Universidad Nacional de Ingeniería

Facultad de Ciencias

TESIS

Análisis de algoritmos para el cálculo de la raíz cuadrada de una matriz no singular y una aplicación a un método en diferencia irracional

Para obtener el título profesional de Licenciado en Matemática

Elaborado por

Diana Carolina Flores Gallo

0000-0001-9524-6780

Asesor

Mg. Jesus Cernades Gomez

0000-0003-3671-4358

LIMA-PERÚ

2024

Citar / How to cite	Flores Gallo [1]
Referencia / Reference Estilo / Style: IEEE	[1] D. Flores Gallo, "*Análisis de algoritmos para el cálculo de la raíz cuadrada de una matriz y una aplicación a un método en diferencia irracional*" [Tesis de pregrado]. Lima (Perú): Universidad Nacional de Ingeniería, 2024.

Citar / How to cite	(Flores, 2024)
Referencia / Reference Estilo / Style: APA (7ma ed.)	Flores, D. (2024). *Análisis de algoritmos para el cálculo de la raíz cuadrada de una matriz y una aplicación a un método en diferencia irracional* [Tesis de pregrado, Universidad Nacional de Ingeniería]. Repositorio Institucional UNI.

Agradecimientos

A mis amados abuelitos María Teresa Gómez Mendoza y Víctor Eladio Gallo Córdova por el amor incondicional que siempre me brindaron, por ser mis segundos padres.

A mi madre, Teresa del Rosario Gallo Gómez, por el apoyo y amor infinito a lo largo de mi vida, en especial de mis años universitarios.

A mi asesor y amigo, Jesús Cernades por sus consejos, paciencia y dedicación durante el desarrollo del presente trabajo.

Al profesor William Echegaray por sus valiosas observaciones y recomendaciones para culminar esta investigación, y mis amigos más cercanos, por sus constantes palabras de aliento.

Resumen

En este trabajo se desarrolla el cálculo de la raíz cuadrada de una matriz no singular y se comparan con algunos algoritmos, realizando previamente algunas evaluaciones de estabilidad y rendimiento computacional. Entre los algoritmos estudiados, se muestran algunos ejemplos y se realizan comparaciones con el método de Newton, método de Newton modificado 1 y 2, Denman Beavers modificado, Padé y Padé a escala, que muestran buenos resultados a pesar del mal condicionamiento de algunas matrices. Luego, se explora un método en diferencia irracional llamado RT-ω, estudiando previamente su estabilidad y su convergencia. A continuación se resuelven ecuaciones diferenciales rígidas oscilatorias y altamente oscilatorias con el método irracional RT-ω, contrastándolo con la solución exacta, el cual tuvo muy buenos resultados, a diferencia de los métodos Runge Kutta-4 explícito, Euler explícito y Runge Kutta-4 implícito con diversos tamaños de mallas, verificando que, de acuerdo al problema y tipo de rigidez presentada, el método RT-ω fue más estable y su convergencia fue más rápida que los otros métodos. Como conclusiones de este trabajo pudimos verificar que el método estudiado RT-ω, con $\omega = 2$ es muy eficiente para aproximar la solución de ecuaciones diferenciales rígidas altamente oscilatorias, ya que ofrece una mejora significativa en la estabilidad numérica.

Palabras Claves: Problemas rígidos, Métodos racionales e irracionales, Método en diferencia irracional, estabilidad

Abstract

In this work, the calculation of the square root of the non-singular matriz is developed and compared with some algorithms, after performing some stability and computational performance evaluations. Among the algorithms studied, some examples are shown and comparisons are made with Newton's method, modified Newton method 1 and 2, modified Denman Beavers, Pade and scaled Pade, which show good results despite the bad conditioning of some matrices. Then, an irrational difference method called RT-ω is explored, previously studying its stability and convergence. Next, rigid oscillatory and highly oscillatory differential equations are solved with the irrational RT-ω method, contrasting it with the exact solution, which had very good results, unlike the explicit Runge Kutta-4, explicit Euler and implicit Runge Kutta-4 methods with various mesh sizes, verifying that, according to the problem and type of rigidity presented, the RT-ω method was more stable and its convergence was faster than the other methods. As conclusions of this work we were able to verify that the studied RT-ω method, with $\omega = 2$ is very efficient to approximate the solution of highly oscillatory rigid differential equations, since it offers a significant improvement in numerical stability.

Palabras Claves: Stiff problems, Rational and irrational methods, Irrational difference method, Stability

Tabla de contenido

Lista de Tabla

Lista de Figuras

Introducción

En el ámbito de las matemáticas aplicadas y la ingeniería, la raíz cuadrada de una matriz es una operación que se utiliza en diversas aplicaciones, como la teoría de control, la resolución de sistemas de ecuaciones diferenciales y la física cuántica. El cálculo de esta operación no es una tarea trivial, ya que implica resolver problemas numéricos complejos, especialmente cuando se trata de matrices grandes o mal condicionadas. La búsqueda de algoritmos eficientes, precisos y estables para calcular la raíz cuadrada de una matriz es, por lo tanto, un área de investigación activa y de gran relevancia.

En el capítulo dedicado a los ejemplos numéricos veremos que hay muchas definiciones diferentes para explicar lo que es un problema rígido, pero la mayoría incluye alguna referencia a que en este tipo de problemas, para la mayoría de los métodos explícitos, la mayor longitud de paso h que garantiza la estabilidad numérica es mucho menor que la mayor longitud numérica que nos permite obtener un error de discretización normal. Por otro lado, los problemas rígidos en ecuaciones diferenciales ordinarias (EDO) presentan desafíos numéricos significativos debido a la presencia de soluciones que varían en diferentes escalas de tiempo. Estos problemas requieren métodos de integración robustos que puedan manejar la rigidez sin comprometer la precisión ni la estabilidad. Tradicionalmente, se han utilizado métodos específicos para abordar estos problemas, pero siempre hay

espacio para explorar nuevos enfoques que puedan ofrecer mejoras sustanciales. Entre los métodos convergentes racionales, aquí se van a tratar métodos de tipo de un paso, como el de Euler, Runge Kutta, trapezoidal y el método RT-ω el cual es un método en diferencia irracional.

El método en diferencia irracional es una técnica relativamente novedosa que promete ventajas en términos de precisión y estabilidad para la solución de problemas numéricos complejos. Este método, que se basa en la utilización de diferencias finitas irracionales, tiene el potencial de mejorar la eficiencia y la exactitud en la solución de problemas rígidos en EDO, que lo convierte en un candidato prometedor para ser investigado y aplicado.

En este trabajo, se propone un análisis exhaustivo de algoritmos existentes para el cálculo de la raíz cuadrada de una matriz, evaluando su desempeño en términos de precisión, estabilidad y rendimiento computacional. Además, se explorará la aplicación del método en diferencia irracional a problemas rígidos, evaluando su efectividad y comparándola con los métodos tradicionales. El objetivo principal es desarrollar una comprensión profunda de estos algoritmos y métodos, y proporcionar soluciones que puedan ser aplicadas de manera efectiva en contextos prácticos.

Esta investigación se estructura de la siguiente manera: en primer lugar, se presenta una revisión de la literatura sobre algoritmos para el cálculo de la raíz cuadrada de una matriz y métodos para resolver problemas rígidos. Luego, se describen los métodos y algoritmos propuestos, junto con el fundamento teórico. Posteriormente, se realizan experimentos numéricos para evaluar el desempeño de los algoritmos y métodos propuestos. Finalmente, se discuten los resultados obtenidos y se presentan las conclusiones y recomendaciones para futuras investigaciones.

A continuación se presenta la descripción del contenido por capítulos:

El capítulo muestra un marco teórico general dedicado exclusivamente a fundamentar la necesidad de nuestra investigación, señalando los antecedentes de la misma. Además, se exhiben los fundamentos teóricos de nuestra propuesta.

El capítulo I analiza los métodos numéricos para el cálculo de la raíz cuadrada de una matriz no singular. En la sección 2.1 se realiza la evolución de los métodos, en la sección 2.2 se indaga sobre el problema de la raíz cuadrada de una matriz, en la sección 2.3 se analiza los métodos para hallar la raíz cuadrada de una matriz, entre ellos el método de Newton y otros.

El capítulo H expone el método en diferencias finitas para problemas rígidos. En la sección 3.1 se presenta el problema de valor inicial; en la sección 3.2 se presenta a los sistemas de ecuaciones no lineales; en la sección 3.3 se estudia los métodos de un paso como el de Euler, Taylor, Euler modificado, Runge Kutta; en la sección 3.4 se realiza la descripción global de los métodos de un paso, se estudia la estabilidad, la convergencia, el error global asintótico, la estimación del error global; en la sección 3.5 se analiza y estudia la estimación del error de truncamiento, en la sección 3.6 se expone las técnicas para el control de paso; en la sección 3.7 se expone el problema de rigidez, y en las secciones siguientes se analiza la A-estabilidad, la aproximación de Padé y finalmente en la sección 3.10 se establecen las regiones de absoluta convergencia.

El capítulo H se ha dedicado a presentar los resultados numéricos de los diferentes métodos para hallar la raíz cuadrada de una matriz no singular, y también se pretende llegar a la conclusión de que la familia de métodos RT-ω son muy eficaces sobre un conjunto amplio

de problemas tests que se encuentran clasificados en función del tipo de problemas y las dificultades que presenten: problemas que proceden de ecuaciones donde un valor propio sea complejo puro y con parte real altamente negativa, en relación con otro valor propio que son problemas rígidos oscilatorios o altamente oscilatorios. Finalmente en el capítulo B se muestran las conclusiones y las sugerencias.

Planteamiento de la realidad problemática

Los problemas rígidos en ecuaciones diferenciales ordinarias (EDO) representan un desafío significativo debido a sus propiedades numéricas particulares, que requieren métodos de integración robustos y estables. En este contexto, surge la necesidad de:

1. Desarrollar y analizar algoritmos eficientes para el cálculo de la raíz cuadrada de una matriz. Este análisis incluirá la evaluación de la precisión, la estabilidad y el rendimiento computacional de estos algoritmos.

2. Explorar la aplicación de un método en diferencia irracional a problemas rígidos en EDO. Dado que los problemas rígidos presentan dificultades numéricas específicas, es fundamental investigar cómo este método puede mejorar la precisión y estabilidad en la solución de tales problemas.

Objetivos

Objetivo general de la investigación

Desarrollar y analizar algoritmos eficientes y estables para el cálculo de la raíz cuadrada de una matriz, y aplicar un método en diferencia irracional a la solución de problemas rígidos en ecuaciones diferenciales ordinarias, con el fin de mejorar la precisión y estabilidad en la solución de estos problemas.

Objetivos específicos de la investigación

1. Analizar y comparar algoritmos para el cálculo de la raíz cuadrada de una matriz en términos de precisión, estabilidad y eficiencia.

2. Desarrollar e implementar un método en diferencia irracional para resolver problemas rígidos en EDO, evaluando su efectividad y comparándolo con métodos tradicionales.

I. Fundamento teórico

En este primer capítulo se presentan los fundamentos teóricos necesarios para abordar el estudio de matrices y su aplicación en la resolución de problemas matemáticos y científicos. Las matrices son estructuras matemáticas que permiten representar y manipular datos de manera eficiente, siendo esenciales en campos como la álgebra lineal, la estadística y la teoría de sistemas dinámicos.

Comenzaremos con los conceptos básicos de matrices, donde exploraremos la notación matricial y vectorial, así como las operaciones fundamentales que pueden realizarse con estas estructuras. La comprensión de la matriz inversa y el cálculo del determinante son aspectos críticos que facilitan la solución de sistemas de ecuaciones lineales. También se abordará la matriz de Vandermonde, una herramienta clave en la interpolación polinómica, y las normas matriciales y vectoriales, que proporcionan un marco para medir la magnitud y la distancia en el espacio.

El capítulo continuará con un análisis de los valores y vectores propios de una matriz, que juegan un papel crucial en la descomposición espectral, permitiendo la simplificación de problemas complejos. Además, se introducirá la función analítica y la transformación de Möbius, que enlazan el álgebra lineal con conceptos de análisis complejo.

Finalmente, se discutirán los errores relativos y absolutos en los cálculos, así como el

método de Newton para resolver sistemas de ecuaciones no lineales, centrándonos en su convergencia local y su aplicación a problemas de valor inicial. Este marco teórico servirá como base para el desarrollo de temas más avanzados en los capítulos posteriores, brindando al lector las herramientas necesarias para comprender y aplicar los conceptos en contextos prácticos. Para mayor detalle se puede consultar en Chávez [5] y Golub [14].

A. Conceptos básicos de matrices

El cálculo matricial se basa en una jerarquía de operaciones en álgebra lineal. Veamos las principales propiedades matriciales y sus operaciones respectivas.

1. Notación matricial

Sea $\mathbb{K}$ el conjunto de los números reales o número complejos. Denotamos el espacio vectorial de todas las matrices en $\mathbb{K}$ de orden $m \times n$ por $\mathbb{K}^{m\times n}$:

$$A \in \mathbb{K}^{m\times n} \quad \Leftrightarrow \quad A = (a_{ij}) = \begin{bmatrix} a_{11} & \cdots & a_{in} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix}, \quad a_{ij} \in \mathbb{K}$$

donde, las letras mayúsculas son usadas para denotar matrices (por ejemplo A, B) y las correspondientes letras minúsculas con los subíndices (i, j), por ejemplo a_{ij}, b_{ij} son los elementos de dichas matrices.

2. Notación vectorial

Sea $\mathbb{K}^n$ el espacio vectorial real de los $n-$vectores:

$$x \in \mathbb{K}^n \qquad \Longleftrightarrow \qquad x = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}, \quad x_i \in \mathbb{K}.$$

Nos referimos a x_i como el i-ésimo componente de x.

Observemos que estamos identificando a $\mathbb{K}^n$ con $\mathbb{K}^{n\times 1}$ y los elementos de $\mathbb{K}^n$ son vectores columnas. Por otro lado, los elementos de $\mathbb{K}^{1\times n}$ son vectores fila:

$$x \in \mathbb{K}^{1\times n} \qquad \Longleftrightarrow \qquad x = (x_1, \cdots, x_n).$$

3. Operaciones con matrices

El espacio vectorial tiene como operaciones usuales o canónica de suma y multiplicación por un escalar.

1. La adición $(\mathbb{K}^{m\times n} \times \mathbb{K}^{m\times n} \to \mathbb{K}^{m\times n})$

$$C = A + B \qquad \Longrightarrow \qquad c_{ij} = a_{ij} + b_{ij}$$

2. La multiplicación por un escalar $(\mathbb{K} \times \mathbb{K}^{m\times n} \to \mathbb{K}^{m\times n})$

$$C = \alpha A \qquad \Longrightarrow \qquad c_{ij} = \alpha a_{ij}$$

3. La multiplicación de matriz con matriz $(\mathbb{K}^{m\times p} \times \mathbb{K}^{p\times n} \to \mathbb{K}^{m\times n})$

$$C = AB \qquad \Longrightarrow \qquad c_{ij} = \sum_{k=1}^{p} a_{ik} b_{kj}$$

4. Las operaciones básicas matriciales incluyen la transposición ($\mathbb{K}^{m\times n} \to \mathbb{K}^{n\times m}$),

$$C = A^*(\text{o } A^T) \qquad \Longrightarrow \qquad c_{ij} = \bar{a}_{ji} \ (\text{o } c_{ij} = a_{ji}),$$

5. Si $A = B + iC \in \mathbb{C}^{m\times n}$, entonces designamos las partes real e imaginaria de A como $Re(A) = B$ e $Im(A) = C$, respectivamente.

6. El producto interno denotado por $\langle\ ,\ \rangle$ es definido:

$$\langle x, y\rangle = \sum_{i=1}^{n} x_i y_i, \ \ \forall x, y \in \mathbb{K}^{n\times 1}$$

Para el espacio vectorial complejo de n - vectores designado por $\mathbb{C}^n$, se define el producto interno de los x e y complejo como:

$$s = x^H y = \sum_{i=1}^{n} \overline{x}_i y_i$$

Finalmente, diremos que una matriz $A \in \mathbb{K}^{n\times n}$ es definida semidefinida positiva (definida positiva) si

$$\langle Ax, x\rangle \geq 0 \ (\langle Ax, x\rangle > 0), \ \forall x \in \mathbb{K}^{n\times 1} - \{0\} \ (\forall x \in \mathbb{K}^{n\times 1})$$

y lo denotaremos por $A \geq 0$ ($A > 0$).

Se dice que una matriz B de orden n es la raíz cuadrada de una matriz A de orden n si $B^2 = A$, denotado por $B = \sqrt{A}$.

4. Matriz inversa

La matriz identidad de orden $n \times n$ denotada por I_n es definida por el particionamiento de la columna

$$I_n = [\ e_1, \ \ \cdots, \ \ e_n\ ]$$

donde e_k es el k ésimo vector canónico:

$$e_k = (\underbrace{0, \cdots, 0}_{k-1}, 1, \underbrace{0, \cdots, 0}_{n-k})^T$$

Si A y X están en $\mathbb{K}^{n\times n}$ y satisface $AX = XA = I$, entonces X es la inversa de A y se denota por A^{-1}.

Varias propiedades de la matriz inversa tienen un papel importante en el cálculo de matrices. A continuación, veamos algunas:

1. La inversa de un producto de matrices invertibles está dado por:

$$(AB)^{-1} = B^{-1}A^{-1}$$

2. La transposición de la inversa es la inversa de la transposición:

$$(A^{-1})^T = (A^T)^{-1}$$

3. La identidad

$$B^{-1} = A^{-1} - B^{-1}(B - A)A^{-1}$$

   muestra cómo la inversa cambia si la matriz cambia.

4. Así también se muestra la fórmula de Sherman-Morrison:

$$(P + UU^T)^{-1} = P^{-1} - P^{-1}U(I + U^T P^{-1}U)^{-1}U^T P^{-1}$$

   donde P es invertible de orden n y U una matriz de orden $n \times k$.

5. Determinante

Sea $A = (a_{ij}) \in \mathbb{K}^{n\times n}$, entonces el determinante, denotado por $\det(A)$, es definido en términos de determinantes de orden $n-1$:

$$\det(A) = \sum_{j=1}^{n}(-1)^{j+1}a_{ij}\det(A_{ij}) = \sum_{j=1}^{n}(-1)^{j+1}a_{ij}\det(A_{ij})$$

donde i, j pueden ser tomados en $I_n = \{1, 2, \ldots, n\}$ y A_{ij} es la matriz de orden $(n-1)\times(n-1)$, obtenida al eliminar la fila i con la columna j de la matriz A. Las propiedades más utilizadas de los determinantes son:

$$\begin{aligned}
\det(AB) &= \det(A)\det(B) && A, B \in \mathbb{K}^{n\times n}\\
\det(A^T) &= \det(A) && A \in \mathbb{K}^{n\times n}\\
\det(cA) &= c^n\det(A) && c \in \mathbb{K}, A \in \mathbb{K}^{n\times n}
\end{aligned}$$

Diremos que una matriz $A \in \mathbb{K}^{n\times n}$ es no singular si $\det(A) \neq 0$, caso contrario diremos que es singular.

Si A^{-1} existe, entonces A se puede probar que es ella no singular.

6. Matriz de Vandermonde

La matriz:

$$V(x_1, x_2, \cdots, x_n) = \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ x_1 & x_2 & x_3 & \cdots & x_n \\ \vdots & \vdots & \vdots & & \vdots \\ x_1^{n-1} & x_2^{n-1} & x_3^{n-1} & \cdots & x_n^{n-1} \end{bmatrix}$$

es llamada la **matriz de Vandermonde**, y su determinante es dado por:

$$\det V(x_1, \cdots, x_n) = \prod_{1\leq i<j\leq n}(x_j - x_i)$$

Para otras propiedades, se puede consultar en Kalman [18].

7. Norma matricial

Desde que $\mathbb{K}^{m\times n}$ es isomorfo a $\mathbb{K}^{m\times n}$, la definición de la norma matricial debe ser equivalente a la definición de norma vectorial. En particular, $f : \mathbb{K}^{m\times n} \to \mathbb{R}$ es la norma matricial si se cumplen las siguientes propiedades:

1. $f(A) \geq 0 \qquad A \in \mathbb{K}^{m\times n}, \qquad (f(A) = 0 \text{ si y solo si } A = 0)$

2. $f(A + B) \leq f(A) + f(B) \qquad A, B \in \mathbb{K}^{m\times n}$

3. $f(\alpha A) = |\alpha| f(A) \qquad \alpha \in \mathbb{K}, A \in \mathbb{K}^{m\times n}$

Al igual que con las normas vectoriales, usamos la notación de la doble barra para designar las normas matriciales, es decir, $\|A\| = f(A)$.

La norma matricial más frecuente usada es la norma de Frobenius.

$$\|A\|_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} |a_{ij}|^2}$$

y la p-norma

$$\|A\|_p = \sup_{x\neq 0} \frac{\|Ax\|_p}{\|x\|_p}$$

Observación 1. *Dada la caracterización de una matriz A de orden n, con norma $\|A\| < 1$, entonces $I - A$ es una matriz invertible.*

8. Norma vectorial

La norma vectorial en $\mathbb{K}^n$ es la función $f : \mathbb{K}^n \to \mathbb{R}$ que satisface las siguientes propiedades:

1. $f(x) \geq 0 \qquad x \in \mathbb{K}^n, \qquad (f(x) = 0 \text{ si y solo si } x = 0)$

2. $f(x+y) \leq f(x) + f(y) \qquad x, y \in \mathbb{K}^n$

3. $f(\alpha x) = |\alpha| f(x) \qquad \alpha \in \mathbb{K}, x \in \mathbb{K}^n$

Denotamos tal función como $f(x) = \|x\|$. Una clase de normal muy usada es la p-norma, definida por:

$$\|x\|_p = (|x_1|^p + \cdots + |x_n|^p)^{\frac{1}{p}}, \qquad p \geq 1$$

También, enunciaremos las siguientes normas importantes:

$$\begin{aligned} \|x\|_1 &= |x_1| + \cdots + |x_n| \\ \|x\|_2 &= (|x_1|^2 + \cdots + |x_n|^2)^{\frac{1}{2}} = (x^T x)^{\frac{1}{2}} \\ \|x\|_\infty &= \max_{1 \leq i \leq n} |x_i| \end{aligned}$$

Observación 2. *Un vector x es llamado unitario con respecto a la norma $\|\cdot\|$ si satisface $\|x\| = 1$.*

El análisis de algoritmos matriciales requiere con frecuencia el uso de normas matriciales. Por ejemplo, la calidad de un solucionador de sistemas lineales puede ser pobre si la matriz de coeficientes es "casi singular". Para cuantificar la noción de casi singularidad necesitamos una medida de distancia en el espacio de matrices. Las normas matriciales proporcionan esa medida.

B. Valores y vectores propios de una matriz

Definición 1 (Ver [15])**.** *Sea A una matriz de orden $n \times n$ con componentes reales. El número λ (real o complejo) se llama valor característico de A si hay un vector v distinto de cero en $\mathbb{K}^n$ tal que:*

$$Av = \lambda v$$

El vector $v \neq 0$ se llama un vector característico de A correspondiente al valor característico λ.

Nota: La palabra *eigen* significa "propio" o "apropiado" en alemán. Los valores característicos se llaman también *valores propios* o autovalores, y los vectores característicos, *vectores propios* o autovectores.

Teorema 1. *Sea A una matriz de orden $n \times n$. Entonces λ es un valor propio de A si y solo si*

$$p(\lambda) = \det(A - \lambda I) = 0$$

Definición 2 (Matriz diagonalizable)**.** *Una matriz A de orden $n \times n$ es diagonalizable si existe una matriz diagonal D tal que la matriz invertible C de orden $n \times n$ cumple*

$$D = C^{-1}AC$$

Teorema 2. *Una matriz A de orden $n \times n$ es diagonalizable si y solo si tiene n vectores propios linealmente independientes. En este caso, la matriz diagonal D equivalente a A está dada por*

$$D = \begin{bmatrix} \lambda_1 & 0 & 0 & \cdots & 0 \\ 0 & \lambda_2 & 0 & \cdots & 0 \\ 0 & 0 & \lambda_3 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \lambda_n \end{bmatrix}$$

donde $\lambda_1, \lambda_2, \cdots, \lambda_n$ son los valores propios de A. Si C es una matriz cuyas columnas son vectores propios linealmente independientes de A, entonces

$$D = C^{-1}AC$$

Teorema 3. *Dadas las matrices $A \in \mathbb{C}^{n\times n}$ y $B \in \mathbb{C}^{m\times m}$, la ecuación de Sylvester $AX + XB = C$ tiene única solución $X \in \mathbb{C}^{n\times m}$ para cualquier $C \in \mathbb{C}^{n\times m}$ si y solo si A y $-B$ no comparten ningún valor propio.*

1. Descomposición espectral

Sea A una matriz simétrica diagonalizada ortogonalmente por

$$P = [u_1 \quad u_2 \quad \cdots \quad u_n]$$

y sea $\lambda_1, \lambda_2, \cdots, \lambda_n$ los valores propios de A asociados a los vectores unitarios $u_1, u_2, \cdots, u_n$. Si se sabe que $D = P^T AP$, donde D es una matriz diagonal con los valores propios ubicados en la diagonal, entonces, la matriz A puede ser expresada como:

$$A = PDP^T = [u_1 \quad u_2 \quad \cdots \quad u_n] \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix} \begin{bmatrix} u_1^T \\ u_2^T \\ \vdots \\ u_n^T \end{bmatrix}$$

$$= [\lambda_1 u_1 \quad \lambda_2 u_2 \quad \cdots \lambda_n u_n] \begin{bmatrix} u_1^T \\ u_2^T \\ \vdots \\ u_n^T \end{bmatrix}$$

Multiplicando esas matrices, obtenemos la fórmula:

$$A = \lambda_1 u_1 u_1^T + \lambda_2 u_2 u_2^T + \cdots + \lambda_n u_n u_n^T$$

que es la denominada descomposición espectral de A. La terminología *descomposición espectral* hace referencia al espectro de una matriz, que como muchas veces es denominado, el conjunto de todos los valores propios de una matriz.

La forma canónica de Jordan

Sea una matriz $A \in \mathbb{C}^{n\times n}$ con p valores propios distintos $\lambda_1, \lambda_2, \cdots, \lambda_p$ de multiplicidades $m_1, m_2, \cdots, m_p$ de modo que

$$m_1 + m_2 + \cdots + m_p = n$$

Entonces existe una matriz P no singular de tal manera que

$$A = PJP^{-1} \in \mathbb{C}^{n\times n} \tag{1}$$

donde

$$J = diag(J_1, J_2, \cdots, J_p),\ J_i = J_i(\lambda_i) = \begin{bmatrix} \lambda_i & 1 & & & \\ & \ddots & \ddots & & \\ & & & \lambda_i & 1 \\ & & & & \lambda_i \end{bmatrix} \in \mathbb{C}^{n\times n} \tag{2}$$

Además también se cumple que:

$$f(A) = Pf(J)P^{-1} \in \mathbb{C}^{m_i\times m_i} \tag{3}$$

donde $f(\lambda) = \det(\lambda I - A)$.

C. Función analítica

Definición 3. *Sea $G \subset \mathbb{C}$ un conjunto abierto. Se dice que una función $f : G \to \mathbb{C}$ es analítica si para todo $z_0 \in G$ existe $\rho > 0$ y una sucesión compleja $\{a_n\}$, que depende de z_0, tales que:*

$$f(z) = \sum_{n=0}^{\infty} a_n (z - z_0)^n, \qquad |z - z_0| < \rho$$

1. Transformación de Möbius

Definición 4 (Ver [1]). *Sean $a, b, c, d \in \mathbb{C}$. Se denomina transformación lineal fraccionada si:*

$$w = T(z) = \frac{az + b}{cz + d}$$

Si además se verifica la condición $ad - bc \neq 0$, dicha expresión recibe el nombre de transformación de Möbius.

Observaciones 1. *La condición $ad - bc \neq 0$ permite garantizar lo siguiente:*

1. *Las expresiones $az + b$ y $cz + d$ no se anulan para los mismos valores de z.*

2. *La transformación T no puede ser constante, ya que a y c no pueden ser ambas cero, al igual que b y d no pueden ser ambas cero.*

3. *En general, el denominador no puede ser un múltiplo constante del numerador, es decir que $az + b$ y $cz + d$ no tienen un factor común.*

Proposición 1. *Se cumplen las siguientes propiedades:*

1. *Toda transformación de Möbius es una biyección. En particular, la inversa de una transformación de Möbius es también una transformación de Möbius.*

2. *Toda transformación de Möbius se puede expresar como la composición de transformaciones lineales (homotecias, rotaciones, traslaciones) y la inversión.*

D. Error relativo y absoluto

Supongamos que $\widehat{x} \in \mathbb{K}^n$ es una aproximación para $x \in \mathbb{K}^n$. Para una norma vectorial dada $\|\cdot\|$, decimos que:

$$\epsilon_{abs} = \|\widehat{x} - x\|$$

es el error absoluto en $\widehat{x}$. Si $x \neq 0$, entonces:

$$\epsilon_{rel} = \frac{\|\widehat{x} - x\|}{\|x\|}$$

es el error relativo en $\widehat{x}$.

Observación 3. *Decimos que la sucesión $\{x^{(k)}\} \subset \mathbb{K}^n$ de vectores converge a x, denotado por $\lim_{k\to\infty} \|x^{(k)} - x\| = 0$, si*

$$\forall \epsilon > 0, \exists\, k_0 \in \mathbb{N}, \forall k \geq k_0, \|x^k - x\| < \epsilon.$$

Definición 5. *La función $F : V \to W$, donde V y W son espacios de Banach (espacio vectorial normado y completo), es diferenciable (en el sentido de Frechet) en el punto $x \in V$ si existe una aplicación lineal y continua $J : V \to W$ tal que*

$$\lim_{h\to 0} \frac{\|F(x+h) - F(x) - J(h)\|}{\|h\|} = 0$$

Se deduce que $J(h) = F'(x)h$, donde $F'(x)$ representa la jacobiana de F definida de la siguiente manera:

$$F'(x) = \begin{bmatrix} \partial_1 f_1(x) & \cdots & \partial_n f_1(x) \\ \vdots & \ddots & \vdots \\ \partial_1 f_n(x) & \cdots & \partial_n f_n(x) \end{bmatrix}$$

usamos $\partial_i f_j(x)$ para denotar la derivada parcial de f_j con respecto a la i-ésima variable y evaluada en x, es decir

$$\partial_i f_j(x) = \frac{\partial f_j}{\partial x_i}(x) = \lim_{h\to 0} \frac{f_j(x + he_i) - f_j(x)}{h}$$

y $e_k = (\underbrace{0, \cdots, 0}_{k-1}, 1, \underbrace{0, \cdots, 0}_{n-k})^T$.

Observación 4. *Note que si $n = 1$, entonces $\|F'(x^* + h) - F'(x^*)\|$ se reduce a la definición usual de diferenciabilidad. Note también que si F es diferenciable en x, entonces F es continua en x; esto se sigue de la desigualdad*

$$\|F(x + h) - F(x)\| \leq \|F(x + h) - F(x) - F'(x)h\| + \|F'(x)h\|$$

Finalmente, notamos que es posible mostrar que si la matriz jacobiana es continua en x, entonces F es diferenciable en x.

E. Método de Newton para sistemas de ecuaciones no lineales

Consideremos el problema de resolver el sistema de ecuaciones no lineales

$$f_i(x_1, \cdots, x_n) = 0 \qquad i = 1, \cdots, n$$

que solemos escribir en la forma vectorial como

$$F(x) = 0 \tag{4}$$

donde $F : \mathbb{R}^n \to \mathbb{R}^n$ es la función cuyos componentes son funciones no lineales $f_i : \mathbb{R}^n \to \mathbb{R}$, para $i = 1, 2, \ldots, n$, esto es, $F(x) = (f_1(x), \cdots, f_n(x))^T$.

Uno de los procedimientos básicos de iteración para aproximar una solución de (4) es usando el método de Newton:

$$F(x^k) + F'(X^k)(x^{k+1} - x^k) = 0, \;\; k = 0, 1, \ldots,$$

en el caso que $F'(x^k)^{-1}$ exista, se tiene

$$x^{k+1} = x^k - F'(x^k)^{-1}F(x^k), \quad k = 0, 1, \cdots, \tag{5}$$

En la práctica, por supuesto, no se invierte $F'(x^k)$ para realizar (5), sino que se resuelve el sistema lineal

$$F'(x^k)y = -F(x^k)$$

y se agrega la "corrección"y por x^k y obtenemos

$$x^{k+1} = x^k - y, \quad k = 0, 1, \cdots,$$

En el análisis del método de Newton, es necesario asumir que la matriz jacobiana es, al menos continua en la solución x^*; esto es, $\|F'(x^* + h) - F'(x^*)\| \to 0$ cuando $h \to 0$. Es fácil ver que este será el caso en cualquier norma, si y solo si las derivadas parciales $\partial_i f_i$ son continuas en x^*.

Para empezar nuestro análisis del método de Newton, consideremos la primera iteración general

$$x^{k+1} = G(x^k), \qquad k = 0, 1, \cdots \tag{6}$$

donde $G : \mathbb{R}^n \to \mathbb{R}^n$. La solución de la ecuación $x = G(x)$ es llamada **punto fijo** de G. Cuando G surge como una función de iteración por la ecuación $F(x) = 0$, entonces, la solución de $F(x) = 0$ será siempre un punto fijo de G. Por ejemplo, para el método de Newton, G es dada por:

$$G(x) \equiv x - F'(x)^{-1}F(x)$$

y asumiendo que $F'(x)$ es no singular, x^* es un punto fijo de G si y solo si $F(x^*) = 0$.

Definición 6. *Un punto fijo x^* de $G : \mathbb{R}^n \to \mathbb{R}^n$ es un* ***punto de atracción*** *de la iteración (6) (alternativamente, decimos que la iteración es* ***localmente convergente*** *en x^*) si hay una vecindad abierta S de x^* tal que cuando $x^0 \in S$, las iteraciones de (6) están bien definidas y convergen a x^*.*

F. Convergencia local del método de Newton

Asumimos que $F : \mathbb{R}^n \to \mathbb{R}^n$ es diferenciable en cada punto de una vecindad abierta de una solución x^* de $F(x) = 0$, que F' es continua en x^*, y que $F'(x^*)$ es no singular. Entonces, x^* es un punto de atracción de la iteración (5) y

$$\lim_{k\to\infty} \frac{\|x^{k+1} - x^*\|}{\|x^k - x^*\|} = 0 \tag{7}$$

además, si

$$\|F'(x) - F'(x^*)\| \leq \alpha\|x - x^*\| \tag{8}$$

para todo x en alguna vecindad abierta de x^*, entonces hay una constante $c < +\infty$ tal que

$$\|x^{k+1} - x^*\| \leq c\|x^k - x^*\|^2 \tag{9}$$

para todo $k \geq k_0$, donde k_0 depende de x^0.

La expresión (7) es conocida como **convergencia superlineal**, mientras que (9) es llamada **convergencia cuadrática**. Notemos que (8) es asegurado si las funciones componentes f_i de F son todas dos veces continuamente diferenciables en una vecindad de x^*. Por lo tanto, bajo estos supuestos de diferenciabilidad suave junto con la no singularidad de $F'(x^*)$, la definición de la convergencia local del método de Newton es siempre local y con convergencia cuadrática; esto es, las iteraciones de Newton deben converger a x^* y (9) debe cumplirse, tan pronto como algún x^k esté suficientemente cerca a x^*.

Teorema 4 (Convergencia local del método de Newton)**.** *Sea $F : \mathbb{R}^n \to \mathbb{R}^n$ diferenciable en una vecindad de un punto $\bar{x} \in \mathbb{R}^n$, con derivada continua en ese punto. Sea $\bar{x}$ una solución de la ecuación $F(x) = 0$, tal que la matriz $F'(\bar{x})$ es no singular. Entonces, para cualquier punto inicial $x^0 \in \mathbb{R}^n$ suficientemente próximo a $\bar{x}$, el algoritmo del método de Newton genera una secuencia $\{x^k\}$ bien definida, que converge a $\bar{x}$. La convergencia es superlineal, y si la derivada de F es Lipschitz-continua en una vecindad de $\bar{x}$, entonces la convergencia es cuadrática.*

Demostración. Ver [17] □

G. El problema de valor inicial

Definición 7. *Una función diferenciable $\psi : I \to \mathbb{R}$ es llamada solución de la ecuación*

$$\frac{dy}{dt} = f(t, y) \tag{10}$$

en el intervalo I, donde $f : \Omega \subset \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^n$, si:

1. *el gráfico de ψ, esto es, $\{(t, \psi(t)) \mid t \in I\}$ está contenido en Ω y*
2. *$\frac{d\psi}{dt}(t) = f(t, \psi(t))$ para todo $t \in I$. Si t es un extremo del intervalo, la derivada es la derivada lateral respectiva.*

Definición 8. *La ecuación diferencial, es un problema de Cauchy si*

$$\frac{dy}{dt} = f(t, y),\ y(t_0) = y_0. \tag{11}$$

en el intervalo I, donde $f : \Omega \subset \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^n$.

Una aplicación $f : \Omega \subset \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^n$ se llama lipschitziana en Ω relativamente a la segunda variable si existe una constante $k \in \mathbb{R}$ tal que:

$$\|f(t,x) - f(t,y)\| \leq k\|x - y\|,$$

para todo $(t, y) \in \Omega$, k es llamada constante de Lipschitz de f.

Lema 1 (Lema de contracción)**.** *Sean $X \subset \mathbb{R}^n$ y $F : X \to X$ una contracción, esto es:*

$$d(F(x), F(y)) \leq k\, d(x, y), 0 \leq k < 1$$

Existe un único punto fijo p para F, esto es, $F(p) = p$.

Teorema 5 (Teorema de Picard)**.** *Sea f un función continua y lipschitziana en $\Omega = I_0 \times B_b$, donde $I_0 = \{t \mid |t - t_0| \leq a\}$, $B_b = \{x \mid \|x - x_0\| \leq b\}$. Si $\|f\| \leq M$ en Ω, existe una y solo una solución de*

$$y'(t) = f(t,x),\ y(t_0) = y_0$$

en I_0, donde $\alpha = \min\{a, b/M\}$.

Demostración. Ver [27] □

Definición 9. *El problema de valor inicial*

$$\frac{dy}{dt} = f(t,y),\ \ a \leq t \leq b, y(a) = \alpha, \tag{12}$$

*se dice que es un **problema bien planteado** si:*

- *Existe una única solución, $y(t)$.*
- *Existen constantes $\epsilon_0 > 0$ y $k > 0$, tales que, para cualquier $\epsilon \in (0, \epsilon_0)$, siempre que $\delta(t)$ es continua con $|\delta(t)| < \epsilon$ para toda $t \in [a, b]$, y cuando $|\delta_0| < \epsilon$, el*

problema de valor inicial

$$\frac{dz}{dt} = f(t,z) + \delta(t), \;\; a \leq t \leq b, \;\; z(a) = \alpha + \delta_0,$$

tiene una única solución $z(t)$ *que satisface* $\|z(t) - y(t)\| < k\epsilon$*, para todo* $t \in [a,b]$.

II. Métodos de un paso y funciones estabilizadoras para problemas rígidos

I. Fundamento teórico En este capítulo se abordarán los métodos de un paso, que son fundamentales para la resolución numérica de ecuaciones diferenciales, especialmente en el contexto de problemas rígidos. La rigidez en sistemas de ecuaciones puede llevar a inestabilidades en los métodos de solución tradicionales, por lo que es crucial entender y aplicar técnicas adecuadas que garanticen resultados precisos y estables.

Comenzaremos con una revisión de los métodos de un paso más utilizados, como el método de Euler, el método de expansión de Taylor, el método de Euler mejorado y el método de Runge-Kutta. Cada uno de estos métodos será analizado en términos de su formulación, ventajas y desventajas, destacando su aplicación a diferentes tipos de problemas.

A continuación, se discutirá la descripción global de estos métodos, poniendo énfasis en conceptos clave como estabilidad, convergencia y el error global asintótico. La comprensión de estos aspectos es esencial para seleccionar el método más adecuado según las características del problema a resolver. También se presentarán técnicas para estimar el error global, lo cual es fundamental para garantizar la precisión en los resultados.

El capítulo incluirá una sección dedicada a la estimación del error de truncamiento y a

la extrapolación local de Richardson, así como a los métodos integrados, que ofrecen enfoques alternativos para mejorar la precisión y la eficiencia computacional. Además, se abordará el control del paso, una estrategia que permite ajustar el tamaño del paso en función de la dinámica del sistema, lo que resulta crucial en problemas rígidos.

Finalmente, se explorarán los problemas de rigidez en profundidad, así como el concepto de A-estabilidad y la aproximación de Padé, junto con las regiones de absoluta estabilidad. Estos temas son vitales para entender las limitaciones y capacidades de los métodos de un paso al abordar ecuaciones diferenciales desafiantes. Este capítulo proporcionará una base sólida para el desarrollo de técnicas avanzadas en los siguientes capítulos, contribuyendo a una mejor comprensión de la resolución numérica de ecuaciones diferenciales. Para mayor detalle, consultar los libros de Gautschi [12] y Burden [3].

Sea el problema de valor inicial

$$\frac{dy}{dt} = y' = f(t, y), \quad a \leq t \leq b, \quad y(a) = \alpha. \tag{I.1}$$

Definimos la malla de $[a, b]$, como el conjunto de puntos $P = \{t_0, t_1, \ldots, t_n\}$, donde $a = t_0 < t_1 < t_1 < \cdots < x_{N-1} < x_N = b$, el cual se pretende aproximar la solución exacta de la ecuación diferencial (I.1) (con $y(x)$) para los puntos $t = t_i$, con $i = 0, 1, 2, \ldots, N$. En el caso que los puntos estén igualmente espaciados tendremos que $h = t_{i+1} - t_i$ (que recibe el nombre tamaño de paso), obteniendo los puntos $t_i = a + ih$, para cada $i = 0, 1, \ldots, N$. Dados los puntos $x \in [a, b]$, $y \in \mathbb{R}^n$, se definen los métodos de un solo paso o métodos Φ por

$$\begin{aligned} y(x + h) &= y(x) + h\Phi(x, y; h), h > 0 \\ y(a) &= \alpha. \end{aligned} \tag{I.2}$$

donde $\Phi : [a, b] \times \mathbb{R}^n \times \mathbb{R}_+ \to \mathbb{R}$.

Definición 10. *El error de truncamiento del método* Φ *en el punto* (x, y) *es definido por*

$$\tau(x, y; h) = \frac{1}{h}[y(x + h) - u(x + h)]. \tag{I.3}$$

donde u *es la solución de la ecuación diferencial (I.1).*

De la definición, tenemos que:

$$\tau(x, y; h) = \Phi(x, y; h) - \frac{1}{h}[u(x + h) - u(x)],$$

donde $u(x) = \alpha$. Una descripción cada vez más fina de la precisión local es proporcionada por las siguientes definiciones, basadas en el concepto de error de truncamiento.

Definición 11. *El método* Φ *es llamado consistente si:*

$$\tau(x, y; h) \to 0, \textit{cuando } h \to 0, \tag{I.4}$$

para cualquier $(x, y) \in [a, b] \times \mathbb{R}^n$.

Definición 12. *El método* Φ *se dice que tiene orden* p*, si:*

$$\|\tau(x, y; h)\| \leq Ch^p,$$

para cualquier $(x, y) \in [a, b] \times \mathbb{R}^n$*, con una constante* C *que no depende de* h, x *e* y*.*

Nosotros entenderemos que Φ es orden p si $\tau(x, y; h) = O(h^p)$, $h \to 0$.

Definición 13. *Una función* $T : [a, b] \times \mathbb{R}^n \to \mathbb{R}^n$ *que satisface* $T(x, y) \neq 0$ *y*

$$\tau(x, y; h) = T(x, y)h^p + O(h^{p+1}), \quad h \to 0$$

es llamada función de error principal.

La función de error principal determina el término principal en el error de truncamiento. Veamos ahora algunos métodos de un paso.

A. Métodos de un paso

1. Método de Euler

Euler propuso su método en el año 1768. Consiste simplemente en seguir la pendiente en el punto genérico (x, y) en un intervalo de longitud h:

$$y(x+h) = y(x) + hf(x, y).$$

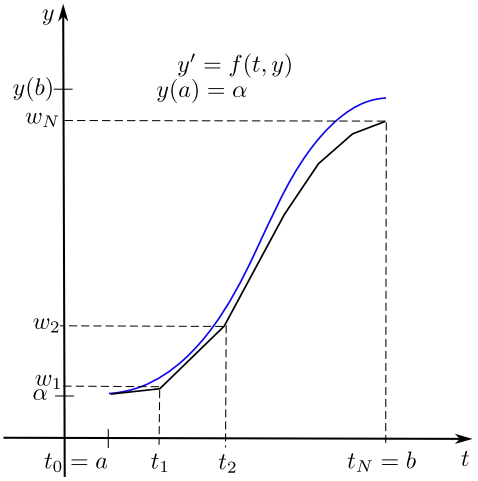

Figura I: Interpretación geométrica de Euler.
Adaptado de [3].

Entonces, $\Phi(x, y, h) = f(x, y)$ no depende de h y desde que $\Phi(x, y, h) = f(x, y)$ el método es evidentemente consistente. El error de truncamiento,

$$\tau(x, y, h) = f(x, y) - \frac{1}{h}[u(x+h) - u(x)]$$

donde $u(x)$ es la solución referente definida en (I.1). Desde que $u'(x) = f(x, u(x)) =$

$f(x, y)$, podemos escribir, usando el teorema de Taylor.

$$\begin{aligned}\tau(x, y, h) &= u'(x) - \frac{1}{h}[u(x + h) - u(x)] \\ &= u'(x) - \frac{1}{h}\left[u(x) + hu'(x) + \frac{1}{2}h^2u''(\epsilon) - u(x)\right] \\ &= -\frac{1}{2}hu''(\epsilon), \qquad x < \epsilon < x + h\end{aligned}$$

Asumiendo que $u \in C^2[x, x + h]$, luego, derivando la ecuación (I.1) con respecto a x y finalmente, haciendo $x = \epsilon$ se tiene:

$$\tau(x, y, h) = -\frac{1}{2}h[f_x + f_yf](\epsilon, u(\epsilon))$$

Si del teorema (5), asumimos que f y todas las primeras derivadas parciales son acotadas en $[a, b] \times \mathbb{R}^d$, entonces existe una constante C independiente de x, y y h tal que

$$\|\tau(x, y, h)\| \leq Ch$$

Así, el método de Euler tiene orden $p = 1$. Si asumimos lo mismo sobre las segundas derivadas parciales de f, tenemos $u''(\epsilon) = u''(x) + O(h)$ y, entonces, de (I.1)

$$\tau(x, y, h) = -\frac{1}{2}h[f_x + f_yf](x, y) + O(h^2), \qquad h \to 0$$

Vemos que el error principal de la función es dado por:

$$T(x, y) = -\frac{1}{2}[f_x + f_yf](x, y)$$

A menos que $f_x + f_yf = 0$, el orden del método de Euler es exactamente $p = 1$.

2. Método de expansión de Taylor

Hemos visto que el método de Euler consiste básicamente en truncar la expansión de Taylor de la solución de referencia después de su segundo término. En cambio, el método

de Taylor usa más términos de expansión. Esto requiere del cálculo de derivadas sucesivas de f.

$$f^{[0]}(x,y) = f(x,y)$$

$$f^{[k+1]}(x,y) = f_x^{[k]}(x,y) + f_y^{[k]}(x,y), \qquad k = 0, 1, 2, \cdots$$

que determinan las derivadas sucesivas de la solución de referencia $u(x)$ de (I.1), en virtud de:

$$y^{(k+1)}(x) = f^{[k]}(x,y), \qquad k = 0, 1, 2, \cdots \tag{I.5}$$

y usamos la forma de la serie de aproximación de Taylor de acuerdo a:

$$y(x+h) = y(x) + h\left[f^{[0]}(x,y) + \frac{1}{2}hf^{[1]}(x,y) + \cdots + \frac{1}{p!}h^{[p-1]}f^{[p-1]}(x,y)\right] \tag{I.6}$$

Esto es:

$$\Phi(x,y,h) = f^{[0]}(x,y) + \frac{1}{2}hf^{[1]}(x,y) + \cdots + \frac{1}{p!}h^{p-1}f^{[p-1]}(x,y)$$

Para el error de truncamiento, asumimos que $f \in C^p$ en $[a,b] \times \mathbb{R}^d$ y usando (I.5) y (I.6) obtenemos:

$$\begin{aligned}
\tau(x,y,h) &= \Phi(x,y,h) - \frac{1}{h}[u(x+h) - u(x)] \\
&= \Phi(x,y,h) - \sum_{k=0}^{p-1} u^{(k+1)}(x)\frac{h^k}{(k+1)!} - u^{(p+i)}(\epsilon)\frac{h^p}{(p+1)!} \\
&= -u^{(p+1)}(\epsilon)\frac{h^p}{(p+1)!}, \quad x < \epsilon < x+h,
\end{aligned}$$

de modo que:

$$\|\tau(x,y,h)\| \leq \frac{C_p}{(p+1)!}h^p$$

donde C_p es el límite de la p-ésima derivada de f. Así, este método tiene exactamente orden p (a menos que $f^{[p]}(x,y) = 0$), y el error principal de la función es:

$$T(x,y) = -\frac{1}{(p+1)!}f^{[p]}(x,y)$$

La necesidad de calcular muchas derivadas parciales fue un factor desalentador para este método. Es por ello que se trata de buscar métodos que no utilice muchas derivadas y que mejoren las aproximaciones.

3. Método de Euler mejorado

Debido a que el método de Euler sigue una dirección de la recta tangente a la solución, se observa que las aproximaciones a la solución de la ecuación se van alejando. Debido a esto, se sugiere reevaluar la pendiente a la mitad del segmento de recta generado, retomando una aproximación a la solución de la ecuación diferencial y luego continuar con esta pendiente en todo el intervalo, es decir:

$$y(x+h) = y(x) + hf\left(x + \frac{1}{2}h, y + \frac{1}{2}hf(x,y)\right)$$

ó

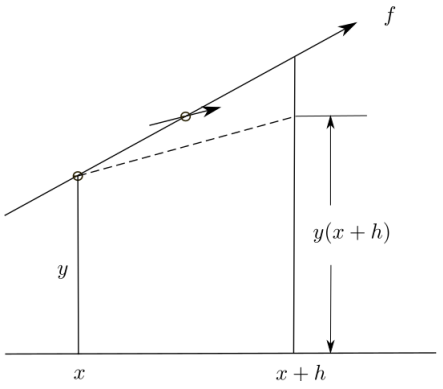

Figura II: Método de Euler mejorado.
Adaptado de [12].

$$\Phi(x,y,h) = f\left(x + \frac{1}{2}h, y + \frac{1}{2}hf(x,y)\right)$$

Otra forma de ver este método es la siguiente:

$$k_1(x,y) = f(x,y),$$
$$k_2(x,y;h) = f\left(x + \frac{1}{2}h, y + \frac{1}{2}hk_1\right),$$
$$y(x+h) = y(x) + hk_2$$

Si tomamos como la segunda pendiente en el punto $(x+h, y+hf(x,y))$, y haciendo un promedio de ellas obtenemos:

$$k_1(x,y) = f(x,y)$$
$$k_2(x,y;h) = f(x+h, y+hk_1),$$
$$y(x+h) = y(x) + \frac{1}{2}h(k_1+k_2)$$

Esto, a veces se conoce como el método de Heun o la regla trapezoidal.

Podemos tomar un enfoque más sistemático y modificar el método de Euler, y obtener los métodos de dos etapas. Escribimos

$$\Phi(x,y;h) = \alpha_1 k_1 + \alpha_2 k_2$$

donde

$$k_1(x,y) = f(x,y)$$
$$k_2(x,y;h) = f(x+\mu h, y+\mu h k_1)$$

Tenemos ahora tres parámetros α_1, α_2 y μ a nuestra disposición, y podemos intentar elegir para maximizar el orden. Una forma sistemática de determinar el orden máximo p es expandir ambos $\Phi(x,y;h)$ y $\frac{1}{h}[u(x+h) - u(x)]$ en términos de h, y unir tantos términos como podamos.

Para expandir Φ, necesitamos la expansión de Taylor,

$$f(x+\triangle x, y+\triangle y) = f+f_x\triangle x+f_y\triangle y+\frac{1}{2}[f_{xx}(\triangle x)^2+2f_{xy}\triangle x\triangle y+(\triangle y)^T f_{yy}(\triangle y)]+\cdots \tag{I.7}$$

donde f_y denota la jacobiana de f y $f_{yy} = [f'_{yy}]$ el vector de la matriz hessiana de f (para un estudio más detallado, ver [21]). En (I.7), tomamos $\triangle x = \mu h, \triangle y = \mu h f$. Entonces, se tiene:

$$k_2(x, y; h) = f + \mu h(f_x + f_y f) + \frac{1}{2}\mu^2 h^2(f_{xx} + 2f_{xy}f + f^T f_{yy} f) + O(h^3) \tag{I.8}$$

Similarmente

$$\frac{1}{h}[u(x + h) - u(x)] = u'(x) + \frac{1}{2}hu''(x) + \frac{1}{6}h^2u'''(x) + O(h^3) \tag{I.9}$$

donde

$$\begin{aligned}
u'(x) &= f, \\
u''(x) &= f^{[1]} = f_x + f_y f, \\
u'''(x) &= f^{[2]} = f_x^{[1]} + f_y^{[1]} f \\
&= f_{xx} + f_{xy}f + f_y f_x + (f_{xy} + (f_y f)_y)f \\
&= f_{xx} + 2f_{xy}f + f^T f_{yy} f + f_y(f_x + f_y f)
\end{aligned}$$

y donde, en la última ecuación hemos usado:

$$(f_y f)_y f = f^T f_{yy} f + f_y^2 f$$

Ahora

$$\tau(x, y; h) = \alpha_1 k_1 + \alpha_2 k_2 - \frac{1}{h}[u(x + h) - u(x)]$$

donde sustituimos las expansiones (I.8) y (I.9). Encontramos:

$$\tau(x,y;h) = (\alpha_1+\alpha_2-1)f + \left(\alpha_2\mu - \frac{1}{2}\right)h(f_x+f_yf) + \frac{1}{2}h^2\left[\alpha_2\mu^2 - \frac{1}{3}\right.$$
$$\left.\times(f_{xx}+2f_{xy}f+f^Tf_{yy}f) - \frac{1}{3}f_y(f_x+f_yf)\right] + O(h^3)$$

Por lo tanto, se obtiene $p=2$, siempre y cuando:

$$\alpha_1+\alpha_2 = 1$$
$$\alpha_2\mu = \frac{1}{2}$$

Esto tiene una familia de soluciones de un parámetro,

$$\alpha_1 = 1-\alpha_2$$
$$\mu = \frac{1}{2\alpha_2}, \qquad (\alpha_2 \neq 0, \text{ arbitrario})$$

En el caso que $\alpha_2 = 1$ se obtiene el método de Euler mejorado y cuando $\alpha_2 = \frac{1}{2}$ se obtiene el Método de Heun. Hay otras elecciones naturales; uno de ellos sería la función de error principal

$$T(x,y) = \frac{1}{2}\left[\left(\frac{1}{4\alpha_2} - \frac{1}{3}\right)(f_{xx}+2f_{xy}f+f^Tf_{yy}f) - \frac{1}{3}f_y(f_x+f_yf)\right]$$

vemos que consiste en una combinación lineal de dos agregados de derivadas parciales.

4. Método de Runge Kutta

Los métodos de Runge Kutta son una extensión directa de los métodos de dos etapas a métodos de r-etapas, las cuales se definen de la siguiente forma:

$$\begin{aligned}
\Phi(x,y;h) &= \sum_{s=1}^{r}\alpha_s k_s \\
k_1(x,y) &= f(x,y) \\
k_s(x,y;h) &= f\left(x+\mu_s h, y+h\sum_{j=1}^{s-1}\lambda_{sj}k_j\right), \quad s=2,3,\ldots,r.
\end{aligned} \tag{I.10}$$

Es natural imponer en (I.10) las condiciones

$$\mu_s = \sum_{j=1}^{s-1} \lambda_{sj},\ s = 2, 3, \ldots, r;\ \sum_{s=1}^{r} \alpha_s = 1,$$

donde la última no es más que la condición de consistencia. Llamamos a (I.10) un método explícito de Runge Kutta en r-etapas. De manera más general, podemos considerar el método implícito de Runge Kutta de r etapas:

$$\begin{aligned} \Phi(x, y; h) &= \sum_{s=1}^{r} \alpha_s k_s(x, y; h) \\ k_s(x, y; h) &= f\left(x + \mu_s h, y + h \sum_{j=1}^{r} \lambda_{sj} k_j\right),\ s = 1, 2, \ldots, r, \end{aligned}$$

en el que las últimas r ecuaciones forman un sistema de ecuaciones (en general no lineales) con incógnitas $k_1, k_2, \ldots, k_r$. La razón de los métodos implícitos es que no solo pueden tener un orden más alto que los métodos explícitos, sino que también tienen mejores propiedades de estabilidad.

En el año 1901, Kutta demostró que el orden del método explícito de Runge Kutta de r-etapas es r. Por ejemplo, el método de Runge Kutta explícito de orden $p = 4$ es dado:

$$\begin{aligned} \Phi(x, y; h) &= \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4) \\ k_1(x, y) &= f(x, y) \\ k_2(x, y; h) &= f\left(x + \frac{1}{2}h, y + \frac{1}{2}hk_1\right) \\ k_3(x, y; h) &= f\left(x + \frac{1}{2}h, y + \frac{1}{2}hk_2\right) \\ k_4(x, y; h) &= f\left(x + h, y + hk_3\right). \end{aligned}$$

B. Descripción global de los métodos de un paso

Pasamos ahora a la solución numérica del problema de valor inicial (I.1) con los métodos de un paso desarrollados en la sección anterior. Para ello, definamos una cuadrícula

en el intervalo $[a, b]$ como un conjunto de puntos $\{x_n\}_{n=0}^{N}$ tal que:

$$a = x_0 < x_1 < \cdots < x_{N-1} < x_N = b,$$

con longitud de malla h_n definido por:

$$h_n = x_{n+1} - x_n, \;\; n = 0, 1, \ldots, N-1.$$

Se define la norma de la malla como:

$$|h| = \max_{0 \leq n \leq N-1} h_n.$$

A menudo, usamos la letra h para designar la colección de longitudes $h = \{h_n\}$. Si $h_1 = h_2 = \cdots = h_{N-1} = \dfrac{b-a}{N}$, llamaremos malla uniforme, en otro caso, malla no uniforme. Un vector $v = \{v_n\}$, $v_n \in \mathbb{R}^n$, definido sobre la malla de $[a, b]$ es llamada una función malla. Así, v_n es el valor de v un punto de malla x_n. Cada función $v(x)$ definido sobre $[a, b]$ induce una malla de restricción. Denotemos el conjunto de funciones malla $[a, b]$ por $\Gamma_h[a; b]$ y para la función malla $v = \{v_n\}$ se define la norma:

$$\|v\|_\infty = \max_{0 \leq n \leq N} \|v_n\|, \quad v \in \Gamma_h[a, b]. \tag{I.11}$$

Dado una secuencia $\{\Phi_n\}$ el método de un paso procede como sigue,

$$\begin{aligned} x_{n+1} &= x_n + h_n, \\ u_{n+1} &= u_n + h_n \Phi_n(x_n, u_n; h_n), \quad n = 0, 1, \ldots, N-1, \end{aligned} \tag{I.12}$$

donde $x_0 = a$, $u_0 = y_0$. Ahora, introducimos un operador R y R_h sobre $C^1[a, b]$ y $\Gamma_h[a, b]$, respectivamente. Estos son los operadores residuales

$$(Rv)(x) = v'(x) - f(x, v(x)), \quad v \in C^1[a, b], \tag{I.13}$$

$$\begin{aligned} (R_h v)_n &= \frac{1}{h_n}(v_{n+1} - v_n) - \Phi(x_n, v_n; h_n), \;\; n = 0, 1, \ldots, N-1; \\ v &= \{v_n\} \in \Gamma_h[a, b]. \end{aligned} \tag{I.14}$$

Entonces, el problema de valor inicial y el método un paso es dado por:

$$\begin{aligned} Ry &= 0, \text{ sobre } [a,b], \quad y(a) = y_0, \\ R_h u &= 0, \text{ sobre } [a,b], \quad u_0 = y_0. \end{aligned}$$

Luego, efectivamente la solución referente $u(t)$ coincide con la solución $y(t)$, y

$$(R_h y)_n = \frac{1}{h_n}[y(x_{n+1}) - y(x_n)] - \Phi(x_n, y(x_n); h; n) = -\tau(x_n, y(x_n); h_n).$$

Para más detalles, ver [12].

1. Estabilidad

La estabilidad es una propiedad solamente del esquema numérico (I.12) y a priori no tiene nada que ver con su poder de aproximación. Caracteriza lo fuerte de un esquema con respecto a pequeñas perturbaciones.

Definimos la estabilidad en términos de un operador discreto residual R_h en (I.14). Como es usual, asumimos $\Phi(x, y; h)$ definido sobre $[a, b] \times \mathbb{R}^n \times [0, h_0]$, donde $h_0 > 0$ es algún número positivo adecuado.

Definición 14. *El método (I.12) es llamado estable sobre* $[a, b]$ *si existe una constante* $K > 0$ *que no depende de* h*, tal que para una malla arbitraria* h *sobre* $[a, b]$*, y para dos funciones malla arbitrarias* $v, w \in \Gamma_h[a, b]$*, se mantiene*

$$\|v - w\|_\infty \leq K(\|v_0 - w_0\| + \|R_h v - R_h w\|_\infty), \ v, w \in \Gamma_h[a, b], \tag{I.15}$$

para todo h *con* $|h|$ *suficientemente pequeño. En (I.15), la norma infinita para una función malla es la norma definida en (I.11).*

Referimos a (I.15) como la desigualdad estable. Supongamos que tuviéramos dos funciones mallas u, w satisfaciendo

$$R_h u = 0, u_0 = y_0 \tag{I.16}$$

$$R_h w = \epsilon, w_0 = y_0 + \nu_0, \tag{I.17}$$

donde $\epsilon = \{\epsilon_n\} \in \Gamma_h[a,b]$ es una función malla con $\|\epsilon_n\|$, pequeño, y $\|\nu_0\|$ es también pequeño. Podemos interpretar $u \in \Gamma_h[a,b]$ como el resultado de aplicar el esquema numérico (I.12) como buena aproximación, mientras $w \in \Gamma_h[a,b]$ podría ser la solución de (I.12) en la aritmética de punto flotante. Entonces, si se mantiene la estabilidad, tenemos

$$\|u - w\|_\infty \leq K(\|\nu_0\| + \|\epsilon\|_\infty),$$

es decir, el cambio global en u es del mismo orden de magnitud como el error residual $\{\epsilon_n\}$ y error inicial ν_0 debe ser despreciado. Sin embargo, la primera ecuación en (I.17) dice que $w_{n+1} - w_n - h_n\Phi(x, y; h_n) = h_n\epsilon_n$, lo que significa que los errores de redondeo deben ir a cero cuando $|h| \to 0$.

Teorema 6. *Si $\Phi(x, y; h)$ satisface la condición de Lipschitz con respecto a la variable y, además:*

$$\|\Phi(x, y; h) - \Phi(x, y^*; h)\| \leq M\|y - y^*\|, \text{ sobre } [a,b] \times \mathbb{R}^n \times [0, h_0],$$

entonces el método (I.12) es estable.

La demostración de este teorema recae del siguiente resultado.

Lema 2. *Sea $\{e_n\}$ una secuencia de números $e_n \in \mathbb{R}$ satisfaciendo*

$$e_{n+1} \leq a_n e_n + b_n, \quad n = 0, 1, \ldots, N-1,$$

donde $a_n > 0$ *y* $b_n \in \mathbb{R}$. *Entonces*

$$e_n \leq E_n,\ E_n = \left(\prod_{k=0}^{n-1} a_k\right) e_0 + \sum_{k=0}^{n-1} \left(\prod_{l=k+1}^{n-1} a_l\right) b_k,\ \ n = 0, 1, \ldots, N.$$

Para la demostración, ver [12].

Lema 3. *Sea* $v \in \Gamma_h[a, b]$ *satisfaciendo*

$$v_{n+1} = v_n + h_n(A_n v_n + b_n), \quad n = 0, 1, \ldots, N-1,$$

donde $A_n \in \mathbb{R}^{n\times n}$, $b_n \in \mathbb{R}^n$, *y* $h = \{h_n\}$ *es una malla arbitraria sobre* $[a, b]$. *Suponiendo que*

$$\|A_n\| \leq M, \qquad \|b_n\| \leq \delta, \quad n = 0, 1, 2, \ldots, N-1$$

donde las constantes M *y* δ *no dependen de* h. *Entonces existe una constante* $K > 0$ *independiente de* h, *pero dependiente de* $\|v_0\|$, *tal que*

$$\|v\|_\infty \leq K.$$

2. Convergencia

La estabilidad es un concepto bastante poderoso, implica una convergencia casi inmediata y también es fundamental para derivar estimaciones de errores globales asintóticos. Definimos la convergencia de un método de un paso de la siguiente manera.

Definición 15. *Sea* $a = x_0 < x_1 < x_2 < \cdots < x_N = b$ *una malla sobre* $[a, b]$ *con longitud de malla* $|h| = \max\limits_{1\leq x\leq N}(x_n - x_{n-1})$. *Sea* $u = \{u_n\}$ *una función malla definida aplicando el método (I.12) sobre* $[a, b]$ *y la función malla* $y = \{y_n\}$ *inducida por la solución exacta del problema de valor inicial (I.1). El método (I.12) se dice que converge sobre* $[a, b]$ *si*

$$\|u - y\|_\infty \to 0,\ \textit{cuando}\ |h| \to 0.$$

Teorema 7. *Si el método (I.12) es consistente y estable sobre* $[a, b]$*, entonces este converge. Además, si* Φ *tiene orden* p*, entonces:*

$$\|u - y\|_\infty = O(|h|^p), \text{ cuando } |h| \to 0.$$

3. Error global asintótico

Así como la función de error principal describe la contribución principal al error de truncamiento local, es de interés identificar el término principal en el error global $u_n - y(x_n)$. Para simplificar las cosas, asumimos una longitud de cuadrícula constante h, aunque no sería difícil tratar con longitudes de cuadrícula variables de la forma $h_n = \mathcal{O}(x_n)h$, donde $\mathcal{O}(x)$ es continuo por partes y $0 < \mathcal{O}(x) \leq \theta$, para $a \leq x \leq b$. Así, consideramos el método de un solo paso

$$\begin{aligned} x_{n+1} &= x_n + h, \\ u_{n+1} &= u_n + h\Phi(x_n, u_n; h), \quad n = 0, 1, \ldots, N-1, \\ x_0 &= a, u_0 = y_0, \end{aligned}$$

definiendo un función malla $u = \{u_n\}$ sobre una malla uniforme en $[a, b]$. Estamos interesados en el comportamiento asintótico de $u_n - y(x_n)$ cuando $h \to 0$, donde $y(x)$ es la solución exacta del problema del valor inicial

$$\frac{dy}{dx} = f(x, y), \; a \leq x \leq b, \;\; y(a) = y_0. \tag{I.18}$$

Teorema 8. *Asumamos que*

1. $\Phi(x, y; h) \in C^2$ *sobre* $[a, b] \times \mathbb{R}^n \times [0, h_0]$.

2. Φ *es un método de orden* $p \geq 1$ *admitiendo una función de error principal* $T(x, y) \in C$ *sobre* $[a, b] \times \mathbb{R}^n$.

3. *$e(x)$ es la función lineal del problema valor inicial*

$$\begin{aligned} \frac{de}{dx} &= f_y(x, y(x))e + T(x, y(x)), \quad a \leq x \leq b \\ e(a) &= 0. \end{aligned} \tag{I.19}$$

Entonces, para $n = 0, 1, \ldots, N$,

$$u_n - y(x_n) = e(x_n)h^p + O(h^{p+1}), \text{ cuando } h \to 0. \tag{I.20}$$

Para la demostración, ver [12]. Veamos algunos comentarios del teorema.

- El significado del resultado es:

$$\|u - y - h^p e\|_\infty = O(h^{p+1}), \text{ cuando } h \to 0,$$

  donde y, e y u son funciones mallas $u = \{u_n\}$, $y = \{y_n\}$, $e = \{e(x_n)\}$ y $\|\cdot\|_\infty$ es la norma definida por (I.11).

- Ya que por consistencia $\Phi(x, y; 0) = f(x, y)$, asumimos que el inciso 1 del teorema 8 implica que $f \in \mathbb{C}^2$ sobre $[a, b] \times \mathbb{R}^n$, el cual es más que suficiente para garantizar la existencia y unicidad de la solución $e(x)$ sobre el intervalo $[a, b]$.

- El hecho de que algunas, pero no todas, las componentes de $T(x, y)$ puede desaparecer no implica que las componentes correspondientes de $e(x)$ también desaparezcan, ya que (I.19) es un sistema acoplado de ecuaciones diferenciales.

4. Estimación del error global

La idea de nuestra estimación es integrar la "ecuación variacional" (I.19) junto con la ecuación principal (I.18). Dado que necesitamos $e(x_n)$ en (I.20) solo con una precisión

de $O(h)$ (cualquier término de error $O(h)$ en $e(x_n)$, multiplicado por h^p, es absorbido por el término $O(h^{p+1})$), podemos usar el método de Euler para ese propósito, el cual proporcionará la aproximación deseada $v_n \approx e(x_n)$.

Teorema 9. *Asumamos que*

1. $\Phi(x, y; h) \in C^2$ *sobre* $[a, b] \times \mathbb{R}^d \times [0, h_0]$,

2. Φ *es un método de orden* $p \geq 1$ *admitiendo una función de error principal* $T(x, y) \in \mathcal{C}^1$ *en* $[a, b] \times \mathbb{R}^n$.

3. *una estimación* $r(x, y; h)$ *es permitida para una función de error principal que satisface*

$$r(x, y; h) = T(x, y) + O(h), \quad h \to 0,$$

uniformemente sobre $[a, b] \times \mathbb{R}^n$,

4. *a lo largo de función malla* $u = \{u_n\}$ *generamos la función malla* $v = \{v_n\}$ *de la siguiente manera,*

$$\begin{aligned} x_{n+1} &= x_n + h, \\ u_{n+1} &= u_n + h\Phi(x_n, u_n; h), \\ v_{n+1} &= v_n + h[f_y(x_n, u_n)v_n + r(x_n, u_n; h)], \\ x_0 &= a, u_0 = y_0, v_0 = 0. \end{aligned}$$

Entonces, para $n = 0, 1, \ldots, N$, *se tiene* $u_n - y(x_n) = v_n h^p + O(h^{p+1})$ *cuando* $h \to 0$.

Para la demostración, ver [12].

C. Estimación del error de truncamiento

1. Extrapolación local de Richardson

Esto funciona para cualquier método de un paso Φ, pero generalmente es considerado caro. Si Φ tiene orden p, el proceso es el siguiente:

$$\begin{aligned} y_h &= y + h\Phi(x, y; h), \\ y_{h/2} &= y + \frac{1}{2}h\Phi\left(x, y; \frac{1}{2}h\right), \\ y_h^* &= y_{h/2} + \frac{1}{2}h\Phi\left(x + \frac{1}{2}h, y_{h/2}; \frac{1}{2}h\right). \end{aligned} \tag{I.21}$$

Notemos que y_h^* es el resultado de aplicar Φ sobre dos pasos consecutivos de longitud $\frac{1}{2}h$, mientras y_h es el resultado de un aplicación sobre la longitud de paso h.

Ahora, verificamos que $r(x, y; h)$ en (I.21) es un estimador aceptable. Para esto, necesitamos asumir que $T(x, y) \in C^1$ sobre $[a, b] \times \mathbb{R}^n$. En términos de solución referente $u(t)$ a través (x, y), tenemos

$$\Phi(x, y; h) = \frac{1}{h}[u(x + h) - u(x)] + T(x, y)h^p + O(h^{p+1}). \tag{I.22}$$

Por lo tanto,

$$\begin{aligned} \frac{1}{h}(y_h - y_h^*) &= \frac{1}{h}(y - y_{h/2}) + \Phi(x, y; h) - \frac{1}{2}\Phi\left(x + \frac{1}{2}h, y_{h/2}; \frac{1}{2}h\right) \\ &= \Phi(x, y; h) - \frac{1}{2}\Phi\left(x, y; \frac{1}{2}h\right) - \frac{1}{2}\Phi\left(x + \frac{1}{2}h, y_{h/2}; \frac{1}{2}h\right). \end{aligned}$$

Aplicando (I.22) a cada uno de los tres términos del lado derecho, encontramos

$$\begin{aligned}\frac{1}{h}(y_h - y_h^*) &= \frac{1}{h}[u(x+h) - u(x)] + T(x,y)h^p + O(h^{p+1}) \\ &\quad -\frac{1}{2}\frac{1}{h/2}\left[u\left(x+\frac{1}{2}h\right) - u(x)\right] - \frac{1}{2}T(x,y)\left(\frac{1}{2}h\right)^p + O(h^{p+1}) \\ &\quad -\frac{1}{2}\frac{1}{h/2}\left[u(x+h) - u\left(x+\frac{1}{2}h\right)\right] \\ &\quad -\frac{1}{2}T\left(x+\frac{1}{2}h, y+O(h)\right)\left(\frac{1}{2}h\right)^p \\ &\quad +O(h^{p+1}) = T(x,y)(1-2^{-p})h^p + O(h^{p+1}).\end{aligned}$$

Consecuentemente,

$$\frac{1}{1-2^{-p}}\frac{1}{h}(y_h - y_h^*) = T(x,y)h^p + O(h^{p+1}) \tag{I.23}$$

entonces la estimación requerida es dada por

$$r(x,y;h) = \frac{1}{1-2^{-p}}\frac{1}{h^{p+1}}(y_h - y_h^*).$$

Restando (I.23) de (I.22) tenemos:

$$\Phi^*(x,y;h) = \Phi(x,y;h) - \frac{1}{1-2^{-p}}\frac{1}{h}(y_h - y_h^*)$$

y así, se define un método de un paso de orden $p+1$.

El procedimiento en (I.21) es bastante costoso. Para un proceso de Runge-Kutta de cuarto orden, requiere un total de 11 evaluaciones de f por paso, casi tres veces el esfuerzo para un solo paso de Runge-Kutta. Por lo tanto, la extrapolación de Richardson normalmente se usa solo después de cada dos pasos de Φ, es decir, se procede según:

$$\begin{aligned}y_h &= y + h\Phi(x,y;h), \\ y_{2h}^* &= y_h + h\Phi(x+h, y_h; h), \\ y_{2h} &= y + 2h\Phi(x,y;h).\end{aligned} \tag{I.24}$$

Entonces, de (I.23) resulta:

$$\frac{1}{2(2^p-1)}\frac{1}{h^{p+1}}(y_{2h}-y^*_{2h})=\tau(x,y)+O(h),$$

así, la expresión del lado izquierdo es un estimador aceptable $r(x,y;h)$. Si los dos pasos en (I.24) producen una precisión aceptable, entonces nuevamente para un cuarto proceso de Runge-Kutta de orden 4, el procedimiento requiere solo tres evaluaciones adicionales de f, ya que y_h y y_{2h} tendría que calcularse de todos modos. Mostramos, sin embargo, que todavía hay esquemas más eficientes.

2. Métodos integrados

La idea básica de este enfoque es muy simple: si el método dado Φ tiene orden p, tomando algún método de un paso Φ^* de orden $p+1$ y definimos:

$$r(x,y;h)=\frac{1}{h^p}[\Phi(x,y;h)-\Phi^*(x,y;h)]. \tag{I.25}$$

Esto es, en efecto, un estimador aceptable, como sigue restando las dos relaciones

$$\begin{aligned}\Phi(x,y;h)-\frac{1}{h}[u(x+h)-u(x)] &= \tau(x,y)h^p+O(h^{p+1}),\\ \Phi^*(x,y;h)-\frac{1}{h}[u(x+h)-u(x)] &= O(h^{p+1})\end{aligned}$$

y dividiendo el resultado por h^p.

La parte difícil es hacer que este procedimiento sea eficiente. Siguiendo una idea de Fehlberg, se puede intentar hacer esto incorporando un proceso de Runge-Kutta (de orden p) en otro (de orden $p+1$). Específicamente, sea Φ algún método explícito de Runge-Kutta

en r-etapas,

$$\begin{aligned} k_1(x,y) &= f(x,y), \\ k_s(x,y;h) &= \left(x+\mu_s h, y+h\sum_{j=1}^{s-1}\lambda_{sj}k_j \right), \quad s=2,3,\ldots,r, \\ \Phi(x,y;h) &= \sum_{s=1}^{r}\alpha_s k_s. \end{aligned}$$

Entonces, para Φ^* escogemos un proceso r^*-etapas similar, con $r^* > r$, de tal manera que:

$$\mu_s^* = \mu_s, \lambda_{sj}^* = \lambda_{sj}, \text{ para } s = 2,3,\ldots,r.$$

Luego, la estimación (I.25) cuesta solo $r^* - r$ evaluaciones extras de f. Si $r^* = r+1$, podríamos incluso intentar guardar la evaluación adicional seleccionando (si es posible)

$$\mu_{r^*} = 1, \lambda_{r^*j} = \alpha_j, \text{ para } j = 1,2,\ldots,r^*-1 (r^* = r+1).$$

Entonces, en efecto, k_{r^*} será idéntica con k_1 para el siguiente paso.

Fehlberg desarrolló pares de fórmulas $(p, p+1)$ de Runge-Kutta integradas de este tipo a fines de la década de 1960. Para más detalles, ver [12].

D. Control del paso

Cualquier estimación $r(x,y;h)$ de una función error principal $T(x,y)$ implica una estimación

$$h^p r(x,y;h) = \tau(x,y;h) + O(h^{p+1})$$

para el error de truncamiento, se puede utilizar para monitorear el error de truncamiento local durante el proceso de integración. Sin embargo, hay que tener en cuenta que el error de truncamiento local es bastante diferente del error global, el error que realmente se

quiere controlar. Para obtener más información sobre la relación entre estos dos errores, veamos el siguiente teorema, que cuantifica la continuidad de la solución de un problema de valor inicial con respecto a los valores iniciales.

Teorema 10. *Sea $f(x, y)$ continua en x para $a \leq x \leq b$ y satisfaciendo una condición de Lipschitz uniforme sobre $[a, b] \times \mathbb{R}^n$ con constante de Lipschitz L. Entonces el problema de valor inicial*

$$\begin{aligned} \frac{dy}{dx} &= f(x, y),\ a \leq x \leq b, \\ y(c) &= y_c, \end{aligned} \tag{I.26}$$

tiene una única solución sobre $[a, b]$ para cualquier c con $a \leq c \leq b$ y para cualquier $y_c \in \mathbb{R}^n$. Sea $y(x, s)$ e $y(x; s^*)$ las soluciones de la ecuación diferencial correspondiente a $y_c = s$ e $y_c = s^*$, respectivamente. Entonces, para cualquier vector norma $\|\cdot\|$,

$$\|y(x; s) - y(x; s^*)\| \leq e^{L|x-c|}\|s - s^*\|.$$

Resolver numéricamente el problema de valor inicial dado por (I.18) mediante un método de un paso (no necesariamente con un paso constante) en realidad significa que se tiene que seguir una secuencia de pistas de solución, en la que en cada punto de la malla x_n se salta de una pista a la siguiente. Luego, por una cantidad determinada por el error de truncamiento en x_n. Esto es así por la propia definición de error de truncamiento, siendo la solución de referencia una de las pistas de solución. Específicamente, la enésima pista, $n = 0, 1, \ldots, N$, es dado por la solución del valor inicial del problema

$$\begin{aligned} \frac{dv_n}{dx} &= f(x, v_n), x_n \leq x \leq b, \\ v_n(x_n) &= u_n, \end{aligned} \tag{I.27}$$

y

$$u_{n+1} = v_n(x_{n+1}) + h_n\tau(x_n, u_n; h_n),\ \ n = 0, 1, \ldots, N - 1.$$

Desde que (I.27) tenemos $u_{n+1} = v_{n+1}(x_{n+1})$, podemos aplicar el teorema 10 a la solución v_{n+1} y v_n, dejando $c = x_{n+1}$, $s = u_{n+1}$, $s^* = u_{n+1} - h_n\tau(x_n.u_n; h_n)$, y la solución obtenida

$$\|v_{n+1}(x) - v_n(x)\| \leq h_n e^{L|x_n - x_{n+1}|}\|\tau(x_n, u_n; h_n)\|, n = 0, 1, \ldots, N-1. \tag{I.28}$$

Ahora

$$\sum_{n=0}^{N-1}[v_{n+1}(x) - v_n(x)] = v_N(x) - v_0(x) = v_N(x) - y(x), \tag{I.29}$$

y desde que $v_N(x_N) = u_N$, dejando $x = x_N$, obtenemos de (I.28) y (I.29) que

$$\|u_N - y(x_N)\| \leq \sum_{n=0}^{N-1}\|v_{n+1}(x_N) - v_n(x)\| \leq \sum_{n=0}^{N-1} h_n e^{L|x_N - x_{n+1}|}\|\tau(x_n, u_n; h_n)\|.$$

Por lo tanto, si hacemos que:

$$\|\tau(x_n, u_n; h_n)\| \leq \epsilon_T,\ n = 0, 1, 2, \ldots, N-1, \tag{I.30}$$

entonces

$$\|u_n - y(x_N)\| \leq \epsilon_T \sum_{n=0}^{N-1}(x_{n+1} - x_n)e^{L|x_N - x_{n+1}|}.$$

Interpretando la suma del lado derecho como una suma de Riemann para una integral definida, finalmente obtenemos, aproximadamente,

$$\|u_N - y(x_N)\| \leq \epsilon_T \int_a^b e^{L(b-x)}\,dx = \frac{\epsilon_T}{L}(e^{L(b-a)} - 1).$$

Así, conocer una estimación de L permitirá establecer un ϵ_T apropiado. Normalmente,

$$\epsilon_T = \frac{L}{e^{L(b-a)} - 1}\epsilon, \tag{I.31}$$

garantiza un error $\|u_N - y(x_N)\| \leq \epsilon$. Lo que vale para toda la malla sobre $[a, b]$, por supuesto, se mantiene para cualquier malla de un subintervalo $[a, x]$, $a \leq x \leq b$. Así,

en principio, dada la precisión deseada ϵ para la solución $y(x)$, podemos determinar una tolerancia leve local ϵ_T (I.31) y lograr la precisión deseada manteniendo el error de truncamiento local por debajo de ϵ_T. Notemos que cuando $L \to 0$ tenemos $\epsilon_T \to \dfrac{\epsilon}{b-a}$. Este valor límite de ϵ_T sería apropiado para un problema de cuadratura, pero definitivamente no para un verdadero problema de ecuaciones diferenciales, donde ϵ_T, en general, debe elegirse considerablemente menor que la tolerancia del error objetivo ϵ.

Consideraciones como estas motivan el siguiente mecanismo de control de pasos, donde cada paso de integración (de x_n o $x_{n+1} = x_n + h_n$) consiste en lo siguiente:

1. Estimación de h_n.

2. Cálculo de $u_{n+1} = u_n + h_n\Phi(x_n, u_n; h_n)$ y $r(x_n, u_n; h_n)$.

3. Test $h_n^p\|r(x_n, u_n; h_n)\| \leq \epsilon_T$ y (I.30).

   Si pasa la prueba, continúe con el siguiente paso; sino, repita el paso con h_n más pequeño, hasta que pase la prueba.

Para estimar h_n, asumimos primero que $n \geq 1$, de manera que el estimador de los pasos previos, $r(x_{n-1}, u_{n-1}; h_{n-1})$, es factible. Entonces, despreciamos los términos de $O(h)$,

$$\|T(x_{n-1}, u_{n-1})\| \approx \|r(x_{n-1}, u_{n-1}; h_{n-1})\|,$$

y desde que $T(x_n, u_n) \approx \tau(x_{n-1}, u_{n-1})$. Asímismo:

$$\|\tau(x_n, u_n)\| \approx \|r(x_{n-1}; h_{n-1})\|.$$

Lo que queremos es:

$$\|T(x_n, u_n)\|h_n^p \approx \theta\epsilon_T,$$

donde θ es el factor de seguridad, es decir, $\theta = 0{,}8$. Eliminando $T(x_n, u_n)$, encontramos:

$$h_n \approx \left\{ \frac{\theta \epsilon_T}{\|r(x_{n-1}, u_{n-1}; h_{n-1})\|} \right\}^{1/p}.$$

Del paso anterior, tenemos $h_{n-1}^p \|r(x_{n-1}, u_{n-1}; h_{n-1})\| \leq \epsilon_T$, así que:

$$h_n \geq \theta^{1/p} h_{n-1},$$

y la tendencia es aumentar el paso.

Si $n = 0$, procedemos similarmente, usando una condición inicial $h_0^{(0)}$ de h_0 y asociando a $r(x_0, y_0; h_0^{(0)})$ obtenemos:

$$h_0^{(1)} = \left\{ \frac{\theta \epsilon_T}{\|r(x_0, y_0; h_0^{(0)})\|} \right\}^{1/p}.$$

El proceso puede repetirse una o dos veces para obtener la estimación final h_0 y $\|r(x_0, y_0; h_0)\|$.

E. Problemas de rigidez

Aunque no existe una definición aceptada de ecuaciones diferenciales rígidas, una característica de estos problemas de rigidez es el cambio "brusco" en la solución de la ecuación diferencial. Esto se manifiesta matemáticamente en la matriz jacobiana de f_y que tiene valores propios con partes reales negativas muy grandes junto con otras de magnitud normal. Los métodos numéricos estándar no pueden hacer frente a tales soluciones a menos que utilicen longitudes de paso largos. Lo que se necesita son métodos que disfruten de una propiedad de estabilidad especial denominada A-estabilidad.

Introducimos este concepto en el contexto de sistemas homogéneos lineales de ecuaciones diferenciales con matriz de coeficientes constantes, esto es

$$\frac{dy}{dx} = Ay, \;\; 0 \leq x < +\infty, \;\; y(a) = y_0, \tag{I.32}$$

donde $A \in \mathbb{C}^{n\times n}$ es una matriz constante de orden n.

Los problemas de valor inicial para los que es probable que esto se presente reciben el nombre de ecuaciones rígidas y son bastante comunes, de modo especial en el estudio de vibraciones, reacciones químicas y circuitos eléctricos.

A continuación presentaremos dos ejemplos de problemas rígidos para el caso lineal dado en [3] y no lineal, dado en [7].

Ejemplo 1. *El problema de valor inicial:*

$$\begin{bmatrix} u_1' \\ u_2' \end{bmatrix} = \begin{bmatrix} 9 & 24 \\ -24 & -51 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} + \begin{bmatrix} 5\cos(t) - \frac{1}{3}\operatorname{sen}(t) \\ -9\cos(t) + \frac{1}{3}\operatorname{sen}(t) \end{bmatrix}$$

con condiciones iniciales $u_1(0) = \frac{4}{3}, u_2(0) = \frac{2}{3}$*, el cual tiene solución única:*

$$u(t) = \begin{bmatrix} 2e^{-3t} - e^{-39t} + \frac{1}{3}\cos(t) \\ -e^{-3t} + 2e^{-39t} - \frac{1}{3}\cos(t) \end{bmatrix}$$

Se puede observar que el término e^{-39t} *en la solución causa que este problema sea rígido. Observemos los resultados para* $h = 0{,}1$ *en la tabla I que son catastróficos.*

Tabla I: Resultados numéricos del método Runge Kutta, con $h = 0{,}05$ y $h = 0{,}1$.

t	$u_1(t)$	$y_1(t)$ $h = 0{,}05$	$y_1(t)$ $h = 0{,}1$	$u_2(t)$	$y_2(t)$ $h = 0{,}05$	$y_2(t)$ $h = 0{,}1$
0.1	1.793061	1.7112219	-2.645169	-1.032001	-0.8703152	7.844527
0.2	1.423901	1.414070	-18.45158	-0.8746809	-0.8550148	38.87631
0.3	1.131575	1.130523	-87.47221	-0.7249984	-0.7228910	176.4828
0.4	0.9094086	0.9092763	-934.0722	-0.6082141	-0.6079475	789.3540
0.5	0.7387877	0.7387506	-1760.016	-0.5156575	-0.5155810	3520.00
0.6	0.6057094	0.6056833	-7848.550	-0.4404108	-0.4403558	15697.84
0.7	0.4998603	0.4998361	-34989.63	-0.3774038	-0.3773540	69979.87
0.8	0.4136714	0.4136490	-155979.4	-0.3229535	-0.3229078	311959.5
0.9	0.3416143	0.3415939	-695332.0	-0.2744088	-0.2743673	1390664
1.0	0.2796748	0.2796568	-3099671.	-0.2298877	-0.2298511	6199352

Nota: Tomado del Burden [3].

Ejemplo 2. *Consideremos el problema rígido de valor inicial dado por:*

$$x_1' = -0{,}013x_1 - 1000x_1x_3, \qquad x_1(0) = 1,$$

$$x_2' = -2500x_2x_3, \qquad x_2(0) = 1,$$

$$x_3' = -0{,}013x_1 - 1000x_1x_3 - 2500x_2x_3, \qquad x_3(0) = 0,$$

donde el jacobiano de f_x tiene valores propios reales, como por ejemplo para $t = 0$ sus valores propios son $0, -0{,}0093$ y -3500.

Observemos los resultados para $h = 0{,}0001$ en la tabla II que son como sigue:

Tabla II: Resultados numéricos del método Runge Kutta, con $h = 0{,}0001$ y $t = 1$.

	Solución exacta	Solución numérica
x_1	$0{,}99073192$	$1{,}47412957$
x_2	$1{,}00926441$	$0{,}52845463$
x_3	$-0{,}00000367$	$0{,}00258421$

Estos ejemplos, nos llevan a definir la caracterización de una ecuación rígida de la siguiente manera.

Definición 16. *Sea la ecuación diferencial dada por I.32, esta ecuación diferencial se caracteriza por ser un problema rígido cuando los valores propios λ_j de A son tales que, si*

$$\max_{1\leq j\leq n} |Re(\lambda_j)| \gg \min_{1\leq i\leq n} |Re(\lambda_i)|$$

para al menos un par de valores propios de módulo grande.

Definición 17. *La ecuación diferencial dada por I.32 se caracteriza por ser un problema altamente oscilatorio cuando los valores propios λ_j de A son puramente imaginarios, $\lambda_j = \mu_j + \nu_j i$, y además $\mu_j < 0$ para todos los j,*

$$\max_{1\leq j\leq n} |\mu_j| \gg \min_{1\leq j\leq n} |\mu_j| \text{ y } |\mu_j| \ll |\nu_j|$$

para al menos un par de valores propios de módulo grande.

A continuación veamos el estudio de la A-estabilidad para ecuaciones diferenciales lineales y algunos métodos A-estables.

F. A-Estabilidad

Supongamos que la matriz A tiene valores propios en el semiplano izquierdo, es decir,

$$\text{Re}(\lambda_i(A)) < 0,\ i = 1, 2, \ldots, n \tag{I.33}$$

donde λ_i, para $i = 1, 2, \ldots, n$, son los valores propios de la matriz de A. Desde que la solución de la ecuación diferencial (I.32) es $y(x) = y_0 e^{A(x-a)}$ estas decaen cuando $x \to +\infty$. Los correspondientes valores propios con partes real negativas muy grandes lo hace particularmente muy rápido, dando lugar al fenómeno de la rigidez.

En particular, para la solución $y(x)$ de (I.32), tenemos

$$y(x) \to 0, \text{ cuando } x \to +\infty. \tag{I.34}$$

Veamos ahora el comportamiento de los métodos de un paso:

$$y(x + h) = y(x) + h\Phi(x, y; h) = \psi(hA)y(x), \tag{I.35}$$

donde ψ es alguna función, llamada la función de estabilidad del método. En lo que sigue asumiremos que la función matriz $\psi(hA)$ está bien definida; mínimamente, requerimos que $\psi : \mathbb{C} \to \mathbb{C}$ sea analítica en una vecindad del origen. Desde que la solución de referencia que pasa por el punto (x_0, y_0) es dado por $u(x) = e^{A(x-x_0)}y_0$, tenemos para el error de truncamiento de Φ en (x_0, y_0)

$$\tau(x_0, y_0; h) = \frac{1}{h}[y(x_0 + h) - u(x_0 + h)] = \frac{1}{h}[\psi(hA) - e^{hA}]y_0.$$

En particular, el método Φ en este caso tiene orden p si y solo si

$$e^z = \psi(z) + O(z^{p+1}),\ \ z \to 0.$$

Esto muestra la relevancia de las aproximaciones a la función exponencial en el contexto de los métodos de un paso aplicados al problema del modelo (I.32).

Desde que la solución aproximada $u = \{u_n\}$ del problema de valor inicial (I.32), y suponiendo por simplicidad una longitud de malla constante h, es dado por

$$u_{n+1} = \psi(hA)u_n, \;\; n = 0, 1, 2, \ldots, \;\; u_0 = y_0,$$

por lo tanto:

$$u_n = (\psi(hA))^n y_0, \;\; n = 0, 1, 2, \ldots, . \tag{I.36}$$

Esto simulará el comportamiento (I.34) de la solución exacta si y solo si

$$\lim_{n\to\infty} (\psi(hA))^n = 0. \tag{I.37}$$

Una condición necesaria y suficiente para que se cumpla (I.37) es que los valores propios de la matriz $\psi(hA)$ estén estrictamente dentro del círculo unitario. Esto a su vez es equivalente a:

$$|\psi(h\lambda_i(A))| < 1, \quad \text{para } i = 1, 2, \ldots, n,$$

donde $\lambda_i(A)$ son los valores propios de A. En vista de (I.33), esto da lugar a la siguiente definición.

Definición 18. *La región $\mathcal{R}$ de estabilidad absoluta para un método de un paso Φ según (I.35) es*

$$\mathcal{R} = \{h\lambda \in \mathbb{C} \mid |\psi(h\lambda)| < 1\}$$

Definición 19. *Un método de paso Φ es llamado A-estable si la función estabilizadora ψ asociada con Φ según (I.35) es definida en el semiplano complejo izquierdo y satisface*

$$|\psi(z)| < 1, \text{ para todo } z \text{ con } Re(z) < 0. \tag{I.38}$$

Nos vemos conducidos al problema de construir una función ψ, el cual sea analítica en el semiplano izquierdo, se aproxime bien a la función exponencial cercano al origen, y satisfaga (I.38). Una herramienta importante para la determinación de una función estabilizadora es la aproximación de Padé a la función exponencial.

A continuación, veamos el estudio de las aproximaciones de Padé a la función exponencial que es fundamental en la construcción de métodos de un paso A-estables.

G. Aproximación de Padé

Para toda función analítica $g(z)$ en una vecindad de z, se define estas aproximaciones de Padé como sigue.

Definición 20. *La aproximación de Padé $R[n,m](z)$ de la función $g(z)$ es la función racional*

$$R[m,n](z) = \frac{P(z)}{Q(z)}, \;\; P \in \mathbb{P}_m, \; Q \in \mathbb{P}_n, \tag{I.39}$$

satisfaciendo

$$g(z)Q(z) - P(z) = O(z^{m+n+1}) \text{ cuando } z \to 0. \tag{I.40}$$

donde $\mathbb{P}_m$ y $\mathbb{P}_n$ son los conjuntos de polinomios de grado m y n respectivamente. Es conocido que la función racional $R[m,n]$ es únicamente determinado por esta definición, aunque en casos excepcionales P y Q pueden tener factores comunes. Si este no es el caso, es decir, P y Q son irreducibles sobre los números complejos, asumimos sin pérdida de generalidad que $Q(0) = 1$.

Nuestro interés aquí es la función $g(z) = e^z$. En este caso, $P = P[m,n]$ y $Q = Q[m,n]$ en (I.39) y (I.40) pueden ser determinado.

Teorema 11. *La aproximación de Padé $R[m,n]$ de la función exponencial $g(z) = e^z$ es dado por*

$$P[m,n](z) = \sum_{k=0}^{m} \frac{m!(n+m-k)!z^k}{(m-k)!(n+m)!k!}, \tag{I.41}$$

$$Q[m,n](z) = \sum_{k=0}^{n} (-1)^k \frac{n!(n+m-k)!z^k}{(n-k)!(n+m)!k!}. \tag{I.42}$$

Además,

$$e^z - \frac{P[m,n](z)}{Q[m,n](z)} = C_{n,m} z^{n+m+1} + \cdots,$$

donde

$$C_{n,m} = (-1)^n \frac{n!m!}{(n+m)!(n+m+1)!}.$$

Demostración. Ver [3] y [12]. □

La aproximación de Padé de la función exponencial tiene algunas propiedades muy útiles e importantes. Enunciaremos los de interés en relación con la A-estabilidad.

1. $P[m,n](z) = Q[n,m](-z)$, el numerador polinomial es el denominador polinomial con índices intercambiados y z reemplazada por $-z$. Esto refleja la propiedad $1/e^z = e^{-z}$ de la función exponencial. La prueba sigue inmediatamente de (I.41) y (I.42).

2. Para cada $n = 0, 1, 2, \ldots$ todos los ceros de $Q[n,n]$ tienen parte real positiva, ver [12].

3. Para todo $t \in \mathbb{R}$, y $n = 0, 1, 2, \ldots$ se mantiene

$$\left| \frac{P[n,n](it)}{Q[n,n](it)} \right| = 1.$$

En efecto, por la propiedad 1, se tiene $P[n,n](it) = \overline{Q[n,n](it)}$.

4. Se mantiene

$$\left|\frac{P[n,n+1](it)}{Q[n,n+1](it)}\right| < 1, \text{ para } t \in \mathbb{R}, t \neq 0, \ n = 0, 1, 2, \ldots.$$

5. Para cada $n = 0, 1, 2, \ldots$ todos los ceros en $Q[n+1, n]$ tienen parte real positiva.

6. Una función racional R satisface $|R(z)| < 1$ para Re $(z) < 0$ si y solamente si R es analítica en Re $(z) < 0$ y $|R(z)| \leq 1$ para Re $(z) = 0$.

   Si $|R(z)| < 1$ en Re$(z) < 0$, no puede haber un polo en Re $(z) \leq 0$ o en $z = \infty$.

   Por continuidad, por lo tanto, $|R(z)| \leq 1$ sobre Re $(z) = 0$.

   Luego, R tiene que ser analítica en Re $(z) \leq 0$ y en $z = \infty$. Claramente, $\lim_{z\to\infty} |R(z)| \leq 1$ y $|R(z)| \leq 1$ sobre el eje imaginario, entonces, por el principio de maximidad, se cumple $|R(z)| < 1$ para Re $(z) < 0$.

7. Como un corolario de la propiedad 6, enunciamos las siguientes propiedades.

   Para cada $n = 0, 1, 2, \ldots$, se tiene

$$\left|\frac{P[n,n](z)}{Q[n,n](z)}\right| < 1, \qquad \text{para Re}(z) < 0 \tag{I.43}$$

$$\left|\frac{P[n+1,n](z)}{Q[n+1,n](z)}\right| < 1, \quad \text{para Re}(z) \leq 0, z \neq 0. \tag{I.44}$$

Teorema 12. *Si la función ψ asociada con el método de un solo paso Φ según (I.35) es la aproximación de Padé $\psi(z) = R[n,n](z)$ de e^z, o la aproximación de Padé $\psi(z) = R[n+1,n](z)$ de e^z, $n = 0, 1, 2, \ldots$, entonces el método Φ es A-estable.*

Ejemplo 3.

1. ***Método de Euler implícito**: Este es un método de solo paso definido por*

$$u_{n+1} = u_n + hf(x_{n+1}, u_{n+1}).$$

Este requiere, en cada paso, la solución de un sistema de ecuaciones no lineales para $u_{n+1} \in \mathbb{R}^n$. En el caso del modelo (I.26), se convierte a $u_{n+1} = u_n + hAu_{n+1}$ y puede ser resuelto explícitamente como $u_{n+1} = (I - hA)^{-1}u_n$. Así, la función asociada ψ aquí es

$$\psi(z) = \frac{1}{1-z} = 1 + z + z^2 + \cdots,$$

la aproximación de Padé $R[1,0](z)$ de la función exponencial e^z. Desde que $\psi(z) - e^z = O(z^2)$ cuando $z \to 0$, el método tiene orden $p = 1$ y por el teorema 12 es A-estable.

2. ***Regla trapezoidal**: Aquí, tenemos que:*

$$u_{n+1} = u_n + \frac{1}{2}h[f(x_n, u_n) + f(x_{n+1}, u_{n+1})],$$

de nuevo una ecuación no lineal en u_{n+1}. Para el problema (I.32), este se convierte en:

$$u_{n+1} = \left(I + \frac{1}{2}hA\right)u_n + \frac{1}{2}hAu_{n+1},$$

por lo tanto $u_{n+1} = \left(I - \frac{1}{2}hA\right)^{-1}\left(I + \frac{1}{2}hA\right)u_n$, *y*

$$\psi(z) = \frac{1 + \frac{1}{2}z}{1 - \frac{1}{2}z} = 1 + z + \frac{1}{2}z^2 + \frac{1}{4}z^3 + \cdots.$$

Este es la aproximación de Padé $\psi(z) = R[1,1](z)$ de la función exponencial e^z y $\psi(z) - e^z = O(z^3)$. De esta manera, el método es A-estable, pero de orden $p = 2$.

3. ***Fórmula implícita de Runge-Kutta**: Como se menciona en la sección (4), donde se describe el método de Runge-Kutta, tiene una etapa r implícita*

$$\begin{aligned} \Phi(x, y; h) &= \sum_{s=1}^{r} \alpha_s k_s(x, y; h) \\ k_s &= f\left(x + \mu_s h, y + h\sum_{j=1}^{r} \lambda_{sj} k_j\right),\ s = 1, 2, \cdots, r \end{aligned} \tag{I.45}$$

es posible demostrar que (I.45) es un método de orden $p, r \leq p \leq 2r$, si $f \in C^p$ en $[a, b] \times \mathbb{R}^n$ y

$$\sum_{j=1}^{r} \lambda_{sj} \mu_j^k = \frac{\mu_s^{k+1}}{k+1}, \quad k = 0, 1, \cdots, r-1;\ s = 1, 2, \cdots, r, \tag{I.46}$$

$$\sum_{s=1}^{r} \alpha_s \mu_s^k = \frac{1}{k+1}, \quad k = 0, 1, \cdots, p-1 \tag{I.47}$$

Para cualquier conjunto de μ_j distintos, y para cada $s = 1, 2, \cdots, r$, las ecuaciones (I.46) representan un sistema de ecuaciones lineales para $\{\lambda_{sj}\}_{j=1}^{r}$, cuya matriz de coeficientes es una matriz de Vandermonde no singular. Por lo tanto, se puede resolver de forma única para $\{\lambda_{sj}\}$. Ambas condiciones (I.46) y (I.47) pueden verse más naturalmente en términos de fórmulas en cuadratura. De hecho, (I.46) es equivalente a

$$\int_0^{\mu_s} p(t)dt = \sum_{j=1}^{r} \lambda_{sj} p(\mu_j), \quad \textit{para todo } p \in \mathbb{P}_{r-1}$$

mientras que (I.47) equivale a:

$$\int_0^1 q(t)dt = \sum_{s=1}^{r} \alpha_s q(\mu_s), \quad \textit{para todo } q \in \mathbb{P}_{p-1}$$

Sabemos que en la ecuación anterior podemos tener $r \leq p \leq 2r$. Luego, se puede obtener una única fórmula de Runge-Kutta en r etapas de orden $p = 2r$. A continuación, se demuestra que los métodos de Runge-Kutta de orden $2r$ son A-estable. En lugar del sistema (I.32), también podemos considerar una ecuación escalar

$$\frac{dy}{dx} = \lambda y, \tag{I.48}$$

a la que (I.32) puede reducirse por descomposición espectral. Aplicando a (I.48), el k_s correspondiente a (I.45) debe satisfacer el sistema lineal

$$k_s = \lambda \left(y + h \sum_{j=1}^{r} \lambda_{sj} k_j \right);$$

esto es (con $z = \lambda h$),

$$k_s - z\sum_{j=1}^{r} \lambda_{sj} k_j = \lambda y, \quad s = 1, 2, \ldots, r$$

Sea

$$d_r(z) = \begin{vmatrix} 1 - z\lambda_{11} & -z\lambda_{12} & \cdots & -z\lambda_{1r} \\ -z\lambda_{21} & 1 - z\lambda_{22} & \cdots & -z\lambda_{2r} \\ \cdots\cdots & \cdots\cdots & \cdots & \cdots\cdots \\ -z\lambda_{r1} & -z\lambda_{r2} & \cdots & 1 - z\lambda_{rr} \end{vmatrix},$$

$$d_{r,s}(z) = \begin{vmatrix} 1 - z\lambda_{11} & \cdots & 1 & \cdots & -z\lambda_{1r} \\ -z\lambda_{21} & \cdots & 1 & \cdots & -z\lambda_{2r} \\ \cdots\cdots & \cdots & \cdot & \cdots & \cdots\cdots \\ -z\lambda_{r1} & \cdots & 1 & \cdots & 1 - z\lambda_{rr} \end{vmatrix}, \; s = 1, 2, \cdots r,$$

donde la columna de unos es la s-ésima columna del determinante. Claramente, d_r y $d_{r,s}$ son polinomios de grado r y $r - 1$, respectivamente. Por la regla de Cramer,

$$k_s = \frac{d_{r,s}(z)}{d_r(z)} \lambda y, \quad s = 1, 2, \cdots, r$$

Así, la función φ asociada al método Φ correspondiente a (I.45) es

$$\varphi(z) = \frac{d_r(z) + z\sum_{s=1}^{r} \alpha_s d_{r,s}(z)}{d_r(z)} \tag{I.49}$$

Vemos que φ es una función racional del tipo $R[r, r]$ y, el método ψ tiene orden $p = 2r$, tenemos

$$e^z = \varphi(z) + O(z^{2r+1}), \quad z \to 0$$

De ello, se deduce que φ en (I.49) es la aproximación de Padé $R[r, r]$ a la función exponencial e^z, y por lo tanto, por el teorema 12, es A-estable.

4. ***Método Ehle**: Ese es un método que involucra derivadas totales de f*

$$\begin{aligned} \Phi(x,y;h) &= k(x,y;h) \\ k &= \sum_{s=1}^{r} h^{s-1}[\alpha_s f^{[s-1]}(x,y) - \beta_s f^{[s-1]}(x+h, y+hk)] \end{aligned} \tag{I.50}$$

También es implícito, ya que requiere la solución de la segunda ecuación en (I.50) para el vector $k \in \mathbb{R}^d$. Un pequeño cálculo mostrará que la función φ asociado con Φ en (I.50) viene dado por

$$\varphi(z) = \frac{1 + \sum_{s=1}^{r} \alpha_s z^s}{1 + \sum_{s=1}^{r} \beta_s z^s}$$

Al elegir este φ para ser una aproximación de Padé a la función exponencial e^z, ya sea $R[r,r]$ o $R[r-1,r]$ (haciendo que $\alpha_r = 0$), nuevamente obtenemos dos métodos A-estables. Este último tiene la propiedad adicional de ser fuertemente A-estable (o L-estable), en el sentido de que

$$\varphi(z) \to 0 \quad \textit{como} \quad Re\,(z) \to -\infty$$

Esto significa, en vista de (I.36), que la convergencia $u_n \to 0$ es $n \to \infty$, es decir, es más rápido para componentes correspondientes a valores propios más a la izquierda en el plano complejo.

H. Regiones de absoluta estabilidad

Para métodos Φ que no son A-estables, es importante saber la región de absoluta estabilidad.

$$\mathcal{D}_A = \{z \in \mathbb{C} : |\varphi(z)| < 1\}$$

Si el método Φ aplicado al problema del modelo (I.32) se aproxima a una solución $u = \{u_n\}$ con $\lim_{n\to\infty} u_n = 0$, es necesario que $h\lambda_i(A) \in \mathcal{D}_A$ para todos los valores propios $\lambda_i(A)$ de A. Si algunos de estos tienen partes reales negativas muy grandes, entonces esta condición impone una restricción severa en la longitud del paso h, a menos que $\mathcal{D}_A$ contenga una gran parte del plano izquierdo. Desafortunadamente, para muchos métodos clásicos, este no es el caso para el método de Euler, por ejemplo tenemos $\varphi(z) = 1 + z$. Por lo tanto:

$$\mathcal{D}_A = \{z \in \mathbb{C} : |1 + z| < 1\},$$

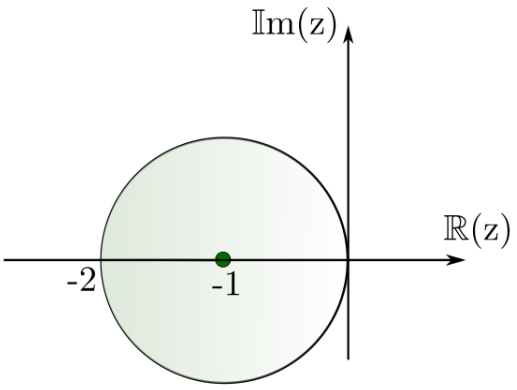

Figura III: Regiones de estabilidad absoluta para el método de Euler.

y la región de estabilidad absoluta es la unidad del disco en $\mathbb{C}$, centrada en -1, como se muestra en la figura III. De manera más general, por la expansión de Taylor de orden $p \geq 1$ y también para todo p estado explícito del método de Runge-Kutta de orden p, con $1 \leq p \leq 4$, se tiene:

$$\varphi(z) = 1 + \frac{1}{1!}z + \frac{1}{2!}z^2 + \cdots + \frac{1}{p!}z^p \tag{I.51}$$

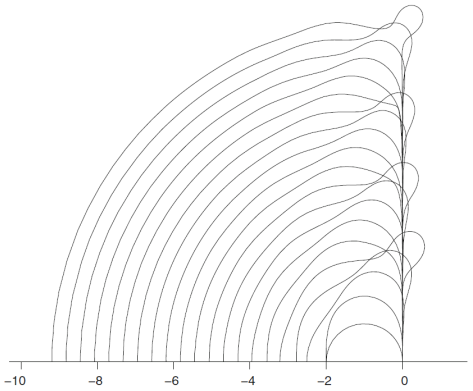

Figura IV: Regiones de estabilidad absoluta para métodos de p-orden con ψ como en

I.51 $p = 1, 2, 3, \ldots, 21$.

Adaptado de [12].

Para calcular la línea de contorno $|\varphi(z)| = 1$, que delimita la región $\mathcal{D}_A$, se puede encontrar una ecuación diferencial para esta línea y usar un método de un paso para resolverla. Es decir, podemos usar un método ψ para analizar su propia región de estabilidad: un caso de autoanálisis, por así decirlo. Los resultados para Φ en (I.51) y $p = 1, 2, \ldots, 21$ son representados en la figura IV. Debido a la simetría, solo las regiones en la parte superior se muestran semiplanos.

En el caso del método trapezoidal, tenemos que su función de estabilidad es $\psi(z) = \dfrac{2+z}{2-z}$. Por lo tanto, su región de estabilidad absoluta son semiplanos izquierdos complejos, como se muestra en la figura V, con lo que tendremos que es un método A-estable.

$$\mathcal{D}_A = \left\{ z \in \mathbb{C} : \left| \frac{2+z}{2-z} \right| < 1 \right\},$$

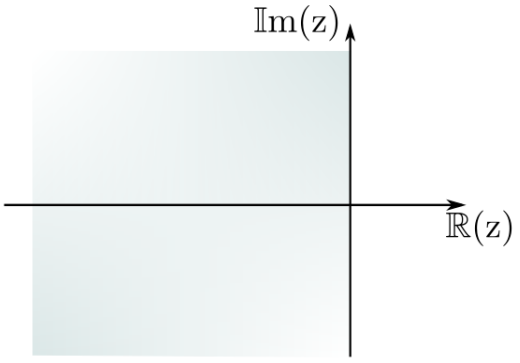

Figura V: Regiones de estabilidad absoluta para el método trapezoidal.

III. Métodos numéricos para calcular la raíz cuadrada de una matriz

El cálculo de la raíz cuadrada de una matriz es un problema fundamental en el ámbito del álgebra lineal y tiene importantes aplicaciones en diversas áreas, como la teoría de control, la estadística y la física. Este capítulo se centra en los métodos numéricos desarrollados para abordar este desafío, explorando tanto su evolución histórica como su aplicación práctica en situaciones modernas.

Comenzaremos con una revisión de la evolución de los métodos para calcular la raíz cuadrada de una matriz, proporcionando un contexto que resalte el progreso en esta área. A continuación, se discutirá la naturaleza de la raíz cuadrada de una matriz, analizando sus propiedades y el significado que tiene en el contexto de operaciones matriciales.

El núcleo del capítulo se dedicará a los métodos numéricos específicos para determinar la raíz cuadrada de una matriz. Se abordará el método de Newton, uno de los enfoques más utilizados, junto con una discusión sobre su convergencia y estabilidad, aspectos críticos para garantizar resultados precisos y fiables. Además, se presentarán otros métodos alternativos, proporcionando una comparación que permita evaluar sus respectivas ventajas y desventajas.

La sección sobre teoremas de convergencia ofrecerá un análisis riguroso de las condiciones bajo las cuales los métodos propuestos convergen, asegurando que el lector comprenda los fundamentos teóricos que sustentan estos procedimientos. También se examinará la estabilidad de los algoritmos, un factor crucial en el tratamiento de problemas numéricos, especialmente en el caso de matrices que pueden presentar características desafiantes.
Por último, se introducirá un álgebra aproximada de la matriz exponencial para problemas rígidos, destacando su relevancia en la computación eficiente y en la resolución de sistemas de ecuaciones diferenciales. Este capítulo proporcionará las herramientas necesarias para entender y aplicar los métodos numéricos en el cálculo de la raíz cuadrada de una matriz, sentando las bases para futuras investigaciones y aplicaciones en este campo.

A. Evolución de los métodos

1. Métodos clásicos: Inicialmente, los métodos para calcular la raíz cuadrada de una matriz se basaban en técnicas analíticas y algebraicas, como la diagonalización y la factorización de matrices. Aunque estos métodos eran efectivos, también eran limitados a matrices específicas, como matrices simétricas o definidas positivas.

    a) Diagonalización: Para matrices simétricas, la diagonalización permite expresar una matriz como el producto de una matriz de vectores propios y una matriz diagonal de valores propios. La raíz cuadrada se obtiene fácilmente al tomar la raíz cuadrada de cada valor propio.

    b) Factorización de Cholesky: Utilizada para matrices definidas positivas, esta técnica factoriza una matriz como el producto de una matriz triangular inferior

y su transpuesta, facilitando el cálculo de la raíz cuadrada.

2. Métodos iterativos:

   a) Método de Newton-Schulz: Uno de los métodos más influyentes es el de Newton, adaptado para matrices. Este método iterativo, también conocido como el método de iteración de Newton-Schulz, es popular debido a su simplicidad y eficacia en la convergencia cuadrática bajo ciertas condiciones. Sin embargo, su implementación puede ser numéricamente inestable si no se maneja adecuadamente.

   b) Métodos basados en la descomposición QR: Son algoritmos que utilizan la descomposición QR para resolver sistemas lineales iterativamente y calcular raíces cuadradas.

3. Algoritmos de Denman-Beavers: Este algoritmo es otro enfoque iterativo que ha ganado reconocimiento. Este método alterna entre dos matrices en cada iteración, el cual converge hacia la raíz cuadrada de la matriz original. Es conocido por su robustez y eficiencia en la convergencia.

4. Métodos basados en descomposiciones:

   a) Descomposición de Schur: Involucra descomponer una matriz en su forma triangular superior, lo que simplifica el cálculo de la raíz cuadrada.

   b) Descomposición espectral: Similar a la diagonalización, esta técnica se utiliza para matrices normales y aprovecha las propiedades espectrales de la matriz.

5. Enfoques numéricos recientes:

a) Aproximación de Padé: Utiliza fracciones continuas para aproximar la función de la raíz cuadrada, mejorando la estabilidad numérica.

b) Series de Taylor: Expande la raíz cuadrada en una serie de Taylor, permitiendo cálculos precisos mediante truncamiento adecuado.

B. Acerca de la raíz cuadrada de una matriz

En esta sección, nos basamos en los trabajos de [22] para fijar algunos conceptos previos que estudian la raíz cuadrada de una matriz.

Se llama raíz cuadrada de una matriz A compleja de orden n, a cualquier matriz X compleja $n \times n$ tal que

$$X^2 = A \tag{II.1}$$

Sobre esta ecuación $(II,1)$ trabajaron matemáticos como Caley, Sylvester y Frobenius a finales del siglo XIX, y algunos como H.F. Baker, L.E. Dickson, W.E. Roth a comienzos del siglo XX.

No toda matriz posee raíz cuadrada, como por ejemplo:

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

es fácil verificar que no existe alguna matriz compleja X de orden 2×2, tal que $X^2 = A$. Sin embargo, puede existir la raíz cuadrada de una matriz y no ser única, como por ejemplo la siguiente matriz:

$$A = \begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix}$$

que tiene como raíz cuadrara a:

$$\begin{bmatrix} \sqrt{2} & 0 \\ 0 & \sqrt{3} \end{bmatrix}, \begin{bmatrix} \sqrt{2} & 0 \\ 0 & -\sqrt{3} \end{bmatrix}, \begin{bmatrix} -\sqrt{2} & 0 \\ 0 & \sqrt{3} \end{bmatrix}, \begin{bmatrix} -\sqrt{2} & 0 \\ 0 & -\sqrt{3} \end{bmatrix}$$

que es un caso finito, o en el caso infinito donde:

$$\begin{bmatrix} a & b \\ b & -a \end{bmatrix}$$

con $a, b \in \mathbb{R}$, tales que $a^2 + b^2 = 1$, son raíces cuadradas de la matriz identidad

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

Teorema 13. *Si $A \in \mathbb{C}^{n\times n}$ es una matriz no singular con la descomposición de Jordan*

$$A = PJP^{-1} = Pdiag(J_{\lambda_1}, J_{\lambda_2}, \cdots, J_{\lambda_r})P^{-1} \tag{II.2}$$

donde $J_{\lambda_r} \in \mathbb{C}^{n\times n}$, r es el número de bloques de Jordan, $\sum_{k=1}^{r} n_i = n$ que es el número de valores propios y s es el número de valores propios diferentes de A. Entonces, si $(s \leq r)$ se verifica lo siguiente:

que la matriz A tiene precisamente 2^s raíces cuadradas que son funciones de A en el sentido 3.

Aunque es frecuente tratar de buscar condiciones para la existencia y unicidad de la raíz cuadrada de una matriz, un resultado frecuente tiene que ver con las matrices *semidefinidas positivas* (s.d.p.). Tal resultado dice que "una matriz A es s.d.p. si y solo si A tiene una raíz cuadrada s.d.p."

Recordemos que se dice que una matriz A compleja hermitiana de orden $n \times n$ es s.d.p. si $\forall X \in \mathbb{C}^n, X^*AX \geq 0$, donde X^* es el conjugado del vector transpuesto X^T de X.

A continuación se presentarán algunos resultados interesantes sobre la raíz cuadrada de una matriz.

Teorema 14. *Toda matriz diagonalizable tiene raíz cuadrada.*

Demostración. Sea A una matriz $n \times n$ sobre $\mathbb{C}$, diagonalizable, y sea P una matriz invertible tal que la matriz $D = P^{-1}AP$ es diagonal. Si X es una raíz cuadrada de D, entonces PXP^{-1} es una raíz cuadrada de A ya que:

$$(PXP^{-1})^2 = PX^2P^{-1} = PDP^{-1} = A$$

□

Es más, toda raíz cuadrada de A es de la forma PXP^{-1} para alguna raíz cuadrada X de la matriz D.

Teorema 15. *Toda matriz hermitiana s.d.p. tiene una raíz cuadrada hermitiana s.d.p.*

Demostración. Sea una matriz A hermitiana de orden $n \times n$ s.d.p. Entonces existe una matriz U unitaria $(U^* = U^{-1})$, tal que:

$$U^*AU = \underbrace{\begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{bmatrix}}_{D},$$

donde $\lambda_1, \cdots, \lambda_n$ son los valores propios de A, los cuales son todos reales; y por ser A s.d.p., tales valores propios son no negativos.

La matriz

$$D^{\frac{1}{2}} = \begin{bmatrix} \sqrt{\lambda_1} & & \\ & \ddots & \\ & & \sqrt{\lambda_n} \end{bmatrix}$$

es una raíz cuadrada s.d.p. de D, y $UD^{\frac{1}{2}}U^*$ (que es s.d.p. como lo es $D^{\frac{1}{2}}$) es una raíz cuadrada de A. □

Se puede probar que $UD^{\frac{1}{2}}U^*$ es la única raíz cuadrada s.d.p. de A.

Corolario 1. *Toda matriz hermitiana definida positiva tiene una única raíz cuadrada hermitiana definida positiva.*

Lema 4. *(Ver [23]) Si $f(x)$ es un polinomio de grado $n > 0$, con coeficientes en $\mathbb{C}$, cuyo término constante no es cero, existe un polinomio $g(x)$ de grado menor que n tal que $(g(x))^2 - x$ es divisible por $f(x)$.*

Teorema 16. *(Ver[23]) Si una matriz compleja A de orden $n \times n$ es invertible, existen matrices complejas B de orden $n \times n$ (necesariamente invertibles) tales que $B^2 = A$, siendo B un polinomio en A, de grado menor que n.*

Demostración. Sea A una matriz invertible de orden $n \times n$. Como A es invertible, su polinomio característico $f(x)$, que es de grado n, tiene término constante no cero; por tanto (según el lema anterior) existe un polinomio $g(x)$, de grado menor que n, tal que

$$(g(x))^2 - x = f(x)h(x)$$

para algún polinomio $h(x)$. Sustituyendo x por A tendremos

$$(g(A))^2 - A = \underbrace{f(A)}_{O} h(A) = O$$

de donde $(g(A))^2 = A$ y así la matriz $B = g(A)$ es una raíz cuadrada de A. □

Ejemplo 4. *Consideremos la matriz*

$$A = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 2 \end{bmatrix}$$

El polinomio característico de A *es* $f(\lambda) = (\lambda - 1)^2(\lambda - 2)$. *Luego, obtenemos por el lema (4)*

$$g(\lambda) = k(\lambda)(\lambda - 1)^2 + h(\lambda)(\lambda - 2) \tag{II.3}$$

donde $k(\lambda) = a_0$ *y* $h(\lambda) = b_0 + b_1(\lambda - 1)$ *son los polinomios a determinar de modo que* $(g(\lambda))^2 - \lambda$ *resulte divisible por* $f(\lambda)$. *Un* $g(\lambda)$ *de la forma II.3 es:*

$$\begin{aligned} g(\lambda) &= \sqrt{2}(\lambda - 1)^2 + \left(1 + \frac{3}{2}(\lambda - 1)\right)(\lambda - 2) \\ &= \sqrt{2}(\lambda - 1)^2 + (\lambda - 2) + \frac{3}{2}(\lambda - 1)(\lambda - 2) \end{aligned}$$

Luego, la matriz:

$$B = g(A) = \begin{bmatrix} -1 & 0 & 0 \\ -1/2 & -1 & 0 \\ \sqrt{2}+1 & 0 & \sqrt{2} \end{bmatrix}$$

es, como se puede verificar, una raíz cuadrada de A.

C. Métodos numéricos para determinar la raíz cuadrada de una matriz

Consideremos la siguiente ecuación matricial no lineal:

$$F(X) = X^2 - A = 0 \tag{II.4}$$

Donde A es una matriz compleja de orden $n \times n$. Una solución de X en (II.4) es llamada raíz cuadrada de A. La raíz cuadrada de una matriz tiene muchas aplicaciones en los problemas de valor frontera, como se menciona en [26] y el cálculo del logaritmo matricial, que se estudia en [6] y [11].

En los últimos años ha habido un interés cada vez mayor en desarrollar la teoría y los métodos numéricos para las raíces cuadradas de una matriz. Se han propuesto varios métodos para calcular la raíz cuadrada de una matriz, como se menciona al inicio de este capítulo, sin embargo, en este trabajo desarrollaremos los métodos iterativos. Las iteraciones de la matriz $X_{k+1} = f(X_k)$, donde f es un polinomio o una función, son alternativas atractivas para calcular las raíces cuadradas. A continuación, estudiaremos el método de Newton, el cual tiene un buen comportamiento numérico, debido a su convergencia cuadrática. Seguidamente veremos algunas variantes del método de Newton aunque ellas tengan poca estabilidad numérica.

Finalmente presentaremos algunos algoritmos para calcular la raíz cuadrada de una matriz, desarrollados en [20].

D. Método de Newton para calcular la raíz cuadrada de una matriz

Sea A una matriz compleja de orden n y supongamos que X es una solución de la ecuación (II.4). Definiendo $F : \mathbb{C}^{n\times n} \to \mathbb{C}^{n\times n}$ podemos entonces aplicar el método de Newton, a la ecuación:

$$F(X) = 0,$$

Tomando como $X_0 \in \mathbb{C}^{n\times n}$ un valor inicial de la iteración de Newton, formamos el siguiente sistema matricial:

$$F(X_k) + F'(X_k)(X_{k+1} - X_k) = 0 \tag{II.5}$$

donde F' denota la derivada de Fréchet de F, que se obtiene de:

$$\begin{aligned} F(X_k + H_k) &= (X_k + H_k)^2 = X_h^2 + (H_k X_k + X_k H_k) + H^2 - A \\ &= X_k^2 - A + (H_k X_k + X_k H_k) + H^2 = F(X_k) + (H_k X_k + X_k H_k) + H^2 \\ F(X_k + H_k) &= F(X_k) + F'(X_k) H_k + H^2 \end{aligned}$$

además $F'(X_k)H_k = X_k H_k + H_k X_k$. Así, el método de Newton para calcular la raíz cuadrada de una matriz es dado por la siguiente ecuación iterativa:

$$(P_k) \quad \begin{cases} X_k H_k + H_k X_k &= -F(X_k) \\ X_{k+1} &= X_k + H_k, \text{para } k = 1, 2, \ldots \end{cases} \tag{II.6}$$

A continuación, veamos su algoritmo respectivo.

Algoritmo 1. *(Método de Newton) Consideremos lo siguiente:*

Paso 0. Dado X_0 y ϵ, fijar $k = 0$.

Paso 1. Sea Res$(X_k) = \dfrac{\|X_k^2 - A\|}{\|A\|}$. Si Res$(X_k) < \epsilon$, parar.

Paso 2. Resolver para H_k en la ecuación de Sylvester:

$$X_k H_k + H_k X_k = -F(X_k) \tag{II.7}$$

Paso 3. Actualizar $X_{k+1} = X_k + H_k, k = k + 1$ y volver al paso 1.

Desde que el método es de convergencia cuadrática, debemos dar una matriz inicial que se encuentre cerca de alguna raíz cuadrada. Como podemos observar en la ecuación (II.6), necesitamos resolver en cada iteración un sistema de Sylvester (como se puede observar en los trabajos [25] y [13]) y desde que el costo computacional es muy caro se necesita entonces efectuar otra manera de iteración. Desde que $A = X^2$, entonces A conmuta con X, y por lo tanto, suponiendo que hemos tomado una matriz inicial X_0

cercana a la raíz cuadrada de A, entonces se puede comprobar que $X_k H_k = H_k X_k$, para $k = 1, 2, \ldots$. En este caso, la ecuación (II.6) se puede escribir de la siguiente manera:

$$\begin{aligned} 2H_k X_k &= 2H_k X_k = A - X_k^2 \\ 2X_k(X_{k+1} - X_k) &= A - X_k^2 \\ 2X_k X_{k+1} - X_k^2 &= A \;\Rightarrow\; 2X_{k+1} = X_k + X_k^{-1} A \\ X_{k+1} &= \frac{1}{2}(X_k + A X_k^{-1}) \end{aligned}$$

De la misma manera para $H_k X_k = A - X_k^2$, se obtiene:

$$X_{k+1} = \frac{1}{2}(X_k + A X_k^{-1})$$

obteniendo dos nuevos algoritmos:

$$Y_{k+1} = \frac{1}{2}(Y_k + Y_k^{-1} A) \tag{II.8}$$

$$Z_{k+1} = \frac{1}{2}(Z_k + A Z_k^{-1}) \tag{II.9}$$

Algoritmo 2. *(Método de Newton modificado, ver [16]). Consideremos lo siguiente:*

Paso 0. Dado Y_0, Z_0 *y* ϵ*, fijar* $k = 0$.

Paso 1. Sea Res$(X_k) = \dfrac{\|Y_k^2 - A\|}{\|A\|}$*. Si Res*$(Y_k) < \epsilon$*, parar.*

Paso 2. Actualizar

$$Y_{k+1} = \frac{1}{2}(Y_k + Y_k^{-1} A) \tag{II.10}$$

$$Z_{k+1} = \frac{1}{2}(Z_k + A Z_k^{-1}) \tag{II.11}$$

$k = k + 1$ *y volver al paso 1.*

1. Convergencia del método de Newton modificado

En esta sección veremos la relación entre el método de Newton y los métodos (II.10) y (II.11) enunciados anteriormente. Para la existencia de la iteración de Newton, necesitamos verificar que X_k y $-X_k$ no posean valores propios en común. Para ello, requerimos que X_k sea no singular.

Teorema 17. *Considerando las iteraciones de P_k, (II.10) y (II.11). Supongamos que $X_0 = Y_0 = Z_0$ conmutan con A y que toda iteración de Newton X_k está bien definida, entonces:*

a) X_k conmuta con A, $\forall k$.

b) $X_k = Y_k = Z_k$, $\forall k$.

Demostración. Procedemos por inducción, para el caso $k = 0$, es trivial. Supongamos que el resultado es válido para k. Desde que consideramos X_k no singular, entonces $F'(X_k)$ también lo será, y definiendo:

$$G_k = \frac{1}{2}\left(X_k^{-1}A - X_k\right)$$

Usando la hipótesis $AX_k = X_kA$, obtenemos:

$$\begin{aligned} F'(X_k)G_k &= F'(X_k)\frac{1}{2}\left(X_k^{-1}A - X_k\right) = \frac{1}{2}F'(X_k)X_k^{-1}(A - X_k^2) \\ &= \frac{1}{2}F'(X_k)X_k^{-1}\left(X_kH_k + H_kX_k\right) \\ &= \frac{1}{2}F'(X_k)X_k^{-1}(2X_kH_k) \\ F'(X_k)G_k &= F'(X_k)H_k \end{aligned}$$

Así, $Gk = H_k$. Luego:

$$\begin{aligned} AX_{k+1} &= A\left(X_k + H_k\right) = AX_k + \frac{1}{2}\left(AX_k^{-1}A - AX_k\right) \\ &= X_kA + \frac{1}{2}\left(AX_k^{-1}A - X_kA\right) = X_{k+1}A \end{aligned}$$

Veamos que $X_{k+1} = Y_{k+1} = Z_{k+1}$. Por hipótesis inductiva:

$$X_{k+1} = X_k + \frac{1}{2}\left(X_k^{-1}A - X_k\right) = \frac{1}{2}\left(X_k + X_k^{-1}A\right) = \frac{1}{2}\left(Y_k + Y_k^{-1}A\right) = Y_{k+1}$$

Desde que $X_k^{-1}A = AX_k^{-1}$, se tiene $X_{k+1} = Z_{k+1}$. □

Siguiendo el análisis iterativo, supongamos que A sea diagonalizable. Esto es, que existe una matriz P invertible tal que $P^{-1}AP = D$, con D matriz diagonal. Desde que A es diagonalizable, entonces $D = diag(\lambda_1, \lambda_2, \ldots, \lambda_n)$, donde $\lambda_1, \lambda_2, \ldots, \lambda_n$ son los valores propios de la matriz A. Definiendo entonces:

$$D_k = P^{-1}Y_kZ$$

luego,

$$\begin{aligned} D_{k+1} &= P^{-1}\frac{1}{2}\left(Y_k + Y_k^{-1}A\right)P \\ &= \frac{1}{2}\left(P^{-1}Y_kP + P^{-1}Y_k^{-1}AP\right) \\ &= \frac{1}{2}\left(P^{-1}Y_kP + \left(P^{-1}Y_kP\right)^{-1}Z^{-1}AZ\right) \\ D_{k+1} &= \frac{1}{2}\left(D_k + D_k^{-1}D\right) \end{aligned} \tag{II.12}$$

Con estas condiciones para la matriz A, supongamos que iniciamos el algoritmo con $Y_0 = mI, m > 0$, y sea X la única raíz cuadrada de A para el cual cada valor propio tiene parte real positiva. Luego, tomando $D_0 = mI$, entonces D_k es una matriz diagonal para cada k. Escribiendo $D_k = diag(d_i^k)$ se tiene que:

$$d_i^{k+1} = \frac{1}{2}(d_i^k + \lambda_i/d_i^k), \;\; 1 \leq i \leq n,$$

el cual es la iteración de Newton para determinar la raíz cuadrada de λ_i, $1 \leq i \leq n$. Por tanto podemos considerar la iteración:

$$z_{k+1} = \frac{1}{2}(z_k + a/z_k).$$

Multiplicando por z_k tenemos:

$$\begin{aligned} 2z_{k+1}z_k &= z_k^2 + (\sqrt{a})^2 \\ 2z_{k+1}z_k \pm 2\sqrt{a}z_k &= z_k^2 \pm 2z_k\sqrt{a} + (\sqrt{a})^2 \end{aligned}$$

obtenemos:

$$z_{k+1} \pm \sqrt{a} = \frac{1}{2}(z_k \pm \sqrt{a})^2/(2z_k). \tag{II.13}$$

$$\frac{z_{k+1} - \sqrt{a}}{z_{k+1} + \sqrt{a}} = \left(\frac{z_0 - \sqrt{a}}{z_0 - \sqrt{a}}\right)^{2^{k+1}} = \gamma^{2^{k+1}}. \tag{II.14}$$

Si a no se mantiene en el eje real no positivo, entonces podemos escoger $\sqrt{a}$ para tener la parte real positivo, ($z_0 > 0$, $|\gamma| < 1$). Consecuentemente, para a y Z_0 de las formas específicas tenemos la forma (II.14). Probamos que la secuencia $\{z_k\}$ está bien definida:

$$\lim_{k\to\infty} z_k = \sqrt{a}, \;\; \text{Re}(\sqrt{a}) > 0$$

Desde que los valores propios λ_i y los valores iniciales $d_i^0 = m > 0$, son de la forma de a y z_0, respectivamente, entonces:

$$\lim_{k\to\infty} D_k = D^{1/2} = diag(\lambda_i^{1/2}), \;\; \text{Re}(\lambda_i)^{1/2} > 0 \tag{II.15}$$

y así $\lim_{k\to\infty} Y_k = \lim_{k\to\infty} PD_kP^{-1} = X$.

La unidad se rige del teorema (15). Luego,

$$\begin{aligned} D_{k+1} - D^{1/2} &= \frac{1}{2}(D_k + D_k^{-1}D) - D^{1/2} \\ &= \frac{1}{2}(D_k + D_k^{-1}D - 2D^{1/2}) \\ &= \frac{1}{2}D_k^{-1}(D_k^2 + D - 2D_kD^{1/2}) \\ &= \frac{1}{2}D_k^{-1}(D_k - D^{1/2})^2 \end{aligned}$$

Multiplicando por Z^{-1} y Z, tenemos:

$$\begin{aligned} Z^{-1}D_{k+1}Z - Z^{-1}D^{1/2}Z &= \frac{1}{2}Z^{-1}D_k^{-1}(D_k - D^{1/2})^2 Z \\ Y_{k+1} - X &= \frac{1}{2}Y_k^{-1}(Z^{-1}(D_k - D^{1/2})^2 Z \\ &= \frac{1}{2}Y_k^{-1}(Y_k - X)^{-1} \end{aligned}$$

De ello obtenemos:

$$\|Y_{k+1} - X\| \leq \frac{1}{2}\|Y_k^{-1}\|\|Y_k - X\|^2. \tag{II.16}$$

Corolario 2. *Sea $A \in \mathbb{C}$ una matriz hermitiana definida positiva. Si $Y_0 = mI$, $m > 0$, entonces las iteraciones $\{Y_k\}$ en (II.10) son todas hermitianas definidas positivas, $\lim_{k\to\infty} Y_k = X$, donde X es la única raíz cuadrada hermitiana definida positiva de A, y (II.16) se mantiene.*

Otra variante del método de Newton es usando la función signo estudiada por Denman y Beavers en [10] de la siguiente manera:

$$\begin{aligned} & P_0 = A,\ Q_0 = I, \\ \left.\begin{aligned} P_{k+1} &= \frac{1}{2}\left(P_k + Q_k^{-1}\right) \\ Q_{k+1} &= \frac{1}{2}\left(Q_k + P_k^{-1}\right) \end{aligned}\right\}, & k = 0, 1, 2, \ldots \end{aligned}$$

Tomando como $Y_0 = I$, tenemos que $Y_1 = \frac{1}{2}(I + A) = P_1$ y $Q_1 = \frac{1}{2}(I + A^{-1}) = \frac{1}{2}A^{-1}Y_1$, suponiendo que $P_k = Y_k$ y $Q_k = A^{-1}Y_k$, entonces:

$$\begin{aligned} P_{k+1} &= \frac{1}{2}\left(P_k + Q_k^{-1}\right) = \frac{1}{2}\left(Y_k + Y_k^{-1}A\right) = Y_{k+1} \\ Q_{k+1} &= \frac{1}{2}\left(Q_k + Y_k^{-1}\right) = \frac{1}{2}\left(A^{-1}Y_k + Y_k^{-1}\right) \\ &= \frac{1}{2}A^{-1}\left(Y_k + AY_k^{-1}\right) = \frac{1}{2}A^{-1}\left(Y_k + AY_k^{-1}\right) \\ &= \frac{1}{2}A^{-1}\left(Y_k + Y_k^{-1}\right) = A^{-1}Y_{k+1}. \end{aligned}$$

Estabilidad del algoritmo

Para garantizar la estabilidad numérica de las iteraciones se requiere que el error del factor de amplificación $\pi_{ij}^{(k)} = \frac{1}{2}\left(1 - \frac{\lambda_j}{d_i^{(k)} d_j^{(k)}}\right)$, sea menor que 1.

Un caso particular es:

$$\frac{1}{2}\left|1 - \left(\frac{\lambda_i}{\lambda_j}\right)^{1/2}\right| \leq 1, \quad 1 \leq i, j \leq n.$$

Por ejemplo en el caso que la matriz A sea hermitiana definida positiva, entonces el condicionamiento es $k(A) \leq 9$. Para más detalles, ver [16].

Un caso particular del algoritmo 1 es la iteración Denman-Beavers, cuyo algoritmo veremos a continuación.

Algoritmo 3. *(la iteración Denman-Beavers, ver [10]). Consideremos lo siguiente:*

Paso 0. Dado $Y_0 = A, Z_0 = I$ *y* ϵ*, fijar* $k = 0$.

Paso 1. Sea Res$(X_k) = \dfrac{\|Y_k^2 - A\|}{\|A\|}$*. Si Res*$(Y_k) < \epsilon$*, parar.*

Paso 2. Actualizar

$$\begin{aligned} Y_{k+1} &= \frac{Y_k + Z_k^{-1}}{2} \\ Z_{k+1} &= \frac{Z_k + Y_k^{-1}}{2}, \end{aligned} \tag{II.17}$$

$k = k + 1$*, y volver al paso 1.*

E. Otros métodos para hallar la raíz cuadrada de una matriz

En esta sección presentamos otros algoritmos para calcular la raíz cuadrada de una matriz no singular A.

La primera idea puede enunciarse de la siguiente manera: si (II.4) tiene una solución no singular X, entonces la expresión (II.4) se puede transformar en una ecuación matricial no lineal equivalente a:

$$F(X) = X - AX^{-1} = 0 \tag{II.18}$$

Luego, se aplica el método de Newton a (II.18) para calcular la raíz cuadrada de una matriz no singular A. Por la definición de la derivada de Frechet, se tiene que si la matriz X es no singular, entonces, la función F es derivable en X y

$$F'(X)H = H + AX^{-1}HX^{-1} \tag{II.19}$$

Así, el método de Newton para (II.8) se puede escribir como:

$$\text{Dado } X_0, \ \ , X_{k+1} = X_k - (F'(X_k))^{-1}(F(X_k)), \quad k = 0, 1, 2, \cdots \tag{II.20}$$

Combinando (II.19), la iteración (II.20) es equivalente al siguiente algoritmo.

Algoritmo 4. *Consideremos lo siguiente.*

Paso 0. Dado X_0 y ε, fijar $k = 0$.

Paso 1. Dar Res$(X_k) = \dfrac{\|X_k^2 - A\|}{\|A\|}$. Si Res$(X_k) < \varepsilon$, parar.

Paso 2. Resolver para H_k en general, la ecuación de Sylvester:

$$AX_k^{-1}H_kX_k^{-1} + H_k = -F(X_k) \tag{II.21}$$

Paso 3. Actualizar $X_{k+1} = X_k + H_k, k = k + 1$, y volver al paso 1 cuando $Res(X_k) = \dfrac{\|X_k^2 - A\|}{\|A\|}$.

También, usando la técnica de Samanskii al método de Newton (II.5) tenemos el siguiente algoritmo.

Algoritmo 5. *Consideremos lo siguiente.*

Paso 0. Dado X_0, m *y* $\varepsilon > 0$*, fijar* $k = 0$*.*

Paso 1. Sea Res$(X_k) = \dfrac{\|X_k^2 - A\|}{\|A\|}$*. Si Res*$(X_k) < \varepsilon$*, parar.*

Paso 2. Sea $X_{k,0} = X_k,\ i = 1$*.*

Paso 3. Si $i > m$*, ir al paso 6.*

Paso 4. Resolver para $H_{k,i-1}$ *en la ecuación generalizada de Sylvester:*

$$AX_k^{-1}H_{k,i-1}X_k^{-1} + H_{k,i-1} = -F(X_{k,i-1}). \tag{II.22}$$

Paso 5. Actualizar $X_{k,i} = X_{k,i-1} + H_{k,i-1},\ i = i + 1$*, e ir al paso 3.*

Paso 6. Actualizar $X_{k+1} = X_{k,m},\ k = k + 1$ *e ir al paso 1.*

F. Teoremas de convergencia

En esta sección, estableceremos los teoremas de convergencia local para los algoritmos 2 y 3. Empezamos con algunos lemas.

Lema 5. *[Ver [19]] Sea* T *un operador no lineal desde un espacio de Banach* E *en sí mismo y sea* $x^* \in E$ *una solución de* $x = Tx$*. Si* T *es Frechet diferenciable en* x^* *con* $\rho(T'_{x^*}) < 1$*, entonces la iteración* $x_{n+1} = Tx_n, n = 0, 1, 2, \cdots$*, converge a* x^**, siempre que* x_0 *esta suficientemente cerca de* x^**.*

Lema 6. *[Ver [24]] Sea* $A, B \in \mathbb{C}^{n\times n}$*. Si* A *es invertible con* $\|A^{-1}\| \leq \alpha$ *y se cumple* $\|A - B\| \leq \beta$ *y* $\alpha\beta < 1$*, entonces* B *es también invertible y:*

$$\|B^{-1}\| \leq \frac{\alpha}{1 - \alpha\beta}. \tag{II.23}$$

Lema 7. *Si la matriz $\widehat{X} \in \mathbb{C}^{n\times n}$ es no singular, entonces existe $\gamma > 0$ y $L > 0$ tal que, para todo $X, Y \in B(\widehat{X}, \gamma)$ se cumple:*

$$\|F'_X - F'_Y\| \leq L\|X - Y\|, \tag{II.24}$$

donde $B(\widehat{X}, r) = \{X \mid \|X - \widehat{X}\| < \gamma\}$ y F'_X, F'_Y son F-derivables definido por (II.19) en X, Y.

Demostración. Sea $\alpha = \|\widehat{X}^{-1}\|$ y seleccionamos $0 < \gamma < \alpha^{-1}$.

Del lema 6 se sigue que X es no singular y $\|X^{-1}\| \leq \alpha/(1-\alpha\gamma)$ para cada $\widehat{X} \in B(X_0, \gamma)$. Entonces F'_X está bien definido, y también F'_Y, donde $Y \in B(\widehat{X}, \gamma)$. De acuerdo con (II.19), tenemos:

$$\begin{aligned}
\|F'_X(H) - F'_Y(H)\| &= \|(H + AX^{-1}HX^{-1}) - (H + AY^{-1}HY^{-1})\| \\
&= \|AX^{-1}HX^{-1} - AY^{-1}HY^{-1}\| \\
&= \|A[(X^{-1}HX^{-1} - X^{-1}HY^{-1}) + (X^{-1}HY^{-1} - Y^{-1}HY^{-1})]\| \\
&= \|A[X^{-1}H(X^{-1} - Y^{-1}) + (X^{-1} - Y^{-1})HY^{-1}]\| \\
&= \|A[X^{-1}HY^{-1}(Y - X)X^{-1} + Y^{-1}(Y - X)X^{-1}HY^{-1}] \\
&\leq \|A\|(\|X^{-1}\|^2\|Y^{-1}\|^2\|Y - X\|\|H\| + \|Y^{-1}\|^2\|X^{-1}\|^2\|Y - X\|\|H\|) \\
&= \|A\|\|X^{-1}\|\|Y^{-1}\|(\|X^{-1}\| + \|Y^{-1}\|)\|X - Y\|\|H\| \\
&\leq \frac{\alpha}{1-\alpha\gamma}\frac{\alpha}{1-\alpha\gamma}\left(\frac{\alpha}{1-\alpha\gamma} + \frac{\alpha}{1-\alpha\gamma}\right)\|A\|\|X - Y\|\|H\| \\
&= 2\left(\frac{\alpha}{1-\alpha\gamma}\right)^3\|A\|\|X - Y\|\|H\| \\
&= L\|X - Y\|\|H\|,
\end{aligned} \tag{II.25}$$

donde $L = 2\left(\dfrac{\alpha}{1-\alpha\gamma}\right)^3\|A\|$. Por lo tanto, tenemos:

$$\|F'_X - F'_Y\| \leq L\|X - Y\|. \tag{II.26}$$

□

Teorema 18. *Si (II.10) tiene una solución no singular X_* y el mapeo $F'_{X_*} : \mathbb{C}^{n\times n} \to \mathbb{C}^{n\times n}$ es invertible entonces existe una bola cerrada $S = B(X_*, \delta)$, tal que, para todo $X_0 \in S$, la sucesión $\{X_k\}$ generada por el algoritmo 2 converge al menos cuadráticamente a la solución X_*.*

Demostración. Sea $\varphi(X) = X - (G'_X)^{-1}(G(X))$. Por la fórmula de Taylor en el espacio de Banach (ver [9], p. 67), tenemos:

$$
\begin{aligned}
&\lim_{\|H\|\to 0} \frac{\|\varphi(X_* + H) - \varphi(X_*)\|}{\|H\|} \\
&= \lim_{\|H\|\to 0} \left(\|[X_* + H - (G'_{X_*+H})] - [X_* - (G'_{X_*})^{-1}(G(X_*))]\| \times \|H\|^{-1}\right) \\
&= \lim_{\|H\|\to 0} \left(\|H + (G'_{X_*})^{-1}(G(X_{X_*})) - (G'_{X_*+H})^{-1} \times (G(X_* + H))\| \times \|H\|^{-1}\right) \\
&= \lim_{\|H\|\to 0} \Big(\Big\|H + (G'_{X_*})^{-1}(G(X_{X_*})) - (G'_{X_*+H})^{-1} \times \\
&\left[G(X_*) + G'_{X_*}(H) + \frac{1}{2}G''_{X_*}(H^2) + \cdots\right]\Big\| \times \|H\|^{-1}\Big) \\
&= \lim_{\|H\|\to 0} (\|H + (G'_{X_*})^{-1}(G(X_{X_*})) - (G'_{X_*+H})^{-1} \times G((X_*)) + (G'_{X_*+H})^{-1} \times \\
&(G'_{X_*}(H)) + \frac{1}{2}(G'_{X_*+H})^{-1}(G''_{x_*}(H^2) + \cdots\| \times \|H\|^{-1})) = 0
\end{aligned}
$$

Por lo tanto, la F-derivada de φ en X_* es 0. Por el lema 5, derivamos la sucesión $\{X_k\}$ generada por la iteración (5) y converge a X_*. Esto también se obtiene de la sucesión $\{X_k\}$ generada por el algoritmo 2 y converge a X_*.

Sea $\|F'_{X_*}\| = \beta$, de acuerdo a $X_k \to X_*(k \to \infty)$ y el lema 6; para un k suficientemente grande, tenemos:

$$\|(F'_{X_k})^{-1}\| \leq \frac{\beta}{1 - \beta(1/2\beta)} = 2\beta. \tag{II.27}$$

Por el lema 7, tenemos:

$$\|F'_{X_k}(X_k - X_*) - F'_{X_*}(X_k - X_*)\| \leq L\|X_k - X_*\|^2. \tag{II.28}$$

Usando la fórmula de Taylor una vez más, para todo $t \in [0, 1]$ tenemos:

$$\begin{aligned}
\|F(X_k) - F(X_*) - F'(X_*)(X_k - X_*)\| &\leq \left\| \int (F'_{X_{k+t}(X_* - X_k)}(X_k - X_*) - F'_{X_0}(X_k - X_*))dt \right\| \\
&\leq \int \|F'_{X_{k+t}(X_* - X_k)} - F'_{X_*}\|dt \|X_k - X_*\| \\
&\leq L\|X_{k+t}(X_* - X_k) - X_*\| \|X_k - X_*\| \\
&= L(1-t)\|X_k - X_*\|^2 \\
&\leq L\|X_k - X_*\|^2. \qquad \text{(II.29)}
\end{aligned}$$

Por lo tanto,

$$\begin{aligned}
&\|X_{k+1} - X_*\| \\
&= \|X_k - (F'_{X_k})^{-1}(F(X_k)) - X_*\| \\
&= \|(F'_{X_k})^{-1}[F'_{X_k}(X_k - X_*) - FX_k]\| \\
&= \|(F'_{X_k})^{-1}[(F'_{X_k}(X_k - X_*) - F'_{X_*}(X_k - X_*)) - (F(X_k) - F(X_*) - F'_{X_*}(X_k - X_*))]\| \\
&\leq \|(F'_{X_k})^{-1}\|[\|F'_{X_k}(X_k - X_*) - F'_{X_*}(X_k - X_*)\| + \|F(X_k) - F(X_*) - F'_{X_*}(X_k - X_*)\|]
\end{aligned} \tag{II.30}$$

Combinando II.27 y II.30, tenemos:

$$\|X_{k+1} - X_*\| \leq 2\beta L\|X_k - X_*\|^2 + 2\beta L\|X_k - X_*\|^2 \tag{II.31}$$

$$= 4\beta L\|X_k - X_*\|^2, \tag{II.32}$$

lo que implica que la sucesión $\{X_k\}$ generada por el algoritmo 2 converge, al menos cuadráticamente a la solución de X_*. □

Teorema 19. *Si (II.19) tiene una solución no singular X_* y el mapeo $F'_{X_*} : \mathbb{C}^{n\times n} \to \mathbb{C}^{n\times n}$ es invertible, entonces existe una bola cerrada $S = B(X_*, \delta)$, tal que, para todo $X_0 \in S$, la sucesión $\{X_k\}$ generada por el algoritmo 3 converge al menos cúbicamente a la solución X_*.*

Demostración. Sea $\varphi(X) = X - (F'_X)^{-1}(F(X))$. Por la fórmula de Taylor en el espacio de Banach (ver [9], p. 67), tenemos:

$$\lim_{\|H\|\to 0} \frac{\|\varphi(X_* + H) - \varphi(X_*)\|}{\|H\|}$$

$$= \lim_{\|H\|\to 0} \left(\|[X_* + H - (F'_{X_*+H})^{-1}(F(X_* + H))] - [X_* - (F'_{X_*})^{-1}(F(X_*))]\| \times \|H\|^{-1}\right)$$

$$= \lim_{\|H\|\to 0} (\|H + (F'_{X_*})^{-1}(F(X_*)) - (F'_{X_*+H})^{-1} \times (F(X_* + H))\| \times \|H\|^{-1}$$

$$= \lim_{\|H\|\to 0} (\|H + (F'_{X_*})^{-1}(F(X_*)) - (F'_{X_*+H})^{-1} \times$$

$$\left[F(X_*) + F'_{X_*}(H) + \frac{1}{2}F''_{X_*}(H^2) + \cdots\right] \| \times \|H\|^{-1})$$

$$= \lim_{\|H\|\to 0} (\|H + (F'_{X_*})^{-1}(F(X_*)) - (F'_{X_*+H})^{-1} \times F((X_*)) + (F'_{X_*+H})^{-1} \times$$

$$(F'_{X_*}(H)) + \frac{1}{2}(F'_{X_*+H})^{-1}(F''_{X_*}(H^2)) + \cdots \| \times \|H\|^{-1})) = 0. \tag{II.33}$$

Por lo tanto, la F-derivada de φ en X_* es 0. Por el lema 5, derivamos la sucesión $\{X_k\}$ generada por la iteración (5) y converge a X_*. Esto también se obtiene de la sucesión $\{X_k\}$ generada por el algoritmo 3 y converge a X_*.

Sea $\|(F'_{X_*})^{-1}\| = \beta$, de acuerdo a $X_k \to X_*(k \to \infty)$ y el lema 6; para un k suficientemente grande, tenemos:

$$\|(F'_{X_k})^{-1}\| \leq \frac{\beta}{1 - \beta(1/2\beta)} = 2\beta. \tag{II.34}$$

Por el lema 7, tenemos:

$$\|F'_{X_k}(X_{k,1} - X_*) - F'_{X_*}(X_{k,1} - X_*)\| \leq L\|X_k - X_*\|\|X_{k,1} - X_*\|. \tag{II.35}$$

Usando la fórmula de Taylor una vez más, para todo $t \in [0,1]$ tenemos:

$$\begin{aligned}
\|F(X_{k,1}) - F(X_*) - F'_{X_*}(X_{k,1} - X_*)\| &\leq \left\| \int (F'_{X_{k,1}+t(X_*-X_{k,1})}(X_{k,1} - X_*) - F'_{X_*}(X_{k,1} - X_*))dt \right\| \\
&\leq \int \|F'_{X_{k,1}+t(X_*-X_{k,1})} - F'_{X_*}\| dt \cdot \|X_{k,1} - X_*\| \\
&\leq L\|X_{k,1} + t(X_* - X_{k,1}) - X_*\| \|X_{k,1} - X_*\| \\
&= L(1-t)\|X_{k,1} - X_*\|^2 \\
&\leq L\|X_{k,1} - X_*\|^2. \qquad \text{(II.36)}
\end{aligned}$$

Por lo tanto,

$$\begin{aligned}
\|X_{k+1} - X_*\| &= \|X_{k,1} - (F'_{X_k})^{-1}(F(X_{k,1})) - X_*\| \\
&= \|(F'_{X_k})^{-1}[F'_{X_k}(X_{k,1} - X_*) - F(X_{k,1})]\| \\
&= \|(F'_{X_k})^{-1} \times [(F'_{X_k}(X_{k,1} - X_*) - F'_{X_*}(X_{k,1} - X_*)) \\
&- (F(X_{k,1}) - F(X_*) - F'_{X_*}(X_{k,1} - X_*))]\| \\
&\leq \|(F'_{X_k})^{-1}\| \times [\|F'_{X_k}(X_{k,1} - X_*) - F'_{X_*}(X_{k,1} - X_*)\| \\
&+ \|F(X_{k,1}) - F(X_*) - F'_{X_*}(X_{k,1} - X_*)\|] \qquad \text{(II.37)}
\end{aligned}$$

Combinando II.34 y II.37 con el teorema 18, tenemos:

$$\begin{aligned}
\|X_{k+1} - X_*\| &\leq 2\beta \left[L\|X_k - X_*\| \cdot \|X_{k,1} - X_*\| + L\|X_{k,1} - X_*\|^2\right] \\
&\leq 2\beta L \left[4\beta L^2 \|X_k - X_*\|^3 + 16\beta^2 L^3 \|X_k - X_*\|^4\right] \\
&= (8\beta^2 L^3 + 32\beta^3 L^4 \|X_k - X_*\|)\|X_k - X_*\|^3 \qquad \text{(II.38)} \\
&\leq (8\beta^2 L^3 + 32\beta^3 L^4 \delta)\|X_k - X_*\|^3 \\
&= M\|X_k - X_*\|,
\end{aligned}$$

donde $M = 8\beta^2 L^3 + 32\beta^3 L^4 \delta$. Por lo tanto, la sucesión $\{X_k\}$ generada por el algoritmo 3 converge al menos cúbicamente a la solución X_*. □

1. Estabilidad del algoritmo

De acuerdo con [6] definimos una iteración $X_{k+1} = f(X_k)$ que se establece en una vecindad cuya solución es $X = f(X)$, si la matriz de error $E_k = X_k - X_*$ satisface

$$E_{k+1} = L(E_k) + O(\|E_k\|^2)$$

donde L es un operador lineal acotado. Este operador implicará que una pequeña perturbación introducida en un determinado paso no se amplifique posteriormente.

Notemos que esta definición de estabilidad es una propiedad asintótica y es diferente del concepto usual de estabilidad numérica, que concierne la propagación de error global, con el objetivo de limitar el error relativo mínimo sobre las iteraciones calculadas.

Considerando en la iteración k, $X_k = E_k + X_*$ sobre los métodos de Newton modificados obtenemos:

$$E_{k+1}X_* + X_*E_{k+1} = 0.$$

Por lo tanto, bajo la condición que X_* y $-X_*$ no tienen valores propios en común, la iteración 5 del algoritmo 2 de Newton modificado tiene estabilidad óptima. Para más detalles ver [20].

Otros algoritmos para determina la raíz cuadrada de una matriz son los siguientes:

Algoritmo 6. *(La iteración de la escala de Denman-Beavers, ver [10]). Consideremos lo siguiente.*

Paso 0. Dado $Y_0 = A, Z_0 = I$, $p \in \mathbb{N}$, n *el orden de la matriz* A, $\varepsilon > 0$ *y fijar* $k = 0$.

Paso 1. Hallando los ϵ_i *y* α_i^2,

$$\xi_i = \frac{1}{2}\left(1 + \cos\frac{(2i-1)\pi}{2p}\right), \quad \alpha_i^2 = \frac{1}{\xi_i} - 1, \qquad i = 1, 2, \cdots, p. \tag{II.39}$$

Paso 2. Sea $Res(X_k) = \dfrac{\|X_k^2 - A\|}{\|A\|}$. *Si* $Res(X_k) < \varepsilon$, *parar.*

Paso 3. Actualizar

$$\begin{aligned} r_k &= |\det(Y_k)\det(Z_k)|^{-1/2n} \\ Y_{k+1} &= \frac{r_k Y_k + r_k^{-1} Z_k^{-1}}{2} \\ Z_{k+1} &= \frac{r_k Z_k + r_k^{-1} Y_k^{-1}}{2}, \end{aligned} \tag{II.40}$$

hacer $k = k + 1$, *e ir al paso 2.*

Algoritmo 7. *(la iteración de Padé, ver [16]). Consideremos lo siguiente.*

Paso 0. Dado $Y_0 = A, Z_0 = I$, $p \in \mathbb{N}$, n *el orden de la matriz* A, $\varepsilon > 0$ *y fijar* $k = 0$.

Paso 1. Hallando los ϵ_i *y* α_i^2,

$$\xi_i = \frac{1}{2}\left(1 + \cos\frac{(2i-1)\pi}{2p}\right), \quad \alpha_i^2 = \frac{1}{\xi_i} - 1, \quad i = 1, 2, \cdots, p. \tag{II.41}$$

Paso 2. Sea $Res(X_k) = \dfrac{\|X_k^2 - A\|}{\|A\|}$. *Si* $Res(X_k) < \varepsilon$, *parar.*

Paso 3. Actualizar

$$\begin{aligned} Y_{k+1} &= \frac{1}{p} Y_k \sum_{i=1}^{p} \frac{1}{\xi_i} (Z_k Y_k + \alpha_i^2 I)^{-1} \\ Z_{k+1} &= \frac{1}{p} Z_k \sum_{i=1}^{p} \frac{1}{\xi_i} (Y_k Z_k + \alpha_i^2 I)^{-1}, \end{aligned} \tag{II.42}$$

hacer $k = k + 1$, *e ir al paso 2.*

Algoritmo 8. *(la iteración en escala de Padé, ver [16]). Consideremos lo siguiente.*

Paso 0. Dado $Y_0 = A, Z_0 = I$, $p \in \mathbb{N}$, n *el orden de la matriz* A, $\varepsilon > 0$ *y fijar* $k = 0$.

Paso 1. Hallar los ϵ_i *y* α_i^2,

$$\xi_i = \frac{1}{2}\left(1 + \cos\frac{(2i-1)\pi}{2p}\right), \quad \alpha_i^2 = \frac{1}{\xi_i} - 1, \quad i = 1, 2, \cdots, p. \tag{II.43}$$

Paso 2. Sea $Res(X_k) = \dfrac{\|X_k^2 - A\|}{\|A\|}$. *Si* $Res(X_k) < \varepsilon$, *parar.*

Paso 3. Actualizar escalonamiento

$$r_k = |\det(Y_k)\det(Z_k)|^{-1/2n}$$

Paso 4. Actualizar

$$\begin{aligned} Y_{k+1} &= \frac{1}{p} Y_k \sum_{i=1}^{p} \frac{1}{\xi_i} (r_k^2 Z_k Y_k + \alpha_i^2 I)^{-1} \\ Z_{k+1} &= \frac{1}{p} Z_k \sum_{i=1}^{p} \frac{1}{\xi_i} (r_k^2 Y_k Z_k + \alpha_i^2 I)^{-1}, \end{aligned} \tag{II.44}$$

hacer $k = k + 1$, e ir al paso 2.

G. Un álgebra aproximada de la matriz exponencial para problemas rígidos

Debido a que la estabilidad involucra la función estabilidad ψ asociada a un método Φ de un paso, entonces se pueden construir funciones de estabilidad que cumplan la condición de Padé y generar nuevo métodos, pero todas esas funciones de estabilidad se limitarían a la forma racional, es por ello que se ve en la necesidad de estudiar otros tipos de funciones de estabilidad que generen una aproximación a la función exponencial y que no sean de la forma racional. A continuación estudiaremos funciones de estabilidad de la forma irracional.

Por razones de cálculo de la función ψ es, en general, restringido a funciones polinómicas racionales, como en el caso de la aproximación de Padé estudiada anteriormente. Sin embargo, estas funciones solo son buenas aproximaciones a la función exponencial en la parte izquierda del plano complejo. Debido a que se consideran ecuaciones diferenciales rígidas, estos rangos no se pueden restringir demasiado. En esta sección se presentará una

familia de esquemas de un paso incondicionalmente estables de orden dos. En una aproximación algebraica a la función exponencial basada en la raíz cuadrada de una matriz se tienen muchas de las propiedades de estabilidad necesarias siempre que el argumento $z = h\lambda$ se mantenga alejado del eje imaginario más allá de los puntos $\pm i$.

Se presenta un esquema desarrollado en [26], para métodos de un paso sobre problemas lineales, rígidos de ecuaciones diferenciales cuya función de estabilidad contiene la raíz cuadrada de una matriz.

Dado el problema rígido:

$$y' = Ay, \quad y(t) \in \mathbb{C}^n \tag{II.45}$$

y sea la función de estabilidad $\psi(z)$, que aproxima a la función e^z. Esta describe la resolución entre dos valores de soluciones y_k y y_{k+1}, en puntos adyacentes en una malla, t_k, t_{k+1} (aquí estudiaremos el caso de una malla uniforme) con distancia $h = t_{k+1} - t_k$ en la forma

$$y_{k+1} = \psi(hA)y_k \tag{II.46}$$

Comencemos entonces definiendo una función de estabilidad ψ y seguidamente una familia de funciones de estabilidad basadas en la función de estabilidad ψ que aproxima a la función exponencial en orden 2. Luego, definiremos los métodos generados por esta familia de funciones de estabilidad.

Definición 21. *Sea* $D = \mathbb{C} - \{z \in \mathbb{C} : |Im(z)| \geq 1 \wedge Re(z) = 0\}$ *y la función de estabilidad*

$$\psi(z) = z + \sqrt[+]{1 + z^2} \tag{II.47}$$

donde $\sqrt[+]{z} = \sqrt{\frac{|z| + Re(z)}{2}} + isgn(Im(z))\sqrt{\frac{|z| - Re(z)}{2}}$.

El siguiente lema muestra el orden de aproximación de la función ψ hacia la función exponencial y algunas propiedades similares a ella.

Lema 8. *Sea $z \in D$, entonces:*

a) $\psi(z) = e^z + O(z^3), z \to 0.$

b) $\psi(-z) = \dfrac{1}{\psi(z)}.$

c) $Re(\psi(z)) > 0,$

d) $|\psi(z)| \leq 1 \Leftrightarrow Re(z) \leq 0, |\psi(z)| \geq 1 \Leftrightarrow Re(z) \geq 0.$

e) $\psi(-|z|) \leq |\psi(z)| \leq \psi(Re(z))$ *si* $Re(z) < 0, \psi(Re(z)) \leq |\psi(z)| \leq \psi(|z|)$, *si* $Re(z) > 0.$

f) $|\psi(z)| \leq \dfrac{1}{|z|}$, *si* $Re(z) < 0, |\psi(z)| \geq |z|$ *si* $Re(z) \geq 0.$

Demostración. Ver [26] □

Comentario: Respecto a la propiedad $a)$ se observa que la aproximación a la función exponencial es de orden 2, mientras que en las propiedades $b), d)$ y $e)$, el comportamiento de la función ψ es similar a la función exponencial. Así, podemos esperar excelentes propiedades de estabilidad si los valores propios de la matriz A se mantienen en dicho eje. Respecto a la propiedad $d)$ garantiza la A-estabilidad.

Lema 9. *Si $\sigma(A)$ denota el espectro de la matriz A. Entonces, bajo la condición $\sigma(h_k A) \subset D$, las expresiones (II.46) y (II.47) están bien definidas.*

Demostración. La raíz cuadrada de la matriz $(I + h^2 A^2)^{1/2}$ es únicamente definida vía la forma de Jordan de la matriz A escogiendo los valores propios de la raíz cuadrada con parte real positiva y excluyendo los puntos $+i$ y $-i$ del espectro de hA. □

Los resultados del lema 8 ya serían suficientes para un análisis de estabilidad en el caso de coeficientes constantes considerados aquí, pero las estimaciones de estabilidad con norma acotada son más fáciles extender a ecuaciones diferenciales generales.

Usaremos la norma $\|\cdot\|_2$, la notación $A^H = -\bar{A}^T$ y:

$$\mathrm{Re}(A) = \frac{1}{2}(A + A^H), \quad \mathrm{Im}(A) = \frac{1}{2i}(A - A^H).$$

El siguiente teorema explica tres estimaciones para la función matricial.

Teorema 20. *Sea A una matriz, $\mu, \nu \in \mathbb{R}$. Entonces*

a) $Re(A) \leq \mu I$, $\mu < 0 \rightarrow \|\psi(hA)\|_2 \leq \psi(h\mu)$, $Re(\psi(hA)) \geq 0$ *para cualquier* $h \geq 0$.

b) $Re(A) \geq \nu I$, $\nu > 0 \rightarrow \|\psi(hA)^{-1}\|_2 \leq \psi(h\nu)^{-1}$, $Re(\psi(hA)) \geq 0$ *para cualquier* $h \geq 0$.

c) $(Im(A))^2 < (Re(A))^2 + k^2 I$, $k \geq 0 \rightarrow \max\{\|\psi(hA)\|_2, \|\psi(hA)^{-1}\|_2\} \leq 1 + h\|A\|_2 + h^2\|A\|_2^2$ *para cualquier* $h \geq 0$ *y* $hk \leq 1$.

En cualquiera de estos casos la raíz cuadrada $X = (I + h^2A^2)^{1/2}$ contenida en $\psi(hA)$ está definida y $Re(X) \geq 0$.

Demostración. Ver [26] □

Veamos ahora una familia de esquemas con una generalización de la función de estabilidad ψ.

$$\psi_\omega(z) = \frac{2 - \omega + z + \sqrt{\omega^2 + z^2}}{2 - \omega - z + \sqrt{\omega^2 + z^2}}, \quad 0 < \omega \leq 2. \tag{II.48}$$

De la definición de la función de estabilidad:

$$
\begin{aligned}
\psi_\omega(z) &= \frac{2-\omega+z+\sqrt{\omega^2+z^2}}{2-\omega-z+\sqrt{\omega^2+z^2}} \\
&= \frac{2-\omega+\omega\left(\frac{z}{\omega}+\sqrt{1+\left(\frac{z}{\omega}\right)^2}\right)}{2-\omega-\omega\left(\frac{z}{\omega}-\sqrt{1+\left(\frac{z}{\omega}\right)^2}\right)} \cdot \frac{\left(\frac{z}{\omega}+\sqrt{1+\left(\frac{z}{\omega}\right)^2}\right)}{\left(\frac{z}{\omega}+\sqrt{1+\left(\frac{z}{\omega}\right)^2}\right)} \\
&= \frac{2-\omega+\omega\psi(z/\omega)}{(2-\omega)\psi(z/\omega)-\omega\left(\left(\frac{z}{\omega}\right)^2-\left(1+\left(\frac{z}{\omega}\right)^2\right)\right)} \cdot \psi(z/\omega) \\
&= \frac{2-\omega+\omega\psi(z/\omega)}{(2-\omega)\psi(z/\omega)+\omega} \cdot \psi(z/\omega)
\end{aligned}
$$

En el caso que $\omega = 1$ tenemos la función de estabilidad estudiada inicialmente. Se cumplen las siguientes propiedades para la función ψ_ω.

Lema 10. *Sea $z \in D$, $0 < \omega \leq 2$. Entonces:*

a) $\psi_\omega(z) = e^z + O(z^3), \quad z \to 0.$

b) $\psi_\omega(-z) = \dfrac{1}{\psi_\omega(z)}, \psi_\omega(z) \neq 0, \psi_\omega(z) \neq \infty.$

c) $|\psi_\omega(z)| \leq 1$ *si y solo si* $Re(z) \leq 0, |\psi(z)| \geq 1$ *si y solo si* $Re(z) \geq 0.$

d) $|\psi_\omega(z)| \leq \psi_\omega(Re(z)), Re(z) \leq 0$, *y con* $x \in \mathbb{R}$, $x \leq 0$, $\psi_\omega(x) \leq \dfrac{1}{1+|x|}, 1 \leq \omega \leq 2, |\psi_\omega(x)| \leq \dfrac{2}{2+|x|}, 0 < \omega \leq 1.$

Demostración. Ver [26] □

H. Regla de la raíz trapezoidal RT-ω

Existe una familia de métodos de un paso con la función estabilidad ψ_ω para la ecuación homogénea:

$$u'(t) = Au(t) + g(t), \quad x \in [a, b]. \tag{II.49}$$

Llamaremos a estos métodos "reglas de raíz trapezoidal" (RT-ω), donde ω denota el parámetro involucrado. La solución u de (II.49) es aproximada sobre una malla $\{x_k\}$, $a = x_0 < x_1 < \ldots < x_N = b$, $h = x_{k+1} - x_k = \dfrac{b-a}{N}$, $k = 0, \ldots, N-1$, y $0 < \omega \leq 2$. Con $\sigma(\omega^{-1}hA) \subset D$ y $X = (\omega^2 I + h^2 A^2)^{1/2}$ la ecuación diferencial (II.49) es reemplazada por la regla $u_{n+1} = \psi_w(hA)u_n$

$$u_{n+1} \;=\; ((2-\omega)I - hA + X)^{-1}\,(((2-\omega)I + hA + X)u_n + G_k) \tag{II.50}$$

donde $G_k = h(g_{k+1} + g_k) - (\omega I + X)^{-1}h^2 A(g_{k+1} - g_k)$, obteniendo los métodos RT-ω. Una propiedad básica de estas reglas se muestran en el siguiente lema.

Lema 11. *La función estabilidad de la regla RT-ω, con $0 < \omega \leq 2$, es $\psi_\omega(z)$. Así, RT-1($\omega = 1$) tiene $\psi(z)$ y RT-2($\omega = 2$) tiene como función estabilidad $\psi(z/2)^2$.*

Demostración. Para la ecuación escalar $u' = \lambda u$, con $z = h_k\lambda$ la regla (II.50) reduce a

$$(2 - \omega - z + \sqrt{\omega^2 + z^2})y_{k+1} - (2 - \omega + z + \sqrt{\omega^2 + z^2})y_k = 0,$$

el cual se muestra en (II.48). Para $\omega = 1$ esto da $[1+\psi(-z)]y_{k+1} = [1+\psi(z)]y_k$. También, el lema 8(b) muestra que $y_{k+1} = \psi(z)y_k$. Para $\omega = 2$ tenemos:

$$(\sqrt{4 + z^2} - z)y_{k+1} - (\sqrt{4 + z^2} + z)y_k = 2[\psi(-z/2)y_{k+1} - \psi(z/2)y_k] = 0,$$

donde $y_{k+1} = \psi(z/2)^2 y_k$. □

El error de consistencia de los esquemas II.50 son estimados mediante las normas locales definidas por

$$\|u\|_{(k)} = \sup\{\|u(x)\|_2 \mid x \in [t_k, t_{k+1})\}. \tag{II.51}$$

Lema 12. *Supongamos que Re*$(A) \geq \lambda I$, $\lambda > 0$ *o Re*$(A) \leq \lambda I$, $\lambda < 0$. *Entonces, el error de truncamiento de la regla RT-*ω*, es dado por*

$$\left\| \tau\left(x, u, k + \frac{1}{2}\right) \right\|_2 \leq \frac{1}{12} h^2 \|u^{(3)}\|_{(k)} + \frac{1}{2} h \,\text{mín}\{1, \omega^{-1} h \|A\|_2\} \|u''\|_{(k)},$$

$k = 0, \ldots, N-1$.

Demostración. Viendo la desviación del esquema RT de la regla trapezoidal, vemos que con $Z = hA$, $X = (\omega^2 I + Z^2)^{1/2}$, $u'_k = u'(t_k)$, el error de truncamiento satisface

$$\begin{aligned} \tau_{k+1/2} &= \frac{1}{2h}((2-\omega)I + X)(u_{k+1} - u_k) - \frac{1}{2}A(u_{k+1} + u_k) - \frac{1}{2}(g_{k+1} - g_k) \\ &= \frac{1}{h}(u_{k+1} - u_k) - \frac{1}{2}(u'_{k+1} + u'_k) + \frac{1}{2}Z(\omega I + X)^{-1}(u'_{k+1} - u'_k) \\ &= -\frac{1}{12}h^2 u^{(3)}(x'_k) + \frac{1}{2}Z(\omega I + X)^{-1}(u'(x_{k+1}) - u'(x_k)). \end{aligned} \tag{II.52}$$

Los primeros términos en la última línea es el error de truncamiento de la regla trapezoidal. Por el teorema 20, $\text{Re}(X) \geq 0$ y $\|(\omega I + X)^{-1}\|_2 \leq \omega^{-1}$. Pero, por otro lado,

$$Z(\omega I + X)^{-1} = [\omega Z^{-1} + sgn(\lambda)(I + \omega^2 Z^{-2})^{1/2}]^{-1} = \psi(sgn(\lambda)\omega Z^{-1})^{-1}.$$

La norma de estas matrices son acotadas por teorema 20 b).

Así, el segundo término en el error de truncamiento puede ser estimado por

$$\frac{1}{2} \,\text{mín}\{1, \omega^{-1} h \|A\|_2\} \|u''(x''_k)\|_2, \quad x''_k \in (x_k, x_{k+1}).$$

□

Desde que la función de estabilidad de RT-2, $f_2(z) = f(z/2)^2$, es la misma como para dos pasos de RT-1 con tamaño semi-escalonado, el término del error principal $f(z/2)^2 - e^z = -\frac{z^3}{24}(z \to 0)$ es una cuarta parte de la de $f_1(z) = f(z)$. Así, el lema 12 prueba una ligera preferencia para los esquemas con los parámetros más grandes ω. En el caso rígido ($h\|A\| \gg 1$) el orden de la consistencia de los esquemas RT es solo 1. Pero por esta pérdida de precisión en la dependencia del tamaño de paso h, tenemos una compensación con respecto a la convergencia con respecto al parámetro ϵ en problemas de perturbación singulares.

Lema 13. *Sea Re*$(\lambda) \neq 0$. *En el límite* $\epsilon/h_k \to 0$, *la regla RT-*ω *aplicado a* $\epsilon u' = \lambda u + g$ *se reduce a:*

$$\lambda y_k + g_k = 0, \quad Re(\lambda) > 0$$

$$\lambda y_{k+1} + g_{k+1} = 0, \quad Re(\lambda) < 0.$$

Demostración. Cuando aplicamos a esta ecuación con $\varphi_k = \dfrac{g_k}{\delta + \sqrt{\delta^2 + \lambda^2}}$, $\delta = \dfrac{\omega\epsilon}{h_k}$ la regla toma la forma:

$$2\delta(\omega^{-1}-1)(y_{k+1}-y_k)+[\delta+\sqrt{\delta^2+\lambda^2}-\lambda](y_{k+1}-\varphi_{k+1})+[\delta+\sqrt{\delta^2+\lambda^2}+\lambda](y_k-\varphi_k) = 0.$$

Pero, para $\delta \to 0$ tenemos la estimación $\sqrt{\delta^2+\lambda^2} = \lambda + O(\delta^2)(\text{Re}(\lambda) > 0)$ y $\sqrt{\delta^2+\lambda^2} = -\lambda + O(\delta^2)(\text{Re}(\lambda) < 0)$. □

Una descripción cuantitativa de esta propiedad se da en la estimación de estabilidad del lema 14. Allí se analiza el comportamiento de los esquemas en problemas de valores iniciales no homogéneos estables, mirando desde el extremo izquierdo o derecho del intervalo. Se dan dos versiones, la primera se concentra en el amortiguamiento exponencial de las perturbaciones locales, el segundo da una descripción más precisa de la dependencia de $|\text{Re}(\lambda)|$, $\lambda \to \infty$. Las bases de estos análisis están en el siguiente lema.

Lema 14. *Sean los valores* v_k, $k = 0, \ldots, N$, *satisfacen (II.50) con los lados derechos reemplazados por* $\tau_{k+\frac{1}{2}}$, *es decir, con* $X = (\omega^2 I + h_k^2 A^2)^{1/2}$,

$$\frac{1}{2h}((2-\omega)I + X)(v_{k+1} - v_k) - \frac{1}{2}A(v_{k+1} + v_k) = \tau\left(x, v, h + \frac{1}{2}\right), \quad k = 0, 1, \ldots, N-1 \tag{II.53}$$

Sea $Re(A) \geq \lambda I$, $\lambda > 0$ *o* $Re(A) \leq \lambda I$, $\lambda < 0$ *y se define* $\mu = |\lambda|$ *(*$1 \leq \omega \leq 2$*), respectivamente,* $\mu = \dfrac{|\lambda|}{2}$ *(*$0 < \omega < 1$*),* $q = (1 + H\mu)^{1/H}$. *Entonces, las estimaciones:*

$$\|v_k\|_2 \leq 4 \sum_{j=0}^{k-1} h \left\| \tau\left(x, y, h + \frac{1}{2}\right) \right\| q^{-|t_j - t_k|} + \max\left\{ \|v_0\|_2 q^{-t_k}, \|v_N\|_2 q^{t_k - 1} \right\}, \tag{II.54}$$

$$\|v_k\|_2 \leq \frac{4}{1 + \mu/2} \max \left\| \tau\left(x, y, h + \frac{1}{2}\right) \right\| + \max\{\|v_0\|_2 q^{-t_k}, \|v_n\|_2 q^{t_k - 1}\}, \tag{II.55}$$

se mantienen para $k = 1, \ldots, N-1$.

Lema 15. *Sea la solución* u *de (II.49) que pertenece a* $C^3[0, 1]$. *Entonces, con la notación del lema 14, el error* $v_k = y_k - u(t_k)$ *del esquema (II.50) en los puntos de la malla satisfacen:*

$$\begin{aligned} \|v_k\|_2 &\leq \max\{\|v_0\|_2 q^{-t_k}, \|v_N\|_2 q^{t_k - 1}\} \\ &+ \frac{4}{2 + \mu} \max_j \left[\frac{1}{6} h^2 \|u^{(3)}\|_{(j)} + \min\{h, \omega^{-1} h^2 \|A\|_2\} \|u''\|_{(j)} \right], \end{aligned}$$

$k = 1, \ldots, N-1$.

Demostración. Ver [26]. □

Para terminar este capítulo terminaremos dando una definición atrevida de métodos en diferencia finita para ecuaciones diferenciales racionales e irracionales.

Definición 22. *Un método* Φ *o un método de un paso es llamado racional si ella proviene una función de estabilidad* ψ *racional y es llamada irracional si la función de estabilidad es una función irracional.*

Para iluminar las propiedades de los esquemas de raíz cuadrada, realizaremos algunos experimentos el capítulo H de los resultados numéricos.

IV. Resultados numéricos

En este capítulo, se explorarán las simulaciones prácticas del cálculo de la raíz cuadrada de matrices definidas positivas, así como la aplicación del método irracional a problemas rígidos. La raíz cuadrada de una matriz definida positiva es un tema de gran relevancia en álgebra lineal y tiene aplicaciones en diversas áreas, desde la optimización hasta la teoría de control y la estadística.

Iniciaremos con las simulaciones de la raíz cuadrada de matrices definidas positivas, donde se presentarán diferentes métodos numéricos en acción. Estas simulaciones no solo permitirán observar la efectividad de los algoritmos propuestos, sino que también proporcionarán una perspectiva visual sobre el comportamiento de las soluciones en diferentes contextos. Analizaremos los resultados obtenidos y discutiremos la influencia de diversos parámetros en la precisión y eficiencia de los métodos utilizados.

A continuación, nos centraremos en la aplicación del método irracional a problemas rígidos. Este enfoque es especialmente interesante debido a la complejidad y los desafíos que presentan las ecuaciones diferenciales rígidas. Discutiremos cómo el método irracional puede ser adaptado y optimizado para resolver estos problemas, analizando sus ventajas y limitaciones en comparación con otros métodos existentes.

Este capítulo no solo proporcionará una comprensión profunda de las técnicas de simu-

lación y su relevancia en el cálculo de la raíz cuadrada de matrices, sino que también permitirá al lector apreciar la versatilidad y aplicabilidad de estos métodos en contextos más amplios. Al final del capítulo, se espera que el lector haya adquirido herramientas valiosas para implementar y evaluar métodos numéricos en problemas prácticos, sentando las bases para futuras investigaciones en el área.

A. Simulaciones de la raíz cuadrada de una matriz definida positiva

En este capítulo presentaremos algunos resultados numéricos del cálculo de la raíz cuadrada de una matriz definida positiva de los métodos desarrollados en el capítulo H. También desarrollaremos algunas aplicaciones a problemas rígidos con los método A-estables, los cuales contienen una función de estabilidad de forma racional, así como el método de la regla de la raíz trapezoidal, estudiadas en los capítulos I y H. Los ejemplos numéricos fueron realizados con el software MATLAB R2023a en una computadora personal Intel® Core™ i5, con una precisión de máquina de $2{,}2204 \times 10^{-16}$.

El criterio de parada para el desarrollo de los algoritmos es dado por el error residual relativo:

$$\text{Res} = \frac{\|X_k^2 - A\|_2}{\|A\|_2} < 10^{-6}, \tag{III.1}$$

donde X_k es el valor actual (k-ésima iteración) de la raíz cuadrada de la matriz A.

El error relativo

$$\text{Error} = \frac{\|X - X_k\|_2}{\|X\|_2}$$

es determinado con el valor de la raíz cuadrada X de la matriz A calculado por Matlab. A continuación, presentamos los siguientes ejemplos.

Ejemplo 5. *Consideremos la matriz*

$$A = \begin{bmatrix} 10 & 7 & 8 & 7 \\ 7 & 5 & 6 & 5 \\ 8 & 6 & 10 & 9 \\ 7 & 5 & 9 & 10 \end{bmatrix}$$

Para hallar la raíz cuadrada de la matriz A usamos los algoritmos 1, 3, 4, 5, 6, 7 y 8. En el caso de los algoritmos 7 y 8 usamos los parámetros $p = 1, 2, 3$ y el valor inicial tomado fue $X_0 = A$, donde A es una matriz simétrica definida positiva, con condicionamiento $k(A) = 2{,}9841 \times 10^3$.

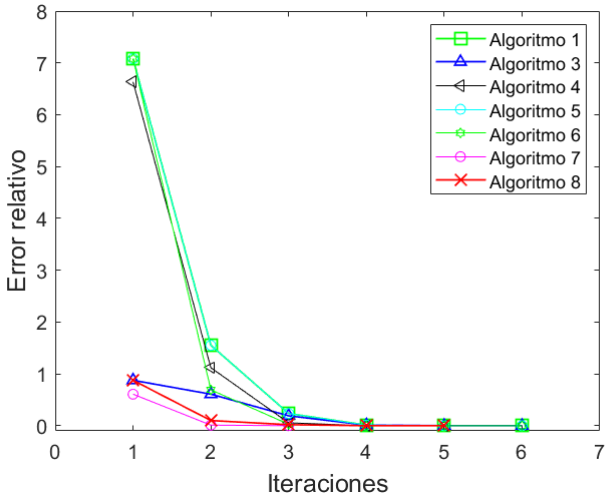

Figura VI: Comportamiento de la convergencia de los algoritmos para el ejemplo 5.

En la figura VI se observa que todos los métodos proporcionan excelentes resultados, a pesar que la matriz no está bien condicionada. Esto es debido a que la matriz es simétrica definida positiva y por el corolario 1 posee una única raíz cuadrada simétrica definida

positiva. Casi todos los métodos presentan un comportamiento similar en las primeras iteraciones, pero la convergencia se logra desde la tercera iteración. Como se puede apreciar en la tabla III, el método de la escala de Padé converge más rápido.

Tabla III: Algoritmos de prueba para el ejemplo 5, donde n representa el número de iteraciones, EAP es el error aproximado y ERR es el error residual relativo.

	n	EAP	ERR
Algoritmo 1	7	$7{,}7153 \times 10^{-15}$	$2{,}1092 \times 10^{-13}$
Algoritmo 3	7	$3{,}2668 \times 10^{-9}$	$8{,}7912 \times 10^{-8}$
Algoritmo 4	7	$7{,}7030 \times 10^{-15}$	$2{,}1102 \times 10^{-13}$
Algoritmo 5	6	$1{,}8969 \times 10^{-15}$	$5{,}1432 \times 10^{-14}$
Algoritmo 6	7	$2{,}3676 \times 10^{-9}$	$2{,}3251 \times 10^{-9}$
Algoritmo 7 con $p = 1$	7	$3{,}2186 \times 10^{-9}$	$8{,}7910 \times 10^{-8}$
Algoritmo 7 con $p = 2$	4	$3{,}2186 \times 10^{-9}$	$8{,}7910 \times 10^{-8}$
Algoritmo 7 con $p = 3$	4	$5{,}3132 \times 10^{-16}$	$6{,}8278 \times 10^{-16}$
Algoritmo 8 con $p = 1$	6	$6{,}9025 \times 10^{-11}$	$3{,}4512 \times 10^{-11}$
Algoritmo 8 con $p = 2$	4	$9{,}4610 \times 10^{-11}$	$4{,}7305 \times 10^{-11}$
Algoritmo 8 con $p = 3$	3	$6{,}1033 \times 10^{-17}$	$2{,}2891 \times 10^{-16}$

Ejemplo 6. *Consideremos la matriz*

$$A = (a_{ij})_{100\times 100} = \begin{cases} \dfrac{j}{20} & , \quad i = j; \\ \dfrac{i+j}{1000} & , \quad i \neq j. \end{cases} \tag{III.2}$$

Para hallar la raíz cuadrada de la matriz A usamos los algoritmos 1, 3, 4, 5, 6, 7 y 8. En

el caso de los algoritmos 7 y 8 usamos los parámetros $p = 1, 2, 3$ *y el valor inicial tomado fue* $X_0 = A$*, donde* A *es una matriz simétrica definida positiva, con condicionamiento* $k(A) = 10{,}7658$.

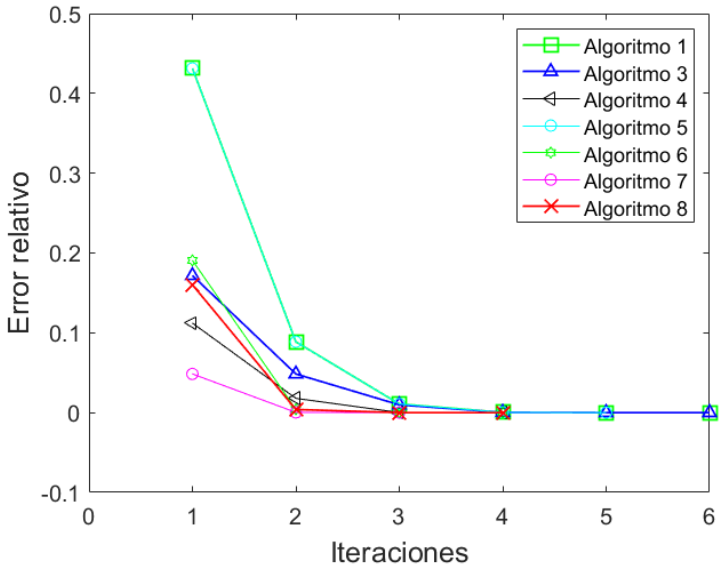

Figura VII: Comportamiento de la convergencia de los algoritmos para el ejemplo 6.

En la figura VII *se observa que todos los métodos proporcionan excelentes resultados, debido a que la matriz posee un buen condicionamiento. Todos los métodos presentan un comportamiento similar en las primeras iteraciones, pero convergen desde la segunda y/ o tercera iteración. Como se puede apreciar en la tabla IV el método de Newton modificado 2 y el método de la escala de Padé convergen más rápido.*

Tabla IV: Algoritmos de prueba para el ejemplo 6, donde n representa el número de iteraciones, EAP es el error aproximado y ERR es el error residual relativo.

	n	EAP	ERR
Algoritmo 1	7	$2{,}1186 \times 10^{-16}$	$1{,}1665 \times 10^{-15}$
Algoritmo 3	7	$1{,}2442 \times 10^{-13}$	$2{,}0413 \times 10^{-13}$
Algoritmo 4	7	$2{,}1206 \times 10^{-16}$	$1{,}1865 \times 10^{-15}$
Algoritmo 5	5	$1{,}9176 \times 10^{-11}$	$3{,}1460 \times 10^{-11}$
Algoritmo 6	5	$2{,}2076 \times 10^{-9}$	$3{,}6217 \times 10^{-9}$
Algoritmo 7 con $p = 1$	6	$2{,}1500 \times 10^{-7}$	$3{,}5271 \times 10^{-7}$
Algoritmo 7 con $p = 2$	4	$1{,}2433 \times 10^{-13}$	$2{,}0421 \times 10^{-13}$
Algoritmo 7 con $p = 3$	3	$3{,}5705 \times 10^{-8}$	$5{,}8576 \times 10^{-8}$
Algoritmo 8 con $p = 1$	5	$2{,}2076 \times 10^{-9}$	$3{,}6217 \times 10^{-9}$
Algoritmo 8 con $p = 2$	3	$2{,}0785 \times 10^{-8}$	$3{,}4106 \times 10^{-8}$
Algoritmo 8 con $p = 3$	3	$3{,}2266 \times 10^{-16}$	$1{,}0997 \times 10^{-15}$

Ejemplo 7. *Consideremos la matriz*

$$A = (a_{ij})_{200\times 200} = \begin{cases} 1 & , \quad i = j; \\ \dfrac{1}{i+j-1} & , \quad i \neq j. \end{cases} \tag{III.3}$$

Para hallar la raíz cuadrada de la matriz A usamos los algoritmos 1, 3, 4, 5, 6, 7 y 8. En el caso de los algoritmos 7 y 8 usamos los parámetros $p = 1, 2, 3$ y el valor inicial tomado fue $X_0 = A$, donde A es una matriz simétrica definida positiva, con condicionamiento $k(A) = 6{,}2304$.

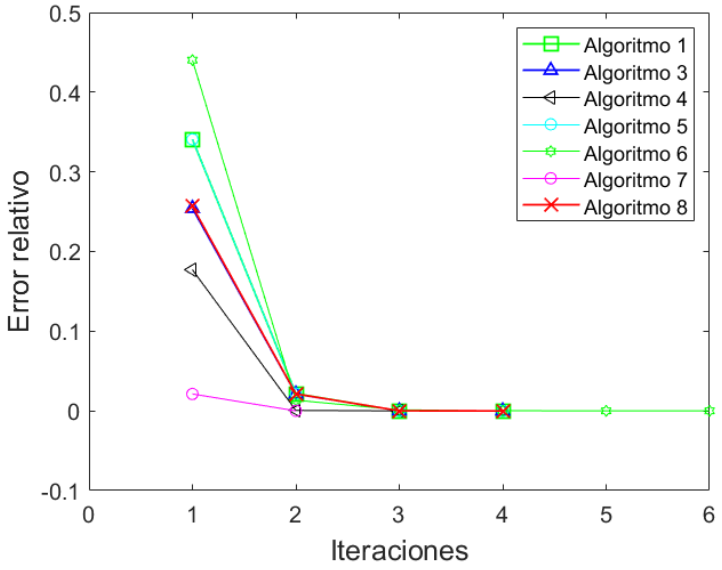

Figura VIII: Comportamiento de la convergencia de los algoritmos para el ejemplo 7.

En la figura VIII se observa que todos los métodos proporcionan excelentes resultados, debido a que la matriz posee un buen condicionamiento. Todos los métodos presentan un comportamiento similar en las primeras iteraciones, pero convergen desde la segunda y tercera iteración. Como se puede apreciar en la tabla V el método de Newton modificado 2 y el método de Padé convergen más rápido.

Tabla V: Algoritmos de prueba para el ejemplo 7, donde n representa el número de iteraciones, EAP es el error aproximado y ERR es el error residual relativo.

	n	EAP	ERR
Algoritmo 1	5	$6{,}2554 \times 10^{-16}$	$3{,}0899 \times 10^{-15}$
Algoritmo 3	5	$3{,}2883 \times 10^{-9}$	$1{,}6442 \times 10^{-9}$
Algoritmo 4	5	$5{,}5006 \times 10^{-16}$	$3{,}1129 \times 10^{-15}$
Algoritmo 5	4	$2{,}0918 \times 10^{-11}$	$1{,}0460 \times 10^{-11}$
Algoritmo 6	7	$2{,}1447 \times 10^{-9}$	$1{,}0724 \times 10^{-9}$
Algoritmo 7 con $p = 1$	5	$3{,}2883 \times 10^{-9}$	$1{,}6442 \times 10^{-9}$
Algoritmo 7 con $p = 2$	3	$3{,}2883 \times 10^{-9}$	$1{,}6442 \times 10^{-9}$
Algoritmo 7 con $p = 3$	3	$1{,}3288 \times 10^{-15}$	$3{,}0917 \times 10^{-15}$
Algoritmo 8 con $p = 1$	5	$3{,}2470 \times 10^{-9}$	$1{,}6225 \times 10^{-9}$
Algoritmo 8 con $p = 2$	3	$3{,}5194 \times 10^{-9}$	$1{,}7197 \times 10^{-9}$
Algoritmo 8 con $p = 3$	3	$1{,}3672 \times 10^{-13}$	$3{,}0858 \times 10^{-15}$

El siguiente ejemplo muestra el cálculo de la raíz cuadrada de una matriz mal condicionada.

Ejemplo 8. *Consideremos la matriz de Hibert, la cual es simétrica definida positiva:*

$$H_n = (h_{ij}) = \left(\frac{1}{i+j-1}\right)_n$$

Para hallar la raíz cuadrada de esta matriz, para el caso cuando $n = 10$, usamos los algoritmos 1, 3, 4, 5, 6, 7 y 8. En el caso de los algoritmos 7 y 8 usamos los parámetros $p = 1, 2, 3$ y el valor inicial tomado fue $X_0 = H_{10}$, donde H_{10} es una matriz simétrica

definida positiva, con condicionamiento $k(H_{10}) = 1{,}6025 \times 10^{13}$.

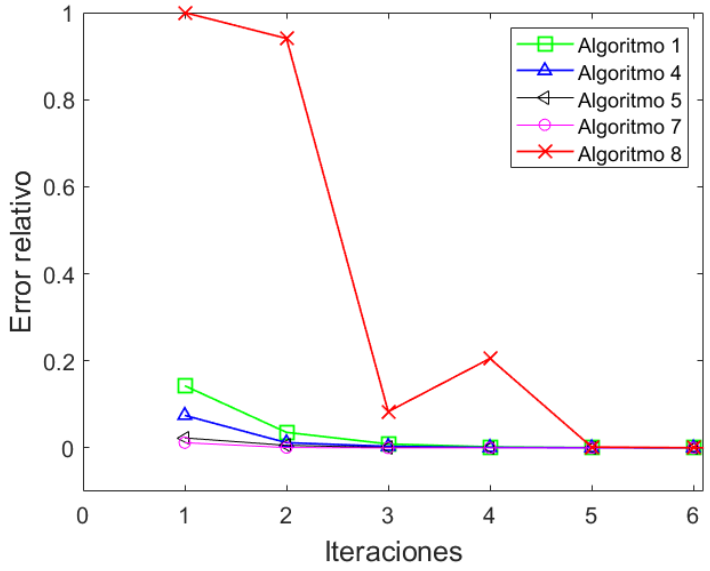

Figura IX: Comportamiento de la convergencia de los algoritmos para el ejemplo 8.

En la figura IX se observa que no todos los métodos proporcionan excelentes resultados, esto se debe a que la matriz de Hilbert no está bien condicionada. El método de escala de Padé al inicio de las iteraciones muestra un comportamiento inestable, pero luego en la iteración 5 converge como los algoritmos 1, 4, 5 y 7, mientras los algoritmos 3 y 6 no convergen a la raíz cuadrada debido a la inestabilidad de la matriz de Hilbert. Finalmente, podemos observar en la tabla VI que los algoritmos 1, 4, 5 y 7 son numéricamente estables y poseen una convergencia del tipo cuadrática.

Tabla VI: Algoritmos de prueba para el ejemplo 8, donde n representa el número de iteraciones, EAP es el error aproximado y ERR es el error residual relativo.

	n	EAP	ERR
Algoritmo 1	13	$8{,}5029 \times 10^{-9}$	$9{,}1962 \times 10^{-5}$
Algoritmo 3	–	–	–
Algoritmo 4	12	$1{,}4690 \times 10^{-8}$	$2{,}8444 \times 10^{-5}$
Algoritmo 5	11	$2{,}3382 \times 10^{-9}$	$2{,}0403 \times 10^{-5}$
Algoritmo 6	–	–	–
Algoritmo 7 con $p = 1$	14	$1{,}0648 \times 10^{-9}$	$2{,}2324 \times 10^{-5}$
Algoritmo 7 con $p = 2$	8	$7{,}2270 \times 10^{-10}$	$1{,}2582 \times 10^{-5}$
Algoritmo 7 con $p = 3$	6	$1{,}1929 \times 10^{-9}$	$2{,}2914 \times 10^{-5}$
Algoritmo 8 con $p = 1$	8	$6{,}3775 \times 10^{-7}$	$3{,}1887 \times 10^{-7}$
Algoritmo 8 con $p = 2$	6	$1{,}0012 \times 10^{-10}$	$5{,}0431 \times 10^{-11}$
Algoritmo 8 con $p = 3$	5	$5{,}6217 \times 10^{-10}$	$2{,}8115 \times 10^{-10}$

B. Aplicación del método irracional a problemas rígidos

En esta sección presentamos algunos resultados numéricos para poner a prueba la estabilidad del método irracional RT-ω (con $\omega = 2$), para problemas rígidos, los cuales como ya vimos en el capítulo H son bastantes inestables en el sentido que si tuviéramos una malla con pocos puntos, dan resultados muy caóticos respecto a la solución verdadera. Para tal fin, pondremos a prueba la convergencia de las soluciones de seis problemas

rígidos tomados en la literatura, que son de tipo oscilatorios y altamente oscilatorios, con algunos métodos numéricos implícitos tradicionales cuyas funciones de estabilidad son racionales estudiadas en el capítulo I. Para el método irracional se aplicarán los métodos que tuvieron mejores resultados en la obtención de la raíz cuadrada, desarrolladas en la sección anterior.

Ejemplo 9. *Consideremos el problema que fue tratado en el capítulo I para estudiar su rigidez:*

$$\begin{bmatrix} u_1 \\ u_2 \end{bmatrix}' = \begin{bmatrix} 9 & 24 \\ -24 & -51 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} + \begin{bmatrix} 5\cos(t) - \frac{1}{3}\operatorname{sen}(t) \\ -9\cos(t) + \frac{1}{3}\operatorname{sen}(t) \end{bmatrix}$$

con condiciones iniciales $u_1(0) = \frac{4}{3}, u_2(0) = \frac{2}{3}$, donde el jacobiano tiene valores propios reales negativos $\lambda_1 = -3$ y $\lambda_2 = -39$ y el cual tiene solución única:

$$u_1(t) = 2e^{-3t} - e^{-39t} + \frac{1}{3}\cos(t), \;\; u_2(t) = -e^{-3t} + 2e^{-39t} - \frac{1}{3}\cos(t).$$

Se puede observar el término e^{-39t} en la solución causa que este problema sea rígido. Para mostrar los efectos de rigidez, realizaremos la prueba numérica con los métodos, Runge Kutta-4 explícito, Euler explicíto, trapezoidal, Runge Kutta-4 implícito y RT-ω, utilizando una longitud de paso constante para $N = 10, 30$ y 50.

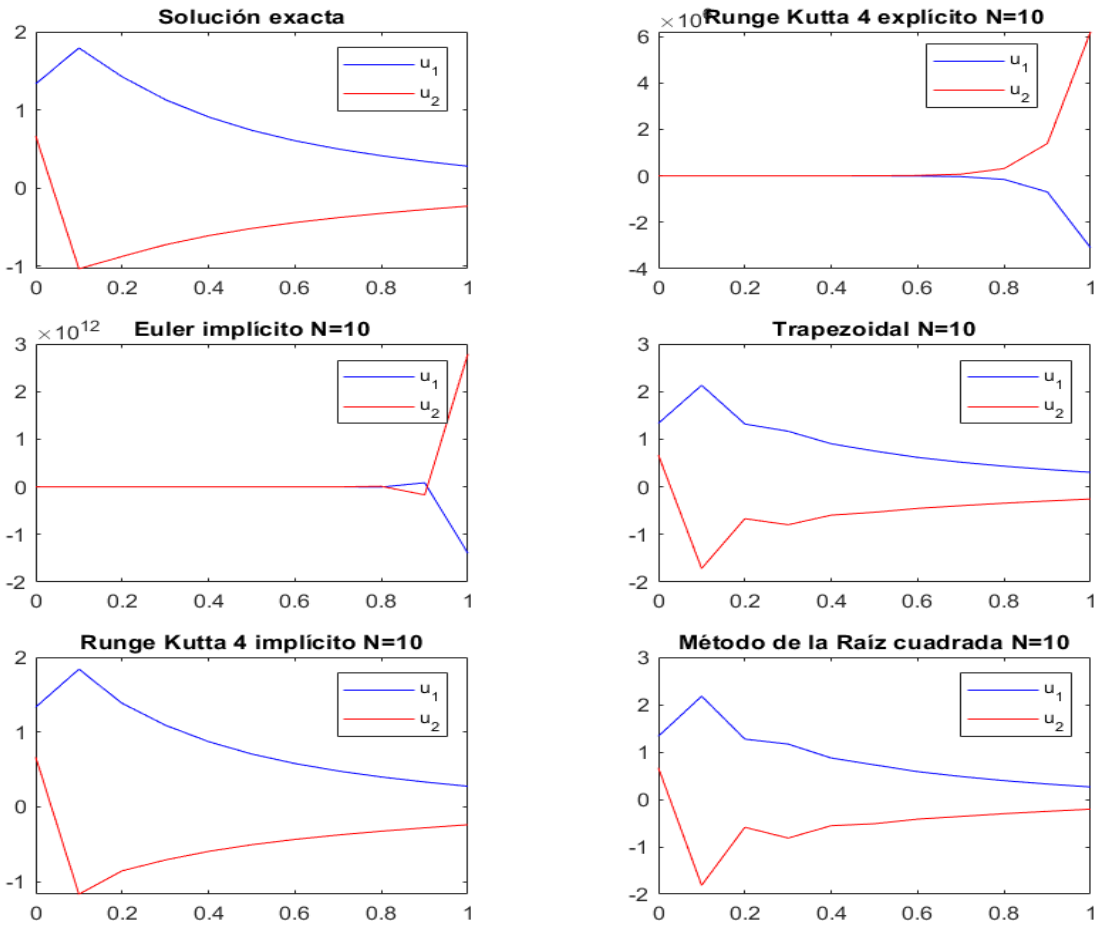

Figura X: Comparación de los métodos RK-4 explícito, Euler implícito, trapezoidal, RK-4 implícito, RT-ω y la solución exacta para $N = 10$.

Como se observa en la figura X los métodos de Runge Kutta-4 explícito y Euler implícito muestran una notable inestabilidad para $N = 10$. Por otro lado, los métodos del trapecio y RT-ω se acercan más a la solución exacta debido a que su región de estabilidad incluye la parte real negativa. Además, el método de Runge Kutta-4 implícito ofrece una aproximación más precisa a la solución exacta gracias a su rápida tasa de convergencia.

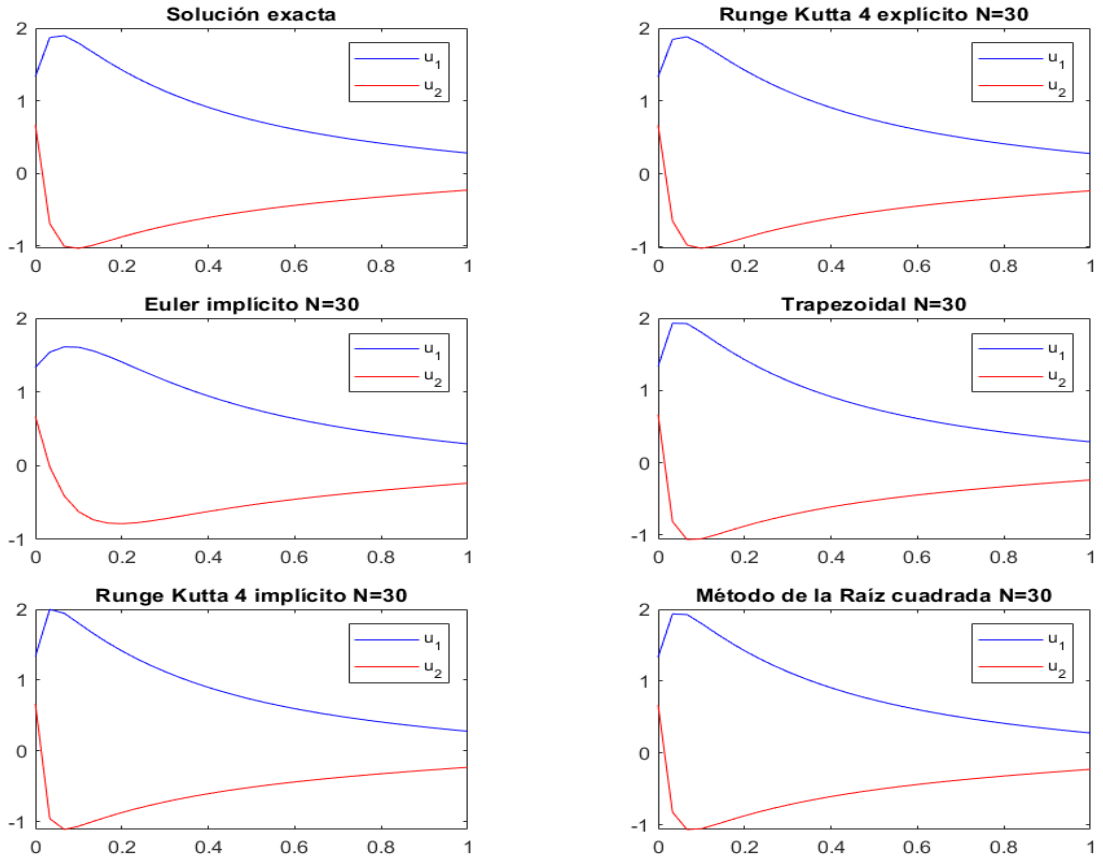

Figura XI: Comparación de los métodos RK-4 explícito, Euler implícito, trapezoidal, RK-4 implícito, RT-ω y la solución exacta para $N = 30$.

Ahora, al considerar $N = 30$, se puede observar en la figura XI que el método de Euler implícito no muestra una aproximación cercana a la solución exacta debido a su falta de estabilidad en su región correspondiente. Por el contrario, los otros métodos muestran una convergencia aparente hacia la solución exacta. Para obtener más detalles sobre el estudio de la convergencia de estos métodos, se analizan los puntos en $t = 0{,}03$ y $0{,}1$, los cuales se presentan en la tabla VII.

Tabla VII: Comparación de resultados, donde RK-4 E: Runge Kutta-4 explícito, EI: Euler implícito, T: trapezoidal, RK-4 I: Runge Kutta-4 implícito y RC: RT-ω para $N = 30$.

	Exacta	RK-4 E	EI	T	RK-4 I	RC
$u_1(0{,}03)$	1,870291	1,844983	1,539188	1,930551	1,998748	1,935617
$u_1(0{,}03)$	−0,692922	−0,642305	−0,016864	−0,813668	−0,954939	−0,824037
$u_2(0{,}1)$	1,793063	1,786881	1,605822	1,804085	1,802393	1,803567
$u_2(0{,}1)$	−1,032002	−1,019638	−0,623372	−1,053997	−1,063091	−1,054154

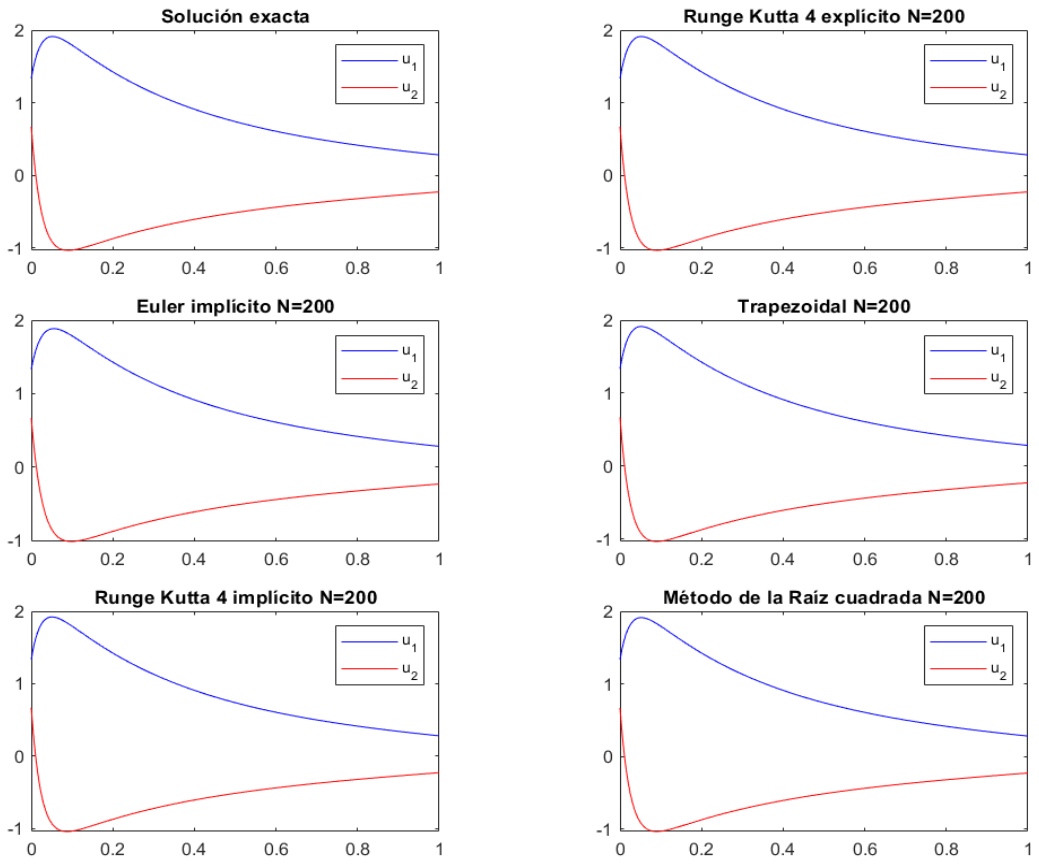

Figura XII: Comparación de los métodos RK-4 explícito, Euler implícito, trapezoidal, RK-4 implícito, RT-ω y la solución exacta para $N = 200$.

Finalmente, al considerar $N = 200$ se puede observar en la figura XII que el método de Euler implícito muestra una mejora en la aproximación a la solución exacta en comparación con el caso cuando $N = 30$. Por otro lado, los métodos de Runge Kutta-4, trapezoidal y RT-ω se acercan aún más a la solución exacta, mostrando una convergencia efectiva hacia ella. Para obtener más detalles sobre el estudio de la convergencia de estos métodos, se analizan los resultados en los puntos $t = 0{,}05$ y $0{,}08$, que se presentan en la tabla VIII.

Tabla VIII: Comparación de resultados,donde RK-4 E: Runge Kutta-4 explícito, EI: Euler implícito, T: trapezoidal, RK-4 I: Runge Kutta-4 implícito y RC: RT-ω para $N = 200$.

	Exacta	RK-4 E	EI	T	RK-4 I	RC
$u_1(0{,}05)$	1,912058	1,912054	1,882286	1,913008	1,921886	1,912996
$u_2(0{,}05)$	$-0{,}909077$	$-0{,}909068$	$-0{,}846651$	$-0{,}910912$	$-0{,}929791$	$-0{,}910967$
$u_1(0,08)$	1,861366	1,861364	1,847409	1,861920	1,865463	1,861820
$u_2(0,08)$	$-1{,}030581$	$-1{,}030577$	$-0{,}998455$	$-1{,}031577$	$-1{,}040317$	$-1{,}031510$

De estos resultados podemos inferir que el problema no exhibe oscilaciones significativas, lo cual se debe al tamaño moderado del N seleccionado. Sin embargo, el método de Euler implícito muestra resultados menos favorables debido a su región de estabilidad limitada.

Ejemplo 10. *El siguiente problema, posee dos valores propios reales negativos donde uno de ellos es considerado negativamente pequeño en comparación del otro valor propio.*

Consideremos el problema rígido de valor inicial:

$$\begin{bmatrix} u_1 \\ u_2 \end{bmatrix}' = \begin{bmatrix} -298 & 99 \\ -594 & 197 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}$$

con condiciones iniciales, $u_1(0) = -0{,}5$ *y* $u_2(0) = 0{,}5$*, donde el jacobiano tiene valores propios reales* $\lambda_1 = -100$ *y* $\lambda_2 = -1$ *y cuya solución exacta es dada por*

$$u_1(t) = 1{,}5e^{-t} - 2e^{-100t}, u_2(t) = 4{,}5e^{-t} - 4e^{-100t}.$$

Se puede observar el término e^{-100t} *en la solución causa que este problema sea rígido. Para mostrar los efectos de rigidez, realizaremos la prueba numérica con los métodos, Runge Kutta-4 explícito, trapezoidal, Runge Kutta-4 implícito y RT-*ω*, utilizando la longitud de paso constante para* $N = 60, 90$ *y* 400*.*

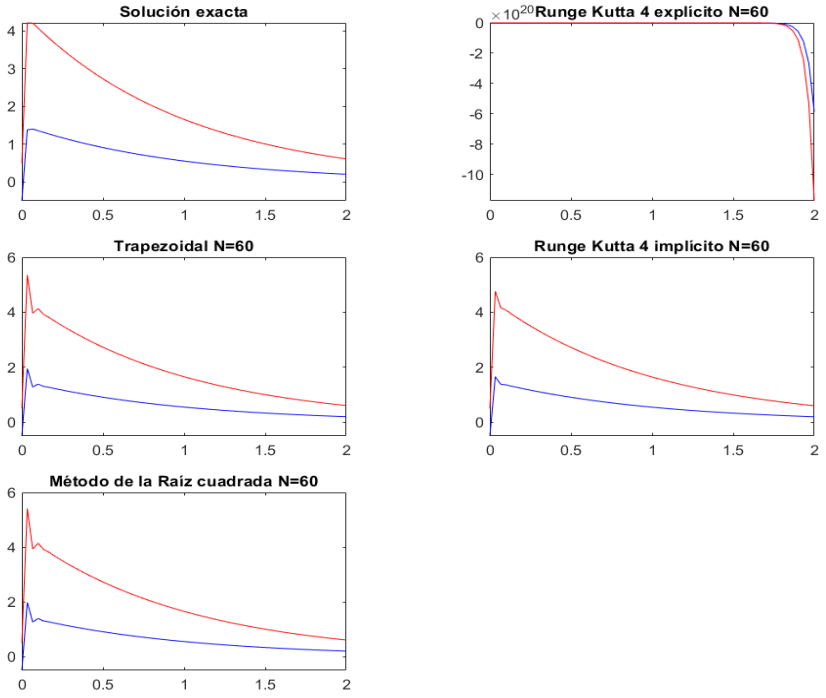

Figura XIII: Comparación de los métodos RK-4 explícito, trapezoidal, RK-4 implícito, RT-ω y la solución exacta para $N = 60$.

Como se puede ver en la figura XIII, los métodos de Runge Kutta-4 explícito muestran una divergencia respecto a la solución exacta debido a su escasa estabilidad con respecto al tamaño del paso. En cambio, los métodos trapezoidal y RT-ω se aproximan a la solución exacta, aunque no convergen debido a su orden de convergencia. Se espera que el método de Runge Kutta-4 implícito, a pesar de su orden de convergencia, logre converger hacia la solución exacta.

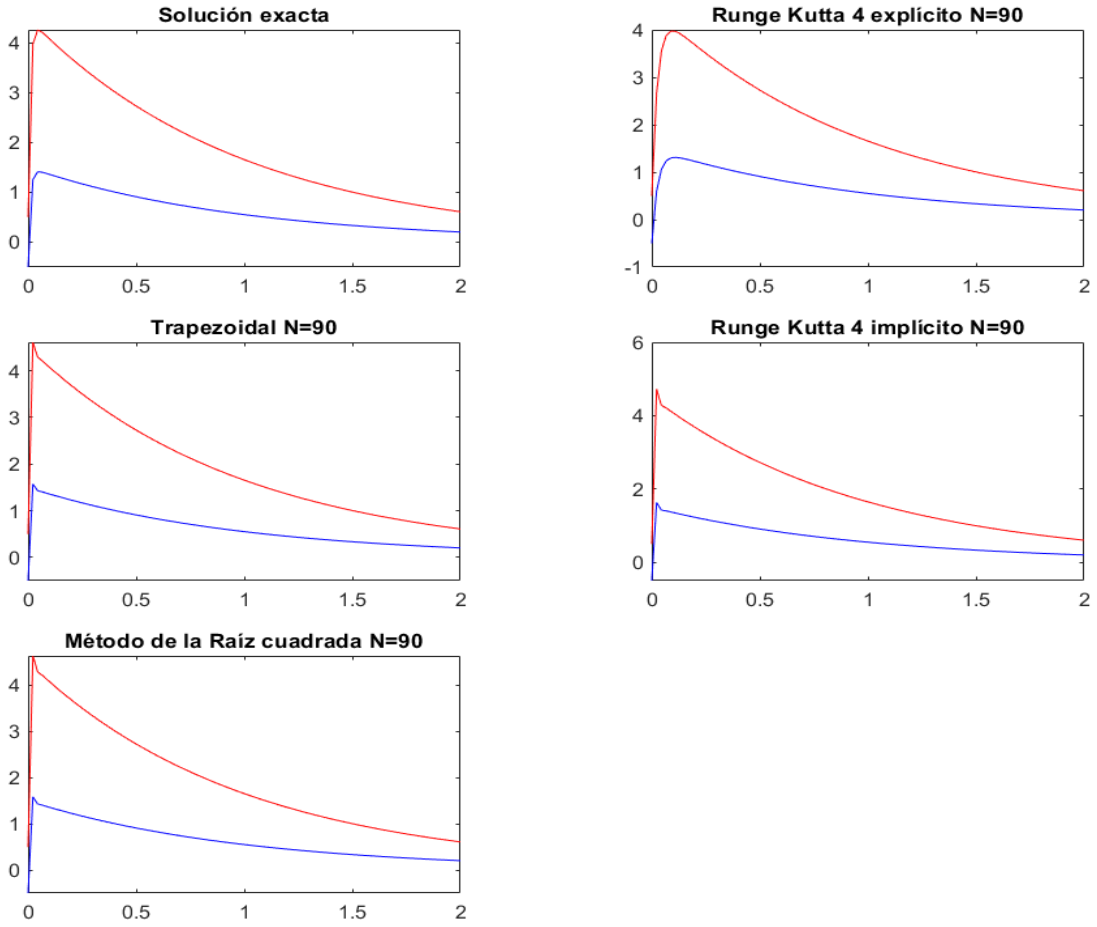

Figura XIV: Comparación de los métodos RK-4 explícito, trapezoidal RK-4 implícito, RT-ω y la solución exacta para $N = 90$.

Para $N = 90$, se observa en la figura XIV que los métodos de Runge Kutta-4 implícito, trapezoidal y RT-ω se acercan a la solución exacta. Debido a su orden de convergencia,

se espera que el método de Runge Kutta-4 implícito converja más rápidamente, mientras que el método de Runge Kutta-4 explícito muestra una mejora en comparación con los otros métodos. Para examinar el estudio de la convergencia de estos métodos, se analizan los resultados en los puntos $t = 0{,}03$ *y* $0{,}1$ *se presentan en la tabla IX.*

Tabla IX: Comparación de resultados, donde RK-4 E: Runge Kutta-4 explícito, T: trapezoidal, RK-4 I: Runge Kutta-4 implícito y RC: RT-ω para $N = 90$.

	Exacta	RK-4 E	T	RK-4 I	RC
$u_1(0{,}044)$	1,41130585	1,05803013	1,42925032	1,42134357	1,42784363
$u_2(0,044)$.	4,25740481	3,55085337	4,29329112	4,27721158	4,29046986
$u_1(0,288)$	1,12364315	1,12360436	1,12362979	1,12227618	1,12358970
$u_2(0,288)$.	3,37092945	3,37085186	3,37088937	3,36682854	3,37076911

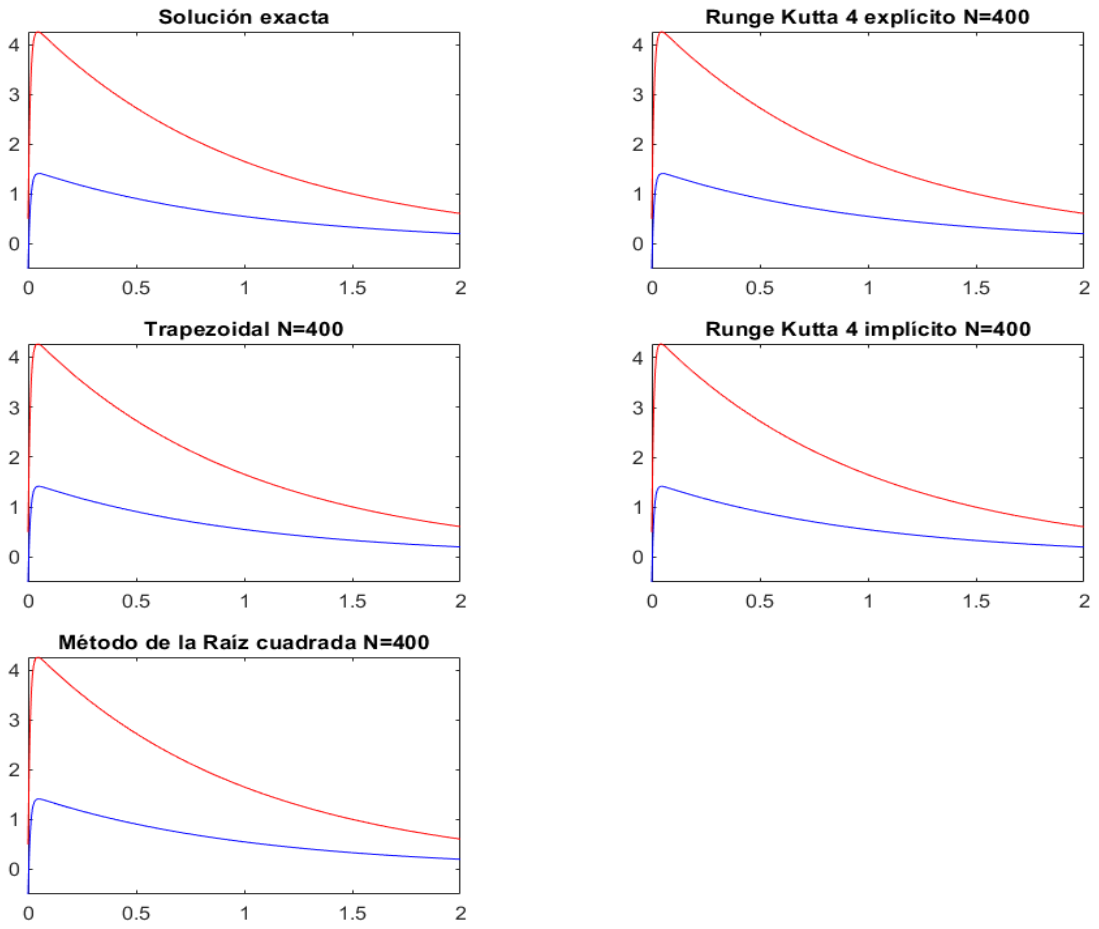

Figura XV: Comparación de los métodos RK-4 explícito, trapezoidal, RK-4 implícito, RT-ω y la solución exacta para $N = 400$.

Para $N = 400$, se puede observar en la figura XV que todos los métodos alcanzan la convergencia hacia la solución exacta. Para obtener más detalles sobre el estudio de la convergencia de estos métodos, se analizan los resultados en los puntos donde $t = 0{,}115$ y $0{,}05$, los cuales se presentan detalladamente en la tabla X.

Tabla X: Comparación de resultados, donde RK-4 E: Runge Kutta-4 explícito, T: trapezoidal, RK-4 I: Runge Kutta-4 implícito y RC: RT-ω para $N = 400$.

	Exacta	RK-4 E	T	RK-4 I	RC
$u_1(0{,}115)$	1,402594	1,402579	1,402989	1,404090	1,402999
$u_2(0{,}115)$	4,210789	4,210759	4,211579	4,213694	4,2115996
$u_1(0{,}05)$	1,4133682	1,4133148	1,4147508	1,4191870	1,4147906
$u_2(0{,}05)$	4,2535807	4,2534737	4,2563454	4,2651509	4,2564247

De estos resultados se puede concluir que, dado que el problema exhibe una alta oscilación, los métodos desarrollados muestran un buen desempeño. Tanto el método trapezoidal como el RT-ω exhiben resultados similares debido a sus regiones de estabilidad y órdenes de convergencia. Por otro lado, el método de Runge-Kutta converge más rápidamente, como se esperaba según la parte teórica del estudio.

Ejemplo 11. *La siguiente ecuación diferencial rígida es un ecuación altamente oscilatorio, llamado B5 tomado de la literatura [4]. Esta ecuación es un problema muy inestable para muchos métodos clásicos, en especial cuando se toman particiones pequeñas, y sus resultados están íntimamente ligados a las propiedades de amortiguamiento. Por ejemplo, tomando para un sistema de seis ecuaciones diferenciales dado por:*

$$\begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \\ u_5 \\ u_6 \end{bmatrix}' = \begin{bmatrix} -10 & w & 0 & 0 & 0 & 0 \\ -w & -10 & 0 & 0 & 0 & 0 \\ 0 & 0 & -4 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & -0{,}5 & 0 \\ 0 & 0 & 0 & 0 & 0 & -0{,}1 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \\ u_5 \\ u_6 \end{bmatrix} + \begin{bmatrix} \epsilon(\cos(t) + 10\,\mathrm{sen}(t)) \\ w\epsilon\,\mathrm{sen}(t) \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

donde se toman los valores de $w = 100$ *y* $\epsilon = 10^{-3}$*, y con valores iniciales* $u_1(0) = 1$*,* $u_2(0) = 1$*,* $u_3(0) = 1$*,* $u_4(0) = 1$*,* $u_5(0) = 1$ *y* $u_6(0) = 1$*, donde el jacobiano tiene valores propios puramente complejos con parte real negativa y otros valores propios no reales negativos* $\lambda_1 = -10 + 50i$*,* $\lambda_2 = -10 - 50i$*,* $\lambda_3 = -4$*,* $\lambda_4 = -1$*,* $\lambda_5 = -0{,}5$ *y* $\lambda_6 = -0{,}1$ *y cuya solución exacta es dada por:*

$$u(t) = \begin{bmatrix} e^{-10t}(\cos(wt) + \mathrm{sen}(wt)) + \epsilon\,\mathrm{sen}(t) \\ e^{-10t}(\cos(wt) - \mathrm{sen}(wt)) \\ e^{-4t} \\ e^{-t} \\ e^{-t/2} \\ e^{-t/10} \end{bmatrix}.$$

Para ilustrar los efectos de rigidez, llevaremos a cabo un análisis numérico utilizando los métodos trapezoidal, Runge Kutta-4 implícito y RT-ω, empleando un tamaño de paso constante para $N = 200, 1000$ y 50000 para $t \in [0, 15]$.

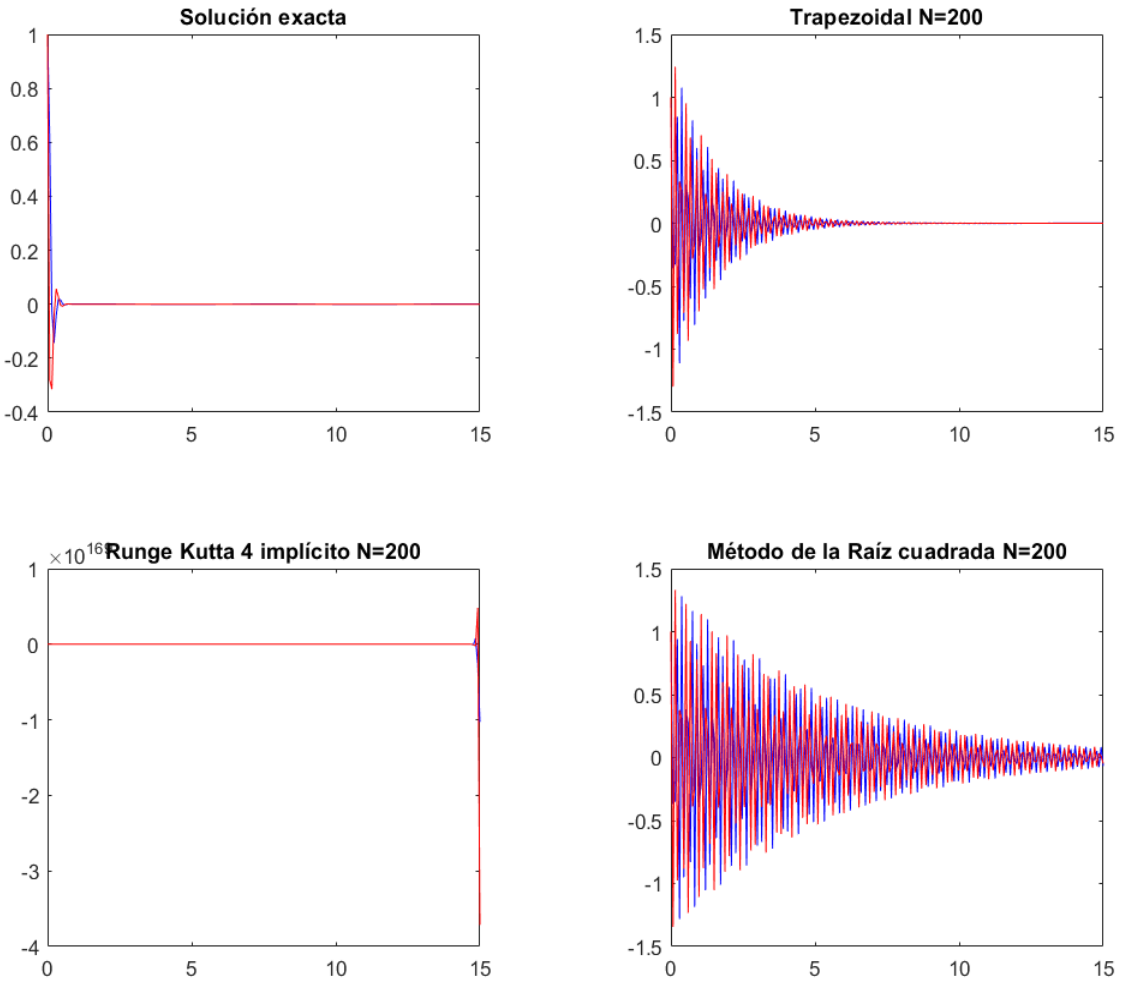

Figura XVI: Comparación de los métodos trapezoidal, RK-4 implícito, RT-ω y la solución exacta para $N = 200$.

Como se puede observar en la figura XVI, el método Runge Kutta-4 implícito mostró una notable inestabilidad con una partición constante de $N = 200$. En contraste, los métodos trapezoidal y RT-ω no se acercaron a la solución exacta, pero no exhibieron un efecto de inestabilidad significativo.

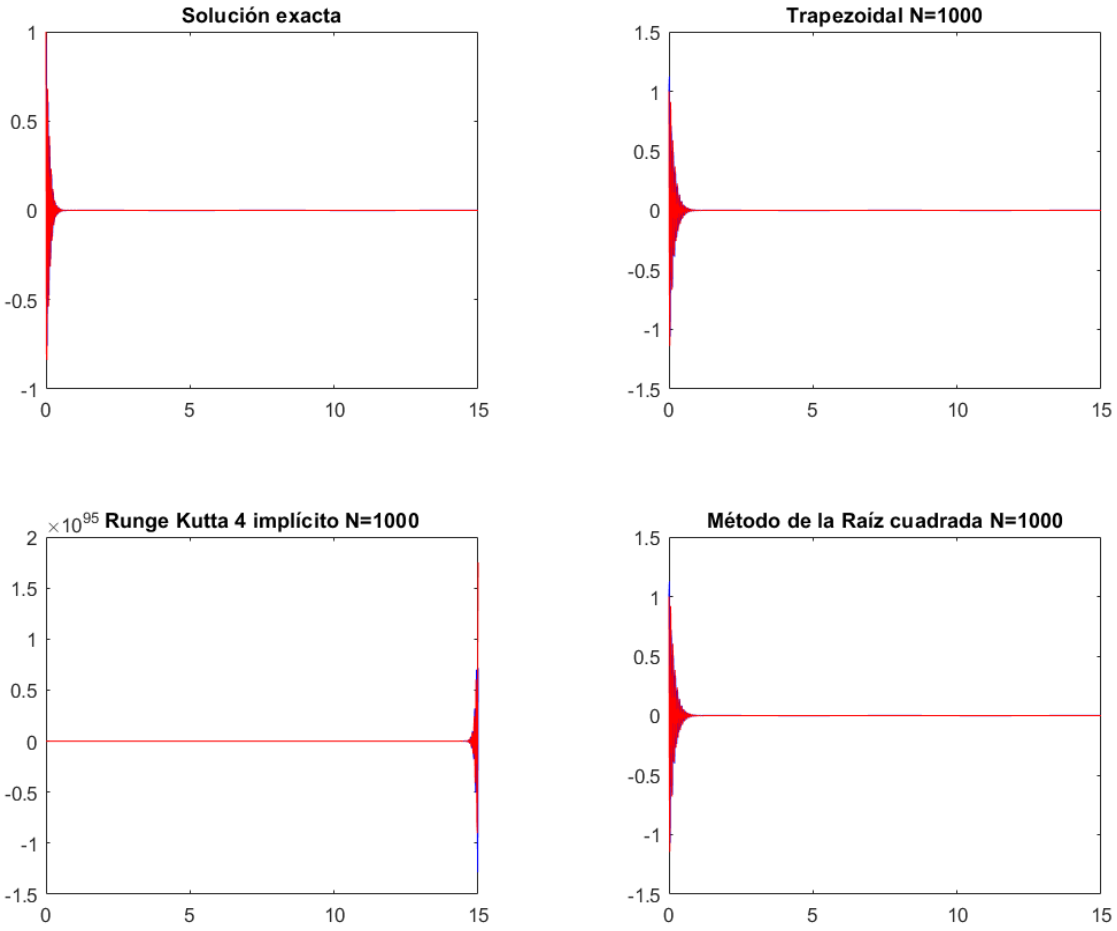

Figura XVII: Comparación de los métodos trapezoidal, RK-4 implícito, RT-ω la solución exacta para $N = 1000$.

Para $N = 1000$ se observa en la figura XVII que el método de Runge Kutta-4 implícito nuevamente muestra resultados poco precisos, con una discrepancia del orden de 10^{95}, debido a las altas oscilaciones presentes en la ecuación y a la escasa estabilidad del método. En contraste, los métodos trapezoidal y RT-ω exhiben un comportamiento estable frente a las grandes oscilaciones de la ecuación, lo cual les permite converger hacia la solución exacta. Para obtener una mejor comprensión de los resultados de convergencia del método trapezoidal y RT-ω, se analizan los resultados en los puntos $t = 0{,}255$ y $0{,}645$, detallados en la tabla XI.

Tabla XI: Comparación de resultados, donde T: trapezoidal, RK-4 I: Runge Kutta-4 implícito y RC: RT-ω para $N = 1000$.

	Exacta	T	RK-4 I	RC
$u_1(0{,}255)$	0,10116292	0,28591421	35,14256669	0,29586809
$u_2(0{,}255)$	0,04483892	0,25069495	46,88495607	0,25927915
$u_1(0{,}645)$	0,00202061	0,00075868	1517,01485295	0,00079891
$u_2(0{,}645)$	$-0{,}00172666$	0,02371843	8555,50196083	0,02653587

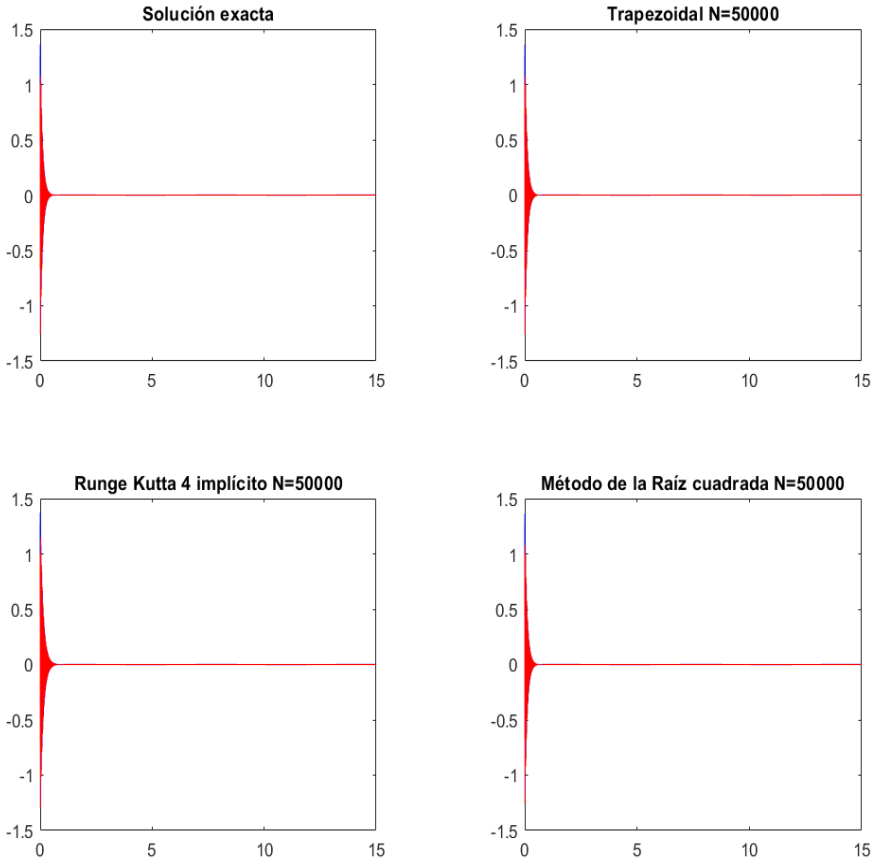

Figura XVIII: Comparación de los métodos trapezoidal, RK-4 implícito, RT-ω y la solución exacta para $N = 50000$.

Para $N = 50000$, se observa en la figura XVIII que los métodos d e R unge Kutta-4 implícito, trapezoidal y RT-ω convergen hacia la solución exacta. Esto se debe a la finura de la malla utilizada, aunque existe el riesgo de encontrar inestabilidades numéricas. Sin embargo, debido a la rápida velocidad de convergencia del método Runge Kutta-4 y la estabilidad de los métodos trapezoidal y RT-ω, este problema no se presenta. Para obtener una comprensión más detallada de los resultados de convergencia en los métodos trapezoidal y RT-ω, se analizan los resultados en los puntos $t = 0{,}0066$ y $0{,}28515$, los cuales se muestran en la tabla XII.

Tabla XII: Comparación de resultados, donde T: trapezoidal, RK-4 I: Runge Kutta-4 impl ícito y RC: RT-ω para $N = 50000$.

	Exacta	T	RK-4 I	RC
$u_1(0{,}0066)$	1,31283659	0,00000257	0,00243750	0,00000281
$u_2(0{,}0066)$	0,18553964	0,00001607	0,00013148	0,00001606
$u_1(0{,}28515)$	$-0{,}06957822$	0,00001080	0,00651247	0,00001017
$u_2(0{,}28515)$	$-0{,}04232651$	0,00004299	0,00232339	0,00004322

Como se observa en la tabla X, tanto el método trapezoidal como el método RT-ω convergen con una precisión de 10^{-5}, mientras que el método Runge Kutta-4 implícito alcanza una precisión de 10^{-2}. A pesar de esto, el método Runge Kutta-4 implícito muestra una convergencia superior al método RT-ω, logrando mejores resultados.

Ejemplo 12. *El siguiente ejemplo muestra una ecuación diferencial rígida altamente oscilante, y es dado por:*

$$\begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix}' = \begin{bmatrix} -1000 & 1000 & 999 \\ -1000 & -1000 & 1000 \\ 0 & 0 & -1 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix} + \begin{bmatrix} 1000\,\mathrm{sen}(t) - 999\cos(t) \\ 999\,\mathrm{sen}(t) + 1000\cos(t) \\ 0 \end{bmatrix}$$

con valores iniciales $u_1(0) = 2$, $u_2(0) = 2$ *y* $u_3(0) = 1$, *donde el jacobiano tiene valores propios puramente complejos con parte real altamente negativa* $\lambda_1 = -1000 + 1000i$, $\lambda_2 = -1000 - 1000i$, *y un valor propio real negativo y en módulo extremadamente pequeño* $\lambda_3 = -1$ *y cuya solución exacta es dada por:*

$$u(t) = \begin{bmatrix} e^{-1000t}(\mathrm{sen}(1000t) + cos(1000t)) + \mathrm{sen}(t) + e^{-t} \\ e^{-1000t}(\cos(1000t) - \mathrm{sen}(1000t)) + cos(t) \\ e^{-t} \end{bmatrix}.$$

Para mostrar los efectos de rigidez, realizaremos la prueba numérica con los métodos, trapezoidal, Runge Kutta-4 implícito y RT-ω*, utilizando una longitud de paso constante para* $N = 40, 300$ *y* 5000. *Para este problema utilizamos el intervalo* $t \in [0, 1]$.

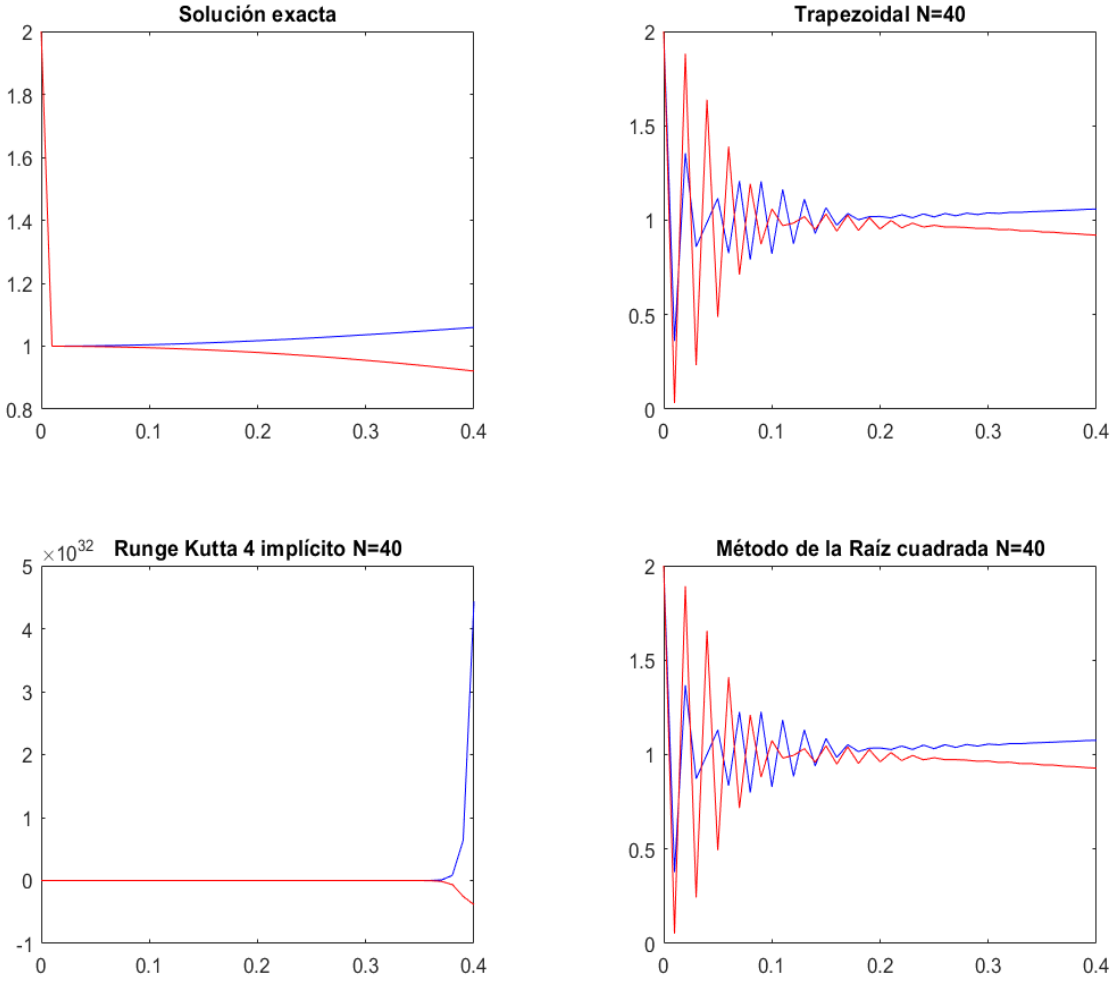

Figura XIX: Comparación de los métodos trapezoidal, RK-4 implícito, RT-ω y la solución exacta para $N = 40$.

Como se observa en la figura XIX, el método de Runge Kutta-4 implícito mostró una gran inestabilidad con una partición constante de $N = 40$, alcanzando una variación de 10^{32}. En contraste, tanto el método trapezoidal como el método RT-ω exhiben un comportamiento oscilatorio similar y no logran converger hacia la solución exacta.

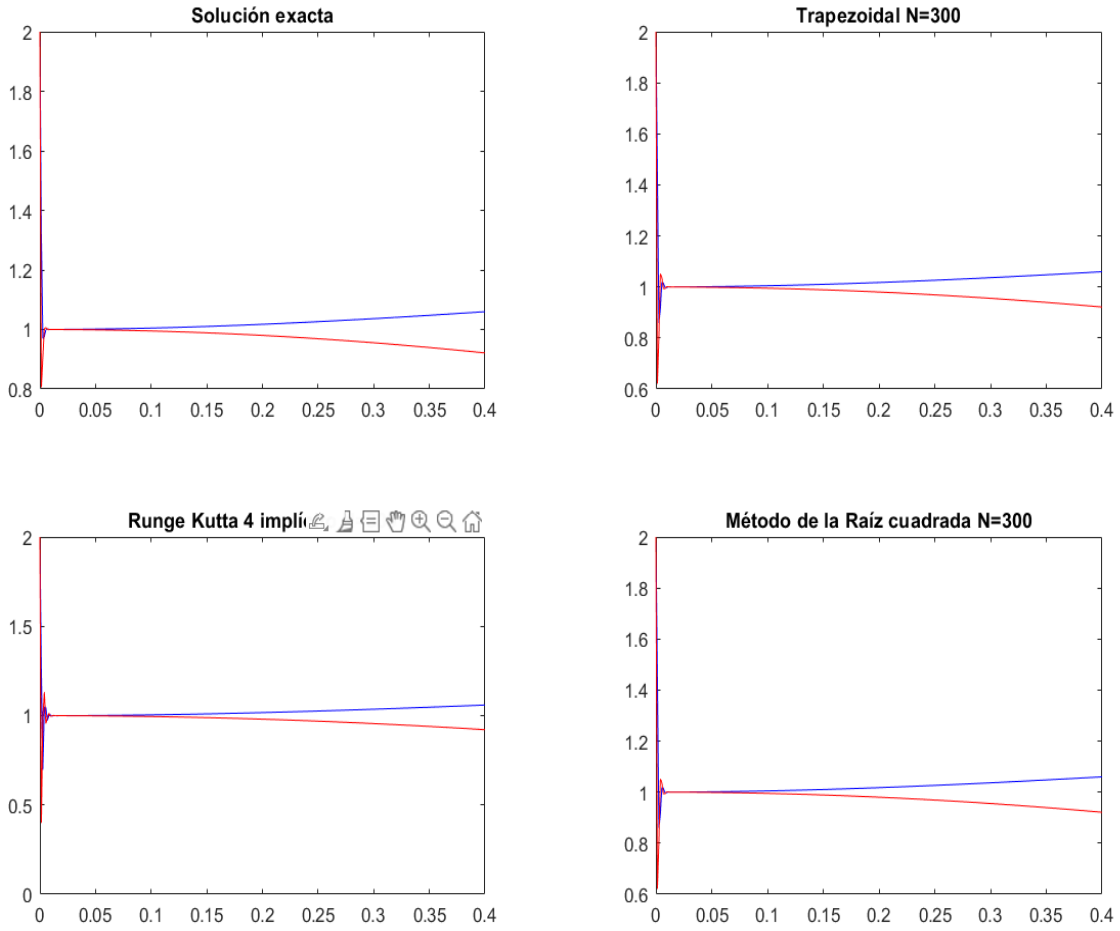

Figura XX: Comparación de los métodos trapezoidal, RK-4 implícito, RT-ω y la solución exacta para $N = 300$.

Para $N = 300$ se observa en la figura XX, el método de Runge Kutta-4 implícito no logra obtener una buena convergencia en comparación con el método trapezoidal y RT-ω, debido a las altas oscilaciones presentes en la ecuación diferencial rígida. Los métodos trapezoidal y RT-ω muestran comportamientos similares y se acercan a la solución exacta. Para un análisis detallado de la convergencia en los métodos trapezoidal y RT-ω, se consideran los resultados en los puntos $t = 0{,}004$ y $0{,}008$, detallados en la tabla XIII.

Tabla XIII: Comparación de resultados, donde T: trapezoidal, RK-4 I: Runge Kutta-4 implícito y RC: RT-ω para $N = 300$.

	Exacta	T	RK-4 I	RC
$u_1(0{,}004)$	0,974174757	0,061214218	0,073491159	0,060997019
$u_2(0{,}004)$	1,00188142070	0,049814216	0,128918247	0,050156230
$u_1(0{,}008)$	1,0003149124	0,002331758	0,001898020	0,0021245334
$u_2(0{,}008)$	0,999587298	0,006571962	0,013967017	0,006358132

Podemos ver de los resultados de la tabla XI que el método trapezoidal y el método RT-ω, convergen a la solución exacta, siendo el problema altamente oscilante.

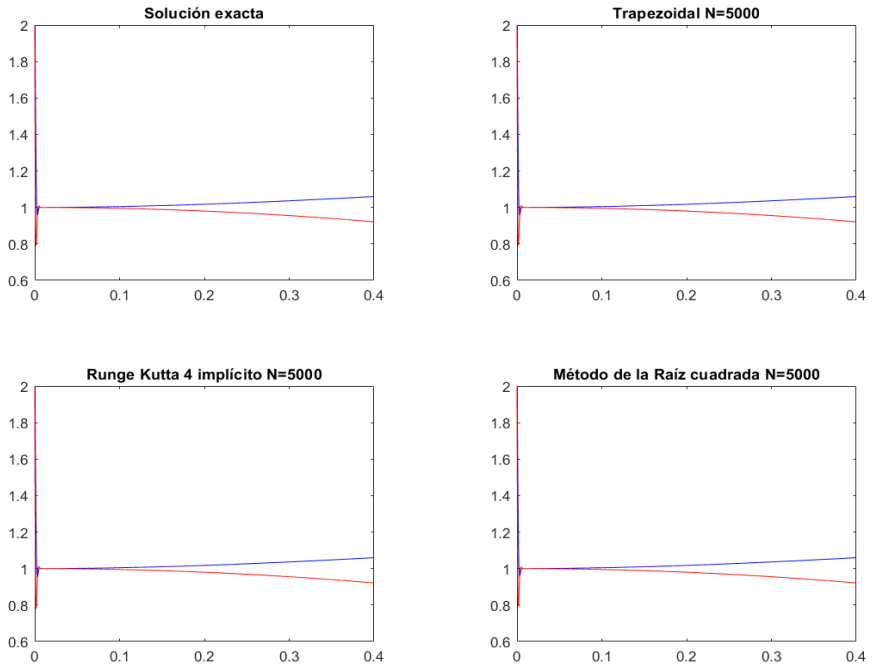

Figura XXI: Comparación de los métodos trapezoidal, RK-4 implícito, RT-ω y la solución exacta para $N = 5000$.

Para $N = 5000$, se puede observar en la figura XXI que todos los métodos muestran buenos resultados de convergencia hacia la solución exacta, gracias a la utilización de una malla muy fina. A pesar de que el método R T-ω tiene un orden de convergencia de uno, lo cual podrá llevar a malos resultados en una malla tan fina su estabilidad permitió obtener buenos resultados en esta situación. Para analizar más detalladamente la convergencia de los métodos trapezoidal y RT-ω, se examinan los resultados en los puntos $t = 0{,}0016$ y $0{,}2588$, que se muestran en la tabla XIV.

Tabla XIV: Comparación de resultados, donde T: trapezoidal, RK-4 I: Runge Kutta-4 implícito y RC: RT-ω para $N = 5000$.

	Exacta	T	RK-4 I	RC
$u_1(0{,}00175)$	$1{,}951065051$	$0{,}000097816$	$0{,}002154108$	$0{,}000097882$
$u_2(0{,}00175)$.	$1{,}577097976$	$0{,}000023891$	$0{,}003459907$	$0{,}000023898$
$u_1(0{,}2588)$	$1{,}027898102$	$0{,}0026141 \cdot 10^{-8}$	$0{,}000013662$	$0{,}24465111 \cdot 10^{-6}$
$u_2(0{,}2588)$.	$0{,}966697779$	$0{,}0000081 \cdot 10^{-8}$	$0{,}000003134$	$0{,}14210442 \cdot 10^{-6}$

Como se puede ver en la tabla XII, tanto el método trapezoidal como el método RT-ω muestran una convergencia con una precisión de 10^{-5}, en la parte más oscilatoria del problema de rigidez, y con una precisión de 10^{-6}, en otra parte. En cambio, el método de Runge Kutta-4 implícito alcanza una precisión de 10^{-2}, en esta misma parte. También es notable que a pesar de que el método RT-ω tiene un orden de convergencia $O(h)$, su buen comportamiento de estabilidad le permite converger hacia la solución exacta incluso con una malla de tamaño $N = 5000$.

Ejemplo 13. *El siguiente ejemplo es un problema altamente rígido debido a que sus valores propios están muy distanciados. Consideremos la ecuación diferencial:*

$$\begin{bmatrix} u_1 \\ u_2 \end{bmatrix}' = \begin{bmatrix} -2000 & 999{,}75 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} + \begin{bmatrix} 1000{,}25 \\ 0 \end{bmatrix}$$

con condiciones iniciales $u_1(0) = 0$, $u_2(0) = -2$, *donde el jacobiano tiene valores propios reales* $\lambda_1 = -0{,}5$ *y* $\lambda_2 = -2000{,}5$ *y cuya solución exacta es dada por:*

$$u(t) = \begin{bmatrix} 1 - 1{,}499875e^{-0{,}5t} + 0{,}499875e^{-2000{,}5t} \\ 1 - 2{,}99975e^{-0{,}5t} - 0{,}00025e^{-2000{,}5t} \end{bmatrix}$$

Se puede observar que el término $e^{-2000{,}5t}$ *en la solución causa que este problema sea rígido y desde que* $\lambda_1 \ll \lambda_2$ *entonces, es altamente rígido. Para mostrar los efectos de rigidez, realizaremos la prueba numérica con el método trapezoidal, Runge Kutta-4 implícito y RT-*ω*, utilizando las longitudes de paso constantes para* $N = 50, 1000$ *y* 6000*. Para este problema utilizamos el intervalo* $t \in [0, 2]$*, con longitud de paso* h *constante.*

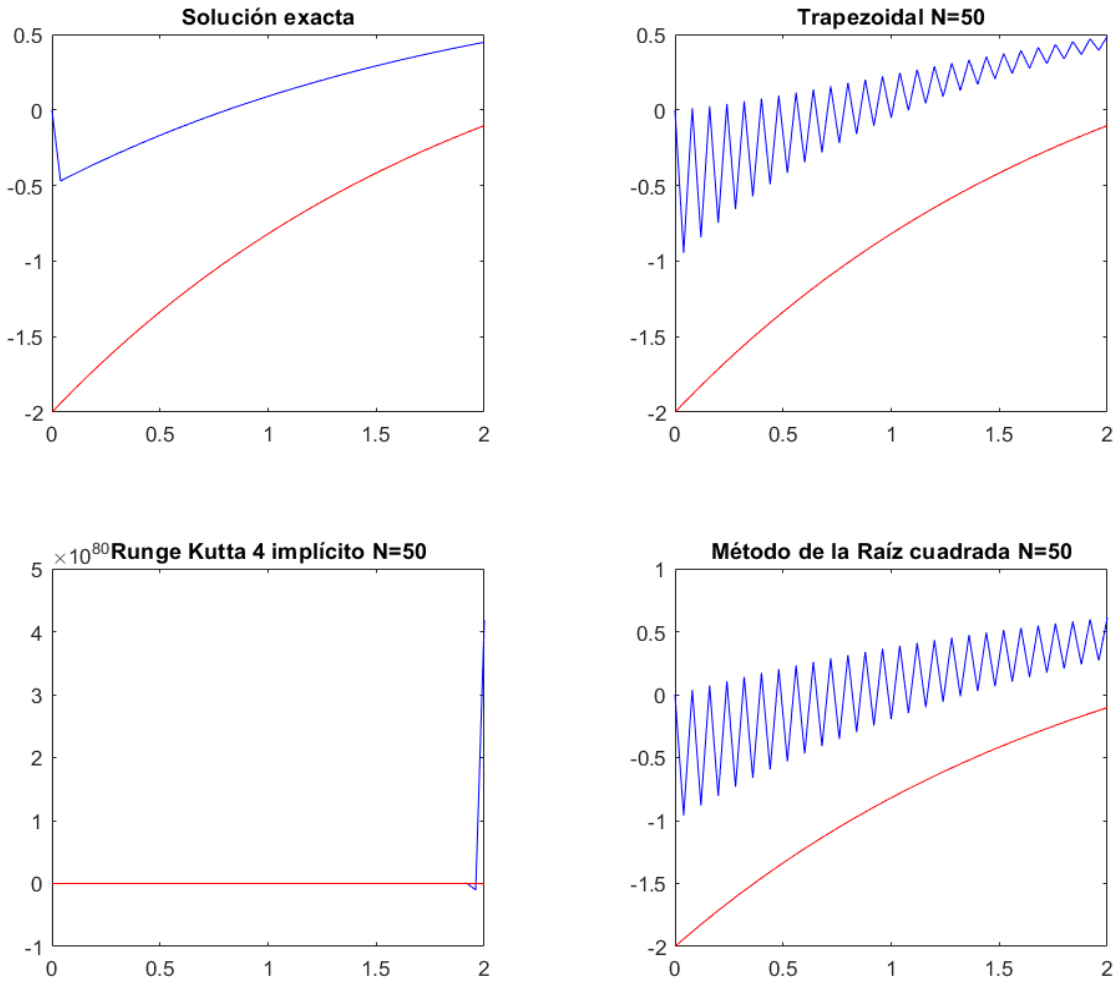

Figura XXII: Comparación de los métodos trapezoidal, RK-4 implícito, RT-ω y la soluci

ón exacta para $N = 50$.

Como se puede ver en la figura XXII el método de Runge Kutta-4 implícito muestra inestabilidad cuando se utiliza una malla con $N = 50$. *En contraste, tanto el método trapezoidal como RT-*ω *muestran resultados en ambas soluciones que están dentro del mismo intervalo respecto a la solución exacta.*

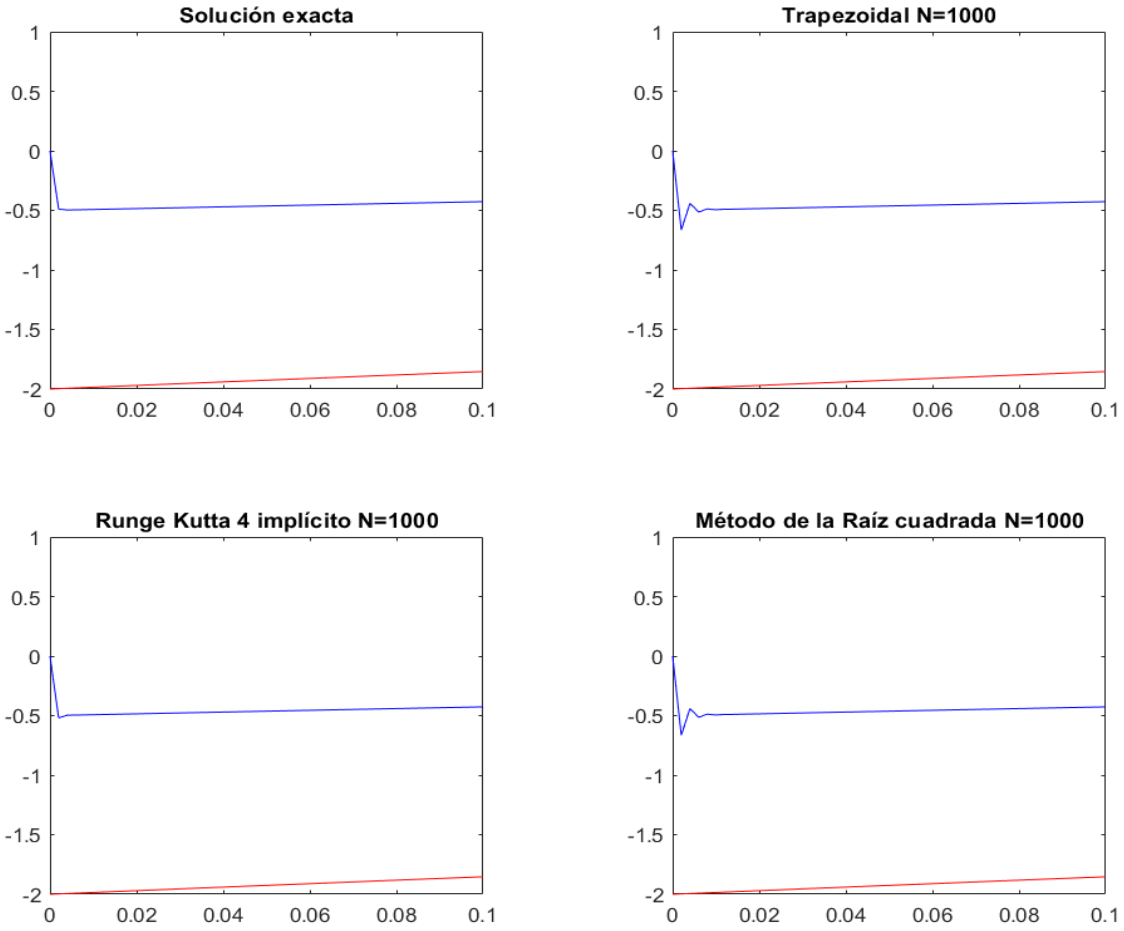

Figura XXIII: Comparación de los métodos trapezoidal, RK-4 implícito, RT-ω y la solución exacta para N=1000.

Para $N = 1000$ vemos en la figura XXIII que el método de Runge Kutta-4 implícito muestra buenos resultados. A pesar de ser inestable, debido a su rápida velocidad de convergencia, logra aproximarse a la solución en comparación con otros métodos. Sin embargo, uno de los conjuntos de soluciones no logra converger adecuadamente debido a la alta rigidez del problema. Para obtener una mejor comprensión de la convergencia de los métodos analizados, examinamos los resultados en $t = 0{,}002$ y $0{,}006$, detallados en la tabla XV.

Tabla XV: Comparación de resultados, donde T: trapezoidal, RK-4 I: Runge Kutta-4 implícito y RC: RT-ω para $N = 1000$.

	Exacta	T	RK-4 I	RC
$u_1(0{,}002)$	$-0{,}489229496$	$0{,}175826911$	$0{,}029878945$	$0{,}176049300$
$u_2(0{,}002)$	$-1{,}996756324$	$0{,}000087936$	$0{,}000015507$	$0{,}000088048$
$u_1(0{,}006)$	$-0{,}495379056$	$0{,}018535468$	$0{,}000037884$	$0{,}018609745$
$u_1(0{,}006)$	$-1{,}990764237$	$0{,}000009271$	$0{,}000001708$	$0{,}000009310$

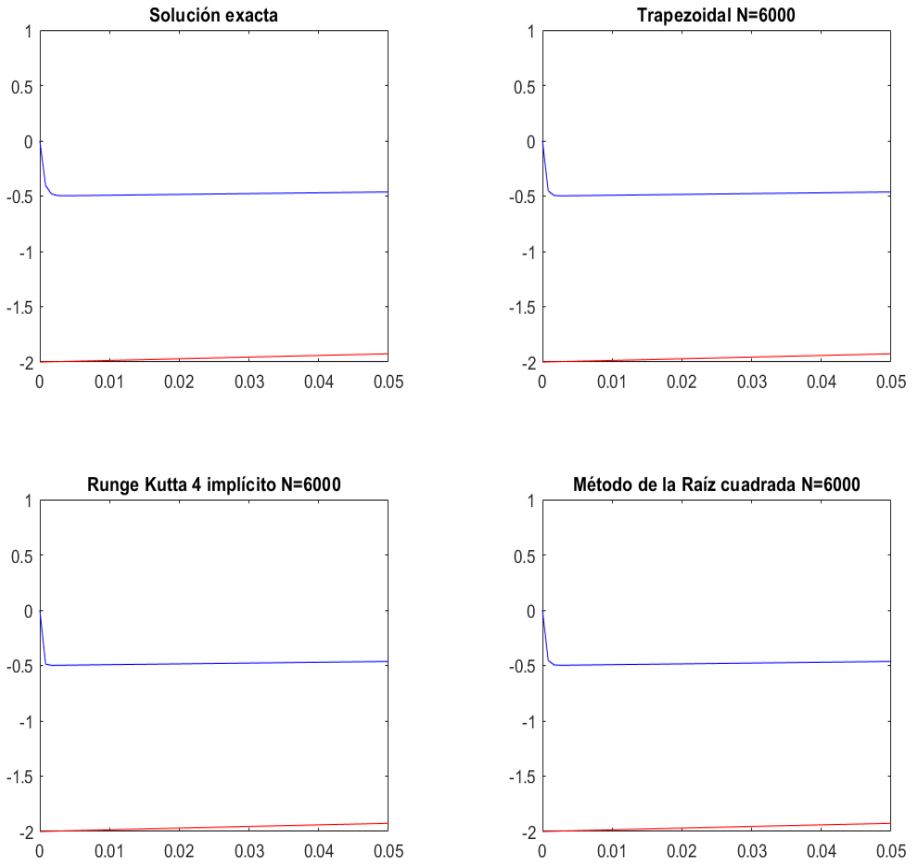

Figura XXIV: Comparación de los métodos trapezoidal, RK-4 implícito, RT-ω y la solución exacta para $N = 6000$.

En la figura XXIV que todos los métodos muestran buenos resultados de convergencia en el intervalo $[0{,}01;\ 2]$ *respecto a la solución exacta. Sin embargo, en el intervalo* $[0;\ 0{,}01]$*, donde hay una mayor inestabilidad, el método de Runge Kutta-4 implícito comienza a perder precisión computacional debido a dos razones principales: su inestabilidad y el tamaño de la malla elegida. A pesar de tener una alta velocidad de convergencia, este método muestra una pérdida de precisión. Mientras tanto, los otros métodos mejoran y se acercan a la solución exacta debido a su mayor estabilidad. Para analizar la con-vergencia de los métodos trapezoidal y RT-*ω*, se consideran los resultados en los puntos* $t = 0{,}002$ *y* $0{,}0325$*, los cuales se muestran en la tabla XVI.*

Tabla XVI: Comparación de resultados, donde T: trapezoidal, RK-4 I: Runge Kutta-4 implícito y RC: RT-ω para $N = 6000$.

	Exacta	T	RK-4 I	RC
$u_1(0{,}002)$	$-0{,}48922949$	$0{,}001344614$	$0{,}004439976$	$0{,}001345590$
$u_2(0{,}002)$	$-1{,}99675632$	$0{,}000000672$	$0{,}000002315$	$0{,}000000673$
$u_1(0{,}0325)$	$-0{,}49189696$	$0{,}0136040 \cdot 10^{-8}$	$0{,}2492559 \cdot 10^{-6}$	$0{,}0080853 \cdot 10^{-8}$
$u_1(0{,}0325)$	$-1{,}98379392$	$0{,}0036906 \cdot 10^{-8}$	$0{,}4990367 \cdot 10^{-6}$	$0{,}0147435 \cdot 10^{-8}$

Como se puede ver en la tabla XIV, tanto el método trapezoidal como el método RT-ω *convergen con una precisión de* 10^{-2} *para la primera solución, que es la que presenta un mayor efecto de rigidez. Para la segunda solución, ambos métodos logran una precisión de* 10^{-8}*. También se puede observar que, a pesar de que el método Runge Kutta-4 implícito tiene un orden de convergencia mayor que los otros dos métodos, su aproximación a la solución exacta no es tan significativa, debido a que no es estable.*

V. Conclusiones y recomendaciones

Conclusiones

En esta trabajo, hemos explorado y evaluado métodos para el cálculo de la raíz cuadrada de matrices no singulares y cómo aplicarlo en ecuaciones diferenciales rígidas de diferentes especies, aproximando la solución con el método irracional RT-ω y con $\omega = 2$. A través de un análisis comparativo con otros métodos, hemos llegado a las siguientes conclusiones.

1. Para el cálculo de la raíz cuadrada de una matriz no singular, tomando como referencia la inversa de la matriz determinada por el software Matlab, se alcanzó buenos resultados en todos los métodos estudiados, para aquellas matrices que tenían un buen condicionamiento. Para las matrices mal condicionadas, como la matriz de Hilbert que al aumentar su dimensión aumentada su condicionamiento, todos los métodos estudiados no dieron buenos resultados, sin embargo, el más rescatable fue el método de Padé a escala y con $p = 3$.

2. Para el método de Padé a escala los errores residuales fueron muy pequeños en comparación de los otros métodos estudiados. Es por ello que se tomó como referencia este método para el desarrollo de la siguiente fase de este trabajo.

3. En el desarrollo del método en diferencia irracional RT-ω para la solución de ecuaciones diferenciales rígidas lineales de diferentes especies, se tomó como referencia $\omega = 2$, tal como el autor del método había mencionado, ya que ese valor obtuvo buenos resultados. Este método ha demostrado ser una herramienta eficiente, para la aproximación de ecuaciones diferenciales rígidas. También demostró ser muy estable en la variedad de problemas rígidos estudiados. Aunque su comportamiento convergente no es muy alto, no presentó ninguna inestabilidad computacional, debido a los pequeños valores que se tomo en la malla.

4. El método irracional RT-ω resultó muy eficiente en comparación de los métodos tradicionales que fueron tomados en este trabajo.

5. Se pretende estudiar este método en aquellos problemas rígidos no lineales, ya que no se vio un desarrollo en el artículo tomado.

6. Por último, se quiere mejorar el orden de convergencia debido a su gran estabilidad que ella posee en el desarrollo de la aproximación de ecuaciones diferenciales rígidas.

Discusiones y recomendaciones

Basado en los hallazgos y resultados obtenidos en esta tesis, se proponen las siguientes recomendaciones para futuras investigaciones y aplicaciones prácticas:

1. Considerar el desarrollo y la evaluación de métodos iterativos para mejorar el cálculo de la raíz cuadrada en matrices con condicionamiento alto, centrando la idea en mejorar la precisión numérica y para el caso de matrices de grandes dimensiones y con gran cantidad de ceros, esto es, sobre las matrices dispersas.

2. Realizar estudio de paralelismo en el cálculo de la raíz cuadrada de matrices con grandes dimensiones.

3. Realizar un profundo estudio en el valor de ω, clasificando lo tipos de ecuaciones rígidas, en el método RT-ω y así mejorar su eficiencia, estabilidad y su rapidez de convergencia.

4. La adaptación del método RT-ω a problemas multidimensionales y no lineales.

Bibliografía

[1] J. Bak, D. J. Newman, and D. J. Newman. *Complex analysis*, volume 8. Springer, 2010.

[2] Å. Björck and S. Hammarling. A schur method for the square root of a matrix. *Linear algebra and its applications*, 52:127–140, 1983.

[3] R. L. Burden and J. D. Faires. *Numerical analysis*. Brooks Cole, 1997.

[4] J. Cash. Modified extended backward differentiation formulae for the numerical solution of stiff initial value problems in odes and daes. *Journal of Computational and Applied Mathematics*, 125(1-2):117–130, 2000.

[5] C. Chávez. Álgebra lineal. *Editorial San Marcos*, 1992.

[6] S. H. Cheng, N. J. Higham, C. S. Kenney, and A. J. Laub. Approximating the logarithm of a matrix to specified accuracy. *SIAM Journal on Matrix Analysis and Applications*, 22(4):1112–1125, 2001.

[7] G. J. Cooper and A. Sayfy. Additive runge-kutta methods for stiff ordinary differential equations. *Mathematics of Computation*, 40(161):207–218, 1983.

[8] F. De Hoog and R. Mattheij. On dichotomy and well conditioning in bvp. *SIAM journal on numerical analysis*, 24(1):89–105, 1987.

[9] K. Deimling. *Nonlinear functional analysis*. Courier Corporation, 2010.

[10] E. D. Denman and A. N. Beavers Jr. The matrix sign function and computations in systems. *Applied mathematics and Computation*, 2(1):63–94, 1976.

[11] L. Dieci, B. Morini, and A. Papini. Computational techniques for real logarithms of matrices. *SIAM Journal on Matrix Analysis and Applications*, 17(3):570–593, 1996.

[12] W. Gautschi. *Numerical analysis*. Springer Science & Business Media, 2011.

[13] G. Golub, S. Nash, and C. Van Loan. A hessenberg-schur method for the problem ax+ xb= c. *IEEE Transactions on Automatic Control*, 24(6):909–913, 1979.

[14] G. H. Golub and C. F. Van Loan. *Matrix computations*. JHU press, 2013.

[15] S. Grossman, I. Stanley, et al. Álgebra lineal. 2019.

[16] N. J. Higham. Stable iterations for the matrix square root. *Numerical Algorithms*, 15:227–242, 1997.

[17] A. Izmailov and M. Solodov. Otimizaçao, volume i. *Condições de Otimalidade, Elementos de Análise Convexa e de Dualidade. IMPA, Rio de Janeiro*, 2, 2007.

[18] D. Kalman. The generalized vandermonde matrix. *Mathematics Magazine*, 57(1):15–21, 1984.

[19] M. A. Krasnosel'skii, G. M. Vainikko, R. Zabreyko, Y. B. Ruticki, and V. V. Stet'senko. *Approximate solution of operator equations*. Springer Science & Business Media, 2012.

[20] C.-M. Li and S.-Q. Shen. Newton's method for the matrix nonsingular square root. *Journal of Applied Mathematics*, 2014(1):267042, 2014.

[21] E. L. Lima. *Análise real*, volume 1. Impa Rio de Janeiro, 2004.

[22] J. Martín Vaquero et al. Métodos exponential fitting y adaptados para problemas stiff. 2006.

[23] J. T. M. Mendoza. Acerca de la raíz cuadrada de una matriz. *Rev. Fac. Cienc. Univ. Nal. Colomb. Medellín (Colombia)*, 5(1):89–95, 1995.

[24] J. M. Ortega and W. C. Rheinboldt. *Iterative solution of nonlinear equations in several variables*. SIAM, 2000.

[25] F. Piao, Q. Zhang, and Z. Wang. The solution to matrix equation ax+ xtc= b. *Journal of the Franklin Institute*, 344(8):1056–1062, 2007.

[26] B. A. Schmitt. An algebraic approximation for the matrix exponential in singularly perturbed boundary value problems. *SIAM journal on numerical analysis*, 27(1):51–66, 1990.

[27] J. Sotomayor. *Lições de equações diferenciais ordinárias*, volume 11. Instituto de Matemática Pura e Aplicada, CNPq, 1979.

[28] A. D. Wunsch. Variable compleja con aplicaciones. 1997.