Métodos numéricos para ecuaciones diferenciales ordinarias
Métodos Runge–Kutta explícitos

Numerical methods on ordinary differential equations
Explicit Runge–Kutta methods

Patricia De León Camejo

Trabajo de Fin de Grado

Departamento de Análisis Matemático

Sección de Matemáticas

Facultad de Ciencias

Universidad de La Laguna

La Laguna, 14 de julio de 2015

Dra. Dña. **Soledad Pérez Rodríguez**, con N.I.F. 45.441.325-H profesora Contratado Doctor adscrita al Departamento de Análisis Matemático de la Universidad de La Laguna

C E R T I F I C A

Que la presente memoria titulada:

"*Métodos numéricos para ecuaciones diferenciales ordinarias. Métodos Runge-Kutta implícitos*"

ha sido realizada bajo su dirección por Dña. **Patricia De León Camejo**, con N.I.F. 78.533.340-R.

Y para que así conste, en cumplimiento de la legislación vigente y a los efectos oportunos firman la presente en La Laguna a 10 de julio de 2015

Agradecimientos

A mi familia por su apoyo incondicional
durante mi etapa universitaria y siempre.

A mis amigos, en especial a Arantxa.

Y a mi tutora Soledad por enseñarme
que tipo de profesional quiero ser.

Resumen

El objetivo de este trabajo es el estudio de los métodos para la resolución de problemas de valor inicial (PVI) en ecuaciones diferenciales ordinarias (EDO), en especial una clase de métodos de un paso, los métodos Runge-Kutta explícitos. Estudiaremos el orden de dichos métodos desarrollando la teoría de las series de Butcher, así como su convergencia y su estabilidad lineal. Finalmente llevamos a cabo un experimento numérico con la ecuación del calor aplicando el método de líneas donde haremos uso de varios Runge-Kutta explícitos, entre los que se encuentran los métodos Runge-Kutta-Chebychev.

Palabras clave: Métodos numéricos, Métodos Runge-Kutta, Series de Butcher

Abstract

This work deals with numerical methods for the solution of initial value problems (IVP) in ordinary differential equations (ODE). In particular, a class of one-step methods is considered, the explicit Runge-Kutta (ERK) methods. We study the order of these methods by deriving the Butcher series theory, their convergence and their linear stability. Finally, a numerical experiment is presented where the method of lines is applied to the well-known heat equation. Several explicit Runge-Kutta methods are implemented such as the Runge-Kutta-Chebychev methods.

Keywords: *Numerical methods, Runge-Kutta methods, Butcher series*

Índice general

Capítulo 1

Introducción

En el estudio de las ecuaciones diferenciales las herramientas numéricas han jugado un papel importante debido a que la mayor parte de las ecuaciones que aparecen en los problemas no se pueden resolver exactamente y, por tanto, hay que recurrir a algún tipo de aproximación de la solución. Sin embargo, durante el siglo XIX y buena parte del XX no se usaron muchos de los métodos que se desarrollaron teóricamente ya que no existían máquinas en los que se pudieran computar. Este hecho cambió a mediados del siglo XX con la aparición de ordenadores que ya poseían una cierta capacidad de cálculo y de almacenamiento de datos.
Los métodos numéricos usados para la resolución de problemas de valor inicial (PVI) en ecuaciones diferenciales ordinarias (EDO)

$$y' = f(t, y), \quad y(0) = y_0, \quad t \in [0, T], \quad y,\ f \in I\!R^m \tag{1.1}$$

pueden clasificarse en dos grupos:

(a) **Métodos de un paso**: Se usa la información de la solución en un instante t para obtener una aproximación de la solución en un instante $t + h$. Más formalmente, si conocemos una aproximación y_n a la solución en un punto t_n, el método lo que nos proporcionará será una nueva aproximación y_{n+1} en el punto $t_{n+1} = t_n + h$.

(b) **Métodos multipaso**: Se usa la información calculada en varios puntos previos $\{t_{n-k}, t_{n-k+1}, ..., t_n\}$ para conseguir la aproximación del siguiente punto t_{n+1}.

En esta memoria nos centramos en el estudio de una clase de métodos de un paso, los **métodos Runge-Kutta explícitos** (RKE). Ilustramos estos métodos con un ejemplo clásico de la Astronomía, el **problema de los tres cuerpos restringido** [5, 12]:

Ejemplo 1.0.1 *Se contemplan dos cuerpos de masas 1-μ y μ en rotación circular en un plano y un tercer cuerpo de masa despreciable moviéndose en el mismo plano, como por ejemplo el movimiento descrito por un cuerpo en el campo gravitatorio creado por la Tierra y la Luna. Suponemos que la Luna describe una órbita circular plana alrededor de la Tierra y que el cuerpo se mueve en el plano Tierra-Luna. Tomamos un sistema de coordenadas cartesianas de tal manera que la Tierra aparezca fija en el punto $(-\mu, 0)$ y*

Cuadro 1.1: Errores globales de los métodos

N	h	Euler	Runge orden 3	Kutta orden 4
6000	2.84420e-03	7.91523e+02	7.453224e-01	2.59667e-01
12000	1.42210e-03	2.05898e+01	1.46501e-01	1.22188e-02
24000	7.11051e-04	1.88980e+00	2.02286e-02	1.15963e-03
48000	3.55525e-04	5.80318e-01	2.90717e-03	6.55000e-05

la Luna en el punto $(1-\mu, 0)$*, y las coordenadas* (y_1, y_2) *del tercer cuerpo verifican las siguientes ecuaciones:*

$$\begin{aligned} y_1'' &= y_1 + 2y_2' - \mu' \frac{y_1+\mu}{D_1} \\ y_2'' &= y_2 - 2y_1' - \mu' \frac{y_2}{D_1} - \mu \frac{y_2}{D_2} \end{aligned} \tag{1.2}$$

Las distancias del cuerpo a la Tierra y a la Luna son:

$$\begin{aligned} D_1 &= ((y_1+\mu)^2 + y_2^2)^{3/2} \qquad \mu = 0{,}012277471 \\ D_2 &= ((y_1-\mu')^2 + y_2^2)^{3/2} \qquad \mu' = 1-\mu \end{aligned}$$

Existen valores iniciales como, por ejemplo,

$$\begin{aligned} &y_1(0) = 0{,}994 \qquad y_1'(0) = 0 \qquad y_2(0) = 0 \\ &y_2'(0) = -2{,}00158510637908252240537862224 \\ &t_{end} = 17{,}06521656015796255889172062 49 \end{aligned}$$

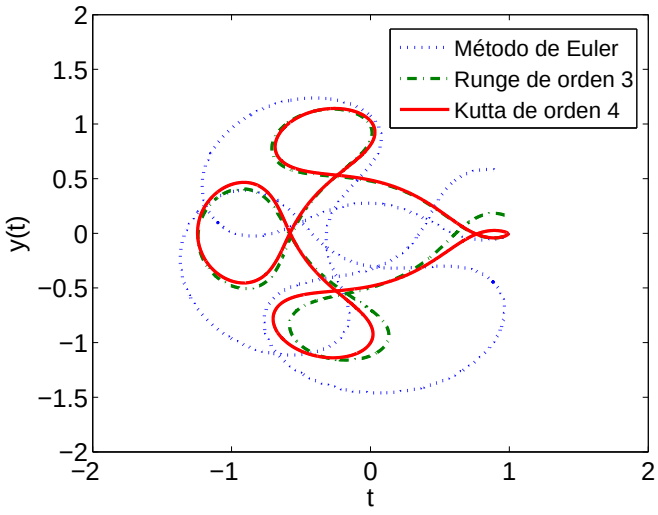
para los que la solución es periódica de periodo t_{end}*. El problema es* C^∞ *con la excepción de las singularidades*

$$y_1 = -\mu \qquad y_1 = 1-\mu' \qquad y_2 = 0$$

Considerando las derivadas primeras $y_3 = y_1'$*,* $y_4 = y_2'$ *transformamos este PVI en uno de tipo* (1.1) *de dimensión* 4 *de la forma usual. En el Cuadro 1.1, presentamos una comparación de los resultados obtenidos por tres métodos numéricos de un paso que detallaremos posteriormente en el Capítulo 2: Euler, Runge de orden 3 y Kutta de orden 4 integrados a paso fijo, es decir, prefijado un número de pasos* N*, los métodos calculan aproximaciones* y_n *a la solución en* $t_n = nh$*,* $n = 1, 2, ..., N$ *donde* $h = t_{end}/N$ *es el tamaño de paso. En dicho cuadro mostramos los errores globales del método que hemos obtenido en* t_{end} *para los pasos que hemos elegido. Los* **errores globales** *de un método de un paso se definen como*

$$e_n := \|y(t_n) - y_n\| \quad \textit{para} \quad n = 0, 1, ..., N$$

Figura 1.1: Runge y Kutta $N = 12000$, Euler $N = 48000$



donde $y(t)$ es la solución del problema de valor inicial e $\{y_n\}_{n=0}^{N}$ son los valores numéricos dados por el método en $\{t_n\}_{n=0}^{N}$. Se dice que **un método de un paso es convergente** *cuando los errores globales se aproximan a cero si $h \to 0$.*

En el Capítulo 2 veremos que el método de Euler es bastante simple de programar y sólo hace una evaluación de $f(t, y)$ en cada paso, pero su principal inconveniente es que para lograr buenas aproximaciones debemos tomar h muy pequeños, en cuyo caso los errores son proporcionales a h. Como vemos en el Cuadro 1.1 Euler empieza a dar errores aceptables a partir de $N = 48000$, obteniendo malos resultados para un número menor de pasos.

Los métodos Runge de orden 3 y Kutta de orden 4 hacen cuatro evaluaciones de $f(t, y)$ en cada paso y convergen de forma mas rápida con una longitud de paso no tan pequeña como el caso de Euler. Los errores en el caso del método de Runge son proporcionales a h^3 y en el caso del método de Kutta son proporcionales a h^4, por lo que su convergencia es mucho mejor. Esta diferencia en la velocidad a la que convergen los distintos métodos nos llevará al estudio del concepto de **orden de convergencia** *en el Capítulo 3. Así veremos que el método de Runge tiene orden 3 mientras que el de Kutta es 4.*

Si estudiamos los errores obtenidos por el método de Euler frente a los errores obtenidos por el método de Runge y el de Kutta, para que esta comparación sea justa tenemos que comparar el error obtenido para un número de paso N de Runge y Kutta con el obtenido por Euler con $4N$ pasos, pues así todos tienen un coste computacional similar es decir, el mismo número de evaluaciones de $f(t, y)$. Si comparamos los datos $N = 24000$ para Euler y $N = 6000$ para los otros dos métodos, vemos que aún así los errores obtenidos por los métodos de Runge y Kutta son mucho mejores que el de Euler. Completando el estudio de la convergencia obteniendo cotas del error local y global de los métodos en el Capítulo 4.

En la gráfica 1.1 dibujamos la solución obtenida para los tres métodos también con

un coste computacional equivalente. Ahí se ve que el único que es capaz de simular el comportamiento periódico de la solución exacta es el de Kutta. Este método no sólo tiene mayor velocidad de convergencia sino que también aporta una solución numérica de mejor calidad a lo largo de toda la integración. Por ello es necesario estudiar otras características de los métodos como la de su estabilidad. Dedicaremos el Capítulo 5 al estudio de la estabilidad de los métodos sobre problemas lineales.

Finalmente en el Capítulo 6 realizamos un experimento numérico con la ecuación del calor aplicando el **método de líneas** *con el que vemos una aplicación práctica de lo estudiado en los capítulos anteriores.*

Capítulo 2

Métodos Runge-Kutta

2.1. Método de Euler

El método de Euler fue uno de los primeros en usarse para la integración numérica de problemas de valor inicial (PVI) del tipo:

$$y' = f(t,y), \quad y(0) = y_0, \quad t \in [0,T], \quad y, f \in \mathbb{R}^m \tag{2.1}$$

Supondremos en lo que sigue que f es C^1 ($[0,T] \times \mathbb{R}^m$) y verifica una condición de Lipschitz respecto de y, es decir, $\exists L > 0$ tal que $\|f(t,y) - f(t,z)\| \leq L\|y - z\|$, $\forall y, z \in \mathbb{R}^m, \forall t \in [0,T]$.
Para poder aplicar el método se toma una partición cualquiera del intervalo de integración $[0,T]$ con $N+1$ puntos, que denotamos

$$P = \{0 = t_0 < t_1 < t_2 < ... < t_N = T\} \tag{2.2}$$

donde las cantidades $h_j = t_{j+1} - t_j$, $j = 0, 1, ..., N-1$ se denominan **tamaños de paso** del método.
Supongamos, cuando $m = 1$, que tenemos la curva solución de la ecuación diferencial $y(t)$ y trazamos la recta tangente a la curva en el punto dado por la condición inicial y_0. La ecuación de esta recta tangente sería $y = y_0 + m(t - t_0)$ con $m = y'(t_0) = f(t_0, y_0)$, esto es $y = y_0 + f(t_0, y_0)(t - t_0)$. Supongamos que t_1 es un punto cercano a t_0, $t_1 = t_0 + h$, y consideramos la aproximación $y(t_1) \approx y_1 = y_0 + h_0 f(t_0, y_0)$. Para obtener una aproximación $y_2 \approx y(t_2)$ repetimos el mismo proceso considerando el punto (t_1, y_1). Por tanto, si tenemos el punto (t_n, y_n) obtenemos la fórmula general del **método de Euler** como

$$y_{n+1} = y_n + h_n f(t_n, y_n) \tag{2.3}$$

Así el método de Euler va obteniendo aproximaciones $\{y_n\}_{n=0}^N$, a la solución en los puntos de la partición P partiendo de $y_0 = y(0)$.

Teorema 2.1.1 ***Teorema de la convergencia del método de Euler:*** *Consideremos el PVI* (2.1)*. Si* $\{y_n\}_{n=0}^N$ *es la solución numérica dada por el método de Euler sobre la partición* P (2.2) *para dicho PVI, y la solución exacta* $y(t)$ *verifica*

$$\|y''(t)\| \leq Y_2, \quad \forall t \in [0,T] \tag{2.4}$$

entonces,

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

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

Para la demostración de este Teorema necesitamos el siguiente resultado:

Lema 2.1.1 *La solución de la inecuación*

$$e_0 = 0, \quad e_n \leq C_n e_{n-1} + D_n, \quad C_n, D_n \in \mathbb{R}, \quad n = 1, 2, 3, ... \tag{2.5}$$

verifica $\quad e_n \leq \sum_{j=1}^{n} \left[\prod_{k=j+1}^{n} C_k \right] D_j \quad$ *denotando* $\prod_{k=n+1}^{n} C_k = 1.$

Demostración: Aplicamos inducción. Para $n = 1$, tenemos que $e_1 \leq C_1 e_0 + D_1 = D_1$, por lo que es trivial. Supongamos que $e_{n-1} \leq \sum_{j=1}^{n-1} \left[\prod_{k=j+1}^{n-1} C_k \right] D_j$. Probémoslo para n

$$e_n \leq C_n e_{n-1} + D_n \leq C_n \left[\sum_{j=1}^{n-1} \prod_{k=j+1}^{n-1} C_k D_j \right] + D_n \leq \left[\sum_{j=1}^{n} \prod_{k=j+1}^{n} C_k \right] D_j.$$

□

Veamos ahora la demostración del Teorema 2.1.1:
Demostración: Denotamos por $e_n := \|y_n - y(t_n)\|$ a los errores globales del método. Recordemos el desarrollo de Taylor de orden 1 con resto integral

$$g(x) = \quad g(a) + \frac{g'(a)}{1!}(x - a) + \int_a^x \frac{g(t)}{1!}(x - t)dt.$$

En nuestro caso, aplicándolo a la solución exacta $y(t)$ con $a = t_n$, $x = t_n + h_n$

$$y(t_n + h_n) = y(t_n) + y'(t_n)h_n + \int_{t_n}^{t_{n+1}} y''(s)(t_n + h_n - s)ds.$$

Tenemos que $y'(t_n) = f(t_n, y(t_n))$ y llamamos $l_n = \int_{t_n}^{t_{n+1}} y''(s)(t_n + h_n - s)ds$ por lo que nos queda,

$$y(t_n + h_n) = y(t_n) + h_n f(t_n, y(t_n)) + l_n. \tag{2.6}$$

Veamos qué ocurre con l_n. Hacemos el cambio de variable

$$s = t_n + \theta h_n, \quad ds = h_n d\theta \quad \theta \in (0, 1).$$

Como, $t_n + h_n - s = h_n(1 - \theta)$

$$l_n = \int_0^1 y''(t_n + \theta h_n)(h_n(1 - \theta))h_n d\theta = h_n^2 \int_0^1 y''(t_n + \theta h_n)(1 - \theta)d\theta.$$

Tomando normas y recordando (2.4)

$$\|l_n\| \leq h_n^2 Y_2 \int_0^1 (1-\theta)d\theta = h_n^2 \frac{Y_2}{2}$$

y de (2.6)

$$\begin{aligned} e_{n+1} \leq & \quad \|y_n - y(t_n)\| + h_n \underbrace{\|f(t_n, y_n) - f(t_n, y(t_n))\|}_{prop.Lipschitz} + \|l_n\| \\ & \leq e_n + h_n L \|y_n - y(t_n)\| + h_n^2 \frac{Y_2}{2} \leq e_n(1 + h_n L) + h_n^2 \frac{Y_2}{2}. \end{aligned}$$

Luego,

$$e_{n+1} \leq e_n(1 + h_n L) + h_n^2 \frac{Y_2}{2}.$$

Aplicando la propiedad de la función exponencial $1 + x \leq \exp(x),\ \forall x \in I\!R$, tenemos que $1 + h_n L \leq \exp(h_n L)$ y, por tanto, los errores globales satisfacen

$$e_0 = 0, \quad e_{n+1} \leq e_n \exp(h_n L) + h_n^2 \frac{Y_2}{2}, \quad n = 0, 1, ..., n-1. \tag{2.7}$$

Aplicando el Lema 1 con $C_k = \exp(L h_{k-1}),\ D_k = h_{k-1}^2 \dfrac{Y_2}{2}$

$$e_n \leq \sum_{j=0}^{n-1} \left(\prod_{k=j+1}^{n-1} \exp(L h_k) \right) \frac{Y_2}{2} h_j^2 \quad n = 1, 2, ..., N.$$

Como $\displaystyle\sum_{k=j+1}^{n-1} h_k = t_n - t_{j+1}$ y $h_j \leq h_{max}$, tenemos que

$$e_n \leq \frac{Y_2}{2} \sum_{j=0}^{n-1} h_j^2 \exp(L(t_n - t_{j+1}) \leq \underbrace{\left(\sum_{j=0}^{n-1} h_j \exp(-L t_{j+1}) \right)}_{(*)} \frac{Y_2}{2} \exp(L t_n) h_{max}.$$

Basta observar que $(*)$ es la suma inferior de Riemman de $\exp(-Lt)$ en $[0, t_n]$, que siempre será menor o igual que la integral que aproxima, $\displaystyle\int_0^{t_n} e^{-Lt} dt = \frac{1 - e^{-L t_n}}{L}$, por lo que

$$e_n \leq \left(\frac{1 - \exp(-L t_n)}{L} \right) \frac{Y_2}{2} \exp(L t_n) h_{max} = \left(\frac{\exp(L t_n) - 1}{L} \right) \frac{Y_2}{2} h_{max}.$$

Por tanto, el máximo resulta

$$\max_{t_n \in P} e_n \leq \max_{t_n \in P} \left(\frac{\exp(L t_n) - 1}{L} \right) \frac{Y_2}{2} h_{max} = \left(\frac{\exp(LT) - 1}{L} \right) \frac{Y_2}{2} h_{max}.$$

□

Observación 2.1.1 *En* (2.6) l_n *es lo que se suele llamar el* **error local del método**, *es decir, es el error que cometería el método de Euler si sólo diésemos un paso partiendo de la solución exacta en* t_n.

Resaltar la importancia de la demostración anterior ya que implícitamente nos dice que el error global en el método de Euler se comporta como Ch, donde C es una constante que dependerá del problema y h es el tamaño máximo del paso. Si se quisiera dar una precisión de 6 decimales por ejemplo, se necesitaría dar alrededor de un millón de pasos. Esto explica los resultados obtenidos en el problema de los tres cuerpos restringido del Ejemplo 1.0.1 del capítulo anterior.
En [5, Cap.I] se puede ver una versión de este teorema local, esto es, que demuestra la convergencia del método en un entorno $D = \{(t, y)/t_0 \leq t \leq X, |y - y_0| \leq b\}$ cuando el (2.1) es escalar y que además demuestra la existencia y unicidad de solución de dicho PVI.

2.2. Formulación general de los métodos Runge-Kutta (RK)

Se define un método **Runge-Kutta de s etapas** como un método numérico que dada una aproximación y_n a la solución del PVI (2.1) en un punto $t_n \in [0, T]$, nos da una aproximación a dicha solución en el punto $t_n + h \in [0, T]$, que denotamos por y_{n+1}, mediante las siguientes fórmulas:

$$\begin{cases} K_1 = f(t_n + c_1 h, y_n + h \sum_{j=1}^{s} a_{1j} K_j) \\ K_2 = f(t_n + c_2 h, y_n + h \sum_{j=1}^{s} a_{2j} K_j) \\ \vdots \\ K_i = f(t_n + c_i h, y_n + h \sum_{j=1}^{s} a_{ij} K_j) \quad 1 \leq i \leq s \end{cases} \tag{2.8}$$

$$y_{n+1} = y_n + h \sum_{i=1}^{s} b_i K_i \tag{2.9}$$

donde los vectores K_1,K_2,...,K_s se llaman **etapas del método RK**.
Se define la **tabla de Butcher** asociada al RK (2.9)-(2.8) como

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

donde la matriz $A = (a_{ij})_{i,j=1}^{s}$ se llama **matriz de coeficientes** del RK, el vector $c = (c_1, c_2, ..., c_s)^T$ es el **vector de nodos** o vector nodal del RK, y el vector $b = (b_1, b_2, ..., b_s)^T$ es el **vector de pesos** del RK. Con la ayuda de esta notación podemos denotar a un método RK como RK(A, b, c).
Con la representación de la tabla de Butcher trabajamos matricialmente con los coeficientes del método.
Según la forma de la matriz A de los métodos RK se suelen dividir en dos grandes grupos:

- Cuando la matriz A es triangular inferior estricta, el método RK se dice **explícito** (RKE), obteniéndose sus etapas de forma recursiva.

$$\begin{cases} K_1 = f(t_n, y_n) \\ K_2 = f(t_n + c_2 h, y_n + hK_1) \\ \vdots \\ K_i = f(t_n + c_i h, y_n + h\sum_{j=1}^{i-1} K_j), \quad 1 \leq i \leq s \end{cases} \tag{2.10}$$

$$y_{n+1} = y_n + h\sum_{i=1}^{s} b_j K_i \tag{2.11}$$

- Cuando $a_{ij} \neq 0$ para algún $j \geq i$, el método se dice **implícito** (RKI), y para calcular sus etapas tendremos que resolver un sistema implícito (2.8) de dimensión $s \times m$.

En esta memoria sólo se estudiarán los métodos RKE tomando como guía el Capitulo II.1 de [5] pero los resultados de esta sección son válidos para todos los métodos RK, incluidos los implícitos.

Definición 2.2.1 *Un método Runge-Kutta es de* **orden p** *si para problemas suficientemente regulares, se verifica que*

$$\|y(x_0 + h) - y_1\| \leq Ch^{p+1} \tag{2.12}$$

esto es, si la serie de Taylor para la solución exacta $y(x_0 + h)$ y para y_1 coincide hasta el término h^p.

Usualmente se supone que

$$\sum_{i=1}^{s} b_i = 1 \quad \text{y} \quad \sum_{i=1}^{s} a_{ij} = c_i \quad i = 1, 2, ..., s. \tag{2.13}$$

Estas condiciones, que ya asumió Kutta en su primera formulación de estos métodos en 1901, expresan que en todos los puntos donde f es evaluada se tienen aproximaciones de primer orden y simplifican enormemente la deducción de condiciones para los métodos de mayor orden.

Para ver esto en más detalle, podemos escribir las condiciones anteriores vectorialmente como $b^T e = 1,\ Ae = c$ donde $e = (1, 1, ..., 1)^T$, y el RK(A, b, c) se puede denotar RK(A, b). Las demostraciones de los siguientes teoremas son variantes simplificadas de las dadas en [4].

Teorema 2.2.1 *Un RK(A, b) tiene al menos orden 1* $\Leftrightarrow b^T e = 1$.

Demostración: Para tener al menos orden 1 debe verificarse que

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

esto es, si la serie de Taylor de $y(t_0 + h)$ y de y_1 coinciden hasta orden h.
El desarrollo de Taylor de la solución exacta, cuando $h \longrightarrow 0$, es

$$y(t_0 + h) = y(t_0) + hy'(t_0) + \mathcal{O}(h^2) = y_0 + hf(t_0, y_0) + \mathcal{O}(h^2). \tag{2.14}$$

Por otra parte, la solución numérica en $t_0 + h$ es $y_1 = y_0 + h\sum_{i=1}^{s} b_i K_i$ donde las etapas del método se desarrollan en (t_0, y_0) por Taylor como

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

Luego,

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

Comparando con (2.14)

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

Por tanto para que coincida al menos hasta orden h es necesario y suficiente que

$$\left(1 - \sum_{i=1}^{s} b_i\right) = 0 \Rightarrow \sum_{i=1}^{s} b_i = 1.$$

□

Teorema 2.2.2 *Un RK(A, b) da exactamente la misma aproximación cuando se aplica al PVI no autónomo*

$$y' = f(t, y), \quad y(t_0) = y_0 \quad y, f \in \mathbb{R}^m \tag{2.15}$$

que cuando se aplica sobre el problema autónomo asociado:

$$z' = g(z), \quad z(t_0) = z_0 \quad z, g \in \mathbb{R}^{m+1} \tag{2.16}$$

$$z = \begin{pmatrix} t \\ y \end{pmatrix}, \quad g(z) = \begin{pmatrix} 1 \\ f(t, y) \end{pmatrix}, \quad z_0 = \begin{pmatrix} t_0 \\ y_0 \end{pmatrix}$$

si y sólo si $b^T e = 1$ *y* $Ae = c$

Demostración: Aplicamos al RK(A,b) a ambos problemas,

$$K_i = f(t_n + c_i h, y_n + h\sum_{j=1}^{i-1} a_{ij}K_j), \quad y_{n+1} = y_n + h\sum_{i=1}^{s} b_i K_i. \tag{2.17}$$

$$U_i = g(z_n + h\sum_{j=1}^{i-1} a_{ij}U_j), \quad z_{n+1} = z_n + h\sum_{i=1}^{s} b_i U_i. \tag{2.18}$$

Tenemos que probar que

$$z_n = \begin{pmatrix} t_n \\ y_n \end{pmatrix} \quad \forall n \geq 0.$$

Lo probamos por inducción sobre n. Para $n = 0$ ya lo tenemos por definición de z_0.
Supongamos cierto para n y probémoslo para $n+1$. Denotamos

$$U_i = \begin{pmatrix} x_i \\ u_i \end{pmatrix} \quad z_{n+1} = \begin{pmatrix} \tau_{n+1} \\ w_{n+1} \end{pmatrix} \quad x_i,\ \tau \in I\!R, \quad u_i,\ w_{n+1} \in I\!R^m$$

Por la hipótesis de inducción

$$z_n + h\sum_{j=1}^{i-1} a_{ij}U_j = \begin{pmatrix} t_n \\ y_n \end{pmatrix} + h\sum_{j=1}^{i-1} a_{ij}\begin{pmatrix} x_j \\ u_j \end{pmatrix} = \begin{pmatrix} t_n + h\sum_{j=1}^{i-1} a_{ij}x_j \\ y_n + h\sum_{j=1}^{i-1} a_{ij}u_j \end{pmatrix}$$

Por tanto,

$$U_i = g\begin{pmatrix} t_n + h\sum_{j=1}^{i-1} a_{ij}x_j \\ y_n + h\sum_{j=1}^{i-1} a_{ij}u_j \end{pmatrix} = \begin{pmatrix} 1 \\ f\left(t_n + h\sum_{j=1}^{i-1} a_{ij}x_j, y_n + h\sum_{j=1}^{i-1} a_{ij}u_j\right) \end{pmatrix}$$

De aquí tenemos inmediatamente que $x_i = 1$ y $u_i = f\left(t_n + h\sum_{j=1}^{i-1} a_{ij}, y_n + h\sum_{j=1}^{i-1} a_{ij}u_j\right)$
para $1 \leq i \leq s$.
Como $u_1 = f(t_n, y_n) = K_1$, iterativamente sale que

$$u_i = K_i \Longleftrightarrow \sum_{j=1}^{i-1} a_{ij} = c_i \text{ para todo } i = 1, ..., s \Longleftrightarrow Ae = c$$

En este caso, análogamente por (2.18)

$$z_{n+1} = \begin{pmatrix} t_n \\ y_n \end{pmatrix} + h\sum_{i=1}^{s} b_i\begin{pmatrix} 1 \\ u_i \end{pmatrix} = \begin{pmatrix} t_n + h\sum_{i=1}^{s} b_i \\ y_n + h\sum_{i=1}^{s} b_i K_i \end{pmatrix}$$

Luego,

$$z_{n+1} = \begin{pmatrix} t_{n+1} \\ y_{n+1} \end{pmatrix} \iff t_{n+1} = t_n + h\sum_{i=1}^{s} b_i \iff \sum_{i=1}^{s} b_i = 1.$$

□

Ejemplo 2.2.1 *Veamos los métodos RKE que se aplican al problema del Ejemplo 1.0.1 con su tabla de Butcher:*
Método de Euler:

Formulación 1 etapa:

$$\begin{cases} K_1 = f_n \\ y_{n+1} = y_n + hK_1 \end{cases}$$

Tabla de Butcher

0	0
	1

Método de Runge (1905) de orden 3:

Formulación 4 etapas:

$$\begin{cases} K_1 = f_n \\ K_2 = f(t_n + \frac{h}{2}, y_n + \frac{h}{2}K_1) \\ K_3 = f(t_n + h, y_n + hK_2) \\ K_4 = f(t_n + h, y_n + hK_3) \\ y_{n+1} = y_n + h(\frac{k_1}{6} + \frac{2}{3}K_2 + \frac{1}{6}K_4) \end{cases}$$

Tabla de Butcher

0	0	0	0	0
$\frac{1}{2}$	$\frac{1}{2}$	0	0	0
1	0	1	0	0
1	0	0	1	0
	$\frac{1}{6}$	$\frac{2}{3}$	0	$\frac{1}{6}$

(2.19)

Método de kutta (1905) de orden 4:

Formulación 4 etapas:

$$\begin{cases} K_1 = f_n \\ K_2 = f(t_n + \frac{h}{2}, y_n + \frac{h}{2}K_1) \\ K_3 = f(t_n + \frac{h}{2}, y_n + \frac{h}{2}K_2) \\ K_4 = f(t_n + h, y_n + hK_3) \\ y_{n+1} = y_n + h(\frac{K_1}{6} + \frac{K_2}{3} + \frac{K_3}{3} + \frac{K_4}{6}) \end{cases}$$

Tabla de Butcher

0	0	0	0	0
$\frac{1}{2}$	$\frac{1}{2}$	0	0	0
$\frac{1}{2}$	0	$\frac{1}{2}$	0	0
1	0	0	1	0
	$\frac{1}{6}$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{6}$

(2.20)

Capítulo 3

Estudio del orden de los métodos Runge-Kutta. Series de Butcher

En este capítulo estudiaremos el orden de convergencia de los métodos Runge-Kutta (2.8)-(2.9) que verifican las condiciones (2.13). Como ya hemos visto en el Teorema 2.2.2 anterior, al exigir estas condiciones aseguramos que si un RK alcanza un orden p sobre **problemas autónomos**:

$$y' = f(y), \quad y(t_0) = 0, \quad y \in [t_0, T] \tag{3.1}$$

entonces alcanza el mismo orden sobre cualquier PVI (2.1). Por ello estudiaremos el orden sobre estos PVIs (3.1).
Expresando la Definición 2.2.1 de forma más general, sabemos que el RK(A, b) tiene orden $p \geq 1$, si y sólo si el **error local** verifica

$$l(t_0, h) := y(t_0 + h; t_0, y_0) - y_{RK}(t_0 + h; t_0, y_0) = \mathcal{O}(h^{p+1}) \tag{3.2}$$

donde $y(t_0 + h; t_0, y_0)$ e $y_{RK}(t_0 + h : t_0, y_0)$ denotan respectivamente la solución exacta y numérica en $t_0 + h$ que tienen el mismo valor inicial $y(t_0) = y_0$.
Por el desarrollo de Taylor de la solución exacta local $y(t; t_0, y_0)$ tenemos

$$y(t_0 + h; t_0, y_0) = \sum_{q=0}^{\infty} \frac{y^{(q}(t_0)}{q!} h^q = \sum_{q \geq 0} C_q h^q, \qquad C_q = \frac{y^{(q}(t_0)}{q!} \tag{3.3}$$

Si consiguiésemos un desarrollo parecido para la solución numérica,

$$y_{RK}(t_0 + h; t_0, y_0) = \sum_{q \geq 0} C_q^{RK} h^q$$

tendríamos una condición inmediata para calcular el orden del método RK:

$$l(t_0, h) = \mathcal{O}(h^{p+1}), \text{ si y sólo si, } C_q = C_q^{RK}, 0 \leq q \leq p.$$

Esta fue la idea que llevó a J.C. Butcher (Nueva Zelanda, 1933) a inventar un nuevo tipo de desarrollo en serie: **el desarrollo en serie de Butcher**. Para explicar este desarrollo

introducimos varias notaciones y conceptos nuevos.

Notación. Cuando estemos tratando con funciones vectoriales de la forma $f : U \subseteq \mathbb{R}^m \to \mathbb{R}^m$, denotaremos las derivadas parciales de cada componente de f_i de f como:

$$f_i^j := \frac{\partial f_i}{\partial x_j}, \quad f_i^{jk} := \frac{\partial^2 f_i}{\partial x_j \partial x_k}, ..., \quad f_i^{i_1,i_2,...,i_n} := \frac{\partial^n f_i}{\partial x_{i_1} ... \partial x_{i_n}}.$$

Cuando estemos trabajando con vectores en lugar de funciones vectoriales, los superíndices no denotarán derivadas, sino componentes.

Definición 3.0.2 *Consideramos una función cualquiera $f : U \subseteq \mathbb{R}^m \to \mathbb{R}^m$ e $y_0 \in U$. Si f admite derivadas k-ésimas en y_0, se define la* **derivada de Frechet k-ésima en** y_0 *como la aplicación*

$$\begin{aligned} F_i = f_{[y_0]}^{(k)} \equiv f^{(k)} : \quad & \mathbb{R}^m \times \mathbb{R}^m \times ...^{(k} \times \mathbb{R}^m \to \mathbb{R}^m \\ & \left(u^1, u^2, ..., u^k\right) \to f^k\left(u^1, u^2, ..., u^k\right) = \begin{pmatrix} F_1 \\ \vdots \\ F_m \end{pmatrix} \end{aligned} \tag{3.4}$$

donde si $k \geq 1$

$$F_i = f_i^{(k)}\left(u^1, u^2, ..., u^k\right) = \sum_{j_1, j_2, ..., j_k = 1}^{m} f_i^{j_1 j_2 ... j_k}(y_0) u_{j_1}^1 u_{j_2}^2 ... u_{j_k}^k, \quad 1 \leq i \leq m \tag{3.5}$$

y $f_{[y_0]}^0 = f(y_0)$.

3.1. Derivadas sucesivas y árboles ordenados monótonamente

Consideramos que la solución exacta de $y(t)$ del PVI local autónomo (3.1) es suficientemente derivable. En el punto $t_1 = t_0 + h$ para un tamaño de paso dado $h > 0$ tiene el desarrollo de Taylor (3.3). El problema de este desarrollo en serie de potencias es que depende de las derivadas de la solución exacta $y(t)$. En la práctica, lo que conocemos del PVI es la función derivada f. Haciendo uso de las derivadas de Frechet resolvemos este problema, como se hace en [3].

Proposición 3.1.1 *Sea y la solución exacta de PVI local autónomo* (3.1)*. Se tiene, en función de las derivadas de Frechet de f en y_0:*

$$\begin{aligned} & y'(t_0) = f \\ & y''(t_0) = f'(f) \\ & y'''(t_0) = f''(f, f) + f'(f'(f)) \\ & y''''(t_0) = f'''(f, f, f) + f''(f'(f), f) + f''(f, f'(f)) + f''(f'(f), f) + f'(f''(f, f)) + f'(f'(f'(f))) \end{aligned} \tag{3.6}$$

Demostración: Por definición tenemos $y'(t_0) = f(y_0)$. Para simplificar la notación vamos a denotar $y^{(n}(t) = y^{(n}_{i=1}(t_0)$ a la componente i-ésima de $y^{(n}(t_0)$ y obviaremos la dependencia de y_0 en las derivadas de f. Derivando $y_i'(t_0)$:

$$y_i''(t_0) = \sum_{j_1=1}^{m} f_i^{j_1} y'_{j_1} = \sum_{j_1=1}^{m} f_i^{j_1} f_{j_1} = f_i'(f).$$

Volviendo a derivar

$$\begin{aligned} y_i''' &= \sum_{j_1=1}^{m} \left[\left(\sum_{j_2=1}^{m} f_i^{j_1 j_2} y'_{j_2} \right) f_{j_1} + f_i^{j_1} \sum_{j_2=1}^{m} f_{j_1}^{j_2} y'_{j_2} \right] \\ &= \sum_{j_1 j_2}^{m} f_i^{j_1 j_2} f_{j_2} f_{j_1} + \sum_{j_1 j_2=1}^{m} f_i^{j_1} f_{j_1}^{j_2} f_{j_2} = f_i''(f,f) + f'(f'(f)). \end{aligned}$$

Derivando de nuevo cada sumando

$$\begin{aligned} y_i'''' &= \sum_{j_1 j_2=1}^{m} \left(\sum_{j_3=1}^{m} f_i^{j_1 j_2 j_3} f_{j_3} f_{j_2} f_{j_1} + f_i^{j_1 j_2} f_{j_2}^{j_3} f_{j_3} f_{j_1} + f_i^{j_1 j_2} f_{j_2} f_{j_1}^{j_3} f_{j_3} \right) + \\ &\sum_{j_1 j_2=1}^{m} \left(\sum_{j_3=1}^{m} f_i^{j_2 j_3} f_{j_3} f_{j_1}^{j_2} f_{j_2} + f_i^{j_2} f_{j_1}^{j_2 j_3} f_{j_3} f_{j_2} + f_i^{j_2} f_{j_1}^{j_2} f_{j_2}^{j_3} f_{j_3} \right) \\ &= \sum_{j_1 j_2 j_3=1}^{m} f_i^{j_1 j_2 j_3} f_{j_3} f_{j_2} f_{j_1} + \sum_{j_1 j_2 j_3=1}^{m} f_i^{j_1 j_2} f_{j_2}^{j_3} f_{j_3} f_{j_1} + \sum_{j_1 j_2 j_3=1}^{m} f_i^{j_1 j_2} f_{j_2} f_{j_1}^{j_3} f_{j_3} + \\ &\sum_{j_1 j_2 j_3=1}^{m} f_i^{j_2 j_3} f_{j_3} f_{j_1}^{j_2} f_{j_2} + \sum_{j_1 j_2 j_3=1}^{m} f_i^{j_2} f_{j_1}^{j_2 j_3} f_{j_3} f_{j_2} + \sum_{j_1 j_2 j_3=1}^{m} f_i^{j_2} f_{j_1}^{j_2} f_{j_2}^{j_3} f_{j_3}. \end{aligned}$$

Esta última expresión es la componente i-ésima de la suma de derivadas de Frechet dada en (3.6). □

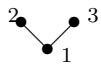
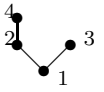
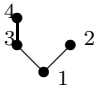
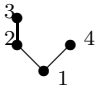
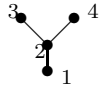
Para simplificar estos sumandos y poder generalizarlos a cualquier derivada se usan un tipo de aplicaciones, que llamaremos **árboles**, que se asocian a cada derivada de Frechet como veremos a continuación.

Definición 3.1.1 *Se llama árbol ordenado monótonamente de orden $q \geq 2$ a cualquier aplicación*

$$\tau : \{2, 3, ..., q\} \to \{1, 2, ..., q-1\} \qquad con\ \tau(i) < i, \quad i = 2, 3, ..., q.$$

Llamaremos $\rho(\tau) = q$ al orden del árbol τ. Por definición, $\tau_0 : 1 \to 0$ es el único **árbol de orden 1** y se denota $\emptyset$ al único **árbol de orden** 0, que es el árbol asociado a la derivada de orden 0, esto es, $y_i(t_0)$. LT_q denota el conjunto de todos los árboles de orden q y $LT = \{\emptyset, \tau_0\} \cup \{\cup_{q \geq 2} LT_q\}$ es el conjunto de todos los árboles ordenados monótonamente.

Cuadro 3.1: Relación entre las derivadas de Frechet y su árbol

Derivadas	**Orden**	**Árboles**	**Grafos**
$y_i' = f_i$	orden 1 $\tau : \{1\} \longrightarrow \{0\}$	τ_0	
$y_i'' = \sum_{j_2} f_i^{j_2} f_{j_2}$	orden 2 $\tau : \{2\} \longrightarrow \{1\}$	$\tau(2) = 1$	
$y_i''' = \sum_{j_2 j_3} f_i^{j_2 j_3} f_{j_2} f_{j_3}$	orden 3	$\tau(2) = \tau(3) = 1$	
$+ \sum_{j_2 j_3} f_i^{j_2} f_{j_2}^{j_3} f_{j_3}$	$\tau : \{2,3\} \longrightarrow \{1,2\}$	$\tau(2) = 1,\ \tau(3) = 2$	
$y_i'''' = \sum_{j_2 j_3 j_4} f_i^{j_2 j_3 j_4} f_{j_2} f_{j_3} f_{j_4}$	orden 4	$\tau(2) = \tau(3) = \tau(4) = 1$	
$+ \sum_{jkl} f_i^{j_2 j_3} f_{j_2}^{j_4} f_{j_3} f_{j_4}$	$\tau : \{2,3,4\} \longrightarrow \{1,2,3\}$	$\tau(2) = \tau(3) = 1,\ \tau(4) = 2$	
$+ \sum_{j_2 j_3 j_4} f_i^{j_3 j_2} f_{j_3}^{j_4} f_{j_2} f_{j_4}$		$\tau(2) = \tau(3) = 1,\ \tau(4) = 3$	
$+ \sum_{j_2 j_3 j_4} f_i^{j_2 j_4} f_{j_2}^{j_3} f_{j_4} f_{j_3}$		$\tau(2) = \tau(4) = 1,\ \tau(3) = 2$	
$+ \sum_{j_2 j_3 j_4} f_i^{j_2} f_{j_2}^{j_3 j_4} f_{j_3} f_{j_4}$		$\tau(2) = 1,\ \tau(3) = \tau(4) = 2$	
$+ \sum_{j_2 j_3 j_4} f_i^{j_2} f_{j_2}^{j_3} f_{j_3}^{j_4} f_{j_4}$		$\tau(2) = 1,\ \tau(3) = 2,\ \tau(4) = 3$	

Estas aplicaciones se representan gráficamente mediante los grafos dados en el Cuadro 3.1. En estos grafos, los puntos numerados se llaman **nodos** y el correspondiente al número 1 se llama **raíz** del árbol.
A continuación, relacionamos formalmente los árboles definidos anteriormente con las derivadas de la solución exacta del PVI para poder generalizar a cualquier derivada sucesiva. Como hemos visto, inductivamente cada derivada sucesiva $y_i^{(q}$ de la solución se va a descomponer en la suma de un número de derivadas de Frechet de f. Cada uno de los sumandos va a asociarse a un árbol $\tau \in LT_q$ y se va a llamar **la diferencial elemental de f asociada a τ en** y_0, y se denotara por $F_i(\tau)(y_0)$.
Por definición, se considera

$$F_i(\emptyset)(y_0) = y_0, \quad F_i(\tau_0) = f_i, \quad 1 \leq i \leq m, \quad \text{y si} \quad F(\tau)(y_0) = \begin{pmatrix} F_1(\tau)(y_0) \\ \vdots \\ F_m(\tau)(y_0) \end{pmatrix} \text{ se tiene:}$$

Teorema 3.1.1 *Existe una relación biunívoca entre todos y cada uno de los términos de la derivada q-ésima de y(t) con los árboles ordenados monótonamente de orden q, o sea,*

$$y^{(q}(t_0) = \sum_{\tau \in LT_q} F(\tau)(y_0). \tag{3.7}$$

Demostración: Para $q = 0$ es trivial. Para el caso $q = 1$ tenemos sólo un árbol de orden 1. Para cada componente $j_1 = 1, ..., m$:

$$y_{j_1}^{(1}(t_0) = f_{j_1} = F_{j_1}(\tau)(y_0)$$

Si añadimos una rama con un nodo obtenemos el caso $q = 2$, sólo un árbol de orden 2.

$$y_{j_1}^{(2}(t_0) = \sum_{j_2=1}^{m} f_{j_1}^{j_2} f_{j_2} = F_{j_1}(\tau)(y_0)$$

Para obtener $q = 3$, partimos del árbol de orden 2 y le añadimos una rama con un nuevo nodo a cada nodo así tenemos dos árboles de orden 3

$$\sum_{j_2,j_3=1}^{m} f_{j_1}^{j_2 j_3} f_{j_2} f_{j_3} = F_{j_1}(\tau)(y_0)$$

$$\sum_{j_2,j_3=1}^{m} f_{j_1}^{j_2} f_{j_2}^{j_3} f_{j_3} = F_{j_1}(\tau)(y_0)$$

Así,

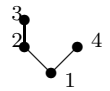
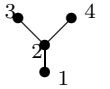
$$y_{j_1}^{(3}(t_0) = \sum_{j_2,j_3=1}^{m} f_{j_1}^{j_2 j_3} f_{j_2} f_{j_3} + \sum_{j_2,j_3=1}^{m} f_{j_1}^{j_2} f_{j_2}^{j_3} f_{j_3} = \sum_{\tau \in LT_3} F_{j_1}(\tau)(y_0)$$

Para obtener $q = 4$, se añade una nueva rama con un nuevo nodo a cada nodo de cada árbol de orden 3. Del primer árbol de orden 3 obtenemos los siguientes árboles de orden

4:

 $\sum_{j_2j_3j_4=1}^{m} f_{j_1}^{j_2j_3j_4} f_{j_2} f_{j_3} f_{j_4}$	 $\sum_{j_2j_3j_4=1}^{m} f_{j_1}^{j_2j_3} f_{j_2}^{j_4} f_{j_3} f_{j_4}$	 $\sum_{j_2j_3j_4=1}^{m} f_{j_1}^{j_2j_3} f_{j_3}^{j_4} f_{j_2} f_{j_4}$

Del segundo árbol de orden 3:

 $\sum_{j_2j_3j_4=1}^{m} f_{j_1}^{j_2j_4} f_{j_2} f_{j_3} f_{j_4}$	 $\sum_{j_2j_3j_4=1}^{m} f_{j_1}^{j_2} f_{j_2}^{j_3j_4} f_{j_3} f_{j_4}$	 $\sum_{j_2j_3j_4=1}^{m} f_{j_1}^{j_2} f_{j_2}^{j_3} f_{j_3}^{f_4} f_{j_4}$

Así,

$$y^{(4}(t_0) = \sum_{\tau \in LT_4} F(\tau)(y_0)$$

y no hay más árboles ordenados monótonamente de orden 4.
Inductivamente para un q general suponemos cierto (3.7) y tenemos que verlo para $q+1$. Para obtener los $y^{(q+1}(t_0)$ tenemos que derivar cada uno de los $F(\tau)$. De esta manera, para cada τ_q de orden q, se van a generar al derivar q árboles de orden $q+1$ añadiendo una rama y un nodo a cada nodo del árbol τ_q. Es decir, si $\tau_q \in LT_q$, $F_{j_1}(\tau_q)(y_0) = \sum_{j_2 \dots j_q}^{m} f_{j_1}^{I_1} f_{j_2}^{I_2} \cdots f_{j_q}^{I_q}$ donde cada I_k es el conjunto de todos los índices correspondientes a las derivadas de f_{j_k} definidas por τ_q. Entonces derivando

$$\frac{d}{dt} F_{j_1}(\tau_q)(y_0) = \frac{d}{dt} \left(\sum_{j_2 \dots j_q=1}^{m} f_{j_1}^{I_1} f_{j_2}^{I_2} \dots f_{j_q}^{I_q} \right)$$

$$= \left(\sum_{j_2 \dots j_q=1}^{m} \left(\sum_{j_{q+1}=1}^{m} (f_{j_1}^{I_1})^{j_{q+1}} f_{j_{q+1}} \right) f_{j_2}^{I_2} \dots f_{j_q}^{I_q} \right) + \dots + \left(\sum_{j_2 \dots j_q=1}^{m} f_j^{I_1} f_j^{I_2} \dots \left(\sum_{j_{q+1}=1}^{m} (f_{j_q}^{I_q})^{j_{q+1}} f_{j_{q+1}} \right) \right)$$

$$= \sum_{j_2 \dots j_{q+1}} (f_{j_1}^{I_1})^{j_{q+1}} f_{j_2}^{I_2} \dots f_{j_q}^{I_q} f_{j_{q+1}} + \dots + \sum_{j_2 \dots j_{q+1}} f_{j_1}^{I_1} f_{j_2}^{I_2} \dots (f_{j_q}^{I_q})^{j_{q+1}} f_{j_{q+1}}$$

y se obtienen los árboles de orden $q+1$ correspondientes a añadir una nueva rama y un nodo a cada nodo de τ_q. Además, al estar monótonamente ordenados cuando se hace con todos los árboles posibles de LT_q se obtienen todos los posibles de LT_{q+1}. □

Aplicando (3.7), obtenemos directamente el siguiente resultado:

Corolario 3.1.1.1 *Si $y(t)$ es analítica en y_0, se tiene el desarrollo en potencias de h:*

$$y(t_0 + h) = \sum_{\tau \in LT} F(\tau)(y_0) \frac{h^{\rho(\tau)}}{\rho(\tau)!}.$$

Por tanto hemos obtenido la relación que buscábamos, es decir, se tiene el desarrollo de la solución $y(t)$ del PVI, en función de las diferenciales de f y en potencias de h. Este desarrollo todavía se puede simplificar más, ya que muchos de los sumandos se repiten.

Definición 3.1.2 *Dos árboles τ_1 y τ_2 se dice que son* **equivalentes** *y se denota por $\tau_1 \sim \tau_2$, si y sólo si son del mismo orden q y existe una permutación σ del conjunto $\{1, 2, ..., q\}$ con $\sigma(1) = 1$ tal que*

$$\sigma \circ \tau_1 = \tau_2 \circ \sigma$$

sobre el conjunto $\{2, 3, ..., q\}$.

Teorema 3.1.2 *La relación " $\sim$ " entre árboles de LT definida anteriormente es una relación de equivalencia en LT, es decir, verifica las siguientes condiciones:*

$$a)\ \tau \sim \tau. \quad b)\ \tau_1 \sim \tau_2 \Rightarrow \tau_2 \sim \tau_1. \quad c)\ \tau_1 \sim \tau_2 \ y\ \tau_2 \sim \tau_3 \Rightarrow \tau_1 \sim \tau_3.$$

Demostración:

a) Tomando $\sigma =$ identidad, $\sigma(i) = i,\ \forall i$ es inmediato.

b) Si σ es la permutación que hace $\tau_1 \sim \tau_2$ basta hacer $\bar{\sigma} = \sigma^{-1}$ como permutación para tener $\tau_2 \sim \tau_1$ o sea, $\bar{\sigma} \circ \tau_2 = \tau_1 \circ \bar{\sigma}$.

c) Si σ_1 es la permutación que hace $\tau_1 \sim \tau_2$ y σ_2 es la permutación que hace $\tau_2 \sim \tau_3$ considerando $\bar{\sigma} = \sigma_2 \circ \sigma_1$ tenemos $\tau_1 \sim \tau_3$, es decir, $\bar{\sigma} \circ \tau_1 = \tau_3 \circ \bar{\sigma}$ □

Por tanto, podemos definir el conjunto cociente de esta relación de equivalencia, esto es, el conjunto de las clases de equivalencia. Se define el **conjunto de los árboles de raíz** como el conjunto cociente $T = LT/\sim$. Se define también

$$T_q = \{\tau \in T \text{ tal que } \ \rho(\tau) = q\}.$$

Además, para cada $\tau \in T$ se define el cardinal de $\tau = \alpha(\tau) = card(\tau)$ =número de árboles de LT equivalentes a τ. La representación gráfica de estas clases de equivalencia $\tau \in T$ es igual que para los $\tau \in LT$ pero sin etiquetas en los nodos.
A partir de esta relación, es inmediato que

$$\tau_1 \sim \tau_2, \text{ entonces } F(\tau_1)(y_0) = F(\tau_2)(y_0).$$

Ejemplo 3.1.1 *Sean dos árboles*

$$\tau_1: \quad F_i(\tau_1)(y_0) = \sum_{j_2 j_3 j_4 = 0} f_i^{j_2 j_4} f_{j_2}^{j_3} f_{j_3} f_{j_4} \quad \tau_1(2) = \tau_1(4) = 1,\ \tau_1(3) = 2$$
$$\tau_2: \quad F_i(\tau_2)(y_0) = \sum_{k_2 k_3 k_4 = 0} f_i^{k_2 k_3} f_{k_2}^{k_4} f_{k_3} f_{k_4} \quad \tau_2(2) = \tau_2(3) = 1,\ \tau_2(4) = 2$$

es claro que $F(\tau_1)(y_0) = F(\tau_2)(y_0)$ simplemente haciendo la permutación de los contadores $k_2 = j_2,\ k_3 = j_4,\ k_4 = j_3$.

Corolario 3.1.2.1 *Si la solución exacta $y(t)$ del PVI es analítica en t_0:*

1. $y^{(q}(t_0) = \sum_{\tau \in T_q} \alpha(\tau) F(\tau)(y_0)$.

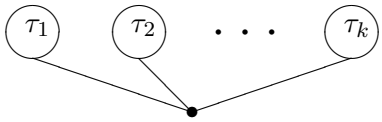
2. $y(t_0 + h) = \sum_{\tau \in T} \alpha(\tau) F(\tau)(y_0) \frac{h^{\rho(\tau)}}{\rho(\tau)!}$.

Con este último resultado ya tenemos un desarrollo más simple que el dado en el Corolario 3.1.1.1 . Pero todavía este desarrollo se puede simplificar más.

Definición 3.1.3 *Dados k árboles no vacíos $\tau_1, \tau_2, ..., \tau_k \in T$, denotaremos como*

$$\tau = \{\tau_1, \tau_2, ..., \tau_k\}$$

al nuevo árbol construido de la forma



cuyo orden es $\rho(\tau) = \rho(\tau_1) + \rho(\tau_2) + ... + \rho(\tau_k) + 1$. En este caso, a los árboles $\tau_1, \tau_2, ..., \tau_k$ se les llama **árboles hijos de** τ.

Nótese que el orden de colocación de los árboles hijos es indiferente. También se puede descomponer la diferencial elemental de τ en función de las derivadas de Frechet de las diferenciales elementales de sus hijos de forma unívoca, como vemos en el siguiente Teorema que se demuestra inductivamente (ver [3]):

Teorema 3.1.3 *Si $\tau = \{\tau_1, \tau_2, ..., \tau_k\}$, entonces*

$$F(\tau)(y_0) = f_0^k(F(\tau_1)(y_0), ..., F(\tau_k)(y_0)) = \left(\sum_{j_1,...,j_k=1}^{m} f_i^{j_1...j_k}(y_0) F_{j_1}(\tau_1)(y_0) ... F_{j_k}(\tau_k)(y_0) \right)_{i=1}^{m}.$$

3.2. Series de Butcher y condiciones de orden

En la sección anterior hemos visto un desarrollo en serie de la solución exacta del PVI (2.1) en función de las diferenciales elementales de f. En esta estudiaremos un desarrollo similar de la solución numérica de un RK(A, b).

Definición 3.2.1 *Sea a una aplicación $a : LT \to \mathbb{R}$ que verifica $a(\tau_1) = a(\tau_2)$ si $\tau_1 \sim \tau_2$. Se define la* **serie de Butcher asociada a** a **en el punto** y_0 *como la serie formal en potencias de h:*

$$B(a, y_0)(h) := \sum_{\tau \in LT} a(\tau) F(\tau)(y_0) \frac{h^{\rho(\tau)}}{\rho(\tau)!} = \sum_{\tau \in T} \alpha(\tau) a(\tau) F(\tau)(y_0) \frac{h^{\rho(\tau)}}{\rho(\tau)!}.$$

Definición 3.2.2 *Dada $g : I_0 \subseteq \mathbb{R} \to \mathbb{R}^m$ de clase $C^\infty(I_0)$ siendo I_0 un entorno en el origen, se dice que g es* **representable en serie de Butcher** *si y sólo si existe una aplicación*

$$a : T \to \mathbb{R} \text{ tal que } g^{(q}(0) = \sum_{\tau \in T_q} \alpha(\tau) a(\tau) F(\tau)(y_0), \quad q = 0, 1, 2, ... \tag{3.8}$$

entendiéndose que $\emptyset$ es el único árbol de orden 0 con $F(\emptyset)(y_0) = y_0$ y τ_0 el único árbol de orden 1 con $F(\tau_0)(y_0) = y_0$ y τ_0. Es decir, si

$$g(h) = B(a, y_0)(h), \quad h \in I_0.$$

Teorema 3.2.1 *Sea U abierto de $\mathbb{R}^m$, $y_0 \in U$, $f : u \subset \mathbb{R}^m \to \mathbb{R}^m$ analítica y sea $g(h)$ representable como B-serie con $g(h) = B(a, y_0)(h)$, $a(\emptyset) = 1$. Entonces $hf(g(h))$ es una B-serie*

$$hf(g(h)) = B(\bar{a}, y_0)(h)$$

donde $\bar{a} : LT \to \mathbb{R}$ está definida recurrentemente por

$$\bar{a}(\emptyset) = 0, \quad \bar{a}(\tau_0) = 1, \quad \bar{a}(\tau) = \rho(\tau) a(\tau_1) a(\tau_2) ... a(\tau_k) \quad \textit{para} \quad \tau = \{\tau_1, ..., \tau_k\}.$$

Demostración: Desarrollaremos aquí la demostración dada en [3]. Sea

$$c(h) = hf(g(h)).$$

Queremos ver que $c(h)$ es una B-serie, esto es existe una aplicación $\bar{a} : LT \to \mathbb{R}$ tal que

$$c^{(i}(0) = \sum_{\tau \in LT_i} \overline{a}(\tau) F(\tau)(y_0) \text{ para } i = 0, 1, ... \tag{3.9}$$

- Para $i = 0$, $c(0) = 0$, $LT_0 = \{\emptyset\}$. Por tanto necesariamente se verifica (3.9) si y sólo si $\bar{a}(\emptyset) = 0$ ya que $F(\emptyset)(y_0) = y_0$.

- Para $i = 1$. Derivando $c'(h) = f(g(h)) + h\dfrac{d}{dh}(f(g(h)))$. Por hipótesis $g(0) = y_0$. Por tanto, $c'(0) = f(g(0)) = f(y_0)$. Por otro lado $LT_1 = \{\tau_0\}$ por lo que (3.9) se verifica sii $f(y_0) = \overline{a}(\tau_0) F(\tau_0)(y_0)$. Por tanto necesariamente $\bar{a}(\tau_0) = 1$.

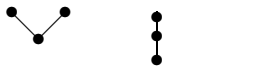
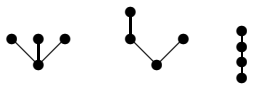
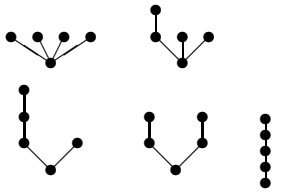
- Para derivar en los casos $i \geq 2$, aplicamos la regla de Leibnitz

$$\begin{aligned} c^{(i}(h) = \quad & \frac{d^i}{dh^i}\left[hf(g(h))\right] = \sum_{k=0}^{i} \left(\begin{array}{c} i \\ k \end{array} \right) \frac{d^k}{dh^k}(h) \frac{d^{i-k}}{dh^{i-k}}(f(g(h))) \\ & = \left(\begin{array}{c} i \\ 0 \end{array} \right) h \frac{d^i}{dh^i}(f(g(h))) + \left(\begin{array}{c} i \\ 1 \end{array} \right) 1 \frac{d^{i-1}}{dh^{i-1}}(f(g(h))) \end{aligned}$$

  lo que implica que

$$c^{(i}(0) = i \frac{d^{i-1}}{dh^{i-1}}(f(g(0))).$$

Cuadro 3.2: Correspondencia entre derivadas y árboles

$i=1$	$(f\circ g)(0)=f^{(0)}_{[y_0]}$	
$i=2$	$(f\circ g)'(0)=f'_{[y_0]}(g'_0)$	
$i=3$	$(f\circ g)''(0)=f''_{[y_0]}(g'_0,g'_0)+f'_{[y_0]}(g''_0)$	
$i=4$	$(f\circ g)'''(0)=f'''_{[y_0]}(g'_0,g'_0,g'_0)+3f''_{[y_0]}(g''_0,g'_0)+f'_{[y_0]}g'''_0$	
$i=5$	$(f\circ g)''''(0)=f''''_{[y_0]}(g'_0,g'_0,g'_0,g'_0)+6f'''_{[y_0]}(g''_0,g'_0,g'_0)$ $+4f''_{[y_0]}(g'''_0,g'_0)+3f''_{[y_0]}(g''_0,g''_0)+f'_{[y_0]}g''''_0$	

Aplicando la regla de la cadena, comenzamos a calcular las derivadas sucesivas $(f\circ g)'(0),(f\circ g)''(0),...$ correspondientes a $c''(0),c'''(0),...$ y usando las derivadas de Frechet obtenemos la columna central del Cuadro 3.2, teniendo en cuenta que $g(0)=y_0$ y denotando $g_0^{(k}=g^{(k}(0)$.

Aplicando un razonamiento análogo al que se vio en la sección anterior para calcular las derivadas sucesivas, se observa que las derivadas sucesivas de la función están en correspondencia biunívoca con los árboles en los que sólo hay ramificaciones en la raíz correspondientes al número de veces que se deriva f. Esta correspondencia que se ve claramente en la columna del Cuadro 3.2, donde el coeficiente por el que aparece cada derivada de Frechet es el cardinal del árbol considerado.

Para formalizar esto denotaremos $u=\{u_1,u_2,...,u_k\}\in LT_i$ los árboles cuyos hijos son de la forma

$$u_j=\ \vdots\ \ \text{con } \rho(u_j)=I_j,\quad \sum_{j=1}^{k}I_j+1=i \tag{3.10}$$

se denominan **árboles de tipo raíz** de orden i

Llamamos SLT al conjunto de todos los árboles de este tipo y se caracterizan por

$$SLT=\left\{u\in LT/\ card(u^{-1}(l))\leq 1,\ l=2,3,...\right\}$$

esto es, cualquier $l \geq 2$ puede ser imagen sólo de uno o de ningún elemento de $\{2, 3, ..., \rho(u)\}$. Ver ejemplo en la Observación 3.

Análogamente al caso de LT, denotamos por $SLT_i = \{u \in SLT / \rho(u) = i\}$ y a cada $u \in SLT$, como vimos en el Cuadro 3.2, si $u = \{u_1, ..., u_k\}$ le asociamos la derivada de Frechet

$$G(u)(h) = f^{(k)}_{[g(h)]}(g^{(I_1}(h), ..., g^{(I_k}(h)). \tag{3.11}$$

Por tanto, la suma de todas las derivadas de Frechet (3.11) de todos los árboles de SLT_i, $i \geq 2$ nos da la derivada

$$(f \circ g)^{(i-1)}(0) = \sum_{u \in SLT_i} G(u)(0) = \sum_{u \in SLT_i} f^{(k)}_{y_0}(g^{(I_1}(0), ..., g^{(I_k}(0)).$$

Como $g(h) = B(a, y_0)(h)$, entonces por (3.8), denotando $F(\tau_j) = F(\tau_j)(y_0)$ tenemos por linealidad de derivadas de Frechet que

$$\begin{aligned}(f \circ g)^{(i-1)}(0) &= \sum_{u \in SLT_i} f^{(k)}_{[y_0]} \left(\sum_{\tau_1 \in LT_{I_1}} a(\tau_1)F(\tau_1), ..., \sum_{\tau_k \in LT_{I_k}} a(\tau_k)F(\tau_k) \right) \\ &= \sum_{u \in SLT_i} \sum_{\tau_1 \in LT_{I_1}} ... \sum_{\tau_k \in LT_{I_k}} a(\tau_1)...a(\tau_k) f^{(k)}_{[y_0]}(F(\tau_1), ..., F(\tau_k)).\end{aligned}$$

Por el Teorema 3.1.3, $f^{(k)}_{[y_0]}(F(\tau_1)(y_0), F(\tau_2)(y_0), ..., F(\tau_k)(y_0)) = F(\{\tau_1, ..., \tau_k)\}(y_0)$. Luego,

$$c^{(i}(0) = i(f \circ g)^{(i-1)}(0) = i \sum_{u \in SLT_i} \sum_{\tau_1 \in LT_{I_1}} ... \sum_{\tau_k \in LT_{I_k}} a(\tau_1)...a(\tau_k)F(\{\tau_1, ..., \tau_k\}).$$

Teniendo en cuenta que $\rho(\{\tau_1, ..., \tau_k\}) = \sum_{i=1}^{k} \rho(I_j) + 1 = i$, tenemos

$$c^{(i}(0) = \sum_{u \in SLT_i} \sum_{\tau_1 \in LT_{I_1}} ... \sum_{\tau_k \in LT_{I_k}} \rho(\{\tau_1, ..., \tau_k\})a(\tau_1)a(\tau_2)...a(\tau_k)F(\{\tau_1, ..., \tau_k\}).$$

Si llamamos $\bar{a}(\{\tau_1...\tau_k\}) = \rho(\{\tau_1...\tau_k\})a(\tau_1)...a(\tau_k)$ tenemos

$$c^{(i}(0) = \sum_{u \in SLT_i} \sum_{\tau_1 \in LT_{I_1}} ... \sum_{\tau_k \in LT_{I_k}} \bar{a}(\{\tau_1, ..., \tau_k\}) F(\{\tau_1, ..., \tau_k\})(y_0). \tag{3.12}$$

Por tanto para demostrar el teorema tenemos que ver que el sumatorio (3.12) es igual a

$$\sum_{\tau \in LT_i} \overline{a}(\tau)F(\tau)(y_0). \tag{3.13}$$

Esto se tiene si y sólo si las dos sumas (3.12) y (3.13) tienen los mismos sumandos. Veamos que todos los sumandos de (3.12) están contenidos en la suma (3.13) y viceversa:

$''\subseteq''$ Sea $u \in SLT_i \Rightarrow u = \{u_1, ..., u_k\}, \rho(u_j) = I_j, j = 1, ..., k$ para ciertos $\{I_j\}$. Cualquier $\tau = \{\tau_1, ..., \tau_k\}$ con $\tau_j \in LT_{I_j}$ tiene orden $\rho(\tau) = \sum_{j=1}^{k} \rho(\tau_j) + 1 = i \Rightarrow \tau \in LT_i \Rightarrow$ todos los sumandos están en (3.13).

$''\supseteq''$ Sea $\tau \in LT_i$, $\tau = \{\tau_1, ..., \tau_k\}\, (i \geq 2)$ con $\rho(\tau_j) = I_j, j = 1, ..., k$. Para cada I_j fijo consideramos $u = \{u_1, ..., u_k\}$ con $\rho(u_j) = I_j$ nodos de la forma (3.10) y, por tanto, $\bar{a}(\tau)F(\tau)(y_0)$ está entre los sumandos de (3.12). □

Observación 3.2.1 *En la demostración anterior hemos introducido el conjunto SLT. Para aclarar mejor esta definición tomemos por ejemplo dos árboles u y v*

$$u : 2 \to 1 \quad 3 \to 1 \quad 4 \to 3 \quad 5 \to 4 \qquad v : 2 \to 1 \quad 3 \to 1 \quad 4 \to 3 \quad 5 \to 3.$$

Es claro que $u \in SLT$ *ya que* $card(u^{-1}(l)) \leq 1$ *con* $l = 2, 3, 4$*, mientras que* $v \in LT$ *pues* $card(v^{-1}(3)) = 2$*.*

Para aplicar esta serie de Butcher a los RKE (2.10) hay que reescribirlos. Estos métodos (2.10) sobre el PVI autónomo (3.1) llamando $g_i = hK_i$, dando un paso de t_0 a $t_0 + h$ resulta:

$$\begin{cases} g_i &= hf(y_0 + \sum_{j=1}^{i-1} a_{ij} g_j) \qquad i = 1, ..., s. \\ y_1 &= y_0 + \sum_{j=1}^{s} b_j g_j. \end{cases} \tag{3.14}$$

Lema 3.2.1 *Cada* g_i *es representable como serie de Butcher en la forma*

$$g_i = B(\varphi_i, y_0)(h)$$

donde $\varphi_1(\tau_0) = 1$, $\varphi_1(\tau) = 0$, $\forall \tau \neq \tau_0$ *y para todo* $i = 2, ..., s$ *por* $\varphi_i(\emptyset) = 0$, $\varphi_i(\tau_0) = 1$

$$\varphi_i(\tau) = \rho(\tau) \sum_{j_1, ..., j_k = 1}^{i-1} a_{ij_1} a_{ij_2} ... a_{ij_k} \varphi_{j_1}(\tau_1) ... \varphi_{j_k}(\tau_k) \quad si \quad \tau = \{\tau_1, ..., \tau_k\}. \tag{3.15}$$

Demostración: Denotamos $\varphi_0 : LT \longrightarrow \mathbb{R}$ tal que $\varphi_0(\emptyset) = 1, \varphi_0(\tau) = 0\ \forall \tau \neq \emptyset$ por lo que $y_0 = B(\varphi_0, y_0)(h)$.

- Para $i = 1$ es claro que $g_1 = hf(y_0) = B(\varphi_1, y_0)(h)$ donde $\varphi_1(\tau_0) = 1, \varphi_1(\tau) = 0\ \forall \tau \neq \tau_0$.
- Suponemos cierto para todos los g_j con $j \leq i - 1$, $i \geq 2$

$$g_j = B(\varphi_j, y_0)(h) \quad j = 1, 2, ..., i - 1.$$

- Veamos que vale para i. Por un lado

$$g_i = hf(y_0 + \sum_{j=1}^{i-1} a_{ij} g_j) = hf(Y_i)$$

donde $Y_i = y_0 + \sum_{j=1}^{i-1} a_{ij} g_j$. Por hipótesis de inducción

$$Y_i = B(\varphi_0, y_0)(h) + \sum_{j=1}^{i-1} a_{ij} B(\varphi_j, y_0)(h).$$

Por linealidad de las derivadas,

$$Y_i = B(\bar{\varphi}_i, y_0)(h), \quad \bar{\varphi}_i = \varphi_0 + \sum_{j=1}^{i-1} a_{ij}\varphi_j.$$

Por el Teorema 3.2.1, $g_i = B(\varphi_i, y_0)$ donde $\varphi_i(\emptyset) = 0, \quad \varphi(\tau_0) = 1$

$$\varphi_i(\tau) = \rho(\tau)\bar{\varphi}_i(\tau_1)...\bar{\varphi}_i(\tau_k) \text{ si } \tau = \{\tau_1, ..., \tau_k\}.$$

En consecuencia de lo anterior para $\tau = \{\tau_1, ..., \tau_k\}$

$$\varphi_i(\tau) = \rho(\tau) \left[\varphi_0(\tau_1) + \sum_{j_1=1}^{i-1} a_{ij_1}\varphi_{j_1}(\tau_1)\right] ... \left[\varphi_0(\tau_k) + \sum_{j_k=1}^{i-1} a_{ij_k}\varphi_{j_k}(\tau_k)\right]$$

donde $\varphi_0(\tau_j) = 0, \forall j = 1, ..., k$. Por tanto, se obtiene el sumatorio (3.15). □

Para expresar estas funciones de forma matricial necesitamos introducir un nuevo producto vectorial:

Definición 3.2.3 *Dados dos vectores cualesquiera* $u, v, \in \mathbb{R}^s$, *se define el* **producto directo de u y v** *como el vector de* $\mathbb{R}^s$ $u \cdot v = (u_1 v_1 ... u_m v_m)^T = (u_i v_i)_{i=1}^m$.

Observemos que definiendo la aplicación $\varphi : T \to \mathbb{R}^m$ por $\varphi(\tau) = (\varphi_1(\tau)...\varphi_m(\tau))^T$ las ecuaciones (3.15) las podemos escribir matricialmente como

$$\varphi(\emptyset) = 0, \quad \varphi(\tau_0) = e, \quad \varphi(\tau) = \rho(\tau)A\varphi(\tau_1) \cdot ... \cdot A\varphi(\tau_k) \text{ si } \tau = \{\tau_1, ..., \tau_k\}. \tag{3.16}$$

Corolario 3.2.1.1 *Podemos representar la solución numérica de un método* $RK(A, b)$ *como*

$$y_1 = y_0 + \sum_{i=1}^{s} b_i B(\varphi_i, y_0)(h) = B(\omega, y_0)(h) \tag{3.17}$$

donde $\omega : T \to \mathbb{R} \quad \omega = \varphi_0 + \sum_{i=1}^{s} b_i\varphi_i$ *o, matricialmente,*

$$\omega(\emptyset) = 1, \ \omega(\tau) = b^T\varphi(\tau), \ \forall\tau \text{ con } \rho(\tau) \geq 1.$$

Cuadro 3.3: Condiciones de orden

Orden p	1	2	3	4	
Condiciones	$b^T e = 1$	$b^T = 1/2$	$b^T c^2 = 1/3$ $b^T Ac = 1/6$	$b^T c^3 = 1/4$ $b^T Ac^2 = 1/12$	$b^T (c \cdot Ac) = 1/8$ $b^T A^2 c = 1/24$

Todavía hay una forma más sencilla y manejable de expresar las etapas del método RK como B-serie. Para ello definimos la siguiente aplicación, $\gamma : T \to I\!R$

$$\gamma(\emptyset) = \gamma(\tau_0) = 1, \quad \gamma(\tau) = \rho(\tau)\gamma(\tau_1)\cdots\gamma(\tau_k) \ \text{ si } \ \tau = \{\tau_1, \cdots, \tau_k\}. \tag{3.18}$$

Lema 3.2.2 *La función* $\Phi(\tau) = \dfrac{1}{\gamma(\tau)}\varphi(\tau)$ *queda determinada recurrentemente por:*

$$\Phi(\emptyset) = 0, \quad \Phi(\tau_0) = e, \quad \Phi(\tau) = (A\Phi(\tau_1)) \cdot (A\Phi(\tau_2)) \cdot \ldots \cdot (A\Phi(\tau_k)) \quad si \ \ \tau = \{\tau_1, \cdots, \tau_k\}. \tag{3.19}$$

Demostración: Por (3.16) y (3.18):

$$\Phi(\emptyset) = \frac{1}{\gamma(\emptyset)}\varphi(\emptyset) = 0, \quad \Phi(\tau_0) = \frac{1}{\gamma(\tau_0)}\varphi(\tau_0) = e$$

$$\Phi(\tau) = \frac{\rho(\tau)}{\rho(\tau)\gamma(\tau_1)...\gamma(\tau_k)}(A\varphi(\tau_1)) \cdot \ldots \cdot (A\varphi(\tau_k)) = A\left(\frac{\varphi(\tau_1)}{\gamma(\tau_1)}\right) \cdot \ldots \cdot A\left(\frac{\varphi(\tau_k)}{\gamma(\tau_k)}\right).$$

□

A partir de estas aplicaciones podemos enunciar el siguiente teorema:

Teorema 3.2.2 *Un método RK*(A, b) *tiene orden* $p \geq 1$ *si y sólo si*

$$b^T \Phi(\tau) = \frac{1}{\gamma(\tau)}, \quad \forall \tau \ con \ \rho(\tau) \leq p$$

donde $\gamma(\tau)$ *y* $\Phi(\tau)$ *están definidas por* (3.18) *y* (3.19) *respectivamente.*

Demostración: Por el Corolario 3.1.2.1, tenemos que la solución exacta verifica

$$y(t_0 + h) = y_0 + \sum_{\tau \in T, \rho(\tau) \geq 1} \alpha(\tau) F(\tau)(y_0) \frac{h^{\rho(\tau)}}{\rho(\tau)!}$$

y la solución numérica del método RK por (3.17)

$$y_{RK}(t_0 + h) = y_1 = y_0 + \sum_{\tau \in T, \rho(\tau) \geq 1} \alpha(\tau)\omega(\tau) F(\tau)(y_0) \frac{h^{\rho(\tau)}}{\rho(\tau)!}$$

Comparando las soluciones obtenemos que el método tiene orden $p \geq 1$ si y sólo si $\omega(\tau) = 1,\ \forall \tau \in T$ con $1 \leq \rho(\tau) \leq p$. Por (3.18) y el Lema 3.19, es claro que $\omega(\tau) = b^T \Phi(\tau)\varphi(\tau)$, de lo que se tiene directamente el Teorema. □

Así podemos establecer el orden de convergencia de un método numérico haciendo uso de condiciones algebraicas con los coeficientes A, b y c de la tabla de Butcher. Por ejemplo, las condiciones que debe satisfacer un método para tener orden p desde 1 hasta 4 están dadas en el Cuadro 3.3 donde $c^k = \left(c_s^k\right)_{i=1}^s$ entendiendo que para tener orden p los coeficientes del método deben verificar todas las condiciones hasta orden p. Es decir, en el caso del método de Euler (2.3) se verifica sólo la primera condición por lo que es de orden 1, el método de Runge de orden 3 (2.19) verifica las condiciones 1, 2 y 3. El método Kutta de orden 4 (2.20) las cumple todas hasta orden 4.

Capítulo 4

Convergencia de métodos Runge-Kutta

En el Teorema 3.2.2 del capítulo anterior se ha visto un criterio para determinar de forma práctica qué orden de convergencia puede alcanzar un método RK dado basado en la serie de Butcher. Pero en ningún momento se consideró la convergencia de dichas series o qué ocurre cuando la solución del PVI no es analítica. En este capítulo abordaremos el problema de dar cotas rigurosas del error de estos métodos. De forma análoga a la demostración del Teorema 2.1.1, para acotar el error global primero hay que acotar el error local, como se ve en [5, Cap.II.3]. Consideremos el PVI

$$y' = f(t,y), \quad y(t_0) = y_0 \quad t \in [t_0, t_f\,], \quad y, f \in I\!R^m \tag{4.1}$$

suponiendo que tiene solución única $y(t)$ en $[t_0, t_f]$. Denotemos por

$$T_\delta = \{(t,y) : \|y - y(t)\| \leq \delta\}$$

un tubo de amplitud δ alrededor de la solución exacta $y(t)$. Para integrar (4.1), usamos un método RK(A,b) (2.8)-(2.9) de s etapas de orden $p \geq 1$, es decir, que verifica el Teorema 3.2.2.

Teorema 4.0.3 Acotación del error local: *Si el RK(A,b) es de orden p y $f \in C^p(T_\delta)$, existe $h^* > 0$ tal que para todo $|h| \leq h^*$ se tiene*

$$\|y(t+h) - y_{RK}(t+h;t,y(t))\| \leq \frac{h^{p+1}}{(p+1)!}\left[\max_{s\in[0,1]} \|y^{(p+1}(t+sh)\| + (p+1)\sum_{i=1}^{s} |b_i| \max_{s\in[0,1]} \|K_i^{(p}(sh)\|\right]$$

donde

$$K_i^{(p}(h) = \frac{\partial^p K_i}{\partial h^p}(h).$$

Demostración: Como $f \in C^p(T_\delta)$, la solución exacta $y(t)$ del PVI (4.1) es derivable hasta orden $p+1$, por lo que aplicando el desarrollo de Taylor con resto integral hasta

orden p de $y(t)$ se llega a que

$$y(t+h) = \sum_{j=0}^{p} \frac{h^j}{j!} y^{(j}(t) + \frac{h^{p+1}}{p!} \int_0^1 (1-\theta)^p y^{(p+1}(t+\theta h) d\theta \tag{4.2}$$

Por otra parte,

$$y_{RK}(t+h; t, y(t)) = y(t) + h \sum_{i=1}^{s} b_i K_i(h) \tag{4.3}$$

donde las etapas vienen dadas por

$$K_i(h) = f(t + c_i h, y(t) + h \sum_{j=1}^{i-1} a_{ij} K_j(h)), \quad 1 \leq i \leq s.$$

Así, existe un $h^* \geq 0$ tal que $K_i \in C^p([0, h^*])$ pues $f \in C^p(T_\delta)$. Aplicando Taylor de orden $p-1$ a cada $K_i(h)$ en 0, obtenemos:

$$K_i(h) = \sum_{j=0}^{p-1} K_i^{(j}(0) \frac{h^j}{j!} + \frac{h^p}{(p-1)!} \int_0^1 (1-\theta)^{p-1} K_i^{(p}(\theta h) d\theta. \tag{4.4}$$

Llevándolo a (4.3) tenemos:

$$\begin{aligned} y_{RK}(t+h; t, y(t)) = y(t) &+ \sum_{j=1}^{p} \frac{h^j}{(j-1)!} \left(\sum_{i=1}^{s} b_i K_i^{(j-1}(0) \right) \\ &+ \frac{h^{p+1}}{(p-1)!} \sum_{i=1}^{s} b_i \int_0^1 (1-\theta)^{p-1} K_i^{(p}(\theta h) d\theta. \end{aligned} \tag{4.5}$$

Como el RK es de orden p, las potencias de h en (4.2) y (4.5) hasta orden p coinciden, por tanto, restando ambas ecuaciones y tomando normas queda

$$\begin{aligned} &\|y(t+h) - y_{RK}(t+h; t, y(t))\| \leq \\ &h^{p+1} \left\{ \frac{1}{p!} \int_0^1 |1-\theta|^p \max_{\zeta \in [0,1]} \|y^{(p+1}(t+\zeta h)\| d\theta + \frac{1}{(p-1)!} \sum_{i=1}^{s} |b_i| \int_0^1 |1-\theta|^{p-1} \max_{\zeta \in [0,1]} \|K_i^{(p}(\zeta h)\| d\theta \right\} \\ &= h^{p+1} \left\{ \frac{1}{(p+1)!} \max_{\zeta \in [0,1]} \|y^{(p+1}(t+\zeta h)\| + \frac{1}{p!} \sum_{i=1}^{s} |b_i| \max_{\zeta \in [0,1]} \|K_i^{(p}(\zeta h)\| \right\}. \end{aligned}$$

Obteniendo finalmente el resultado del teorema. □

Para acotar al error global se necesita del siguiente resultado:

Definición 4.0.4 *Se llama* **función incremento** *de un método RKE* (2.10) *a la función*

$$\phi(t, y, h) = \sum_{i=1}^{s} b_i K_i(t, y, h)$$

Teorema 4.0.4 *La función incremento $\phi(t,y,h)$ tiene una constante de Lipschitz Ω respecto de y en $[0,T]\times T_\delta\times[0,h_0]$, o sea, $\forall\ t\in[0,T],\forall y,\bar{y}\in T_\delta,\forall h\in[0,h_0]$*

$$\|\phi(t,y,h)-\phi(t,\bar{y},h)\|\leq\Omega\|y-\bar{y}\|,\quad \Omega=L|b|^T(I-h_0L|A|)^{-1}e$$

siendo L es la constante de Lipschitz de f respecto de y en T_δ y $|b|^T=(|b_1|,...,|b_s|)$, $|A|=(|a_{ij}|)_{i,j=1}^s$, $h_0>0$.

Demostración: Denotamos $\phi=\sum_{i=1}^s b_iK_i$ y $\bar{\phi}=\sum_{i=1}^s b_i\bar{K}_i$ siendo $K_i=K_i(t,y,h)$, $\bar{K}_i=K_i(t,\bar{y},h)$. Usando que f es Lipschitz respecto de y, veamos que $\|K_i-\bar{K}_i\|\leq L_i\|y-\bar{y}\|$ para ciertas constantes L_i.
Para $i=1,\|K_1-\bar{K}_1\|=\|f(t,y)-f(t,\bar{y})\|\leq L_1\|y-\bar{y}\|$, con $L_1=L$. Supongamos cierto $\|K_j-\bar{K}_j\|\leq L_j\|y-\bar{y}\|$ para $\forall 1\leq j\leq i-1$. Veámoslo para i.

$$\|K_i-\bar{K}_i\|\leq L\|y-\bar{y}\|+Lh\sum_{j=1}^{i-1}|a_{ij}|\|K_j-\bar{K}_j\|\leq\left(L+Lh\sum_{j=1}^{i-1}|a_{ij}|L_j\right)\|y-\bar{y}\|$$

$$\leq\left(L+Lh_0\sum_{j=1}^{i-1}|a_{ij}|L_j\right)\|y-\bar{y}\|\leq L_i\|y-\overline{y}\|.$$

Lo pasamos a notación vectorial

$$V=\begin{pmatrix}L_1\\L_2\\L_3\\\vdots\\L_s\end{pmatrix}=\begin{pmatrix}L\\L+Lh_0|a_{21}|L_1\\L+Lh_0(|a_{31}|L_1+|a_{32}|L_2)\\\vdots\\L+Lh_0\sum_{j=1}^{s-1}|a_{sj}|L_j\end{pmatrix}=\begin{pmatrix}L\\L\\L\\\vdots\\L\end{pmatrix}+Lh_0|A|\begin{pmatrix}L_1\\L_2\\L_3\\\vdots\\L_s\end{pmatrix}.$$

Así, $V=Le+Lh_0|A|V$ donde e es el vector unitario y A es la matriz de coeficientes del método. Es decir $(I-Lh_0|A|)V=Le$. Como $\rho(A)=0$ por ser una matriz triangular estricta existe la inversa de $(I-Lh_0|A|)$ [1]. Luego,

$$V=L(I-Lh_0|A|)^{-1}e.$$

Por tanto,

$$\|\phi-\bar{\phi}\|\leq\sum_{i=1}^s|b_i|\|K_i-\bar{K}_i\|\leq\left(\sum_{i=1}^s|b_i|L_i\right)\|y-\bar{y}\|$$
$$=|b|^TV\|y-\bar{y}\|=L(|b|^T(I-Lh_0\|A\|)^{-1}e)\|y-\bar{y}\|.$$

□

Teorema 4.0.5 Acotación del error global: *Si un $RK(A,b)$ es de orden p y $f \in C^p(T_\delta)$, y consideramos una partición cualquiera $P = \{t_0 < t_1 < ... < t_N = t_f\}$ de $[t_0, t_f]$ con $h_{max} := \max_j h_j \leq h^*$ donde h^* es el valor dado en el Teorema 4.0.3 y L la constante de Lipschitz de f, entonces se tiene que*

$$\|y(t_n) - y_n\| \leq h_{max}^p \frac{C}{\Omega}\left[e^{\Omega(t_n - t_0)} - 1\right], \quad n = 0, 1, ..., N$$

donde

$$C = \frac{1}{(p+1)!}\left\{\max_{t\in[t_0,t_f]} \|y^{(p+1}(t)\| + (p+1)\sum_{i=1}^{s} |b_i| \max_{t\in[t_0,t_f]\ h\in[0,h^*]} \|K_i^{(p}(t,h)\|\right\} \tag{4.6}$$

$$K_i(t,h) = f(t + c_i h, y(t) + h\sum_{j=1}^{s} a_{ij} K_j(t,h)),\ 1 \leq i \leq s$$

y Ω es la constante de Lipshitz de la función incremento de ϕ del RK dada en el Teorema 4.0.4.

Demostración: Esta demostración sigue un razonamiento análogo al del Teorema 2.1.1. Denotemos

$$e_n := \|y(t_n) - y_n\|, \qquad \bar{y}_n := y_{RK}(t_n; t_{n-1}, y(t_{n-1})).$$

Claramente el error local en t_n es $l_n = \|y(t_n) - \bar{y}_n\|$. Por tanto, sumando y restando en e_n el vector $\bar{y}_n$:

$$\begin{aligned} e_n &\leq l_n + \|\bar{y}_n - y_n\| \\ &\leq l_n + \|y(t_{n-1}) + h_{n-1}\phi(t_{n-1}, y(t_{n-1}), h_{n-1})) - (y_{n-1} + h_{n-1}\phi(t_{n-1}, y_{n-1}, h_{n-1}))\|. \end{aligned}$$

Aplicando el Teorema 4.0.4 tenemos que

$$e_n \leq l_n + [\|y_{n-1} - y(t_{n-1})\| + \Omega h_{n-1}\|y_{n-1} - y(t_{n-1})\|] \leq l_n + (1 + \Omega h_{n+1})\, e_{n-1}.$$

Como $1 + \Omega h_{n-1} \leq \exp(\Omega h_{n-1})$, obtenemos la inecuación en diferencias

$$e_0 = 0 \quad e_n \leq \exp(\Omega h_{n-1}) e_{n-1} + l_n \quad n \leq 1.$$

Resolviéndola como hicimos en la demostración del Teorema 2.1.1 usando en Lema 2.1.1 con $C_n = exp(\Omega h_{n-1})$, $D_n = l_n$

$$e_n \leq \sum_{j=0}^{n-1} \exp\left(\Omega \sum_{k=j+1}^{n-1} h_k\right) l_{j+1} \leq \exp(\Omega t_n) \sum_{j=0}^{n-1} \exp(-\Omega t_{j+1}) l_{j+1}.$$

Aplicando el Teorema 4.0.3, se tiene que $l_n \leq C h_{n-1}^{p+1} \leq C h_{n-1} h_{max}^p, \forall n$ con C dado por (4.6) por lo que

$$e_n \leq C \exp(\Omega t_n)\left(\sum_{j=0}^{n-1} \exp(-\Omega t_{j+1}) h_j\right) h_{max}^p.$$

El paréntesis anterior engloba a la suma inferior de Riemann de $\exp(-\Omega t)$ en $[t_0, t_n]$, por lo que se acota por su integral y se consigue finalmente el resultado del Teorema. $\square$

Capítulo 5

Estabilidad lineal de los métodos Runge-Kutta

A continuación vamos a realizar un experimento numérico. Consideramos dos problemas de valor inicial lineales con coeficientes constantes no homogéneos propuestos en [9, Cap.6]:
Problema 1:

$$\begin{pmatrix} y_1' \\ y_2' \end{pmatrix} = \begin{pmatrix} -2 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} + \begin{pmatrix} 2\sin t \\ 2(\cos t - \sin t) \end{pmatrix}, \quad \begin{pmatrix} y_1(0) \\ y_2(0) \end{pmatrix} = \begin{pmatrix} 2 \\ 3 \end{pmatrix}$$

Problema 2:

$$\begin{pmatrix} y_1' \\ y_2' \end{pmatrix} = \begin{pmatrix} -2 & 1 \\ 998 & -999 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} + \begin{pmatrix} 2\sin t \\ 999(\cos t - \sin t) \end{pmatrix}, \quad \begin{pmatrix} y_1(0) \\ y_2(0) \end{pmatrix} = \begin{pmatrix} 2 \\ 3 \end{pmatrix}$$

Observamos que los problemas son muy similares. Ambos presentan la misma solución exacta:

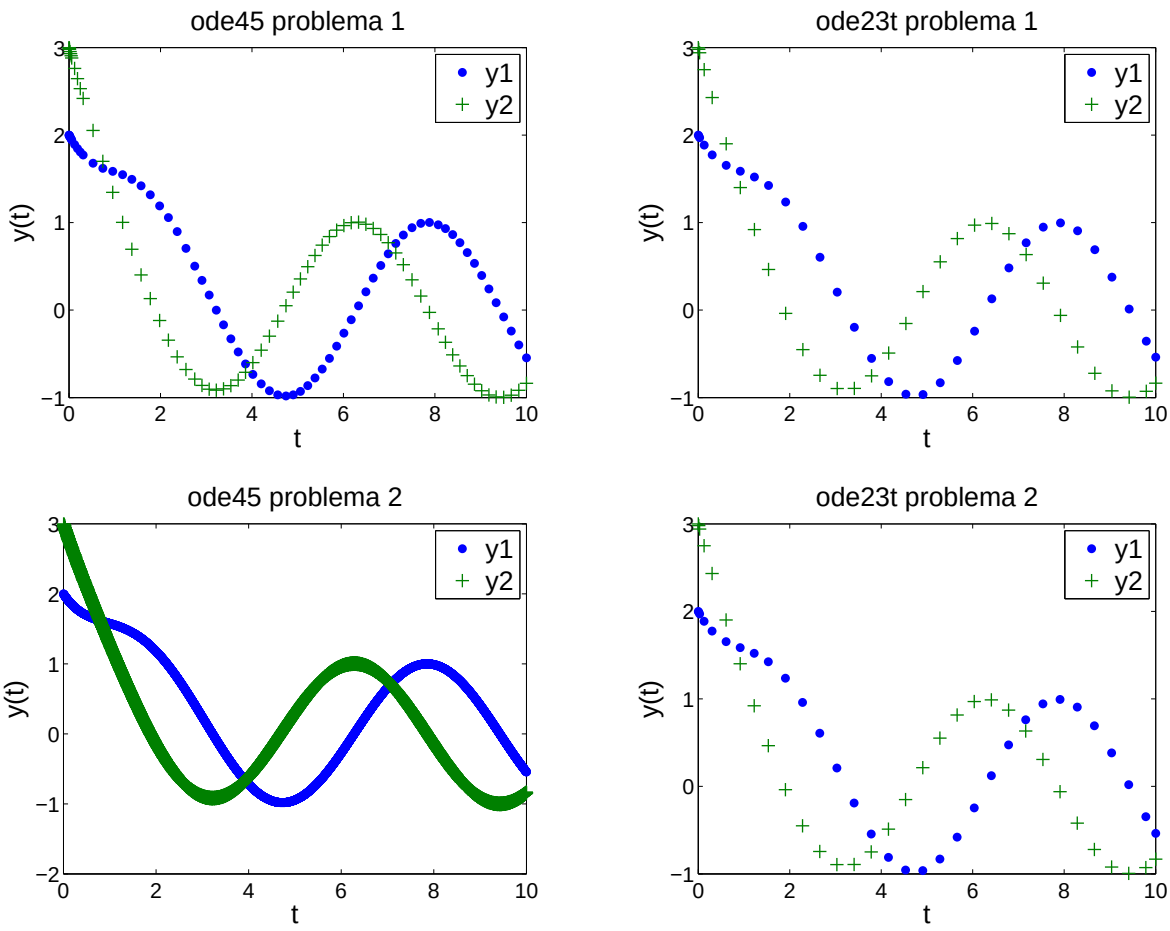
$$y_1(t) = 2e^{-t} + \sin t, \quad y_2(t) = 2e^{-t} + \cos t.$$

Integramos los dos problemas con dos rutinas de Matlab [7] diferentes:

- Rutina **ode45**: función que implementa a paso variable el par explícito RK DOPRI(4,5). Es un método explícito de 7 etapas y orden 5 que usa estimación del error local con el par encajado de orden 4.
- Rutina **ode23t**: es una implementación de la regla trapezoidal, es decir, esta basado en un RK implícito de 2 etapas y orden 2, que usa un método de orden 3 encajado para estimar el error local.

Para ver los detalle de ambas rutinas, recomendamos ver la documentación del Matlab [7]. Usamos para los dos métodos la misma tolerancia del error absoluto y relativo $ABSTOL = RELTOL = 10^{-2}$ y paso inicial $h_0 = 0{,}1$, $t \in [0, 10]$. En la Figura 5.1 dibujamos las componentes de las aproximaciones numéricas de ambos métodos.
Con la rutina **ode45** los resultados obtenidos para los problemas son muy diferentes. En el primer problema da 17 pasos y 115 evaluaciones de la función, mientra que para el

Figura 5.1: Soluciones numéricas de ode45 y ode23t



segundo problema da 3011 pasos y 19,309 evaluaciones de la función. Por tanto, vemos que para conseguir una buena solución para el segundo problema necesitamos un coste computacional exagerado, lo que se refleja en la gráfica en la que los puntos se superponen. Con la rutina **ode23t** los resultados obtenidos sobre los dos problemas son iguales, da 31 pasos y 40 evaluaciones de la función. En este método conseguimos una solución buena y barata computacionalmente en ambos problemas.
Así tenemos dos PVIs muy similares que se comportan de manera muy distinta cuando los abordamos numéricamente. Vemos que el problema 1 es resuelto sin mayor dificultad con ambos métodos mientras que para la resolución del problema 2 el esfuerzo computacional del **ode45** lo hace impracticable. Este fenómeno que aparece en el problema 2 se llama **stiffness**, y el problema se dice **stiff**. El problema 1 no lo es.
No hay una definición matemática estricta de lo que es un problema stiff. La definición más aceptada es que **un problema stiff es un problema sobre los que los métodos explícitos no van bien**.
Un modelo que explica la diferencia del comportamiento de los métodos explícitos e implícitos es el **problema lineal escalar test**

$$y'(t) = \lambda y(t), \quad y(0) = y_0, \quad t \geq 0, \quad \lambda \in \mathbb{C}. \tag{5.1}$$

cuya solución exacta es bien conocida, $y(t) = \exp(\lambda t) y_0$. Para que el problema (5.1) sea estable es necesario que $Re\lambda \leq 0$ pues si $Re\lambda > 0$ y partimos de dos valores iniciales diferentes y_0 e $\bar{y}$:

$$|y(t) - \bar{y}|(t) = |\exp(\lambda t)(y_0 - \bar{y}_0)| = \exp(Re\lambda t)|y_0 - \bar{y}_0| \longrightarrow \infty \text{ si } t \longrightarrow \infty$$

donde $y(t), \bar{y}(t)$ son las soluciones de (5.1) correspondientes a y_0 e $\bar{y}_0$ respectivamente.
Por tanto, vamos a aplicar al problema (5.1) con $Re\lambda \leq 0$ dos métodos sencillos:

- **Euler explícito**: a paso fijo h, $t_n = nh$

$$y_{n+1} = y_n + hf(t_n, y_n) = y_n + h\lambda y_n \quad n = 0, 1, 2, ...$$

Si llamamos $z = \lambda h$ se obtiene $y_{n+1} = (1 + z)y_n$. Si partimos de otro valor inicial $\bar{y}_0$ tenemos $\bar{y}_{n+1} = (1 + z)y_0$. Así

$$|y_{n+1} - \bar{y}_{n+1}| = |1 + z||y_n - \bar{y}_n| = |1 + z|^{n+1}|y_0 - \bar{y}_0|.$$

Para que no se vaya a ∞ cuando $n \longrightarrow \infty$ ($t_n \longrightarrow \infty$) tiene que ocurrir $|1+z| \leq 1$, o sea z tiene que estar en el interior del círculo de centro -1 y radio 1. Por ejemplo, si $\lambda = -1000$, esta condición sólo se da si $0 \leq h \leq 2/1000$. Por tanto, para integrar este problema en $t \in [0, 10]$ se tendría que dar al menos 5000 pasos.

- **Euler implícito**:

$$y_{n+1} = y_n + hf(t_{n+1}, y_{n+1}) = y_n + h\lambda y_{n+1}.$$

Igual que en el caso anterior, llamamos $z = \lambda h$ y despejando obtenemos $y_{n+1} = \left(\frac{1}{1-z}\right) y_n$.
Si tomamos otro valor inicial $\bar{y}_0$ tenemos $\bar{y}_{n+1} = \left(\frac{1}{1-z}\right) \bar{y}_n$, luego,

$$|y_{n+1} - \bar{y}_{n+1}| = \frac{1}{|1-z|}|y_n - \bar{y}_n| = \frac{1}{|1-z|^{n+1}}|y_0 - \bar{y}_0|.$$

Así para que este método vaya bien sobre este problema se necesitará que

$$\frac{1}{|1-z|} \leq 1 \Longleftrightarrow |1-z| \geq 1$$

es decir, z tiene que estar en el exterior del círculo de centro 1 y de radio 1. Es evidente que para cualquier λ con $Re\lambda \leq 0$ y cualquier $h \geq 0$ se verifica esta condición y por tanto el método funcionará bien sobre este problema para cualquier h.
Este diferente comportamiento de muchos problemas sobre métodos implícitos y explícitos, que no tiene que ver con su orden, es la principal característica de los **problemas stiff**.

5.1. Problemas stiff

A pesar de la falta de definición estricta, lo que sí se ha estudiado en detalle son las características más habituales de dichos problemas destacando como una de las más determinantes la siguiente:

Si los autovalores $\lambda_i(t)$ de la matriz jacobiana $\partial f/\partial y$ de la función derivada de f de un PVI

$$y' = f(t,y), \qquad y(t_0) = y_0, \quad t \in [t_0, t_f]$$

sobre la solución $y(t)$ verifican las dos condiciones siguientes:

1. *Existe k tal que $Re\lambda_k(t) = \mu$ con $|\mu|$ pequeño de modo que $Re\lambda_i(t) \leq \mu$, $\forall i$.*

2. *Existe algún j tal que $Re\lambda_j(t)$ es negativo con módulo grande.*

Entonces el PVI es stiff.

En el ejemplo numérico visto en la sección anterior, el problema 1 tiene autovalores -1 y -3 mientras que el problema 2 los autovalores son -1 y -1000. Por eso el segundo es stiff, mientras que el problema 1 no lo es.
El estudio del comportamiento de los métodos RK en general sobre el problema lineal test (5.1) se llama **teoría de la estabilidad lineal** o **A-estabilidad** de los métodos RK.

5.2. A-estabilidad de los métodos Runge-Kutta

Siguiendo las ideas de [6, Cap.IV.2 y Cap.IV.3], aplicamos un método RK(A,b) de s etapas general

$$\begin{cases} K_i = f(t_n + c_i h, y_n + h\sum_{j=1}^{s} a_{ij}K_j), \quad 1 \leq i \leq s \\ y_{n+1} = y_n + h\sum_{i=1}^{s} b_i K_i \end{cases} \tag{5.2}$$

a la ecuación lineal test (5.1). Para cada $1 \leq i \leq s$:

$$K_i = \lambda \left[y_n + h \sum_{j=1}^{s} a_{ij} K_j \right].$$

Considerando el vector $K = (K_1, ..., K_s)^T \in \mathbb{R}^s$ es fácil ver que estas ecuaciones se escriben de forma compacta como:

$$K = \lambda e y_n + \lambda h A K \Rightarrow (I - \lambda h A) K = \lambda e y_n. \tag{5.3}$$

Considerando h suficientemente pequeño para que exista $(I - \lambda h A)^{-1}$ y llamando $z = \lambda h$, se tiene

$$y_{n+1} = (1 + z b^T (I - zA)^{-1} e) y_n = R(z) y_n \tag{5.4}$$

donde $R(z)$ se llama **función de estabilidad de un método RK**(A, b).
Aplicándolo reiteradas veces obtenemos $y_{n+1} = R(z)^{n+1} y_0$ y tomando otro valor inicial $\bar{y}_0$, resulta

$$|y_{n+1} - \bar{y}_{n+1}| = |R(z)|^{n+1} |y_0 - \bar{y}_0|.$$

Para evitar que tienda a ∞ cuando $n \longrightarrow \infty$ es necesario que $|R(z)| \leq 1 \ \forall z$. Por tanto, a partir de esta restricción para la función $R(z)$ llamamos al conjunto

$$\mathcal{S} = \{z \in \mathbb{C} \ : |R(z)| \leq 1\}$$

dominio de A-estabilidad de un RK(A, b). Denotando $\mathbb{C}^- = \{z \in \mathbb{C} : Re z \leq 0\}$ se dice que un RK(A, b) es **A-estable** si y sólo si $\mathbb{C}^- \subseteq \mathcal{S}$.
Por otra parte, se define $[-\beta_R, 0]$ como el mayor segmento del eje real negativo contenido en $\mathcal{S}$ y β_R se llama **frontera de estabilidad real del método**, valor que cobra mayor importancia cuando se integran problemas stiff con autovalores reales negativos como se ve en [8] y [11].
Veamos algunas propiedades de esta función $R(z)$:

Teorema 5.2.1 *Si un RK(A, b) tiene orden p entonces*

$$R(z) = \exp(z) + \mathcal{O}(z^{p+1}) \quad (z \longrightarrow 0).$$

Demostración: Consideramos el problema lineal test (5.1) con $y_0 = 1$. Por tanto, $y(h) = \exp(z)$. Aplicamos un paso del RK de longitud de paso h. Por (5.4), $y_1 = R(z) y_0 = R(z)$. Como es de orden p, $y(h) - y_1 = \mathcal{O}(h^{p+1}) \Longrightarrow \exp(z) - R(z) = \mathcal{O}(z^{p+1}) \quad z \to 0.$ □

Teorema 5.2.2 *[6, Cap.IV.3] Dado un RK(A, b) de s etapas, se tiene*

$$R(z) = \frac{det(I - z(A - e b^T))}{det(I - zA)}$$

y por tanto $R(z) \in \prod_{s/s} := \{p(z)/q(z) : p(z), q(z)$ *polinomios de grado* $\leq s\}$.

Demostración: Si en (5.2) llamamos $K_i = f(t_n + c_i h, Y_i)$, el RK se puede expresar entonces como

$$\begin{cases} Y_i = y_n + h\sum_{j=1}^{s} a_{ij} f(t_n + c_j h, Y_j), \quad 1 \le i \le s. \\ y_{n+1} = y_n + h\sum_{j=1}^{s} b_j f(t_n + c_j h, Y_j). \end{cases}$$

Lo aplicamos al problema (5.1) con $y_0 = 1$ y un paso h, $z = \lambda h$, obteniendo

$$Y_i = 1 + z\sum_{j=1}^{s} a_{ij} Y_j \quad 1 \le i \le s, \qquad y_1 = 1 + z\sum_{j=1}^{s} b_j Y_j, \tag{5.5}$$

Igual que hicimos con K_i en (5.3), matricialmente tenemos $(I - zA)Y = e$, donde $Y = (Y_1, Y_2, ..., Y_s)^T \in I\!R^s$. Así podemos expresar este sistema (5.5) como

$$\underbrace{\begin{pmatrix} I - zA & 0 \\ -zb^T & 1 \end{pmatrix}}_{M} \begin{pmatrix} Y \\ y_1 \end{pmatrix} = \begin{pmatrix} e \\ 1 \end{pmatrix}$$

Para calcular y_1, aplicamos la regla Cramer

$$y_1 = \frac{det\begin{pmatrix} I - zA & e \\ -zb^T & 1 \end{pmatrix}}{det(M)}$$

Calculando del determinante de M por bloques obtenemos $detM = det(I - zA)$. Por otra parte, para el numerador, aplicamos operaciones elementales sobre la última fila para obtener ceros en la última columna.

$$det\begin{pmatrix} I - zA & e \\ -zb^T & 1 \end{pmatrix} = det\begin{pmatrix} I - zA + zeb^T & 0 \\ -zb^T & 1 \end{pmatrix} = det(I - zA + zeb^T)$$

Como $y_1 = R(z)$ por (5.4) tiene el cociente dado por el Teorema. Como ambos determinantes son de dimensión s, al desarrollarlos obtendremos potencias de z hasta orden z^s y por tanto $R(z) \in \prod_{s/s}$. □

El objetivo de este Teorema 5.4 es obtener el siguiente corolario:

Corolario 5.2.2.1 *Si un RK(A, b) de s etapas es explícito entonces $R(z)$ es un polinomio de grado $\le s$. Por tanto, ningún RK explícito puede ser A-estable.*

Demostración: Sabemos que la matriz A de los métodos RK explícitos es triangular inferior estricta, así $det(I - zA) = 1$. Por el Teorema 5.2.2, necesariamente $R(z)$ es un polinomio de grado $\le s$. Precisamente por eso cuando $z \longrightarrow \infty$ no estará acotado, esto es,

$$|R(z)| > 1, \quad |z| > C, \quad z \in \mathbb{C}^-$$

para una cierta C. Por tanto, no puede ser A-estable. □

Capítulo 6

Experimento numérico con la ecuación del calor

Una de las ecuaciones diferenciales parciales clásica es la ecuación que describe la conducción del calor en un cuerpo sólido [2]. La primera investigación importante sobre la conducción del calor fue llevada a cabo por Joseph B. Fourier $(1768 - 1830)$.
Esta ecuación modeliza la conducción de calor a lo largo una barra recta de sección uniforme y material homogéneo. Elijamos el eje x a lo largo del eje de la barra y denotemos por $x = 0$ y $x = l$ los extremos de la barra. Supongamos además que los lados de la barra, excepto posiblemente los extremos $x = 0, l$, están perfectamente aislados de modo que no pase calor a través de ellos y que las dimensiones de la sección perpendicular son tan pequeñas que la temperatura u puede considerarse constante sobre cualquier sección recta dada. Entonces, u es una función sólo de la coordenada axial x y el tiempo t.
La variación de temperatura en la barra se expresa por una ecuación diferencial parcial, denominada **ecuación del calor** y tiene la forma

$$u_t(x,t) = \alpha^2 u_{xx}(x,t), \quad x \in [0,l], \quad t \geq 0. \tag{6.1}$$

donde α^2 es una constante que se conoce como disipación térmica. El parámetro α^2 depende únicamente del material de la barra y se define por $\alpha^2 = k/\rho S$ donde k es la conductividad térmica, ρ es la densidad y S es el calor específico del material de la barra. Además, supongamos que se da la distribución inicial de la temperatura en la barra, es decir, conocemos lo que se llama como condición inicial: $u(x,0) = u_0(x)$, $x \in [0,l]$ donde u_0 es una función dada.
Por último, supongamos que los extremos de la barra se mantienen a temperatura fija igual a 0^o, obteniendo así las condiciones de frontera de Dirichlet homogéneas: $u(0,t) = 0$, $u(l,t) = 0$, $t \geq 0$.
Por tanto, tenemos el siguiente problema:

$$\begin{cases} u_t = \alpha^2 u_{xx}, \quad x \in [0,l], \quad t \geq 0. \\ u(x,0) = u_0(x), \quad u(0,t) = 0, \quad u(l,t) = 0. \end{cases} \tag{6.2}$$

El problema fundamental de la conducción del calor es encontrar $u(x,t)$ que satisfaga la ecuación diferencial (6.1), la condición inicial y las condiciones de frontera.
Para aproximar numéricamente la solución $u(x,t)$ vamos a aplicar el **método de líneas** [8, 9], que permite resolver la ecuación parabólica en derivadas parciales (6.2) como un PVI de ecuaciones diferenciales ordinarias. Veamos en qué consiste este método.
Definimos en $[0,l]$ los nodos $x_j = j\Delta x$ con $\Delta x = l/(m+1)$. Para cada x_j vamos a denotar

$$u_j(t) = u(x_j,t), \qquad t \geq 0.$$

Nótese que los valores en los extremos de la barra vienen dados por las condiciones de frontera, $u_0(t) = u_{m+1}(t) = 0$, pero no son conocidos los que corresponden a los otros índices. Se aproximan las derivadas segundas $u_{xx}(x_j,t)$ mediante diferencias centrales de segundo orden

$$u_{xx}(x_j,t) = \frac{u(x_{j-1},t) - 2u(x_j,t) + u(x_{j+1},t)}{(\Delta x)^2} + \mathcal{O}((\Delta x)^2) \qquad \Delta x \geq 0.$$

Por otra parte, $u_t(x_j,t) = u_j'(t),\ j = 1,...,m$. Si denotamos por $v_j(t)$ la aproximación buscada de $u(x_j,t)$, nos queda

$$v_j'(t) = \frac{\alpha^2}{(\Delta x)^2}\left(v_{j-1}(t) - 2v_j(t) + v_j(t)\right) \qquad 1 \leq j \leq m. \tag{6.3}$$

Definiendo los vectores
$v(t) = (v_1(t)\ ,...,\ v_m(t))^T$, $v'(t) = (v_1'(t)\ ,...,\ v_m'(t))^T$, $v^0(t) = (u_0(x_1)\ ...\ u_0(x_m))^T$
podemos expresar el sistema de ODEs (6.3) con las condiciones de frontera y los valores iniciales como el PVI de dimensión m lineal de coeficientes constantes:

$$v'(t) = \mathcal{A}v(t), \quad v(0) = v^0, \quad t \geq 0 \tag{6.4}$$

donde la matriz $\mathcal{A}$ del problema es la matriz tridiagonal

$$\mathcal{A} = \frac{\alpha^2}{(\Delta x)^2}\begin{pmatrix} -2 & 1 & 0 & 0 & \cdots & 0 & 0 & 0 \\ 1 & -2 & 1 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 1 & -2 & 1 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1 & -2 & 1 \\ 0 & 0 & 0 & 0 & \cdots & 0 & 1 & -2 \end{pmatrix}$$

Por tanto, para obtener una aproximación numérica a la solución de la ecuación del calor (6.1) a los puntos $\{x_j\}_{j=1}^m$ basta con aplicar un método RK al PVI (6.4) resultante.
Para seleccionar el RK a utilizar es importante estudiar el carácter más o menos stiff que presenta este sistema lineal, ya que esto nos condicionará a usar unos métodos numéricos u otros. Este carácter depende de los autovalores de la matriz $\mathcal{A}$, que vienen dados por

$$\lambda_j = \frac{-4\alpha^2}{(\Delta x)^2}\sin^2\left(\frac{j\pi}{2(m+1)}\right), \quad j = 1,...,m$$

por ser una matriz Toeplitz [10]. Obsérvese que

$$\frac{-4\alpha^2}{(\Delta x)^2} \backsimeq \lambda_m \leq \lambda_j \leq \lambda_1 \backsimeq 0, \quad 1 \leq j \leq m$$

y que todos son reales y negativos. Cuanto más pequeño es el tamaño de paso Δx más negativos serán los autovalores, por tanto el PVI (6.4) será más stiff. Como vamos a aplicar métodos RK explícitos para su resolución, por el Corolario 5.2.2.1 sabemos que el dominio de estabilidad $\mathcal{S}$ de dichos métodos va a ser una región acotada de $\mathbb{C}^-$. En la Figura 6.1 (a) dibujamos el borde de dichos dominios para los métodos de Euler (2.3), Runge de orden 3 (2.19) y Kutta de orden 4 (2.20), y en el Cuadro 6.1 representamos sus fronteras de estabilidad real β_R.
Para que estos tres métodos vayan bien con un tamaño de paso temporal h fijo, por la teoría de estabilidad lineal vista en la sección anterior, cuando m es grande necesariamente

$$z = -\frac{4\alpha^2}{(\Delta x)^2} h \in [-\beta_R, 0]$$

o lo que es lo mismo

$$0 \leq h \leq \frac{(\Delta x)^2}{4\alpha^2}\beta_R \tag{6.5}$$

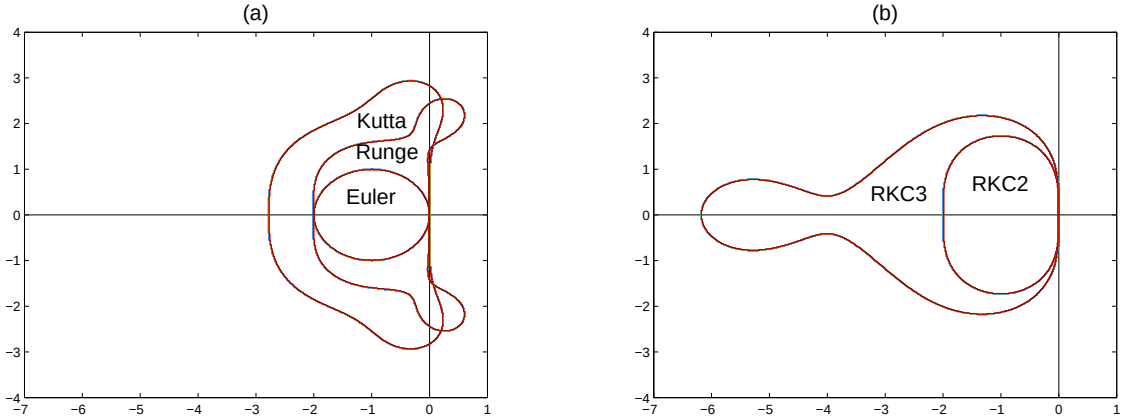
lo que supone un gran restricción de paso para estos métodos. Este comportamiento numérico se refleja claramente en las primeras columnas del Cuadro 6.2. Para elaborar dicho cuadro hemos computado diferentes particiones de m nodos espaciales con los citados métodos con diferentes tamaños de paso temporales h fijos. Lo que aparece en cada entrada del cuadro para cada (m, h) y método fijo es el error obtenido cuando $t = 1$. Para calcular este error, se ha computado una solución de referencia $\{u_R(x_j, 1)\}_{j=1}^m$ con la rutina **ode23t** de Matlab [7] con mucha precisión y luego hemos computado

$$error = \max_{1 \leq j \leq m} |u_{RK}(x_j, 1) - u_R(x_j, 1)|$$

Así podemos observar que la restricción (6.5) es determinante, lo que hace que para los valores de $m \geq 160$ ninguno de los tres sea capaz de dar una solución con tamaño de paso h relativamente razonables.
Una alternativa para mejorar esto es diseñar métodos explícitos que tengan una función de estabilidad con β_R mayor. Esta es la idea que está detrás de los llamados **métodos RK explícitos estabilizados** [6], y en particular, una clase muy eficiente de estos llamados **Runge-Kutta-Chebychev (RKC)** [6, 8, 11]. Un RKC es un método s etapas de $s \geq 2$ y orden 2 que tiene una región de estabilidad formada por una franja estrecha que puede extenderse a lo largo del eje real negativo aumentando s. La frontera de estabilidad de dicho métodos es $\beta_R \approx 0,653s^2$. Su fórmula general de $s \geq 2$ etapas sobre $PVIs$ (4.1) es

$$\begin{cases} Y_0 = y_n, \ Y_1 = Y_0 + \bar{\mu}_1 h F_0 \\ Y_j = (1 - \mu_j - v_j)Y_0 + \mu_j Y_{j-1} v_j Y_{j-2} + \bar{\mu}_j h F_{j-1} + \gamma_j h F_0, \ j = 2, ..., s. \\ y_{n+1} = Y_s \end{cases} \tag{6.6}$$

Figura 6.1: Dominio de A-estabilidad



Cuadro 6.1: Frontera de estabilidad real de los métodos

	Euler	Runge	Kutta	RKC2	RKC3
β_R	2	2	2.7852	2	6.1802

donde los coeficientes tiene las siguientes expresiones:

$$\varepsilon = 2/13,\ w_0 = 1+\varepsilon/s^2,\ w_1 = \frac{T_s'(w_0)}{T_s''(w_0)},\ b_j = \frac{T_j''(w_0)}{(T_j'(w_0))^2}\ 2 \le j \le s,\ b_0 = b_2,\ b_1 = b_2$$

$$\bar{\mu}_1 = b_1 w_1,\ \mu_1 = \frac{2b_j w_0}{b_{j-2}},\ v_j = \frac{-b_j}{b_{j-2}}, \bar{\mu}_j = \frac{2b_j w_1}{b_{j-1}},\ \gamma_j = -\left(1 - b_{j-1}T_{j-1}(w_0)\right)\bar{\mu}_j\ 2 \le j \le s.$$

donde los T_j son los polinomios de Chebychev de primera especie de grado j esto es, $T_j(x) = \cos(j \arc\cos x)$.
Estos métodos pertenecen a la clase de métodos RKE (2.10). En particular, para $s = 2$ (RKC2) y $s = 3$ (RKC3) se tienen las siguientes tablas de Butcher:

0	0	0
$\frac{13}{54}$	$\frac{13}{54}$	0
	$\frac{-14}{13}$	$\frac{27}{13}$

0	0	0	0
$\frac{5025735}{53925088}$	$\frac{5025735}{53925088}$	0	0
$\frac{42955}{113288}$	$\frac{-5197555}{13254696}$	$\frac{42955}{55692}$	0
	$\frac{-70817}{42471}$	$\frac{26962544}{15077205}$	$\frac{113288}{128865}$

Cuadro 6.2: Errores de los métodos

m	h	Euler	Runge	Kutta	RKC2	RKC3
40	1/80	1.167e+47	1.816e+131	5.722e+94	8.124e+74	1.163e-02
	1/160	1.411e+30	1.398e+66	2.724e-03	5.896e+37	4.225e-03
	1/320	4.990e-03	6.311e-04	9.967e-05	1.582e-03	8.187e-04
	1/640	1.889e-03	4.213e-05	3.826e-06	2.711e-04	1.536e-04
	1/1280	9.075e-04	3.820e-06	1.931e-07	5.521e-05	3.288e-05
80	1/320	1.441e+193	–	–	3.070e+301	5.687e-03
	1/640	7.552e+119	4.867e+250	1.228e-03	2.675e+146	2.031e-03
	1/1280	2.420e-03	2.886e-04	4.481e-05	7.481e-04	3.890e-04
	1/2560	9.312e-04	1.954e-05	1.742e-06	1.292e-04	7.343e-05
160	1/1280	–	8.859e+303	–	–	2.810e-03
	1/2560	–	–	5.829e-04	–	9.954e-04
320	1/5121	–	1.338e+303	2.409e+303	–	1.396e-03
	1/10241	–	–	2.839e-04	–	4.927e-04

Sin embargo, se suele preferir la notación (6.6) pues fue la que usaron sus diseñadores para mejorar otras características de estos métodos, como la llamada *estabilidad interna* (ver [8, Cap.IV.1]).
En la Figura 6.1 (b) y en las dos últimas columnas del Cuadro 6.1 se presentan los dominios de estabilidad $\mathcal{S}$ y las fronteras β_R del RKC2 y RKC3 respectivamente. En el Cuadro 6.2 vemos los resultados obtenidos con estos dos métodos sobre la ecuación del calor (6.4).
Evidentemente RKC2 no mejora respecto de los otros pues $\beta_R = 2$, pero la ventaja de usar RKC3 es clara pues su frontera β_R es bastante mayor, integrando sin ninguna dificultad este problema para los valores de (m, h) considerados.
Hay que observar que, a pesar de que RKC3 es un método sólo de orden 2, se comporta mucho mejor que el Runge de orden 3 y el de Kutta de orden 4. Por tanto, la mejora de la estabilidad de los métodos RK es fundamental para que puedan aplicarse de forma eficiente, independientemente del orden que tengan.
Por otra parte, también se puede ver en el Cuadro 6.2 que una vez que la restricción de estabilidad (6.5) se ha superado, el orden determina la velocidad de convergencia y la precisión de los métodos.

Bibliografía

[1] K.E. Atkinson, John Wiley & Sons, *An introduction to numerical analysis, (2^a edición).*

[2] W.E. Boyce, Richard C. Diprima, *Ecuaciones diferenciales y problemas con valores en la frontera, 3^a edición*, Instituto Politecnico Rensseler.

[3] M. Calvo, J.I. Montijano, L. Rández, *Métodos de Runge-Kutta para la resolución numérica de ecuaciones diferenciales ordinarias*, Secretariado de Publicaciones, Universidad de Zaragoza, 1998.

[4] P. Deuflhard, F. Bornemann, *Scientific computing with ordinary differential equations*, Texts in Aplied Mathematics, Spinger Verlag New York, Inc. 2002.

[5] E. Hairer, S.P. Nørsett, G. Wanner, *Solving ordinary differential equations I (nonstiff problems)*, Springer Verlag, 1993.

[6] E. Hairer, G. Wanner, *Solving ordinary differential equations II (stiff and differencial algebraic problems)*, Springer Verlag, 1996.

[7] D.J. Higham, N,J. Higham, *Matlab guide*, Second Edition, SIAM 2005.

[8] W.Hundsdorfer, J.G Verwer, *Numerical solution of time-dependent advection-diffusion reaction equations*, Springer Verlag, 2003.

[9] J.D. Lambert, J. Wiley, *Numerical methods for ordinary differential systems, (the initial value problem)*, 1991.

[10] C.D.Mayer, *Matrix analysis and applied linear algebra*, SIAM, 2000.

[11] B.P Sommeijer, L.F. Shampine, J.G.Verwer, *RKC: An explicit solver for parabolic PDEs*, Journal of Computational and Applied Mathematics, 88, pp. $315 - 326$, 1997.

[12] V. Szebetiely, *Theory of orbits. The restricted problem of three bodies*, Academic Press, New York, 1967.