Sección de Matemáticas
Universidad de La Laguna

Yaiza Pérez Tejera

Teoría de las estrellas de orden

Order star theory

Trabajo Fin de Grado
Grado en Matemáticas
La Laguna, Julio de 2023

Dirigido por
María Soledad Pérez Rodriguez

María Soledad Pérez Rodriguez
Análisis Matemático
Universidad de La Laguna
38200 La Laguna, Tenerife

Agradecimientos

En primer lugar agradecerle a mi tutora María Soledad Pérez por su dedicación, por todo lo que me ha enseñado, ayudado y guiado en este TFG.
También agradecerle a mi familia que han estado a lo largo de toda mi carrera apoyándome en todo momento y animándome a seguir adelante. Por último pero no menos importante, a mis amigos que me han tendido una mano cuando lo he necesitado y me han hecho pasar momentos inolvidables en mi etapa universitaria.
Gracias a todos, sin ustedes no hubiera sido posible.

Yaiza Pérez Tejera
La Laguna, 10 de julio de 2023

Resumen · Abstract

Resumen

En el presente Trabajo de Fin de Grado se ha estudiado la A-estabilidad de los métodos Runge-Kutta y, en particular, el estudio de la teoría de las estrellas de orden, que determina condiciones necesarias y suficientes para que la función de estabilidad de un método sea A-estable.

Para ello, hemos dividido este trabajo en tres capítulos. En el primero introducimos los métodos Runge-Kutta y algunas de sus propiedades. En el segundo capítulo estudiamos su estabilidad y definimos la función de estabilidad y la A-estabilidad. En último lugar, desarrollamos la teoría de las estrellas de orden, con la que garantizamos la A-estabilidad de los aproximantes de Padé bajo ciertas condiciones.

Palabras clave: *Métodos Runge-Kutta – Estabilidad – A-estabilidad – Aproximantes de Padé – Estrellas de orden.*

Abstract

In this Bachelor's thesis, the A-stability of Runge-Kutta methods has been studied, with a particular focus on the theory of order stars, which determines necessary and sufficient conditions for the stability function of a method to be A-stable.

To do this, we have divided this work into three chapters. In the first chapter, we introduce the Runge-Kutta methods and some of their properties. In the second chapter, we study their stability and define the stability function and A-stability. Finally, we develop the theory of order stars, which guarantees the A-stability of Padé approximants under certain conditions.

Keywords: *Runge-Kutta methods – Stability – A-stability – Padé approximants – Order stars.*

Índice general

Agradecimientos III

Resumen/Abstract V

Introducción IX

1. Introducción a los métodos Runge-Kutta 1
- 1.1. Introducción 1
- 1.2. Formulación de los métodos de Runge-Kutta 2
- 1.3. Método de Euler 4
- 1.4. Ejemplos de métodos RK 5
- 1.5. Existencia de la solución numérica de los RKI 6
- 1.6. Orden de los métodos Runge-Kutta 8

2. Estabilidad de los métodos RK 11
- 2.1. Ecuación test de Dahlquist 11
- 2.2. Función de estabilidad 12
- 2.3. A-estabilidad 15
- 2.4. I-estabilidad 16
- 2.5. L-estabilidad 18
- 2.6. Ejemplos 19

3. Estrellas de orden (Order Stars) 21
- 3.1. Aproximación de Padé a la función exponencial 21
- 3.2. Estrellas de orden (Order Stars) 23
- 3.3. Order Star cuando $z \to 0$ 24
- 3.4. Order Star cuando $z \to \infty$ 30
- 3.5. Ceros y polos de $R(z)$ 35
- 3.6. Orden y estabilidad para $R(z)$ 41
- 3.7. Estabilidad de las aproximaciones de Padé 44
- 3.8. Ejemplos de aproximantes de Padé 45

Bibliografía 49

Poster 51

Introducción

En el campo del Análisis Numérico, uno de los temas fundamentales es el estudio de los métodos de resolución de ecuaciones diferenciales que tienen aplicaciones en diversas áreas, como la física, la ingeniería, la biología, entre otras. En particular, los métodos de Runge-Kutta (RK) han demostrado ser una herramienta muy importante para aproximar soluciones númericas con alta precisión.

Este Trabajo de Fin de grado tiene como principal objetivo estudiar las estrellas de orden, una herramienta que permite clasificar y comparar la A-estabilidad de los métodos.

En esta memoria comenzamos examinando los métodos RK y una serie de propiedades que los caracterizan. En el primer capítulo, se introducen los métodos RK, tanto explícitos como implícitos, y su orden de convergencia, viendo ejemplos de diferentes órdenes. Para ello principalmente utilizaremos [4], [5] y [6].

En el segundo capítulo se estudia la estabilidad de los métodos numéricos siguiendo [4] y [5]. La estabilidad de un método numérico tiene que ver con la manera en que los errores numéricos se propagan. Introduciremos la función de estabilidad, así como ciertas propiedades que satisface ésta demostrando que es una función racional que aproxima a la función exponencial. Se define entonces el concepto de la A-estabilidad de los métodos o, lo que es lo mismo, que la función de estabilidad está acotada por 1 en el semiplano complejo negativo, lo que garantiza que el método será estable sobre problemas lineales de coeficientes constantes. También se definen otros conceptos relacionados como la I-estabilidad y la L-estabilidad, viendo ejemplos de métodos que satisfacen estas propiedades.

En el último capítulo, abordamos el estudio de la teoría de las estrellas de orden *(order stars)*, donde principalmente desarrollamos parte del Capítulo

IV.4 de [5]. Esta teoría se estudia en general para profundizar en la relación entre los aproximantes racionales a la exponencial y dicha exponencial. Con ello, por una parte, si tenemos una aproximación racional a e^z podremos deducir si es A-estable o no y, por otro lado, ayuda a construir aproximaciones racionales que sean A-estables. En particular, se estudian los aproximantes de Padé a e^z, que alcanzan el orden de aproximación mayor posible (alrededor de $z = 0$), entre todas las aproximaciones racionales a la exponencial con los grados del numerador y del denominador prefijados.

El objetivo de la teoría desarrollada permitirá demostrar el principal teorema de este trabajo que garantiza la A-estabilidad de los aproximantes de Padé a la exponencial.

1

Introducción a los métodos Runge-Kutta

En este primer capítulo introduciremos los métodos Runge-Kutta, una familia de algoritmos que se utilizan ampliamente para la solución numérica de EDOs. Haremos hincapié en algunas de sus propiedades y visualizaremos algunos ejemplos. Por último, se demostrará el teorema que garantiza la definición de los métodos RK implícitos.

Las principales propiedades sobre ecuaciones diferenciales las hemos extraído de [6] y [1]. Para introducir los métodos RK y sus propiedades hemos seguido [4] y [5].

1.1. Introducción

Consideremos el problema de valor inicial (PVI)

$$y'(t) = f(t, y(t)), \quad y(t_0) = y_0, \quad t \in [t_0, t_f] \tag{1.1}$$

con $f : [t_0, t_f] \times \Omega \to \mathbb{R}^m$, donde $\Omega \subset \mathbb{R}^m$ es un abierto no vacío.

Una **solución** de (1.1) es una función continuamente diferenciable en algún intervalo I que contiene a t_0 y que verifica (1.1).

Teorema 1 *Teorema de existencia[6]*
Sea f continua en $[t_0, t_f] \times \Omega$. Para cada $y_0 \in \Omega$ existe una solución de (1.1).

Definición 1 *Una función $f : [t_0, t_f] \times \Omega \to \mathbb{R}^m$ se dice **Lipschitz** en $[t_0, t_f] \times \Omega$ si existe una constante $L > 0$ tal que*

$$|f(s, u) - f(s, v)| \leq L|u - v|, \forall s \in I \subset [t_0, t_f], \quad \forall u, v \in U \subset \Omega.$$

Definición 2 *Una función $f : [t_0, t_f] \times \Omega \to \mathbb{R}^m$ se dice* ***localmente Lipschitz*** *si para cada $(t, y) \in [t_0, t_f] \times \Omega$ existen entornos abiertos $I \subset [t_0, t_f], U \subset \Omega$ y una constante $L > 0$ tal que $(t, y) \in I \times U$ y*

$$||f(s, u) - f(s, v)|| \leq L||u - v||, \forall s \in I, \forall u, v \in U \tag{1.2}$$

Observación 1 Si f es una función lipschitz entonces f es localmente lipschitz.

Teorema 2 *[6] Sea $f : [t_0, t_f] \times \Omega \to \mathbb{R}^m$ continua y localmente Lipschitz con respecto a su segunda componente. Entonces, para cada $y_0 \in \Omega$, existe una única solución de (1.1).*

En este trabajo estamos interesados en la aproximación numérica de las soluciones de PVI's (1.1). En particular, la aproximación dada por los métodos Runge-Kutta.

Para simplificar la explicación, en lo que sigue consideraremos que la función derivada f está definida, es continua y Lipschitz respecto de y en $I \times \mathbb{R}^m$ donde $I \subset \mathbb{R}$ es cualquier intervalo cerrado.
En estos casos utilizaremos como normas en $\mathbb{R}^m$,

$$||v||_\infty = \max_{1 \leq i \leq m} \{|v_i|\} \quad \text{ó} \quad ||v||_2 = \sqrt{\sum_{i=1}^{m} |v_i|^2}$$

donde son la norma uniforme y la norma l_2 respectivamente.

1.2. Formulación de los métodos de Runge-Kutta

Dado el PVI (1.1), para resolverlo numéricamente se toma una partición del intervalo de integración $[t_0, t_f]$

$$P_N = \{t_0 < t_1 < \ldots < t_N = t_f\} \tag{1.3}$$

y se calculan una serie de aproximantes $y_0, y_1, \ldots, y_N$ tal que

$$y_n \approx y(t_n), \quad n = 0, 1, \ldots, N$$

mediante el siguiente algoritmo.

Definición 3 *Dado un entero $s \geq 1$ (número de etapas) el método que, dada una aproximación y_n a la solución del PVI (1.1) en t_n nos da una aproximación y_{n+1} a la solución en $t_{n+1} = t_n + h$ mediante la fórmula:*

$$k_i = f(t_n + c_i h, y_n + h \sum_{j=1}^{s} a_{ij} k_j), \quad i = 1, \ldots, s \tag{1.4}$$

$$y_{n+1} = y_n + h \sum_{i=1}^{s} b_i k_i \tag{1.5}$$

se llama método ***Runge-Kutta de s etapas*** *(RK).*

Para manejar los coeficientes que definen el método, se define la **tabla de Butcher** asociada al RK (1.4) como:

$$\begin{array}{c|cccc} c_1 & a_{11} & a_{12} & \ldots & a_{1s} \\ c_2 & a_{21} & a_{22} & \ldots & a_{2s} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ c_s & a_{s1} & a_{s2} & \ldots & a_{ss} \\ \hline & b_1 & b_2 & \ldots & b_s \end{array}$$

donde la matriz $A = (a_{ij})_{i,j=1}^{s}$ se llama **matriz de coeficientes**, el vector $c = (c_1, c_2, \ldots, c_s)^T$ es el vector de **nodos** y $b = (b_1, b_2, \ldots, b_s)^T$ es el vector de **pesos**.

Como es usual, a partir de ahora supondremos que los pesos y los nodos satisfacen:

$$\sum_{i=1}^{s} b_i = 1, \quad c_i = \sum_{j=1}^{s} a_{ij}, \quad i = 1, \ldots, s. \tag{1.6}$$

Estas expresiones también pueden escribirse de forma vectorial como $b^T e = 1$ y $Ae = c$ donde $e = (1, 1, \ldots, 1)^T$, por lo que denotaremos como $RK(A, b)$ al método RK con matriz A, pesos b y nodos $c = Ae$.

Según la forma de la matriz A los métodos RK se pueden clasificar en dos grupos:

1. Cuando la matriz A es triangular inferior estricta, el método se dice **explícito (RKE)**, pues sus etapas $\{k_i\}_{i=1}$ se obtienen de forma recursiva, es decir,

$$\begin{cases} k_i = f(t_n + c_i h, y_n + h \sum_{j=1}^{i-1} k_j), \quad 1 \leq i \leq s \\ y_{n+1} = y_n + h \sum_{i=1}^{s} b_j k_i \end{cases} \tag{1.7}$$

2. Cuando $a_{ij} \neq 0$ para algún $j \geq i$, el método se llama **implícito (RKI)**. En estos métodos necesitaremos resolver el sistema implícito (1.4) de dimensión $s \times m$ para calcular sus etapas $\{k_i\}_{i=1}^{s}$.

1.3. Método de Euler

Euler fue el primero que propuso aproximar la solución de un PVI. El método de Euler también es el más simple de los métodos numéricos para resolver un PVI's (1.1).

Consideramos la partición P_N dada en (1.3) y denotamos $h_j = t_{j+1} - t_j$, $j = 0, \ldots, N-1$. Por la definición de derivada para cada, $k = 0, 1, \ldots, N-1$,

$$y'(t_k) = \lim_{k \to \infty} \frac{y(t_k + h_k) - y(t_k)}{h_k},$$

por lo que esperamos que para h_k "suficientemente pequeño"

$$y'(t_k) = f(t_k, y(t_k)) \simeq \frac{y(t_k + h_k) - y(t_k)}{h_k}.$$

Por tanto, como $t_k + h_k = t_{k+1}$:

$$y(t_{k+1}) \simeq y(t_k) + h_k f(t_k, y(t_k)), \quad k = 0, 1, \ldots, N-1.$$

Como estamos buscando unos aproximantes $y_k \approx y(t_k)$, esto nos sugiere el cálculo de los y_k mediante la recurrencia,

$$y_0 = y(t_0), \quad y_{k+1} = y_k + h f(t_k, y_k), \quad k = 0, 1, \ldots, N-1, \tag{1.8}$$

Esta fórmula (1.8) se conoce como **método de Euler**, y es un caso particualr de RK (1.4)-(1.5) con $s = 1$, $A = 0$ y $b = 1$.

Teorema 3 *Teorema de convergencia del método de Euler:[1]*
Consideramos el PVI (1.1) con f continua y Lipschitz respecto de y. Si $\{y_n\}_{n=0}^{N}$ es la solución dada por el método de Euler (1.8) sobre la partición (1.3), y la solución exacta $y(t)$ verifica

$$||y''(t)|| \leq Y_2, \quad \forall t \in [t_0, t_f]$$

entonces

$$\max_{t_n \in P} ||y_n - y(t_n)|| \leq \left(\frac{Y_2(e^{L(t_f - t_0)} - 1)}{2L} \right) h_{max}$$

donde $h_{max} = \max h_j$ y L es la constante de Lipschitz de f.

1.4. Ejemplos de métodos RK

A partir del método de Euler y con el desarrollo de las fórmulas de cuadratura, a finales del siglo XIX varios autores comenzaron a proponer otros métodos de un paso para resolver PVIs, que más tarde se incluyeron en la clase de métodos RK. A continuación, pondremos algunos ejemplos:

1. Método de Runge (1905): (4 etapas)

$$\begin{cases} k_1 = f_n \\ k_2 = f(t_n + \frac{h}{2}, y_n + \frac{h}{2}k_1) \\ k_3 = f(t_n + h, y_n + hk_2) \\ k_4 = f(t_n + h, y_n + hk_3) \\ y_{n+1} = y_n + h\left(\frac{k_1}{6} + \frac{2}{3}k_2 + \frac{1}{6}k_4\right) \end{cases} \qquad \begin{array}{c|cccc} 0 & 0 & 0 & 0 & 0 \\ 1/2 & 1/2 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ \hline & 1/6 & 2/3 & 0 & 1/6 \end{array} \tag{1.9}$$

2. Método de Kutta (1905):

$$\begin{cases} k_1 = f_n \\ k_2 = f(t_n + \frac{h}{2}, y_n + \frac{h}{2}k_1) \\ k_3 = f(t_n + \frac{h}{2}, y_n + \frac{h}{2}k_2) \\ k_4 = f(t_n + h, y_n + hk_3) \\ y_{n+1} = y_n + h\left(\frac{k_1}{6} + \frac{k_2}{3} + \frac{k_3}{3} + \frac{k_4}{6}\right) \end{cases} \qquad \begin{array}{c|cccc} 0 & 0 & 0 & 0 & 0 \\ 1/2 & 1/2 & 0 & 0 & 0 \\ 1/2 & 0 & 1/2 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ \hline & 1/6 & 1/3 & 1/3 & 1/6 \end{array} \tag{1.10}$$

3. Método de Euler implícito

$$y_1 = y_0 + hf(x_1, y_1) \qquad \begin{array}{c|c} 1 & 1 \\ \hline & 1 \end{array} \tag{1.11}$$

4. Regla del punto medio implícito

$$\begin{aligned} k_1 &= f\left(x_0 + \frac{h}{2}, y_0 + \frac{h}{2}k_1\right) \\ y_1 &= y_0 + hk_1 \end{aligned} \qquad \begin{array}{c|c} 1/2 & 1/2 \\ \hline & 1 \end{array} \tag{1.12}$$

5. Regla trapezoidal

$$\begin{aligned} k_1 &= f(t_n, y_n) \\ k_2 &= f(t_n + h, y_n + \frac{h}{2}(k_1 + k_2)) \\ y_{n+1} &= y_n + \frac{h}{2}(k_1 + k_2) \end{aligned} \qquad \begin{array}{c|cc} 0 & 0 & \\ 1 & 1/2 & 1/2 \\ \hline & 1/2 & 1/2 \end{array} \tag{1.13}$$

Como se puede observar los dos primeros son explícitos mientras que los demás son implícitos.

1.5. Existencia de la solución numérica de los RKI

Para calcular la aproximación y_{n+1} en cada método implícito (1.4)-(1.5), necesitamos asegurar que el sistema implícito (1.4) de las etapas tiene solución única.

Teorema 4 *[4] Sea $f : [t_0, t_f] \times \mathbb{R}^m \to \mathbb{R}^m$ una función continua que satisface una condición de Lipschitz respecto a y, siendo L la constante de Lipschitz. Si tomamos*

$$\tilde{h} < \frac{1}{L \operatorname{máx}_i \sum_j |a_{ij}|}, \tag{1.14}$$

para todo $h \in [0, \tilde{h}]$ existe una solución única del sistema implícito (1.4) que se puede obtener por iteración funcional. Si f es p veces continuamente diferenciable, las funciones $k_i = k_i(h)$ también son de $C^p([0, \tilde{h}])$.

Demostración: Sea $K = (k_1, k_2, \ldots, k_s)^T \in \mathbb{R}^{sm}$ y consideramos la norma $||K|| = \operatorname{máx}_{1 \le i \le s} ||k_i||$ donde $||\cdot||$ es la norma considerada en $\mathbb{R}^m$. El sistema implícito (1.4) se puede reescribir como $K = F(K)$ donde $F(K) = (F_1(K), \ldots, F_s(K))^T$, siendo

$$F_i(K) = f(t_n + c_i h, y_n + h \sum_{j=1}^{s} a_{ij} k_j), \quad 1 \le i \le s, \quad h \in [0, \tilde{h}].$$

Considérense $K^1, K^2 \in \mathbb{R}^{sm}$. Para cada $i = 1, 2, \ldots, s$,

$$||F_i(K^1) - F_i(K^2)|| = ||f(t_n + c_i h, y_n + h \sum_{j=1}^{s} a_{ij} k_j^1) - f(t_n + c_i h, y_n + h \sum_{j=1}^{s} a_{ij} k_j^2)||$$

$$\le L||h \sum_{j=1}^{s} a_{ij}(k_j^1 - k_j^2)|| \le hL \sum_{j=1}^{s} |a_{ij}| ||k_j^1 - k_j^2||$$

Por tanto,

$$||F(K^1) - F(K^2)|| = \operatorname{máx}_{1 \le i \le s} ||F_i(K^1) - F_i(K^2)|| \le$$

$$\leq \max_{1\leq i\leq s} hL \sum_{j=1}^{s} |a_{ij}| ||k_j^1 - k_j^2|| \leq$$

$$\leq \tilde{h}L \left[\max_{1\leq i\leq s} \sum_{j=1}^{s} |a_{ij}| \right] ||K^1 - K^2||$$

Por hipótesis tenemos que $\tilde{h} \cdot L \cdot \max \sum_{j=1}^{s} |a_{ij}| < 1$, por lo que para todo $h \in [0, \tilde{h}]$,

$$||F(K^1) - F(K^2)|| \leq M||K^1 - K^2||, \quad M = \tilde{h}L \max \sum |a_{ij}| < 1.$$

En consecuencia, F es contractiva. Por el Teorema del Punto Fijo, existe solución única del sistema $K = F(K)$ y dicha solución se puede aproximar por iteración funcional

$$K^{(\nu+1)} = F(K^{(\nu)}), \quad \nu = 0, 1, \ldots, \quad \text{dado } K^{(0)}.$$

En otras palabras, tomando un valor inicial $K^{(0)} = (k_1^{(0)}, \ldots, k_s^{(0)})$ se obtiene para cada $i = 1, \ldots, s$, la aproximación a la solución de (1.4) mediante la iteración:

$$k_i^{(\nu+1)} = f\left(t_n + c_i h, y_n + h \sum_{j=1}^{s} a_{ij} k_j^{(\nu)}\right), \quad \nu = 0, 1, 2, \ldots$$

Para demostrar que la ecuación (1.4) define a las etapas k_i como funciones derivables de h, vamos a aplicar el Teorema de la Función Implícita [7]. Para ello reescribimos (1.4) como

$$\phi(h, K) = K - F(K) = 0$$

En consecuencia,

$$\frac{\partial \phi(h, K)}{\partial K} = I - \frac{\partial F}{\partial K}, \qquad \frac{\partial F}{\partial K} = \frac{\partial (F_1(K), \ldots, F_s(K))}{\partial (k_1, k_2, \ldots, k_s)}$$

Para cada $i, l = 1, 2, \ldots, s$:

$$\frac{\partial F_i(k_1, \ldots, k_2)}{\partial k_l} = \frac{\partial}{\partial k_l}\left(f(t_n + c_i h, y_n + h \sum_{j=1}^{s} a_{ij} k_j)\right) \tag{1.15}$$

$$= \frac{\partial f}{\partial y}(t_n + c_i h, y_n + h \sum_{j=1}^{s} a_{ij} k_j) \cdot h a_{il} \tag{1.16}$$

Entonces, si $h = 0$, $\partial F_i / \partial k_l = 0, \forall i, l = 1, \ldots, s$, o bien $\partial F / \partial K = 0$, lo que implica que

$$\frac{\partial \phi}{\partial K}(0, K) = I$$

En consecuencia, por el Teorema de la Función Implícita $\phi(h, K) = 0$ define implícitamente a K como función de h, $K = K(h)$, y además, como f es p veces diferenciable con continuidad, $K(h)$ es de clase $C^p([0, \tilde{h}])$. □

1.6. Orden de los métodos Runge-Kutta

Hemos visto que los métodos RK están bien definidos. De forma natural, tenemos que ver si las aproximaciones que dan los métodos realmente aproximan a la solución del PVI (1.1).

Definición 4 *Un método $RK(A, b)$ es de* ***orden p*** ≥ 1 *si para problemas (1.1) con $f \in C^p([t_0, t_f] \times \Omega)$ y lipschitz, se verifica*

$$||y(t_0 + h) - y_1|| \leq K h^{p+1}, \quad h \to 0^+, \tag{1.17}$$

es decir, si la serie de Taylor para la solución exacta $y(t_0 + h)$ y la aproximación y_1 coinciden hasta el término h^p, cuando $h \to 0^+$.

Teorema 5 *Un $RK(A, b)$ tiene al menos orden $1 \Leftrightarrow b^T e = 1$*

Demostración: Para tener orden 1 debe verificarse que

$$||y(t_0 + h) - y_1|| \leq K h^2, \quad h \to 0$$

El desarrollo de Taylor cuando $h \to 0$, es

$$y(t_0 + h) = y(t_0) + h y'(t_0) + \mathcal{O}(h^2) = y_0 + h f(t_0, y_0) + \mathcal{O}(h^2) \tag{1.18}$$

Además la solución en $t_0 + h$ es $y_1 = y_0 + \sum_{i=1}^{s} b_i K_i$ donde las etapas del método son:

$$K_i = f(t_0 + c_i h, y_0 + h \sum_{j=1}^{i-1} a_{ij} K_j) = f(t_0, y_0) + \mathcal{O}(h), \quad i = 1, 2, \ldots, s.$$

Por tanto,

$$y_1 = y_0 + h \sum_{i=1}^{s} b_i (f(t_0, y_0) + \mathcal{O}(h)) = y_0 + h \left(\sum_{i=1}^{s} b_i \right) f(t_0, y_0) + \mathcal{O}(h^2)$$

Comparando con (1.18)

$$y(t_0 + h) - y_1 = h\left(1 - \sum_{i=1}^{s} b_i\right) f(t_0, y_0) + \mathcal{O}(h^2)$$

Luego para tener orden 1 es necesario y suficiente que

$$1 - \sum_{i=1}^{s} b_i = 0 \Rightarrow \sum_{i=1}^{s} b_i = 1.$$

□

Existen formas de saber si un método RK determinado va a converger y con qué orden. Un requerimiento fundamental para poder garantizar dicha convergencia es que el método propague bien los errores a medida que se van calculando los aproximantes, es decir, garantizar que el método es estable. En los siguientes capítulos nos centraremos en el estudio de la estabilidad.

2

Estabilidad de los métodos RK

La estabilidad es un aspecto crucial al trabajar con métodos numéricos para resolver ecuaciones diferenciales. En el caso de los métodos Runge-Kutta, la estabilidad se convierte en un factor determinante para garantizar la fiabilidad de las soluciones aproximadas. Lo primero que se ve cuando se estudia dicha estabilidad es la diferencia que hay entre los métodos explícitos y los implícitos.

Como ejemplo lustrativo de dicha diferencia de estabilidad entre los métodos explícitos e implícitos comenzaremos comparando la aplicación del método de Euler explícito con el implícito sobre un problema que parece sencillo, la ecuación test de Dahlquist. Para el desarrollo de este capítulo hemos extraído principalmente la información de [5, Cap.IV.3].

2.1. Ecuación test de Dahlquist

Consideremos la **ecuación test de Dahlquist**

$$y' = \lambda y, \quad Re\lambda \leq 0 \tag{2.1}$$

El método Euler explícito sobre (2.1) con paso fijo $h = h_j, \forall j$, nos da las aproximaciones

$$y_n = y_{n-1} + hf(t_{n-1}, y_{n-1}) = y_{n-1} + h\lambda y_{n-1} = R(h\lambda)y_{n-1}, \quad R(z) = 1 + z$$

por lo que, iterando,

$$y_n = R(h\lambda)^n y_0$$

Por tanto, si h es tal que $|R(h\lambda)| > 1$, la solución numérica explotará a medida que calculemos más pasos. Así que para que el método dé una aproximación satisfactoria, necesariamente se tiene que tomar h suficientemente pequeño para que

$$|R(h\lambda)| \leq 1 \Leftrightarrow |1 + h\lambda| \leq 1, \tag{2.2}$$

o, lo que es lo mismo, que $h\lambda$ caiga dentro del círculo centrado en -1 de radio 1. Si por ejemplo, $\lambda = -10^3$, esto exige una restricción de paso

$$|1 - \lambda \cdot 10^3| \leq 1 \Leftrightarrow -1 \leq 1 - h \cdot 10^3 \leq 1 \Leftrightarrow h \leq 2 \cdot 10^{-3}$$

Esto supone que, si integramos el problema para $t \in [0, 1]$, el método tendrá que dar al menos 500 pasos para obtener una aproximación razonable.

En cambio, sobre Euler implícito (2.1)

$$y_n = y_{n-1} + hf(t_n, y_n) = y_{n-1} + h\lambda y_n \Rightarrow$$

$$\Rightarrow (1 - h\lambda)y_n = y_{n+1} \Rightarrow y_n = R(h\lambda)y_{n-1}, \quad R(z) = \frac{1}{1-z}$$

En este caso,

$$|R(h\lambda)| = \frac{1}{|1 - h\lambda|} \leq 1 \Leftrightarrow |1 - h\lambda| \geq 1$$

Es decir, $h\lambda$ tiene que estar fuera del círculo de centro 1 y radio 1, lo que se da para cualquier $h \geq 0$, ya que $Re\lambda \leq 0$. Por tanto, no hay ninguna restricción de paso por estabilidad en este método.

Como ya veremos, lo que le pasa al método de Euler explícito se dará también para los métodos RKE en mayor o menor medida. Para verlo, tendremos que estudiar como son las funciones $R(z)$ de los RK.

2.2. Función de estabilidad

Consideremos un método RK(A, b) de s etapas:

$$\begin{cases} k_i = f\left(t_n + c_i h, y_n + h\sum_{j=1}^{s} a_{ij}k_j\right), & 1 \leq i \leq s \\ y_{n+1} = y_n + h\sum_{i=1}^{s} b_i k_i \end{cases} \tag{2.3}$$

Proposición 1 *El método RK(A, b) (2.3) aplicado a la ecuación de Dalhquist (2.1) resulta:*

$$y_{n+1} = R(h\lambda)y_n, \quad R(z) = 1 + zb^T(I - zA)^{-1}e \tag{2.4}$$

donde $e = (1, \ldots, 1)^T$ e I es la matriz identidad de dimensión s.

La función $R(z)$ se llama **función estabilidad** del método (2.3).
Demostración: Como $f(t,y) = \lambda y \in \mathbb{R}$, las etapas k_i de (2.3) resultan

$$k_i = \lambda \left(y_n + h \sum_{j=1}^{s} a_{ij} k_j \right), \quad i = 1, \ldots, s$$

Matricialmente, si $K = (k_1, \ldots, k_s)^T$

$$K = \lambda y_n e + \lambda h A K \Rightarrow (I - \lambda h A) K = \lambda y_n e$$

Luego para h suficientemente pequeño, $I - \lambda h A$ es invertible y

$$K = \lambda (I - \lambda h A)^{-1} y_n e \tag{2.5}$$

Por tanto,

$$y_{n+1} = y_n + h \sum b_i K_i = y_n + h b^T K \Rightarrow$$

$$y_{n+1} = y_n + h b^T \left(\lambda (I - h\lambda A)^{-1} y_n e \right) = y_n + h\lambda \left(b^T (I - h\lambda A)^{-1} e \right) y_n$$

Luego, denotando $z = h\lambda$ nos queda:

$$y_{n+1} = (1 + z b^T (I - zA)^{-1} e) y_n = R(h\lambda) y_n$$

con lo que se demuestra (2.4). □

Proposición 2 *La función de estabilidad (2.4) satisface*

$$R(z) = \frac{det(I - zA + zeb^T)}{det(I - zA)} \tag{2.6}$$

Demostración: Si llamamos

$$g_i = y_n + h \sum_{j=1}^{s} a_{ij} k_j, \quad i = 1 \ldots, s, \quad g = (g_1, \ldots, g_s)^T,$$

usando la misma notación matricial que en la proposición anterior, se tiene que $g = e y_n + h A K$.
Por (2.5) tenemos,

$$g = e y_n + h\lambda A (I - h\lambda A)^{-1} y_n e = (I + h\lambda A (I - h\lambda A)^{-1}) e y_n$$

$$= ((I - h\lambda A) + h\lambda A)(I - h\lambda A)^{-1} e y_n = (I - h\lambda A)^{-1} e y_n$$

Entonces,

$$(I - zA) g = e y_n \tag{2.7}$$

Por otro lado,

$$y_{n+1} = y_n + hb^T K = y_n + h\lambda b^T (I - h\lambda A)^{-1} e y_n =$$
$$= y_n + h\lambda b^T (I - h\lambda A)^{-1} e y_n = y_n + z b^T g$$

Luego,

$$-zb^T g + y_{n+1} = y_n \tag{2.8}$$

Agrupando (2.7) y (2.8) obtenemos un sistema lineal de $s+1$ ecuaciones con incógnitas $\{g, y_{n+1}\}$ de la forma:

$$\begin{pmatrix} I - zA & \mathbf{0} \\ -zb^T & 1 \end{pmatrix} \begin{pmatrix} g \\ y_{n+1} \end{pmatrix} = \begin{pmatrix} ey_n \\ y_n \end{pmatrix} \tag{2.9}$$

Como la matriz de coeficientes del sistema es triangular inferior el determinante será $\det(I - zA)$, por lo que, resolviendo el sistema para y_{n+1}, como $y_n \in \mathbb{R}$,

$$y_{n+1} = \frac{\begin{vmatrix} I - zA & ey_n \\ -zb^T & y_n \end{vmatrix}}{\det(I - zA)} = \frac{\begin{vmatrix} I - zA & e \\ -zb^T & 1 \end{vmatrix}}{\det(I - zA)} \cdot y_n \tag{2.10}$$

Por otro lado, si denotamos $M = I - zA = (m_{ij})_{ij=1}^s$, aplicando operaciones elementales,

$$\begin{vmatrix} I - zA & e \\ -zb^T & 1 \end{vmatrix} = \begin{vmatrix} m_{11} & m_{12} & \dots & m_{1s} & 1 \\ m_{21} & m_{22} & \dots & m_{2s} & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ m_{s1} & m_{s2} & \dots & m_{ss} & 1 \\ -zb_1 & -zb_2 & \dots & -zb_s & 1 \end{vmatrix} = \tag{2.11}$$

$$= \begin{vmatrix} m_{11} + zb_1 & m_{12} + zb_2 & \dots & m_{1s} + zb_s & 0 \\ m_{21} + zb_1 & m_{22} + zb_2 & \dots & m_{ss} + zb_s & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ m_{s1} + zb_1 & m_{s2} + zb_2 & \dots & m_{ss} + zb_s & 0 \\ -zb_1 & -zb_2 & \dots & -zb_s & 1 \end{vmatrix} = \begin{vmatrix} m_{11} + zb_1 & \dots & m_{1s} + zb_s \\ \vdots & \ddots & \vdots \\ m_{s1} + zb_1 & \dots & m_{ss} + zb_s \end{vmatrix} = \det(M + zeb^T) \tag{2.12}$$

y se obtiene (2.6). □

Como consecuencia, desarrollando los determinantes de orden s de (2.6) se tiene que $R(z)$ es una <u>función racional</u> de z, esto es

$$R(z) = \frac{P(z)}{Q(z)} \tag{2.13}$$

cuyos grados, $\deg P = k$ y $\deg Q = j$ respectivamente, son ambos menores o iguales a s.

Además, esta función racional es una aproximación a la función exponencial.

En lo que sigue $P(z)$ y $Q(z)$ denotarán el numerador y el denominador, respectivamente, de una función racional $R(z)$ y (k, j) denotarán sus grados respectivos.

Proposición 3 *Si el método es de orden $p \geq 1$ entonces,*

$$e^z - R(z) = Cz^{p+1} + \mathcal{O}(z^{p+2}) \quad \textit{cuando}\, z \to 0, \quad C \neq 0. \tag{2.14}$$

Como consecuencia, $R(z)$ es una aproximación racional de e^z de orden p.

Demostración: Consideramos el PVI $y' = \lambda y, \quad y(0) = 1$. Su solución exacta en $t_1 = h$ es $y(h) = 1 \cdot e^{\lambda h} = e^z$ y la solución numérica en $t_1 = h$ dada por el método $RK(A, b)$ es $y_1 = R(\lambda h) \cdot y_0 = R(z)$
Si el método es de orden p:

$$y(t_1) - y_1 = \widetilde{C} h^{p+1} + \mathcal{O}(h^{p+2}), \quad h \to 0$$

Por tanto,

$$e^z - R(z) = \widetilde{C} h^{p+1} + \mathcal{O}(h^{p+2}) = Cz^{p+1} + \mathcal{O}(z^{p+2}), \quad z \to 0,$$

para cierto C independiente de z. □

2.3. A-estabilidad

Definición 5 *Un método RK, cuyo* ***dominio de estabilidad***

$$S = \{z \in \mathbb{C} / |R(z)| \leq 1\} \tag{2.15}$$

satisface

$$\mathbb{C}^- := \{z : \, Rez \leq 0\} \subset S$$

se denomina ***A-estable****.*

Proposición 4

$$\textit{un RK(A,b) es A-estable} \Leftrightarrow \begin{cases} R(z) \textit{ es analítica en} \, Re\, z < 0 \\ |R(iy)| \leq 1, \forall y \in \mathbb{R} \end{cases}$$

Demostración: Si existe $\widetilde{z} \in \mathbb{C}^-$ con $Q(\widetilde{z}) = 0$, $R(z)$ no está acotada. Por tanto para ser A-estable $R(z)$ tiene que ser analítica en $\mathbb{C}^-$. Además, por el principio del máximo, si $R(z)$ es analítica en $\mathbb{C}^-$, su máximo se alcanza en su frontera, es decir, el eje imaginario. □

Proposición 5 *Ningún método RK explícito puede ser A-estable.*

Demostración: Sabemos que la matriz A en los métodos explícitos es triangular inferior estricta, por lo tanto $\det(I - zA) = 1$, entonces, por (2.6) necesariamente $R(z)$ es un polinomio de grado $\leq s$. Por eso cuando $z \to \infty$ no puede estar acotado, así que no puede ser A-estable. □

2.4. I-estabilidad

La condición $|R(iy)| \leq 1$, $\forall y \in \mathbb{R}$, vista en la Proposición (4), suele llamarse **I-estabilidad**.

Estudiando mejor esta condición, tenemos

$$|R(iy)|^2 \leq 1 \Leftrightarrow \frac{|P(iy)|^2}{|Q(iy)|^2} \leq 1 \Leftrightarrow |P(iy)|^2 - |Q(iy)|^2 \leq 0 \Leftrightarrow |Q(iy)|^2 - |P(iy)|^2 \geq 0$$

Por tanto, si denotamos

$$E(y) = |Q(iy)|^2 - |P(iy)|^2$$

la I-estabilidad es equivalente a que

$$E(y) \geq 0, \forall y \in \mathbb{R}. \tag{2.16}$$

Proposición 6 *E(y) es un polinomio de potencias pares con*

$$degE \leq 2\,\text{máx}(degP, degQ).$$

Si R(z) es una aproximación del orden de p a la exponencial, entonces

$$E(y) = \mathcal{O}(y^{p+1}), \quad y \to 0$$

Demostración:

$$E(y) = Q(iy)Q(-iy) - P(iy)P(-iy) = E(-y), \quad \forall y \in \mathbb{R},$$

por lo que $E(y)$ es un polinomio par, es decir, solo tiene potencias pares. Por otro lado, por (2.14) para $z = iy$ y usando que $|e^{iy}| = 1$,

$$|e^{iy}| - \frac{|P(iy)|}{|Q(iy)|} = 1 - \frac{|P(iy)|}{|Q(iy)|} = \mathcal{O}(y^{p+1})$$

de donde se deduce que

$$|Q(iy)| - |P(iy)| = \mathcal{O}(y^{p+1})$$

En consecuencia,

$$E(y) = |Q(iy)|^2 - |P(iy)|^2 = (|Q(iy)| + |P(iy)|) \cdot (|Q(iy)| - |P(iy)|) = \mathcal{O}(y^{p+1})$$

ya que $(|Q(iy)| + |P(iy)|)$ está acotado. □

La siguiente propiedad, que aplicaremos en el siguiente capítulo, se deduce de lo anterior.

Proposición 7 *Una función racional $R(z)$ (2.13) de orden $p \geq 2j - 2$ es I-estable si y solo si $|R(\infty)| \leq 1$.*

Demostración: Por la Proposición 6, sabemos que

$$E(y) = a_{2l}y^{2l} + a_{2l-2}y^{2l-2} + \cdots + a_2y^2 + a_0 \geq 0 \tag{2.17}$$

con $l = \text{máx}\{j, k\}$ y, como tiene orden p, $E(y) = \mathcal{O}(y^{p+1})$, $y \to 0$, por lo que tendrá que ser

$$E(y) = a_{2l}y^{2l} + \cdots + a_{p+1}y^{p+1} \geq 0,\ y \to 0 \tag{2.18}$$

" $\Rightarrow$ " Como $R(z)$ es I-estable, entonces si $y \to \infty$, $E(y) \simeq a_{2l}y^{2l} \geq 0$, luego $a_{2l} \geq 0$. Por tanto,

$$\begin{aligned} &|Q(iy)|^2 - |P(iy)|^2 \simeq a_{2l}y^{2l} \geq 0 \Rightarrow |Q(iy)|^2 \simeq |P(iy)|^2 + a_{2l}y^{2l} \geq 0 \Rightarrow \\ &\Rightarrow |R(iy)|^2 = \frac{|P(iy)|^2}{|Q(iy)|^2} \simeq \frac{|P(iy)|^2}{|P(iy)|^2 + a_{2l}y^{2l}} \leq 1 \Rightarrow |R(\infty)| \leq 1 \end{aligned} \tag{2.19}$$

" $\Leftarrow$ " Tenemos por hipótesis que $|R(\infty)| \leq 1$, entonces necesariamente, el grado del numerador ha de ser menor o igual que el del denominador, esto es, $k \leq j$. Por tanto, se obtiene (2.18) con $l = j$. Además, cuando $y \to \infty$, $|R(iy)| \leq 1$ entonces $E(y) \geq 0$, $y \to \infty$. Por otra parte,

$$p \geq 2j - 2 \Rightarrow p + 1 \geq 2j - 1$$

luego en (2.18) necesariamente tiene que ser

$$E(y) = a_{2j}y^{2j},\ \forall y \in \mathbb{R}$$

y $E(y) \geq 0$ cuando $y \to \infty$. Entonces, $a_{2j} \geq 0$ y $E(y) \geq 0, \forall y \in \mathbb{R}$. Por lo tanto, R es I-estable. □

2.5. L-estabilidad

Definición 6 *Un método se dice L-estable si es A-estable y*

$$\lim_{z\to\infty} R(z) = 0$$

Proposición 8 *Si un método implícito Runge-Kutta con A no singular satisface una de las siguientes condiciones:*

$$\begin{array}{lll} a_{sj} = b_j & j = 1,\dots,s & \\ a_{i1} = b_1 & i = 1,\dots,s & (b_1 \neq 0) \end{array} \tag{2.20}$$

entonces $R(\infty) = 0$.

Demostración: Sabemos por (2.4) que $R(z)$ es

$$R(z) = 1 + zb^T(I - zA)^{-1}e = 1 + b^T\left(z^{-1}(I - zA)\right)^{-1} e = 1 + b^T(z^{-1}I - A)^{-1}e$$

Luego, si tomamos $z \to \infty$ nos queda que

$$R(\infty) = 1 + b^T(-A)^{-1}e = 1 - b^T A^{-1}e$$

Por tanto, si comprobamos cada una de las condiciones (2.20) nos queda:

- Si $a_{sj} = b_j$, $j = 1,\dots,s$, y denotamos la última fila de A como $A_s^T = e_s^T A$, $e_s^T = (0,\dots,0,1)$, tenemos que $e_s^T A = b^T$, y, por lo tanto,

$$R(\infty) = 1 - b^T A^{-1}e = 1 - e_s^T A A^{-1}e = 1 - e_s^T e = 0$$

- Si $a_{i1} = b_1$, $i = 1,\dots,s$, $b_1 \neq 0$ vectorialmente podemos denotarlo como que $Ae_1 = b_1 e$ donde $e_1 = (1,0,\dots,0)^T$. Por tanto,

$$e = \frac{1}{b_1} Ae_1$$

  Por consiguiente,

$$R(\infty) = 1 - b^T A^{-1}e = 1 - b^T A^{-1}(\frac{1}{b_1}Ae_1) =$$

$$= 1 - \frac{1}{b_1} b^T A^{-1} A e_1 = 1 - \frac{1}{b_1} b^T e_1 = 0$$

Obsérvese que si $a_{i1} = b_1 = 0$, $i = 1,\dots,s$, realmente la primera etapa sería superflua y el método RK sería reducible.

□

En consecuencia, los métodos A-estables que cumplen alguna de las condiciones (2.20) son L-estables.

2.6. Ejemplos

A continuación, aplicaremos lo visto en este capítulo mediante una serie de ejemplos.

(i) θ-método [4, p.204]

$$\begin{array}{c|cc} 0 & 0 & 0 \\ 1 & 1-\theta & \theta \\ \hline & 1-\theta & \theta \end{array} \qquad R(z) = \frac{1+z(1-\theta)}{1-z\theta},\ \theta \in [0,1] \tag{2.21}$$

Si $\theta = 0$ tenemos el Euler explícito, si $\theta = 1$ Euler implícito y si $\theta = 1/2$ tenemos la regla trapezoidal. En este caso, $R(z)$ es analítica en $\mathbb{C}^-$ dado que para $\theta > 0$, $z = 1/\theta \notin \mathbb{C}^-$ y si $\theta = 0$, $R(z)$ es un polinomio. Así,

$$E(y) = |1+\theta iy|^2 - |1+(1-\theta)iy|^2 = (-1+2\theta)y^2 \geq 0 \Leftrightarrow \theta \geq \frac{1}{2}$$

Luego, tenemos que el método de Euler implícito y la regla trapezoidal sí son A-estables pero el Euler explícito no lo es. Entre los dos implícitos el único L-estable es el Euler implícito ya que

$$\lim_{z\to\infty} \frac{1+z(1-\theta)}{1-z\theta} = 0 \Leftrightarrow \theta = 1.$$

Obsérvese que $R(z)$ es la misma para la regla trapezoidal que para la regla del punto medio implícito.

(ii) Método de Hammer-Hollingsworth [4, p.207] de orden 4:

$$\begin{array}{c|cc} \frac{1}{2}-\frac{\sqrt{3}}{6} & \frac{1}{4} & \frac{1}{4}-\frac{\sqrt{3}}{6} \\ \frac{1}{2}+\frac{\sqrt{3}}{6} & \frac{1}{4}-\frac{\sqrt{3}}{6} & \frac{1}{4} \\ \hline & 1/2 & 1/2 \end{array} \qquad R(z) = \frac{1+z/2+z^2/12}{1-z/2+z^2/12} \tag{2.22}$$

En este caso,

$$E(y) = \left|1 - \frac{iy}{2} + \frac{(iy)^2}{12}\right|^2 - \left|1 + \frac{iy}{2} + \frac{(iy)^2}{12}\right|^2 =$$

$$= \left|\left(1 - \frac{y^2}{12}\right) - \frac{iy}{2}\right|^2 - \left|\left(1 - \frac{y^2}{12}\right) + \frac{iy}{2}\right|^2 =$$

$$= \left(1 - \frac{y^2}{12}\right)^2 + \left(-\frac{y}{2}\right)^2 - \left(1 - \frac{y^2}{12}\right)^2 - \left(\frac{y}{2}\right)^2 = 0$$

Además, los ceros de $Q(z)$ son $z = 3 \pm \sqrt{3}i$, luego es analítica en $\mathbb{C}^-$, por lo tanto es A-estable. Como $\lim_{z\to\infty} R(z) = 1$ no es L-estable.

(iii) SDIRK orden 3 [4, p.207]

$$\begin{array}{c|cc} \gamma & \gamma & 0 \\ 1-\gamma & 1-2\gamma & \gamma \\ \hline & 1/2 & 1/2 \end{array} \qquad R(z) = \frac{1 + z(1-2\gamma) + z^2(1/2 - 2\gamma + \gamma^2)}{(1-\gamma z)^2} \qquad \gamma = \frac{3 \pm \sqrt{3}}{6} \tag{2.23}$$

En este caso

$$E(y) = (\gamma - 1/2)^2(4\gamma - 1)y^4$$

Por lo tanto, cuando $\gamma \geq 1/4$ el método es A-estable ya que además $R(z)$ es analítica en $\mathbb{C}^-$ dado que el cero de $Q(z)$ es $z = 1/\gamma > 0$. Esto implica que será A-estable cuando $\gamma = (3 + \sqrt{3})/6 \simeq 0{,}7887$ pero no cuando $\gamma = (3 - \sqrt{3})/6 \simeq 0{,}2113$. Por otro lado, como $\lim_{z\to\infty} R(z) \neq 0$ no es L-estable.

(iv) Método de Kuntzmann y Butcher [4, p.209] de orden 6

$$\begin{array}{c|ccc} \frac{1}{2} - \frac{\sqrt{15}}{10} & \frac{5}{36} & \frac{2}{9} - \frac{\sqrt{15}}{15} & \frac{5}{36} - \frac{\sqrt{15}}{30} \\ \frac{1}{2} & \frac{5}{36} + \frac{\sqrt{15}}{24} & \frac{2}{9} & \frac{5}{36} - \frac{\sqrt{15}}{24} \\ \frac{1}{2} + \frac{\sqrt{15}}{10} & \frac{5}{36} + \frac{\sqrt{15}}{30} & \frac{2}{9} + \frac{\sqrt{15}}{15} & \frac{5}{36} \\ \hline & 5/18 & 4/9 & 5/18 \end{array} \qquad R(z) = \frac{1 + z/2 + z^2/10 + z^3/120}{1 - z/2 + z^2/10 - z^3/120} \tag{2.24}$$

En este ejemplo ya vemos que desde que los grados del numerador y denominador aumentan garantizar la A-estabilidad es mucho más complicado. Por eso se desarrolló una teoría más amplia que intenta estudiar cómo deben ser las aproximaciones racionales a e^z para que sean A-estables, la teoría de las estrellas de orden que veremos en el próximo capítulo.

3

Estrellas de orden (Order Stars)

Como hemos visto en el capítulo anterior, las funciones de estabilidad $R(z)$ de los métodos RK son aproximantes racionales a la exponencial, que aproximan su desarrollo de Taylor al menos hasta el orden p del método. Si queremos obtener aproximaciones a e^z con orden alto, lo que nos permitirá deducir métodos RK con órdenes altos, los grados (k, j) de $R(z)$ tendrán que tener un tamaño relativamente grande, lo que a su vez hace que sea muy difícil saber si es A-estable o no.

Las estrellas de orden van a ayudar a deducir la A-estabilidad de un aproximante racional a e^z en general y, en particular, determinan la A-estabilidad de la clase de aproximantes con los que se puede alcanzar el mayor orden posible, los aproximantes de Padé.

Hemos seguido principalmente la teoría de las estrellas de orden desarrollada en el [5, Cap.IV.4].

3.1. Aproximación de Padé a la función exponencial

Los aproximantes de Padé a la exponencial aportan el orden máximo que se puede obtener con una función racional con grados (k, j), del numerador y denominador respectivamente, prefijados.

Definición 7 *[2] Dada una función f y dos enteros $m \geq 0$ y $n \geq 0$,* ***el aproximante de Padé*** *de orden $[m/n]$ es la función racional*

$$R(z) = \frac{a_0 + a_1 z + a_2 z^2 + \ldots + a_m z^m}{1 + b_1 z + b_2 z^2 + \ldots + b_n z^n},$$

que verifica

$$f^{(l)}(0) = R^{(l)}(0), \quad k = 0, 1, \ldots, m + n \tag{3.1}$$

Si hacemos el desarrollo en serie de Maclaurin de $R(z)$, (3.1) es equivalente a

$$f(z) - R(z) = \mathcal{O}(z^{m+n+1}), \quad z \to 0$$

y es el mayor orden de aproximación posible. Además, el aproximante de Padé es único para m y n.

En lo que sigue llamaremos **[k/j]-Padé** al aproximante de Padé de orden $[k/j]$.

Teorema 6 *Dado dos enteros $k \geq 0$ y $j \geq 0$, la aproximación $[k, j]$-Padé a e^z está dada por*

$$R_{kj}(z) = \frac{P_k(z)}{Q_j(z)}$$

donde

$$P_k(z) = 1 + \sum_{l=1}^{k} \frac{(k)_l}{(j+k)_l} \cdot \frac{z^l}{l!}, \qquad Q_j(z) = 1 + \sum_{l=1}^{j} (-1)^l \cdot \frac{(j)_l}{(k+j)_l} \cdot \frac{z^l}{l!} \tag{3.2}$$

denotando $(k)_l = k(k-1)\cdots(k-(l-1))$.

Ejemplo 1 Veamos algunos ejemplos de aproximantes de Padé.

- [1/0]-Padé:

  $$R_{10}(z) = \frac{1+z}{1}$$

  Vemos que este aproximante coincide con tiene la función de estabilidad del método de Euler (1.8).

- [0/1]-Padé:

  $$R_{01}(z) = \frac{1}{1-z}$$

  En este caso, vemos que esta es la función de estabilidad del método de Euler implícito (1.11).

- [1/1]-Padé:

  $$R_{11}(z) = \frac{1+\frac{1}{2}z}{1-\frac{1}{2}z}$$

  Este coincide con la función la función de estabilidad de la regla implícita del punto medio (1.12) y de la regla trapezoidal (1.13).

- [3/3]-Padé

$$R_{33}(z) = \frac{1 + \frac{1}{2}z + \frac{1}{5}\frac{z^2}{2!} + \frac{1}{20}\frac{z^3}{3!}}{1 - \frac{1}{2}z + \frac{1}{5}\frac{z^2}{2!} - \frac{1}{20}\frac{z^3}{3!}}$$

En este caso, coincide con la función de estabilidad del método de Kuntzmann y Butcher de orden 6 (2.24).

3.2. Estrellas de orden (Order Stars)

Las estrellas de orden aparecen al comparar un aproximante racional a e^z con la propia e^z.

Sea $R(z) = P(z)/Q(z)$ un aproximante a e^z de orden $p \geq 1$ con deg $P = k \geq 0$, deg $Q = j \geq 0$ y coeficientes reales.

Definición 8 *El conjunto*

$$A = \{z \in \mathbb{C} : |R(z)| > |e^z|\} = \{z \in \mathbb{C} : |q(z)| > 1\}, \quad q(z) = R(z)/e^z \tag{3.3}$$

se denomina ***estrella de orden (order star)*** *de R.*

Proposición 9 *La estrella de orden es simétrica respecto al eje real.*

Demostración: Veamos que $\forall z \in A$, entonces $\bar{z} \in A$. Al ser $R(z)$ un cociente de polinomios con coeficientes reales, $\overline{R(z)} = R(\bar{z})$, $\forall z \in A$.
Por tanto, $|R(z)| = |\overline{R(z)}| = |R(\bar{z})|$, así

$$|R(\bar{z})| \geq |e^z| = e^{Re\, z} = e^{Re\bar{z}} = |e^{\bar{z}}|,$$

por lo que $\bar{z} \in A$. □

Proposición 10 $A \cap i\mathbb{R} = S^c \cap i\mathbb{R}$.

Demostración: Para todo $z \in A \cap i\mathbb{R}$, $z = iy$ para cierto $y \in \mathbb{R}$ y $|R(z)| > |e^z| = |e^{iy}| = 1 \Leftrightarrow z \notin S \Leftrightarrow z \in S^c \cap i\mathbb{R}$. □

Lema 1 *$R(z)$ es I-estable si y solo si*

$$A \cap i\mathbb{R} = \emptyset \tag{3.4}$$

Demostración: $R(z)$ es I-estable si y solo si $i\mathbb{R} \subset S$, lo que es equivalente a que $S^c \cap i\mathbb{R} = \emptyset = A \cap i\mathbb{R}$, por la Proposición 10. □

Lema 2 *$R(z)$ es A-estable si y solo si $A \cap i\mathbb{R} = \emptyset$ y todos los polos de $R(z)$ están en el semiplano positivo $\mathbb{C}^+$.*

Demostración: Sabemos por la Proposición 4 que $R(z)$ es A-estable sii es analítica en $\mathbb{C}^-$ y $|R(iy)| \leq 1, \ \ \forall y \in \mathbb{R}$.

Por tanto, si $R(z)$ es A-estable, todos sus polos (o ceros de $Q(z)$) están en $\mathbb{C}^+$. Además, $|R(iy)| \leq 1, \forall y \in \mathbb{R}, i\mathbb{R} \subset S$ por lo que $S^c \cap i\mathbb{R} = \emptyset = A \cap i\mathbb{R}$ por la proposición 10.

Por otra parte, si todos los polos de $R(z)$ están en $\mathbb{C}^+$ entonces $R(z)$ es analítica en $\mathbb{C}^-$. Además, $A \cap i\mathbb{R} = S^c \cap i\mathbb{R} = \emptyset$ implica que $i\mathbb{R} \subset S$, es decir, $|R(iy)| \leq 1, \forall y \in \mathbb{R}$. Por tanto, por la Proposición 4, se tiene la A-estabilidad. $\square$

Conocer cómo son las estrellas de orden asociadas a cada aproximación $R(z)$ nos dará información para determinar si dicha $R(z)$ es A-estable o no. Como estamos tratando con funciones racionales, hay dos casos de $\mathbb{C}$ en los que es fundamental saber cómo son dichas estrellas: cuando estamos cerca del origen y cuando $z \to \infty$.

3.3. Order Star cuando $z \to 0$

Para deducir cómo es A cerca del origen, es necesario que previamente veamos ciertas propiedades.

Lema 3 *(a) Dada una constante $0 < k < 1$;*

$$\cos x > k \Leftrightarrow x \in D_\gamma^+ := \bigcup_{k \in \mathbb{Z}} \left(-\gamma + 2k\pi, \gamma + 2k\pi\right), \ \gamma = \arccos(k) \in \left(0, \frac{\pi}{2}\right].$$

(b) Si $-1 < k \leq 0$,

$$\cos x < k \Leftrightarrow x \in D_\gamma^- := \bigcup_{k \in \mathbb{Z}} \left(2k\pi + \gamma, (2k+2)\pi - \gamma\right), \ \gamma = \arccos(k) \in \left[\frac{\pi}{2}, \pi\right].$$

Demostración: Este lema se puede ver facilmente de forma gráfica. En la figura 3.1 se observa que si elegimos un $k \in (0,1)$ la recta $y = k$ corta a la función $y = \cos x$ en dos puntos γ y $-\gamma$ donde $\gamma = \arccos k$ de tal forma que la función siempre queda por encima de la recta $y = k$.

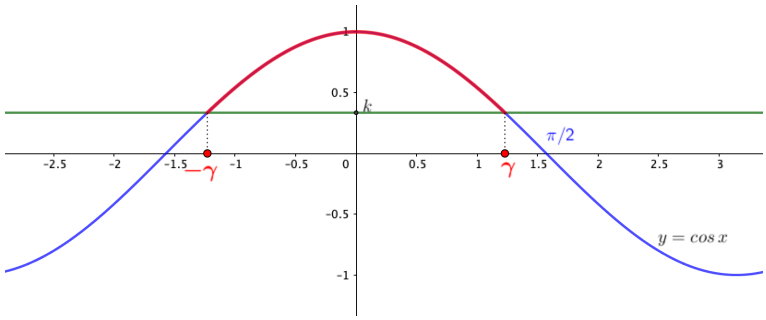


Figura 3.1: Área donde $\cos x > k$ en rojo

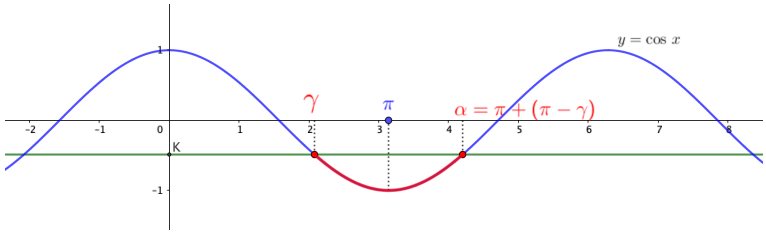


Figura 3.2: Área donde $\cos x < k$ en rojo

En cambio, si tomamos $k \in (-1, 0)$ en la Figura 3.2 corta a $y = \cos x$ en dos puntos $\gamma = \arccos k \in [\pi/2, \pi]$ y $\alpha = \pi + (\pi - \gamma)$ de tal forma que el $\cos x < k$, para todo $x \in (\gamma, 2\pi - \gamma)$. Por la periodicidad del $\cos x$, esto que ocurre cuando $x \in [0, 2\pi]$ se repite en todos los intervalos $[2k\pi, (2k+2)\pi]$, $k \in \mathbb{Z}$, de donde se tiene lo que dice el lema. $\square$

Observación 2 De lo anterior se deduce también que:

- Si $0 < k < 1$: $\cos x = k \Leftrightarrow x \in \partial D_\gamma^+$ y $\cos x < k \Leftrightarrow x \in \mathbb{R} - \overline{(D_\gamma^+)}$
- Si $-1 < k < 0$: $\cos x = k \Leftrightarrow x \in \partial D_\gamma^-$, $\cos x > k \Leftrightarrow x \in \mathbb{R} - \overline{(D_\gamma^-)}$

Lema 4 *Dado $p \in \mathbb{N}$ y una constante $C \in \mathbb{R}$, $C \neq 0$, consideramos la función $f(z) = 1 - Cz^{p+1}$ definida para $z = re^{i\theta} \in B_M = \{z \in \mathbb{C}/|z| < M\}$ donde se toma $M > 0$ de forma que $0 \leq \frac{|C|}{2} r^{p+1} < 1$ para todo $0 \leq r < M$, en cuyo caso consideramos los siguientes subconjuntos de $\mathbb{R}$:*

(a) *Si $C > 0$:* $D_\gamma := \bigcup_{k \in \mathbb{Z}} \left(-\frac{\gamma}{p+1} + \frac{2k\pi}{p+1}, \frac{\gamma}{p+1} + \frac{2k\pi}{p+1} \right)$

donde $\gamma = \arccos\left(\frac{Cr^{p+1}}{2}\right) \in \left[0, \frac{\pi}{2}\right]$.

(b) *Si* $C < 0$*:* $D_\gamma := \bigcup_{k\in\mathbb{Z}} \left(\frac{\gamma}{p+1} + \frac{2k\pi}{p+1}, \frac{2\pi-\gamma}{p+1} + \frac{2k\pi}{p+1} \right)$

donde $\gamma = \arccos\left(\frac{Cr^{p+1}}{2}\right) \in \left[\frac{\pi}{2}, \pi\right]$.

Entonces, se tiene
(i) Si $\theta \in D_\gamma \Rightarrow |f(z)| < 1$
(ii) Si $\theta \in \partial D_\gamma \Rightarrow |f(z)| = 1$
(iii) Si $\theta \in \mathbb{R} - (D_\gamma \cup \partial D_\gamma) \Rightarrow |f(z)| > 1$

Demostración: Primero veamos cómo es el módulo de $f(z) = 1 - Cz^{p+1}$, $z = re^{i\theta}$:

$$|f(z)|^2 = |1 - Cr^{p+1}e^{i(p+1)\theta}|^2 = |1 - Cr^{p+1}(\cos(p+1)\theta + i\sin(p+1)\theta)|^2 =$$

$$= (1 - Cr^{p+1}\cos(p+1)\theta)^2 + (Cr^{p+1}\sin(p+1)\theta)^2 =$$

$$= 1 - 2Cr^{p+1}\cos(p+1)\theta + C^2r^{2(p+1)}\cos^2(p+1)\theta + C^2r^{2(p+1)}\sin^2(p+1)\theta$$

Sacando factor común nos queda:

$$|f(z)|^2 = 1 - 2Cr^{p+1}\cos(p+1)\theta + C^2r^{2(p+1)}$$

Por tanto,

$$\begin{aligned} |f(z)| < 1 \Leftrightarrow & 1 - 2Cr^{p+1}\cos(p+1)\theta + C^2r^{2(p+1)} < 1 \Leftrightarrow \\ \Leftrightarrow & -2Cr^{p+1}\cos(p+1)\theta + C^2r^{2(p+1)} < 0 \end{aligned} \tag{3.5}$$

(a) Si $C > 0$, (3.5) es equivalente a

$$Cr^{p+1} - 2r^{p+1}\cos(p+1)\theta < 0 \quad \text{ó} \quad \cos(p+1)\theta > \frac{Cr^{p+1}}{2}$$

Aplicando el apartado (i) del Lema 3 para $x = (p+1)\theta$ y $k = Cr^{p+1}/2$, esto se da si y solo si $(p+1)\theta \in \cup_{k\in\mathbb{Z}}(-\gamma + 2k\pi, \gamma + 2k\pi)$ donde $\gamma = \arccos(Cr^{p+1}/2)$, o lo que es lo mismo, sii $\theta \in D_\gamma$.
De la Observación 3.3 se deducen los puntos (ii) y (iii) del lema.

(b) Si $C < 0$, (3.5) es equivalente a que

$$\cos(p+1)\theta < \frac{Cr^{p+1}}{2}$$

y aplicando el apartado (b) del Lema 3 igual que antes se concluye la demostración.

□

Lema 5 *Suponemos que tenemos las mismas hipótesis del Lema 4 y* $\gamma = \arccos(Cr^{p+1}/2)$, *entonces:*

(a) Si $C > 0$ *y* $r \to 0^+$, *se tiene que* $D_\gamma \subset D^+$ *y*

$$D_\gamma \to D^+ := \bigcup_{k\in\mathbb{Z}} \left(\frac{-\pi/2}{p+1} + \frac{2k\pi}{p+1}, \frac{\pi/2}{p+1} + \frac{2k\pi}{p+1} \right).$$

(b) Si $C < 0$ *y* $r \to 0^+$, *se tiene que* $D_\gamma \subset D^-$ *y*

$$D_\gamma \to D^- := \bigcup_{k\in\mathbb{Z}} \left(\frac{\pi/2}{p+1} + \frac{2k\pi}{p+1}, \frac{3\pi/2}{p+1} + \frac{2k\pi}{p+1} \right).$$

Demostración:

(a) Como $C > 0$, $r \to 0^+$, $\gamma = \arccos(Cr^{p+1}/2) \to \pi/2$ con $\gamma < \pi/2$ por lo que $\forall \varepsilon > 0, \exists r_0 > 0$ tal que $\forall r : 0 < r < r_0$, $\gamma = \arccos(Cr^{p+1}/2) < \pi/2$ y $\pi/2 - \gamma < \varepsilon$, lo que, aplicando el apartado (a) del Lema 4, demuestra que $D_\gamma \subset D^+$ y $D_\gamma \to D^+$.

(b) Se obtiene de forma similar cuando $C < 0$ teniendo en cuenta que ahora $\gamma = \arccos(Cr^{p+1}/2) \to \pi/2$ con $\gamma > \pi/2$.

□

Obsérvese que, independientemente de que $C > 0$ o bien $C < 0$ la recta real se puede dividir como se ve en la Figura 3.3

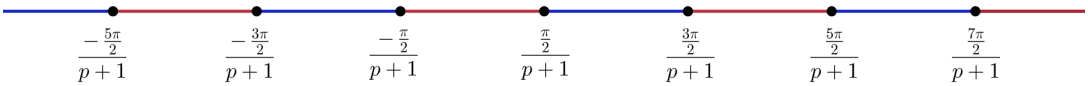


Figura 3.3: Subintervalos de D^+ en rojo y de D^- en azul.

Lema 6 *Si* $R(z)$ *es una aproximación a* e^z *de orden* p, *es decir,*

$$e^z - R(z) = Cz^{p+1} + \mathcal{O}(z^{p+2}), \quad z \to 0, \quad C \neq 0, \tag{3.6}$$

en un entorno de $z = 0$ *el conjunto* A *consiste en los puntos de* $p+1$ *sectores circulares de igual amplitud* $\pi/(p+1)$ *alternados con* $p+1$ *sectores circulares de la misma amplitud de* A^c.

Demostración: Dividiendo (3.6) por e^z,obtenemos

$$\frac{R(z)}{e^z} = 1 - Cz^{p+1} + \mathcal{O}(z^{p+2}) = f(z) + \mathcal{O}(z^{p+2}), \quad z \to 0^+ \tag{3.7}$$

donde $f(z) = 1 - Cz^{p+1}$. Sea $z = re^{i\theta}$ con r suficientemente cerca de 0 para que se verifiquen las hipótesis de los Lemas 4 y 5.
Al igual que en ambos Lemas, tenemos que distinguir si $C > 0$ o bien $C < 0$.

(1) Si $C > 0$

- Si $\theta \in D^+$, por el Lema 5(a) y r suficientemente pequeño se tiene que $\theta \in D_\gamma \subset D^+$. Aplicando el Lema 4(a)-(i), se tiene que $|f(z)| < 1$. Por (3.7), se tiene que $|q(z)| = |R(z)/e^z| < 1$ y, por tanto, $z \notin A$.

- Si $\theta \in D^-$, entonces $\theta \in \mathbb{R} - (D_\gamma \cup \partial D_\gamma)$ para todo $\gamma = \arccos(Cr^{p+1}/2)$, por lo que, por el Lema 4(a)-(iii) se tiene que $|f(z)| > 1$ y, por (3.7), $|q(z)| > 1$ y $z \in A$.

- Si $\theta \in \partial D^+$ no podemos afirmar nada, pero lo anterior nos dice que si consideramos los interiores de los sectores circulares $\{z = re^{i\theta}/\theta \in D^+\}$ están contenidos en A^c y se alternan con los sectores circulares $\{z = re^{i\theta}/\theta \in D^-\}$ que están en A.

(2) Si $C < 0$, aplicando los apartados (b) de los Lemas 4 y 5 obtenemos de forma similar los mismo sectores circulares que en el caso $C > 0$ con la diferencia de que altera el interior de los sectores $\{z = re^{i\theta}/\theta \in D^-\}$ están en A^c y los sectores $\{z = re^{i\theta}/\theta \in D^+\}$ están en A.

Gráficamente

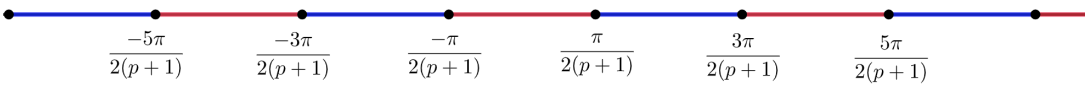


Figura 3.4: Subintervalos de D^+ en rojo y los de D^- en azul.

□

Observación 3 Estudiemos con un poco más de detalle cómo son estos sectores de la estrella de orden cuando $z \to 0$.
Por simetría, el semieje real positivo tiene que ser la bisectriz de un sector y el semieje real negativo la bisectriz de otro.
Por tanto, en el primer cuadrante la primera de las aristas de la "estrella"será la recta

$$\theta = \theta_0 := \frac{\pi}{2(p+1)}$$

Las siguientes aristas serán las rectas

$$\theta = \theta_l := \theta_0 + l\frac{\pi}{p+1}, \quad l = 1, 2, \ldots, 2(p+1) - 1.$$

Teniendo en cuenta

$$\theta_0 + l\frac{\pi}{p+1} \leq \frac{\pi}{2} \Leftrightarrow \frac{\pi}{2}\left(\frac{1+2l}{p+1}\right) \leq \frac{\pi}{2} \Leftrightarrow \frac{1+2l}{p+1} \leq 1 \Leftrightarrow l \leq \frac{p}{2}$$

se tiene que

- **Si p es par**: el semieje imaginario positivo es una arista de la estrella (l=p/2) y, por simetría, el semieje imaginario negativo también lo será, por lo que habrá 4 sectores cuyo borde es el eje imaginario. Por tanto, tenemos que hay al menos

$$\begin{gathered}\frac{p}{2} - 1 \text{ sectores de } A \text{ y } \frac{p}{2} - 1 \text{ sectores de } A^c \\ \text{completamente contenidos en } \mathbb{C}^+.\end{gathered} \tag{3.8}$$

  y que, al menos,

$$\begin{gathered}\frac{p}{2} \text{ sectores de } A \text{ y } \frac{p}{2} \text{ sectores de } A^c \\ \text{arranquen en } \mathbb{C}^+.\end{gathered} \tag{3.9}$$

- **Si p es impar**: el semieje imaginario positivo es una bisectriz de un sector de la estrella y el semieje imaginario negativo la bisectriz de otro. Por simetría, ambos pertenecen a A o ambos pertenecen a A^c.
  Por tanto, tendríamos que al menos hay

$$\begin{gathered}\frac{p-1}{2} \text{ sectores de } A \text{ y } \frac{p-1}{2} \text{ sectores de } A^c \\ \text{completamente contenidos en } \mathbb{C}^+.\end{gathered} \tag{3.10}$$

  y, además, hay, como mínimo,

$$\begin{gathered}\frac{p+1}{2} \text{ sectores de } A \text{ y } \frac{p+1}{2} \text{ sectores de } A^c \\ \text{que comienzan en } \mathbb{C}^+.\end{gathered} \tag{3.11}$$

Luego, de forma general tenemos que al menos hay

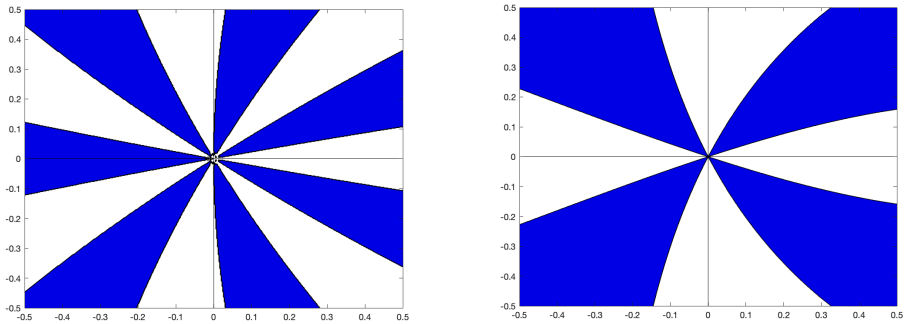
$$\begin{gathered}E\left[\frac{p+1}{2}\right] - 1 \text{ sectores de } A \text{ y } E\left[\frac{p+1}{2}\right] - 1 \text{ sectores de } A^c \\ \text{completamente contenidos en } \mathbb{C}^+.\end{gathered} \tag{3.12}$$

Además hay, al menos,

$$\begin{gathered}E\left[\frac{p+1}{2}\right] \text{ sectores de } A \text{ y } E\left[\frac{p+1}{2}\right] \text{ sectores de } A^c \\ \text{que comienzan en } \mathbb{C}^+.\end{gathered} \tag{3.13}$$

Todo lo anterior se da de la misma forma para $\mathbb{C}^-$.

Por ejemplo, en la Figura 3.5 podemos ver las estrellas de orden del [0/6]-Padé y de la función de estabilidad del método SDIRK de orden 3 (2.6) con $\gamma = (3+\sqrt{3})/6$ cerca del origen.



(a) [0/6]-Padé orden 6 (b) SDIRK orden 3, $\gamma = (3+\sqrt{3})/6$

Figura 3.5: Estrellas de orden cerca del origen (A en azul y A^c en blanco).

3.4. Order Star cuando $z \to \infty$

Como ya dijimos, otro estudio relevante para entender las estrellas de orden es saber qué ocurre cuando $z \to \infty$.

Lema 7 *Sea $z = re^{i\theta}$, entonces*

(i) $\forall \theta \in \left(\frac{\pi}{2}, \frac{3\pi}{2}\right), \exists K_1 > 0, \forall r > K_1$ *tal que* $z \in A$.

(ii) $\forall \theta \in \left(\frac{-\pi}{2}, \frac{\pi}{2}\right), \exists K_2 > 0, \forall r > K_2$ *tal que* $z \notin A$.

Demostración:

(i) $\forall \theta \in \left(\frac{\pi}{2}, \frac{3\pi}{2}\right)$, $\cos\theta < 0$, entonces $e^{-r\cos\theta} \to \infty$, $r \to \infty$.

Por tanto,

$$\lim_{\substack{r \to \infty \\ \theta \in (\pi/2, 3\pi/2)}} \frac{|R(z)|}{|e^z|} = \lim_{\substack{r \to \infty \\ \theta \in (\pi/2, 3\pi/2)}} |R(z)| \cdot e^{-r\cos\theta} = +\infty$$

Luego, $\forall M > 0, \exists K_1 > 0, \forall r > K_1 : \left|\frac{R(z)}{e^z}\right| > M, \forall \theta \in \left(\frac{\pi}{2}, \frac{3\pi}{2}\right)$

Eligiendo $M > 1$, $\forall r > K_1, \forall \theta \in \left(\frac{\pi}{2}, \frac{3\pi}{2}\right), z \in A$.

(ii) $\forall \theta \in \left(\frac{-\pi}{2}, \frac{\pi}{2}\right)$, $\cos\theta > 0$, entonces $e^{-r\cos\theta} \to 0, r \to \infty$

En consecuencia,

$$\lim_{\substack{r \to \infty \\ \theta \in (-\pi/2, \pi/2)}} \frac{|R(z)|}{|e^z|} = \lim_{\substack{r \to \infty \\ \theta \in (-\pi/2, \pi/2)}} |R(z)| \cdot e^{-r\cos\theta} = 0$$

Luego, $\forall \varepsilon > 0, \exists K_2 > 0, \forall r > K_2 : \left|\dfrac{R(z)}{e^z}\right| < \varepsilon, \forall \theta \in \left(\dfrac{-\pi}{2}, \dfrac{\pi}{2}\right)$

Tomando $\varepsilon < 1$, $\forall r > K_2, \forall \theta \in \left(\dfrac{-\pi}{2}, \dfrac{\pi}{2}\right) : z \notin A$.

□

Observación 4 El lema anterior nos garantiza que para todo $z \in \mathbb{C}$ con $|z| > \text{máx}\{K_1, K_2\}$ y $Re\, z \neq 0$,

$$Re\, z < 0 \Leftrightarrow z \in A$$

$$Re\, z > 0 \Leftrightarrow z \in A^c.$$

Lema 8 $\lim\limits_{z\to\infty} \left|\dfrac{R'(z)}{R(z)}\right| = 0$

Demostración: Derivando $R(z)$:

$$\frac{R'(z)}{R(z)} = \frac{\dfrac{P'(z)Q(z) - P(z)Q'(z)}{Q(z)^2}}{\dfrac{P(z)}{Q(z)}} = \frac{P'(z)Q(z) - P(z)Q'(z)}{P(z)Q(z)}.$$

Por tanto, como $\deg P' = k - 1$, $\deg Q' = j - 1$, el numerador tendrá grado $k + j - 1$ a lo sumo y el denominador grado $k + j$. Entonces, $|R'(z)/R(z)| \to 0, z \to \infty$ ya que el grado del numerador es menor que el del denominador. □

Lema 9 *Consideramos $z = re^{i\theta} \in \mathbb{C}$, con $r \to \infty$.*
Si

$$\varphi_1(\theta) = |e^z|^2 = e^{2r\cos\theta}, \quad \varphi_2(\theta) = |R(z)|^2 = R(re^{i\theta})R(re^{-i\theta}), \tag{3.14}$$

entonces:

(i) $\dfrac{\varphi_1'(\theta)}{\varphi_1(\theta)} = -2r\sin(\theta)$ *y* $\dfrac{\varphi_2'(\theta)}{\varphi_2(\theta)} = 2rRe(w)$, *donde* $w = ie^{i\theta} \cdot \dfrac{R'(re^{i\theta})}{R(re^{i\theta})}$.

(ii) $\forall \varepsilon \in (0, \pi/2)$, $\dfrac{\varphi_1'(\theta)}{\varphi_1(\theta)} < \dfrac{\varphi_2'(\theta)}{\varphi_2(\theta)}$, $\forall \theta \in [\varepsilon, \pi - \varepsilon]$.

Demostración:

(i) Derivando cada función, obtenemos:

$$\varphi_1'(\theta) = -2re^{2r\cos(\theta)}\sin(\theta)$$
$$\varphi_2'(\theta) = R'(re^{i\theta}) \cdot re^{i\theta}i \cdot R(re^{-i\theta}) + R(re^{i\theta}) \cdot R'(re^{-i\theta}) \cdot re^{-i\theta}(-i)$$

Luego,

$$\frac{\varphi_1'(\theta)}{\varphi_1(\theta)} = \frac{-2re^{2r\cos(\theta)}\sin(\theta)}{e^{2r\cos\theta}} = -2r\sin(\theta)$$

$$\begin{aligned}\frac{\varphi_2'(\theta)}{\varphi_2(\theta)} &= \frac{R'(re^{i\theta})\cdot re^{i\theta}i\cdot R(re^{-i\theta}) + R(re^{i\theta})\cdot R'(re^{-i\theta})\cdot re^{-i\theta}(-i)}{R(re^{i\theta})R(re^{-i\theta})} = \\ &= \frac{R'(re^{i\theta})\cdot re^{i\theta}i}{R(re^{i\theta})} + \frac{R'(re^{-i\theta})\cdot re^{-i\theta}(-i)}{R(re^{-i\theta})}\end{aligned}$$

Dado que los coeficientes de $R(z)$ son reales, entonces $R(\bar{z}) = \overline{R(z)}$ y $R'(\bar{z}) = \overline{R'(z)}$. Por tanto, si

$$u = \frac{R'(re^{i\theta})\cdot re^{i\theta}i}{R(re^{i\theta})} \Rightarrow \bar{u} = \frac{R'(re^{-i\theta})\cdot re^{-i\theta}(-i)}{R(re^{-i\theta})} \tag{3.15}$$

En consecuencia,

$$\frac{\varphi_2'(\theta)}{\varphi_2(\theta)} = u + \bar{u} = 2Re\,u = 2rRe(w), \text{donde}\, w = ie^{i\theta}\cdot\frac{R'(re^{i\theta})}{R(re^{i\theta})}$$

(ii) Consideremos $0 < \varepsilon < \pi/2$ y veamos que $\dfrac{\varphi_1'(\theta)}{\varphi_1(\theta)} < \dfrac{\varphi_2'(\theta)}{\varphi_2(\theta)}$, $\forall\theta \in [\varepsilon, \pi - \varepsilon]$.

Por (i) demostrar (ii) es lo mismo que ver que $-\sin\theta < Re(w)$, $\forall\theta \in [\varepsilon, \pi-\varepsilon]$, teniendo en cuenta que $w = w(r,\theta)$.

Sabemos por el Lema 8 que si $z \to \infty$, $Re(w) \to 0$, por tanto, $\forall\alpha > 0, \exists K > 0, \forall r > K : |Re(w(r,\theta))| < \alpha, \forall\theta \in \mathbb{R}$.
Aplicándolo para $0 < \alpha = \sin\varepsilon < 1$, existe $K > 0, \forall r > K, |Re(w(r,\theta))| < \alpha, \forall\theta \in [\varepsilon, \pi - \varepsilon]$.
Además, $\forall\theta \in [\varepsilon, \pi - \varepsilon] : \sin\theta \geq \alpha$, (ver Figura 3.6). Por tanto,

$$\left.\begin{array}{l} -\alpha < Re\,w < \alpha \\ -\sin\theta \leq -\alpha \end{array}\right\} \Rightarrow -\sin\theta \leq -\alpha < Re\,w(r,\theta), \forall\theta \in [\varepsilon, \pi - \varepsilon],$$

con lo que se demuestra (ii).

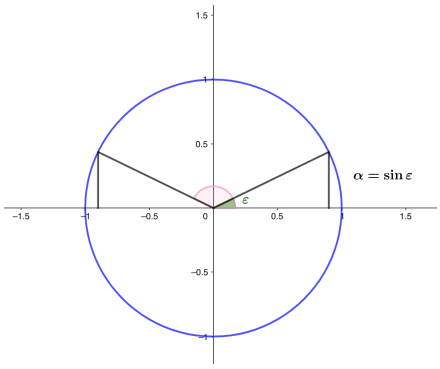


Figura 3.6: Función $\alpha = \sin\varepsilon$ para $0 < \varepsilon < \pi/2$.

□

Lema 10 *El borde ∂A posee solo dos ramas que van al infinito.*

Demostración: Para demostrar el lema utilizaremos las funciones definidas en el Lema 9.

Por el Lema 7 sabemos que existen $K_1, K_2 > 0$ tal que

$$\forall\theta \in (\pi/2, \pi), r > K_1 : z \in A \Rightarrow \varphi_2(\theta) - \varphi_1(\theta) > 0$$

$$\forall\theta \in (0, \pi/2), r > K_2 : z \in A^c \Rightarrow \varphi_2(\theta) - \varphi_1(\theta) < 0$$

Entonces, por continuidad de φ_1 y φ_2 en $\mathbb{R}$, $\exists\tilde{\theta} \in (0, \pi) : \varphi_1(\tilde{\theta}) = \varphi_2(\tilde{\theta})$, $\forall r > \text{máx}\{K_1, K_2\}$, lo que implica que $z = re^{i\tilde{\theta}} \in \partial A$.
Es decir, hay al menos una rama de ∂A que se va al infinito en $Im\, z > 0$.

Por el apartado (ii) del Lema 9, $\forall\varepsilon \in [0, \pi/2]$, $\forall\theta \in [\varepsilon, \pi - \varepsilon]$:

$$\begin{aligned}\frac{\varphi_1'(\theta)}{\varphi_1(\theta)} < \frac{\varphi_2'(\theta)}{\varphi_2(\theta)} &\Rightarrow \varphi_1'(\theta)\varphi_2(\theta) < \varphi_1(\theta)\varphi_2'(\theta) \Rightarrow \\ &\Rightarrow \varphi_1'(\theta)\varphi_2(\theta) - \varphi_1(\theta)\varphi_2'(\theta) < 0 \Rightarrow \\ &\Rightarrow \left(\frac{\varphi_1(\theta)}{\varphi_2(\theta)}\right)' = \frac{\varphi_1'(\theta)\varphi_2(\theta) - \varphi_1(\theta)\varphi_2'(\theta)}{\varphi_2^2(\theta)} < 0\end{aligned}$$

Luego, $\varphi_1(\theta)/\varphi_2(\theta)$ es estrictamente decreciente en $[\varepsilon, \pi - \varepsilon], \forall\varepsilon \in (0, \pi/2)$. Supongamos ahora que hay dos ramas de ∂A en $Im\, z > 0$ que se van a infinito, esto es, existen dos valores $0 < \theta_1 < \theta_2 < \pi$ tales que $\varphi_1(\theta_1) = \varphi_2(\theta_1)$, $\varphi_1(\theta_2) = \varphi_2(\theta_2)$. Entonces, existe $\varepsilon > 0$ tal que $0 < \varepsilon < \theta_1 < \theta_2 < \pi - \varepsilon < \pi$ y φ_1/φ_2 es estrictamente decreciente en $[\varepsilon, \pi - \varepsilon]$ por lo que

$$\frac{\varphi_1(\theta_1)}{\varphi_2(\theta_1)} > \frac{\varphi_1(\theta_2)}{\varphi_2(\theta_2)}$$

o, lo que es lo mismo, que $1 > 1$, lo que es una contradicción que viene de suponer que existían dos valores en los que φ_1 y φ_2 son iguales. Así que solo existe un valor θ en $[\varepsilon, \pi - \varepsilon]$ tal que $\varphi_1(\theta) = \varphi_2(\theta)$.
Consecuentemente, existe una única rama de ∂A en $Im\, z > 0$ que se va al infinito.
Por simetría, habrá otra rama de ∂A en $Im\, z < 0$ que se va al infinito. □

Lema 11 *En las condiciones de los lemas anteriores, si*

$$R(z) = Kz^l + \mathcal{O}(z^{l-1}), \quad z \to \infty \quad (l < 0) \tag{3.16}$$

estas ramas se acercan asintóticamente a la curva

$$x = \log|K| + l \cdot \log|y| \tag{3.17}$$

Demostración: Del Lema 10 tenemos que existe un único $\tilde{\theta} \in (0, \pi)$ y $\widehat{K} > \text{máx}\{K_1, K_2\} > 0$ tal que $\forall r > \widehat{K}$ $z = re^{i\tilde{\theta}} \in \partial A$.
Luego, si $z = x + iy$, por hipótesis, $|R(z)| = |e^z|$ y

$$|R(z)| = |Kz^l + \mathcal{O}(z^{l+1})| \simeq |K||z|^l = |K|(x^2 + y^2)^{l/2} \simeq |e^z| = e^x \Rightarrow$$

$$\Rightarrow \log|K| + \frac{l}{2}\log(x^2 + y^2) \simeq x$$

Como $z \in \partial A$ podemos acercarnos a z tanto como queramos con puntos A y de A^c, es decir, $\forall \varepsilon > 0$, $\exists z_1^\varepsilon \in A$, $\exists z_2^\varepsilon \in A^c$ con $z_1^\varepsilon, z_2^\varepsilon \notin i\mathbb{R}$, tal que

$$|z - z_1^\varepsilon| < \varepsilon, \quad \arg z_1^\varepsilon \sim \tilde{\theta}, \quad |z_1^\varepsilon| > \text{máx}\{K_1, K_2\}$$

$$|z - z_2^\varepsilon| < \varepsilon, \quad \arg z_2^\varepsilon \sim \tilde{\theta}, \quad |z_2^\varepsilon| > \text{máx}\{K_1, K_2\}$$

Por la Observación 4, $Re\, z_1^\varepsilon < 0$ y $Re\, z_2^\varepsilon > 0$, y se tiene

$$\left.\begin{array}{l} \arg z_1^\varepsilon \in \left(\dfrac{\pi}{2}, \dfrac{3\pi}{2}\right) \\[2ex] \arg z_2^\varepsilon \in \left(\dfrac{-\pi}{2}, \dfrac{\pi}{2}\right) \end{array}\right\} \Rightarrow \text{necesariamente} \quad \tilde{\theta} \simeq \tfrac{\pi}{2},$$

por lo que $x = r\cos\tilde{\theta} \simeq 0$ e $y = r\sin\tilde{\theta} \simeq r$. En consecuencia, $x/y \simeq 0$. Por tanto, dado que

$$x^2 + y^2 = y^2(x^2/y^2 + 1) \simeq y^2,$$

las ramas se acercan asintóticamente a

$$x = \log|K| + \frac{l}{2}\log(|y|^2) = \log|K| + l \cdot \log|y|$$

□

3.5. Ceros y polos de $R(z)$

Como $R(z) = P(z)/Q(z)$ es una función racional y $e^z \neq 0$, $\forall z \in \mathbb{C}$ se tiene que los ceros de $q(z)$ son los ceros de $P(z)$ y los polos de $q(z)$ son los ceros de $Q(z)$. Además,

$$\begin{aligned} &\text{Si } w \text{ es cero de q entonces, } q(w) = 0 < 1 \text{ por lo tanto, } w \in A^c. \\ &\text{Si } w \text{ es polo de q entonces } q(w) = \infty \text{ por lo que, } |q(w)| > 1 \text{ y } w \in A. \end{aligned} \tag{3.18}$$

Combinando las propiedades de las estrellas de orden cerca del origen y cuando $z \to \infty$ podremos localizar los ceros y polos de la función $R(z)$.

Lema 12 *Cada subconjunto $F \subset A$ acotado con borde común $\partial F \subset \partial A$ debe contener al menos un polo de $q(z)$.*

Demostración: Sea $F \subset A$ acotado con $\partial F \subset \partial A$. Supongamos que no hay polos de $q(z)$ en F. Como F es acotado por el principio del máximo tenemos que

$$\max_{z \in F} |q(z)| = \max_{z \in \partial F \subset \partial A} |q(z)| = 1$$

Pero si $z \in F \subset A$ entonces $|q(z)| > 1$, por lo que, hemos llegado a una contradicción al suponer que no hay polos de $q(z)$ en F. □

Para estudiar más profundamente el comportamiento de los ceros y polos de $q(z)$ necesitamos recordar algunos conceptos de análisis complejo, que hemos extraído de [3] y [5].

Definición 9 *Se dice que f es **meromorfa** en un abierto U si es holomorfa en U salvo en un conjunto de puntos que son polos de f.*

Definición 10 *Dadas dos curvas τ_0 y $\tau_1 : [0,1] \longrightarrow D$ cerradas y simples en una región D. Se dice que τ_0 es **homotópica** a τ_1 en D ($\tau_0 \sim \tau_1$) si existe una función continua $\gamma : [0,1] \times [0,1] \longrightarrow D$ tal que*

$$\begin{cases} \gamma(s,0) = \tau_0(s), \quad \gamma(s,1) = \tau_1(s), \quad 0 \leq s \leq 1 \\ \gamma(0,t) = \gamma(1,t), \quad 0 \leq t \leq 1 \end{cases}$$

Es decir, si $\tau_1 \sim \tau_2$, una se puede "deformar" continuamente en la otra.

Definición 11 *Si τ es una curva simple cerrada en $\mathbb{C}$ y $a \in \mathbb{C} - \tau$ se define el **índice** de τ con respecto al punto a como*

$$Ind(a, \tau) = \frac{1}{2\pi i} \oint_\tau \frac{dw}{w - a}$$

Este valor también se llama ***número de giros*** *de la curva τ alrededor del punto a, ya que si $Ind(a,\tau) = n$ entonces τ es homotópica a la curva*

$$z(t) = a + e^{2\pi int}, \quad t \in [0,1].$$

El Principio del Argumento nos permite contar el número de los ceros y polos de una función meromorfa f que están rodeados por un contorno simple cerrado C.

Teorema 7 *(Principio del Argumento).[3] Sea C un contorno simple cerrado contenido en un dominio D. Supongamos que la función f es meromorfa en D, y que $f(z) \neq 0$, $\forall z \in C$. Entonces,*

$$\frac{1}{2\pi i}\oint_C \frac{f'(z)}{f(z)}dz = N_0 - N_p, \tag{3.19}$$

donde N_0 es el número total de ceros de f y N_p es el número total de polos de f en la región de D rodeada por C. En la determinación de N_0 y N_p, tanto los ceros como los polos se cuentan según sus respectivas multiplicidades.

Si hacemos el cambio $w = f(z)$, $dw = f'(z)dz$ nos queda

$$\frac{1}{2\pi i}\oint_C \frac{f'(z)}{f(z)}dz = \frac{1}{2\pi i}\oint_{f(C)} \frac{dw}{w} = Ind(0, f(C))$$

Por tanto, el Teorema 9 nos dice que el número de ceros menos el número de polos de f en la región rodeada por C es igual al número de giros que la curva $f(C)$ da alrededor del origen.

Lema 13 *Cada subconjunto acotado $F \subset A$ con borde común $\partial F \subset \partial A$ que reúne m sectores de A en el origen debe contener al menos m polos de $q(z)$ (contados según su multiplicidad). Análogamente, cada subconjunto acotado $F \subset A^c$ con m sectores de A^c en el origen debe contener al menos m ceros de $q(z)$.*

Demostración: Primero, vamos a demostrarlo en el caso en que $m = 1$ y ∂F es una curva cerrada parametrizada orientada positivamente $c(t)$, $t_0 \leq t \leq t_1$.
Sea el vector velocidad $\vec{a} = (c_1'(t), c_2'(t))$, que es tangente a la curva, y $\vec{n} = (c_2'(t), -c_1'(t))$ un vector normal exterior a ∂F.
Escribimos $q(z) = R(z)/e^z = r(x,y) \cdot e^{i\varphi(x,y)}$, donde $z = x + iy$, es decir, $r(x,y) = |q(z)|$ y $\varphi(x,y) = \arg(q(z))$. Luego

$$\log q(z) = \log(r(x,y)) + i\varphi(x,y) \tag{3.20}$$

Por otro lado, para cualquier incremento $\Delta > 0$, como el borde ∂A está rodeado exteriormente por puntos de A^c

$$\begin{cases} (x+\Delta\, c_2', y-\Delta\, c_1') \in A^c \Rightarrow r(x+\Delta\, c_2', y+\Delta\, c_1') < 1, \\ (x,y) \in \partial F \subset \partial A \Rightarrow r(x,y) = 1, \end{cases}$$

entonces,

$$\frac{\partial r}{\partial \vec{n}}(x,y) = \lim_{\Delta \to 0^+} \frac{r(x+\Delta\, c_2', y-\Delta\, c_1') - r(x,y)}{\Delta} \leq 0$$

Por tanto, r se incrementa hacia dentro de F, es decir,

$$\frac{\partial r}{\partial \vec{n}}(x,y) \leq 0 \Rightarrow \frac{\partial r}{\partial(-\vec{n})}(x,y) \geq 0$$

Además,

$$\begin{aligned} &\frac{\partial}{\partial \vec{n}}(\log r(x,y)) = \nabla(\log r)\cdot \vec{n} = \left(\frac{r_x}{r}, \frac{r_y}{r}\right)\cdot \vec{n} = \\ &= \frac{1}{r}((r_x, r_y)\cdot \vec{n}) = \frac{1}{r}(\nabla r(x,y)\cdot \vec{n}) = \frac{1}{r}\cdot \frac{\partial r(x,y)}{\partial \vec{n}} \leq 0 \end{aligned} \tag{3.21}$$

Por otro lado, por las ecuaciones de Cauchy-Riemann para $\log q(z)$ (3.20),

$$\frac{\partial(\log r)}{\partial x} = \frac{\partial \varphi}{\partial y}, \qquad \frac{\partial(\log r)}{\partial y} = -\frac{\partial \varphi}{\partial x} \tag{3.22}$$

de modo que

$$\begin{aligned} \frac{\partial(\log r)}{\partial \vec{n}} &= \left(\frac{\partial(\log r)}{\partial x}, \frac{\partial(\log r)}{\partial y}\right)(c_2'(t), -c_1'(t)) = \frac{\partial(\log r)}{\partial x}\cdot c_2'(t) - \frac{\partial(\log r)}{\partial y}\cdot c_1'(t) = \\ &= \frac{\partial \varphi}{\partial y}\cdot c_2'(t) - \left(-\frac{\partial \varphi}{\partial x}\right)\cdot c_1'(t) = \frac{\partial \varphi}{\partial x}\cdot c_1'(t) + \frac{\partial \varphi}{\partial y}\cdot c_2'(t) = \frac{\partial \varphi}{\partial \vec{a}} \end{aligned}$$

Por tanto, por (3.21),

$$\frac{\partial \varphi}{\partial \vec{a}} \leq 0. \tag{3.23}$$

Sea $\widetilde{q}(t) = q(c_1(t), c_2(t)) = r(c_1(t), c_2(t))e^{i\varphi(c_1(t),c_2(t))}$. Entonces, derivando,

$$\begin{aligned} \widetilde{q}\,'(t) &= \left(\frac{\partial r}{\partial x}\cdot c_1'(t) + \frac{\partial r}{\partial y}\cdot c_2'(t)\right) e^{i\varphi(c_1(t),c_2(t))} \\ &+ r(c_1(t), c_2(t))\cdot i\cdot e^{i\varphi(c_1(t),c_2(t))}\cdot \left(\frac{\partial \varphi}{\partial c_1}\cdot \frac{\partial c_1}{\partial t} + \frac{\partial \varphi}{\partial c_2}\cdot \frac{\partial c_2}{\partial t}\right) = \\ &= \frac{\partial r}{\partial \vec{a}}(c(t))e^{i\varphi(c(t))} + i\cdot q(c(t))\cdot \frac{\partial \varphi}{\partial \vec{a}}(c(t)) \end{aligned}$$

Como $c(t) \in \partial F$, $\forall t \in [t_0, t_1]$, $r(c(t)) = |q(c(t))| = 1$, $\forall t \in [t_0, t_1]$, por lo que $\dfrac{\partial r}{\partial \vec{a}}(c(t)) = 0$, se tiene,

$$\widetilde{q}'(t) = i \cdot q(c(t)) \cdot \frac{\partial \varphi}{\partial \vec{a}}(c(t))$$

Por otra parte, por la regla de la cadena,

$$\widetilde{q}'(t) = \frac{d}{dt} q(c(t)) = q'(c(t)) \cdot c'(t)$$

donde $c'(t) = c_1'(t) + ic_2'(t)$. Entonces,

$$iq(c(t)) \frac{\partial \varphi}{\partial \vec{a}}(c(t)) = q'(c(t))c'(t)$$

Luego,

$$\frac{\partial \varphi}{\partial a}(c(t)) = -i \frac{q'(c(t))}{q(c(t))} c'(t), \ \forall t \in [t_0, t_1]$$

Como q' tiene un número finito de ceros, esto implica que $\dfrac{\partial \varphi}{\partial a}(c(t)) = 0$ solo en un número finito de puntos y $\dfrac{\partial \varphi}{\partial \vec{a}}(c(t)) < 0$ en el resto.
El argumento de $q(z)$ solo se anula en un número finito de puntos de ∂F y decrece cuando $z \in \partial F$ entonces, $\tau = q(\partial F)$ se recorre en sentido negativo.

Es decir, el argumento de $q(z)$ decrece (salvo en un número finito de puntos) a medida que se recorre la curva ∂F en sentido positivo. En otras palabras, la curva $\tau = q(\partial F)$ se recorre en sentido negativo a medida que ∂F se recorre en sentido positivo y tiene que dar al menos una vuelta completa.

Además, $|q(z)| = 1$, $\forall z \in \partial F$, así que $\tau = q(\partial F)$ tiene que caer sobre la circunferencia unidad y darle al menos una vuelta en sentido negativo. Por tanto, $Ind(0, \tau) \leq -1$.

Por el Principio del Argumento $Ind(0, \tau) = N_0 - N_p$ donde N_0 es el número de ceros y N_p el número de polos de q en F. Como $F \subset A$, por (3.18), en F no hay ceros de q, $N_0 = 0$. Por tanto, $Ind(0, \tau) = -N_p \leq -1$, luego $N_p \geq 1$, así que tiene que haber al menos un polo de q en F.
Por otro lado, si $F \subset A^c$ acotado, $\partial F \subset \partial A$, con ∂F una curva simple cerrada parametrizada $c(t)$, $t \in [t_0, t_1]$ igual que en el caso anterior, si $q(z) = r(x, y)e^{i\varphi(x,y)}$ ahora tendremos que

$$\frac{\partial \varphi}{\partial \vec{a}} = \frac{\partial (\log r)}{\partial \vec{n}} \geq 0$$

y solo se da la igualdad en un número finito de puntos de ∂F, por lo que φ es creciente, y $\tau = q(\partial F)$ se recorre en sentido positivo, girando sobre la circunferencia unidad y, por el Principio del Argumento,

$$N_0 - N_p = Ind(0, \tau)$$

donde N_0 es el número de ceros de q y N_p el número de polos en $F \subset A^c$. Como $F \subset A^c$, no hay polos de q en F, entonces $N_p = 0$ y $Ind(0, \tau) \geq 1$ porque la circunferencia unidad se recorre al menos una vez, luego $N_0 \geq 1$, y F contiene al menos un cero de q.

Si $F \subset A$ con $\partial F \subset \partial A$ reúne $m \geq 1$ sectores del origen,

$$F = \cup_{i=1}^{m} F_i \tag{3.24}$$

donde cada uno de los F_i incluye un solo sector del origen, $F_i \cap F_j = \emptyset$ y $\partial F_i \subset \partial A$ es una curva cerrada simple, aplicando lo anterior sobre cada F_i, se demuestra que hay al menos m polos de q en F.

Si $F \subset A^c$ es de la forma (3.24) lo que se obtienen son m ceros de q en F.

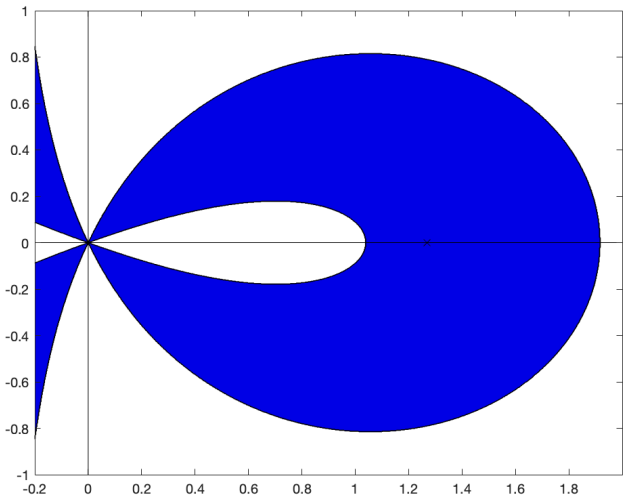


Figura 3.7: Subconjunto $F \subset A$ del SDIRK $\gamma = (3 + \sqrt{3})/6$ en color azul.

Si $F \subset A$ acotado, $\partial F \subset \partial A$ pero la curva es cerrada no simple, la demostración para un m general es muy complicada. Veámoslo para el caso $m = 2$, es decir, que F contiene 2 sectores del origen, pues ∂F no es una curva simple. Este es el caso, por ejemplo, del subconjunto acotado $F \subset A$ en $Re\, z > 0$ de la estrella de orden de la función de estabilidad del método SDIRK de orden 3 (2.6) con $\gamma = (3 + \sqrt{3})/6$ que se puede ver en la Figura 3.7.

Así $\partial F \subset \partial A$ rodea a un conjunto $F_2 \subset A^c$. Consideremos, además, el subconjunto de $\mathbb{C}$

$$F_1 = F \cup F_2$$

Entonces, $\partial F = \partial F_1 \cup \partial F_2$ y ∂F_1 y ∂F_2 son dos curvas cerradas simples orientadas positivamente.

Sean N_0^i y N_p^i el número de ceros en F_i y $\tau_i = q(\partial F_i)$ para $i = 1$ y 2. Sabemos que $N_0^i - N_p^i = Ind(0, \tau_i)$, $i = 1, 2$. Además, al igual que vimos en el caso $m = 1$, como ∂F_1 rodea puntos de A, τ_1 recorre al menos una vez la circunferencia unidad en sentido negativo, por lo que $Ind(0, \tau_1) \leq -1$. En cambio, como ∂F_2 rodea a $F_2 \subset A^c$, τ_2 recorre en sentido positivo la circunferencia uniddad, al menos una vez al completo, por lo que $Ind(0, \tau_2) \geq 1$.
Como sabemos que en $F_2 \subset A^c$ no puede haber polos de q, $N_p^2 = 0$, por lo que $N_0^2 = Ind(0, \tau_2)$, y si hay polos de q en F, han de estar en F_1.

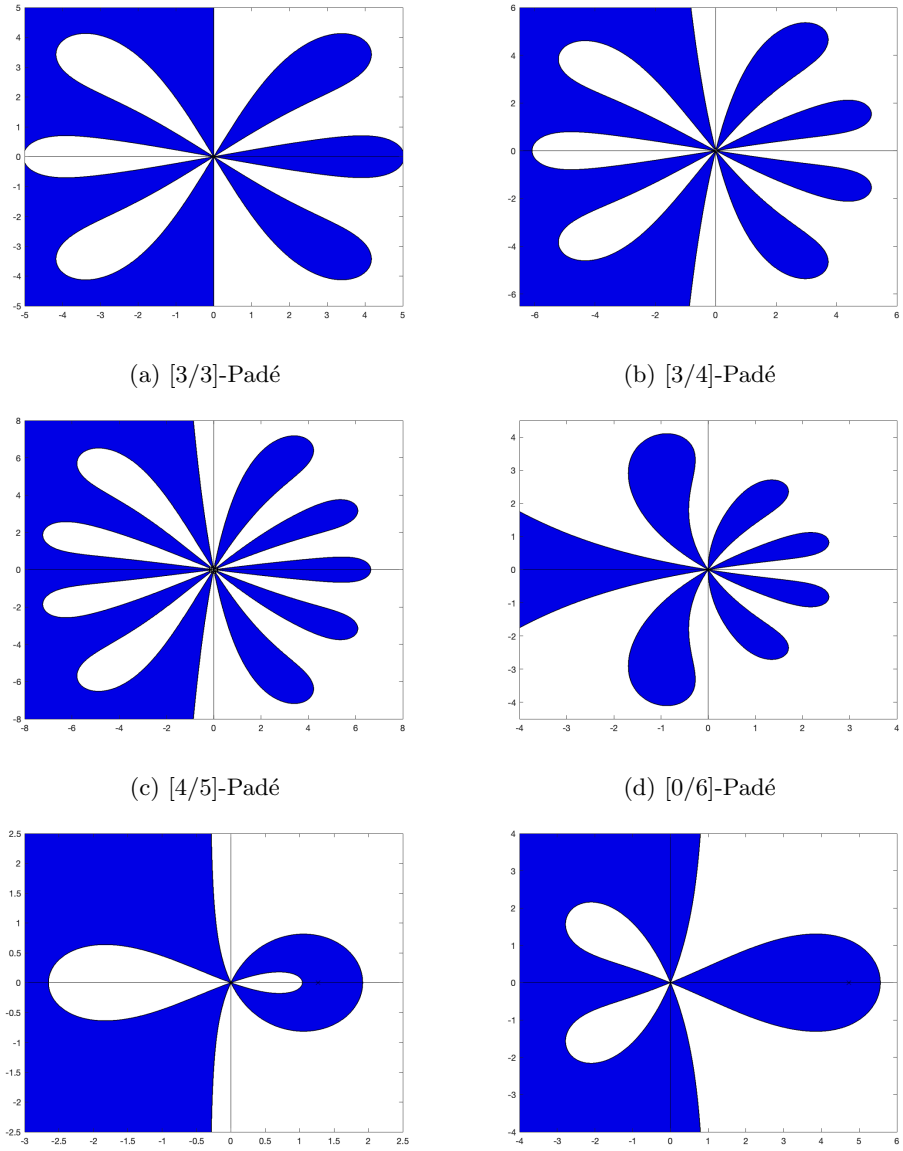
Además, como en $F \subset A$ no puede haber ceros de q, si F_1 tiene ceros, tienen que estar en F_2, por tanto, $N_0^1 = N_0^2$. Por tanto,

$$N_p^1 = Ind(0, \tau_2) - Ind(0, \tau_1) \geq 1 + 1 = 2$$

con lo que se concluye que debe de haber al menos dos polos de $F \subset A$.

Se procede análogamente si $F \subset A^c$ acotado conteniendo 2 sectores del origen. $\square$

Para acabar esta sección, hemos creado un programa en Matlab que dibuja las estrellas de orden. En la Figura 3.8 pueden ver algunas estrellas de orden de aproximantes de Padé y de las funciones de estabilidad de los métodos SDIRK de orden 3 que no son Padé.



(a) [3/3]-Padé

(b) [3/4]-Padé

(c) [4/5]-Padé

(d) [0/6]-Padé

(e) SDIRK $\gamma = (3+\sqrt{3})/6$

(f) SDIRK $\gamma = (3-\sqrt{3})/6$

Figura 3.8: Order star

3.6. Orden y estabilidad para $R(z)$

Combinando las propiedades vistas anteriormente, podemos deducir algunas condiciones para que una función racional sea A-estable.

Teorema 8 *Si $R(z)$ es A-estable, entonces $p \leq 2k_1 + 2$, donde k_1 es el número diferentes de ceros de $R(z)$ en $\mathbb{C}^-$.*

Demostración: Sabemos por la Observación 3 que hay al menos

$$\begin{cases} E\left[\dfrac{p+1}{2}\right] & \text{sectores de } A \text{ que arrancan en } \mathbb{C}^- \\ \\ E\left[\dfrac{p+1}{2}\right]-1 & \text{sectores de } A^c \text{ completamente contenidos en } \mathbb{C}^- \end{cases}$$

Los sectores de A que arrancan en $\mathbb{C}^-$ no pueden estar acotados, pues si lo estuviera, habría al menos un polo de $R(z)$ (o de $q(z)$) en $\mathbb{C}^-$ por el Lema 13, por lo que no sería A-estable. Además, los sectores de A^c completamente contenidos en $\mathbb{C}^-$ tienen que estar acotados, pues si fueran infinitos se contradeciría el Lema 7, y cada uno de ellos contrendía al menos un cero de $R(z)$. Por tanto, habría al menos $E\left[\dfrac{p+1}{2}\right]-1$ ceros diferentes de $R(z)$ en $\mathbb{C}^-$, lo que implica que

$$E\left[\frac{p+1}{2}\right]-1 \leq k_1$$

En consecuencia,

$$\begin{cases} p \leq 2k_1+1 & \text{si p es impar} \\ \\ p \leq 2k_1+2 & \text{si p es par} \end{cases}$$

Por tanto, solo podemos garantizar que $p \leq 2k_1+2$. □

Teorema 9 *Si $R(z)$ es I-estable, entonces $p \leq 2j_1$, donde j_1 es el número de polos de $R(z)$ en $\mathbb{C}^+$.*

Demostración: De igual manera que antes sabemos por la Observación 3 que al menos $E\left[\dfrac{p+1}{2}\right]$ sectores de A empiezan en $\mathbb{C}^+$.

Como $R(z)$ es I-estable, los sectores de A que arrancan en $\mathbb{C}^+$ no pueden cruzar el eje imaginario, es decir, tienen que estar completamente contenidos en $\mathbb{C}^+$, y tienen que estar acotados pues, si no, se contradecirá el Lema 7.

Por el Lema 13, cada uno de esos sectores acotados de A deben contener al menos un polo de $R(z)$, por lo que se garantiza que hay al menos $E\left[\dfrac{p+1}{2}\right]$ polos de $R(z)$ en $\mathbb{C}^+$.

Si j_1 es el número de polos de $R(z)$ en $\mathbb{C}^+$, entonces nos queda

$$E\left[\frac{p+1}{2}\right] \leq j_1$$

o, lo que es lo mismo,

$$\begin{cases} \text{p es impar} & p \leq 2j_1 - 1 \\ \text{p es par} & p \leq 2j_1 \end{cases}$$

Entonces, en general solo podemos garantizar que $p \leq 2j_1$. □

Teorema 10 *Supongamos que $p \geq 2j - 1$ y $|R(\infty)| \leq 1$. Entonces, $R(z)$ es A-estable.*

Demostración: Sabemos que $R(z)$ es A-estable si y solo si, es I-estable y todos los polos están en $\mathbb{C}^+$.

Como $p \geq 2j - 1 > 2j - 2$ entonces es I-estable por la Proposición 7 del Capítulo 2. Por el Teorema 9, $p \leq 2j_1 \leq 2j$ ya que $j_1 \leq j$. Pero además,

$$2j - 1 \leq p \leq 2j_1 \Rightarrow 2j - 1 \leq 2j_1 \Rightarrow j \leq j_1 + \frac{1}{2}$$

Por tanto,

$$j_1 \leq j \leq j_1 + \frac{1}{2} \Rightarrow j = j_1,$$

lo que implica que todos los polos de $R(z)$ están en $\mathbb{C}^+$. Aplicando la Proposición 4 del Capítulo 2, se tiene que $R(z)$ es A-estable. □

Teorema 11 *Suponemos que $p \geq 2j - 2$, $|R(\infty)| \leq 1$, y los coeficientes del denominador $Q(z)$ tienen signos alternos. Entonces, $R(z)$ es A-estable.*

Demostración: Si $p \geq 2j - 1$ entonces por el teorema anterior $R(z)$ es A-estable. Nos falta ver el caso $p = 2j - 2$. Por la Proposición 7 del Capítulo 2 sabemos que $R(z)$ es I-estable. Luego, por el Teorema 9, $p \leq 2j_1$ con j_1 el número de polos de $R(z)$ en $\mathbb{C}^+$. Por tanto,

$$2j - 2 \leq p \leq 2j_1 \quad y \quad j_1 \leq j$$

entonces nos queda o que $j_1 = j$ y por tanto A-estable, o bien $j_1 = j - 1$. Si $j_1 = j - 1$ entonces hay un único polo de $R(z)$ en $\mathbb{C}^-$, o sea existe $w \in \mathbb{C}^- : Q(w) = 0$.
Como $Q(z)$ tiene coeficientes reales, w tiene que ser real, es decir, $w = -x \in \mathbb{R}^-, x > 0$. Sin embargo, por hipótesis

$$Q(z) = b_0 - b_1 z + b_2 z^2 - b_3 z^3 + \cdots + (-1)^j b_j z^j$$

con $b_0 \neq 0$ y $b_l \geq 0$. Así que para $x > 0$,

$$\widetilde{Q}(x) := Q(-x) = b_0 + b_1 x + b_2 x^2 + \cdots + b_j x^j.$$

Podemos ver que no hay cambios de signos en los coeficientes de $\widetilde{Q}(x)$. Por tanto, no puede tener raíces reales positivas (regla de Descartes) lo que nos lleva a una contradicción, así que no es posible que haya un único polo de $R(z)$ en $\mathbb{C}^-$, lo que demuestra la A-estabilidad. □

Acabamos esta sección con un teorema extraído de [5, p.57], que no demostraremos ya que su demostración requiere más herramientas de análisis complejo de las que no disponemos en un nivel de grado.

Teorema 12 *Sea $R(z)$ con k_0 ceros distintos y j_0 polos distintos. Entonces, $p \leq k_0 + j_0$.*

3.7. Estabilidad de las aproximaciones de Padé

Aplicando las propiedades de las secciones anteriores podemos deducir la A-estabilidad de los aproximantes de Padé.

Teorema 13 *La aproximación $[k/j]$-Padé, dada en (3.2), es A-estable si y solo si $k \leq j \leq k+2$. Además, todos sus ceros y polos son simples.*

Demostración: " $\Rightarrow$ " Tenemos que ver que $k \leq j \leq k+2$, y tenemos que $p = k + j$. Además, por el Teorema 8 sabemos que

$$p \leq 2k_1 + 2 \Rightarrow k + j \leq 2k_1 + 2 \leq 2k + 2, \quad (k_1 \leq k).$$

Por tanto, $j \leq k+2$ y, como es A-estable, sabemos que $j \geq k$.
" $\Leftarrow$ " Tenemos que $k \leq j \leq k+2$ y $p = k + j$. Por una parte,

$$k \geq j - 2 \Rightarrow p = k + j \geq 2j - 2.$$

Por otra parte, por (3.2), si $j = k$, $|R(\infty)| = 1$ y si $j = k+1$ ó $k+2$, $|R(\infty)| = 0$, así que $|R(\infty)| \leq 1$ para $k \leq j \leq k+2$. Además, $Q(z)$ tiene signos alternos, así que por el Teorema 11, es A-estable.
Por otro lado, para ver que los polos y ceros son simples tenemos que ver que los ceros y polos tienen multiplicidad 1. Por el Teorema 12 tenemos que si k_0 y j_0 son, respectivamente, el número de ceros y de polos distintos de $R(z)$,

$$k + j = p \leq k_0 + j_0 \leq k + j \Rightarrow k = k_0 \quad \text{y} \quad j = j_0$$

con lo que se concluye la demostración. □

3.8. Ejemplos de aproximantes de Padé

En el capítulo 2 vimos que cuando las funciones de estabilidad de los métodos es simple, como en el caso del θ-método, detectar si es A-estable o no es sencillo.

Aplicando la teoría de las estrellas de orden desarrollada en este capítulo podemos deducir la A-estabilidad de métodos de alto orden como el caso del método de Kuntzmann y Butcher de orden 6, donde

$$R(z) = \frac{1 + z/2 + z^2/10 + z^3/120}{1 - z/2 + z^2/10 - z^3/120}$$

Como podemos observar la función de estabilidad coincide con el aproximante [3/3]-Padé cuya estrella de orden podemos ver en la Figura 3.8a. Por tanto, podemos aplicar el Teorema 13, y vemos que dicha función es A-estable dado que $3 \leq 3 \leq 5$. Por otro lado, como el numerador y el denominador son del mismo grado entonces, $R(z) \to -1$ cuando $z \to \infty$, por tanto, no es L-estable.

Una aplicación inmediata de esta teoría es que permite construir métodos RK de alto orden A-estable, eligiendo sus coeficientes de forma que su función de estabilidad sea un aproximante de Padé verificando el Teorema 13. Esta es la continuación natural del trabajo realizado en esta memoria.

Conclusiones

Este Trabajo de Fin de Grado se ha centrado en el estudio de los métodos Runge-Kutta, su estabilidad, en particular su A-estabilidad y la teoría de las estrellas de orden.

La estabilidad de los métodos numéricos en la resolución de ecuaciones diferenciales es fundamental para obtener aproximaciones fiables y eficientes. Garantizar que un método es A-estable asegura que propagará concretamente los errores sobre problemas lineales de coeficientes constantes. En la práctica también mejora la estabilidad de los métodos sobre muchos problemas no lineales. Hemos visto que no hay métodos explícitos A-estables por lo que uno de los tópicos de investigación principales en esta área es la de construir métodos RK implícitos A-estables con orden alto.

El principal y último resultado de esta memoria es el Teorema 13, que garantiza que, bajo ciertas condiciones, los aproximantes de Padé producen métodos A-estables con el orden mayor posible. Hemos demostrado las principales propiedades de las estrellas de orden que se usan en la demostración de dicho Teorema.

Bibliografía

[1] U.M. Ascher and Linda.R. Petzold, *Computer Methods for Ordinary Differential equations and Differential-Algebraic Equations*, SIAM,1997.

[2] George A. Baker, P. Graves-Morris and P.A. Carruthers *Padé Approximants*, Addison-Wesley Publishing Company, London, 1981.

[3] John B. Conway, *Functions of One Complex Variable*, Springer-Verlag, New York, 1973.

[4] E. Hairer, S.P. NØrsett and G. Wanner, *Solving ordinary diffrential Equations I*, Springer-verlag, 1993.

[5] E. Hairer and G. Wanner, *Solving ordinary diffrential Equations II*, Springer-verlag, 1996.

[6] H. Logemann and E.P. Ryan, *ordinary differential equations*, Springer Verlag, London, 2014.

[7] G. Söderlind, L. Jay and M. Calvo, *Stiffness 1952-2012: Sixty years in search of a definition*, BIT Number Math, 2015, 55, pp.531-558.

Order star theory

Yaiza Pérez Tejera
Facultad de Ciencias • Sección de Matemáticas
Universidad de La Laguna
alu0101266473@ull.edu.es

Abstract

In this Bachelor's thesis, the A-stability of Runge-Kutta methods has been studied, with a particular focus on the theory of order stars, which determines necessary and sufficient conditions for the stability function of a method to be A-stable.
To do this, we have divided this work into three chapters. In the first chapter, we introduce the Runge-Kutta methods and some of their properties. In the second chapter, we study their stability and define the stability function and A-stability. Finally, we develop the theory of order stars, which guarantees the A-stability of Padé approximants under certain conditions.

1. Introduction to Runge-Kutta Methods

Let us consider the initial value problem

$$y'(t) = f(t, y(t)), \quad y(t_0) = y_0, \quad t \in [t_0, t_f] \tag{1}$$

with $f : [t_0, t_f] \times \Omega \to \mathbb{R}^m$, where $\Omega \subset \mathbb{R}^m$ is an openset.
A solution to IVP (1) is a continuously differentiable function on some interval I that contains t_0 and it satisfies the IVP.
Given an integer $s \geq 1$, the method that, given an approximation y_n to the solution of the IVP (1) at t_n, provides an approximation y_{n+1} to the solution at $t_{n+1} = t_n + h$ using the formula:

$$\begin{aligned} k_i &= f(t_n + c_i h, y_n + h \sum_{j=1}^{s} a_{ij} k_j), \quad i = 1, \dots, s \\ y_{n+1} &= y_n + h \sum_{i=1}^{s} b_i k_i \end{aligned} \tag{2}$$

is called **s-stage Runge-Kutta method**. $A = (a_{ij})_{i,j=1}^{s}$ is called the **coefficient matrix**, the vector $c = (c_1, c_2, \dots, c_s)^T$ is the vector of **nodes**, and $b = (b_1, b_2, \dots, b_s)^T$ is the vector of **weights**.

Theorem 1 *Let $f : [t_0, t_f] \times \mathbb{R}^m \to \mathbb{R}^m$ be a continuous function that satisfies a Lipschitz condition with respect to y, where L is the Lipschitz constant. If we take*

$$\tilde{h} < \frac{1}{L \max_i \sum_j |a_{ij}|}, \tag{3}$$

then for every $h \in [0, \tilde{h}]$, there exists a unique solution of the implicit system (2) that can be obtained through functional iteration. If f is p times continuously differentiable, the functions $k_i = k_i(h)$ are also $C^p([0, \tilde{h}])$.

Definition 1 *A method $RK(A, b)$ is of **order** $p \geq 1$ if for IVPs (1) with $f \in C^p([t_0, t_f] \times \Omega)$ and Lipschitz, there exists $K > 0$ independent of h such that*

$$||y(t_0 + h) - y_1|| \leq K h^{p+1}, \quad h \to 0^+. \tag{4}$$

2. Stability of Implicit Runge-Kutta Methods

Applying the RK (2) to the Dahlquist test equation

$$y' = \lambda y, Re\, \lambda \leq 0. \tag{5}$$

it yields

$$y_{n+1} = R(h\lambda) y_n, \quad R(z) = 1 + z b^T (I - zA)^{-1} e \tag{6}$$

where $e = (1, \dots, 1)^T$ and I is the identity matrix of size s. The function $R(z)$ is called the **stability function** of the method.

Proposition 1

$$R(z) = \frac{det(I - zA + zeb^T)}{det(I - zA)} \tag{7}$$

Proposition 2 *If the method is of order $p \geq 1$, then,*

$$e^z - R(z) = C z^{p+1} + \mathcal{O}(z^{p+2}) \quad \textit{as}, z \to 0, \quad C \neq 0. \tag{8}$$

Definition 2 *A method, whose **stability domain** $S = z \in \mathbb{C} / |R(z)| \leq 1$ satisfies $\mathbb{C}^- \subset S$ is called **A-stable**.*

3. Order Stars

Given (k, j) the $[k/j]$-Padé approximant to e^z is the only rational function that satisfies

$$R_{kj}(z) = \frac{P_k(z)}{Q_j(z)} = e^z + \mathcal{O}(z^{k+j+1}), \quad z \to 0 \tag{9}$$

with $\deg P_k = k$, $\deg Q_j = j$.

Definition 3 *The set*

$$A = z \in \mathbb{C} : |R(z)| > |e^z| = z \in \mathbb{C} : |q(z)| > 1, \quad q(z) = \frac{R(z)}{e^z} \tag{10}$$

*is called the **order star** of R.*

Lemma 1 *If $R(z)$ is an approximation to e^z of order p, that is,*

$$e^z - R(z) = C z^{p+1} + \mathcal{O}(z^{p+2}), \quad z \to 0, \quad C \neq 0, \tag{11}$$

In a neighborhood of $z = 0$, the set A consists of the points in $p+1$ circular sectors of equal width $\pi/(p+1)$ alternating with $p+1$ circular sectors of the same width of A^c.

Lemma 2 *Every bounded subset $F \subset A$ with a common boundary $\partial F \subset \partial A$ that includes m sectors of A at the origin must contain at least m poles of $q(z)$ (counted with their multiplicities). Similarly, every bounded subset $F \subset A^c$ with m sectors of A^c at the origin must contain at least m zeros of $q(z)$.*

Theorem 2 *The $[k/j]$-Padé approximation, given in (9), is A-stable if and only if $k \leq j \leq k+2$. Moreover, all its zeros and poles are simple.*

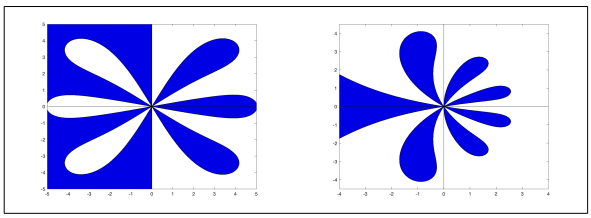


Figure 1: *Order star for Padé approximations*

References

[1] E.Hairer, S.P.Norsett and G.Wanner, *"Solving ordinary diffrential Equations I"*, Springer-verlag.

[2] E.Hairer and G.Wanner, *"Solving ordinary diffrential Equations II"*, Springer-verlag.