Universidad del Bío-Bío
Facultad de Educación y Humanidades
Escuela de Pedagogía en Educación Matemática

MÉTODOS ITERATIVOS PARA SISTEMAS DE ECUACIONES LINEALES

Autor: Darío Andrés Burgos Alarcón

Profesor Guía: Roberto Carlos Cabrales

MEMORIA PARA OPTAR AL TÍTULO DE PROFESOR DE EDUCACIÓN MEDIA EN EDUCACIÓN MATEMÁTICA

CHILLÁN, ABRIL 2016

Índice general

Agradecimientos III

Introducción VI

1. Preliminares **1**
1.1. Vectores y Matrices 1
1.1.1. Tipos especiales de matrices 6
1.2. Sistemas de Ecuaciones 8
1.3. Normas vectoriales y matriciales 11
1.4. Matrices por bloques y sucesiones de matrices 17

2. Métodos Iterativos **21**
2.1. Convergencia General de Métodos 22
2.2. Métodos de Jacobi, Gauss-Seidel y SOR 23
2.2.1. Convergencia de los métodos de Jacobi, Gauss-Seidel y SOR 28

3. Resultados Numéricos **34**
3.1. Discretización del problema de Poisson en una dimensión 36
3.2. Discretización del problema de Poisson en dos dimensiones 40

Conclusiones **46**

Bibliografía **47**

Agradecimientos

Le agradezco a Dios por haberme acompañado y guiado a lo largo de mi carrera, por ser mi fortaleza en los momentos de debilidad; brindarme nuevas experiencias y sobre todo felicidad.

Le doy gracias a mis padres, Daniel y Elizabeth, por apoyarme en todo momento, por los valores que me han inculcado, y por haberme dado la oportunidad de educarme.

A mis hermanos, Daniel, Miriam, Israel y Abigail, por ser parte importante de mi vida y llenarla de alegrías y amor, cuando más lo he necesitado.

Le agradezco la confianza, apoyo y dedicación de tiempo a mi profesor Roberto Cabrales; por haber compartido conmigo sus conocimientos, permitiendo llevar a cabo esta tesis.

Finalmente, agradezco a cada una de las personas que, de una u otra manera han contribuido a mi crecimiento personal e instructivo.

Darío.

Introducción

En el álgebra lineal, uno de los objetivos principales es resolver sistemas de ecuaciones lineales. Su importancia se debe a que diferentes problemas en ingeniería y modelos matemáticos se reducen a resolver un sistema de este tipo. Comunmente, en la resolución de estos sistemas de ecuaciones, se utilizan matrices, una herramienta fundamental para realizar cálculos en el álgebra lineal. En este caso el problema es el siguiente: dados una matriz $A \in \mathcal{M}_{m\times n}(\mathbb{R})$ (llamada matriz de coeficientes) y un vector $\boldsymbol{b} \in \mathbb{R}^m$ (llamado vector de términos independientes), se desea saber si existe al menos un vector $\boldsymbol{x} \in \mathbb{R}^n$ (llamado vector de incógnitas), tales que

$$A\boldsymbol{x} = \boldsymbol{b}. \tag{1}$$

Asociados al sistema (1), hay varias preguntas:

1. Existe al menos un vector $\boldsymbol{x}$ tal que (1) sea válida?
2. Si existe el vector $\boldsymbol{x}$, cuántos hay?
3. Si existe un único vector $\boldsymbol{x}$, cómo lo calculamos?
4. Bajo qué condiciones sobre la matriz A el sistema tiene solución?

En la presente actividad de titulación, nos concentraremos en la tercera pregunta: el estudio de algunos métodos para calcular el vector $\boldsymbol{x}$ más especificamente, dichos métodos calculan en general una aproximación del vector $\boldsymbol{x}$, llamado solución del sistema (1). Consideramos además el caso donde $m = n$.

Para resolver un sistema de ecuaciones lineales existen dos grandes familias de métodos. Están los métodos directos y los métodos iterativos. Estos últimos son más adecuados cuando trabajamos con matrices de gran tamaño y que muchas veces son esparsas, entendiendo por matrices esparsas, aquellas matrices que tienen una gran cantidad de entradas nulas.

Los cursos de álgebra lineal, se centran, en su mayoría, en el estudio de los métodos directos para la resolución del sistema (1), como el método de eliminación de Gauss que consiste en la reducción de la matriz A a su forma escalonada. Sin embargo, quedarse con el estudio de solo estos métodos, considerando que la resolución de sistemas de ecuaciones tiene una gran cantidad de aplicaciones en las que la matriz A tiene características que hacen difícil el uso de los métodos directos, nos parece un poco limitado. Por esta razón creemos que es interesante ampliar el estudio a los métodos iterativos para resolver sitemas lineales.

En los métodos iterativos, se realizan iteraciones para aproximarse a las solución $\boldsymbol{x}$ aprovechando las características propias de la matriz A, tratando de usar el menor número de operaciones matemáticas posible. Rara vez se usan para resolver sistemas de ecuaciones lineales de dimensión pequeña, ya que el tiempo necesario para conseguir una exactitud satisfactoria rebasa el requerido en los métodos directos. Sin embargo, en el caso de sistemas con un gran número de ecuaciones y con un alto procentaje de coeficientes nulos, donde los métodos directos no suelen resultar ventajosos (ya que a lo largo del proceso de eliminación muchos de los coeficientes nulos de A dejan de serlo, elevando notablemente el gasto de memoria en el ordenador), los métodos iterativos sí son eficientes, tanto en almacenamiento en la computadora como en el tiempo que se invierte en su solución.

Los métodos iterativos para resolver sistemas de ecuaciones lineales, a su vez, se dividen en dos tipos los métodos iterativos estacionarios y los métodos iterativos no estacionarios, estos están basados en la construcción de una sucesión de vectores que convergen a la solución exacta de $A\boldsymbol{x} = \boldsymbol{b}$. Este estudio se basa en los métodos estacionarios.

La base de los métodos iterativos consiste en la construcción de una sucesión de vectores que convergen a $\boldsymbol{x}$, la solución del problema (1). Que esta sucesión converja, depende de las características de la matriz A, la que es dividida de alguna forma adecuada.

En esta actividad de titulación se tiene como objetivo general estudiar métodos iterativos para resolver sistemas de ecuaciones lineales, sus aplicaciones y hacer comparaciones entre los diferentes métodos. Este está dividido en los siguientes tres objetivos específicos:

1. Comprender la teoría de convergencia de los métodos iterativos estacionarios.

2. Realizar la implementación computacional de los métodos iterativos estacionarios.

3. Diseñar experimentos numéricos para comparar orden de convergencia y rapidez de convergencia de algunos los métodos estacionarios.

La actividad de titulación está dividida en tres capítulos. El primero, de Preliminares, es para introducir al lector en el estudio de los métodos iterativos. Como es un capítulo introductorio, solamente están los conceptos e ideas más importantes para el estudio de los métodos iterativos, asumiendo un conocimiento básico del lector respecto a los temas a tratar. Se comienza con la teoría de vectores y matrices, sus propiedades, operaciones básicas y reconociéndoles como espacios vectoriales. Además, se presentan tipos especiales de matrices como lo son las simétricas, las definidas positivas y las diagonal dominante, incluyendo algunas de sus propiedades. También se incluye los sitemas de ecuaciones, su consistencia y el concepto de rango y determinate de una matriz y la importancia que tienen al momento de resolver un sistema. En la sección de normas vectoriales y matriciales, se presentan sus definiciones y propiedades, además se presentan resultados que establecen relaciones con las secciones anteriores y que sirven para demostrar teoremas en el capítulo 2. El capítulo concluye con matrices por bloques y sucesiones de matrices donde se presentan resultados importantes y útiles para demostrar teoremas en el capítulo de métodos iterativos.

El segundo, de Métodos Iterativos, se centra en el estudio de los métodos iterativos estacionarios, la forma en que contruyen una sucesión de vectores y resultados que verifiquen que esta suceción converja a la solución exacta del sitema. Particularmente se presentan tres métodos, los de Jacobi, Gauss-Seidel en sus dos formas y SOR. Se estudia su forma de contrucción en forma matricial y también algunos resultados de convergencia de cada uno de ellos. Además se desarrolla un ejemplo para observar el comportamiento de estos tres métodos.

En el tercero, de Resultados Numéricos, se presentan dos ejemplos basados en el problema de Poisson en 1 y 2 dimensiones, incluye el modelamiento del problema usando diferencias finitas y los resultados del sistema de ecuaciones lineales que se obtiene. En la resolución de estos sitemas se utilizan los métodos iterativos estudiados en el segundo capítulo, se verifica el cumplimiento de la teoria del capítulo 2 y se establecen comparaciones entre ellos del tipo rapidez de convergencia y comportamiento del error.

Capítulo 1
Preliminares

Comenzaremos con la teoría que sustenta el estudio de los métodos iterativos y que serán utilizados en la presente actividad de titulación. Esta puede ser encontrada en cualquier libro de álgebra lineal, como por ejemplo [4] y [7]. Para simplificar la presentación hay cosas que se han obviado, asumiendo un conocimiento básico del lector.

1.1. Vectores y Matrices

Iniciamos con la definición de lo qué es un vector y qué es una matriz.

Definición 1.1.1 *Un vector de n componentes se define como un conjunto ordenado de n números escritos de la siguiente manera:*

$$(x_1, x_2, x_3, ..., x_n). \tag{1.1}$$

Al vector (1.1) se le llama vector fila y se le llama vector columna a un conjunto ordenado de n números escritos de la siguiente manera:

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{pmatrix}. \tag{1.2}$$

Usaremos $\mathbb{R}^n$ para denotar al conjunto de todos los vectores con n componentes reales y $\mathbb{C}^n$ para denotar al conjunto de todos los vectores con n componentes complejas. Los vectores se resaltarán con letras minúsculas negritas como $\boldsymbol{a}$, $\boldsymbol{b}$, $\boldsymbol{c}$, etc. El vector cero (todas sus componentes son ceros) se denota por $\mathbf{0}$.

Definición 1.1.2 *Sean m, n enteros positivos. Una matriz A de tamaño $m \times n$ es un arreglo rectangular de $m \cdot n$ números dispuestos en m filas y n columnas de la siguiente forma:*

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1j} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2j} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots & & \vdots \\ a_{i1} & a_{i2} & \cdots & a_{ij} & \cdots & a_{in} \\ \vdots & \vdots & & \vdots & & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mj} & \cdots & a_{mn} \end{pmatrix}. \tag{1.3}$$

El elemento ij de A, denotado por a_{ij}, es el número que aparece en la fila i y la columna j de A. En ocasiones la matriz A se escribe como $A = (a_{ij})$. Las matrices las denotaremos por letras mayúsculas. Para referirnos a una matriz A de tamaño $m \times n$, simplemete diremos que $A \in \mathcal{M}_{m\times n}(\mathbb{R})$ si sus componentes son reales o $A \in \mathcal{M}_{m\times n}(\mathbb{C})$ si sus componentes son complejas.

Si A es una matriz de tamaño $m \times n$ con $m = n$, entonces A se llama matriz cuadrada y diremos que $A \in \mathcal{M}_n(\mathbb{R})$ si sus componentes son reales o $A \in \mathcal{M}_n(\mathbb{C})$ si sus componentes son complejas. Por último una matriz de tamaño $m \times n$ con todos los elementos iguales a cero se denomina matriz cero de tamaño $m \times n$.

Cabe mencionar que cada vector es un tipo especial de matriz. Por ejemplo, el vector fila (1.1) de n componentes es una matriz de tamaño $1 \times n$, mientras que el vector columna (1.2) de n componentes es una matriz de tamaño $n \times 1$.

El conjunto de todos los vectores o matrices del mismo tamaño, tienen la estructura de espacio vectorial, como veremos más adelante en esta sección.

Definición 1.1.3 *Un espacio vectorial real V es un conjunto de objetos, denominados vectores, junto con dos operaciones binarias llamadas suma y multiplicación por un escalar y que satisfacen los diez axiomas enumerados a continuación.*

1. Si $\boldsymbol{x} \in V$ y $\boldsymbol{y} \in V$, entonces $\boldsymbol{x} + \boldsymbol{y} \in V$.
2. Para todo $\boldsymbol{x}$, $\boldsymbol{y}$ y $\boldsymbol{z}$ en V, $(\boldsymbol{x} + \boldsymbol{y}) + \boldsymbol{z} = \boldsymbol{x} + (\boldsymbol{y} + \boldsymbol{z})$.
3. . Existe un vector $\mathbf{0} \in V$tal que para todo $\boldsymbol{x} \in V$, $\boldsymbol{x} + \mathbf{0} = \mathbf{0} + \boldsymbol{x} = \boldsymbol{x}$.
4. Si $\boldsymbol{x} \in V$, existe un vector $-\boldsymbol{x} \in V$ tal que $\boldsymbol{x} + (-\boldsymbol{x}) = \mathbf{0}$.
5. Si $\boldsymbol{x}$ y $\boldsymbol{y}$ están en V, entonces $\boldsymbol{x} + \boldsymbol{y} = \boldsymbol{y} + \boldsymbol{x}$.
6. Si $\boldsymbol{x} \in V$ y α es un escalar, entonces $\alpha\boldsymbol{x} \in V$.
7. Si $\boldsymbol{x}$ y $\boldsymbol{y}$ están en V y α es un escalar, entonces $\alpha(\boldsymbol{x} + \boldsymbol{y}) = \alpha\boldsymbol{x} + \alpha\boldsymbol{y}$.
8. Si $\boldsymbol{x} \in V$ y α y β son escalares, entonces $(\alpha + \beta)\boldsymbol{x} = \alpha\boldsymbol{x} + \beta\boldsymbol{x}$.

9. Si $\boldsymbol{x} \in V$ y α y β son escalares, entonces $\alpha(\beta\boldsymbol{x}) = (\alpha\beta)\boldsymbol{x}$.

10. Para cada vector $\boldsymbol{x} \in V$, $1\boldsymbol{x} = \boldsymbol{x}$.

Como los vectores son un tipo de matrices podemos definir la suma para matrices. Lo haremos considerando las matrices de componentes reales.

Definición 1.1.4 *Sean $A = (a_{ij})$ y $B = (b_{ij}) \in \mathcal{M}_{m\times n}(\mathbb{R})$. Entonces la suma de A y B es la matriz $C = (c_{ij})$ tal que*

$$c_{ij} = (a_{ij} + b_{ij}).$$

Es decir, $C \in \mathcal{M}_{m\times n}(\mathbb{R})$ se obtiene al sumar las componentes correspondientes de A y B. Notemos que para poder sumar dos matrices estas deben ser de igual tamaño.

Otra operación con matrices es la multiplicación de una matriz por un escalar.

Definición 1.1.5 *Si $A = (a_{ij}) \in \mathcal{M}_{m\times n}(\mathbb{R})$ y si α es un escalar, entonces la matriz $\alpha A \in \mathcal{M}_{m\times n}(\mathbb{R})$ está dada por*

$$\alpha A = (\alpha a_{ij}).$$

Es decir, αA es la matriz obtenida al multiplicar cada componente de A por el escalar α.

Se presentan a continuación las propiedades básicas sobre la suma de matrices y la multiplicación por escalares.

Teorema 1.1.6 *Sean A, B y $C \in \mathcal{M}_{m\times n}(\mathbb{R})$ y sean α y β dos escalares. Entonces*

i. $A + 0 = A$.

ii. $0A = 0$.

iii. $A + B = B + A$. *(Ley conmutativa para la suma de matrices).*

iv. $(A + B) + C = A + (B + C)$. *(Ley asociativa para la suma de matrices).*

v. $\alpha(A+B) = \alpha A+\alpha B$. *(Ley distributiva para la multiplicación por un escalar).*

vi. $1A = A$.

vii. $(\alpha + \beta)A = \alpha A + \beta A$.

Para poder definir la multiplicación entre matrices, primero debemos tener noción de lo que se conoce como producto interno entre dos vectores.

Definición 1.1.7 *Sean $\boldsymbol{a} = (a_1, a_2, ..., a_n)$ y $\boldsymbol{b} = (b_1, b_2, ..., b_n)$ dos vectores en $\mathbb{R}^n$. Entonces el producto escalar de a y b denotado por $\langle \boldsymbol{a}, \boldsymbol{b}\rangle$ está dado por*

$$\langle \boldsymbol{a}, \boldsymbol{b}\rangle = a_1b_1 + a_2b_2 + ... + a_nb_n. \tag{1.4}$$

El producto escalar a menudo es llamado producto interno o producto punto de los vectores y también puede ser denotado como $\boldsymbol{a} \cdot \boldsymbol{b}$. Notemos que $\boldsymbol{a}$ y $\boldsymbol{b}$ deben tener el mismo número de componentes, además observemos que el producto escalar siempre es un número.

El teorema que se presenta a continuación se deduce directamente de la definición del producto escalar.

Teorema 1.1.8 *Sean $\boldsymbol{a}$, $\boldsymbol{b}$ y $\boldsymbol{c}$ tres vectores en $\mathbb{R}^n$ y sean α y β dos escalares. Entonces*

i. $\boldsymbol{a} \cdot 0 = 0$.

ii. $\boldsymbol{a} \cdot \boldsymbol{b} = \boldsymbol{b} \cdot \boldsymbol{a}$. *(Ley conmutativa del producto escalar).*

iii. $\boldsymbol{a} \cdot (\boldsymbol{b} + \boldsymbol{c}) = \boldsymbol{a} \cdot \boldsymbol{b} + \boldsymbol{a} \cdot \boldsymbol{c}$. *(Ley distributiva del producto escalar).*

iv. $(\alpha \boldsymbol{a}) \cdot \boldsymbol{b} = \alpha(\boldsymbol{a} \cdot \boldsymbol{b})$.

Ahora podemos definir la multiplicación de matrices.

Definición 1.1.9 *Sea $A = (a_{ij}) \in \mathcal{M}_{m\times n}(\mathbb{R})$, y sea $B = (b_{ij}) \in \mathcal{M}_{n\times p}(\mathbb{R})$. Entonces el producto de A y B es una matriz $C = (c_{ij}) \in \mathcal{M}_{m\times p}(\mathbb{R})$, en donde*

$$c_{ij} = a_{i1}b_{1j} + a_{i2}b_{2j} + ... + a_{in}b_{nj}. \tag{1.5}$$

Es decir, el elemento ij de AB es el producto punto de la fila i de A y la columna j de B. Notemos que dos matrices se pueden multiplicar solo si el número de columnas de la primera matriz es igual al número de filas de la segunda matriz, ya que de otra forma el vector de la fila i de A y el vector de la columna j de B no tendrán la misma cantidad de componentes por lo que no se podrá calcular el producto punto entre ellos. Ahora, si el número de columnas de A es igual al número de filas de B, entonces se dice que A y B son compatibles bajo la multiplicación.

Consideremos el siguiente teorema donde encontramos algunas propiedades para la multiplicación de matrices.

Teorema 1.1.10 *Sea $A \in \mathcal{M}_{m\times n}(\mathbb{R})$, $B \in \mathcal{M}_{n\times p}(\mathbb{R})$ y $C \in \mathcal{M}_{p\times q}(\mathbb{R})$. Entonces se cumple la ley asociativa*

$$A(BC) = (AB)C.$$

La matriz ABC es una matriz que pertenece a $\mathcal{M}_{m\times q}(\mathbb{R})$. Es de importancia mencionar que la multiplicación de matrices, en general, no es conmutativa, esto quiere decir que, puede existir algún caso donde sí la multiplicación de dos matrices sea conmutativa.

Ejemplo 1.1.11 Consideremos las siguientes matrices de tamaño 2×2,

$$A = \begin{pmatrix} 2 & 0 \\ 3 & 1 \end{pmatrix} \quad y \quad B = \begin{pmatrix} 0 & 5 \\ 3 & 4 \end{pmatrix}.$$

Consideremos la multiplicación AB y BA.

$$AB = \begin{pmatrix} 0 & 10 \\ 3 & 19 \end{pmatrix} \neq \begin{pmatrix} 15 & 5 \\ 18 & 4 \end{pmatrix} = BA.$$

Así $AB \neq BA$, por lo que la multiplicación de matrices no es conmutativa. □

Como se puede observar, los conjuntos de todos los vectores $\in \mathbb{R}^n$ y de todas las matrices $\in \mathcal{M}_{m\times n}(\mathbb{R})$ cuentan con diversas propiedades particulares respecto a la suma y a la multiplicación por escalar, por ello son espacios vectoriales.

Considerando que nuestro estudio se centra en los métodos iterativos y estos están basados en la construcción de una sucesión de vectores que convergen a la solución del sistema de ecuaciones, será necesario entonces, estudiar cuál es la posibilidad de que estos métodos converjan. Para ello, necesitaremos tener conocimientos acerca de los valores y vectores propios y del polinomio característico de una matriz.

Definición 1.1.12 *Sea $A \in \mathcal{M}_n(\mathbb{C})$. Un número $\lambda \in \mathbb{C}$ se denomina valor propio o característico de A si existe un vector $\boldsymbol{v}$ diferente de cero en $\mathbb{C}^n$ tal que*

$$A\boldsymbol{v} = \lambda\boldsymbol{v}.$$

El vector $\boldsymbol{v}$ se denomina vector propio o característico de A correspondiente al valor propio λ.

Sea λ un valor propio de A. Por la Definición 1.1.12 tenemos que

$$(A - \lambda I)\boldsymbol{v} = 0. \tag{1.6}$$

La ecuación (1.6) corresponde a un sistema homogéneo de n ecuaciones y de n incógnitas dadas por las componentes del vector $\boldsymbol{v}$. Como el sistema cuenta con soluciones no triviales ($\boldsymbol{v} \neq 0$), se concluye que $\det(A - \lambda I) = 0$. Ahora, si $\det(A - \lambda I) = 0$, entonces la ecuación (1.6) tiene soluciones no triviales y λ es el valor propio de A. Es posible probar que $\det(A-\lambda I)$ es un polinomio de grado n, llamado polinomio característico de A.

Definición 1.1.13 *Si A es una matriz cuadrada, el polinomio definido por*

$$p(\lambda) = \det(A - \lambda I),$$

*recibe el nombre de **polinomio característico** de A.*

Los n ceros de p (polinomio característico de A), son valores propios de esa matriz A. Un concepto de mucha utilidad en los métodos iterativos tiene relación directa con los valores propios de una matriz.

Definición 1.1.14 *El **radio espectral** de una matriz A está definido por*

$$\rho(A) = \max\{|\lambda| : \lambda \text{ es un valor propio de } A\}.$$

Este concepto será de gran importancia al momento de demostrar si un método iterativo es o no convergente.

Ahora si consideramos una matriz $A \in \mathcal{M}_n(\mathbb{R})$ con n considerablemete grande, notaremos que encontrar su polinomio característico y posteriormente sus valores y vectores característicos será cada vez más difícil y si a esto agregamos la importancia que tendrá el radio espectral de la matriz para establecer condiciones de convergencia de un método iterativo, será necesario, como veremos más adelante en este capítulo, establecer una forma de relacionar el radio espectral con algo más factible de calcular.

1.1.1. Tipos especiales de matrices

La convergencia de un método iterativo no se puede asegurar para todas las matrices en general. Sin embargo, hay teoremas que demuestran que existen matrices especiales en las que sí se puede asegurar su convergencia. Para ello será importante tener el concepto de matrices definidas positivas.

Definición 1.1.15 *Una matriz $A \in \mathcal{M}_n(\mathbb{R})$ es definida positiva en $\mathbb{R}^n$ si $\langle A\boldsymbol{x}, \boldsymbol{x}\rangle > 0$, para todo $\boldsymbol{x} \in \mathbb{R}^n$, con $\boldsymbol{x} \neq 0$.*
Si la desigualdad estricta es sustituida por $(\geq)$ la matriz A se llama definida semi-positiva.

Una matriz definida positiva y definida semi-positiva también se denotan como $A > 0$ y $A \geq 0$ respectivamente. Las matrices definidas positivas cumplen con las siguientes propiedades, que nos serán útiles para demostrar proposiciones y teoremas en el capítulo 2.

Proposición 1.1.16 *Sean A, $B \in \mathcal{M}_n(\mathbb{R})$.*

i. Si A es definida positiva entonces es invertible (su determinante es positivo), y su inversa es definida positiva.

ii. Si A y B son matrices definidas positivas, entonces la suma $A + B$ también lo es. Además si $AB = BA$, entonces AB es también definida positiva.

En los textos de álgebra lineal podemos encontar otros resultados para las matrices definidas positivas, ver [4] y [8]. Algunos importantes para nuestro estudio son los que presentaremos en las siguientes definiciones.

Definición 1.1.17 *Una matriz B se dice que es la raíz cuadrada de una matriz A, si el producto matricial BB es igual a A. La raíz cuadrada de una matriz es denotada por*

$$B = A^{1/2},$$

y su inversa es denotada por $A^{-1/2}$.

Respecto a la definición anterior, en el texto [9] en la página 40, podemos encontrar algunas características de la raíz cuadrada de una matriz.

Proposición 1.1.18 *Sea $A \in \mathcal{M}_n(\mathbb{R})$.*

i. Si A es una matriz definida positiva, entonces existe al menos una matriz definida positiva B tal que $B^2 = A$.

ii. Si A es definida positiva entonces $A^{1/2}$ también es definida positiva. Si A es semidefinida positiva entonces también lo es $A^{1/2}$.

iii. $A^{1/2}$ conmuta con A y cualquier polinomio de A.

iv. $A^{1/2}$ es la única solución positiva de la ecuación matricial $X^2 = A > 0$.

En algunos casos el que una matriz sea definida positiva no será condición suficiente para que un método iterativo sea convergente y será necesario pedir que también sea simétrica. Pero, para establecer qué es una matriz simétrica debemos introducir primero el concepto de matriz transpuesta.

Definición 1.1.19 *Sea una matriz $A \in \mathcal{M}_{m\times n}(\mathbb{R})$, entonces la transpuesta de A, que se denota por A^T, es la matriz en $\mathcal{M}_{n\times m}(\mathbb{R})$ que se obtiene al intercambiar las filas por las columnas de A. Es decir, si $A = (a_{ij})$ entonces $A^T = (a_{ji})$.*

Definición 1.1.20 *Una matriz $A \in \mathcal{M}_n(\mathbb{R})$ se llama simétrica si $A = A^T$.*

Es necesario mencionar que si $A = -A^T$ se llama antisimétrica y se llama ortogonal si $A^T A = AA^T = I$, es decir, $A^{-1} = A^T$.

Ahora si consideramos $A \in \mathcal{M}_{m\times n}(\mathbb{C})$, denotamos por $\overline{A}$ la matriz conjugada compleja, es decir, si $B = \overline{A}$ entonces $b_{ij} = \overline{a}_{ij}$.

Definición 1.1.21 *Sea $A \in \mathcal{M}_{m\times n}(\mathbb{C})$, la matriz $B = A^H \in \mathcal{M}_{m\times n}(\mathbb{C})$, se llama la matriz transpuesta conjugada o adjunta de A si*

$$B = A^H = \overline{A}^T.$$

Ampliando la definición de una matriz simétrica al conjunto de las matrices complejas, presentamos la definición de una matriz hermitiana.

Definición 1.1.22 *Una matriz $A \in \mathcal{M}_n(\mathbb{C})$ se llama hermitiana o auto-adjunta, si $A^T = \overline{A}$, es decir,*

$$A^H = A.$$

También cabe mencionar que, $A \in \mathcal{M}_{m\times n}(\mathbb{C})$ es llamada unitaria si $A^H A = AA^H = I$, como consecuencia para que la matriz A sea unitaria se debe cumplir que $A^{-1} = A^H$. Finalmente, si $AA^H = A^H A$, A se llama normal.

Ahora, intoduciremos la noción de matrices similares y su relación con los valores propios de una matriz.

Definición 1.1.23 *Sean A y $B \in \mathcal{M}_n(\mathbb{C})$, son llamadas similares si existe una matriz T regular tal que*

$$A = T^{-1}BT.$$

La importancia de estas matrices radica en las propiedades que tienen. En particular, si A y B son similares, los valores propios de ellas coinciden, contando multiplicidades, por lo tanto $\rho(A) = \rho(B)$.

Otras matrices especiales para las que se pueden establecer criterios de convergencias de los métodos iterativos son las llamadas matrices diagonal dominante.

Definición 1.1.24 *Una matriz $A \in \mathcal{M}_n(\mathbb{R})$ se llama diagonalmente dominante por filas si*

$$|a_{ii}| \geq \sum_{j=1, j\neq i}^{n} |a_{ij}|, \qquad con\ \ i = 1, ..., n.$$

A se llama diagonalmente dominante por columnas si

$$|a_{ii}| \geq \sum_{j=1, j\neq i}^{n} |a_{ji}|, \qquad con\ \ i = 1, ..., n.$$

Si las desigualdades anteriores se cambian a un sentido estricto, A se llama estrictamente diagonal dominante por filas y por columnas, respectivamente.

1.2. Sistemas de Ecuaciones

Como ya hemos comentado anteriormente, uno de los objetivos centrales del álgebra lineal es el estudio de los sistemas de ecuaciones lineales y una de las preguntas más importantes es si existe el vector $\boldsymbol{x}$, solución del sistema. En relación a esto podemos decir que un sistema de ecuaciones lineales de la forma $A\boldsymbol{x} = \boldsymbol{b}$, puede no tener soluciones y que la existencia de soluciones está relacionada directamente con las características de la matriz de los coeficientes A.

Consideremos una matriz $A \in \mathcal{M}_{m\times n}(\mathbb{R})$. En el caso donde $m \neq n$, la existencia de solución está dado por el estudio del rango de la matriz $(A|\boldsymbol{b})$, y que se puede calcular al escalonarla.

Definición 1.2.1 *Una matriz se encuentra en la forma escalonada si se cumplen las siguientes condiciones:*

i. Todos las filas nulas (si las hay) aparecen en la parte inferior de la matriz.

ii. El primer número diferente de cero (comenzando por la izquierda) en cualquier fila cuyos elementos no todos son cero es 1.

iii. Si dos filas sucesivas tienen elementos distintos de cero, entonces el primer 1 en la fila de abajo está más hacia la derecha que el primer 1 en la fila de arriba.

iv. Cualquier columna que contiene el primer 1 en una fila, tiene ceros en el resto de sus elementos. El primer número diferente de cero en una fila (si lo hay) se llama pivote para esa fila.

Para escalonar una matriz se usan las siguientes operaciones elementales por filas:

i. Multiplicar una fila por un número diferente de cero.

ii. Sumar un múltiplo de una fila a otra fila.

iii. Intercambiar dos filas.

Definición 1.2.2 *Supongamos que una matriz $A \in \mathcal{M}_{m\times n}(\mathbb{R})$ se reduce mediante operaciones elementales por filas a una forma escalonada E. El **rango de** A es el número de pivotes o número de filas no nulas de E y se denota como $r(A)$.*

Un importante resultado que se encuentra en el texto [4], es el siguiente.

Teorema 1.2.3 *Si $(A|\boldsymbol{b})$ está en su forma escalonada se cumple que:*

i. $r(A) = r(A|\boldsymbol{b}) = n$, el sistema tiene solución única.

ii. $r(A) = r(A|\boldsymbol{b}) < n$, el sistema tiene infinitas soluciones.

iii. $r(A) \neq r(A|\boldsymbol{b})$, el sistema no tiene solución.

Definición 1.2.4 *Un sistema lineal que no tiene solución se dice **inconsistente**. En otro caso, se dice que el sistema lineal es **consistente**.*

En el caso particular, en que $m = n$, obtenemos una matriz cuadrada $A \in \mathcal{M}_n(\mathbb{R})$. Para la existencia de solución necesitamos conocer el concepto de inversa de una matriz.

Definición 1.2.5 *Sea $A \in \mathcal{M}_n(\mathbb{R})$. A es invertible si existe $B \in \mathcal{M}_n(\mathbb{R})$ tal que*

$$AB = BA = I_n,$$

*donde I_n es la matriz identidad de orden n. B es la **inversa** de A y se denota por A^{-1}.*

Una matriz cuadrada que tiene inversa se llama **invertible o regular**. Una matriz cuadrada que no tiene inversa se llama **singular**.

Ahora bien, si consideramos el sistema de ecuaciones $A\boldsymbol{x} = \boldsymbol{b}$, y lo multiplicamos por izquierda por A^{-1} obtenemos

$$A^{-1}\boldsymbol{b} = A^{-1}(A\boldsymbol{x}) = (A^{-1}A)\boldsymbol{x} = I_n\boldsymbol{x} = \boldsymbol{x},$$

ya que A^{-1} es la inversa de A. Finalmente tenemos

$$\boldsymbol{x} = A^{-1}\boldsymbol{b}.$$

De lo anterior podemos concluir que la unicidad de la solución de un sistema de ecuaciones lineales $A\boldsymbol{x} = \boldsymbol{b}$, está condicionado por la existencia de la matriz A^{-1} inversa de la matriz $A \in \mathcal{M}_n(\mathbb{R})$.

Definición 1.2.6 *En un sistema de ecuaciones $A\boldsymbol{x} = \boldsymbol{b}$, donde $A \in \mathcal{M}_n(\mathbb{R})$, el sistema tiene solución si existe la matriz A^{-1}.*

A continuación definiremos el determinante de una matriz, que guarda una estrecha relación con la unicidad de la solución y que será necesario más adelante en este capítulo y en los siguientes.

Definición 1.2.7 *Sea $A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$ una matriz en $\mathcal{M}_2(\mathbb{R})$. El determinante de A se define como*

$$\det A = a_{11}a_{22} - a_{12}a_{21}.$$

Con frecuencia el determinante de una matriz A se denota como $|A|$.

Para extender la definición del determinante de una matriz al conjunto de $\mathcal{M}_n(\mathbb{R})$, será necesario conocer primero lo que es un menor y un cofactor.

Definición 1.2.8 *Sea $A \in \mathcal{M}_n(\mathbb{R})$ y sea $M_{ij} \in \mathcal{M}_{n-1}(\mathbb{R})$ la matriz que se obtiene de A eliminando la fila i y la columna j. M_{ij} se llama el menor ij de A.*

Sea $A \in \mathcal{M}_n(\mathbb{R})$. El cofactor ij de A, denotado por A_{ij}, está dado por

$$A_{ij} = (-1)^{i+j}|M_{ij}|.$$

El cofactor ij de A se obtiene tomando el determinante del menor ij y multiplicándolo por $(-1)^{i+j}$. Observemos que si $i+j$ es par $(-1)^{i+j} = 1$ y si $i+j$ es impar $(-1)^{i+j} = -1$.

Definición 1.2.9 *Sea $A \in \mathcal{M}_n(\mathbb{R})$. Entonces el determinante de A, está dado por*

$$\det A = a_{11}A_{11} + a_{12}A_{12} + ... + a_{1n}A_{1n} = \sum_{k=1}^{n} a_{1k}A_{1k}. \tag{1.7}$$

En la ecuación (1.7) se define el determinante mediante los menores y cofactores del primer renglón de A. Es necesario mencionar que también se puede obtener utilizando cualquier fila o columna.

Para finalizar el tema de sistemas de ecuaciones lineales, considerando conceptos básicos que no fueron incluidos en los preliminares, pero que pueden ser encontrados en cualquier libro de álgebra lineal, podemos concluir entonces, respecto a si un sistema de ecuaciones tiene solución, con el siguiente teorema.

Teorema 1.2.10 *Sea $A \in \mathcal{M}_n(\mathbb{R})$. Entonces las ocho afirmaciones siguientes son equivalentes; es decir, cada una implica a las otras siete (de manera que si una es cierta, todas son ciertas).*

i. A es invertible.

ii. *La única solución al sistema homogéneo* $A\boldsymbol{x} = \mathbf{0}$ *es la solución trivial (*$\boldsymbol{x} = \mathbf{0}$*).*

iii. *El sistema* $A\boldsymbol{x} = \boldsymbol{b}$ *tiene una solución única para cada* $\boldsymbol{b} \in \mathbb{R}^n$*.*

iv. *A es equivalente por filas a la matriz identidad de orden n,* I_n*.*

v. *A es el producto de matrices elementales.*

vi. *La forma escalonada por filas de A tiene n pivotes.*

vii. $\det(A) \neq 0$*.*

viii. *Las columnas (y filas) de A son linealmente independientes.*

1.3. Normas vectoriales y matriciales

Para poder comenzar el estudio de los métodos iterativos para resolver sistemas lineales de ecuaciones, es necesario contar con un medio para medir distancias entre vectores, y así determinar si una sucesión de vectores converge a la solución del sistema. Para ello consideraremos $\mathbb{R}^n$ el conjunto de todos los vectores columna n-dimensionales con componentes reales. Para definir una distancia en $\mathbb{R}^n$ utilizaremos la noción de norma.

Definición 1.3.1 *Una norma vectorial en* $\mathbb{R}^n$ *es una función de* $\mathbb{R}^n$ *a* $\mathbb{R}$ *denotada como* $\|\cdot\|$*, con las siguientes propiedades*

i. $\|\boldsymbol{x}\| \geq 0$ *para todo* $\boldsymbol{x} \in \mathbb{R}$*.*

ii. $\|\boldsymbol{x}\| = 0$ *si y solo si* $\boldsymbol{x} = 0$*.*

iii. $\|\alpha\boldsymbol{x}\| = |\alpha|\|\boldsymbol{x}\|$ *para todo* $\alpha \in \mathbb{R}^n$*.*

iv. $\|\boldsymbol{x} + \boldsymbol{y}\| \leq \|\boldsymbol{x}\| + \|\boldsymbol{y}\|$ *para todo* $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$*.*

En este estudio utilizaremos principalmente dos normas específicas en $\mathbb{R}^n$, la norma euclideana denotada como l_2 y la norma infinito denotada como l_∞.

Definición 1.3.2 *Las normas* l_2 *y* l_∞ *de un vector* $\boldsymbol{x} \in \mathbb{R}^n$ *están definidas respectivamente como*

$$\|\boldsymbol{x}\|_2 = \left\{ \sum_{i=1}^{n} x_i^2 \right\}^{1/2} \quad y \quad \|\boldsymbol{x}\|_\infty = \max_{1 \leq i \leq n} |x_i|.$$

Usando la definición de norma de un vector podemos definir la distancia entre dos vectores como la norma de su diferencia.

Definición 1.3.3 *Si* $\boldsymbol{x}$*,* $\boldsymbol{y}$ *son vectores de* $\mathbb{R}^n$*, las distancias* l_2 *y* l_∞ *entre* $\boldsymbol{x}$ *e* $\boldsymbol{y}$ *están definidas respectivamente como*

$$\|\boldsymbol{x} - \boldsymbol{y}\|_2 = \left\{ \sum_{i=1}^{n} (x_i - y_i)^2 \right\}^{1/2} \quad y \quad \|\boldsymbol{x} - \boldsymbol{y}\|_\infty = \max_{1 \leq i \leq n} |x_i - y_i|.$$

Utilizando el concepto de distancia, podemos definir la convergencia de una sucesión de vectores en $\mathbb{R}^n$.

Definición 1.3.4 *Se dice que una sucesión $\{\boldsymbol{x}^{(k)}\}_{k=1}^{\infty}$ de vectores en $\mathbb{R}^n$* ***converge*** *a $\boldsymbol{x}$ respecto a la norma $\|\cdot\|$ si dado cualquier $\varepsilon > 0$, existe un natural $N(\varepsilon)$ tal que*

$$\|\boldsymbol{x}^{(k)} - \boldsymbol{x}\| < \varepsilon, \qquad \textit{para todo } k \geq N(\varepsilon).$$

Lo que necesitamos es que $\|\boldsymbol{x}^{(k)} - \boldsymbol{x}\|$ tienda a cero cuando k tiende a infinito, es decir $x_i^{(k)}$, la i-ésima componente del vector $\boldsymbol{x}^{(k)}$, tienda a x_i, la i-ésima componente del vector $\boldsymbol{x}$. Lo anterior lo podemos resumir en el siguiente teorema.

Teorema 1.3.5 *La sucesión de vectores $\{\boldsymbol{x}^{(k)}\}$ converge a $\boldsymbol{x}$ en $\mathbb{R}^n$ respecto a $\|\cdot\|_\infty$ si y solo si*

$$\lim_{k\to\infty} x_i^{(k)} = x_i,$$

para cada $i = 1, 2, ..., n$.

Demostración: Supongamos que $\{\boldsymbol{x}^{(k)}\}$ converge a $\boldsymbol{x}$ con respecto a la norma l_∞. Dado cualquier $\varepsilon > 0$, existe un entero $N(\varepsilon)$ tal que para todo $k \geq N(\varepsilon)$,

$$\max_{i=1,2,...,n} |x_i^{(k)} - x_i| = \|\boldsymbol{x}^{(k)} - \boldsymbol{x}\| < \varepsilon.$$

Este resultado implica que $|x_i^{(k)} - x_i| < \varepsilon$, para cada $i = 1, 2, ..., n$, de modo que $\lim_{k\to\infty} x_i^{(k)} = x_i$ para cada i.

En sentido contrario, supongamos que $\lim_{k\to\infty} x_i^{(k)} = x_i$, para cada $i = 1, 2, ..., n$. Para un $\varepsilon > 0$ dado, sea $N_i(\varepsilon)$ para cada i un entero con la propiedad de que

$$|x_i^{(k)} - x_i| < \varepsilon,$$

siempre que $k \geq N_i(\varepsilon)$.
Definimos $N(\varepsilon) = \max_{i=1,2...,n} N_i(\varepsilon)$. Si $k \geq N(\varepsilon)$, entonces

$$\max_{i=1,2,...,n} |x_i^{(k)} - x_i| = \|\boldsymbol{x}^{(k)} - \boldsymbol{x}\| < \varepsilon.$$

Esto implica que $\{\boldsymbol{x}^{(k)}\}$ converge a $\boldsymbol{x}$ con respecto a la norma l_∞. ■

Cabe mencionar que el teorema anterior es válido para cualquier norma de $\mathbb{R}^n$, ya que todas las normas son equivalentes. Ahora bien, ampliando la definición de norma para vectores al conjunto de las matrices, podemos determinar la distancia entre matrices de $\mathcal{M}_n(\mathbb{R})$ definiendo una vez más el concepto de norma, pero ahora para matrices.

Definición 1.3.6 *Una norma matricial sobre $\mathcal{M}_n(\mathbb{R})$ es una función de valor real, $\|\cdot\|$, definida en este conjunto y que satisface para todas las matrices A y $B \in \mathcal{M}_n(\mathbb{R})$ y todos los números reales α, las siguientes propiedades:*

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

(ii) $\|A\| = 0$, *si y solo si* A *es* 0, *la matriz con todas las entradas cero.*

(iii) $\|\alpha A\| =\leq |\alpha|\|A\|$.

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

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

Al igual que con los vectores, la distancia entre dos matrices A y B de $\mathcal{M}_n(\mathbb{R})$ respecto a esta norma matricial se define como $\|A - B\|$.

Una definición importante a considerar, por su importancia en teoremas posteriores, es la de una norma matricial consistente.

Definición 1.3.7 *Diremos que una norma matricial* $\|\cdot\|$ *es compatible o consistente con una norma vectorial* $\|\cdot\|$ *si*

$$\|A\boldsymbol{x}\| \leq \|A\|\|\boldsymbol{x}\|, \qquad \textit{para todo } \boldsymbol{x} \in \mathbb{R}^n.$$

Que una norma matricial sea consistente, será crucial al momento de verificar si un método iterativo es convergente, por la relación que se puede establecer con el radio espectral de una matriz.

Definición 1.3.8 *Sea* $\|\cdot\|$ *una norma vectorial de* $\mathbb{R}^n$. *Entonces podemos definir una norma matricial (llamada natural o inducida) como*

$$\|A\| = \max_{\|\boldsymbol{x}\|=1} \|A\boldsymbol{x}\|.$$

Como podemos ver normas normas matriciales naturales son definidas por las normas de vectores. Una forma alternativa de escribir $\|A\|$.

Teorema 1.3.9 *Sea* $\|\cdot\|$ *una norma vectorial. La función*

$$\|A\| = \max_{\boldsymbol{x}\neq 0} \frac{\|A\boldsymbol{x}\|}{\|\boldsymbol{x}\|},$$

es una matricial inducida o natural.

Demostración: Notemos que para cualquier $\boldsymbol{x} \neq 0$, el vector $\boldsymbol{y} = \boldsymbol{x}/\|\boldsymbol{x}\|$ es un vector unitario. Por lo que

$$\max_{\|\boldsymbol{y}\|=1} \|A\boldsymbol{y}\| = \max_{\boldsymbol{x}\neq 0} \left\| A\left(\frac{\boldsymbol{x}}{\|\boldsymbol{x}\|}\right)\right\| = \max_{\boldsymbol{x}\neq 0} \frac{\|A\boldsymbol{x}\|}{\|\boldsymbol{x}\|}.$$

Lo que demuestra el teorema anterior. ■

De acuerdo a la definición 1.3.8 podemos establecer definiciones para las normas matriciales l_∞ y l_2 respectivamente:

$$\|A\|_\infty = \max_{\|\boldsymbol{x}\|_\infty=1} \|A\boldsymbol{x}\|_\infty, \tag{1.8}$$

$$\|A\|_2 = \max_{\|\boldsymbol{x}\|_2=1} \|A\boldsymbol{x}\|_2. \tag{1.9}$$

La norma l_∞, también pude ser calculada fácilmente con las entradas de la matriz A.

Teorema 1.3.10 *Si $A = (a_{ij}) \in \mathcal{M}_n(\mathbb{R})$, entonces*

$$\|A\|_\infty = \max_{1\le i\le n} \sum_{j=1}^{n} |a_{ij}|.$$

Demostración: En primer lugar mostraremos que $\|A\|_\infty \le \max_{1\le i\le n} \sum_{j=1}^{n} |a_{ij}|$.

Sea $\boldsymbol{x}$ un vector n-dimensional con $\|\boldsymbol{x}\|_\infty = 1$. Así $A\boldsymbol{x}$ también es un vector n-dimensional,

$$\|A\boldsymbol{x}\|_\infty = \max_{1\le i\le n} |(A\boldsymbol{x})_i| = \max_{1\le i\le n} \left| \sum_{j=1}^{n} a_{ij}x_j \right| \le \max_{1\le i\le n} \sum_{j=1}^{n} |a_{ij}| \max_{1\le j\le n} |x_j|.$$

Como $\|\boldsymbol{x}\|_\infty = 1$ y por la Definición 1.3.8 se tiene que

$$\|A\|_\infty = \max_{\|\boldsymbol{x}\|_\infty=1} \|A\boldsymbol{x}\|_\infty \le \max_{1\le i\le n} \sum_{j=1}^{n} |a_{ij}|.$$

Ahora vamos a mostrar la desigualdad opuesta. Sea p un entero con

$$\sum_{j=1}^{n} |a_{pj}| = \max_{1\le i\le n} \sum_{j=1}^{n} |a_{ij}|,$$

y $\boldsymbol{x}$ un vector con componentes

$$x_j = \begin{cases} 1 & si \quad a_{pj} \ge 0, \\ \\ -1 & si \quad a_{pj} < 0. \end{cases}$$

Donde $\|\boldsymbol{x}\|_\infty = 1$ y $a_{pj}x_j = |a_{pj}|$, para todo $j = 1, ..., n$

$$\|A\boldsymbol{x}\|_\infty = \max_{1\le i\le n} \left| \sum_{j=1}^{n} a_{ij}x_j \right| \ge \left| \sum_{j=1}^{n} a_{ij}x_j \right| = \left| \sum_{j=1}^{n} |a_{ij}| \right| = \max_{1\le i\le n} \sum_{j=1}^{n} |a_{ij}|.$$

Este resultado implica que

$$\|A\|_\infty = \max_{\|\boldsymbol{x}\|_\infty=1} \|A\boldsymbol{x}\|_\infty \ge \max_{1\le i\le n} \sum_{j=1}^{n} |a_{ij}|.$$

Entonces considerando que $\|A\|_\infty \leq \max\limits_{1\leq i\leq n} \sum\limits_{j=1}^{n} |a_{ij}|$ y también que $\|A\|_\infty \geq \max\limits_{1\leq i\leq n} \sum\limits_{j=1}^{n} |a_{ij}|$, tenemos que

$$\|A\|_\infty = \max_{1\leq i\leq n} \sum_{j=1}^{n} |a_{ij}|.$$

■

Otra forma de escribir la norma matricial l_2, es la que presentamos en el siguiente teorema.

Teorema 1.3.11 *Sea $A \in \mathcal{M}_n(\mathbb{R})$, entonces*

$$\|A\|_2 = \max\{|\langle A\boldsymbol{x}, \boldsymbol{y}\rangle|/(\|\boldsymbol{x}\|_2\|\boldsymbol{y}\|_2), \qquad \boldsymbol{x} \neq 0, \boldsymbol{y} \in \mathbb{R}^n\}.$$

Ahora, dada una norma vectorial $\|\cdot\|$ y una matriz T, también podemos definir una norma vectorial adicional, $\|\cdot\|_T$ como

$$\|\boldsymbol{x}\|_T = \|T\boldsymbol{x}\|.$$

De la misma forma denotamos la norma matricial

$$\|A\|_T = \|TAT^{-1}\|, \qquad A \in \mathcal{M}_n(\mathbb{R}).$$

Del texto [9] en la página 41, consideraremos la siguiente definición que será de utilidad al momento de demostrar algunas proposiciones en el capítulo 2.

Definición 1.3.12 *Sea A una matriz definida positiva. La norma $\|\cdot\|_A$ es generada por el escalar.*

$$\langle \boldsymbol{x}, \boldsymbol{y}\rangle_A = \langle A\boldsymbol{x}, \boldsymbol{y}\rangle .$$

Una representación equivalente de la norma $\|\cdot\|_A$ es

$$\|\boldsymbol{x}\|_A = \langle A\boldsymbol{x}, \boldsymbol{x}\rangle^{1/2} .$$

Por último, la norma matricial correspondiente para $\|\cdot\|_A$ está relacionada con l_2 por

$$\|B\|_A = \|A^{1/2}BA^{-1/2}\|_2, \qquad B \in \mathcal{M}_n(\mathbb{R}). \tag{1.10}$$

En la ecuación (1.10), el hecho que A sea definida positiva es importante, ya que de esto depende que exista $A^{1/2}$ (Proposición 1.1.18).

Anteriormente comentamos que sería necesario establecer alguna relación del radio espectral de una matriz con algo más factible de calcular. La siguiente definición relaciona el radio espectral de una matriz con la norma de esta.

Teorema 1.3.13 *Si $A \in \mathcal{M}_n(\mathbb{C})$ y $\|\cdot\|$ es una norma matricial consistente, entonces*

$$\rho(A) \leq \|A\|.$$

Demostración: Sabemos que $\rho(A) = \text{máx}\{|\lambda_k|$, con λ_k un valor propio de $A\}$, entonces existe $\boldsymbol{x}_\lambda \neq \mathbf{0}$ tal que $A\boldsymbol{x}_\lambda = \lambda_k \boldsymbol{x}_\lambda$. Ahora tenemos $\|\lambda_k \boldsymbol{x}_\lambda\| = |\lambda_k| \|\boldsymbol{x}_\lambda\| = \|A\boldsymbol{x}_\lambda\|$, como $\|\cdot\|$ es norma consistente y $\boldsymbol{x}_k \neq 0$ tenemos que

$$|\lambda_k| \|\boldsymbol{x}_\lambda\| = \|A\boldsymbol{x}_\lambda\| \leq \|A\| \|\boldsymbol{x}_\lambda\|,$$

de donde obtenemos que

$$|\lambda_k| \leq \|A\|,$$

para todo λ_k valor propio de A. Por lo tanto $\rho(A) \leq \|A\|$. ■

El siguiente teorema relaciona la norma l_2 de una matriz con su radio espectral y es el siguiente.

Teorema 1.3.14 *Sea A una matriz normal. La norma l_2 satisface*

$$\|A\|_2 = \rho(A). \tag{1.11}$$

Demostración: Dada una matriz normal A, podemos encontar una matriz unitaria Q y una matriz diagonal D tales que $A = QDQ^H$. Ahora notemos que, como $\|Q\| = \|Q^H\| = 1$, como muestra el Lema 2.9.1 del texto [9], tenemos que

$$\|Q^H D\|_2 = \|DQ^H\|_2 = \|Q^H DQ\|_2 = \|QDQ^H\|_2 = \|D\|_2,$$

y de esta forma $\|A\|_2 = \|QDQ^H\|_2 = \|D\|_2$. Ahora bien, $\|D\|_2 = \rho(D)$, ya que D es diagonal. Por la similaridad de A y D, $\rho(A) = \rho(D)$, entonces

$$\|A\|_2 = \|D\|_2 = \rho(D) = \rho(A).$$

■

Siguiendo la relación de la norma l_2 de una matriz con su radio espectral, tenemos la siguiente proposición.

Proposición 1.3.15 *Sean A y $B \in \mathcal{M}_n(\mathbb{R})$*

(i) $0 \leq A \leq B$ *implica que,* $\|A\|_2 \leq \|B\|_2$ *y* $\rho(A) \leq \rho(B)$.

(ii) $0 \leq A < B$ *implica que,* $\|A\|_2 < \|B\|_2$ *y* $\rho(A) < \rho(B)$.

Demostración: Para demostrar esta proposición mencionaremos un resultado que podemos encontrar en el texto [9] en la página 37 (Lema 2.9.11), y dice lo siguiente

$$\rho(A) \leq r(A) \leq \|A\|_2, \tag{1.12}$$

donde $r(A)$ se llama radio numérico de una matriz y se define como

$$r(A) = \text{máx}\{|\langle A\boldsymbol{x}, \boldsymbol{x}\rangle|/\|x\|_2^2, \quad \boldsymbol{x} \neq 0, \ \boldsymbol{x} \in \mathbb{C}^n\}.$$

Como $\|A\|_2 = \rho(A)$ y $\|B\|_2 = \rho(B)$, es suficiente mostrar solamente que $\rho(A) \leq \rho(B)$. $A \geq 0$, tiene un valor propio $\lambda = \rho(A)$ y su vector propio correspondiente $\boldsymbol{x}$ con $\|\boldsymbol{x}\|_2 = 1$ y por la ecuación (1.12), tenemos que

$$\rho(A) = \langle A\boldsymbol{x}, \boldsymbol{x}\rangle \leq \langle B\boldsymbol{x}, \boldsymbol{x}\rangle = \rho(B).$$

Lo que demuestra la afirmación (i) y de forma análoga se demuestra (ii). ■

Un resultado interesante, que será útil para demostrar un teorema en el próximo capítulo, y es similar al Teorema 1.3.13 es el siguiente.

Proposición 1.3.16 *Sea $A \in \mathcal{M}_n(\mathbb{C})$ y $\varepsilon > 0$. Entonces, existe una norma matricial inducida $\|\cdot\|_{A,\varepsilon}$ (dependiendo de ε) tal que*

$$\rho(A) < \|A\|_{A,\varepsilon} \leq \rho(A) + \varepsilon.$$

1.4. Matrices por bloques y sucesiones de matrices

En los métodos iterativos se estudia la velocidad de convergencia de cada uno de estos, algunos se le introducen parámetros que aceleran la velocidad de convergencia de un método. La siguiente definición será de utilidad para establecer el valor de un parámetro óptimo.

Definición 1.4.1 *Diremos que una matriz A está descompuesta o particionada propiamente en bloques si se puede organizar como una matriz de bloques o submatrices en la forma*

$$A = \begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1r} \\ A_{21} & A_{22} & \cdots & A_{2r} \\ \vdots & \vdots & \vdots & \vdots \\ A_{p1} & A_{p2} & \cdots & A_{pr} \end{bmatrix}$$

Los bloques se obtienen trazando rectas verticales y horizontales imaginarias entre los elementos de la matriz A. Los bloques se designan de la forma A_{ij} . Notemos que el número de filas en el bloque A_{ij} depende solo de j, siendo el mismo para todos los i, por otro lado el número de columnas en el bloque A_{ij} depende solo de i y es el mismo para todos los j.

También podemos definir una matriz diagonal por bloques de la siguiente manera:

Definición 1.4.2 *Sean $B_1, B_2, ..., B_n$ matrices cuadradas, se define una matriz diagonal por bloques como*

$$\begin{bmatrix} B_1 & 0 & \cdots & 0 \\ 0 & B_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & B_n \end{bmatrix},$$

y se denota por $diag(B_1, B_2, ..., B_n)$.

A continuación presentamos un ejemplo de una matriz diagonal por bloques.

Ejemplo 1.4.3 Consideremos las matrices $B_1 = \begin{pmatrix} 2 & 3 \\ 4 & 1 \end{pmatrix}$, $B_2 = \begin{pmatrix} 7 \end{pmatrix}$ y $B_3 = \begin{pmatrix} 3 & 4 & 2 \\ 8 & 7 & 5 \\ 1 & 1 & 1 \end{pmatrix}$.

Entonces $diag(B_1, B_2, B_3)$ está dada por

$$\begin{bmatrix} 2 & 3 & 0 & 0 & 0 & 0 \\ 4 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 7 & 0 & 0 & 0 \\ 0 & 0 & 0 & 3 & 4 & 2 \\ 0 & 0 & 0 & 8 & 7 & 5 \\ 0 & 0 & 0 & 1 & 1 & 1 \end{bmatrix}.$$

□

Para concluir este capítulo presentamos el concepto de convergencia de una sucesión de matrices, así como algunos resultados que nos serán de utilidad para demostrar teoremas presentados en el siguiente capítulo.

En el estudio de las técnicas iterativas para resolver un sistema de ecuaciones, es de particular importancia saber cuándo la potencia de una matriz tiende a la matriz nula, es decir, cuando todas las entradas se acercan a cero. Matrices de este tipo se llaman convergentes.

Una sucesión de matrices $\{A^{(k)}\} \in \mathcal{M}_n(\mathbb{R})$ se dice que es convergente a $A \in \mathcal{M}_n(\mathbb{R})$ si

$$\lim_{k\to\infty} \|A^{(k)} - A\| = 0.$$

En este caso la elección de la norma no influye, ya que en $\mathcal{M}_n(\mathbb{R})$ todas las normas son equivalentes (ver texto [5]).

Como ya hemos mencionado, en el estudio de los métodos iterativos nos interesa las matrices convergentes para las cuales

$$\lim_{k\to\infty} A^k = 0 \ , \quad \text{donde } 0 \text{ es la matriz nula.}$$

Definición 1.4.4 *Sea $A \in \mathcal{M}_n(\mathbb{R})$, la llamaremos convergente si*

$$\lim_{k\to\infty} (A^k)_{ij} = 0,$$

para cada $i = 1, 2, ..., n$ y cada $j = 1, 2, ..., n$.

Veamos ahora un ejemplo de una matriz convergente.

Ejemplo 1.4.5 Consideremos

$$A = \begin{pmatrix} \frac{1}{2} & 0 \\ \\ \frac{1}{4} & \frac{1}{2} \end{pmatrix}.$$

Notemos que A es convergente, ya que las potencias A^2, A^3 y A^4 son, respectivamente,

$$\begin{pmatrix} \frac{1}{4} & 0 \\ \\ \frac{1}{4} & \frac{1}{4} \end{pmatrix}, \qquad \begin{pmatrix} \frac{1}{8} & 0 \\ \\ \frac{3}{16} & \frac{1}{8} \end{pmatrix}, \qquad \begin{pmatrix} \frac{1}{16} & 0 \\ \\ \frac{1}{8} & \frac{1}{16} \end{pmatrix}.$$

En general, se puede probar por inducción que

$$A^k = \begin{pmatrix} \left(\frac{1}{2}\right)^k & 0 \\ \\ \frac{k}{2^{k+1}} & \left(\frac{1}{2}\right)^k \end{pmatrix}.$$

Podemos ver que A es una matriz convergente ya que

$$\lim_{k\to\infty} \left(\frac{1}{2}\right)^k = 0 \quad y \quad \lim_{k\to\infty} \frac{k}{2^{k+1}} = 0.$$

□

La proposición que se presenta a continuación, nos será útil para demostrar el teorema que sigue, el cual guarda una estrecha relación con el radio espectral de una matriz.

Proposición 1.4.6 *Sea $A \in \mathcal{M}_n(\mathbb{R})$, se tiene que*

$$\lim_{k\to\infty} A^k = 0 \quad \textit{si y solo si} \quad \lim_{k\to\infty} A^k \boldsymbol{y} = 0,$$

para cualquier $\boldsymbol{y} \in \mathbb{R}^n$.

Demostración: Considerando que $\lim_{k\to\infty} A^k = 0$ y que $\|A^k\boldsymbol{y}\| \leq \|A^k\|\|\boldsymbol{y}\|$, así cuando k tiende a ∞ entonces $\|A^k\|\|\boldsymbol{y}\|$ tiende a 0, ya que $\|A^k\|$ tiende a 0.

Por el contrario, si $\lim_{k\to\infty} A^k\boldsymbol{y} = 0$ para todo $\boldsymbol{y} \in \mathbb{R}^n$, entonces $\lim_{k\to\infty} A^k = 0$. Si consideramos $L = \lim_{k\to\infty} A^k$, es decir, que para todo $\boldsymbol{y} \in \mathbb{R}^n$, $L\boldsymbol{y} = \boldsymbol{0}$ si y solo si $\|L\boldsymbol{y}\| = 0$.
Ahora bien, $L\boldsymbol{y} = A^k\boldsymbol{y} - A^k\boldsymbol{y} + L\boldsymbol{y} = (L - A^k)\boldsymbol{y} + A^k\boldsymbol{y}$, por lo tanto,

$$\begin{aligned} 0 &\leq \|L\boldsymbol{y}\| = \|(L - A^k)\boldsymbol{y} + A^k\boldsymbol{y}\| \\ &\leq \|(L - A^k)\boldsymbol{y}\| + \|A^k\boldsymbol{y}\| \\ &\leq \|(L - A^k)\|\|\boldsymbol{y}\| + \|A^k\boldsymbol{y}\|. \end{aligned}$$

Así tenemos $0 \leq \|L\boldsymbol{y}\| \leq \|(L - A^k)\|\|\boldsymbol{y}\| + \|A^k\boldsymbol{y}\|$, luego si k tiende a ∞, entonces $0 \leq \|L\boldsymbol{y}\| \leq \|(L - L)\|\|\boldsymbol{y}\| + 0$, ya que $\|A^k\boldsymbol{y}\|$ por hipótesis tiende a 0.

Luego $0 \leq \|L\boldsymbol{y}\| \leq 0$, entonces $\|L\boldsymbol{y}\| = 0$ si y solo si $L\boldsymbol{y} = 0$, para todo $\boldsymbol{y} \in \mathbb{R}^n$. Así $L = 0$, por lo tanto

$$L = \lim_{k\to\infty} A^k = 0.$$

■

El siguiente teorema relaciona el límite de una sucesión de matrices con el radio espectral de la matriz, que como ya hemos comentado nos permitirá establecer si un método iterativo es convergente, y nos será de utilidad para demostrar teoremas presentados en siguiente capítulo.

Teorema 1.4.7 *Sea $A \in \mathcal{M}_n(\mathbb{R})$, entonces*

$$\lim_{k\to\infty} A^k = 0 \quad \textit{si y solo si} \quad \rho(A) < 1.$$

Demostración: Si $\rho(A) < 1$, entonces existe $\varepsilon > 0$ tal que $\rho(A) < 1 - \varepsilon$. Por lo tanto $\rho(A) + \varepsilon < 1$ y gracias a la Proposición 1.3.16, existe una norma matricial inducida tal que $\|A\| \leq \rho(A) + \varepsilon < 1$. Del hecho de que $\|A^k\| \leq \|A\|^k < 1$ y de la definición de convergencia, se tiene que cuando k tiende a ∞ la sucesión $\{A^k\}$ tiende a 0.

En sentido contrario, considerando que $\lim_{k\to\infty} A^k = 0$ y λ un valor propio de A, se tiene que $A^k\boldsymbol{x} = \lambda^k\boldsymbol{x}$, con $\boldsymbol{x} \neq 0$ un vector propio asociado al valor propio λ. Así el $\lim_{k\to\infty} \lambda^k = 0$. De esta forma, se concluye que $|\lambda| < 1$ y, debido a que λ es un valor propio cualquiera, se deduce que $\rho(A) < 1$. ■

Capítulo 2
Métodos Iterativos

Un método iterativo para resolver un sistema de ecuaciones lineales $A\boldsymbol{x} = \boldsymbol{b}$ comienza con una aproximación inicial $\boldsymbol{x}^{(0)}$, a partir de la cual se construye una sucesión $\{\boldsymbol{x}^{(k)}\}_{k=0}^{\infty}$ de vectores que se espera converja a la solución exacta $\boldsymbol{x} = A^{-1}\boldsymbol{b}$, es decir:

$$\boldsymbol{x} = \lim_{k\to\infty} \boldsymbol{x}^{(k)}. \tag{2.1}$$

Consideramoslos métodos iterativos estacionarios donde la sucesión $\{\boldsymbol{x}^{(k)}\}$ se construye de la forma

$$\boldsymbol{x}^{(k+1)} = B\boldsymbol{x}^{(k)} + \boldsymbol{c}, \qquad \text{para cada } k \geq 0, \tag{2.2}$$

donde $B \in \mathcal{M}_n(\mathbb{R})$, se llama matriz de iteración y $\boldsymbol{c} \in \mathbb{R}^n$ un vector fijo.

Si denotamos por $\boldsymbol{e}^{(k)} = \boldsymbol{x}^{(k)} - \boldsymbol{x}$ el error en el paso k-ésimo de la iteración, la condición de convergencia (2.1), equivale a exigir que

$$\lim_{k\to\infty} \boldsymbol{e}^{(k)} = \boldsymbol{0}.$$

En la pŕactica el proceso se va a detener en el valor mínimo de k tal que $\|\boldsymbol{x}^{(k)} - \boldsymbol{x}\| < \varepsilon$, donde ε es una tolerancia fija dada previamente y $\|\cdot\|$ una norma vectorial seleccionada.

Desafortunadamente el error es tan inaccesible como lo es la solución del sistema, ya que utiliza esta última. Sin embargo podemos calcular otra medida que nos permite observar que tan cerca está la aproximación $\boldsymbol{x}^{(k)}$ de la solución $\boldsymbol{x}$, lo llamaremos residuo y está dado por

$$\boldsymbol{r}^{(k)} = \boldsymbol{b} - A\boldsymbol{x}^{(k)},$$

y simplemente es la cantidad por la que el vector $\boldsymbol{x}^{(k)}$ deja de satisfacer el sitema $A\boldsymbol{x} = \boldsymbol{b}$. Como el residuo es un vector, al igual que el error, puede ser medido con cualquier norma. También notemos que $\boldsymbol{r}^{(k)} = 0$ si y solo si $\boldsymbol{e}^{(k)} = 0$.

Ahora reescribiendo la ecuación del residuo de la forma, $A\boldsymbol{x}^{(k)} = \boldsymbol{b} - \boldsymbol{r}$, luego la restamos a $A\boldsymbol{x} = \boldsymbol{b}$, tenemos

$$A\boldsymbol{e}^{(k)} = A(\boldsymbol{x} - \boldsymbol{x}^{(k)}) = A\boldsymbol{x} - A\boldsymbol{x}^{(k)} = \boldsymbol{b} - (\boldsymbol{b} - r^{(k)}) = \boldsymbol{r}^{(k)}$$

Esto lo podemos hacer ya que el error es la distancia entre dos vectores la cual siempre es positiva. La ecuación $A\boldsymbol{e}^{(k)} = \boldsymbol{r}^{(k)}$, se llama ecuación residual, y dice que el error satisface el mismo conjunto de ecuaciones que $\boldsymbol{x}$, cuando reeplazamos $\boldsymbol{b}$ por $\boldsymbol{r}^{(k)}$.

2.1. Convergencia General de Métodos

A continuación se presentan algunos resultados de convergencia general de los métodos iterativos.

Definición 2.1.1 *Un método iterativo de la forma* (2.2) *se dice que es consistente con* $A\boldsymbol{b} = \boldsymbol{x}$ *si* $\boldsymbol{c}$ *y* B *son tales que* $\boldsymbol{x} = B\boldsymbol{x} + \boldsymbol{c}$*. De manera equivalente*

$$\boldsymbol{c} = (I - B)A^{-1}\boldsymbol{b}.$$

La consistencia por si sola no es suficiente para asegurar la covergencia del método iterativo (2.2). Por lo tanto debemos considerar el siguiente teorema.

Teorema 2.1.2 *Sea* (2.2) *un método consistente. La sucesión de vectores* $\{\boldsymbol{x}^{(k)}\}$ *converge a la solución de* $A\boldsymbol{x} = \boldsymbol{b}$ *para cualquier elección de* $\boldsymbol{x}^{(0)}$ *si y solo si* $\rho(B) < 1$*.*

Demostración: A partir de que $\boldsymbol{e}^{(k)} = \boldsymbol{x}^{(k)} - \boldsymbol{x}$ y la suposición de consitencia, tenemos que, si a (2.2) le restamos $\boldsymbol{x} = B\boldsymbol{x} + \boldsymbol{c}$, obtenemos la siguente relación recursiva.

$$\boldsymbol{e}^{(k+1)} = \boldsymbol{x}^{(k+1)} - \boldsymbol{x} = B\boldsymbol{x}^{(k)} - B\boldsymbol{x} + \boldsymbol{c} - \boldsymbol{c} = B(\boldsymbol{x}^{(k)} - \boldsymbol{x}) = B\boldsymbol{e}^{(k)}.$$

Por lo tanto, $\boldsymbol{e}^{(k)} = B^k\boldsymbol{e}^{(0)}$. Así gracias a la Proposición 1.4.6 se deduce que, si $\boldsymbol{e}^{(k)} = 0$ el $\lim_{k\to\infty} B^k\boldsymbol{e}^{(0)} = 0$ si y solo si $\lim_{k\to\infty} B^k = 0$ y por Teorema 1.4.7 el $\lim_{k\to\infty} B^k = 0$ si y solo si $\rho(A) < 1$.

Ahora en sentido contrario, lo haremos por contradicción suponiendo que $\rho(A) < 1$, entonces existe al menos un valor propio λ de B con módulo superior a 1. Sea $\boldsymbol{e}^{(0)}$ un vector propio asociado a λ, entonces $B\boldsymbol{e}^{(0)} = \lambda\boldsymbol{e}^{(0)}$ y por lo tanto $\boldsymbol{e}^{(k)} = \lambda^k\boldsymbol{e}^{(0)}$ y como consecuencia, $\boldsymbol{e}^{(k)}$ no puede tender a cero cuando k tiende a ∞, ya que $|\lambda| > 1$. ■

Calcular el radio espectral de una matriz será más complejo mientras mayor sea el tamaño de esta. La proposición siguiente relaciona los resultados de los Teoremas 2.1.2 y 1.3.13. De este modo podremos encontrar una forma más factible de verificar método iterativo converge.

Proposición 2.1.3 *Sea* (2.2) *un método consistente. Una condición suficiente para la convergencia de la sucesión de vectores* $\{\boldsymbol{x}^{(k)}\}$ *construida usando* (2.2)*, con cualquier elección de* $\boldsymbol{x}^{(0)}$*, es que* $\|B\| < 1$*, para cualquier norma matricial.*

Demostración: Por el Teorema 2.1.2 tenemos que una sucesión converge para cualquier elección de $\boldsymbol{x}^{(0)}$ si y solo si $\rho(B) < 1$ y por Teorema 1.3.13 tenemos que $\rho(B) \leq \|B\|$, con $\|\cdot\|$ una norma consistente. Así que una condición suficiente para que una sucesión de vectores converja es que

$$\|B\| < 1.$$

■

2.2. Métodos de Jacobi, Gauss-Seidel y SOR

Una técnica para construir métodos iterativos lineales consistentes se basa en la división de la matriz A, de la forma $A = M - N$, con M y N matrices adecuadas y M no singular (M se llama precondicionador).

Partiendo de $A\boldsymbol{x} = \boldsymbol{b}$, y dado $\boldsymbol{x}^{(0)}$, se puede calcular $\boldsymbol{x}^{(k+1)}$ para $k \geq 0$, resolviendo el sistema,

$$M\boldsymbol{x}^{(k+1)} = N\boldsymbol{x}^{(k)} + \boldsymbol{b}, \tag{2.3}$$

donde la matriz de iteración es dada por $B = M^{-1}N$ y $\boldsymbol{c} = M^{-1}\boldsymbol{b}$.

Aquí podemos notar el porqué M debe ser no singular y además ser facilmente invertible, con el fin de mantener un bajo coste computacional. Ahora si $M = A$ y $N = 0$ el método (2.3) sería convergente en una iteración, pero con el mismo coste que un método directo.

Proposición 2.2.1 *Sea $A = M - N$, con A y M simétrica y definida positiva. Si la matriz $2M - A$ es definida positiva, entonces el método iterativo definido en* (2.3) *es convergente para cualquier elección del dato inicial $\boldsymbol{x}^{(0)}$ y*

$$\rho(B) = \|B\|_A = \|B\|_M < 1.$$

Demostración: Las siguientes matrices son similares a la matriz de iteración B.

$$B' = A^{1/2}BA^{-1/2} = I - A^{1/2}M^{-1}A^{1/2},$$

$$B'' = M^{1/2}BM^{-1/2} = I - M^{1/2}AM^{1/2},$$

esto implica que $\rho(B) = \rho(B') = \rho(B'')$. Como B' y B'' son simétricas, $\rho(B') = \|B'\|_2 = \|B\|_A$ y $\rho(B'') = \|B''\|_2 = \|B\|_M$. Así entonces,

$$\rho(B) = \|B\|_A = \|B\|_M < 1.$$

■

Proposición 2.2.2 *Sea $A = M - N$, con A simétrica y definida positiva. Si la matriz $M + M^T - A$ es definida positiva, entonces M es invertible, el método iterativo* (2.3) *converge monotónicamente con respecto a la norma $\|\cdot\|_A$ y $\rho(B) \leq \|B\|_A < 1$.*

Que un método iterativo converja de forma monótona, siginifica que el error de aproximación decrece en todo momemento, es decir

$$\|\boldsymbol{e}^{(k)}\| \leq \|\boldsymbol{e}^{(k-1)}\|.$$

Demostración: Para demostrar la regularidad de M, debemos asumir que $M\boldsymbol{x} = 0$,

$$0 = \langle M\boldsymbol{x}, \boldsymbol{x}\rangle + \langle \boldsymbol{x}, M\boldsymbol{x}\rangle = \left\langle (M + M^T)\boldsymbol{x}, \boldsymbol{x}\right\rangle,$$

implica que $\boldsymbol{x} = 0$, ya que $M + M^T > 0$, por lo tanto M es invertible.

Por el Teorema 1.3.13 tenemos que $\rho(B) \leq \|B\|_A$, entonces solo falta mostrar $\|B\|_A < 1$. Por (1.10) tenemos que $\|B\|_A = \|A^{1/2}BA^{-1/2}\|_2 = \|\hat{B}\|_2$, para $\hat{B} = I - A^{1/2}MA^{1/2}$. Se verifica que

$$\begin{aligned} \hat{B}^T\hat{B} &= (I - A^{1/2}M^{-T}A^{1/2})(I - A^{1/2}M^{-1}A^{1/2}) \\ &= I - A^{1/2}(M^{-T} + W^{-1})A^{1/2} + A^{1/2}M^{-T}AM^{-1}A^{1/2} \\ &= I - A^{1/2}M^{-T}(M + M^T)M^{-1}A^{1/2} + A^{1/2}M^{-T}AM^{-1}A^{1/2} \\ &< I - A^{1/2}M^{-T}AM^{-1}A^{1/2} + A^{1/2}M^{-T}AM^{-1}A^{1/2} = I. \end{aligned}$$

Entonces por la Proposición 1.3.15

$$\|B\|_A = \|\hat{B}\|_2 = \rho(\hat{B}^H\hat{B})^{1/2} < \rho(I)^{1/2} = 1.$$

■

Las proposiciones anteriores serán de mucha utilidad al momento de demostrar si un método iterativo propuesto es convergente.

A continuación presentamos tres métodos iterativos estacionarios particulares y algunos resultados de convergencia para cada uno de ellos.

Método de Jacobi

En el método de Jacobi la división de $A = M - N$ es la siguiente

$$M = D, \quad N = D - A = L + U,$$

donde D es la matriz diagonal principal de A, L es una matriz triangular inferior de entradas $l_{ij} = -a_{ij}$ si $i > j$, $l_{ij} = 0$ si $i \leq j$ y U es una matriz triangular superior de entradas $u_{ij} = -a_{ij}$ si $j > i$, $u_{ij} = 0$ si $j \leq i$.

Entonces el método de Jacobi escrito en forma matricial está dado por

$$D\boldsymbol{x}^{(k+1)} = M\boldsymbol{x}^{(k+1)} = N\boldsymbol{x}^{(k)} + \boldsymbol{b} = (L + U)\boldsymbol{x}^{(k)} + \boldsymbol{b},$$

o de forma equivalente

$$\boldsymbol{x}^{(k+1)} = D^{-1}(L + U)\boldsymbol{x}^{(k)} + D^{-1}\boldsymbol{b}, \tag{2.4}$$

donde la matriz de iteración B_J y el vector $\boldsymbol{c}$ del método de Jacobi son, respectivamente

$$B_J = D^{-1}(L + U) = I - D^{-1}A, \qquad \boldsymbol{c} = D^{-1}\boldsymbol{b}.$$

También podemos encontrar una ecuación para calcular cada componente del vector de aproximación. En este caso, las componentes del vector $\boldsymbol{x}^{(k+1)}$ se calculan mediante la ecuación

$$x_i^{(k+1)} = \frac{1}{a_{ii}}\left[b_i - \sum_{j=1, j\neq i}^{n} a_{ij}x_j^{(k)}\right], \qquad \text{para cada } i \in \{1, ..., n\}. \tag{2.5}$$

Método de Gauss-Seidel

Este método se diferencia al de Jacobi en que en la etapa $(k+1) - ésima$ los valores disponibles de $\boldsymbol{x}^{(k+1)}$ se utilizan para actualizar la solución. La división de $A = M - N$ correpondiente es

$$M = D - L, \quad N = U,$$

donde D es la matriz diagonal principal de A, L es una matriz triangular inferior y U es una matriz triangular superior.

De esta forma el método de Gauss-Seidel escrito en forma matricial es dado por

$$(D-L)\boldsymbol{x}^{(k+1)} = M\boldsymbol{x}^{(k+1)} = N\boldsymbol{x}^{(k)} + \boldsymbol{b} = U\boldsymbol{x}^{(k)} + \boldsymbol{b}$$

o de forma equivalente

$$\boldsymbol{x}^{(k+1)} = (D-L)^{-1}U\boldsymbol{x}^{(k)} + (D-L)^{-1}\boldsymbol{b}, \tag{2.6}$$

donde la matriz de iteración asociada B_{GS} y el vector $\boldsymbol{c}$ son, respectivamente

$$B_{GS} = (D-L)^{-1}U, \qquad \boldsymbol{c} = (D-L)^{-1}\boldsymbol{b}.$$

Al igual que en el método de Jacobi, podemos encontrar la ecuación para calcular las componentes del vector de aproximación, $x_i^{(k+1)}$ en este caso se calcula mediante

$$x_i^{(k+1)} = \frac{1}{a_{ii}}\left[b_i - \sum_{j=1}^{i-1} a_{ij}x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij}x_j^{(k)}\right], \qquad \text{para cada } i \in \{1,...,n\}. \tag{2.7}$$

En este método, estamos asumiendo que los componentes del vector $\boldsymbol{x}^{(k+1)}$ son actualizados hacia adelante. Pero también podemos formular otra versión donde los componentes del vector de aproximación son actualizados hacia atrás. La forma matricial de esta versión es la siguiente

$$M = D - U, \quad N = L.$$

De esta forma el método es dado por

$$(D-U)\boldsymbol{x}^{(k+1)} = M\boldsymbol{x}^{(k+1)} = N\boldsymbol{x}^{(k)} + \boldsymbol{b} = L\boldsymbol{x}^{(k)} + \boldsymbol{b}$$

o de forma equivalente

$$\boldsymbol{x}^{(k+1)} = (D-U)^{-1}L\boldsymbol{x}^{(k)} + (D-U)^{-1}\boldsymbol{b}, \tag{2.8}$$

donde la matriz de iteración asociada $B_{GS(2)}$ y el vector $\boldsymbol{c}$ son, respectivamente

$$B_{GS(2)} = (D-U)^{-1}L, \qquad \boldsymbol{c} = (D-U)^{-1}\boldsymbol{b}.$$

Método de Sobre-Relajación Sucesiva (SOR)

Una generalización del método de Gauss-Seidel es el método de sobre-relajación sucesiva (SOR), en el que después de introducir un parámetro w de relajación en la ecuación (2.6), se sustituye por la siguiente

$$\boldsymbol{x}^{(k+1)} = (D - wL)^{-1}[(1-w)D + wU]\boldsymbol{x}^{(k)} + w(D - wL)^{-1}\boldsymbol{b}. \tag{2.9}$$

En este caso la matriz de iteración $B(w)$ y el vector $\boldsymbol{c}$ están son, respectivamente

$$B(w) = (D - wL)^{-1}[(1-w)D + wU], \qquad \boldsymbol{c} = w(D - wL)^{-1}\boldsymbol{b}.$$

Este método es consistente para cualquier $w \neq 0$. Cuando $w = 1$ coincide con el método de Gauss-Seidel. En particular cuando $0 < w < 1$ el método se llama sub-relajación mientras que si $w > 1$ se llama sobre relajación.

Tal como en los métodos anteriores, para encontrar las componentes del vector de aproximación, debemos utilizar la siguiente igualdad

$$x_i^{(k+1)} = \frac{w}{a_{ii}}\left[b_i - \sum_{j=1}^{i-1} a_{ij}x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij}x_j^{(k)}\right] + (1-w)x_i^{(k)}, \qquad \text{para cada } i \in \{1, ..., n\}. \tag{2.10}$$

Presentamos un ejemplo que muestra cómo funcionan los método de Jacobi, de Gauss-Seidel y SOR presentaremos el siguiente ejemplo. Los cálculos y resultados se obtuvieron utilizando el software matemático MATLAB. Haremos notar que los resultados estarán aproximados con 4 decimales.

Ejemplo 2.2.3 Consideremos el sistema lineal dado por

$$\begin{array}{rcrcrcrcrcrcl}
10x_1 & + & 3x_2 & + & x_3 & + & 4x_4 & + & x_5 & & & = & 1 \\
3x_1 & - & 10x_2 & + & x_3 & - & x_4 & + & 2x_5 & - & x_6 & = & 1 \\
x_1 & + & 3x_2 & + & 10x_3 & - & x_4 & - & 2x_5 & + & x_6 & = & 1 \\
-3x_1 & - & x_2 & + & x_3 & + & 10x_4 & + & 2x_5 & - & 2x_6 & = & 1 \\
x_1 & + & x_2 & + & x_3 & - & 2x_4 & - & 10x_5 & + & 3x_6 & = & 1 \\
2x_1 & - & x_2 & & & + & 3x_4 & + & x_5 & + & 10x_6 & = & 1
\end{array}$$

Este sistema cuenta como solución única

$$\boldsymbol{x} = \begin{pmatrix} 0{,}0779 \\ -0{,}1041 \\ 0{,}1112 \\ 0{,}1317 \\ -0{,}1043 \\ 0{,}0450 \end{pmatrix}.$$

Consideraremos $\boldsymbol{x}^{(0)} = (0,0,0,0,0,0)$ y el proceso de iteración se detendrá cuando el error relativo sea menor que 10^{-3}, es decir,

$$\frac{\|\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)}\|_\infty}{\|\boldsymbol{x}^{(k+1)}\|_\infty} < 10^{-3}.$$

En primer lugar, para encontrar la solución usaremos el método de Jacobi. La solución se encuentra en la iteración número 11, y la solución arrojada es

$$\boldsymbol{x}^{(11)} = \begin{pmatrix} 0{,}0778 \\ -0{,}1041 \\ 0{,}1112 \\ 0{,}1317 \\ -0{,}1044 \\ 0{,}0450 \end{pmatrix}.$$

La aproximación de la solución encontrada depende claramente de la tolerancia del error relativo, este también infuye en la cantidad de iteraciones que se deben realizar antes de encontar la solución.

Ahora usaremos el método de Gauss-Seidel. La solución se encuentra en la iteración número 6, y la solución arrojada es

$$\boldsymbol{x}^{(6)} = \begin{pmatrix} 0{,}0779 \\ -0{,}1040 \\ 0{,}1112 \\ 0{,}1317 \\ -0{,}1043 \\ 0{,}0449 \end{pmatrix}.$$

También utilizaremos Gauss-Seidel hacia atrás. La solución se encuentra en la iteración número 6, y la solución arrojada es

$$\boldsymbol{x}^{(6)} = \begin{pmatrix} 0{,}0778 \\ -0{,}1040 \\ 0{,}1112 \\ 0{,}1317 \\ -0{,}1043 \\ 0{,}0450 \end{pmatrix}.$$

Podemos notar en este ejemplo, que el método de Gauss-Seidel en sus dos formas es más rápido que Jacobi, ya que se necesitan menos iteraciones, la efectividad es la misma ya que las soluciones aproximadas son muy cercanas a la solución del sistema.

Para finalizar usaremos el método SOR. Para este ejemplo utilizaremos $w = \frac{1}{2}$ y $w = \frac{3}{4}$. La solución se encuentra en la iteración número 12 y la iteración número 7 respectivamente

y la solución arrojada son respectivamente

$$\boldsymbol{x}^{(12)} = \begin{pmatrix} 0{,}0778 \\ -0{,}1039 \\ 0{,}1112 \\ 0{,}1316 \\ -0{,}1042 \\ 0{,}0450 \end{pmatrix} \quad y \quad \boldsymbol{x}^{(7)} = \begin{pmatrix} 0{,}0778 \\ -0{,}1040 \\ 0{,}1112 \\ 0{,}1316 \\ -0{,}1043 \\ 0{,}0450 \end{pmatrix}.$$

Notemos que a pesar que para $w = \frac{1}{2}$ el método es más lento que el de Jacobi y que el de Gauss-Seidel, ya en con $w = \frac{3}{4}$ la cantidad de iteraciones disminuye lo que nos hace pensar en la idea de un valor para w con el cual el método SOR funcione de manera óptima. □

Es necesario destacar que, en este ejemplo si ahora aumentamos la tolerancia para el error relativo a 10^{-6}, el método de Gauss-Seidel hacia atrás es más rápido que el mismo método hacia adelante, usando respectivamente 11 y 14 iteraciones.

2.2.1. Convergencia de los métodos de Jacobi, Gauss-Seidel y SOR

La convegencia de los métodos iterativos, como ya hemos mencionado anteriormente, no la podemos asegurar para todas la matrices, por lo que a continuación se presentan algunos resultados de convergencia para algunas matrices especiales de los métodos de Jacobi, Gauss-Seidel y SOR.

El primer resultado importante a considerar, será la convergencia para una matriz estrictamente diagonal dominante.

Teorema 2.2.4 *Si A es una matriz estrictamente diagonal dominante por filas, los métodos de Jacobi y Gauss-Seidel son convergentes.*

Demostración: Veamos la demostración para el método de Jacobi. Dado que A es estrictamente diagonal dominante por filas se tiene que

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

De aquí obtenemos que

$$\sum_{j=1, j\neq i}^{n} \frac{|a_{ij}|}{|a_{ii}|} < 1, \qquad i = 1, ..., n.$$

Ahora, sea B_J la matriz de iteración del método de Jacobi entonces por la Proposición 2.1.3, si $\|B_J\| < 1$ el método converge, para cualquier norma consistente. Consideremos la norma

$$\|A\|_\infty = \max_{1\leq i\leq n} \sum_{j=1}^{n} |a_{ij}|.$$

Ahora, tenemos que $B_J = (b_{ij})$ tiene entradas $b_{ij} = -\frac{a_{ij}}{a_{ii}}$ cuando $i \neq j$, y $b_{ij} = 0$ cuando $i = j$, es decir

$$B_J = \begin{pmatrix} 0 & -\frac{a_{12}}{a_{11}} & \cdots & -\frac{a_{1n}}{a_{11}} \\ -\frac{a_{21}}{a_{22}} & 0 & \cdots & -\frac{a_{2n}}{a_{22}} \\ \vdots & \vdots & \ddots & \vdots \\ -\frac{a_{n1}}{a_{nn}} & -\frac{a_{n2}}{a_{nn}} & \cdots & 0 \end{pmatrix}.$$

Entonces por la hipótesis el método de Jacobi converge ya que

$$\|B_J\|_\infty = \max_{1\leq i\leq n} \sum_{j=1}^{n} \frac{|a_{ij}|}{|a_{ii}|} < 1.$$

■

La demostración para Gauss-Seidel, como lo menciona el texto [5], se encuentra en el texto *Iterative Solution Methods* (Axelsson O. 1994).

Otras matrices especiales son las matrices simétricas y las matrices definidas positivas, los dos teoremas siguientes muestran la convergencia de los métodos de Jacobi y de Gauss-Seidel, cuando una matriz cuenta con estas dos características.

Teorema 2.2.5 *Si A y $2D - A$ son matrices simétricas y definidas positivas entonces el método de Jacobi es convergente y $\rho(B_J) = \|B\|_A = \|B\|_D$.*

Demostración: Podemos utilizar la Proposición 2.2.1, considerando $M = D$. ■

Teorema 2.2.6 *Si A es simétrica definida positiva, el método Gauss-Seidel converge monotónicamente con respecto a la norma $\|\cdot\|_A$.*

Demostración: Usando la Proposición 2.2.2, aplicada a la matriz $M = D - L$. En primer lugar, observemos que $(D - L)^T = D - U$, ya que A es simétrica, así $M^T = D - U$. Como $A = D - L - U$, entonces

$$M + M^T - A = D - L + D - U - A = 2D - L - U - A = D.$$

Así $M + M^T - A = D$ es definida positiva, entonces el método de Gauss-Seidel converge monotónicamente con respecto a la norma $\|\cdot\|_A$. ■

El teorema que sigue establece una relación de convergencia entre los métodos de Jacobi y Gauss-Seidel. Pero antes presentaremos un lema, que se encuentra en el texto [6], que será de utilidad para demostrar el teorema.

Lema 2.2.7 *Sea A una matriz tridiagonal por bloques de oreden N.*

$$A = \begin{bmatrix} A_{11} & A_{12} & 0 & \cdots & 0 \\ A_{21} & A_{22} & A_{23} & \ddots & \vdots \\ 0 & A_{32} & A_{33} & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & A_{p-1p} \\ 0 & \cdots & 0 & A_{pp-1} & A_{pp} \end{bmatrix},$$

donde cada bloque A_{ii} es una matriz de orden n_i y $\sum_{i=1}^{p} n_i = N$. Sea ahora la descomposición por bloques $A = D - L - U$, donde D es la diagonal de bloques

$$\begin{bmatrix} A_{11} & 0 & \cdots & 0 \\ 0 & A_{22} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & A_{pp} \end{bmatrix}$$

y donde $-L$ y $-U$ son la parte triangular inferior y superior respectivamente excluyendo la diagonal. Sea μ un número no nulo y llamemos $A(\mu) = D - \mu L - \frac{1}{\mu}U$ entonces

$$\det(A(\mu)) = \det(A).$$

Teorema 2.2.8 *Sea A una matriz tridiagonal con bloques diagonales no singulares. Entonces se cumple que:*

$$\rho(B_{GS}) = [\rho(B_J)]^2. \tag{2.11}$$

Demostración: En primer lugar tenemos la matriz de iteración del método de Jacobi $B_J = D^{-1}(L+U)$ y sus valores propios son las raíces del polinomio

$$P(\lambda) = \det(\lambda I - D^{-1}(L+U)) = 0,$$

y también

$$\det(D)P(\lambda) = \det(D)\det(\lambda I - D^{-1}(L+U)) = \det(\lambda D - L - U) = 0,$$

por lo tanto los valores propios de B_J son la raíces de la ecuación

$$\det(\lambda D - L - U) = 0. \tag{2.12}$$

Ahora, la matriz de iteración del método de Gauss-Seidel es $B_{GS} = (D-L)^{-1}U$ y sus valores propios son las raíces del polinomio

$$P_{GS}(\lambda) = \det(\lambda I - (D-L)^{-1}U) = 0,$$

y también

$$\det(D-L)P_{GS}(\lambda) = \det(D-L)\det(\lambda I - (D-L)^{-1}U) = \det(\lambda D - \lambda L - U) = 0,$$

de donde

$$\det(D-L)P_{GS}(\lambda) = \det\left(\sqrt{\lambda}\left[\sqrt{\lambda}D - \sqrt{\lambda}L - \frac{1}{\sqrt{\lambda}}\right]\right) = 0,$$

es decir los valores propios de B_{GS} son las raíces de la ecuación

$$(\sqrt{\lambda})^N \det(\sqrt{\lambda}D - L - U) = 0, \tag{2.13}$$

donde hemos utilizado el lema anterior.

Ahora comparando (2.12) y (2.13), notaremos que si λ es un valor propio de B_{GS} entonces $\pm\sqrt{\lambda}$ lo es de B_J y que si α es valor propio de B_J, α^2 lo es de B_{GS}, entonces

$$\rho(B_{GS}) = [\rho(B_J)]^2.$$

■

La importancia de la igualdad (2.11) radica en que se puede concluir que tanto el método de Gauss-Seidel como el método de Jacobi convergen o no convergen simultáneamente. Un resultado particular es que si A es tridiagonal, simétrica y definida positiva por el Teorema 2.2.6 el método de Gauss-Seidel converge y por (2.11) también podemos asegurar la convergencia del método de Jacobi. También, en (2.11) podemos notar que el método de Gauss-Seidel converge más rápido que el método de Jacobi, ya que la velocidad de convergencia está asociada al radio espectral.

Ahora consideraremos una definición que nos será de utilidad para asegurar convergencia y la velocidad de convergencia del método SOR.

Definición 2.2.9 *Una matriz $M \in \mathcal{M}_n(\mathbb{R})$ (es decir, una matriz tal que $\alpha D^{-1}L + \alpha^{-1}D^{-1}U$, para $\alpha \neq 0$,tiene valores propios que no dependen de α. Donde $M = D - L - U$, con $D = diag(m_{11}, ..., m_{nn})$, L y U son matrices triangulares estrictamente inferior y superior respectivamente), disfruta de la A-propiedad si puede ser dividida en la forma de bloques 2×2.*

$$M = \begin{bmatrix} \widetilde{D}_1 & M_{12} \\ M_{21} & \widetilde{D}_2 \end{bmatrix},$$

donde $\widetilde{D}_1$ y $\widetilde{D}_2$ son matrices diagonales.

También, para estudiar la convergencia del método SOR vamos a analizar el radio espectral de la matriz $B(w)$.

Teorema 2.2.10 *Para cualquier $w \in \mathbb{R}$ tenemos que $\rho(B(w)) \geq |1 - w|$; por lo tanto, el método SOR converge solo si $w \in (0, 2)$.*

Demostración: Para la matriz $B(w)$ se tiene que

$$\begin{aligned} B(w) &= (D - wL)^{-1}[(1-w)D + wU] \\ &= [D(I - wD^{-1}L)]^{-1}[(1-w)D + wU] \\ &= (I - wD^{-1}L)^{-1}D^{-1}[(1-w)D + wU] \\ &= (I - wD^{-1}L)^{-1}[(1-w)I + wD^{-1}U]. \end{aligned}$$

Por lo tanto

$$\det(B(w)) = \det(I - wD^{-1}L)^{-1} \det[(1-w)I + wD^{-1}U].$$

Como $D^{-1}L$ es una matriz tringular inferior con ceros en la diagonal principal, tenemos que

$$\det(I - wD^{-1}L) = 1,$$

por lo tanto el determinante de $(I - wD^{-1}L)^{-1}$ también será igual a 1. Por otra parte la matriz $(1-w)I + wD^{-1}U$ es a su vez una matriz triangular superior y los elementos de su diagonal principal son todos iguales a $1 - w$. Así

$$\det[(1-w)I + wD^{-1}U] = (1-w)^n.$$

Ahora considerando que el determinante de una matriz es el producto de los valores propios de esa matriz y denotando por λ_i a los valores propios de la matriz $B(w)$, tenemos que

$$\prod_{i=1}^{n} \lambda_i = (1-w)^n.$$

De la definición de radio espectral se tiene que

$$\rho(B(w)) \geq |\lambda_i|, \qquad i = 1, ..., n.$$

De esta manera tenemos que

$$[\rho(B(w))]^n \geq \prod_{i=1}^{n} |\lambda_i| = |1-w|^n.$$

Así tomando raiz n-ésima resulta la desigualdad

$$\rho(B(w)) \geq |1-w|.$$

Ahora bien, por Teorema 2.1.2 sabemos que si $\rho(B(w)) < 1$ el método converge. Por lo tanto, utilizando la desigualdad anterior

$$|w - 1| < 1,$$

siendo w un parámetro real. Esto equivale a decir que

$$0 < w < 2.$$

■

Particularmente, el método SOR para algunas matrices específicas, cumple la siguiente proposición.

Proposición 2.2.11 *Si A es simétrica y definida positiva, entonces el método SOR es convergente si y solo si $0 < w < 2$. Por otra parte su convergencia es monótona con respecto a $\|\cdot\|_A$. Finalmente, si A es estríctamente diagonal dominante por filas el método SOR converge si $0 < w \leq 1$.*

Podemos notar que el método SOR converge dependiendo el valor del parametro w y que de la elección de este, como mencionamos en el Ejemplo 2.2.3, depende la rapidez con que el método converge. Entonces es lógico que nos hagamos la pregunta de cuál es el valor óptimo de w, w_{opt}, con el cual la velocidad de convergencia es la más alta posible. No podemos asegurar un valor de w_{opt} en general, pero presentaremos resultados para algunos casos especiales. La siguiente proposición se encuentra en el texto [6] en las páginas 145 y 146.

Proposición 2.2.12 *Si la matriz A disfruta de la A-propiedad y si B_J tiene valores propios reales, el método SOR converge para cualquier elección de $\boldsymbol{x}^{(0)}$ si y solo si $\rho(B_J) < 1$ y $0 < w < 2$. Además*

$$w_{opt} = \frac{2}{1 + \sqrt{1 - [\rho(B_J)]^2}}. \tag{2.14}$$

En el texto [2] mencionan, para que se cumpla (2.14), que la matriz A debe ser definida positiva y tridiagonal. Con esta elección de w se tiene que

$$\rho(B(w)) = w - 1. \tag{2.15}$$

La importancia de esta proposición es clara, ya que ahora contamos con un valor para w que mencionabamos en el Ejemplo 2.2.3 y que podremos usar en el siguiente capítulo donde realizaremos algunos cálculos para comprobar los teoremas anteriores y el funcionamiento de los métodos en estudio, utilizando matrices de mayor tamaño.

Para ilustrar la Proposición 2.2.12 presentaremos el siguiente ejemplo.

Ejemplo 2.2.13 Para calcular w_{opt} consideraremos la siguiente matriz

$$A = \begin{pmatrix} 5 & 2 & 0 \\ 2 & 5 & 1 \\ 0 & 1 & 5 \end{pmatrix}.$$

Notemos que la matriz A es definida positiva y tridiagonal, por lo que para encontrar w_{opt}, debemos calcular el radio espectral de B_J, $\rho(B_J)$. Así tenemos que

$$B_J = \begin{pmatrix} 0 & \frac{2}{5} & 0 \\ \frac{2}{5} & 0 & \frac{1}{5} \\ 0 & \frac{1}{5} & 0 \end{pmatrix},$$

de donde $\rho(B_J) = 0,4472$. Así tenemos que

$$w_{opt} = \frac{2}{1 + \sqrt{1 - (0,4472)^2}} \approx 1,0557.$$

Entonces el radio espectral de la matriz de iteración del método SOR es $\rho(B(w)) = 0,0557$. □

Aprovecharemos el ejemplo para comprobar lo que dice el Teorema 2.2.8. El radio espectral de la matriz de iteración del método Gauss-Seidel es $\rho(B_{GS}) = 0,2000$ y haremos notar que,

$$[\rho(B_J)]^2 = (0,4472)^2 \approx 0,1999.$$

Capítulo 3
Resultados Numéricos

En este capítulo presentamos ejemplos que permiten analizar numéricamente el comportamiento de los métodos iterativos estudiados, y comprobar los teoremas del capítulo anterior. Los cálculos se realizarán mediante el software MATLAB, que es una herramienta que ofrece un entorno de desarrollo integrado con un lenguaje de programación propio. La configuración del computador utilizado es:

- Notebook HP Pavilion g4.
- Procesador Inter(R) Core(TM) i3-3110M CPU @ 2.40 GHz 2.40 GHz.
- RAM 8 GB.
- Sistema operativo de 64 bits, procesador x64.
- Windows 8.1 Single Language.

Para comenzar, como ya hemos comentado y estudiado, en un método iterativo a partir de una aproximación inicial $\boldsymbol{x}^{(0)}$, se construye una secesión de vectores $\{\boldsymbol{x}^{(k)}\}_{k=0}^{\infty}$ que esperamos converja a la solución exacta de un sistema de ecuaciones $A\boldsymbol{x} = \boldsymbol{b}$. Al aplicar un método iterativo para resolver un sistema lineal de ecuaciones obtendremos una sucesión de vectores que se aproximan a la solución del sistema. Este proceso tendrá una velocidad que depende del método que utilicemos y como siempre, existirá un error de aproximación.

En este proceso de aproximación es posible que necesitemos un gran número de iteraciones lo que no es deseable en el momento de hacer los cálculos, ya que el tiempo que se empleará puede ser excesivo. Por esta razón, el proceso se detendrá en aquel valor de k tal que,

$$\|\boldsymbol{x}^{(k)} - \boldsymbol{x}\| < \varepsilon, \tag{3.1}$$

donde ε es una tolerancia fija, cercana a cero y $\|\cdot\|$ es una norma vectorial. Es decir, donde la norma del error absoluto sea menor que la tolerancia ε.

Ahora, considerando que lo que buscamos es el vector $\boldsymbol{x}$ solución del sistema de ecuaciones, entonces notaremos que, en general, no contamos con $\boldsymbol{x}$ en la ecuación (3.1) para que el proceso de iteraciones se detenga. Entonces una forma de saber cuándo detener el proceso

de iteraciones es calculando la norma del error absoluto entre $\boldsymbol{x}^{(k)}$ y $\boldsymbol{x}^{(k-1)}$ y que eso sea menor que una tolerancia prescrita, es decir

$$\|\boldsymbol{x}^{(k)} - \boldsymbol{x}^{(k-1)}\| < \varepsilon. \tag{3.2}$$

Como lo que se pretende es encontrar la mejor aproximación a $\boldsymbol{x}$, el error absoluto lo reemplazaremos por el error relativo, debido a que el error absoluto en varias situaciones no nos permite decidir si una aproximación es mejor que otra, lo que si es posible establecer con la norma del error relativo, tal como dice en el texto [2], como una medida de precisión, el error absoluto puede ser engañoso y el error relativo más significativo, debido a que el error relativo toma en consideración el tamaño del valor. Por esto en lugar de (3.2) usaremos

$$\frac{\|\boldsymbol{x}^{(k)} - \boldsymbol{x}^{(k-1)}\|}{\|\boldsymbol{x}^{(k)}\|} < \varepsilon, \tag{3.3}$$

donde ε sique siendo una tolerancia fija. Como en $\mathbb{R}^n$ todas la normas son equivalentes podemos utilizar cualquier norma, pero en este estudio usaremos la norma l_∞.

A continuación se presenta un ejemplo sencillo que muestra la diferencia entre el error absoluto y el error relativo.

Ejemplo 3.0.14 *Consideremos las aproximaciones*

(i) 25 *a* 30

(ii) 95 *a* 100

Notemos que el error absoluto en ambos casos es 5. Pero si tomamos el error relativo, en (i) es igual a $\frac{1}{6}$ y en (ii) es igual a $\frac{1}{20}$.

Podemos ver que en primera instancia las aproximaciones son iguales, pero al calcular el error relativo, que considera el tamaño del valor, nos damos cuenta que la segunda aproxiamación es mucho mejor que la primera. □

En este capítulo presentaremos dos ejemplos en los cuales se utilizan sistemas de ecuaciones lineales que se obtienen al resolver la ecuación de Poisson usando diferencia finitas. La ecuación de Poisson es una ecuación en derivadas parciales con amplias aplicaciones en electrostática, ingeniería mecánica y física teórica. Su nombre se debe al matemático, geómetra y físico francés Siméon-Denis Poisson.

De hecho, sea $\Omega \subseteq \mathbb{R}^M$, con $M = 1, 2, 3$, un conjunto abierto y acotado. Y sea $f : \Omega \to \mathbb{R}$ dada y $k > 0$. La ecuación de Poisson en un sistema de coordenadas cartesianas tridimensional, toma la forma:

$$(P) \begin{cases} -k\left(\dfrac{\partial^2 u}{\partial x^2} + \dfrac{\partial^2 u}{\partial y^2} + \dfrac{\partial^2 u}{\partial z^2}\right) &= f(x,y,z), \quad \text{en } \Omega, \\ u &= 0, \quad \text{en } \partial\Omega. \end{cases}$$

En los ejemplos consideramos $M = 1$ y $M = 2$.

Por simplicidad y por razones históricas, estos problemas son muy utilizados como ejemplos en los métodos iterativos. Como nuestro interés es ver el funcionamiento de los métodos iterativos hemos elegido estos problemas por las características de la matriz que se forma al buscar la solución de estos.

3.1. Discretización del problema de Poisson en una dimensión

Consideremos un problema de valores en la frontera que aparece en muchas aplicaciones físicas. El problema está dado por la siguiente ecuación diferencial de segundo orden. En el caso que $M = 1$, tomamos $\Omega = (0, 1)$ y (P) se escribe como:

$$(P_1) \begin{cases} -k\dfrac{d^2u}{dx^2} &= f(x), \quad 0 < x < 1, \\ u(0) = u(1) &= 0. \end{cases}$$

Aunque, en general, se puede resolver este problema analíticamente, es instructivo para ilustrar los métodos iterativos presentados en el capítulo anterior.

Utilizando el método de diferencias finitas, el dominio (Ω) de este problema, se divide en N subintervalos cuyos extremos son los puntos $x_j = jh$, con $j = 0, ..., N$, donde $h = 1/N$ es la anchura de los subintervalos (Figura 3.1).

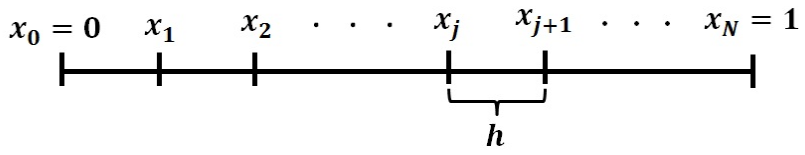

Figura 3.1: División del dominio Ω.

En cada uno de los $N - 1$ puntos interiores, se sustituye la ecuación diferencial original por una aproximación de diferencias finitas de segundo orden. Al hacer esta sustitución, también introducimos v_j como una aproximación a la solución exacta $u(x_j)$. Esta solución aproximada puede ser representada por un vector $\boldsymbol{v} = (v_1, ..., v_{N-1})^T$ cuyos componentes satisfacen la ecuación lineal

$$f(x_j) = \frac{k}{h^2}[-v_{j-1} + 2v_j - v_{j+1}], \tag{3.4}$$

donde $1 \leq j \leq N - 1$ y además $v_o = v_N = 0$.

Definiendo el vector de valores del lado derecho, $\boldsymbol{f} = (f(x_1), ..., f(x_{N-1}))^T = (f_1, ..., f_{N-1})$ y considerando $k = 1$, obtenemos el sistema de ecuaciones lineales en forma matricial $A\boldsymbol{v} = \boldsymbol{f}$, dado por

$$\begin{pmatrix} 2 & -1 & 0 & \cdots & \cdots & 0 \\ -1 & 2 & -1 & \ddots & & \vdots \\ 0 & -1 & 2 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & -1 & 2 & -1 \\ 0 & \cdots & \cdots & 0 & -1 & 2 \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ \vdots \\ \vdots \\ v_{N-1} \end{pmatrix} = \begin{pmatrix} f_1 \\ f_2 \\ \vdots \\ \vdots \\ \vdots \\ f_{N-1} \end{pmatrix} h^2,$$

donde la matriz de coeficientes A, es una matriz tridiagonal, simétrica definida positiva, con entradas $a_{ij} = 2$ cuando $i = j$ y entradas $a_{ij} = -1$ cuando $|i - j| = 1$ y de dimensión $N - 1 \times N - 1$. Además, notemos que por las características de la matriz A, el sistema de ecuaciones anterior tiene solución única por (Teorema 1.2.10).

Ahora bien, volviendo a la notación de los capítulos anteriores, $\boldsymbol{v}$ lo reeplazaremos por $\boldsymbol{x}$ y $\boldsymbol{f}$ por $\boldsymbol{b}$. Comenzaremos analizando la cantidad de iteraciones que los métodos iterativos estudiados, tienen que hacer para encontrar la solución exacta. Para ello, tomamos $\boldsymbol{x} = (x_j)$ donde $x_j = 1$ para todo $j \in \{1, ..., N - 1\}$, el vector $\boldsymbol{b}$ será construido a partir de $A\boldsymbol{x} = \boldsymbol{b}$ y la aproximación inicial, $\boldsymbol{x}^{(0)}$, será el vector nulo.

En los cuadros de comparación 3.1 y 3.2 de los métodos iterativos Jacobi, Gauss-Seidel y SOR, se muestra la cantidad de iteraciones que se deben realizar para alcanzar la tolerancia declarada. Para ello utilizamos tres valores para N (dimensión de la matriz) y dos valores para tolerancia del error relativo de 10^{-5} y 10^{-8}.

En primer lugar podemos ver que la cantidad de iteraciones necesarias, para hallar la solución del sistema, depende claramente de la tolerancia que se le exija al error relativo, mientras más pequeña sea esta, el proceso de iteraciones será más largo ya que el número de iteraciones requerido será mayor.

Ahora si queremos hacer comparaciones entre los métodos, notemos que con ambas tolerancias del error relativo y en las diferentes dimensiones de la matriz A, el método Gauss-Seidel requiere menos iteraciones, para llegar a la solución del sistema, que el método Jacobi, lo que nos muestra la veracidad del Teorema 2.2.8, donde la comparación entre los radios espectrales de B_J y B_{GS}, muestra que el método de Gauss-Seidel converje más rapidamente que el método de Jacobi.

A la vez, en el método SOR, para ciertos valores de w el número de iteraciones para encontrar la solución es menor que en el método Gauss-Seidel, notemos que solo es para algunos valores de w y que no para todos los valores la cantidad de iteraciones es la misma. Dos ideas están representadas aquí, la primera es la idea de sub-relajación y de sobre-relajación, donde el método SOR, dependiendo el valor del parámetro w frena o acelera la rapidez de

convergencia del método Gauss-Seidel y la segunda es, que nos hace pensar en que existe un valor para w el cual optimiza el proceso de iteración, tal como lo señala la Proposición 2.2.12.

Notemos, que aunque la convergencia de estos métodos para esta matriz A está asegurada (por las características de la matriz), podemos también comprobar la convergencia con el radio espectral de la matriz de iteración, sabiendo por el Teorema 2.1.2 que un método iterativo converge si y solo si $\rho(B) < 1$. El cuadro 3.3 muestra el radio espectral de la matriz de iteración de los tres métodos estudiados considerando $N = 10$, $N = 100$ y $N = 1000$. Notemos que, como también se muestra en el texto [3], cuando N aumenta el radio espectral de la matriz de iteración tiende a 1, lo que no significa que el método no converge, sino que le cuesta más converger. Esto lo podemos ver comparando los cuadros 3.1 y 3.2 con el cuadro 3.3, ya que mientras más cercano a 1 sea radio espectral de la matriz de iteración, la cantidad de iteraciones requeridas es mayor, es decir, el proceso de iteración se hace más lento.

A partir el radio espectral de la matriz de iteración en el método de Jacobi y utilizando la Proposición 2.2.12 encontramos el valor óptimo del parámetro w.

$$w_{opt} = \frac{2}{1+\sqrt{1-(0,9595)^2}} \approx 1,5604, \qquad \text{para } N = 10.$$

$$w_{opt} = \frac{2}{1+\sqrt{1-(0,9995)^2}} \approx 1,9387, \qquad \text{para } N = 100.$$

$$w_{opt} = \frac{2}{1+\sqrt{1-(0,9999)^2}} \approx 1,9801, \qquad \text{para } N = 1000.$$

Si ahora utilizamos el método SOR para resolver el sistema de ecuaciones para $N = 10$, $N = 100$ y $N = 1000$, pero ahora usamos w_{opt} para cada uno, obtenemos que el número de iteraciones necesarias para hallar la solución exacta es 27, 212 y 4897 respectivamente, con una tolerancia de 10^{-5} y si consideramos una tolerancia de 10^{-8} el números de iteraciones necesarias son 40, 345 y 11769 para las respectivas dimensiones. Claramente es menor que en cualquier otro valor de w y que en los métodos de Jacobi y Gauss-Seidel.

Por otra parte, analizaremos la Figura 3.2, con el fin de comparar el comportamiento del error relativo en cada uno de los métodos estudiados. En el caso del método SOR se utiliza $w_{opt} = 19387$.

Observemos que en los métodos de Jacobi y Gauss-Seidel, el comportamiento del error es parecido a pesar de que las iteraciones requeridas son más en el método de Jacobi que en Gauss-Seidel, notaremos que el error desciende rápidamente al comienzo, pero luego se detiene y el descenso comienza a ser más lento y constante.

A diferencia de los anteriores, en el método SOR, el error relativo disminuye no de forma tan brusca al comiezo, sino que lo hace de una forma suave pero constante, a excepción de algunos sectores donde el descenso es muy rápido, pero en general se nota la influencia del parámetro de relajación.

Método/N	10	100	1000
Jacobi	208	8519	48570
Gauss	113	4977	24734
$SOR_{\frac{1}{4}}$	630	20766	24472
$SOR_{\frac{1}{2}}$	302	11524	24413
$SOR_{\frac{3}{4}}$	180	7414	24398
$SOR_{\frac{5}{4}}$	70	3303	44498
$SOR_{\frac{3}{2}}$	37	2036	44877
$SOR_{\frac{7}{4}}$	45	994	31517

Cuadro 3.1: Ejemplo 1. Número de iteraciones para las diferentes dimensiones de la matriz de iteración, considerando una tolerancia para el error relativo de 10^{-5}.

Método/N	10	100	1000
Jacobi	375	22794	1307841
Gauss	197	12115	724292
$SOR_{\frac{1}{4}}$	1227	70731	3687164
$SOR_{\frac{1}{2}}$	557	32941	1838273
$SOR_{\frac{3}{4}}$	321	19313	1120720
$SOR_{\frac{5}{4}}$	118	7584	465691
$SOR_{\frac{3}{2}}$	58	4411	278607
$SOR_{\frac{7}{4}}$	67	2001	131684

Cuadro 3.2: Ejemplo 1. Número de iteraciones para las diferentes dimensiones de la matriz de iteración, considerando una tolerancia para el error relativo de 10^{-8}.

Método/N	10	100	1000
Jacobi	0,9595	0,9995	0,9999
Gauss	0,9339	0,9990	0,9999
$SOR_{\frac{1}{4}}$	0,9885	0,9999	0,9999
$SOR_{\frac{1}{2}}$	0,9738	0,9997	0,9999
$SOR_{\frac{3}{4}}$	0,9551	0,9994	0,9999
$SOR_{\frac{5}{4}}$	0,8663	0,9984	0,9999
$SOR_{\frac{3}{2}}$	0,7180	0,9971	0,9999
$SOR_{\frac{7}{4}}$	0,7500	0,9932	0,9999

Cuadro 3.3: Ejemplo 1. Radios espectrales de la matriz de iteración.

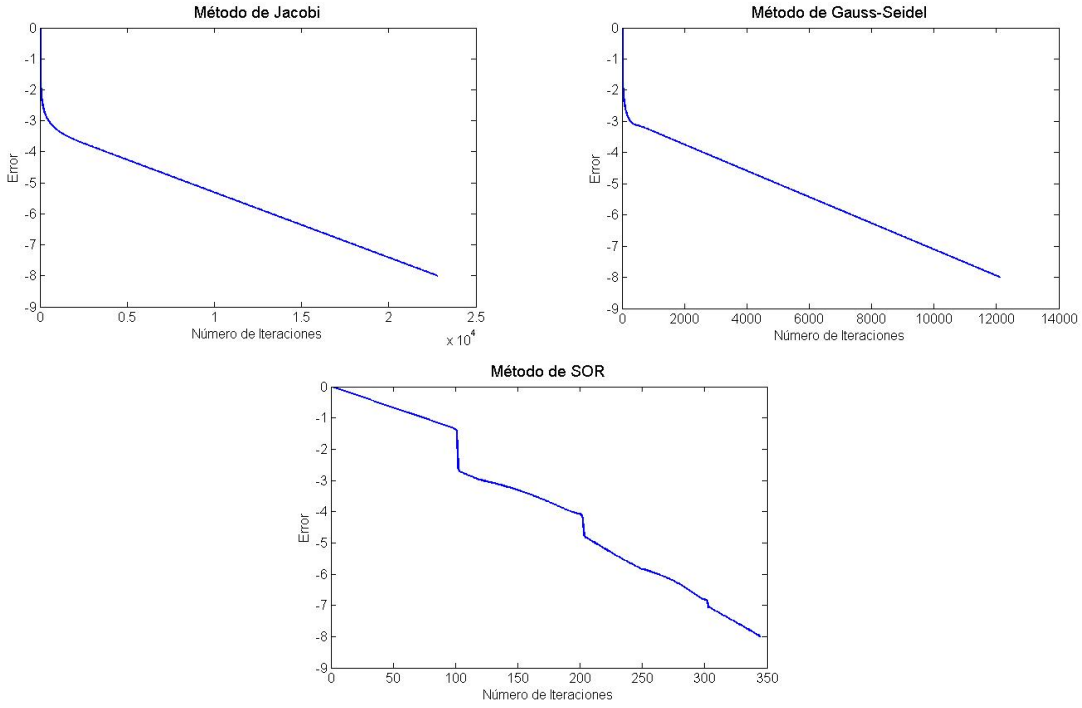

Figura 3.2: Ejemplo 1. $N = 100$.

3.2. Discretización del problema de Poisson en dos dimensiones

Tomamos ahora $M = 2$ y $\Omega = (0,1) \times (0,1)$. El problema es:

$$(P_2) \begin{cases} -k\left(\dfrac{\partial^2 u}{\partial x^2} + \dfrac{\partial^2 u}{\partial y^2}\right) &= f(x,y), \quad (x,y) \in \Omega, \\ u &= 0, \qquad \text{en } \partial\Omega. \end{cases}$$

Notemos que la ecuación está sujeta a la condición que $u = 0$ en el borde del cuadrado.

Como en el problema anterior, podemos formular la versión discreta de (P_2) usando diferencias finitas. Para ello definimos un punto de la cuadrícula $(x_i, y_j) = (ih, jh)$, donde $h = 1/N$. Consideramos v_{ij} una aproximación a la solución exacta $u(x_i, y_j)$ y denotamos por f_{ij} el valor de f en (x_i, y_j). Considerando $k = 1$ y reemplazando las derivadas de (P_2) por las diferencias finitas de segundo orden, nos conduce al sistema de ecuaciones lineales

$$\frac{-v_{i-1,j} + 2v_{ij} - v_{i+1,j}}{h^2} + \frac{-v_{i,j-1} + 2v_{ij} - v_{i,j+1}}{h^2} = f_{ij}, \tag{3.5}$$

igualdad que se puede escribir como

$$\frac{-v_{i-1,j} - v_{i,j-1} + 4v_{ij} - v_{i+1,j} - v_{i,j+1}}{h^2} = f_{ij}, \tag{3.6}$$

donde $v_{ij} = 0$ si $i \in \{0, N\}$ y $j \in \{1, ..., N-1\}$ o $j \in \{0, N\}$ e $i \in \{1, ..., N-1\}$.

El número de elementos de la solución exacta u_{ij} es $n = (N-1)^2$ y corresponde al número de los puntos de rejilla interna. Si queremos representar el sistema de ecuaciones en su forma matricial $A\boldsymbol{v} = \boldsymbol{f}$, queda una matriz A de $n \times n$ y los vectores n-dimensionales $\boldsymbol{v}$ y $\boldsymbol{f}$ con $n = (N-1)^2$. La forma de la matriz A implica que los puntos (interior) de la cuadrícula deben ser enumeradas de alguna manera, para objeto de este estudio será con un orden lexicográfico.

La representación del sistema en forma matricial, A debe escribirse como una matriz de bloques, el vector $\boldsymbol{v}_i = (v_{i1}, ..., v_{i,N-1})^T$, que recoge las incognitas de la fila i de la red, se descompone naturalmente en $N-1$ bloques y de igual forma $\boldsymbol{f}_i = (f_{i1}, ..., f_{i,N-1})^T$.

$$\frac{1}{h^2}\begin{pmatrix} C & -I & 0 & \cdots & \cdots & 0 \\ -I & C & -I & \ddots & & \vdots \\ 0 & -I & C & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & -I & C & -I \\ 0 & \cdots & \cdots & 0 & -I & C \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ \vdots \\ \vdots \\ v_{N-1} \end{pmatrix} = \begin{pmatrix} f_1 \\ f_2 \\ \vdots \\ \vdots \\ \vdots \\ f_{N-1} \end{pmatrix}$$

Donde C es una matriz de la siguiente forma.

$$C = \begin{pmatrix} 4 & -1 & 0 & \cdots & \cdots & 0 \\ -1 & 4 & -1 & \ddots & & \vdots \\ 0 & -1 & 4 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & -1 & 4 & -1 \\ 0 & \cdots & \cdots & 0 & -1 & 4 \end{pmatrix}$$

De esta forma, A toma la forma de una matriz tridiagonal por bloques de dimensión $(N-1) \times (N-1)$, la diagonal principal está formada por $(N-1)$ bloques de C, que a su vez es una matriz tridiagonal de $(N-1) \times (N-1)$ y la matriz I es la matriz identidad de $(N-1) \times (N-1)$. Además de ser tridiagonal A es simétrica por bloques y dispersa.

Retomando la notación de los capítulos anteriores, el vector $\boldsymbol{v}$ pasará a ser $\boldsymbol{x}$ y el vector $\boldsymbol{f}$ será $\boldsymbol{b}$. Para efecto de nuestro estudio A será una matriz de dimensión N^2, el vector $\boldsymbol{x}$ será el vector unitario de dimensión N^2 y como $\boldsymbol{b}$ es construido a partir de $A\boldsymbol{x} = \boldsymbol{b}$, por lo tanto la dimensión de $\boldsymbol{b}$ también es N^2. La aproximación inicial será el vector nulo de dimansión N^2.

Para comenzar el análisis de cómo es el comportamiento de los métodos iterativos en estudio, respecto a esta matriz A, cabe señalar que la convergencia de estos métodos está asegurada por las características de la matriz. Notando que A es estrictamente diagonal dominante por filas, entonces por el Teorema 2.2.4 los métodos de Jacobi y Gauss-Seidel convergen y

como A también es una matriz definida positiva, por la Proposición 2.2.11 el método SOR converge.

Lo primero que haremos será analizar la cantidad iteraciones que requiere cada método para hallar la solución exacta. Antes mencionaremos que, por la forma que está construida la matriz A (por bloques) y porque su dimención es N^2, los valores para N serán más pequeños que en el ejemplo anterior, con el fin de disminuir el gasto computacional. En los cuadros 3.4 y 3.5 encontramos la cantidad de iteraciones que demora cada método en hallar la solución exacta, utilizamos distintos valores para N y una tolerancia para el error relativo de 10^{-5} y 10^{-8}.

Al igual que el ejemplo anterior, el método SOR es el que requiere menos iteraciones para hallar la solución exacta, siendo más rápido que los métodos de Gauss-Seidel y de Jacobi. Notemos que mientras mayor es w el número de iteraciones necesarias es cada vez menor, por lo que nos es de interés conocer el valor óptimo de w para este problema en las diferentes dimensiones de la matriz de iteración.

Con el fin de encontrar w_{opt}, calculamos el radio espectral de la matriz de iteración del método de Jacobi (Cuadro 3.6). Están también incluidos los radios espectrales de las matrices de iteracón del método de Gauus-Seidel y del método SOR, para algunos valores de w.

Los valores óptimos de w, en el método SOR, para las diferentes dimensiones de la matriz son

$$w_{opt} = \frac{2}{1+\sqrt{1-(0,8660)^2}} \approx 1,3333, \qquad \text{para } N = 5.$$

$$w_{opt} = \frac{2}{1+\sqrt{1-(0,9595)^2}} \approx 1,5604, \qquad \text{para } N = 10.$$

$$w_{opt} = \frac{2}{1+\sqrt{1-(0,9927)^2}} \approx 1,7847, \qquad \text{para } N = 25.$$

Más adelante analizaremos el comportamiento del error, considerando $N = 100$ por lo que, incluiremos el valor óptimo de w para esta dimensión.

$$w_{opt} = \frac{2}{1+\sqrt{1-(0,9995)^2}} \approx 1,9387, \qquad \text{para } N = 100.$$

Utilizando estos valores de w en su respectivos valores de N, obtenemos que el número de iteraciones necesarias para hallar la solución usando el método SOR es 16 para $N = 5$, 28 para $N = 10$, 62 para $N = 25$ y 364 para $N = 100$, con una tolerancia para el error relativo de 10^{-5}. Podemos ver que la cantidad de iteraciones requeridas es menor que en los métodos de Jacobi, Gauss-Seidel e incluso para otros valores de w en SOR. Haremos notar también que, w_{opt} no es el mismo para todas las dimensiones de la matriz de iteración, por lo que no sería correcto usar solo uno para todas las dimensiones de ella, un ejemplo de ello es que, si utilizamos $w = 1,7847$ en la matriz de dimensión $N = 5$ la cantidad de iteraciones

requeridas para hallar la solución aciende a 51.

Por otra parte, podemos analizar cómo es el comportamiento del error relativo en los tres métodos (Figura 3.3) considerando $N = 100$. En el caso del método SOR se utiliza $w_{opt} = 19387$.

El comportamiento de error en los métodos de Jacobi y Gauss-Seidel, al igual que en (P_1) es muy parecido, descendiendo muy rápido solo al comienzo y luego viene un descenso más lento y constante.

En el método SOR se nota la clara influencia de w_{opt}, haciendo que el error, en general, disminuya más rápido y de una forma más suave.

Método/N	5	10	25	100
$Jacobi$	70	213	968	9016
$Gauss$	39	117	532	5226
$SOR_{\frac{1}{4}}$	228	650	2795	22497
$SOR_{\frac{1}{2}}$	107	310	1372	12270
$SOR_{\frac{3}{4}}$	63	184	830	7830
$SOR_{\frac{3}{2}}$	21	38	200	2120

Cuadro 3.4: Ejemplo 2. Número de iteraciones para las diferentes dimensiones de la matriz de iteración, considerando una tolerancia para el error relativo de 10^{-5}.

Método/N	5	10	25	100
$Jacobi$	118	380	1912	23293
$Gauss$	63	200	1004	12365
$SOR_{\frac{1}{4}}$	408	1246	6110	72478
$SOR_{\frac{1}{2}}$	184	566	2793	33690
$SOR_{\frac{3}{4}}$	105	326	1618	19729
$SOR_{\frac{3}{2}}$	30	60	353	4495

Cuadro 3.5: Ejemplo 2. Número de iteraciones para las diferentes dimensiones de la matriz de iteración, considerando una tolerancia para el error relativo de 10^{-8}.

Método/N	5	10	25	100
Jacobi	0,8660	0,9595	0,9927	0,9995
Gauss	0,7500	0,9206	0,9855	0,9990
$SOR_{\frac{1}{4}}$	0,9624	0,9885	0,9979	0,9999
$SOR_{\frac{1}{2}}$	0,9140	0,9733	0,9951	0,9997
$SOR_{\frac{3}{4}}$	0,8482	0,9522	0,9913	0,9994
$SOR_{\frac{3}{2}}$	0,5000	0,7280	0,9557	0,9971

Cuadro 3.6: Ejemplo 2. Radios espectrales de la matriz de iteración.

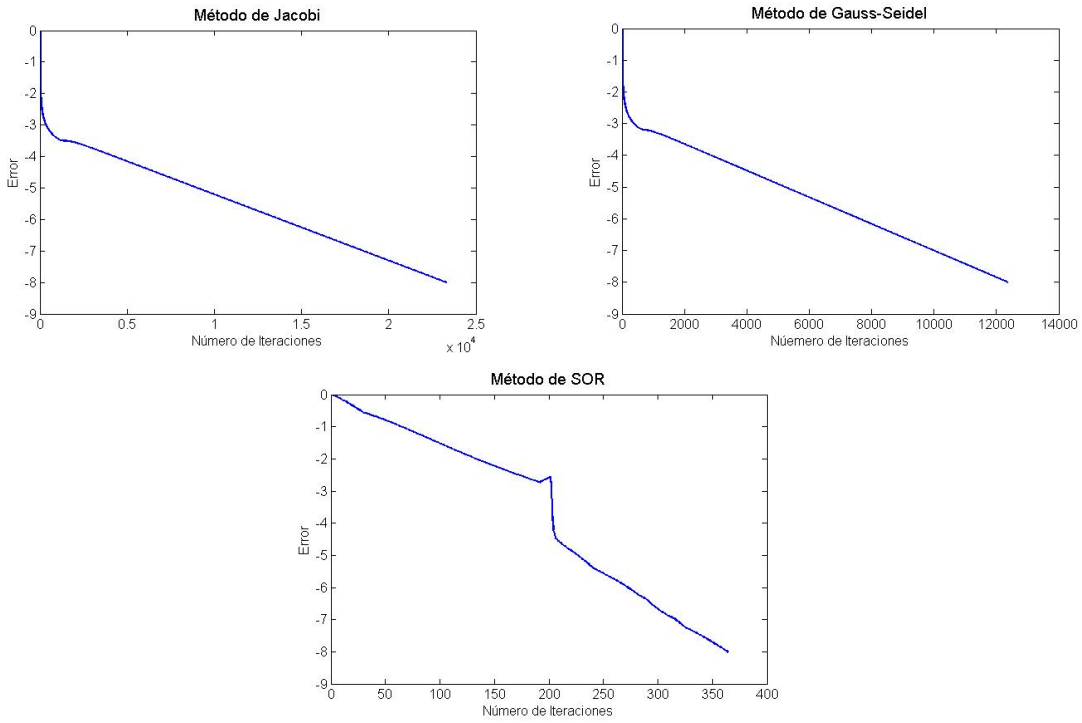

Figura 3.3: Ejemplo 2. $N = 100$.

Conclusiones

Antes de comenzar esta actividad de titulación, la única forma conocida para resolver un sitema de ecuaciones era utilizando los métodos directos, pero en este actividad aprendimos una nueva forma de resolverlos, los métodos iterativos para resolver sistema de ecuaciones lineales, por cierto, no todos ni toda su teoría, nos centramos en tres métodos iterativos estacionarios, los cuales consisten en resolver el sistema $A\boldsymbol{x} = \boldsymbol{b}$ utilizando la descomposición $A = M - N$, más en concreto $A = D - L - U$.

El primer método para resolver sitemas de ecuaciones lineales en estudio fue el de Jacobi, que consiste básicamente en despejar en cada ecuación la incógnita correspondiente, considerando que todos los elementos de la daigonal son distinto de cero.

El método de Gauss-Seidel incorpora un simple cambio al método de Jacobi. Los componentes de la nueva aproximación son usados tan pronto como están calculados, eso quiere decir que los componentes del vector $\boldsymbol{x}^{(k+1)}$ (de aproximación) se sobrescriben al momento de ser actualizados. Este pequeño cambio reduce considerablemente el almacenamiento requerido para el vector de aproximación.

Debido a este cambio parece lógico que el método de Gauss-Seidel converja más rápido que el método de Jacobi, esto se vió reflejado , en el capítulo 3 (ambos ejemplos), en la cantidad de iteraciones que debió realizar cada método para hallar la solución exacta, siendo el método de Gauss-Seidel más rápido que el de Jacobi. También notemos, aunque el error relativo deciende más rápido en Gauss-Seidel que en Jacobi, el comportamiento es muy parecido y se debe a que tanto en ambos métodos el proceso de iteración es el mismo, solo que en el de Gauss-Seidel las componentes del vector $\boldsymbol{x}^{(k+1)}$ son actualizadas en cuanto están disponibles y esto hace que solo el némero de iteraciones sea menor.

Al igual que el método de Gauss-Seidel al método de Jacobi, el método SOR introduce un cambio al de Gauss-Seidel, por medio de un parámetro de relajación (w). Como comentamos anteriormente, dependiendo del valor de este parámetro el método converge más rápido o más lento, incluso puede hasta no converger como lo muestran los resultados del capítulo 2. A que converja más rápido o más lento, se le asigna los nombres de sobre-relajación ($w > 1$) y sub-relajación ($0 < w < 1$) respectivamente. Esto se pudo comprobar en los ejemplos del capítulo 3, donde para los valores de w menores que 1, el método SOR necesitó más iteraciones que el método de Gauss-Seidel y para los valores de w mayores que 1 la solución fue hallada en menos iteraciones que el método de Gauss-Seidel (recordando que para $w = 1$,

SOR corresponde al mismo Gauss-Seidel).

Al mismo tiempo hicimos notar la existencia de un valor óptimo para el parámetro de relajación (w_{opt}), se mencionó en la teoría del capítulo 2 y se comprobó en los resultados numéricos de los ejemplos del capítulo 3, que al utilizar este valor para w el método SOR converje más rápido que los otros métodos y que el mismo método SOR para otros valores de w.

Otro concepto importante en la convergencia de los diferentes métodos, es el radio espectral de la matriz de iteración, en el capítulo 2 se mencionó que para que un método iterativo converja el radio espectral de la matriz de iteración debe ser menor que 1. Esto se comprobó en el capítulo de Resultados Numéricos, donde se pudo observar que el radio espectral de la matriz de iteración nos dice que tan rápido será el proceso de convergencia. El radio espectral es afectado si el tamaño de la matriz aumenta o disminuye y también en el método SOR es afectado por el valor que se le asigna a w. Obviamente si esto ocurre, también se ve reflejado en la cantidad de iteraciones necesarias para hallar la solución exacta del sistema.

Finalmente, en el orden Jacobi $\rightarrow$ Gauss-Seidel $\rightarrow$ SOR, en cada uno se introduce algún cambio para mejorar la rapidez y exactitud de convergencia, entendiendo así que el mejor método es el método SOR. No por esto podemos quedarnos solamente con el método SOR para resolver un sistema de ecuaciones, porque esto dependerá de las características de la matriz A y también recordemos que el valor óptimo de w necesita el radio espectral de la matriz de iteración del método de Jacobi.

El trabajo con un software matemático, en nuetro caso MATLAB, ha sido esencial para el trabajo con los métodos iterativos para resolver sistemas de ecuaciones, reduce el tiempo destinado a los cálculos y nos asegura que, estando bien la programación, el resultado es correcto y que no hemos cometido algún error en medio del proceso.

Bibliografía

[1] Axelsson O. Iterative Solution Methods. Cambridge University Press, New York.1994.

[2] Burden R. y Faires J. Numerical Analysis. Ninth ed. Brooks-Cole Cengage learning. 2011.

[3] Briggs W. A Multigrid Tutorial. Society for Industrial and Applied Mathematics. 1987.

[4] Fraleigh J. y Beauregard R. Álgebra Lineal. Addison-Wesley Iberoamericana, S. A. Wilmington. 1987.

[5] Quarteroni A. Sacco R. y Saleri F. Numerical Mathematics. Second Ed. Texts in applied mathematics. Springer. 2007.

[6] Skiba Y. Métodos y esquemas numéricos: Un análisis computacional. Dirección General de Publicaciones y Fomento Editorial. México. 2005.

[7] Stanley I. Grossman. Álgebra Lineal. Sexta Edición. McGraw-Hill. 2008.

[8] Strang G. Introduction to Linear Algebra. Fourth Edition. Wellesley-Cambridge Press. 2009.

[9] Wolfgang H. Iterative Solution of Large Sparse Systems of Equations. Springer-Verlag. 1994.