado es evidente.
2. Ahora, supongamos que $k_A > 1$ de modo que $0 < \gamma < 1$. Denotamos $M := \lambda_{\max}(A)$ y $m := \lambda_{\min}(A)$ con lo cual $\sigma[A] \subset [m, M]$. Consideramos el siguiente polinomio

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

Es claro que $p \in \prod_n$ con $p(0) = 1$ y si $t \in [m, M]$, entonces $-1 \leq \frac{M+m-2t}{M-m} \leq 1$. Por tanto, puesto que $k_A = \frac{M}{m}$, usando (2.16)

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

Aplicando el Teorema 2.25 sigue que $\|x_n - x_*\|_A \leq 2\gamma^n\|x_0 - x_*\|_A$. Finalmente, de (2.15) obtenemos que

$$\|x_n - x_*\|_2 \leq \frac{1}{\sqrt{m}}\|x_n - x_*\|_A \leq \frac{2\gamma^n}{\sqrt{m}}\|x_0 - x_*\|_A \leq 2\sqrt{k_A}\gamma^n\|x_0 - x_*\|_2. \qquad \square$$

Observación 2.29. Este teorema refleja que la convergencia del método GC puede verse ralentizada si $k_A \gg 1$.

2.2.2. 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. Observamos que $A^\top A$ es una matriz regular, $det(A^\top A) = det(A)^2 > 0$, es simétrica y también definida positiva, puesto que $x^\top A^\top Ax = (Ax)^\top (Ax) = \|Ax\|_2^2 > 0$ para todo $x \in \mathbb{R}\backslash\{0\}$.

El método del Gradiente Conjugado para las ecuaciones normales (GCNR) consiste en aplicar el método GC al sistema

$$A^\top Ax = A^\top b$$

para obtener $x_* = A^{-1}b$. Así pues, el método GCNR considera los subespacios de Krylov $\mathcal{K}_n(A^\top A, A^\top b) = span\left\{A^\top b, (A^\top A)(A^\top b), \ldots, (A^\top A)^{n-1}(A^\top b)\right\}$.

Antes de estudiar la estimación del error para el método GCNR presentamos su correspondiente algoritmo:

Algoritmo 2 Método GC para Ecuaciones Normales (GCNR)

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. Denotamos $r_j := Ax_j - b$ y $\widehat{r}_j = A^\top r_j$, $j \geq 0$.

1: Definimos $x_0 := 0$, $r_0 := -b$, $\widehat{r}_0 := A^\top r_0$ y $\widehat{d}_0 := A^\top b$.
2: Dados x_n, r_n, $\widehat{r}_n$ y $\widehat{d}_n$ con $n \geq 0$, mientras $r_n \neq 0$ calcular
$x_{n+1} = x_n + \alpha_n \widehat{d}_n$, con $\alpha_n = \frac{\|\widehat{r}_n\|_2^2}{\langle A^\top A\widehat{d}_n, \widehat{d}_n\rangle_2} = \frac{\|\widehat{r}_n\|_2^2}{\|A\widehat{d}_n\|_2^2}$
$r_{n+1} = r_n + \alpha_n A\widehat{d}_n$
$\widehat{r}_{n+1} = \widehat{r}_n + \alpha_n A^\top A\widehat{d}_n = A^\top r_{n+1}$
$\widehat{d}_{n+1} = -\widehat{r}_{n+1} + \beta_n \widehat{d}_n$, con $\beta_n = \frac{\|\widehat{r}_{n+1}\|_2^2}{\|\widehat{r}_n\|_2^2}$.

Observación 2.30. A diferencia del método GC para el caso simétrico definido positivo, el método GCNR requiere otro producto matriz-vector $A^\top r_j$ para cada $j \geq 0$. Sin embargo, podemos evitar calcular la operación $A^\top A$ que requiere un gran costo computacional.

Observación 2.31. (i) Como consecuencia directa del Teorema 2.8, se da la siguiente propiedad minimal para las iteraciones del método GCNR:

$$\|Ax_n - b\|_2 = \min_{x\in\mathcal{K}_n(A^\top A, A^\top b)} \|Ax - b\|_2.$$

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

(ii) Por el Teorema 2.27, obtenemos la siguiente estimación del error para el método GCNR:

$$\|Ax_n - b\|_2 \leq 2\gamma^n \|Ax_0 - b\|_2, \quad \|x_n - x_*\|_2 \leq 2k_A\gamma^n \|x_n - x_*\|_2, \qquad n \geq 1,$$

siendo $\gamma = \frac{k_A - 1}{k_A + 1}$. Observar que $k_{A^\top A} = cond_2(A^\top A) = cond_2(A)^2 = k_A^2$.

2.3. Método del residual mínimo generalizado

Presentamos otra posibilidad para resolver un sistema lineal regular $Ax = b$, donde $A \in \mathbb{R}^{N\times N}$ es una matriz regular arbitraria, no necesariamente simétrica o definida positiva.

Definición 2.32. *Dada una sucesión de subespacios (2.3), el método del residual mínimo consiste en hallar*

$$\left.\begin{array}{r} x_n \in \mathcal{D}_n \\ \|Ax_n - b\|_2 = \min\limits_{x\in\mathcal{D}_n} \|Ax - b\|_2 \end{array}\right\} \qquad n = 1, 2, \ldots \tag{2.17}$$

Observación 2.33 (Existencia y unicidad de aproximación con residual minimal). Para ver que (2.17) tiene solución única, definimos $\mathcal{S}_n := A\mathcal{D}_n = \{Ad \,/\, d \in \mathcal{D}_n\}$. Claramente, $\mathcal{S}_n$ es un subespacio de $\mathbb{R}^N$. Por la teoría de existencia y unicidad de mejor aproximación en espacios pre-Hilbert aplicada a $\left(\mathbb{R}^N, \langle\cdot,\cdot\rangle_2\right)$ con subespacio finitodimensional $\mathcal{S}_n$, tenemos que existe un único vector $z_n \in \mathcal{S}_n$ tal que $\|z_n - b\|_2 = \min\limits_{z\in S_n} \|z - b\|_2$. Como A es una matriz regular, esto equivale a decir que existe un único $x_n \in \mathcal{D}_n$ tal que $\|Ax_n - b\|_2 = \min\limits_{x\in\mathcal{D}_n} \|Ax - b\|_2$.

Definición 2.34. *Dada una matriz $A \in \mathbb{R}^{N\times N}$ regular y un vector $b \in \mathbb{R}^N$, el método del residual mínimo generalizado (GMRES) es el método (2.17) siendo $\mathcal{D}_n = \mathcal{K}_n(A, b)$ dados por (2.1) para $n = 0, 1, \ldots$.*

Observación 2.35. Observar que si A es simétrica y definida positiva, entonces el método GMRES implica que $\|x_n - x_*\|_{A^2} = \min\limits_{x\in\mathcal{K}_n(A,b)} \|x_n - x_*\|_{A^2}$ por lo que se espera, en este caso, un comportamiento similar al del método GC respecto al error con relación número de iteraciones.

El procedimiento básico que seguiremos para implementar el método GMRES es el siguiente:

(i) Usando el proceso de Arnoldi (ver Algoritmo 3) se genera una base ortogonal de $\mathcal{K}_n(A, b)$ con respecto al producto escalar euclídeo.

(ii) Usando dicha base ortogonal, el problema (2.17) se reduce a un problema de mínimos cuadrados.

Sea $v_1 \in \mathbb{R}^N$ un vector tal que $\|v_1\| = 1$. El proceso de Arnoldi consiste en generar una sucesión de vectores $v_1, v_2, \ldots$ tal que $\|v_i\| = 1, i = 1, 2, ...,$ ortogonales con respecto al producto escalar $\langle\cdot,\cdot\rangle_2$ mediante el proceso de Gram-Schmidt aplicado a los vectores $v_1, Av_1, Av_2, \ldots$ de la siguiente manera:

Definimos $v_1 = \frac{b}{\|b\|_2}$ y consideramos el sistema $\{v_1, Av_1\}$. Aplicando el proceso de Gram-Schmidt, se obtiene

$$\widehat{v}_2 = Av_1 - \langle Av_1, v_1\rangle_2 \, v_1.$$

Si $\widehat{v}_2 \neq 0$, entonces definimos $v_2 := \frac{\widehat{v}_2}{\|\widehat{v}_2\|_2}$ y obtenemos el sistema ortonormal $\{v_1, v_2\}$. A continuación, aplicamos el proceso de Gram-Schmidt a $\{v_1, v_2, Av_2\}$, para obtener

$$\widehat{v}_3 = Av_2 - \langle Av_2, v_1\rangle_2 \, v_1 - \langle Av_2, v_2\rangle_2 \, v_2$$

Luego, si $\widehat{v}_3 \neq 0$, entonces definimos $v_3 := \frac{\widehat{v}_3}{\|\widehat{v}_3\|_2}$, resultando un nuevo sistema ortornormal $\{v_1, v_2, v_3\}$. En general, el proceso de Arnoldi se define en el siguiente algoritmo:

Algoritmo 3 Proceso de Arnoldi.

1: Dado $b \in \mathbb{R}^N$, $b \neq 0$, definimos $\widehat{v}_1 := b$. Mientras $\widehat{v}_n \neq 0$ calcular
2: $v_n = \frac{\widehat{v}_n}{\|\widehat{v}_n\|_2}$.
3: $\widehat{v}_{n+1} = Av_n - \sum\limits_{j=1}^{n} \langle Av_n, v_j\rangle_2 v_j$.

Observación 2.36. El proceso de Arnoldi parará en la primera iteración tal que $Av_n \in span\{v_1, \ldots, v_n\}$.

Observación 2.37. 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 $j = 0, \ldots, n-1$ y Av_k es combinación lineal de $\{v_1, \ldots, v_k, v_{k+1}\}$. Luego, el proceso de Arnoldi se convierte en una recursión a tres términos:

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

Este caso especial del proceso de Arnoldi aplicado a matrices simétricas se llama proceso de Lanczos.

El siguiente lema permite afirmar que el proceso de Arnoldi genera una base ortonormal de los subespacios de Krylov $\mathcal{K}_n(A, b)$.

Lema 2.38. *Los vectores $v_1, v_2, \ldots, v_{n_*}$ generados por el proceso de Arnoldi forman un sistema ortonormal respecto del producto euclídeo $\langle\cdot,\cdot\rangle_2$ en $\mathbb{R}^N$ y*

$$span\{v_1, \ldots, v_n\} = span\{v_1, \ldots, v_{n-1}, Av_{n-1}\} = \mathcal{K}_n(A, b) \text{ para } 1 \leq n \leq n_*.$$

Demostración. La ortonormalidad de los vectores $v_1, v_2, ..., v_{n_*}$ es inmediata por propia construcción, pues es claro que $\|v_j\|_2 = 1$ para $1 \leq j \leq n_*$, y si se asume que $\langle v_j, v_k\rangle_2 = \delta_{jk}$ para $1 \leq j < k \leq n$, entonces dado k tal que $1 \leq k \leq n$:

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

Ahora, por inducción, veamos que

$$span\{v_1, \ldots, v_n\} = span\{v_1, \ldots, v_{n-1}, Av_{n-1}\} = \mathcal{K}_n(A, b) \text{ para } 1 \leq n \leq n_*.$$

La propiedad es obvia para $n = 1$, pues $v_1 = \frac{b}{\|b\|_2}$. Suponemos cierta para un n con $1 \leq n \leq n_* - 1$, y veamos que se cumple para $n+1$. En primer lugar,

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

Además, por hipótesis de inducción, sabemos que $v_1, \ldots, v_n \in \mathcal{K}_n$ y $Av_n \in A\mathcal{K}_n \subseteq \mathcal{K}_{n+1}$. Luego, $span\{v_1, \ldots, v_{n-1}, v_n, v_{n+1}\} \subseteq span\{v_1, \ldots, v_n, Av_n\} \subseteq \mathcal{K}_{n+1}$, siendo $dim(\{v_1, \ldots, v_{n+1}\}) = n+1$ y $dim\,(span\{v_1, \ldots, v_n, Av_n\}) \leq n+1$, $dim(\mathcal{K}_{n+1}) \leq n+1$. Luego, por un argumento de dimensión, los tres subespacios son iguales. □

Observación 2.39. Este lema nos permite afirmar que el número de iteraciones n_* necesarias para que el proceso de Arnoldi pare coincide con el menor valor n_* tal que $\mathcal{K}_{n_*}(A, b) = \mathcal{K}_{n_*+1}(A, b)$. Así pues, en el caso de que A sea simétrica y definida positiva, el número de iteraciones para alcanzar la convergencia con el método del Gradiente Conjugado y GMRES coincide. En otro caso, el número de iteraciones de GMRES y GCNR no tiene porqué coincidir. A continuación, presentamos dos ejemplos mediante los métodos GCNR y GMRES donde observaremos, que dependiendo de las características del sistema lineal, un método puede converger con menos iteraciones que el otro.

Ejemplo 2.40. Sean e_i es el i-ésimo vector canónico de $\mathbb{R}^{N\times N}$, $1 \leq i \leq N$, y

$$A = \begin{bmatrix} 0 & & & 1 \\ 1 & 0 & & \\ & \ddots & \ddots & \\ & & 1 & 0 \end{bmatrix} = \begin{bmatrix} e_N^\top \\ e_1^\top \\ \vdots \\ e_{N-1}^\top \end{bmatrix} \in \mathbb{R}^{N\times N}, \quad b = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \in \mathbb{R}^N, \quad x_* = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix} \in \mathbb{R}^N$$

de modo que x_* es la única solución del sistema lineal $Ax = b$. Demostraremos que:

1. El método GCNR converge exactamente en una sola iteración, $x_1 = x_*$.
   Consideramos el subespacio de Krylov $\mathcal{K}_n(A^\top A, A^\top b)$ donde
   $$A^\top b = \begin{bmatrix} e_2^\top \\ \vdots \\ e_N^\top \\ e_1^\top \end{bmatrix} e_1 = e_N, \quad A^\top A = \begin{bmatrix} e_2^\top \\ \vdots \\ e_N^\top \\ e_1^\top \end{bmatrix} [e_2|...|e_N|e_1] = I_N.$$
   Tenemos que $\mathcal{K}_1(A^\top A, A^\top b) = \mathcal{K}_2(A^\top A, A^\top b) = span\{e_N\}$ y GCNR converge en una iteración.
2. El método GMRES converge exactamente en N iteraciones.
   Tenemos que $b = e_1$, $Ab = e_2$, $A^2 b = e_3$,..., $A^{N-1} b = e_N$, $A^N b = e_1$. Luego, $\mathcal{K}_j(A, b) = span\{e_1, \ldots, e_j\}$ para $1 \le j \le N$, y $\mathcal{K}_{N+1}(A, b) = \mathcal{K}_N(A, b)$. Luego, el método GMRES converge en N iteraciones. Además, si $j = 1, \ldots, N-1$, entonces
   $$\min_{x \in \mathcal{K}_j(A,b)} \|Ax - b\|_2 = \min_{x = \sum_{k=1}^{j} \lambda_k e_k} \left\| \sum_{k=2}^{j+1} \lambda_{k-1} e_k - e_1 \right\|_2 = \min_{\lambda_1, \ldots, \lambda_j \in \mathbb{R}} \sqrt{1 + \lambda_1^2 + \ldots + \lambda_j^2} = 1$$
   que se obtiene con $\lambda_1 = \ldots = \lambda_j = 0$; esto es, $x_j = 0$ si $1 \le j \le N-1$ y $x_N = x_*$.

Este ejemplo muestra una situación en la que el método GCNR converge en una iteración mientras que el método GMRES requiere N iteraciones. A continuación, en el Ejemplo 2.42 ilustraremos la situación recíproca. Antes estudiaremos una propiedad del rango de matrices que nos será de utilidad en el resto del capítulo.

Lema 2.41. *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 $\mathcal{L}_M : \mathbb{R}^q \to \mathbb{R}^p$, $\mathcal{L}_N : \mathbb{R}^r \to \mathbb{R}^q$, $\mathcal{L}_{MN} : \mathbb{R}^r \to \mathbb{R}^p$. Como $dim(Im(\mathcal{L}_M)) = rango(M) = q$, $dim(Ker(\mathcal{L}_M)) = 0$ y $Ker(\mathcal{L}_M) = \{0\}$. Probaremos que $Ker(\mathcal{L}_N) = Ker(\mathcal{L}_{MN})$, de donde se obtendría inmediatamente el enunciado por el primer Teorema de isomorfía.

Es evidente que $Ker(\mathcal{L}_N) \subseteq Ker(\mathcal{L}_{MN})$, pues si $Nv = 0$, entonces $MNv = 0$. Sea $v \in Ker(\mathcal{L}_{MN})$: $MNv = 0$. Entonces, $Nv \in Ker(\mathcal{L}_M) = \{0\}$ y, por tanto, $v \in Ker(\mathcal{L}_N)$. □

Ejemplo 2.42. Sea la matriz tridiagonal $\mathcal{T} = \begin{bmatrix} 1 & \alpha & & \\ \alpha & 1 & \ddots & \\ & \ddots & \ddots & \alpha \\ & & \alpha & 1 \end{bmatrix}$ con $0 < |\alpha| \le \frac{1}{2}$. Los autovalores de $\mathcal{T}$ son $\lambda_j = 1 - 2\sqrt{\alpha^2} \cos\left(\frac{j\pi}{N+1}\right) > 0$ para $j = 1, \ldots, N$, (veáse, p.e., [1, p.349]). Luego, $\mathcal{T}$ es una matriz simétrica y definida positiva. Por lo tanto, admite una descomposición de Cholesky: $LL^\top = \mathcal{T}$ con $l_{11} = 1$, donde L es una matriz triangular inferior inversible. Es fácil ver que L debe ser de la forma $L = \begin{bmatrix} 1 & & & \\ \alpha & * & & \\ \vdots & \ddots & \ddots & \\ 0 & \cdots & \cdots & * \end{bmatrix}$.

Sea $A := L^\top$ y $b := e_1$. Planteamos el sistema $Ax = b$ con solución $x = e_1$ pues $Ae_1 = e_1$. Por tanto,

1. El método GMRES converge en una iteración ya que $Ab = b$ implica $\mathcal{K}_2(A,b) = \mathcal{K}_1(A,b) = span\,\{b\}$.
2. Respecto al método GCNR, debemos considerar los subespacios $\mathcal{K}_n(\underbrace{A^\top A}_{\mathcal{T}}, A^\top b)$, $n \geq 1$.

   Tenemos que

$$
\begin{aligned}
&v_1 := A^\top b = Le_1 = e_1 + \alpha e_2\\
&v_2 := \mathcal{T}v_1 = (1+\alpha^2)e_1 + 2\alpha e_2 + \alpha^2 e_3\\
&v_3 := \mathcal{T}v_2 = \beta_{31}e_1 + \beta_{32}e_2 + \beta_{33}e_3 + \beta_{34}e_4, \text{ con } \beta_{34} = \alpha^3, \beta_{33} = 3\alpha^2\\
&\vdots\\
&v_j := \mathcal{T}v_{j-1} = \beta_{j1}e_1 + \ldots + \beta_{jj}e_j + \beta_{j,j+1}e_{j+1}, \text{ con } \beta_{j,j+1} = \alpha^j, \beta_{jj} = j\alpha^{j-1}\\
&\vdots\\
&v_{N-1} := \mathcal{T}v_{N-2} = \beta_{N-1,1}e_1 + \ldots + \beta_{N-1,N-1}e_{N-1} + \beta_{N-1,N}e_N,\\
&\text{con } \beta_{N-1,N} = \alpha^{N-1}, \beta_{N-1,N-1} = (N-1)\alpha^{N-2}\\
&v_N := \mathcal{T}v_{N-1} = \beta_{N,1}e_1 + \ldots + \beta_{N,N}e_N \text{ con } \beta_{N,N} = N\alpha^{N-1}.
\end{aligned}
$$

   Además, $\mathcal{T}e_1 = e_1 + \alpha e_2 = v_1$. Luego, $\mathcal{T}[e_1|v_1|\ldots|v_{N-1}] = [v_1|v_2|\ldots|v_N]$. Como $\alpha \neq 0$, $\mathcal{T}$ es regular y $rango\,(\mathcal{T}) = N$, por el Lema 2.41, $rango\,([v_1|v_2|\ldots|v_N]) = rango\,([e_1|v_1|\ldots|v_{N-1}]) = N$. Por tanto, $v_1, \ldots, v_N$ son vectores linealmente independientes y el método GCNR converge en N iteraciones.

Observación 2.43. En el resto de esta sección, 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

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

Por tanto, obtenemos una matriz rectangular de Hessenberg superior:

$$H_n = \begin{bmatrix} h_{11} & h_{12} & \ldots & h_{1n} \\ h_{21} & h_{22} & \ldots & h_{2n} \\ \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \ldots & h_{n,n} \\ 0 & 0 & \ldots & h_{n+1,n} \end{bmatrix} \in \mathbb{R}^{(n+1)\times n}. \tag{2.18}$$

Además, según la Observación 2.37, si A es una matriz simétrica, entonces $(H_n)_{i,j} = 0$, si $|i-j| \geq 2$.

Teorema 2.44. *Sea una matriz $A \in \mathbb{R}^{N\times N}$ y un vector $b \in \mathbb{R}^N$ no nulo. Asumiendo la notación (2.18) del proceso de Arnoldi, se tiene que*

$$AV_n = V_{n+1}H_n, \quad n = 1, \ldots, n_* - 1, \quad y \quad AV_{n_*} = V_{n_*}H_{n_*} \tag{2.19}$$

donde

$$V_j = [v_1|\ldots|v_j] \in \mathbb{R}^{N\times j},\ 1 \leq j \leq n_*, \quad y \quad H_{n_*} = \begin{bmatrix} h_{11} & \ldots & \ldots & h_{1n_*} \\ h_{21} & \ddots & & \vdots \\ & \ddots & \ddots & \vdots \\ & & h_{n_*,n_*-1} & h_{n_*n_*} \end{bmatrix} \in \mathbb{R}^{n_*\times n_*}. \tag{2.20}$$

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 $1 \leq n \leq n_*$. □

Observación 2.45. Observar que $AV_{n_*} = V_{n_*}H_{n_*}$ implica que $V_{n_*}^\top AV_{n_*} = H_{n_*}$. Esto es, A es semejante, por transformación ortogonal, a la matriz de Hessenberg cuadrada superior definida en (2.20).

Seguidamente, desarrollaremos un procedimiento para implementar el método GMRES que emplea las bases ortogonales de los subespacios de Krylov generados mediante el proceso de Arnoldi.

Teorema 2.46. *Sean $x_1, \ldots, x_{n_*} \in \mathbb{R}^N$ las iteraciones del método GMRES aplicado al sistema lineal $Ax = b$ con $A \in \mathbb{R}^{N\times N}$ matriz regular y vector $b \in \mathbb{R}^N$ no nulo. Para cada $n = 1, \ldots, n_*$, el problema de mínimos cuadrados*

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

admite solución única $z_n \in \mathbb{R}^n$. Además, $x_n = V_n z_n$ y $\|Ax_n - b\|_2 = \|H_n z_n - c_n\|_2$ para $1 \le n \le n_*$, donde V_n y H_n están dadas por (2.20) y (2.18) respectivamente.

Demostración. Por (2.19), observamos que si $1 \le n \le n_*$, entonces

$$Ax - b = V_{\widehat{n}} H_n V_n^\top x - V_{\widehat{n}} c_n \quad \text{con } \widehat{n} = \min\{n+1, n_*\}.$$

Luego, como $V_j^\top V_j = I_j$ para $1 \le j \le n_*$, sigue que $\|Ax - b\|_2 = \|H_n V_n^\top x - c_n\|_2$. Ahora, por el Lema 2.38, $x \in \mathcal{K}_n(A, b) = span\{v_1, \ldots, v_n\}$ si y sólo si existe un único $z \in \mathbb{R}^n$ tal que $x = V_n z$. Además, puesto que $V_j V_j^\top = I_N$ para $1 \le j \le n_*$, $V_n^\top x = V_n^\top y$ si y sólo si $x = y$. Puesto que existe un único vector $x_n \in \mathcal{K}_n(A, b)$ tal que $\|Ax_n - b\|_2 = \min_{x\in\mathcal{K}_n(A,b)} \|Ax - b\|_2$, existe un único $z_n \in \mathbb{R}^n$ tal que $\|H_n z_n - c_n\|_2 = \min_{z\in\mathbb{R}^n} \|H_n z - c_n\|_2$ y $x_n = V_n z_n$. □

Observación 2.47. La matriz de Hessenberg $H_n \in \mathbb{R}^{\widehat{n}\times n}$, donde $\widehat{n} = min\{n+1, n_*\}$ para $1 \le n \le n_*$, 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, $Q_n^\top = Q_n^{-1}$, 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, $\widehat{R}_n := \left((R_n)_{i,j}\right)_{i,j=1}^{n} \in \mathbb{R}^{n\times n}$ es una matriz regular. Probaremos esta propiedad en el Lema 2.50 y la utilizaremos a continuación para dar una expresión cerrada para las iteraciones del método GMRES y las correspondientes normas de los residuales.

Teorema 2.48. *Para cada $n \in \{1, \ldots, n_*\}$, sea $z_n \in \mathbb{R}^n$ es la única solución del problema de mínimos cuadrados definido en el Teorema 2.46 con $H_n = Q_n R_n$, donde Q_n y R_n son las matrices definidas en la Observación 2.47. Entonces,*

(i) $z_n = \widehat{R}_n^{-1} I_{n\times\widehat{n}} Q_n^\top c_n$, donde $I_{n\times\widehat{n}}$ es la matriz identidad rectangular de dimensión $n\times\widehat{n}$.

(ii) $\|H_n z - c_n\|_2 = \rho_n$ siendo

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

Demostración. Por hipótesis, sabemos que $H_n z - c_n = Q_n\left(R_n z - Q_n^\top c_n\right)$. Por lo tanto, $\|H_n z - c_n\|_2 = \|R_n z - Q_n^\top c_n\|_2$. 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 valor mínimo igual a cero con $z = z_n := \widehat{R}_n^{-1} Q_n^\top c_n$.
2. Si $n < n_*$, $R_n z = \begin{bmatrix} \widehat{R}_n z \\ \hline 0 \end{bmatrix}$ y $Q_n^\top c_n = \left[\begin{array}{c} I_{n\times(n+1)} Q_n^\top c_n \\ \hline e_{n+1}^\top Q_n^\top c_n \end{array}\right]$. Luego, $\|R_n z - Q_n^\top c_n\|_2^2 = |e_{n+1}^\top Q_n^\top c_n|^2 + \|\widehat{R}_n z - I_{n\times(n+1)} Q_n^\top c_n\|_2^2$, y esta expresión se minimiza tomando $z = z_n := \widehat{R}_n^{-1} I_{n\times\widehat{n}} Q_n^\top c_n$, con valor mínimo $\|R_n z - Q_n^\top c_n\|_2 = |e_{n+1}^\top Q_n^\top c_n|$. □

Observación 2.49. El siguiente lema nos dice que las matrices superiores de Hessenberg H_n, para $1 \leq n \leq n_*$, 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:

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

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 $a = \frac{c}{r}$ y $b = \frac{s}{r}$. Si $\begin{pmatrix} c \\ s \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ basta tomar $a=1$, $b=0$.

Lema 2.50. *La matriz H_n (2.18)-(2.20) 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 2.47.*

Demostración. Denotemos a la matriz Hessenberg como $H_n^{(0)} := H_n$. Para cada $i \in \{1,\ldots,n\}$ 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 $R_n := H_n^{(n)} = G_n G_{n-1} \ldots G_1 H_n$ verifica que $(R_n)_{i,j} = 0$ para $i > j$. Además, puesto que $G_i^\top = G_i^{-1}$, $1 \leq i \leq n$, definiendo la matriz $Q_n := G_1^\top \ldots G_{n-1}^\top G_n^\top$ obtenemos $Q_n Q_n^\top = Q_n^\top Q_n = I$ y $H_n = Q_n R_n$.

Finalmente, veamos que la matriz $\widehat{R}_n := \left((R_n)_{i,j}\right)_{i,j=1}^{n}$ es regular. Por (2.19), 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 2.41, 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. □

Corolario 2.51. *Las iteraciones $x_1,\ldots,x_{n_*}$ del método GMRES se pueden expresar como $x_n = \left(V_n \widehat{R}_n^{-1} I_{n\times\widehat{n}} Q_n^\top e_1\right)\|b\|_2$, con $e_1 = (1,0,\ldots,0)^\top \in \mathbb{R}^{\widehat{n}}$, y residual*

$$Ax_n - b = \begin{cases} \left(V_{n+1}Q_n e_{n+1}\right)\left(-e_{n+1}^\top Q_n^\top e_1\right)\|b\|_2 & , 1 \leq n \leq n_*-1, \\ 0 & , n = n_*. \end{cases}$$

Demostración. Por (2.19), $Ax_n - b = V_{n+1}H_n V_n^\top x_n - V_{n+1}c_n = V_{n+1}(H_n z_n - c_n)$, con $z_n = V_n^\top x_n$. Aplicando el Lema 2.50, obtenemos que $V_{n+1}(H_n z_n - c_n) = V_{n+1}Q_n\left(R_n z_n - Q_n^\top c_n\right)$. Finalmente, usando el procedimiento de la demostración del Teorema 2.48, para $1 \leq n \leq n_*-1$

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

□

El siguiente resultado da una estimación de convergencia del método GMRES, en la línea del Teorema 2.25 para el método del Gradiente Conjugado.

Teorema 2.52. *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 $D := diag\,(\lambda_1,\ldots,\lambda_N)$. Entonces, si x_n denota la enésima iteración del método GMRES:*

$$\|Ax_n - b\|_2 \leq cond_2(T)\left(\inf_{\substack{p\in\prod_n \\ p(0)=1}} \max_{k=1,\ldots,N} |p(\lambda_k)|\right)\|b\|_2.$$

Demostración. Para cualquier polinomio $p \in \prod_n$ con $p(0) = 1$, definimos el polinomio $q(t) := \frac{1-p(t)}{t}$ de grado como máximo $n-1$. Luego, definiendo $x := q(A)b \in \mathcal{K}_n(A,b)$ tenemos que $Ax - b = -p(A)b$. La propiedad minimal del método GMRES implica que

$$\|Ax_n - b\|_2 \leq \|Ax - b\|_2 = \|p(A)b\|_2 \leq \|p(A)\|_2 \|b\|_2.$$

Ahora, como A es una matriz diagonalizable

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

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

Corolario 2.53. *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 m iteraciones.*

Demostración. Definimos el polinomio $p(t) = \prod_{i=1}^{m} \frac{\lambda_i - t}{\lambda_i} \in \prod_m$ con $p(0) = 1$. Entonces, $\max_{k=1,..,N} |p(\lambda_k)| = 0$. Y, por el Teorema 2.52, tenemos $\|Ax_m - b\|_2 = 0$. □

Observación 2.54 (Implementación del método GMRES basado en el proceso de Arnoldi y rotaciones de Givens).

Finalmente, trataremos la implementación práctica de las iteraciones método GMRES basado en el proceso de Arnoldi y rotaciones de Givens. Consideramos el sistema $Ax = b$ donde $A \in \mathbb{R}^{N\times N}$ es una matriz regular y $b \in \mathbb{R}^N$ un vector no nulo.

Siguiendo la notación antes introducida:

$\boxed{i=1}$ Definimos $v_1 = \frac{b}{\|b\|_2}$, $h_{11} := h_{11}^{(0)} = \langle Av_1, v_1\rangle_2$, $\widehat{v}_2 = Av_1 - h_{11}v_1$ y $h_{21} := h_{21}^{(0)} = \|\widehat{v}_2\|_2$. Sea $H_1 = H_1^{(0)} = \begin{bmatrix} h_{11}^{(0)} \\ h_{21}^{(0)} \end{bmatrix}$. Obtenemos $G_1^{[2]} = \begin{pmatrix} a_1 & b_1 \\ -b_1 & a_1 \end{pmatrix}$ tal que

$$G_1^{[2]} H_1^{(0)} = \begin{bmatrix} h_{11}^{(1)} \\ 0 \end{bmatrix} =: H_1^{(1)} =: R_1 = \begin{bmatrix} r_{11} \\ 0 \end{bmatrix}.$$

Sea $c_1^{(0)} = \|b\|_2$ y calcular $G_1^{[2]} \begin{bmatrix} c_1^{(0)} \\ 0 \end{bmatrix} = \begin{bmatrix} c_1^{(1)} \\ c_2^{(1)} \end{bmatrix}$. Entonces, $z_1 = \frac{c_1^{(1)}}{r_{11}}$, $x_1 = z_1 v_1$ y $\|Ax_1 - b\|_2 = |c_2^{(1)}|$. Si $|c_2^{(1)}| > TOL$, donde TOL es una cantidad previamente fijada, entonces se sigue iterando.

$\boxed{i=2}$ Definimos $v_2 = \frac{\widehat{v}_2}{\|\widehat{v}_2\|_2}$, $h_{i2} := h_{i2}^{(0)} = \langle Av_2, v_i\rangle_2$ para $1 \leq i \leq 2$, $\widehat{v}_3 = Av_2 - \sum_{i=1}^{2} h_{i2}v_i$ y $h_{32} := h_{32}^{(0)} = \|\widehat{v}_3\|_2$. Sea $H_2 = H_2^{(0)} = \begin{bmatrix} h_{11}^{(0)} & h_{12}^{(0)} \\ h_{21}^{(0)} & h_{22}^{(0)} \\ 0 & h_{32}^{(0)} \end{bmatrix}$. Tomando $G_1^{[3]} = \left(\begin{array}{c|c} G_1^{[2]} & 0 \\ \hline 0 & 1 \end{array}\right)$ se tiene que $G_1^{[3]} H_2^{(0)} = \begin{bmatrix} h_{11}^{(1)} & h_{12}^{(1)} \\ 0 & h_{22}^{(1)} \\ 0 & h_{32}^{(1)} \end{bmatrix} =: H_2^{(1)}$ con $h_{32}^{(1)} = h_{32}^{(0)}$ y $\begin{bmatrix} h_{12}^{(1)} \\ h_{22}^{(1)} \end{bmatrix} = G_1^{[2]} \begin{bmatrix} h_{12}^{(0)} \\ h_{22}^{(0)} \end{bmatrix}$.

Además, $G_1^{[3]} \begin{bmatrix} c_1^{(0)} \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} c_1^{(1)} \\ c_2^{(1)} \\ 0 \end{bmatrix}$. Ahora, elegimos $G_2^{[2]} = \begin{pmatrix} a_2 & b_2 \\ -b_2 & a_2 \end{pmatrix}$ tal que $G_2^{[2]} \begin{bmatrix} h_{22}^{(1)} \\ h_{32}^{(1)} \end{bmatrix} = \begin{bmatrix} h_{22}^{(2)} \\ 0 \end{bmatrix}$. Entonces, con $G_2^{[3]} = \left(\begin{array}{c|c} 1 & 0 \\ \hline 0 & G_2^{[2]} \end{array}\right)$ sigue que $G_2^{[3]} H_2^{(1)} = \begin{bmatrix} h_{11}^{(1)} & h_{12}^{(1)} \\ 0 & h_{22}^{(2)} \\ 0 & 0 \end{bmatrix} =: H_2^{(2)} =:$

$R_2 = \begin{bmatrix} r_{11} & r_{12} \\ 0 & r_{22} \\ 0 & 0 \end{bmatrix}$. Además, $G_2^{[3]} \begin{bmatrix} c_1^{(1)} \\ c_2^{(1)} \\ 0 \end{bmatrix} = \begin{bmatrix} c_1^{(1)} \\ c_2^{(2)} \\ c_3^{(2)} \end{bmatrix}$ con $\begin{bmatrix} c_2^{(2)} \\ c_3^{(2)} \end{bmatrix} := G_2^{[2]} \begin{bmatrix} c_2^{(1)} \\ 0 \end{bmatrix}$. Por tanto, $z_2 = \begin{bmatrix} r_{11} & r_{12} \\ 0 & r_{22} \end{bmatrix}^{-1} \begin{bmatrix} c_1^{(1)} \\ c_2^{(2)} \end{bmatrix}$, $x_2 = \left[\, v_1 \middle| v_2 \,\right] \cdot z_2$ y $\|Ax_2 - b\|_2 = |c_3^{(2)}|$. Si $|c_3^{(2)}| > TOL$, seguir.

$\boxed{i = 3}$ Definimos $v_3 = \frac{\widehat{v}_3}{\|\widehat{v}_3\|_2}$, $h_{i3} := h_{i3}^{(0)} = \langle Av_3, v_i \rangle_2$ para $1 \leq i \leq 3$, $\widehat{v}_4 = Av_3 - \sum_{i=1}^{3} h_{i3} v_i$ y $h_{43} := h_{43}^{(0)} = \|\widehat{v}_4\|_2$. Sea $H_3 = H_3^{(0)} = \begin{bmatrix} h_{11}^{(0)} & h_{12}^{(0)} & h_{13}^{(0)} \\ h_{21}^{(0)} & h_{22}^{(0)} & h_{23}^{(0)} \\ 0 & h_{32}^{(0)} & h_{33}^{(0)} \\ 0 & 0 & h_{43}^{(0)} \end{bmatrix}$. Tomando $G_1^{[4]} = \left(\begin{array}{c|c} G_1^{[2]} & 0 \\ \hline 0 & I_2 \end{array}\right)$ y $G_2^{[4]} = \left(\begin{array}{c|c|c} 1 & 0 & 0 \\ \hline 0 & G_2^{[2]} & 0 \\ \hline 0 & 0 & 1 \end{array}\right)$ se tiene que $G_1^{[4]} H_3^{(0)} = \begin{bmatrix} h_{11}^{(1)} & h_{12}^{(1)} & h_{13}^{(1)} \\ 0 & h_{22}^{(1)} & h_{23}^{(1)} \\ 0 & h_{32}^{(1)} & h_{33}^{(1)} \\ 0 & 0 & h_{43}^{(1)} \end{bmatrix} =: H_3^{(1)}$ con $\begin{cases} h_{32}^{(1)} = h_{32}^{(0)} \\ h_{33}^{(1)} = h_{33}^{(0)} \\ h_{43}^{(1)} = h_{43}^{(0)} \end{cases}$ y $\begin{bmatrix} h_{13}^{(1)} \\ h_{23}^{(1)} \end{bmatrix} = G_1^{[2]} \begin{bmatrix} h_{13}^{(0)} \\ h_{23}^{(0)} \end{bmatrix}$. Además, $G_2^{[4]} H_3^{(1)} = \begin{bmatrix} h_{11}^{(1)} & h_{12}^{(1)} & h_{13}^{(1)} \\ 0 & h_{22}^{(2)} & h_{23}^{(2)} \\ 0 & 0 & h_{33}^{(2)} \\ 0 & 0 & h_{43}^{(2)} \end{bmatrix} =: H_3^{(2)}$ con $h_{43}^{(2)} = h_{43}^{(1)}$ y $\begin{bmatrix} h_{23}^{(2)} \\ h_{33}^{(2)} \end{bmatrix} = G_2^{[2]} \begin{bmatrix} h_{23}^{(1)} \\ h_{33}^{(1)} \end{bmatrix}$. Ahora, elegimos $G_3^{[2]} = \begin{pmatrix} a_3 & b_3 \\ -b_3 & a_3 \end{pmatrix}$ tal que $G_3^{[2]} \begin{bmatrix} h_{33}^{(2)} \\ h_{43}^{(2)} \end{bmatrix} = \begin{bmatrix} h_{33}^{(3)} \\ 0 \end{bmatrix}$. Entonces, con $G_3^{[4]} = \left(\begin{array}{c|c} I_2 & 0 \\ \hline 0 & G_3^{[2]} \end{array}\right)$ sigue que $G_3^{[4]} H_3^{(2)} = \begin{bmatrix} h_{11}^{(1)} & h_{12}^{(1)} & h_{13}^{(1)} \\ 0 & h_{22}^{(2)} & h_{23}^{(2)} \\ 0 & 0 & h_{33}^{(3)} \\ 0 & 0 & 0 \end{bmatrix} =: H_3^{(3)} =: R_3 = \begin{bmatrix} r_{11} & r_{12} & r_{13} \\ 0 & r_{22} & r_{23} \\ 0 & 0 & r_{33} \\ 0 & 0 & 0 \end{bmatrix}$. Por otra parte, $G_3^{[4]} G_2^{[4]} G_1^{[4]} \begin{bmatrix} c_1^{(0)} \\ 0 \\ 0 \\ 0 \end{bmatrix} = G_3^{[4]} G_2^{[4]} \begin{bmatrix} c_1^{(1)} \\ c_2^{(1)} \\ 0 \\ 0 \end{bmatrix} = G_3^{[4]} \begin{bmatrix} c_1^{(1)} \\ c_2^{(2)} \\ c_3^{(2)} \\ 0 \end{bmatrix} = \begin{bmatrix} c_1^{(1)} \\ c_2^{(2)} \\ c_3^{(3)} \\ c_4^{(3)} \end{bmatrix}$ donde $\begin{bmatrix} c_3^{(3)} \\ c_4^{(3)} \end{bmatrix} := G_3^{[2]} \begin{bmatrix} c_3^{(2)} \\ 0 \end{bmatrix}$. En definitiva, $z_3 = \begin{bmatrix} r_{11} & r_{12} & r_{13} \\ 0 & r_{22} & r_{23} \\ 0 & 0 & r_{33} \end{bmatrix}^{-1} \begin{bmatrix} c_1^{(1)} \\ c_2^{(2)} \\ c_3^{(3)} \end{bmatrix}$, $x_3 = \left[\, v_1 \middle| v_2 \middle| v_3 \,\right] \cdot z_3$ y $\|Ax_3 - b\|_2 = |c_4^{(3)}|$. Si $|c_4^{(3)}| > TOL$, seguir.

$\boxed{i = n}$ Suponemos que en las iteraciones previas $1 \leq i \leq n-1$ se han obtenido matrices $G_i^{[2]} = \begin{pmatrix} a_i & b_i \\ -b_i & a_i \end{pmatrix}$ tales que $G_i^{[2]} \begin{bmatrix} h_{ii}^{(i-1)} \\ h_{i+1,i}^{(i-1)} \end{bmatrix} = \begin{bmatrix} h_{ii}^{(i)} \\ 0 \end{bmatrix}$, y que se han calculado los siguientes coeficientes:

$$G_1^{[2]} \begin{bmatrix} h_{1i}^{(0)} \\ h_{2i}^{(0)} \end{bmatrix} = \begin{bmatrix} h_{1i}^{(1)} \\ h_{2i}^{(1)} \end{bmatrix}, \text{ para } 1 \leq i \leq n-1,$$

$$G_2^{[2]} \begin{bmatrix} h_{2i}^{(1)} \\ h_{3i}^{(1)} \end{bmatrix} = \begin{bmatrix} h_{2i}^{(2)} \\ h_{3i}^{(2)} \end{bmatrix}, \text{ para } 2 \leq i \leq n-1,$$

y

$$G_j^{[2]} \begin{bmatrix} h_{ji}^{(j-1)} \\ h_{j+1,i}^{(j-1)} \end{bmatrix} = \begin{bmatrix} h_{ji}^{(j)} \\ h_{j+1,i}^{(j)} \end{bmatrix}, \text{ para } j \leq i \leq n-1,$$

así como también, con $c_1^{(0)} := \|b\|_2$,

$$G_j^{[2]} \begin{bmatrix} c_j^{(j-1)} \\ 0 \end{bmatrix} = \begin{bmatrix} c_j^{(j)} \\ c_{j+1}^{(j)} \end{bmatrix} \text{ para } 1 \leq j \leq n-1.$$

Definimos entonces $v_n = \frac{\widehat{v}_n}{\|\widehat{v}_n\|_2}$, $h_{in} := h_{in}^{(0)} = \langle Av_n, v_i \rangle_2$ para $1 \leq i \leq n$, $\widehat{v}_{n+1} = Av_n - \sum_{i=1}^{n} h_{in} v_i$ y $h_{n+1,n} := h_{n+1,n}^{(0)} = \|\widehat{v}_{n+1}\|_2$.

Con $G_j^{[n+1]} = \left(\begin{array}{c|c|c} I_{j-1} & 0 & 0 \\ \hline 0 & G_j^{[2]} & 0 \\ \hline 0 & 0 & I_{n-j} \end{array}\right)$ para $1 \leq j \leq n-1$, sigue que

$$G_{n-1}^{[n+1]} ... G_2^{[n+1]} G_1^{[n+1]} H_n^{(0)} = G_{n-1}^{[n+1]} ... G_2^{[n+1]} H_n^{(1)} = ... = G_{n-1}^{[n+1]} H_n^{(n-2)} = H_n^{(n-1)},$$

donde se han de calcular

$$G_1^{[2]} \begin{bmatrix} h_{1n}^{(0)} \\ h_{2n}^{(0)} \end{bmatrix} =: \begin{bmatrix} h_{1n}^{(1)} \\ h_{2n}^{(1)} \end{bmatrix}, G_2^{[2]} \begin{bmatrix} h_{2n}^{(1)} \\ h_{3n}^{(1)} \end{bmatrix} =: \begin{bmatrix} h_{2n}^{(2)} \\ h_{3n}^{(2)} \end{bmatrix} \text{ y en general } G_j^{[2]} \begin{bmatrix} h_{jn}^{(j-1)} \\ h_{j+1,n}^{(j-1)} \end{bmatrix} =: \begin{bmatrix} h_{jn}^{(j)} \\ h_{j+1,n}^{(j)} \end{bmatrix}$$

para $1 \leq j \leq n-1$. Observar que

$$H_n^{(n-1)} = \begin{bmatrix} h_{11}^{(1)} & h_{12}^{(1)} & \dots & h_{1,n-1}^{(1)} & h_{1n}^{(1)} \\ 0 & h_{22}^{(2)} & \dots & h_{2,n-1}^{(2)} & h_{2n}^{(2)} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & h_{n-1,n-1}^{(n-1)} & h_{n-1,n}^{(n-1)} \\ 0 & 0 & \dots & 0 & h_{nn}^{(n-1)} \\ 0 & 0 & \dots & 0 & h_{n+1,n}^{(n-1)} \end{bmatrix} \text{ siendo } h_{n+1,n}^{(n-1)} = h_{n+1,n}^{(0)} = h_{n+1,n}^{(j)}, 0 \leq j \leq n-1.$$

Además,

$$G_{n-1}^{[n+1]} ... G_2^{[n+1]} G_1^{[n+1]} \begin{bmatrix} c_1^{(0)} \\ 0 \\ \vdots \\ 0 \end{bmatrix} = G_{n-1}^{[n+1]} ... G_2^{[n+1]} \begin{bmatrix} c_1^{(1)} \\ c_2^{(1)} \\ 0 \\ \vdots \\ 0 \end{bmatrix} = ... = G_{n-1}^{[n+1]} \begin{bmatrix} c_1^{(1)} \\ c_2^{(2)} \\ c_3^{(3)} \\ \vdots \\ c_{n-1}^{(n-2)} \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} c_1^{(1)} \\ c_2^{(2)} \\ c_3^{(3)} \\ \vdots \\ c_{n-1}^{(n-1)} \\ c_n^{(n-1)} \\ 0 \end{bmatrix}$$

que ya es conocido de iteraciones previas. Determinamos una nueva matriz $G_n^{[2]} = \begin{bmatrix} a_n & b_n \\ -b_n & a_n \end{bmatrix}$ tal que $G_n^{[2]} \begin{bmatrix} h_{nn}^{(n-1)} \\ h_{n+1,n}^{(n-1)} \end{bmatrix} = \begin{bmatrix} h_{nn}^{(n)} \\ 0 \end{bmatrix}$. Finalmente, con la matriz $G_n^{[n+1]} = \left(\begin{array}{c|c} I_{n-1} & 0 \\ \hline 0 & G_n^{[2]} \end{array}\right)$ se obtiene que

$$G_n^{[n+1]} H_n^{(n-1)} = \begin{bmatrix} h_{11}^{(1)} & h_{12}^{(1)} & \dots & h_{1n}^{(1)} \\ 0 & h_{22}^{(2)} & \dots & h_{2n}^{(2)} \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \dots & h_{nn}^{(n)} \\ 0 & 0 & \dots & 0 \end{bmatrix} =: H_n^{(n)} =: R_n = \begin{bmatrix} r_{11} & r_{12} & \dots & r_{1n} \\ 0 & r_{22} & \dots & r_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & r_{nn} \\ 0 & 0 & \dots & 0 \end{bmatrix}$$

$$\text{y } G_n^{[n+1]} \begin{bmatrix} c_1^{(1)} \\ c_2^{(2)} \\ c_3^{(2)} \\ \vdots \\ c_{n-1}^{(n-1)} \\ c_n^{(n-1)} \\ 0 \end{bmatrix} = \begin{bmatrix} c_1^{(1)} \\ c_2^{(2)} \\ c_3^{(3)} \\ \vdots \\ c_{n-1}^{(n-1)} \\ c_n^{(n)} \\ c_{n+1}^{(n)} \end{bmatrix} \text{ siendo } G_n^{[2]} \begin{bmatrix} c_n^{(n-1)} \\ 0 \end{bmatrix} = \begin{bmatrix} c_n^{(n)} \\ c_{n+1}^{(n)} \end{bmatrix}. \text{ Por tanto, } z_n = \begin{bmatrix} r_{11} & r_{12} & \dots & r_{1n} \\ 0 & r_{22} & \dots & r_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & r_{nn} \end{bmatrix}^{-1} \begin{bmatrix} c_1^{(1)} \\ \vdots \\ c_n^{(n)} \end{bmatrix}, \; x_n = \left[v_1 | \dots | v_n \right] \cdot z_n \text{ y } \|Ax_n - b\|_2 = |c_{n+1}^{(n)}|.$$

El desarrollo previo para el método GMRES se reduce al siguiente algoritmo:

Algoritmo 4 Método GMRES con proceso de Arnoldi y rotaciones de Givens.

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.

1: Definir $v_1 := \frac{b}{\|b\|_2} \in \mathbb{R}^N$, $x_0 := 0$ y $c := \|b\|_2 e_1 \in \mathbb{R}^N$. 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: Aplicar la rotación de Givens $G_j = \begin{bmatrix} a_j & b_j \\ -b_j & a_j \end{bmatrix}$ a $(h_{jn}, h_{j+1,n})^\top$ 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 $|c_{n+1}| < TOL$, entonces resolver $(h_{ij})_{i,j=1}^n z = (c_i)_{i=1}^n$ y tomar $x_n = x_0 + \sum_{j=1}^{n} z_j v_j$.
- En otro caso, calcular $v_{n+1} := \frac{\widehat{v}_{n+1}}{\|\widehat{v}_{n+1}\|_2}$ y volver al paso 2 con $n = n+1$.

Observación 2.55 (GMRES con reinicialización, o GMRES de m pasos). La implementación del método GMRES requiere almacenar los vectores $v_1, \dots, v_n$ generados por el proceso de Arnoldi en una matriz $V_n = \left[v_1 | \dots | v_n \right] \in \mathbb{R}^{N \times n}$ para el cálculo de la iteración x_n. Si $N \gg 1$, este almacenamiento puede ser poco práctico o inviable si el número de iteraciones n es grande. Por ello, en la práctica, se considera una variante del método GMRES con reinicialización del proceso de Arnoldi. Sea $m \in \mathbb{N}$ con $1 \leq m \leq N$ fijo. Se llama método GMRES de m pasos, o GMRES(m), al siguiente algoritmo:

1. Dada una aproximación inicial x_0, dar m iteraciones del método GMRES para obtener x_m, con $r_m = Ax_m - b$.
2. Si $\|r_m\|_2 \leq TOL$ parar. En caso contrario, volver al punto 1 con $x_0 := x_m$.

Claramente este proceso limita a dimensión $N \times m$ el tamaño de la matriz para almacenar los vectores generados por el proceso de Arnoldi. La elección de m en el método GMRES(m) es una cuestión empírica que depende de cada problema concreto. Incluso para un mismo problema, los comportamientos de GMRES(m) y GMRES($m+1$) pueden ser muy diferentes.

A diferencia del método GMRES en el cual siempre se alcanza la convergencia en el peor de los casos dando $n = N$ iteraciones, el método GMRES(m), con $m < N$, no tiene porqué converger. De hecho, sus iteraciones se pueden estancar en un valor concreto diferente a la solución. Un ejemplo de esta situación se da al aplicar GMRES(m), con $m < N$, al Ejemplo 2.40 con $x_0 = 0$, de modo que se obtendría $x_n = 0$, para todo $n \geq 0$.

2.4. Iteraciones precondicionadas.

En las secciones 2.2 y 2.3, hemos obtenido estimaciones de convergencia para los métodos GC, GCNR y GMRES que indican que la tasa de convergencia puede verse afectada por el número de condición, respecto de la norma euclídea, de cierta matriz relacionada con la matriz de coeficientes del sistema $Ax = b$. Esto puede verse en el Teorema 2.27, la Observación 2.31(ii) y el Teorema 2.52, respectivamente. Así pues, los métodos GC, GCNR y GMRES pueden experimentar lenta convergencia si este número de condición es grande. De hecho, en el próximo capítulo mostraremos que esta lenta convergencia es habitual en problemas prácticos.

Para mitigar los problemas de condicionamiento, una alternativa consiste en *precondicionar* el sistema lineal $Ax = b$, esto es, considerar una matriz regular M, denominada precondicionador, para transformar 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 y en [7, Cap. 9] pueden verse otras variantes de precondicionamiento.

(i) 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 $\langle u, v\rangle_M = v^\top Mu$, para $u, v \in \mathbb{R}^N$, siendo M una matriz simétrica y definida positiva. Considerando los subespacios de Krylov $\mathcal{K}_n\left(M^{-1}A, M^{-1}b\right)$ y el método del residual ortogonal (2.4) respecto de $\langle\cdot,\cdot\rangle_M$ se obtiene el siguiente algoritmo:

Algoritmo 5 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 $r_j = Ax_j - b$, $\widetilde{r}_j = M^{-1}r_j$ para $j \geq 0$.

1: Definimos $x_0 := 0$, $r_0 := -b$, $\widetilde{r}_0 := M^{-1}r_0$ y $\widetilde{d}_0 := -\widetilde{r}_0$.
2: Dados x_n, r_n, $\widetilde{r}_n$ y $\widetilde{d}_n$, mientras $r_n \neq 0$, calcular

$$x_{n+1} = x_n + \alpha_n \widetilde{d}_n, \text{ con } \alpha_n = \frac{\langle \widetilde{r}_n, \widetilde{r}_n\rangle_M}{\langle M^{-1}A\widetilde{d}_n, \widetilde{d}_n\rangle_M} = \frac{\langle r_n, \widetilde{r}_n\rangle_2}{\langle A\widetilde{d}_n, \widetilde{d}_n\rangle_2}$$

$$r_{n+1} = r_n + \alpha_n A\widetilde{d}_n$$

$$\widetilde{r}_{n+1} = \widetilde{r}_n + \alpha_n \left(M^{-1}A\right)\widetilde{d}_n = M^{-1}\left[r_n + \alpha_n A\widetilde{d}_n\right] = M^{-1}r_{n+1}$$

$$\widetilde{d}_{n+1} = -\widetilde{r}_{n+1} + \beta_n \widetilde{d}_n, \text{ con } \beta_n = \frac{\langle \widetilde{r}_{n+1}, \widetilde{r}_{n+1}\rangle_M}{\langle \widetilde{r}_n, \widetilde{r}_n\rangle_M} = \frac{\langle r_{n+1}, \widetilde{r}_{n+1}\rangle_2}{\langle r_n, \widetilde{r}_n\rangle_2}.$$

Observación 2.56. Observar que cada iteración del método PGC requiere hallar $A\widetilde{d}_n$ y resolver el sistema lineal $M\widetilde{r}_{n+1} = r_n + \alpha_n A\widetilde{d}_n$.

(ii) El método GCNR Precondicionado (PGCNR): análogamente consiste en aplicar el método GC al sistema precondicionado $M^{-1}A^\top Ax = M^{-1}A^\top b$ respecto del producto interior $\langle\cdot,\cdot\rangle_M$ siendo M una matriz simétrica y definida positiva. Considerando ahora los subespacios de Krylov $\mathcal{K}_n\left(M^{-1}A^\top A, M^{-1}A^\top b\right)$ y el método del residual ortogonal (2.4) respecto de $\langle\cdot,\cdot\rangle_M$ se obtiene el siguiente algoritmo:

Algoritmo 6 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 $r_j = Ax_j - b$, $\widehat{r}_j = A^\top r_j$ y $\widetilde{r}_j = M^{-1}\widehat{r}_j$ para $j \geq 0$.

1: Definimos $x_0 := 0$, $r_0 := -b$, $\widehat{r}_0 = A^\top r_0$, $\widetilde{r}_0 := M^{-1}\widehat{r}_0$ y $\widetilde{d}_0 := -\widetilde{r}_0$.
2: Dados x_n, r_n, $\widehat{r}_n$, $\widetilde{r}_n$ y $\widetilde{d}_n$, mientras $r_n \neq 0$, calcular

$$x_{n+1} = x_n + \alpha_n\widetilde{d}_n, \text{ con } \alpha_n = \frac{\langle\widetilde{r}_n,\widetilde{r}_n\rangle_M}{\langle M^{-1}A^\top A\widetilde{d}_n,\widetilde{d}_n\rangle_M} = \frac{\langle\widehat{r}_n,\widetilde{r}_n\rangle_2}{\|A\widetilde{d}_n\|_2^2}$$

$$r_{n+1} = r_n + \alpha_n A\widetilde{d}_n$$

$$\widehat{r}_{n+1} = A^\top r_{n+1}$$

$$\widetilde{r}_{n+1} = \widetilde{r}_n + \alpha_n\left(M^{-1}A^\top A\right)\widetilde{d}_n = M^{-1}\left[\widehat{r}_n + \alpha_n A^\top A\widetilde{d}_n\right] = M^{-1}A^\top\left[r_n + \alpha_n A\widetilde{d}_n\right]$$

$$\widetilde{d}_{n+1} = -\widetilde{r}_{n+1} + \beta_n\widetilde{d}_n, \text{ con } \beta_n = \frac{\langle\widetilde{r}_{n+1},\widetilde{r}_{n+1}\rangle_M}{\langle\widetilde{r}_n,\widetilde{r}_n\rangle_M} = \frac{\langle\widehat{r}_{n+1},\widetilde{r}_{n+1}\rangle_2}{\langle\widehat{r}_n,\widetilde{r}_n\rangle_2}.$$

Observación 2.57. En cada iteración de PGCNR se requiere computar dos productos matriz-vector $A\widetilde{d}_n$, $\widehat{r}_{n+1} = A^\top r_{n+1}$ y resolver $M\widetilde{r}_{n+1} = \widehat{r}_{n+1}$.

(iii) 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 $\mathcal{K}_n\left(M^{-1}A, M^{-1}b\right)$, se trata de hallar el vector $x_n \in \mathcal{K}_n\left(M^{-1}A, M^{-1}b\right)$ tal que

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

Se obtiene el siguiente algoritmo:

Algoritmo 7 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.

1: Definimos los vectores $x_0 := 0$, $\widetilde{b} := M^{-1}b$, $v_1 = \frac{\widetilde{b}}{\|\widetilde{b}\|_2}$ y $c = \|\widetilde{b}\|_2 e_1 \in \mathbb{R}^N$.
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 $h_{n+1,n} = \|\widehat{v}_{n+1}\|_2$.
3: Continuar con los pasos 3-5 del Algoritmo 4.

Observación 2.58. Observemos que cada iteración de PGMRES requiere resolver el sistema $M\widetilde{v}_n = Av_n$.

3

Ilustración Numérica

3.1. Aplicación a la discretización espacial de EDPs

Consideramos la comparación de los métodos numéricos presentados en los capítulos previos aplicados a sistemas lineales resultantes de la discretización espacial de la ecuación en derivadas parciales lineal con coeficientes constantes de tipo difusión-convección-reacción

$$-d \cdot \Delta u + a\,(\nabla \cdot \mathbf{1})\,u + ru = f \ , \text{ en } \ (0,1)^n \tag{3.1}$$

con condiciones de frontera de tipo Dirichlet, siendo $\Delta = \sum_{j=1}^{n} \frac{\partial^2}{\partial x_j^2}$ y $\nabla \cdot \mathbf{1} = \sum_{j=1}^{n} \frac{\partial}{\partial x_j}$. En (3.1), d, a y r son constantes, y tomaremos fundamentalmente $n = 2, 3$. Para la discretización espacial de (3.1) consideramos diferencias finitas centrales de segundo orden tanto para $u_{x_j x_j}$ como para u_{x_j}, $1 \leq j \leq n$.

Observamos en primer lugar que para el caso $n = 1$, con $u = u(x)$ suficientemente regular, estas aproximaciones en diferencias centrales verifican para $h > 0$

$$\begin{aligned} u''(x) &= \frac{u(x+h) - 2u(x) + u(x-h)}{h^2} - \frac{1}{12}u^{(iv)}(\xi_1)h^2 \\ u'(x) &= \frac{u(x+h) - u(x-h)}{2h} - \frac{1}{6}u'''(\xi_2)h^2 \end{aligned} \qquad , \xi_1, \xi_2 \in [x-h, x+h]$$

y, en particular, los cocientes en diferencias dan aproximaciones exactas a las derivadas correspondientes si $u(x)$ es un polinomio de grado menor o igual que dos. Considerando $n = 1$ en (3.1) y una partición uniforme de $N+2$ nodos equiespaciados en $[0,1]$ a distancia $h > 0$, $x^{(i)} = ih$, con $0 \leq i \leq N+1$ y $h = \frac{1}{N+1}$, la discretización mediante diferencias centrales de orden dos conduce a un sistema lineal $A^{(1)}U = b$ donde $U = (U_i)_{i=1}^N$ con $U_i \approx u\left(x^{(i)}\right)$, $1 \leq i \leq N$, y $b = (b_i)_{i=1}^N$ contiene la discretización del dato f así como los valores de frontera de $u(x)$. En concreto, $b_i = h^2 f(x^{(i)})$ para $2 \leq i \leq N$, $b_1 = h^2 f(x^{(1)}) + (d + a\frac{h}{2})u(0)$, $b_N = h^2 f(x^{(N)}) + (d - a\frac{h}{2})u(1)$ y

$$A^{(1)} := d \cdot Tridiag(-1, 2, -1) + a\frac{h}{2}Tridiag(-1, 0, 1) + rh^2 I_N \in \mathbb{R}^{N \times N}.$$

Observación 3.1. Considerando la descomposición $A^{(1)} = D^{(1)} + L^{(1)} + R^{(1)}$ tenemos para la matriz de iteración del método de Jacobi

$$H_{\mathcal{J}}^{(1)} = -\left(D^{(1)}\right)^{-1}\left(L^{(1)} + R^{(1)}\right) = \frac{1}{2d + rh^2}Tridiag\left(d + a\frac{h}{2}, 0, d - a\frac{h}{2}\right)$$

cuyos autovalores son * $\frac{2}{2d+rh^2}\sqrt{d^2-\left(\frac{ah}{2}\right)^2}\cos\left(\frac{j\pi}{N+1}\right)$, $1\le j\le N$. Por tanto,

$$\rho_{\mathcal{J}}=\rho\left(\mathcal{H}_J^{(1)}\right)=\frac{2}{|2d+rh^2|}\sqrt{\left|d^2-\left(\frac{ah}{2}\right)^2\right|}\cos\left(\frac{\pi}{N+1}\right).$$

Tomaremos este valor ρ_J a efectos de elegir el parámetro óptimo w^* en el método de relajación SOR.

Análogamente, la discretización de (3.1) en el caso $n=2$ en una partición de nodos $x^{(i)}=ih$, $y^{(j)}=jh$, $0\le i,j\le N+1$ y $h=\frac{1}{N+1}$, conduce al siguiente sistema lineal de dimensión N^2

$$\begin{aligned}&d\left[(-U_{i+1,j}+2U_{ij}-U_{i-1,j})+(-U_{i,j+1}+2U_{ij}-U_{i,j-1})\right]\\&\quad+\frac{ah}{2}\left[(U_{i+1,j}-U_{i-1,j})+(U_{i,j+1}-U_{i,j-1})\right]+rh^2U_{ij}=b_{ij}\text{ para }1\le i,j\le N,\end{aligned}\tag{3.2}$$

donde $U_{ij}\approx u\left(x^{(i)},y^{(j)}\right)$, $1\le i,j\le N$, y b_{ij} contiene la discretización del dato f, así como los valores de frontera de $u(x,y)$ cuando $i,j\in\{1,N\}$. Considerando los vectores

$$\begin{aligned}U&=(U_{11},U_{21},...,U_{N1},U_{12},U_{22},...,U_{N2},...,U_{1N},U_{2N},...,U_{NN})^\top\in\mathbb{R}^{N^2},\\b&=(b_{11},b_{21},...,b_{N1},b_{12},b_{22},...,b_{N2},...,b_{1N},b_{2N},...,b_{NN})^\top\in\mathbb{R}^{N^2},\end{aligned}$$

el sistema (3.2) se puede escribir en forma compacta como $A^{(2)}U=b$, siendo

$$A^{(2)}=d\left(I_N\otimes T_2+T_2\otimes I_N\right)+\frac{ah}{2}\left(I_N\otimes T_1+T_1\otimes I_N\right)+rh^2I_{N^2}$$

donde $T_2=Tridiagonal(-1,2,-1)$, $T_1=Tridiagonal(-1,0,1)$ y $\otimes$ denota el producto de Kronecker de matrices definido como

$$\mathcal{A}\otimes\mathcal{C}=(a_{ij}\mathcal{C})_{\substack{i=1,\dots,p\\j=1,\dots,q}}\in\mathbb{R}^{(pr)\times(qs)}\text{ con }\mathcal{A}=(a_{ij})\in\mathbb{R}^{p\times q},\ \mathcal{C}\in\mathbb{R}^{r\times s}.$$

Observación 3.2. Sean las matrices $A\in\mathbb{R}^{N\times N}$ y $B\in\mathbb{R}^{M\times M}$. Entonces los autovalores de la *suma de Kronecker* $A\otimes I_M+I_N\otimes B$ son $\lambda_i+\mu_j$ con $\lambda_i\in\sigma[A]$ y $\mu_j\in\sigma[B]$ para $1\le i\le N$ y $1\le j\le M$ (ver, p.e., [1, p.146]).

Observamos que

$$A^{(2)}=I_N\otimes\left[dT_2+\frac{ah}{2}T_1+\frac{r}{2}h^2I_N\right]+\left[dT_2+\frac{ah}{2}T_1+\frac{r}{2}h^2I_N\right]\otimes I_N=D^{(2)}+L^{(2)}+R^{(2)}$$

con $D^{(2)}=\left(4d+rh^2\right)I_{N^2}$ y $L^{(2)}+R^{(2)}=I_N\otimes\left[L^{(1)}+R^{(1)}\right]+\left[L^{(1)}+R^{(1)}\right]\otimes I_N$. Por tanto, para $\mathcal{H}_J^{(2)}=-\left(D^{(2)}\right)^{-1}\left(L^{(2)}+R^{(2)}\right)$ se tiene que

$$\sigma\left[\mathcal{H}_J^{(2)}\right]=\left\{\frac{2}{4d+rh^2}\sqrt{d^2-\left(\frac{ah}{2}\right)^2}\left(\cos\left(\frac{i\pi}{N+1}\right)+\cos\left(\frac{j\pi}{N+1}\right)\right)\Big/1\le i,j\le N\right\}.$$

En definitiva

$$\rho_J=\rho(\mathcal{H}_J^{(2)})=\frac{4}{|4d+rh^2|}\sqrt{\left|d^2-\left(\frac{ah}{2}\right)^2\right|}\cos\left(\frac{\pi}{N+1}\right).$$

* Los autovalores de una matriz tridiagonal Toeplitz $Tridiag(b,a,c)\in\mathbb{R}^{N\times N}$ vienen dados por $\lambda_j=a+2\sqrt{bc}\cos\left(\frac{j\pi}{N+1}\right)$ para $1\le j\le N$ (veáse, p.e., [1, p.349]).

En general, la discretización espacial del problema (3.1) en n dimensiones mediante diferencias centrales de segundo orden sobre una malla uniforme de $[0,1]^n$

$$x_j^{(i)} = ih,\ h = \frac{1}{N+1} \text{ para } 0 \le i \le N+1,\ 1 \le j \le n, \tag{3.3}$$

conduce a un sistema lineal

$$A^{(n)}U = b \tag{3.4}$$

de dimensión N^n donde $U \in \mathbb{R}^{N^n}$, con componentes ordenadas de acuerdo al orden lexicográfico, es un vector de aproximaciones a $u(x_1, x_2, ..., x_n)$ en los respectivos puntos interiores de la malla, $b \in \mathbb{R}^{N^n}$ contiene la discretización del dato f así como valores de frontera de la solución exacta y

$$A^{(n)} = d\sum_{i=1}^{n} \overbrace{I_N \otimes ... \otimes \underbrace{T_2}_{(i)} \otimes ... \otimes I_N}^{(n)} + \frac{ah}{2}\sum_{i=1}^{n} \overbrace{I_N \otimes ... \otimes \underbrace{T_1}_{(i)} \otimes ... \otimes I_N}^{(n)} + rh^2 I_{N^n} \in \mathbb{R}^{N^n \times N^n}.$$

Nuevamente, tenemos que

$$A^{(n)} = \sum_{i=1}^{n} I_N \otimes ... \otimes \underbrace{\left[dT_2 + \frac{ah}{2}T_1 + \frac{rh^2}{n}I_N\right]}_{(i)} \otimes ... \otimes I_N = D^{(n)} + L^{(n)} + R^{(n)},$$

con $D^{(n)} = (2nd + rh^2)I_{N^n}$ y $L^{(n)} + R^{(n)} = \sum_{i=1}^{n} I_N \otimes ... \otimes \left[L^{(1)} + R^{(1)}\right] \otimes ... \otimes I_N$. Así pues, para $\mathcal{H}_J^{(n)} = -\left(D^{(n)}\right)^{-1}\left(L^{(n)} + R^{(n)}\right)$, teniendo en cuenta los autovalores de una suma de Kronecker, sigue que

$$\begin{aligned} \rho_J = \rho(\mathcal{H}_J^{(n)}) &= \frac{1}{|2nd + rh^2|}\left(2\sqrt{\left|d^2 - \left(\tfrac{ah}{2}\right)^2\right|}\right)\left(n\cos\left(\tfrac{\pi}{N+1}\right)\right) \\ &= \frac{2n}{|2nd + rh^2|}\sqrt{\left|d^2 - \left(\tfrac{ah}{2}\right)^2\right|}\cos\left(\tfrac{\pi}{N+1}\right). \end{aligned} \tag{3.5}$$

Observación 3.3. En los ejemplos que presentamos a continuación consideramos que la solución $u(x_1, ..., x_n)$ de la EDP (3.1) es un polinomio de grado menor o igual que dos en cada variable de modo que la discretización espacial sea exacta. Esto implica que la restricción de u a la malla (3.3) es precisamente la solución del sistema lineal $A^{(n)}U = b$. A partir de la solución exacta u, el dato f se toma como $f = -d \cdot \Delta u + a\,(\nabla \cdot \mathbf{1})\,u + ru$.

Por otra parte, a partir de la solución exacta y una vez generada la matriz $A^{(n)}$ y el vector U, hemos generado en Matlab el vector b, que contiene la discretización espacial del dato f y las condiciones de frontera, simplemente como $b = A^{(n)}U$. El vector U se usará como solución de referencia para estudiar la precisión de las aproximaciones numéricas provistas por los métodos presentados en los dos primeros capítulos. Para todos los métodos iterativos se ha usado como valor inicial $\widehat{U}^{(0)} = 0$. Los tiempos de CPU referidos en las gráficas siguientes corresponden a un procesador Intel Core i5 a 3.3GHz y 8Gb de memoria RAM.

Ejemplo 3.4. La ecuación de Poisson con condiciones de frontera homogéneas de tipo Dirichlet en $[0,1]^n$:

$$\begin{cases} -\Delta u = f, & \text{en } \Omega = (0,1)^n, \\ u = 0, & \text{en } \Gamma = \partial\Omega, \end{cases} \tag{3.6}$$

con solución exacta $u(x_1, ..., x_n) = 4^n \prod_{j=1}^{n} x_i(1 - x_i)$. Este problema se obtiene de (3.1) tomando $d = 1$ y $a = r = 0$.

Para las dimensiones $n = 2, 3$ consideramos una malla (3.3) con $N = 2^j$ para $1 \le j \le 11$, si $n = 2$, y $N = 10 \cdot j$ para $1 \le j \le 16$, si $n = 3$. Para resolver el sistema lineal $A^{(n)}U = b$ de dimensión N^n consideramos en primer lugar tres métodos directos:

1. Eliminación Gaussiana: $A^{(n)} = L_o U_p$, donde en Matlab se obtiene esta descomposición LU mediante $[L_o, U_p] = lu(A^{(n)})$. La solución numérica correspondiente será
$$U_G = U_p \setminus (L_o \setminus b)\,.$$
2. Factorización de Cholesky: $A^{(n)} = L_C L_C^\top$ obtenida en Matlab mediante $L_C = chol(A^{(n)})$ y
$$U_C = \left(L_C'\right) \setminus (L_C \setminus b)\,.$$
3. Operador Backslash (\): resolvemos directamente con $U_b = A^{(n)} \setminus b$.

Seguidamente aplicamos los métodos de sobrerelajación sucesiva de Gauss-Seidel (SOR) y Gradiente Conjugado (GC) iterados con un máximo de $5 \cdot 10^4$ iteraciones hasta garantizar un error
$$\|U - \widehat{U}\|_\infty \leq errtol,$$
siendo $\widehat{U}$ la solución numérica del método correspondiente y
$$errtol = min\left\{\|U - U_G\|_\infty, \|U - U_C\|_\infty, \|U - U_b\|_\infty, 5 \cdot 10^{-13}\right\}.$$
Para el método SOR se ha elegido el parámetro óptimo $w^* = \frac{2}{1+\sqrt{1-\rho_J^2}}$ con $\rho_J = \cos\left(\frac{\pi}{N+1}\right)$ dado por (3.5).

Respecto al caso bidimensional $n = 2$, observamos en la Figura 3.1 que el comando Backslash (\) de Matlab proporciona la aproximación numérica más eficiente si $N \leq 2^{10}$. No obstante, este comando no pudo realizar la resolución numérica para $N = 2^{11}$ por insuficiencia de memoria de almacenamiento. Lo mismo ocurre con las rutinas internas lu y $chol$ para $N \geq 2^{10}$.

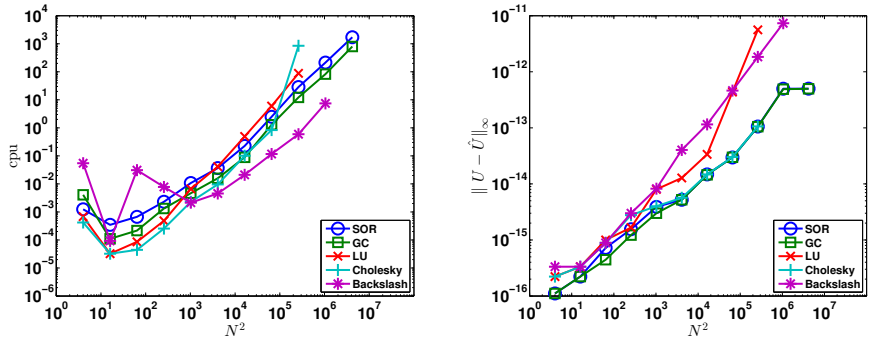


Figura 3.1. Comparación en tiempos cpu (izquierda) y error (derecha) $\|U - \widehat{U}\|_\infty$ respecto a la dimensión de la discretización espacial en el ejemplo (3.6) con $n = 2$ para los métodos directos LU, Cholesky y Backslash y los métodos iterativos SOR y GC.

En lo que respecta al caso tridimensional $n = 3$ en la Figura 3.2, lu y $chol$ dejaron de funcionar para $N \geq 50$, mientras que el comando Backslash lo hizo a partir de $N = 90$. Por otra parte, observamos que GC resulta ligeramente más eficiente que SOR.

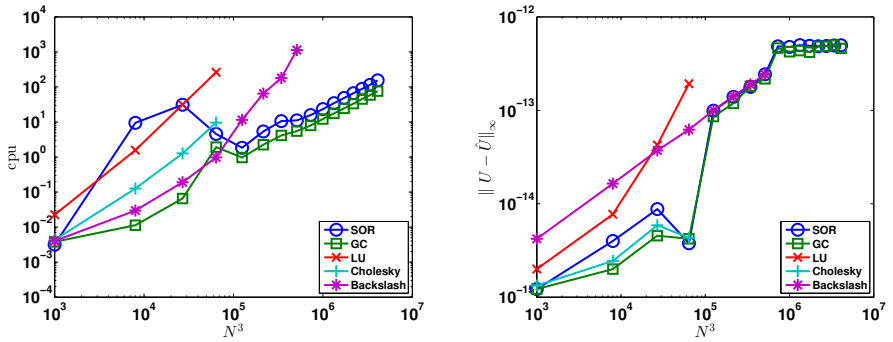


Figura 3.2. Comparación en tiempos cpu (izquierda) y error (derecha) $\|U - \widehat{U}\|_\infty$ respecto a la dimensión de la discretización espacial en el ejemplo (3.6) con $n = 3$ para los métodos directos LU, Cholesky y Backslash y los métodos iterativos SOR y GC.

Notamos respecto a estas dos figuras que GC y SOR se han iterado hasta garantizar un error al menos menor o igual al mínimo que proveen los tres métodos directos.

Ejemplo 3.5. La ecuación de Poisson con condiciones de frontera no homogéneas de tipo Dirichlet en $[0,1]^3$ con solución exacta $u(x,y,z) = x^2 + y^2 + z^2$(ver Figura 3.3):

$$\begin{cases} -\Delta u = -6, & (x,y,z) \in \Omega = (0,1)^3 \\ u(x,y,z) = x^2 + y^2 + z^2, & (x,y,z) \in \Gamma = \partial\Omega \end{cases} \tag{3.7}$$

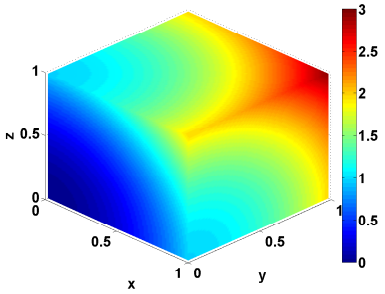


Figura 3.3. Solución exacta de (3.7).

Consideramos la malla (3.3) con $N = 100$ y el sistema lineal $A^{(3)}U = b$ (3.4) de dimensión $N^3 = 10^6$. Comparamos, en primer lugar, los errores y tiempos de CPU cuando se dan Niter $= 200$ iteraciones con los métodos SOR (con parámetro óptimo $w^* = \cos\left(\frac{\pi}{N+1}\right)$), GC y GMRES($m$), para $m = 10, 50, 100$ y 200. Observamos en la Figura 3.4 que el método SOR proporciona mejores aproximaciones respecto tanto al número de iteraciones como al tiempo de CPU, aunque la precisión alcanzada es bastante exigua.

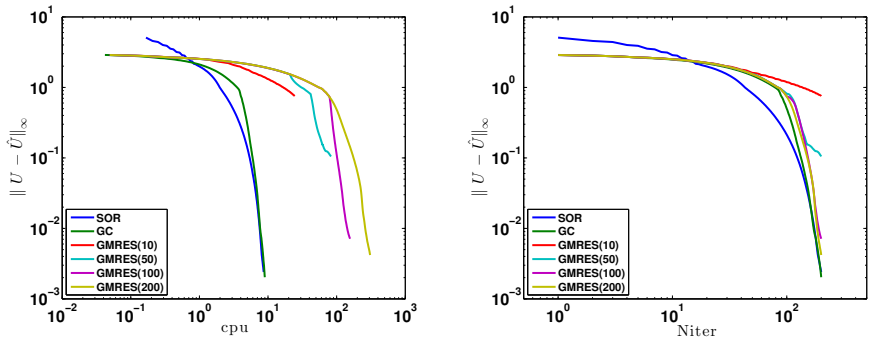


Figura 3.4. Comportamiento del error en función del tiempo cpu (izquierda) y número de iteraciones (derecha) para diversos métodos iterativos aplicadosa la discretización espacial de (3.7).

El comportamiento de GC y GMRES(m) se puede mejorar considerando sus variantes con precondicionamiento PGC y PGMRES(m). Para ello, consideramos en el caso del operador *Laplaciano* el precondicionador $M = S(w^*)$ con

$$S(w) = \frac{w}{2-w}\left(\frac{1}{w}D^{(3)} + L^{(3)}\right)\left(D^{(3)}\right)^{-1}\left(\frac{1}{w}D^{(3)} + L^{(3)\top}\right) \tag{3.8}$$

siendo $A^{(3)} = D^{(3)} + L^{(3)} + R^{(3)}$. La matriz $S(w)^{-1}$ se relaciona con la matriz de iteración del método SOR simétrico SSOR (ver, p.e., [3, p.620]). Con la elección del precondicionador SSOR se logra mitigar el condicionamiento del sistema lineal (3.4). En concreto, $\frac{|\lambda_{\text{máx}}(M^{-1}A^{(3)})|}{|\lambda_{\text{mín}}(M^{-1}A^{(3)})|} = \mathcal{O}(h^{-1}), h \to 0$, mientras que $\frac{|\lambda_{\text{máx}}(A^{(3)})|}{|\lambda_{\text{mín}}(A^{(3)})|} = \mathcal{O}(h^{-2}), h \to 0$ (véase, p.e., [6, p.55]). Observamos ahora en la Figura 3.5 como PGC y PGMRES(m), $m = 10, 50$, convergen a la solución exacta en menos de 100 iteraciones. Además, PGC resulta más eficiente en términos de CPU que PGMRES(m) al tratarse $A^{(3)}$ de una matriz simétrica.

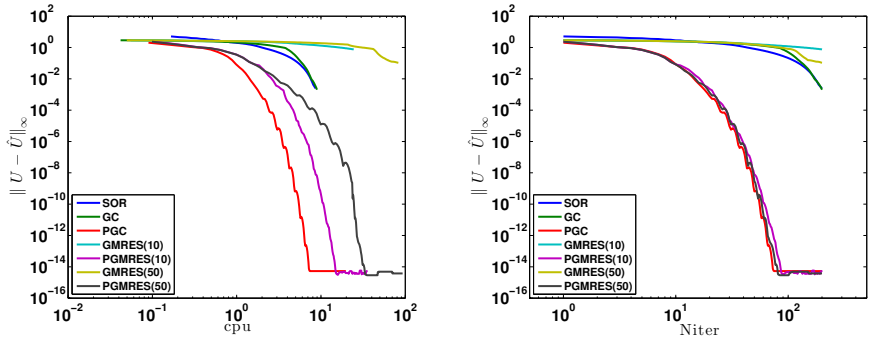


Figura 3.5. Comportamiento del error en función del tiempo cpu (izquierda) y número de iteraciones (derecha) para diversos métodos iterativos precondicionados aplicados a la discretización espacial de (3.7).

Ejemplo 3.6. El modelo de difusión-convección-reacción (3.1) con convección dominante $d = 1$ y $r = -3a$, $a = 10$ o 100, con solución exacta $u(x, y, z) = 1$ en $[0, 1]^3$.

Consideramos nuevamente la malla (3.3) con $N = 100$ y el correspondiente sistema lineal $A^{(3)}U = b$ (3.4) de dimensión $N^3 = 10^6$.

Observamos que los elementos no nulos en la parte triangular inferior de $A^{(3)}$ vienen dados por $-d - a\frac{h}{2}$, mientras que los elementos no nulos en la parte triangular superior son $-d + a\frac{h}{2}$. Luego, la diferencia entre ambos, ah, da una medida de la asimetría de la matriz A. Para $a = 10$: $-d - a\frac{h}{2} \doteq -1'0495$, $-d + a\frac{h}{2} \doteq -0'9505$, $\rho_J \doteq 0'9988$, $w^* \doteq 1'9059$. Para $a = 100$: $-d - a\frac{h}{2} \doteq -1'4951$, $-d + a\frac{h}{2} \doteq -0'5050$, $\rho_J \doteq 0'8827$, $w^* \doteq 1'3439$.

Para ambos valores de a aplicamos en primer lugar los métodos SOR(w^*), GC, GCNR y GMRES(m), $m = 10, 50, 100, 200$, dando en cada caso 200 iteraciones. Observamos que el método GC no tiene porqué converger (ni siquiera estar bien definido) pues A no es simétrica.

En la Figura 3.6 (con $a = 10$ en el bloque superior y $a = 100$ en el inferior) observamos la divergencia de GC, la lenta convergencia de GCNR y GMRES(m), mientras que se aprecia convergencia para el método SOR, siendo ésta más rápida en el caso $a = 100$ por cuanto el radio espectral de la matriz de iteración correspondiente es menor ($\rho_{w^*} \doteq 0'3439 = w^* - 1$).

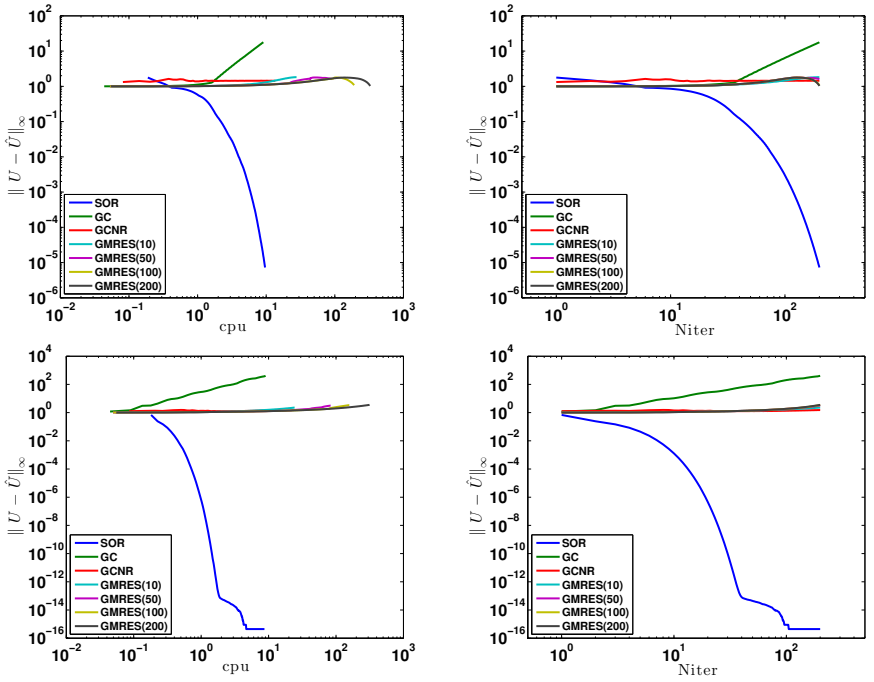


Figura 3.6. Comportamiento del error en función del tiempo cpu (izquierda) y número de iteraciones (derecha) para diversos métodos iterativos aplicados a la discretización espacial de (3.1), con $d = 1$ y $r = -3a$, para $a = 10$ (bloque superior) y $a = 100$ (bloque inferior).

Por otra parte, consideramos también las variantes con precondicionamiento PGC, PGCNR y PGMRES(m). Nuevamente se observa divergencia para PGC en la Figura 3.7, así como una aceleración de la convergencia para PGCNR y PGMRES(m) con respecto a los correspondientes métodos sin precondicionamiento.

Debemos indicar que para PGC y PGMRES(m) hemos considerado precondicionamiento SSOR, tal como en el Ejemplo 3.5, esto es $M = S(w^*)$ dada por (3.8).

Respecto al método PGCNR, puesto que GCNR se obtiene aplicando GC a las ecuaciones normales $(A^{(3)})^\top A^{(3)}U = (A^{(3)})^\top b$, hemos considerado el precondicionador $M = S(w^*)^\top S(w^*)$.

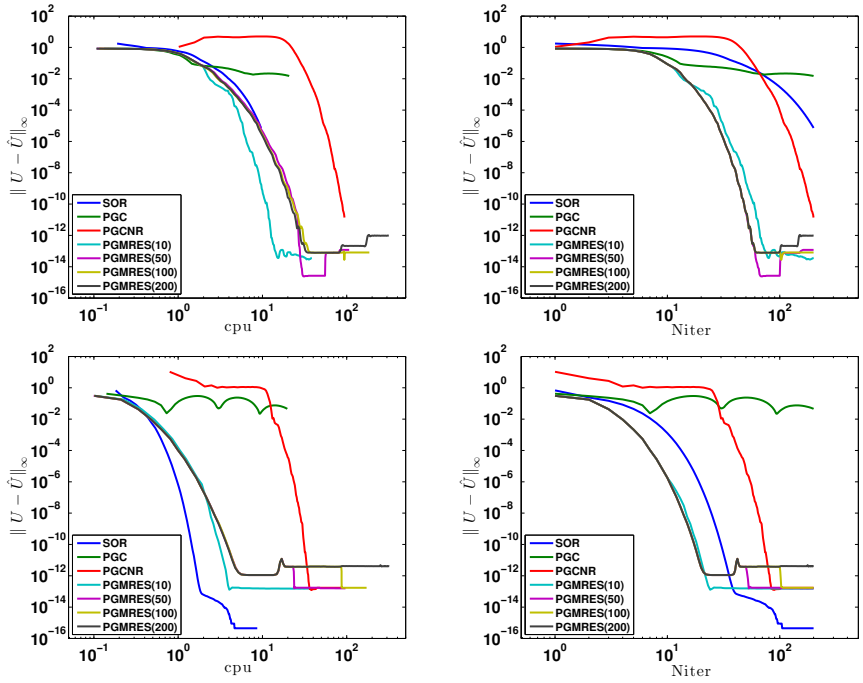


Figura 3.7. Comportamiento del error en función del tiempo cpu (izquierda) y número de iteraciones (derecha) para diversos métodos iterativos precondicionados aplicados a la discretización espacial de (3.1), con $d = 1$ y $r = -3a$, para $a = 10$ (bloque superior) y $a = 100$ (bloque inferior).

En el caso $a = 10$, la asimetría de la matriz $A^{(3)}$ es menos significativa y se observa en la Figura 3.7 (bloque superior) que los métodos precondicionados proveen aproximaciones numéricas más precisas que SOR. No obstante, PGCNR es menos eficiente que PGMRES(m), pues requiere en cada iteración un producto adicional por $(A^{(3)})^\top$ así como el doble de gasto computacional por aplicación del precondicionador $S(w^*)^\top S(w^*)$.

En el caso $a = 100$, por una parte la matriz $A^{(3)}$ presenta una asimetría más destacada, mientras que la matriz de iteración de SOR tiene un radio espectral menor. En este caso, se observa en la Figura 3.7 (bloque inferior) que el método SOR es más eficiente que PGMRES(m) y que PGCNR. No obstante, se puede mejorar la eficiencia de PGMRES(m) y PGCNR considerando precondicionadores más generales basados en descomposiciones LU y de Cholesky incompletas [7, p.301].

Observación 3.7. Para la implementación de los métodos considerados en este capítulo hemos empleado los códigos que aparecen en [2] y que están disponibles en https://es.mathworks.com/matlabcentral/fileexchange/2158-templates-for-the-solution-of-linear-systems.

3.2. Conclusiones

Los métodos directos para resolver sistemas lineales $Ax = b$, tales como eliminación gaussiana o descomposición de Cholesky, resultan eficientes cuando la dimensión N del sistema lineal es moderada. Sin embargo, para grandes sistemas lineales, como los que resultan en la discretización espacial de Ecuaciones en Derivadas Parciales en varias dimensiones espaciales, estos métodos pierden eficiencia, o directamente son inviables, pues requieren un elevado costo computacional así como una gran disponibilidad de memoria de almacenamiento. Observemos que aunque la matriz A sea sparse, tanto el cálculo de su inversa como su descomposición LU requiere almacenar matrices llenas, siendo éstas de tamaño N^2.

En el capítulo 1 de este trabajo hemos comenzado con una revisión de resultados para los métodos iterativos clásicos de Jacobi y de Gauss-Seidel. Hemos extendido el análisis estándar de estos métodos para matrices estrictamente diagonal dominantes al caso de matrices irreduciblemente diagonal dominantes. Esta última propiedad es menos restrictiva y tiene mayor alcance práctico, en particular en la discretización espacial de Ecuaciones en Derivadas Parciales.

También en el capítulo 1 hemos considerado el método de relajación asociado al método de Gauss-Seidel (o método SOR de sobrerelajación sucesiva) estudiando su convergencia en función del parámetro de relajación ω. En particular, para las denominadas matrices consistentemente ordenadas hemos establecido una expresión cerrada para el radio espectral de la matriz de iteración del método SOR que permite seleccionar el parámetro de relajación de modo óptimo. Asimismo, hemos observado que las matrices consistentemente ordenadas surgen de modo habitual en la discretización espacial de Ecuaciones en Derivadas Parciales. En el capítulo 3, observamos que el método SOR es bastante eficiente cuando se conoce de antemano el valor óptimo del parámetro de relación ω. No obstante, en problemas más prácticos, la cuestión de seleccionar este parámetro óptimo no es simple y resulta crucial de cara a la eficiencia del método.

En el capítulo 2 hemos introducido los denominados subespacios de Krylov con el fin de presentar otra clase de métodos iterativos, en particular los métodos del Gradiente Conjugado (GC) y del Residual Mínimo Generalizado (GMRES). En aritmética infinita, estos métodos convergen a la solución del sistema lineal en un número finito de iteraciones menor o igual que la dimensión N del sistema, aunque este número puede ser considerablemente grande en la práctica. El método GC está especialmente indicado para matrices simétricas y definidas positivas, aunque también hemos presentado su variante GCNR para el caso general tratado a través del sistema de ecuaciones normales $A^TAx = A^Tb$. El método GMRES es de propósito general para matrices regulares arbitrarias, siendo su costo computacional ligeramente superior al de GC e inferior al de GCNR. En particular, el algoritmo de Arnoldi para generar bases ortogonales de los subespacios de Krylov, junto con las rotaciones de Givens para resolver problemas de mínimos cuadrados, permiten una implementación bastante simple del método GMRES.

Además en el capítulo 2 hemos establecido resultados de convergencia para los métodos GC, GCNR y GMRES que reflejan que su tasa de convergencia puede verse ralentizada por el número de condición de la matriz A (o de cierta matriz relacionada con A), y en la sección 3.1 de este capítulo hemos ilustrado que, en efecto, ésto es lo que ocurre habitualmente en la práctica. Por ello, para finalizar el capítulo 2 hemos presentado variantes de estos métodos iterativos considerando precondicionamiento (a la izquierda) del sistema lineal. Las variantes con precondicionamiento PGC, PGCNR y PGMRES permiten mitigar el mal condicionamiento de los sistemas lineales y mejorar la eficiencia respecto a los métodos originales sin precondicionamiento. En la sección 3.1 hemos observado el efecto positivo del precondicionamiento en la aplicación de los métodos GC, GCNR y GMRES. No obstante, este efecto positivo depende en gran medida de la selección de un precondicionador adecuado. En esta sección hemos considerado el precondicionador SSOR (SOR simétrico) por ser muy eficiente en la discretización de la ecuación de Laplace. El estudio de la selección eficiente de precondicionadores en problemas más generales queda fuera del alcance y objetivos de este trabajo.

Bibliografía

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

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

[3] Golub, G., Van Loan, C.F. *Matrix computations.* The Johns Hopkins University press, 4rd ed. (2013).

[4] Hämmerlin, G., Hoffmann, K.H. *Numerical mathematics.* Springer-Verlag (1991).

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

[6] Quarteroni, A., Valli, A. *Numerical approximation of partial differential equations.* Springer (1994).

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

Iterative methods for systems of linear equations

Ana Gómez Pérez

Facultad de Ciencias · Sección de Matemáticas

Universidad de La Laguna

alu0100818987@ull.edu.es

Abstract

In this project several iterative methods are studied to solve linear systems of equations. First, a brief introduction about the direct methods are included. Next, the classic iterative methods of Jacobi and Gauss-Seidel are considered. A convergence analysis focusing on strictly diagonally dominant and irreducibly diagonally dominant matrices is presented. Likewise, the relaxation method is considered by studying its convergence as a function of the relaxation parameter, specifically, for consistently ordered matrices. Also, other classes of iterative methods based on Krylov subspaces are introduced: the Conjugate Gradient method, for symmetric and positive definite matrices, and the generalized minimal residual method for any nonsingular matrix. Finally, some examples are illustrated comparing the different methods according to their CPU time, error and number of iterations for linear systems arising from the spatial discretization of diffusion-convection-reaction partial differential equations with constant coefficients.

1. Introduction

In this work we will study different numerical methods to solve large linear systems of equations $Ax = b$ with $A \in \mathbb{R}^{N\times N}$ nonsingular matrix and $b \in \mathbb{R}^N$, with the unique solution $x_* = A^{-1}b \in \mathbb{R}^N$. The most classic direct method is **Gaussian Elimination** that amounts to decompose the matrix A as LU, leading to the solution of the triangular systems $Ly = b$ and $Ux = y$.
On the other hand, if A is positive definite and symmetric, then A admits the **Cholesky descomposition** $A = LL^\top$.
These direct methods use $\mathcal{O}(N^3)$ arithmetic operations, so that they are expensive or non-viable when N is large.

2. Basic iterative methods

An approximation to x_* is obtained by considering a fixed point iteration

$$x^{(n+1)} = \mathscr{H}x^{(n)} + z, \qquad n \geq 0 \qquad (1)$$

with iteration matrix $\mathscr{H} \in \mathbb{R}^{N\times N}$.

Theorem 1 *The iterative method (1) is convergent if and only if* $\rho(\mathscr{H}) < 1$.

By splitting $A = D + L + R$ where D is diagonal, L lower triangular and R upper triangular, the **Jacobi method** is

$$x^{(n+1)} = D^{-1}b - D^{-1}(L+R)x^{(n)} = \mathscr{H}_J x^{(n)} + z.$$

The **Gauss-Seidel method** is

$$x^{(n+1)} = (D+L)^{-1}b - (D+L)^{-1}Rx^{(n)} = \mathscr{H}_{GS}x^{(n)} + z.$$

Theorem 2 *If A is an irreducibly diagonally dominant matrix, then Jacobi and Gauss-Seidel methods are convergent.*

The **relaxation method** is defined as

$$x^{(n+1)} = (D+wL)^{-1}\left[wb + ((1-w)D - wR)\,x^{(n)}\right]$$

with $\mathscr{H}(w) = (D+wL)^{-1}((1-w)D - wR)$.

Theorem 3 *If A symmetric and definite, then the relaxation method is convergent for* $0 < w < 2$.

3. Krylov iterative methods

Krylov subspaces are defined for $n \geq 0$ as

$$\mathscr{K}_n = \mathscr{K}_n(A,b) = span\{b, Ab, \ldots, A^{n-1}b\}.$$

Corollary 4 *There exists a unique* $0 \leq n_* \leq N$ *such that* $\{0\} = \mathscr{K}_0 \subsetneq \mathscr{K}_1 \subsetneq \ldots \subsetneq \mathscr{K}_{n_*-1} \subsetneq \mathscr{K}_{n_*} = \mathscr{K}_{n_*+1}$. *Moreover,* $x_* \in \mathscr{K}_{n_*} \setminus \mathscr{K}_{n_*-1}$.

The **Conjugate Gradient method (GC)**, for symmetric and positive definite matrices, is described by

$$\left.\begin{array}{l} x_n \in \mathscr{K}_n(A,b) \\ \|x_n - x_*\|_A = \min\limits_{x\in\mathscr{K}_n(A,b)} \|x - x_*\|_A \end{array}\right\} \quad n = 1,2,\ldots$$

x_n is the best approximation to x_* in $\mathscr{K}_n$ respect to $\langle\cdot,\cdot\rangle_A$.

Algorithm 1 The GC method
Let $r_j := Ax_j - b$, $j \geq 0$

1: Define $x_0 := 0$, $r_0 := -b$ y $d_0 := -r_0 = b$.
2: Given x_n, r_n y d_n, $n \geq 0$, while $r_n \neq 0$ calculate
$x_{n+1} = x_n + \alpha_n d_n$, with $\alpha_n = \frac{\|r_n\|_2^2}{(Ad_n, d_n)_2}$
$r_{n+1} = r_n + \alpha_n A d_n$
$d_{n+1} = -r_{n+1} + \beta_n d_n$, with $\beta_n = \frac{\|r_{n+1}\|_2^2}{\|r_n\|_2^2}$.

The **GC method for the normal equations (GCNR)** consists in applying the GC method to the system $A^\top Ax = A^\top b$ where A is any non-singular matrix.
The **Generalized minimal residual method (GMRES)**, for any $A \in \mathbb{R}^{N\times N}$ nonsingular matrix, is described by

$$\left.\begin{array}{l} x_n \in \mathscr{K}_n(A,b) \\ \|Ax_n - b\|_2 = \min\limits_{x\in\mathscr{K}_n(A,b)} \|Ax - b\|_2 \end{array}\right\} \quad n = 1,2,\ldots$$

Using the Arnoldi process we generate an orthogonal basis of $\mathscr{K}_n(A,b)$. The method is reduced to a least squares problem which is solved with QR decomposition and Givens rotations.

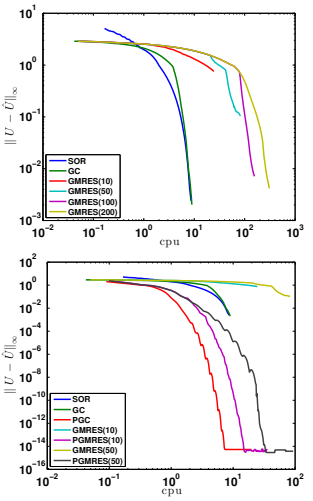
If the condition number of A is large, the convergence might be slow. So, we precondition the linear system $M^{-1}Ax = M^{-1}b$ with M non-singular *close to* A which allows to speed up the convergence.

4. Numerical illustration

We consider the Poisson equation on $[0,1]^3$ with nonhomogeneous boundary conditions according to the exact solution $u(x,y,z) = x^2 + y^2 + z^2$:

$$\begin{cases} -\Delta u = -6, & (x,y,z) \in \Omega = (0,1)^3 \\ u(x,y,z) = x^2 + y^2 + z^2, & (x,y,z) \in \Gamma = \partial\Omega \end{cases}$$

With a grid of 100 interior points along each spatial direction we obtain a linear system $A^{(3)}U = b$ of dimension $N = 10^6$. In the plots below the efficiency of some iterative methods is compared considering the error versus CPU time.



To enhance the performance of Krylov methods we use preconditioning.

References

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

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