Universidad de Valladolid

Facultad de Ciencias

TRABAJO FIN DE GRADO

Grado en Matemáticas

Análisis regresivo de integradores numéricos para ecuaciones diferenciales ordinarias

Autora: Marta Alonso Tubía

Tutora: **María Paz Calvo Cabrero**

Agradecimientos

Me gustaría expresar mi agradecimiento a Mari Paz Calvo Cabrero por su dedicación y su compromiso para dirigir este TFG en las mejores condiciones y con el máximo aprendizaje, por resolver todas mis dudas con paciencia y por darme las pautas para realizar un trabajo riguroso. El resultado de esta memoria se debe en gran medida a ella.

Dar las gracias también a mis padres, por apoyarme siempre en todo lo que hago, y por hacer todos los esfuerzos necesarios para darme una educación.

Índice general

Introducción

Un análisis clásico de los errores, el llamado *análisis progresivo de los errores*, proporciona cotas superiores para la diferencia entre la solución exacta y la solución aproximada del problema en estudio. Este tipo de análisis puede no resultar apropiado cuando al cambiar ligeramente los datos en dicho problema, se pueden obtener soluciones exactas muy distintas. Es probable que en este caso, los métodos numéricos proporcionen soluciones muy diferentes de la solución exacta buscada, y la conclusión de este análisis sería que todos los métodos fracasan. El *análisis regresivo de los errores* surge como una alternativa a las acotaciones clásicas y trata de interpretar la solución numérica de un problema dado como la solución exacta de un problema del mismo tipo, pero con los datos ligeramente modificados. Este tipo de análisis fue introducido por J.H. Wilkinson en los años 70 del siglo pasado en el contexto del Álgebra Lineal Numérica y ha sido muy utilizado desde entonces.

En la resolución de ecuaciones diferenciales ordinarias, es frecuente el uso de métodos numéricos para obtener soluciones aproximadas cuando no es posible hallar soluciones en forma cerrada. En este contexto, una primera alternativa al análisis progresivo de los errores, es el *shadowing*, técnica que compara la solución numérica con condición inicial y_0, con la solución exacta del mismo sistema diferencial, pero con condición inicial ligeramente perturbada $\tilde{y}_0$. Por su parte el análisis regresivo interpreta la solución numérica de un problema de valores iniciales como la solución exacta de otro problema en el que se mantiene la misma condición inicial pero se modifica el lado derecho del sistema diferencial (*ecuaciones modificadas*). El análisis de los errores del método numérico se puede llevar a cabo entonces estudiando la diferencia entre las soluciones exactas de dos sistemas diferenciales próximos que comparten la misma condición inicial.

En este trabajo nos centraremos en esta segunda idea, y mostraremos que el análisis regresivo de los errores puede resultar muy útil, especialmente cuando se está interesado en el comportamiento cualitativo de las soluciones numéricas de ciertos sistemas diferenciales, y cuando se requieren integraciones en intervalos temporales largos.
En el Capítulo 1 se muestran algunos ejemplos para los que se construyen los correspondientes sistemas diferenciales modificados, y se estudia un procedimiento sistemático para la construcción de las ecuaciones modificadas asociadas a un método numérico utilizando las llamadas B-series. En el segundo capítulo se particulariza este análisis al caso de sistemas Hamiltonianos, y se demuestra que usando métodos simplécticos para su integración numérica, las ecuaciones modificadas son también Hamiltonianas. En el Capítulo 3 se incluyen experimentos numéricos que muestran que en integraciones temporales largas los errores ge-

nerados por métodos simplécticos presentan un crecimiento más lento que los generados por métodos convencionales. También se justifican teóricamente los resultados obtenidos haciendo uso del análisis regresivo estudiado en el capítulo anterior. Finalmente, en el Apéndice se incluyen los códigos de los programas que se han implementado para generar las gráficas de los Capítulos 1 y 3.

Capítulo 1

Ecuaciones modificadas para métodos de un paso.

En este primer capítulo se introduce el concepto de ecuación modificada para un integrador temporal y se ilustra a través de varios ejemplos su construcción por medio de los desarrollos de Taylor de la solución numérica y de la solución exacta del sistema diferencial modificado. A través de ellos se motiva la necesidad de diseñar un procedimiento que permita construir de manera sistemática las ecuaciones modificadas. La herramienta para realizar dicha construcción son las llamadas B-series. En la Sección 1.2 se describe brevemente la teoría de árboles con raíz y de B-series necesaria para llegar al resultado fundamental del capítulo: la construcción de las ecuaciones modificadas.

1.1. Planteamiento del problema y ejemplos

Comenzaremos resolviendo numéricamente la ecuación test escalar

$$\dot{y}(t) = \lambda y(t), \qquad y(0) = y_0,$$

con el método de Euler explícito y paso fijo h, cuya definición es

$$y_{n+1} = y_n + h\lambda y_n. \tag{1.1.1}$$

Por inducción se comprueba que $y_n = (1 + h\lambda)^n y_0$. Esto puede reescribirse en la forma

$$y_n = (1 + h\lambda)^n y_0 = e^{nh(ln(1+h\lambda)/h)} y_0,$$

a partir de lo cual es fácil ver que si denotamos por $t_n = nh$, entonces la solución propuesta por el método numérico en tiempo t_n es la solución exacta de la ecuación diferencial

$$\dot{\tilde{y}}(t) = \Lambda_h \tilde{y}(t), \qquad \tilde{y}(0) = y_0, \tag{1.1.2}$$

donde

$$\Lambda_h = \frac{ln(1+h\lambda)}{h} = \lambda\left(1 - \frac{1}{2}\lambda h + \frac{1}{3}\lambda^2 h^2 - \frac{1}{4}\lambda^3 h^3 + O(h^4)\right). \tag{1.1.3}$$

La ecuación (1.1.2)-(1.1.3) se denomina *ecuación modificada* de la ecuación test escalar en el caso particular del método de Euler explícito, y hemos visto que la solución numérica en $t_n = nh$ generada con el método de Euler explícito coincide con la solución exacta de la ecuación modificada (1.1.2)-(1.1.3). Nos planteamos ahora la construcción de dichas ecuaciones modificadas para métodos numéricos de un paso más generales, implementados con paso fijo h.
Consideremos un sistema de ecuaciones diferenciales ordinarias

$$\dot{y}(t) = f(y(t)), \tag{1.1.4}$$

y un método numérico de un paso $\Psi_{h,f}$, que partiendo de $y(t_0) = y_0$ produce aproximaciones $y_1, y_2, ...$ a la solución del problema (1.1.4) en tiempos equiespaciados $t_n = t_0 + nh, n \geq 1$. Denotaremos por $\Phi_{h,f}(y)$ a la solución exacta de (1.1.4) tras h unidades de tiempo, con condición inicial y.
Diremos que el método $\Psi_{h,f}$ tiene orden r si

$$\Phi_{h,f}(y) - \Psi_{h,f}(y) = O(h^{r+1}), \quad h \to 0,$$

para toda condición inicial y y toda función suficientemente regular f. El objetivo ahora es encontrar una ecuación diferencial modificada $\dot{\tilde{y}} = f_h(\tilde{y})$ de la forma

$$\dot{\tilde{y}} = f(\tilde{y}) + hf_2(\tilde{y}) + h^2 f_3(\tilde{y}) + \cdots \tag{1.1.5}$$

con la misma condición inicial $\tilde{y}(t_0) = y_0$ tal que $y_n = \tilde{y}(t_0 + nh), n \geq 1$, es decir, las aproximaciones numéricas de la solución de (1.1.4) en los tiempos $t_0 + nh$ son soluciones exactas para la ecuación modificada (1.1.5).
Para el cálculo de (1.1.5), desarrollamos en serie de Taylor la solución exacta de (1.1.5) en $t + h$

$$\begin{aligned} \tilde{y}(t+h) &= y + h(f(y) + hf_2(y) + h^2 f_3(y) + \cdots)) \\ &\quad + \frac{h^2}{2!}(f'(y) + hf_2'(y) + \cdots)(f(y) + hf_2(y) + \cdots) + \cdots \end{aligned} \tag{1.1.6}$$

y comparamos con el resultado obtenido con el método numérico $\Psi_{h,f}(y)$ tras un paso de longitud h partiendo de y. Obtenemos

$$\Psi_{h,f}(y) = y + hf(y) + h^2 d_2(y) + h^3 d_3(y) + \cdots. \tag{1.1.7}$$

En ambos casos $'$ denota derivación respecto de y. En (1.1.7) se ha supuesto que el método es consistente y que, por tanto, el coeficiente que acompaña a h en $\Psi_{h,f}(y)$

es $f(y)$. Se trata de comparar potencias de h en ambas expresiones (1.1.6)-(1.1.7), obteniendo las siguientes relaciones para las funciones f_j, $2 \leq j \leq 4$,

$$\begin{aligned} f_2(y) &= d_2(y) - \frac{1}{2!}f'f(y), \\ f_3(y) &= d_3(y) - \frac{1}{3!}(f''(f,f)(y) + f'f'f(y)) - \frac{1}{2!}(f'f_2(y) + f_2'f(y)), \\ f_4(y) &= d_4(y) - \frac{1}{4!}(f'''(f,f,f)(y) + 4f''(f'f,f)(y) + f'f'f'f(y)) \\ &- \frac{1}{3!}(f''(f_2,f)(y) + f'f'f_2(y) + f_2''(f,f)(y) + f_2'f'f(y) + f''(f,f_2)(y) \\ &+ f'f_2'f(y)) - \frac{1}{2}(f_3'f(y) + f_2'f_2(y) + f'f_3(y)). \end{aligned} \tag{1.1.8}$$

Ejemplo 1.1.1. *Consideremos en primer lugar el problema de valores iniciales*

$$\dot{y}(t) = [y(t)]^2, \qquad y(0) = 1, \tag{1.1.9}$$

y lo resolvemos inicialmente con el método de Euler explícito para el cual

$$\Psi_{h,f}(y) = y + hf(y). \tag{1.1.10}$$

En este caso se tiene $d_2 = d_3 = \cdots = 0$ en (1.1.7). Usando las ecuaciones obtenidas (1.1.8) y que $f(y) = y^2$, resulta

$$f_2(y) = -y^3, \qquad f_3(y) = \frac{3}{2}y^4, \qquad f_4(y) = -\frac{8}{3}y^5,$$

con lo que el lado derecho de la ecuación modificada (1.1.6) es

$$f_h(y) = y^2 - hy^3 + h^2\frac{3}{2}y^4 - h^3\frac{8}{3}y^5 + O(h^4).$$

En la Figura 1.1 se presenta la solución exacta de la ecuación original (1.1.9) en color azul y la solución exacta de la ecuación modificada truncada tras los términos en h (color amarillo) y en h^2 (color verde), respectivamente. Se puede observar que la diferencia entre la solución numérica (representada con círculos rojos) y la solución de la ecuación modificada es bastante menor que la diferencia entre la solución numérica y la solución exacta de (1.1.9). Además esta diferencia disminuye a medida que incluimos más términos en la ecuación modificada.

Si resolvemos (1.1.9) con la regla implícita del punto medio

$$y_{n+1} = y_n + hf(\frac{y_n + y_{n+1}}{2}),$$

el desarrollo de Taylor de la solución numérica hasta los términos en h^4 es

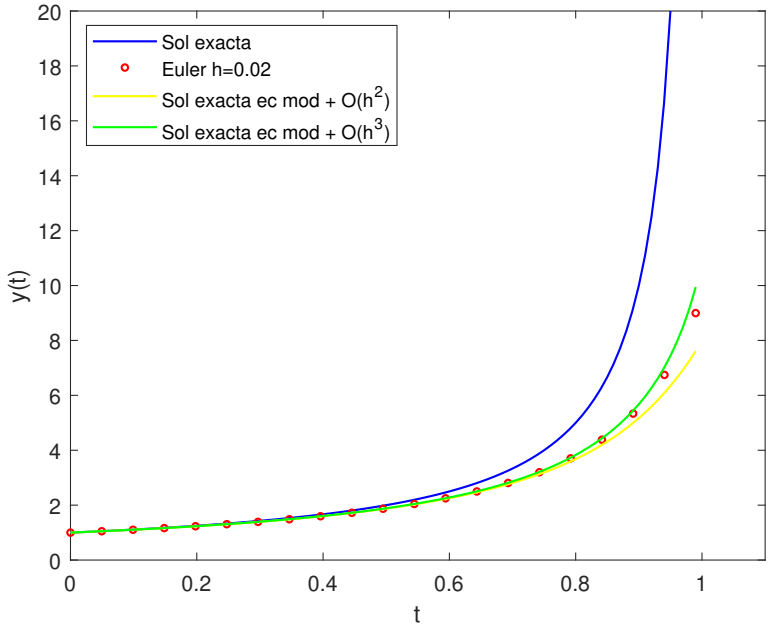


Figura 1.1: Evolución con el tiempo t de la solución exacta de (1.1.9), solución numérica y solución exacta de las ecuaciones modificadas

$$y_{n+1} = y_n + hy_n^2 + h^2y_n^3 + \frac{5}{4}h^3y_n^4 + \frac{7}{4}h^4y_n^5 + O(h^5),$$

es decir, $d_2(y) = y^3$, $d_3(y) = \frac{5}{4}y^4$, $d_4(y) = \frac{7}{4}y^5$, *y utilizando (1.1.8) se llega a*

$$f_2(y) = 0, \qquad f_3(y) = \frac{1}{4}y^4, \qquad f_4(y) = 0.$$

La ecuación modificada queda entonces

$$\dot{\tilde{y}} = \tilde{y}^2 + \frac{1}{4}h^2\tilde{y}^4 + \frac{1}{8}h^4\tilde{y}^6 + O(h^6).$$

Observamos que el lado derecho de la ecuación modificada empieza incorporando términos de tamaño $O(h^2)$ *al lado derecho de la ecuación original (1.1.9), y que solo aparecen potencias pares de* h. *Esto es debido a que el método numérico utilizado es de orden 2 y simétrico [8].*
Por último, para resolver (1.1.9) vamos a usar el método de Runge de orden 4 definido por

$$y_{n+1} = y_n + \frac{h}{6}(k_1 + 2k_2 + 2k_3 + k_4),$$

donde

$$k_1 = y_n^2, \quad k_2 = (y_n + \frac{h}{2}k_1)^2, \quad k_3 = (y_n + \frac{h}{2}k_2)^2, \quad k_4 = (y_n + hk_3)^2.$$

Se procede análogamente, haciendo los desarrollos de Taylor de la solución numérica

$$\begin{aligned} y_{n+1} &= y_n + hf(y_n) + \frac{h^2}{2}f'f(y_n) + \frac{h^3}{6}(f''(f,f)(y_n) + f'f'f(y_n)) \\ &+ \frac{h^4}{24}(3f''(f,f'f)(y_n) + f'f''(f,f)(y_n) + f'f'f'f(y_n)) \\ &= y_n + hy_n^2 + h^2y_n^3 + h^3y_n^4 + h^4y_n^5 + O(h^5). \end{aligned} \tag{1.1.11}$$

Se puede leer directamente

$$d_2(y) = y^3, \qquad d_3(y) = y^4, \qquad d_4(y) = y^5,$$

y a partir de las ecuaciones (1.1.8), se calcula

$$f_2(y) = f_3(y) = f_4(y) = 0.$$

La ecuación modificada es entonces

$$\dot{\tilde{y}} = \tilde{y}^2 - h^4\frac{1}{24}\tilde{y}^6 + h^6\frac{65}{576}\tilde{y}^8 + O(h^7).$$

Observamos en este ejemplo que el lado derecho de la ecuación modificada es una perturbación del lado derecho de la ecuación original de tamaño $O(h^r)$, siendo $r = 4$ el orden del método.

Ejemplo 1.1.2. *Consideremos ahora las ecuaciones de Lotka-Volterra*

$$\dot{p}(t) = p(t)[2 - q(t)], \qquad \dot{q}(t) = q(t)[p(t) - 1]. \tag{1.1.12}$$

Aplicamos en primer lugar el método de Euler explícito.
En este caso,

$$f'(p,q) = \begin{pmatrix} 2-q & -p \\ q & p-1 \end{pmatrix},$$

y como

$$d_i = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \qquad i \geq 2,$$

se tiene

$$f_2 = \begin{pmatrix} 0 \\ 0 \end{pmatrix} - \frac{1}{2}\begin{pmatrix} 2-q & -p \\ q & p-1 \end{pmatrix}\begin{pmatrix} p(2-q) \\ q(p-1) \end{pmatrix}.$$

La ecuación modificada consistente con el método hasta los términos $O(h)$ tiene lado derecho

$$f_h(p,q) = \begin{pmatrix} p(2-q) \\ q(p-1) \end{pmatrix} - \frac{h}{2} \begin{pmatrix} q(p^2-pq+1) \\ p(q^2-3q-pq+4) \end{pmatrix} + O(h^2).$$

Como segundo integrador para (1.1.12) consideramos el método de Euler simpléctico, que aplicado a sistemas diferenciales de la forma

$$\dot{p} = f(p,q), \qquad\qquad \dot{q} = g(p,q),$$

produce aproximaciones

$$\begin{aligned} p_{n+1} &= p_n + hf(p_{n+1}, q_n), \\ q_{n+1} &= q_n + hg(p_{n+1}, q_n). \end{aligned}$$

En el caso particular de (1.1.12), $f(p,q) = p(2-q)$, $g(p,q) = q(p-1)$, y los primeros términos del desarrollo en serie de Taylor de la solución numérica son

$$\begin{pmatrix} p_{n+1} \\ q_{n+1} \end{pmatrix} = \begin{pmatrix} p_n \\ q_n \end{pmatrix} + h \begin{pmatrix} p_n(2-q_n) \\ q_n(p_n-1) \end{pmatrix} + h^2 \begin{pmatrix} p_n(2-q_n)^2 \\ q_n p_n(2-q_n) \end{pmatrix} + O(h^3).$$

Utilizando de nuevo (1.1.8) se tiene

$$f_2 = \begin{pmatrix} p(2-q)^2 \\ qp(2-q) \end{pmatrix} - \frac{1}{2} \begin{pmatrix} p(q^2-3q-pq+4) \\ q(p^2-pq+1) \end{pmatrix} = \begin{pmatrix} \frac{1}{2}pq^2 - \frac{5}{2}pq + \frac{1}{2}p^2q + 2p \\ -\frac{1}{2}q - \frac{1}{2}pq^2 - \frac{1}{2}qp^2 - 2qp \end{pmatrix},$$

es decir, el lado derecho de la ecuación modificada es

$$f_h = \begin{pmatrix} p(2-q) \\ q(p-1) \end{pmatrix} + \frac{h}{2} \begin{pmatrix} +p(q^2-5q+pq+4) \\ -q(1+pq+p^2-4p) \end{pmatrix} + O(h^2).$$

En la Figura 1.2 se comparan las soluciones numéricas aportadas por los métodos de Euler (izquierda) y Euler simpléctico (derecha) junto con el flujo exacto de la ecuación diferencial (1.1.12) en color azul y el flujo exacto de la ecuación modificada truncada tras los términos en h (color amarillo). La condición inicial considerada $(p,q) = (1,3)$ aparece representada en el plano de las fases por un punto negro de mayor tamaño. En la figura de la derecha se presentan también las soluciones correspondientes a la condición inicial $(p,q) = (1,6)$. De nuevo se han utilizado círculos rojos para representar la solución numérica.
Observamos que las soluciones de la ecuación modificada truncada son periódicas para el método de Euler simpléctico, al igual que lo es la solución exacta, puesto que en ambos casos existe una integral primera [7]. La periodicidad de la solución no se hereda, sin embargo, al aplicar el método de Euler. Vemos también que, como ya ocurría en el Ejemplo 1.1.1, la solución numérica está más próxima a la solución exacta de la ecuación modificada que a la solución exacta del problema original (1.1.12).

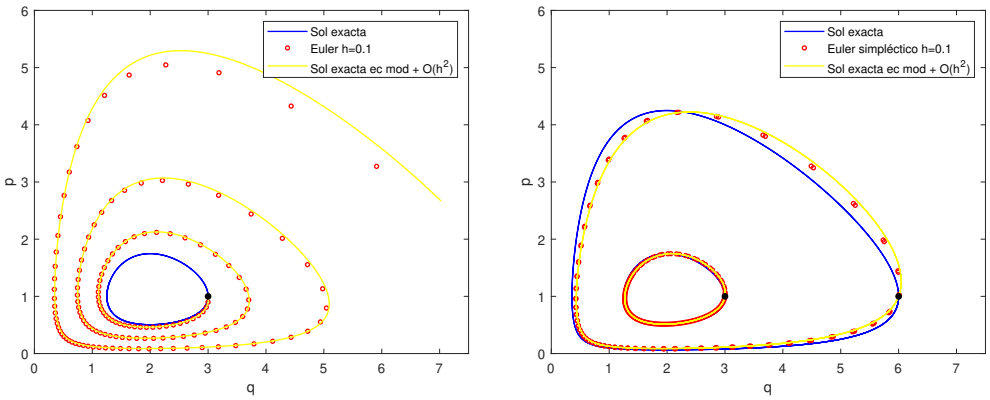


Figura 1.2: Solución exacta del problema de Lotka-Volterra (1.1.12), solución numérica y solución exacta de la ecuación modificada, representadas en el plano de fases.

Ejemplo 1.1.3. *Por último consideremos de nuevo la ecuación test escalar*

$$\dot{y}(t) = \lambda y(t), \qquad y(0) = y_0, \tag{1.1.13}$$

y tomemos ahora como método numérico un método Runge-Kutta de s etapas con tablero de Butcher

$$\begin{array}{c|c} c & A \\ \hline & b^T \end{array}$$

, es decir, con coeficientes

$$A = (a_{ij}) \in \mathbb{R}^{s\times s}, \qquad b = (b_1, ..., b_s)^T \in \mathbb{R}^s. \tag{1.1.14}$$

Los números b_i se llaman pesos del método y los c_i satisfacen

$$c_i = \sum_{j=1}^{s} a_{ij}, \qquad 1 \le i \le s.$$

Las relaciones para avanzar desde (t_n, y_n) un paso de longitud h con un método Runge-Kutta para resolver el problema general (1.1.4) son

$$k_i = f(t_n + c_i h, y_n + h\sum_{j=1}^{s} a_{ij}k_j), \qquad 1 \le i \le s, \tag{1.1.15}$$

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

Aplicando estas relaciones a la ecuación test escalar (1.1.13), se obtiene

$$y_{n+1} = R(h\lambda)y_n,$$

donde $R(z)$ es la llamada función de estabilidad del método Runge-Kutta

$$R(z) = 1 + zb^T(I - zA)^{-1}e, \qquad e = (1, ..., 1)^T \in \mathbb{R}^s.$$

Podemos escribir la solución numérica de (1.1.13) en tiempo $t = nh$ como

$$\tilde{y}(t) = R(h\lambda)^{t/h}y_0 = e^{t(\frac{ln(R(h\lambda)}{h})}y_0. \tag{1.1.16}$$

Es conocido que la solución de la ecuación diferencial (1.1.13) es de la forma $y(t) = e^{\lambda t}y_0$. Por tanto si la expresión del método es (1.1.16), se deduce fácilmente que la solución numérica es solución exacta de la ecuación diferencial

$$\dot{\tilde{y}}(t) = f_h(\tilde{y}(t)) \quad \text{con} \quad f_h(\tilde{y}(t)) = \frac{1}{h}ln(R(h\lambda))\tilde{y}(t).$$

Esta es, por tanto, la ecuación modificada de un método Runge-Kutta aplicado a (1.1.13), que generaliza a la ya encontrada para el método de Euler explícito al principio de esta sección.
Usando desarrollos de Taylor conocidos podemos escribir

$$R(h\lambda) = I + h\lambda b^T\left(\sum_{k=0}^{\infty}(h\lambda)^k A^k\right)e = I + h\lambda + \sum_{k=1}^{\infty}(h\lambda)^{k+1}b^T A^k e,$$

donde se ha usado que $(1-x)^{-1} = \sum_{k=0}^{\infty} x^k$ y que el método tiene orden 1 ($b^T e = 1$).
Por tanto, como $ln(1+x) = x - x^2/2 + x^3/3 + O(x^4)$,

$$\frac{1}{h}ln(R(h\lambda)) = \lambda + h\lambda^2(b^T Ae - \frac{1}{2}) + h^2\lambda^3(b^T A^2 e - b^T Ae + \frac{1}{3}) + \cdots .$$

La ecuación modificada resultante es

$$\dot{\tilde{y}} = (\lambda + hc_2\lambda^2 + h^2c_3\lambda^3 + \cdots)\tilde{y}, \tag{1.1.17}$$

con $c_2, c_3, \ldots$ definidos en términos de los coeficientes del método.

1.2. Árboles con raíz y B-series.

La escritura de las condiciones que tienen que satisfacer los coeficientes de un método Runge-Kutta para alcanzar un orden r puede volverse bastante complicada. A medida que aumenta el número de etapas, el proceso de comparar el

desarrollo en serie de potencias de la solución aproximada proporcionada por el método numérico y de la solución exacta de un problema general (1.1.4) se hace tedioso. Fue Butcher [2] quien desarrolló una teoría que permite una escritura natural de este proceso. Esta teoría se apoya en unos objetos llamados *árboles con raíz*, que se han estudiado en la asignatura optativa "Solución numérica de ecuaciones diferenciales", y que describimos aquí brevemente para utilizarlos en la construcción de las ecuaciones modificadas.

Definicion 1.2.1. *Un árbol con raíz es un grafo que tiene la propiedad de ser conexo, no contiene ciclos, y tiene un nodo o vértice destacado que es la raíz del árbol.*

Si y_1 es la solución numérica del problema (1.1.4) tras un paso de longitud h con un método Runge-Kutta con coeficientes (1.1.14), se puede escribir su desarrollo de Taylor como [2]

$$y_1 = y_0 + \sum_{\rho\tau \in \mathcal{RT}} \frac{h^{n(\rho\tau)}}{n(\rho\tau)!} \alpha(\rho\tau) \left(\gamma(\rho\tau) \sum_{i=1}^{s} b_i \Phi_i(\rho\tau) \right) F(\rho\tau)(y_0) \tag{1.2.1}$$

donde

$\mathcal{RT}$	es el conjunto de árboles con raíz,
$n(\rho\tau)$	es el número de vértices del arbol $\rho\tau$ (orden de $\rho\tau$),
$\alpha(\rho\tau)$	es el número de posibles etiquetados monótonos del árbol $\rho\tau$,
$\gamma(\rho\tau)$	es un coeficiente entero conocido como densidad del árbol $\rho\tau$,
$\Phi_i(\rho\tau)$	es una expresión polinómica dependiente de los coeficientes del método que se conoce como peso elemental de la etapa i-ésima,
$F(\rho\tau)(y)$	es la diferencial elemental asociada al árbol $\rho\tau$ evaluada en y, que depende de la función $f(y)$ y de sus derivadas.

En lo que sigue, el árbol de orden 1 se denota por $\rho\tau_1$, y denotaremos por $\rho\tau = [t_1, ..., t_m]$ el árbol que consta de una raíz y m ramas que parten de ella, a partir de las cuales se unen los árboles con raíz $t_1, ..., t_m$. Las funciones que intervienen en (1.2.1) se pueden definir recursivamente de la siguiente manera:

$$n(\rho\tau_1) = 1, \quad \alpha(\rho\tau_1) = 1, \quad \gamma(\rho\tau_1) = 1, \quad \Phi_i(\rho\tau_1) = 1, \quad F(\rho\tau_1)(y) = f(y),$$

y para $\rho\tau = [t_1, ..., t_m]$,

$$\begin{aligned}
n(\rho\tau) &= 1 + n(t_1) + \cdots + n(t_m), \\
\alpha(\rho\tau) &= \binom{n(\rho\tau) - 1}{n(t_1), \ldots, n(t_m)} \cdot \alpha(t_1) \cdots \alpha(t_m) \frac{1}{\mu_1! \mu_2! \ldots}, \\
\gamma(\rho\tau) &= n(\rho\tau) \cdot \gamma(t_1) \cdots \gamma(t_m), \\
\Phi_i(\rho\tau) &= \sum_{j_1, \ldots, j_m} a_{ij_1} \Phi_{j_1}(t_1) \cdots a_{ij_m} \Phi_{j_m}(t_m), \\
F(\rho\tau)(y) &= f^{(m)}(y) \cdot (F(t_1)(y), ..., F(t_m)(y)).
\end{aligned} \tag{1.2.2}$$

Los enteros $\mu_1, \mu_2, ..$ en (1.2.2) dan cuenta del número de árboles iguales entre $t_1, ..., t_m$.
Se define el peso elemental del método Runge-Kutta asociado al árbol con raíz $\rho\tau$ como

$$\Phi(\rho\tau) = \sum_{i=1}^{s} b_i \Phi_i(\rho\tau). \tag{1.2.3}$$

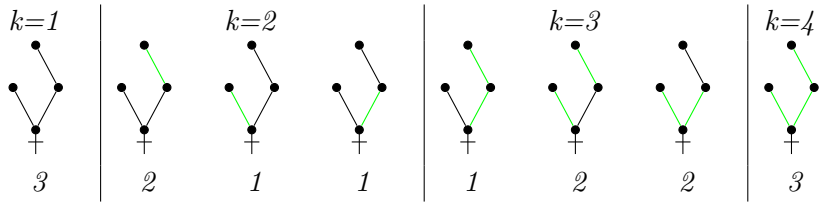
El desarrollo de Taylor de la solución exacta de (1.1.4) es [2]

$$y(t_0 + h) = y_0 + \sum_{\rho\tau \in \mathcal{RT}} \frac{h^{n(\rho\tau)}}{n(\rho\tau)!} \alpha(\rho\tau) F(\rho\tau)(y_0). \tag{1.2.4}$$

Introducimos también la siguiente definición que es necesaria para entender un resultado fundamental posterior.

Definicion 1.2.2. *Sea $\rho\tau \in \mathcal{RT}$ un árbol con raíz. Una partición de $\rho\tau$ en k subárboles con raíz $s_1, ..., s_k$ es un conjunto S de k-1 ramas de $\rho\tau$ tales que los árboles $s_1, ..., s_k$ se obtienen cuando las ramas de S se suprimen de $\rho\tau$. Tal partición se denota por $(\rho\tau, S)$. Denotamos además por $\alpha(\rho\tau, S)$ el número de etiquetados monótonos de $\rho\tau$ para los cuales los vértices de cada subárbol con raíz s_j están etiquetados consecutivamente.*

Ejemplo 1.2.1. *Mostramos a continuación las posibles particiones del árbol $\rho\tau = [\rho\tau_1, [\rho\tau_1]]$ en k subárboles junto con los números $\alpha(\rho\tau, S)$ debajo de cada partición. Las ramas coloreadas en verde constituyen el conjunto S de la definición.*



Hairer y Wanner introdujeron la noción de B-serie [9] que admite como casos particulares tanto a la solución exacta de (1.1.4) como a la solución numérica (1.2.1) calculada tras un paso con un método Runge-Kutta, pero que, como veremos más adelante, permite representar otros objetos de interés en este trabajo.

Definicion 1.2.3. *Dada una aplicación real a definida en la unión de $\mathcal{RT}$ y el conjunto vacío, $\emptyset$, una B-serie $B(a, y)$ es una serie formal de potencias*

$$a(\emptyset)y + \sum_{\rho\tau \in \mathcal{RT}} \frac{h^{n(\rho\tau)}}{n(\rho\tau)!} \alpha(\rho\tau) a(\rho\tau) F(\rho\tau)(y), \tag{1.2.5}$$

donde $n(\rho\tau), \alpha(\rho\tau)$ y $F(\rho\tau)(y)$ están dados por (1.2.2).

Al flujo exacto de (1.1.4), representado en (1.2.4), le corresponde la B-serie con coeficientes $a \equiv 1$, mientras que para la solución (1.2.1) obtenida con un método Runge-Kutta,

$$a(\emptyset) = 1, \quad a(\rho\tau) = \gamma(\rho\tau) \sum_{i=1}^{s} b_i \Phi_i(\rho\tau), \qquad \rho\tau \in \mathcal{RT}. \tag{1.2.6}$$

Aunque hayamos podido expresar nuestro método numérico como una B-serie $B(a, y)$, si la función $f(y)$ solo es N veces continuamente diferenciable, hay que truncar (1.2.5) a una suma finita sobre $\rho\tau \in \mathcal{RT}$ con $n(\rho\tau) \leq N$, y se añade un resto $O(h^{N+1})$.
Supondremos en lo que sigue que el método numérico es al menos de orden 1, es decir, $a(\emptyset) = 1$ y $a(\rho\tau_1) = 1$.

1.3. Construcción de las ecuaciones modificadas utilizando B-series

El objetivo de esta sección es, dado un método de un paso $\Psi_{h,f}$ representable por una B-serie (1.2.5), construir una serie formal de potencias para $f_h(y)$, es decir, encontrar una aplicación $b : \mathcal{RT} \to \mathbb{R}$ con

$$f_h(y) = \sum_{\rho\tau \in \mathcal{RT}} \frac{h^{n(\rho\tau)-1}}{n(\rho\tau)!} \alpha(\rho\tau) b(\rho\tau) F(\rho\tau)(y), \tag{1.3.1}$$

de forma que para cada $N \geq 1$, si denotamos

$$f_{N,h}(y) = \sum_{1 \leq n(\rho\tau) \leq N} \frac{h^{n(\rho\tau)-1}}{n(\rho\tau)!} \alpha(\rho\tau) b(\rho\tau) F(\rho\tau)(y), \tag{1.3.2}$$

el sistema diferencial $\dot{\tilde{y}} = f_{N,h}(\tilde{y})$ es una ecuación modificada de orden N para el método $\Psi_{h,f}$, es decir, $\Phi_{h,f_{N,h}}(y) - \Psi_{h,f}(y) = O(h^{N+1})$ para todo y y para toda f suficientemente regular.

Teorema 1.3.1. *Sea $B(a, y)$ la B-serie asociada a un método de un paso para la integración numérica de (1.1.4) y supongamos que f es N veces diferenciable con continuidad. Entonces la solución numérica satisface*

$$y_1 = \tilde{y}(t_0 + h) + O(h^{N+1}), \tag{1.3.3}$$

donde $\tilde{y}(t)$ es la solución exacta de la ecuación diferencial modificada

$$\dot{\tilde{y}} = f_{N,h}(\tilde{y}), \tag{1.3.4}$$

con lado derecho dado por (1.3.2). Los coeficientes $b(\rho\tau)$ se definen recursivamente por

$$a(\rho\tau) = \sum_{k=1}^{n(\rho\tau)} \frac{1}{k!} \sum_{(\rho\tau,S)} \binom{n(\rho\tau)}{n(s_1),\ldots,n(s_k)} \frac{\alpha(\rho\tau,S)}{\alpha(\rho\tau)} b(s_1)\cdots b(s_k). \tag{1.3.5}$$

Demostración. Suponiendo suficiente regularidad de $f(y)$, probaremos que las derivadas de $\tilde{y}(t)$ satisfacen

$$\frac{h^k}{k!}\tilde{y}^{(k)}(t) = \sum_{\substack{\rho\tau \in \mathcal{LRT},\\ k \le n(\rho\tau) \le N}} \frac{h^{n(\rho\tau)}}{n(\rho\tau)!} b^k(\rho\tau) F(\rho\tau)(\tilde{y}(t)) + O(h^{N+1}), \tag{1.3.6}$$

con coeficientes $b^k(\rho\tau)$ que serán construidos más adelante. Se trabaja con árboles etiquetados $\mathcal{LRT}$, luego puede omitirse el factor $\alpha(\rho\tau)$. Por definición de $\tilde{y}(t)$ esta fórmula (1.3.6) es cierta para $k = 1$ con $b^1(\rho\tau) = b(\rho\tau)$. Para k general diferenciamos k veces (1.3.4) utilizando (1.3.2) para obtener

$$\frac{h^k}{k!}\tilde{y}^{(k)}(t) = \frac{1}{k!} \sum_{n(u)\le N} \frac{h^{n(u)}}{n(u)!} b(u) \cdot h^{k-1} \frac{d^{k-1}}{dt^{k-1}}(F(u)(\tilde{y}(t))). \tag{1.3.7}$$

Por tanto será necesario calcular las derivadas de $F(u)(\tilde{y}(t))$. Fijado un árbol etiquetado $u \in \mathcal{LRT}$, obtendremos $n(u)$ sumandos. Usando la regla de Leibniz y la regla de la cadena y reemplazando $h\dot{\tilde{y}}(t)$ por (1.3.2), se tiene

$$h\frac{d}{dt}(F(u)(\tilde{y}(t))) = \sum_{i=1}^{n(u)} \sum_{n(v)\le N} \frac{h^{n(v)}}{n(v)!} b(v) F(u \circ_i v)(\tilde{y}(t)), \tag{1.3.8}$$

donde $u \circ_i v$ denota el árbol etiquetado que se obtiene añadiendo una nueva rama con v al vértice de u con etiqueta i. Continuamos diferenciando (1.3.8) y cada vez reemplazamos en el lado derecho las primeras derivadas de $F(u)(\tilde{y}(x))$ por (1.3.8).

$$\begin{aligned} h^l \frac{d^l}{dt^l}(F(u_1)(\tilde{y}(t))) = \sum_{u_2,\ldots,u_l \in \mathcal{LRT}} \sum_{i_1=1}^{n_1} \sum_{i_2=2}^{n_1+n_2} \sum_{i_l=1}^{n_1+\cdots+n_l} \frac{h^{n(u_2)+\cdots+n(u_l)}}{n(u_2)!\cdots n(u_l)!} \\ \cdot b(u_2)\cdots b(u_l) F(u_1 \circ_{i_1} u_2 \circ_{i_2} \cdots \circ_{i_{l-1}} u_l)(\tilde{y}(t)), \end{aligned} \tag{1.3.9}$$

donde la primera suma es sobre aquellos $u_2, ..., u_l \in \mathcal{LRT}$ para los cuales $n_i = n(u_i) \leq N$ y donde hemos usado la notación $u \circ_i v \circ_j w := (u \circ_i v) \circ_j w$, etc. Insertando (1.3.9) con $l = k-1$ en (1.3.7) y agrupando los coeficientes de aquellos $(u_1, i_1, u_2, i_2, ..., u_l)$ para los cuales $u_1 \circ_{i_1} u_2 \circ_{i_2} \cdots \circ_{i_{i-1}} u_l$, considerado como un árbol sin etiquetas, es igual a $\rho\tau$, obtenemos (1.3.6) con

$$b^k(\rho\tau) = \frac{1}{k!} \sum_{(\rho\tau, S)} \binom{n(\rho\tau)}{n(s_1), \ldots, n(s_k)} \frac{\alpha(\rho\tau, S)}{\alpha(\rho\tau)} b(s_1) \cdots b(s_k), \tag{1.3.10}$$

donde el sumatorio se extiende a todas las particiones de $\rho\tau$ en k subárboles con raíz $s_1, ..., s_k$.
Del desarrollo en serie de Taylor

$$\tilde{y}(x_0 + h) = y_0 + \sum_{k=1}^{N} \frac{h^k}{k!} \tilde{y}^{(k)}(x_0) + O(h^{N+1}),$$

y de (1.2.5) vemos que podemos obtener (1.3.3) poniendo

$$a(\rho\tau) = \sum_{k=1}^{n(\rho\tau)} b^k(\rho\tau), \qquad \text{para todo } \rho\tau \text{ con } 1 \leq n(\rho\tau) \leq N. \quad \square$$

En la Sección 1.1 construimos las ecuaciones modificadas escribiendo directamente los primeros términos de los desarrollos de Taylor, procedimiento que puede sistematizarse y hacerse menos costoso por medio de las B-series y del Teorema 1.3.1. A continuación procedemos a aplicar esta nueva herramienta a los tres métodos utilizados en el Ejemplo 1.1.1.
Puesto que vamos a usar la representación gráfica de los árboles, incluimos una tabla en la que aparecen los árboles con 4 nodos o menos.

Vamos a construir, en primer lugar, las relaciones (1.3.5) que dicta el Teorema 1.3.1 para los árboles de la Tabla 1.1.
Para el único árbol de orden 1 tenemos

$$a(\rho\tau_1) \quad = \quad b(\rho\tau_1). \tag{1.3.11}$$

Para el árbol de orden 2,

$$a(\rho\tau_2) \quad = \quad b(\rho\tau_2) + b(\rho\tau_1)^2. \tag{1.3.12}$$

Para los dos árboles con raíz de 3 nodos,

$$\begin{aligned} a(\rho\tau_{31}) &= b(\rho\tau_{31}) + \frac{1}{2}\binom{3}{2,1} b(\rho\tau_2) b(\rho\tau_1) + \frac{1}{3!}\binom{3}{1,1,1} b(\rho\tau_1)^3, \\ a(\rho\tau_{32}) &= b(\rho\tau_{32}) + \frac{1}{2}\left(\binom{3}{2,1} b(\rho\tau_2) b(\rho\tau_1) + \binom{3}{2,1} b(\rho\tau_2) b(\rho\tau_1)\right) \\ &\quad + \frac{1}{3!}\binom{3}{1,1,1} b(\rho\tau_1)^3. \end{aligned} \tag{1.3.13}$$

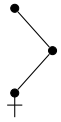
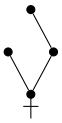
$n(\rho\tau)$	
1	 $\rho\tau_1$
2	 $\rho\tau_2$
3	 $\rho\tau_{31}$  $\rho\tau_{32}$
4	 $\rho\tau_{41}$  $\rho\tau_{42}$  $\rho\tau_{43}$  $\rho\tau_{44}$

Tabla 1.1: Árboles con raíz hasta orden 4

Por último, para los cuatro árboles con raíz de orden 4,

$$\begin{aligned}
a(\rho\tau_{41}) &= b(\rho\tau_{41}) + \frac{1}{2}\binom{4}{3,1} b(\rho\tau_{31})b(\rho\tau_1) + \frac{1}{6}\binom{4}{2} b(\rho\tau_2)b(\rho\tau_1)^2 \\
&\quad + \frac{1}{24}\binom{4}{1} b(\rho\tau_1)^4, \\
a(\rho\tau_{42}) &= b(\rho\tau_{42}) + \frac{1}{2}\left(\binom{4}{3}\frac{2}{3} b(\rho\tau_{31})b(\rho\tau_1) + \binom{4}{3}\frac{1}{3} b(\rho\tau_{32})b(\rho\tau_1)\right. \\
&\quad \left. + \binom{4}{2}\frac{1}{3} b(\rho\tau_2)^2\right) + \frac{1}{3!}\binom{4}{2,1,1}\frac{5}{3} b(\rho\tau_2)b(\rho\tau_2)^2 + b(\rho\tau_1)^4, \\
a(\rho\tau_{43}) &= b(\rho\tau_{43}) + \frac{1}{2}\left(\binom{4}{3,1} b(\rho\tau_{32})b(\rho\tau_1) + \binom{4}{3,1} b(\rho\tau_{31})b(\rho\tau_1)\right) \\
&\quad + \frac{1}{3!}\binom{4}{2,1,1} 2b(\rho\tau_2)b(\rho\tau_1)^2 + b(\rho\tau_1)^4, \\
a(\rho\tau_{44}) &= b(\rho\tau_{44}) + \frac{1}{2}\left(\binom{4}{3,1} 2b(\rho\tau_{32})b(\rho\tau_1) + \binom{4}{2,2} b(\rho\tau_2)^2\right) \\
&\quad + \frac{1}{3!}\binom{4}{2,1,1} 3b(\rho\tau_2)b(\rho\tau_1)^2 + \frac{1}{4!}\binom{4}{1,1,1,1} b(\rho\tau_1)^4). \qquad (1.3.14)
\end{aligned}$$

Una vez conocidas las relaciones entre los coeficientes de las B-series impli-

cadas, volvemos a la ecuación diferencial (1.1.9). Como todos los métodos utilizados para integrar numéricamente (1.1.9) son casos particulares de los métodos Runge-Kutta, los coeficientes de las correspondientes B-series vienen dados por $a(\rho\tau) = \gamma(\rho\tau)\sum_{i=1}^{s} b_i\Phi_i(\rho\tau)$, para cada $\rho\tau \in \mathcal{RT}$. Solo queda despejar de las ecuaciones anteriores (1.3.11)-(1.3.14) los coeficientes b necesarios para construir la B-serie que define el lado derecho de la ecuación modificada (1.3.2) en cada caso particular. Se empieza despejando por la ecuación relativa a $a(\rho\tau_1)$, siguiendo en orden creciente del orden de los árboles. Nótese que en este ejemplo particular (1.1.9), no es necesario calcular $a(\rho\tau_{41})$ ni $b(\rho\tau_{41})$ puesto que la diferencial elemental asociada a dicho árbol es cero, y por tanto no van a aparecer en las B-series (1.2.5) y (1.3.2).

Método de Euler explícito.

De la comparación de (1.1.10) y (1.2.5) es fácil ver que el único coeficiente no nulo es $a(\rho\tau_1) = 1$. Despejando de (1.3.11)-(1.3.14), tenemos

$$\begin{array}{llll} b(\rho\tau_1) = 1, & b(\rho\tau_2) = -1, & b(\rho\tau_{31}) = \frac{1}{2}, & b(\rho\tau_{32}) = 2, \\ b(\rho\tau_{42}) = -\frac{2}{3}, & b(\rho\tau_{43}) = -2, & b(\rho\tau_{44}) = -6. & \end{array}$$

Sustituyendo en (1.3.2) obtenemos la ecuación modificada

$$\begin{aligned} \dot{\tilde{y}} &= f(\tilde{y}) + \frac{h}{2}(-1)f'f(\tilde{y}) + \frac{h^2}{6}(\frac{1}{2}f''(f,f)(\tilde{y}) + 2f'f'f(\tilde{y})) \\ &\quad + \frac{h^3}{24}\left(-\frac{2}{3}\cdot 3f''(f,f'f)(\tilde{y}) - 2f'f''(f,f)(\tilde{y}) - 6f'f'f'f(\tilde{y})\right) + O(h^4). \end{aligned}$$

En el caso particular de (1.1.9) en que $f(\tilde{y}) = \tilde{y}^2$ la ecuación modificada es

$$\tilde{y}' = \tilde{y}^2 - h\tilde{y}^3 + h^2\frac{3}{2}\tilde{y}^4 - h^3\frac{8}{3}y^5 + O(h^4),$$

como ya se vió en la Sección 1.1.

Regla implícita del punto medio.

En este caso,

$$\begin{array}{llll} a(\rho\tau_1) = 1, & a(\rho\tau_2) = 1, & a(\rho\tau_{31}) = \frac{3}{4}, & a(\rho\tau_{32}) = \frac{3}{2}, \\ a(\rho\tau_{42}) = 1, & a(\rho\tau_{43}) = \frac{3}{2}, & a(\rho\tau_{44}) = 3. & \end{array}$$

De nuevo, despejando de (1.3.11)-(1.3.14), se obtiene

$$\begin{array}{llll} b(\rho\tau_1) = 1, & b(\rho\tau_2) = 0, & b(\rho\tau_{31}) = -\frac{1}{4}, & b(\rho\tau_{32}) = \frac{1}{2}, \\ b(\rho\tau_{42}) = 0, & b(\rho\tau_{43}) = 0, & b(\rho\tau_{44}) = 0. & \end{array}$$

Por lo tanto la ecuación modificada queda

$$\dot{\tilde{y}} = f(\tilde{y}) + \frac{h^2}{6}(-\frac{1}{4}f''(f,f)(\tilde{y}) + \frac{1}{2}f'f'f(\tilde{y})) = \tilde{y}^2 + \frac{h^2}{4}\tilde{y}^4 + O(h^4).$$

Método Runge-Kutta de orden 4.

De (1.2.6) y de las definiciones (1.2.2) se deduce

$$\begin{array}{llll} a(\rho\tau_1) = 1, & a(\rho\tau_2) = 1, & a(\rho\tau_{31}) = 1, & a(\rho\tau_{32}) = 1, \\ a(\rho\tau_{42}) = 1, & a(\rho\tau_{43}) = 1, & a(\rho\tau_{44}) = 1. & \end{array}$$

Despejando de (1.3.11)-(1.3.14), se llega a

$$\begin{array}{llll} b(\rho\tau_1) = 1, & b(\rho\tau_2) = 0, & b(\rho\tau_{31}) = 0, & b(\rho\tau_{32}) = 0, \\ b(\rho\tau_{42}) = 0, & b(\rho\tau_{43}) = 0, & b(\rho\tau_{44}) = 0. & \end{array}$$

La ecuación modificada resultante es la misma que se obtuvo en la Sección 1.1

$$\tilde{y}' = \tilde{y}^2 - h^4\frac{1}{24}\tilde{y}^6 + O(h^6).$$

Aunque los términos en h^4 no se pueden calcular con los árboles representados en la tabla, estos se han especificado también.
Los métodos empleados, método de Euler explícito, regla implícita del punto medio y método de Runge de 4 etapas, tienen orden 1, 2 y 4, respectivamente. Como vamos a ver, esto provoca que no aparezcan las funciones f_j en (1.1.5) hasta cierto j. Esto se ha podido observar en los ejemplos. No es casualidad, sino resultado del orden de consistencia del método, como muestra el siguiente teorema.

Teorema 1.3.2. *Supongamos que el método $y_{n+1} = \Psi_{h,f}(y_n)$ es de orden r, es decir,*

$$\Psi_{h,f}(y) = \Phi_{h,f}(y) + h^{r+1}\delta_{r+1}(y) + O(h^{r+2}),$$

donde $\Phi_{h,f}(y)$ denota el flujo exacto de $\dot{y} = f(y)$, y $h^{r+1}\delta_{r+1}$ es el término principal del error de truncación local. La ecuación modificada entonces satisface

$$\dot{\tilde{y}} = f(\tilde{y}) + h^r f_{r+1}(\tilde{y}) + h^{r+1}f_{r+2}(\tilde{y}) + \cdots, \qquad \tilde{y}(t_0) = y_0,$$

con $f_{r+1}(y) = \delta_{r+1}(y)$.

Demostración. Supongamos que el método $\Psi_{h,f}$ tiene orden r y que hemos podido expresarlo como una B-serie (1.2.5). Puesto que tiene orden r, $a(\rho\tau) \equiv 1$ para todo $\rho\tau$ tal que $n(\rho\tau) \leq r$. Queremos probar que $b(\rho\tau) \equiv 0$ para todo árbol con raíz $\rho\tau$ tal que $2 \leq n(\rho\tau) \leq r$ y vamos a razonar por inducción.
De las fórmulas (1.3.11)-(1.3.12) se deduce que

$$\begin{aligned} b(\rho\tau_1) &= a(\rho\tau_1) = 1, \\ b(\rho\tau_2) &= a(\rho\tau_2) - a(\rho\tau_1)^2 = 0. \end{aligned}$$

Supongamos que $b(\rho\tau) \equiv 0$ para todo árbol con raíz $\rho\tau$ tal que $2 \leq n(\rho\tau) \leq N < r$.
Sea $\rho\tau^*$ un árbol con raíz con $n(\rho\tau^*) = N + 1$. Observemos que según (1.3.5)

$$\begin{aligned} b(\rho\tau^*) &= a(\rho\tau^*) - (\text{combinación lineal de productos de } b(s) \text{ con } n(s) < n(\rho\tau^*)) \\ &\quad -b(\rho\tau_1)^{n(\rho\tau^*)}. \end{aligned}$$

Por tanto, puesto que $a(\rho\tau^*) = 1$, $b(\rho\tau_1) = 1$ y por hipótesis de inducción $b(s) = 0$ para todo s con $n(s) \leq N < N + 1 = n(\rho\tau^*)$, se obtiene que $b(\rho\tau^*) = 0$.
De (1.3.1) se sigue que las funciones $f_j(y)$ se construyen del siguiente modo

$$f_j(y) = \sum_{\rho\tau \in \mathcal{RT}, n(\rho\tau)=j} \frac{\alpha(\rho\tau) b(\rho\tau) F(\rho\tau)(y)}{n(\rho\tau)!},$$

de donde se deduce que para $2 \leq j \leq r$, como $b(\rho\tau) = 0$ para todo $\rho\tau$ tal que $n(\rho\tau) = j$, entonces $f_j(y) \equiv 0$. $\square$

Capítulo 2

Ecuaciones modificadas de métodos simplécticos.

Las primeras secciones de este capítulo se dedican a la presentación de la clase de problemas que vamos a tratar, los sistemas Hamiltonianos, y a la definición de aplicación simpléctica. Mostraremos que el flujo de un sistema Hamiltoniano es una transformación simpléctica y veremos las implicaciones de utilizar un integrador simpléctico para aproximar la solución de un sistema Hamiltoniano. Continuaremos con la presentación de algunas familias de integradores simplécticos, estudiando las condiciones que deben satisfacer los coeficientes de un método Runge-Kutta para ser simpléctico. En la siguiente sección se realiza el análisis regresivo de los errores de los integradores simplécticos, presentando el resultado fundamental del capítulo: cuando se integra un sistema diferencial Hamiltoniano con un método simpléctico, las ecuaciones modificadas son también Hamiltonianas.

2.1. Sistemas Hamiltonianos

Empezamos introduciendo la clase de problemas que vamos a tratar así como algo de notación. Sea Ω un dominio (subconjunto no vacío, abierto, conexo) en el espacio Euclídeo $\mathbb{R}^{2d}$ de puntos $(p,q) = (p_1, ..., p_d, q_1, ..., q_d)$. Denotaremos por I a un intervalo abierto de la recta real $\mathbb{R}$, ya sea acotado o no acotado. Si $H = H(p,q,t)$ es una función real suficientemente regular definida en el producto $\Omega \times I$, entonces el *sistema Hamiltoniano* de ecuaciones diferenciales con función *Hamiltoniana* H viene dado por

$$\frac{d}{dt}p_i(t) = -\frac{\partial H}{\partial q_i}, \qquad \frac{d}{dt}q_i(t) = \frac{\partial H}{\partial p_i}, \qquad i = 1, ..., d. \tag{2.1.1}$$

El entero d es el número de grados de libertad del sistema Hamiltoniano, Ω es el *espacio de fases* y el producto $\Omega \times I$ es el *espacio de fases extendido*. Aquí

supondremos al menos continuidad C^2 para la función Hamiltoniana, de forma que el lado derecho de (2.1.1) es C^1 y los teoremas de existencia y unicidad de solución para problemas de valores iniciales pueden aplicarse.
En mecánica, las variables q se llaman normalmente *coordenadas generalizadas*, las variables p son los *momentos generalizados* y H corresponde a la energía mecánica total.

Consideremos la matriz

$$J = \begin{pmatrix} 0_d & \mathbb{I}_d \\ -\mathbb{I}_d & 0_d \end{pmatrix}, \tag{2.1.2}$$

siendo $\mathbb{I}_d$ y 0_d las matrices unidad e indénticamente nulas de dimensión d, y sea $y = (p, q)^T$.
El sistema (2.1.1) puede reescribirse en términos de la matriz (2.1.2) de la siguiente manera

$$\frac{d}{dt}y(t) = (J^{-1}\nabla_y H)y(t). \tag{2.1.3}$$

Cuando el Hamiltoniano presenta una estructura

$$H(p, q, t) = T(p) + V(q, t), \tag{2.1.4}$$

el Hamiltoniano se llama *separable*. Un caso especial es el de Hamiltoniano independiente del tiempo o *autónomo*, en cuyo caso H es una función definida en el espacio de fases Ω. A lo largo del capítulo supondremos que el Hamiltoniano es autónomo salvo que se indique lo contrario.

Denotaremos por $\Phi_{t,H}$ al flujo del sistema Hamiltoniano (2.1.1), visto como una aplicación de Ω en sí mismo, de forma que $(p, q) = \Phi_{t,H}(p_0, q_0)$ es el valor en tiempo t de la solución de (2.1.1) con condición inicial (p_0, q_0) en $t = 0$.

2.2. Transformaciones simplécticas

Definicion 2.2.1. *Una aplicación $\Psi : \mathbb{R}^{2d} \to \mathbb{R}^{2d}$ se llama simpléctica con respecto a una matriz J, constante e invertible, si la matriz jacobiana $\Psi'(y)$ satisface*

$$(\Psi'(y))^T J \Psi'(y) = J \tag{2.2.1}$$

para todo y en el dominio de definición de Ψ. En el caso de que J esté definida por (2.1.2), se usa la denominación aplicación canónica como sinónimo de aplicación simpléctica.

La matriz J de (2.1.2) tiene la propiedad de que para cada par de vectores v, w de $\mathbb{R}^{2d}$, $v^T J w$ representa la suma de las áreas 2-dimensionales de los d paralelogramos resultantes de proyectar el paralelogramo determinado por v y w en

los planos de las variables (p_i,q_i), $1 \leq i \leq d$.

Ahora fijemos un punto (p, q) en Ω y construyamos un paralelogramo $\mathcal{P}$ con un vértice en (p, q) y lados los vectores v y w. Entonces $\Psi(\mathcal{P})$ es un paralelogramo con los lados curvados, que puede aproximarse a un paralelogramo $\mathcal{P}^*$ con un vértice en $\Psi(p, q)$ y lados $\Psi' v, \Psi' w$. Para ello se desprecian términos de orden superior a los lineales en v y w. Entonces $\mathcal{P}$ y $\mathcal{P}^*$ tienen el mismo área si y solamente si

$$v^T \Psi'^T J \Psi' w = v^T J w.$$

Esta relación se satisface para todos los paralelogramos $\mathcal{P}$ en Ω si y solamente si se verifica (2.2.1), es decir, si las transformación Ψ es simpléctica o canónica.

A continuación probamos que el flujo de (2.1.1) es una transformación simpléctica.

Teorema 2.2.1. *Para cada t, el flujo $\Phi_{t,H}$ de un sistema Hamiltoniano (2.1.1) es una transformación simpléctica.*

Demostración. Denotemos al jacobiano del flujo $\Phi_{t,H}$ por $F(t) \in \mathbb{R}^{2d \times 2d}$,

$$F(t) = \frac{\partial}{\partial y} \Phi_{t,H}(y^0).$$

Derivando con respecto al tiempo,

$$\begin{aligned}
\frac{d}{dt} F(t) &= \frac{\partial}{\partial t} \frac{\partial}{\partial y} \Phi_{t,H}(y^0) \\
&= \frac{\partial}{\partial y} \left[\frac{\partial}{\partial t} \Phi_{t,H}(y^0) \right] \\
&= \frac{\partial}{\partial y} \left[J^{-1} \nabla_y H(\Phi_{t,H}(y^0)) \right] \\
&= J^{-1} H_{yy}(y(t)) \left[\frac{\partial}{\partial y} \Phi_{t,H}(y^0) \right] \\
&= J^{-1} H_{yy}(y(t)) F(t),
\end{aligned}$$

donde se ha usado que

$$\frac{d}{dt} y(t) = \frac{\partial}{\partial t} \Phi_{t,H}(y^0) = J^{-1} \nabla_y H(\Phi_{t,H}(y^0)) = J^{-1} \nabla_y H(y(t)).$$

Tenemos por tanto que $F(t)$ es solución de la ecuación diferencial

$$\frac{d}{dt} F = J^{-1} H_{yy}(y(t)) F, \tag{2.2.2}$$

con condición inicial $F(0) = \mathbb{I}_{2d}$.
Hay que probar que

$$F(t)^T J F(t) = J.$$

Como $F(0) = \mathbb{I}_{2d}$, la afirmación es cierta para $t = 0$. Solo hay que probar que

$$\frac{d}{dt}K(t) = 0, \qquad \text{siendo} \qquad K(t) = F(t)^T J F(t). \tag{2.2.3}$$

En efecto, diferenciando en (2.2.3) se obtiene

$$\begin{aligned}
\frac{d}{dt}K(t) &= F(t)^T J \frac{d}{dt}F(t) + [\frac{d}{dt}F(t)]^T J F(t) \\
&= F(t)^T J[J^{-1}H_{yy}(y(t))F(t)] + [F(t)^T H_{yy}(y(t))(J^{-1})^T]JF(t) \\
&= F(t)^T H_{yy}(y(t))F(t) - F(t)^T H_{yy}(y(t))F(t) \\
&= 0. \quad \square
\end{aligned}$$

Además, si Ω es simplemente conexo, el recíproco también es cierto (ver [13]).

Esta propiedad de conservación de áreas del flujo de los sistemas Hamiltonianos tiene un impacto en el comportamiento a largo plazo de las soluciones de dichos sistemas: no puede haber sumideros ni fuentes, ni tampoco ciclos límite, ya que si los hubiera, en las proximidades de ellos el área se contraería o se expandería.
Por otro lado, del teorema de Liouville aplicado a la mecánica Hamiltoniana se sigue [1] que $det(\Phi') \equiv 1$, es decir, el flujo de un sistema Hamiltoniano conserva el volumen orientado en $\mathbb{R}^{2d}$. Para $d = 1$ esto se corresponde con la conservación del área orientada. Sin embargo la generalización correcta al pasar de $d = 1$ a $d > 1$ no es la conservación del volumen sino la simplecticidad, es decir, la conservación de la suma de las áreas orientadas de las proyecciones sobre los d planos coordenados (p_i, q_i). La simplecticidad caracteriza el flujo de los sistemas Hamiltonianos, mientras que la conservación del volumen es una propiedad más débil que comparten otros sistemas no Hamiltonianos.

2.3. Integradores simplécticos

Definicion 2.3.1. *Un método numérico de un paso*

$$(p_{n+1}, q_{n+1}) = \Psi_{h,H}(p_n, q_n) \tag{2.3.1}$$

se llama simpléctico si $\Psi_{h,H}$ es una transformación simpléctica para todos los Hamiltonianos H y todas las longitudes de paso h.

Existen diversas familias de integradores convencionales que contienen métodos que son simplécticos. En este capítulo presentamos los métodos Runge-Kutta, que ya se han mencionado en el Capítulo 1. Posteriormente, en el Capítulo 3, se presentan los métodos Runge-Kutta-Nyström que se utilizarán en los experimentos numéricos.

2.3.1. B-series simplécticas

Ya hemos visto que las B-series constituyen una herramienta poderosa para expresar métodos numéricos para la integración de ecuaciones diferenciales. En el caso de sistemas Hamiltonianos, es de especial interés encontrar métodos para los que $\Psi_{h,H}$ sea una transformación simpléctica para cada tamaño de paso h y cada Hamiltoniano H. Interesa, por tanto, establecer bajo qué condiciones sobre sus coeficientes, una B-serie define una transformación simpléctica. En esta sección se van a establecer las condiciones necesarias y suficientes que debe satisfacer una B-serie para que el método que define sea simpléctico, y se particularizará en el caso de B-series asociadas a métodos Runge-Kutta. Veremos también que el cumplimiento de estas condiciones hará que las condiciones de orden se vean notablemente simplificadas.
Consideremos una B-serie

$$y + \sum_{n=1}^{\infty} \frac{h^n}{n!} \sum_{\rho\tau \in \mathcal{RT}_n} c(\rho\tau) F(\rho\tau)(y) \tag{2.3.2}$$

donde h es un parámetro real, $\mathcal{RT}_n$ es el conjunto de árboles con raíz $\rho\tau$ de n vértices, $c(\rho\tau)$ es un coeficiente real asociado a $\rho\tau$ y $F(\rho\tau)(y)$ es la diferencial elemental correspondiente a $\rho\tau$ evaluada en $y \in \mathbb{R}^{2d}$. La notación utilizada en (2.3.2) se corresponde con la empleada en (1.2.5) tomando $c(\emptyset) = 1$ y $c(\rho\tau) = \alpha(\rho\tau)a(\rho\tau)$, y se han agrupado los términos que corresponden a árboles con el mismo número de nodos. Los coeficientes $c(\rho\tau)$ dependen del método específico, y en el caso de los métodos Runge-Kutta, siguiendo (1.2.6), $c(\rho\tau) = \alpha(\rho\tau)\gamma(\rho\tau)\sum_{i=1}^{s} b_i \Phi_i(\rho\tau)$. Antes de introducir el resultado fundamental de este apartado, definiremos una relación de equivalencia en el conjunto de árboles con raíz, y presentaremos el concepto de árboles *vecinos*.

Definicion 2.3.2. *Dos árboles con raíz $\rho\tau_1$ y $\rho\tau_2$ se identifican si solo difieren en la posición de la raíz, pero constan de los mismos vértices y aristas. Cada clase de equivalencia bajo esta relación se llama árbol, y se denota por τ (sin ρ).*

Definicion 2.3.3. *Dos árboles con raíz $\rho\tau_i$ (con raíz en el vértice i) y $\rho\tau_j$ (con raíz en el vértice j) son vecinos si ambos pertenecen a la misma clase de equivalencia τ y sus respectivas raíces son vértices adyacentes.*

Ejemplo 2.3.1. *En el árbol τ_{31} de la Tabla 2.1 si escogemos el vértice de la izquierda como raíz, obtenemos $\rho\tau_{31}$, mientras que si escogemos como raíz el nodo central, se obtiene $\rho\tau_{32}$. Por lo tanto, según la definición anterior, $\rho\tau_{31}$ y $\rho\tau_{32}$ son árboles vecinos.*

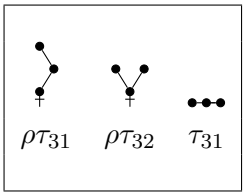


Tabla 2.1: Ejemplo 2.3.1

Antes de presentar un teorema que caracteriza a las B-series simplécticas, es necesario introducir algo más de notación. La Tabla 2.2 contiene un árbol τ, y dos árboles con raíz $\rho\tau_I$ y $\rho\tau_J$, con raíces en i y en j respectivamente, que resultan de eliminar en τ la arista que une el vértice i con el vértice j y de situar las raíces de los dos subárboles que se generan en los vértices con índices i y j, respectivamente.

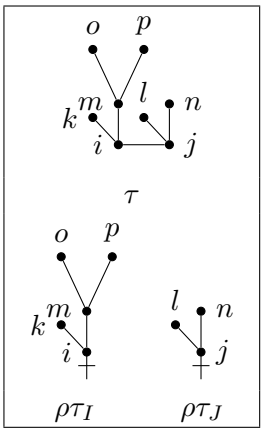


Tabla 2.2: Construcción para ilustrar el Teorema 2.3.1

Teorema 2.3.1. *La B-serie (2.3.2) es canónica, es decir, define una transformación simpléctica para cada h y cada problema Hamiltoniano (2.1.1) si, y solamente si, para cada par de árboles vecinos $\rho\tau_i$ y $\rho\tau_j$ se tiene*

$$\frac{c(\rho\tau_i)}{\alpha(\rho\tau_i)\gamma(\rho\tau_i)} + \frac{c(\rho\tau_j)}{\alpha(\rho\tau_j)\gamma(\rho\tau_j)} = \frac{c(\rho\tau_I)}{\alpha(\rho\tau_I)\gamma(\rho\tau_I)}\frac{c(\rho\tau_J)}{\alpha(\rho\tau_J)\gamma(\rho\tau_J)}. \tag{2.3.3}$$

Demostración. La demostración se puede encontrar en [5].

2.3.2. Métodos Runge-Kutta (RK)

Las ecuaciones para avanzar un paso de longitud h con un método Runge-Kutta con coeficientes (1.1.14) a problemas de valores iniciales generales (1.1.4) quedaron descritas en (1.1.15). Cuando un método RK con coeficientes (1.1.14) se utiliza para la integración de un sistema Hamiltoniano (2.1.1) resulta en las relaciones

$$\begin{aligned} P_i &= p_n - h\sum_{j=1}^{s} a_{ij} H_q(P_j, Q_j), \\ Q_i &= q_n + h\sum_{j=1}^{s} a_{ij} H_p(P_j, Q_j), \qquad (2.3.4) \\ p_{n+1} &= p_n - h\sum_{i=1}^{s} b_i H_q(P_i, Q_i), \\ q_{n+1} &= q_n + h\sum_{i=1}^{s} b_i H_p(P_i, Q_i), \qquad (2.3.5) \end{aligned}$$

donde H_q y H_p denotan las funciones d-dimensionales con componentes $\frac{\partial H}{\partial q_i}$ y $\frac{\partial H}{\partial p_i}$, $1 \le i \le d$, respectivamente, y P_i y Q_i son las etapas intermedias de las variables p y q.

En el caso particular de un método Runge-Kutta, el Teorema 2.3.1 se traduce en

Teorema 2.3.2. *Si los coeficientes (1.1.14) de un método Runge-Kutta de s etapas satisfacen las relaciones*

$$b_i a_{ij} + b_j a_{ji} - b_i b_j = 0, \qquad 1 \le i, j \le s, \qquad (2.3.6)$$

entonces el método es simpléctico.

Demostración. En este caso las cantidades $c(\rho\tau)/(\alpha(\rho\tau)\gamma(\rho\tau))$ son simplemente los pesos elementales $\Phi(\rho\tau)$. Por lo tanto (2.3.3) queda

$$\Phi(\rho\tau_i) + \Phi(\rho\tau_j) = \Phi(\rho\tau_I)\,\Phi(\rho\tau_J). \qquad (2.3.7)$$

Escribamos

$$\rho\tau_I = [\rho\tau_1, \ldots, \rho\tau_m],$$

donde $\rho\tau_\alpha, \alpha = 1, ..., m$, son los subárboles con raíz que se generan al suprimir de $\rho\tau_I$ su raíz y sus lados adyacentes, y de forma similar

$$\rho\tau_J = [\rho\tau^1, \ldots, \rho\tau^n],$$

siendo $\rho\tau^{\beta}, \beta = 1, ..., n$, los subárboles con raíz que se generan al suprimir de $\rho\tau_J$ su raíz y sus lados adyacentes.
Entonces

$$\begin{aligned} \rho\tau_i &= [\rho\tau_J, \rho\tau_1, \ldots, \rho\tau_m], \\ \rho\tau_j &= [\rho\tau_I, \rho\tau^1, \ldots, \rho\tau^n], \end{aligned}$$

y, de acuerdo a la definición de pesos elementales (1.2.3),

$$\begin{aligned} \Phi(\rho\tau_i) &= \sum_{i=1}^{s} b_i \Phi_i(\rho\tau_J) \prod_{\alpha=1}^{m} \Phi_i(\rho\tau_\alpha) = \sum_{i,j=1}^{s} b_i a_{ij} \prod_{\alpha=1}^{m} \Phi_i(\rho\tau_\alpha) \prod_{\beta=1}^{n} \Phi_j(\rho\tau^\beta), \\ \Phi(\rho\tau_j) &= \sum_{j=1}^{s} b_j \Phi_j(\rho\tau_I) \prod_{\beta=1}^{n} \Phi_j(\rho\tau^\beta) = \sum_{i,j=1}^{s} b_j a_{ji} \prod_{\alpha=1}^{m} \Phi_i(\rho\tau_\alpha) \prod_{\beta=1}^{n} \Phi_j(\rho\tau^\beta), \end{aligned}$$

donde Φ_i denota el peso elemental de la etapa i. Tras estos cálculos, (2.3.7) queda

$$\sum_{i,j=1}^{s} b_i a_{ij} \prod_{\alpha=1}^{m} \Phi_i(\rho\tau_\alpha) \prod_{\beta=1}^{n} \Phi_j(\rho\tau^\beta) + \sum_{i,j=1}^{s} b_j a_{ji} \prod_{\alpha=1}^{m} \Phi_i(\rho\tau_\alpha) \prod_{\beta=1}^{n} \Phi_j(\rho\tau^\beta) =$$

$$\left(\sum_{i=1}^{s} b_i \prod_{\alpha=1}^{m} \Phi_i(\rho\tau_\alpha) \right) \left(\sum_{j=1}^{s} b_j \prod_{\beta=1}^{n} \Phi_j(\rho\tau^\beta) \right).$$

Reorganizando lo anterior,

$$\sum_{i,j=1}^{s} (b_i a_{ij} + b_j a_{ji} - b_i b_j)(\prod_{\alpha=1}^{m} \Phi_i(\rho\tau_\alpha))(\prod_{\beta=1}^{n} \Phi_j(\rho\tau^\beta)) = 0. \tag{2.3.8}$$

De (2.3.8) se concluye que la condición

$$b_i a_{ij} + b_j a_{ji} - b_i b_j = 0, \qquad 1 \leq i, j \leq s, \tag{2.3.9}$$

es suficiente para garantizar la canonicidad de la B-serie del método Runge-Kutta. □

Aunque no incluimos la demostración en este trabajo, se puede probar que si el método Runge-Kutta no tiene etapas redundantes, la condición del teorema anterior tambien es una condición necesaria para la simplecticidad [13].

Condiciones de orden para métodos Runge-Kutta simplécticos

Las condiciones que un método Runge-Kutta debe satisfacer para alcanzar orden $\geq r$ son conocidas [8]: para cada árbol con raíz $\rho\tau$ con número de vértices $n(\rho\tau) \leq r$, debe cumplirse

$$\Phi(\rho\tau) = \frac{1}{\gamma(\rho\tau)}, \tag{2.3.10}$$

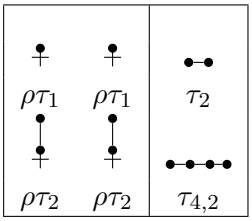


Tabla 2.3: Ejemplos de árboles superfluos

siendo $\gamma(\rho\tau)$ la densidad y $\Phi(\rho\tau)$ el peso elemental del árbol $\rho\tau$.
Si se consideran el número de etapas s, y los coeficientes a_{ij}, b_j (1.1.14) como parámetros libres, entonces las condiciones (2.3.10) son independientes unas de otras. Sin embargo, cuando se imponen las condiciones (2.3.6), no todos los coeficientes del método son parámetros libres y aparecen ciertas redundancias entre las condiciones de orden (2.3.10). Por tanto, para conseguir orden $\geq r$, no es necesario satisfacer una ecuación para cada árbol con raíz de orden $\leq r$. El número de ecuaciones viene dado por el número de árboles no superfluos, que se definen a continuación.

Definicion 2.3.4. *Se llaman árboles superfluos a aquellos que resultan cuando dos copias del mismo árbol con raíz de N vértices se unen por sus respectivas raíces añadiendo una nueva rama, dando lugar a un árbol de 2N vértices.*

Por ejemplo, el árbol τ_2 en la Tabla 2.3 es superfluo porque es la yuxtaposición de dos copias de $\rho\tau_1$. Análogamente, $\tau_{4,2}$ es superfluo pues se puede construir como yuxtaposición de dos copias de $\rho\tau_2$.

Teorema 2.3.3. *Supongamos que el método Runge-Kutta satisface la condición de simplecticidad (2.3.6) y tiene orden de consistencia $\geq r \geq 1$. Entonces tiene orden de consistencia $\geq r+1$ si y solamente si para cada árbol no superfluo τ con r+1 vértices, hay un árbol con raíz $\rho\tau$ asociado con τ para el cual se cumple (2.3.10).*

Demostración. Ver [12].

El Teorema 2.3.3 implica que como el único árbol de dos vértices es suplerfluo, cada método Runge-Kutta simpléctico consistente (de orden 1) tiene orden al menos 2. Por otra parte, para asegurar orden ≥ 3, será suficiente con imponer la condición de orden asociada a $\rho\tau_{31}$, o bien la asociada a $\rho\tau_{32}$, y para tener orden 4 solo hay que imponer la condición de orden asociada a $\rho\tau_{4,1}$ (ver Tabla 1.1).

En la Tabla 2.4, tomada de [13], se ha recogido el número de condiciones de orden independientes que hay que imponer para garantizar un orden determinado tanto en el caso de métodos Runge-Kutta generales como en el caso de

Order	RK general	RK simpléctico
1	1	1
2	2	1
3	4	2
4	8	3
5	17	6
6	37	10
7	85	21
8	200	40

Tabla 2.4: Número de condiciones de orden

métodos Runge-Kutta simplécticos. Podemos observar la reducción en el número de condiciones de orden para órdenes altos.

2.4. Análisis regresivo de métodos simplécticos

En esta sección vamos a mostrar cómo afecta la propiedad de simplecticidad a las ecuaciones modificadas asociadas a un método de un paso. Vamos a probar que cuando un método Runge-Kutta simpléctico se utiliza para integrar un sistema Hamiltoniano, las ecuaciones modificadas construidas en el Capítulo 1 corresponden también a un sistema Hamiltoniano, pero con una función Hamiltoniana modificada. Más precisamente, vamos a probar que para cualquier Hamiltoniano independiente del tiempo y cualquier entero positivo N, se puede encontrar un sistema Hamiltoniano modificado (también autónomo) con función Hamiltoniana $H_{N,h}$ tal que $\Psi_{h,H}$ difiere del flujo $\Phi_{h,H_{N,h}}$ en términos $O(h^{N+1})$. La diferencia entre el verdadero Hamiltoniano H y el Hamiltoniano modificado $H_{N,h}$ es $O(h^r)$, siendo r el orden del método.

Veamos primero un lema que se va a utilizar para la prueba de la existencia de un Hamiltoniano modificado local.

Lema 2.4.1. *Sea $D \subset \mathbb{R}^n$ abierto y $f : D \to \mathbb{R}^n$ diferenciable con continuidad, y supongamos que el jacobiano $f'(y)$ es simétrico para todo $y \in D$. Entonces para cada $y_0 \in D$ existe un entorno y una función $H(y)$ tal que*

$$f(y) = \nabla H(y) \tag{2.4.1}$$

en dicho entorno.

Demostración. Supongamos que $y_0 = 0$, y consideramos una bola en torno a y_0

contenida en D. En esta bola definimos

$$H(y) = \int_0^1 y^T f(ty)\, dt + C,$$

con C una constante.
Derivando respecto a y_k y usando la hipótesis de simetría $\frac{\partial f_i}{\partial y_k} = \frac{\partial f_k}{\partial y_i}$ se tiene

$$\frac{\partial H}{\partial y_k}(y) = \int_0^1 \left(f_k(ty) + y^T \frac{\partial f}{\partial y_k}(ty)t \right) dt = \int_0^1 \frac{d}{dt}\left(t f_k(ty)\right) dt = f_k(y),$$

lo que prueba el lema. □
Para $D = \mathbb{R}^{2d}$ o para conjuntos estrellados, la demostración anterior muestra que la función H del Lema 2.4.1 está globalmente definida.

Teorema 2.4.1. ***Existencia de un Hamiltoniano modificado local***
Si se aplica un método simpléctico Ψ_h a un sistema Hamiltoniano con función Hamiltoniana suficientemente regular, $H : D \to \mathbb{R}$, siendo $D \subset \mathbb{R}^{2d}$ un abierto, entonces la ecuación modificada (1.1.5) es localmente Hamiltoniana con Hamiltoniano H_h. De forma más precisa, localmente existen funciones reales H_j para $j = 2, 3, ...$ tales que $f_j(y) = J^{-1}\nabla_y H_j(y)$, es decir,

$$H_h(y) = H(y) + hH_2(y) + h^2 H_3(y) + \cdots .$$

Demostración. Lo probaremos por inducción. Supongamos que $f_j(y) = J^{-1}\nabla_y H_j(y)$ para $j = 1, 2, ..., N$ (para $N = 1$ se satisface puesto que $f_1(y) = f(y) = J^{-1}\nabla_y H(y)$).
La ecuación modificada truncada

$$\dot{\tilde{y}} = f(\tilde{y}) + hf_2(\tilde{y}) + \cdots + h^{N-1} f_N(\tilde{y}),$$

es un sistema Hamiltoniano con Hamiltoniano $H(y) + hH_2(y) + \cdots + h^{N-1}H_N(y)$. Su flujo $\Phi_{t,f_{N,h}}(y_0)$ comparado con el de (1.1.5) satisface

$$\Psi_{h,H}(y_0) = \Phi_{h,f_{N,h}}(y_0) + h^{N+1} f_{N+1}(y_0) + O(h^{N+2}),$$

y los jacobianos satisfacen

$$\Psi'_{h,H}(y_0) = \Phi'_{h,f_{N,h}}(y_0) + h^{N+1} f'_{N+1}(y_0) + O(h^{N+2}).$$

Utilizando que el método es simpléctico, junto con la hipótesis de inducción, se concluye que $\Psi_{h,H}$ y $\Phi_{h,f_{N,h}}$ son transformaciones simplécticas. Además como $\Phi'_{h,f_{N,h}}(y_0) = I + O(h)$, se tiene que

$$J = \Psi'_{h,H}(y_0)^T J \Psi'_{h,H}(y_0) = J + h^{N+1}\left(f'_{N+1}(y_0)^T J + J f'_{N+1}(y_0) \right) + O(h^{N+2}).$$

Por consiguiente, el término entre paréntesis tiene que anularse y, por tanto, la matriz $Jf'_{N+1}(y)$ es simétrica. La existencia de $H_{N+1}(y)$ satisfaciendo $f_{N+1} = J^{-1}\nabla H_{N+1}(y)$ se sigue del Lema 2.4.1. □

Ejemplo 2.4.1. *En este ejemplo se pretende ilustrar la pérdida del carácter Hamiltoniano del sistema modificado al usar fórmulas no simplécticas: el proceso de integración numérica perturba el modelo de forma que lo saca de la clase Hamiltoniana. Para ello consideremos el Hamiltoniano con un grado de libertad $H = T(p) + V(q)$ y su correspondiente sistema diferencial*

$$\frac{dp}{dt} = f(q), \qquad \frac{dq}{dt} = g(p), \tag{2.4.2}$$

donde $f = -V'$ y $g = T'$. Integramos (2.4.2) con el método de Euler simpléctico

$$p_{n+1} = p_n + hf(q_n), \qquad q_{n+1} = q_n + hg(p_{n+1}). \tag{2.4.3}$$

Siguiendo la línea de las ecuaciones diferenciales modificadas, nos gustaría encontrar un sistema cuya solución esté más cerca de nuestro método numérico, que es de orden 1, y difiere en términos de $O(h^2)$ del verdadero flujo $\Phi_{h,H}$. Usando (1.1.8) es fácil comprobar que (2.4.3) es consistente de orden 2 con el sistema

$$\frac{dp}{dt} = f(q) - \frac{h}{2}f'(q)g(p), \tag{2.4.4a}$$

$$\frac{dq}{dt} = g(p) + \frac{h}{2}g'(p)f(q). \tag{2.4.4b}$$

Puesto que el método es simpléctico, el Teorema 2.4.1 garantiza que existe una función $H_2 : \mathbb{R}^2 \to \mathbb{R}$ tal que $f_2 = J^{-1}\nabla H_2(p,q)$, es decir

$$\begin{aligned} \frac{1}{2}f'(q)g(p) &= \frac{\partial H_2}{\partial q}, \\ \frac{1}{2}g'(p)f(q) &= \frac{\partial H_2}{\partial p}. \end{aligned}$$

Basta tomar $H_2 = \frac{1}{2}T'(p)V'(q)$ y $H_{2,h}(p,q) = T(p) + V(q) + \frac{h}{2}T'(p)V'(q)$ es el Hamiltoniano de la ecuación modificada truncada.

Si integramos ahora (2.4.2) con el método de Euler explícito, la ecuación modificada queda

$$\frac{dp}{dt} = f(q) - \frac{h}{2}f'(q)g(p), \tag{2.4.5a}$$

$$\frac{dq}{dt} = g(p) - \frac{h}{2}g'(p)f(q). \tag{2.4.5b}$$

La solución numérica es consistente hasta los términos $O(h^2)$ con la solución del sistema de ecuaciones modificadas que acabamos de calcular. Sin embargo, este nuevo sistema de ecuaciones modificadas (2.4.5a)-(2.4.5b) ya no es un sistema Hamiltoniano debido al cambio de signo en (2.4.4b) respecto de (2.4.5b). Esto se debe a que no es un integrador simpléctico.

A continuación veremos que un método numérico (2.3.1) simpléctico puede caracterizarse en términos de una función real S en lugar de en términos de las $2d$ componentes de $\Psi_{h,H}$. Dicha función S, se llama *función generatriz* de $\Psi_{h,H}$ (para simplificar la notación hemos suprimido la dependencia explícita respecto de h y H).
Dada una función $S(p,q)$, usaremos la siguiente notación

$$dS(p,q) = dS = S_p dp + S_q dq = \sum_{i=1}^{d} \left(\frac{\partial S}{\partial p_i}(p,q) dp_i + \frac{\partial S}{\partial q_i}(p,q) dq_i \right),$$

donde S_p y S_q son vectores fila con las derivadas parciales, y $dp = (dp_1, \ldots, dp_d)^T$, $dq = (dq_1, \ldots, dq_d)^T$.

Teorema 2.4.2. *Una aplicación $\Psi : (p,q) \to (P,Q)$ es simpléctica si y solamente si existe localmente una función S(p,q) tal que*

$$P^T dQ - p^T dq = dS. \tag{2.4.6}$$

Esto significa que $P^T dQ - p^T dq$ es una diferencial total.

Demostración. Tenemos el Jacobiano de Ψ

$$\frac{\partial(P,Q)}{\partial(p,q)} = \begin{pmatrix} P_p & P_q \\ Q_p & Q_q \end{pmatrix}.$$

Teniendo en cuenta la estructura de la matriz J, la condición de simplecticidad (2.2.1), da lugar a

$$P_p^T Q_p = Q_p^T P_p, \qquad P_p^T Q_q - I = Q_p^T P_q, \qquad Q_q^T P_q = P_q^T Q_q. \tag{2.4.7}$$

Si ahora insertamos $dQ = Q_p dp + Q_q dq$ en el lado izquierdo de (2.4.6) obtenemos

$$\left(P^T Q_p, P^T Q_q - p^T\right) \begin{pmatrix} dp \\ dq \end{pmatrix} = \begin{pmatrix} Q_p^T P \\ Q_q^T P - p \end{pmatrix}^T \begin{pmatrix} dp \\ dq \end{pmatrix}.$$

Para aplicar el Lema (2.4.1) tenemos que verificar primero la simetría del jacobiano del vector de coeficientes,

$$\begin{pmatrix} Q_p^T P_p & Q_p^T P_q \\ Q_q^T P_p - I & Q_q^T P_q \end{pmatrix} + \sum_i P_i \frac{\partial^2 Q_i}{\partial (p,q)^2}. \tag{2.4.8}$$

Como la matriz Hessiana de Q_i es simétrica, la simetría de la matriz (2.4.8) es equivalente a las condiciones de simplecticidad (2.4.7). $\square$.

Supongamos ahora que q y Q son funciones independientes en Ω, es decir, cada punto en Ω puede ser descrito unívocamente por los correspondientes valores de q y Q (por ejemplo, si $\frac{\partial Q}{\partial p}$ es invertible y p puede expresarse como función de q y Q). Entonces podemos expresar $S(p,q)$ en (2.4.6) como una función de q y Q. Comparando los coeficientes de

$$dS = \frac{\partial S(q,Q)}{\partial q}dq + \frac{\partial S(q,Q)}{\partial Q}dQ,$$

con los de (2.4.6) es fácil ver que

$$p = -\frac{\partial S}{\partial q}, \qquad P = \frac{\partial S}{\partial Q}, \tag{2.4.9}$$

relaciones que permiten reconstruir la transformación simpléctica $\Psi : (p,q) \to (P,Q)$ a partir de la función escalar $S(q,Q)$. Otra caracterización de una transformación simpléctica equivalente a (2.4.6) es la siguiente

$$Q^T dP + p^T dq = d(P^T q + S^1(P,q)). \tag{2.4.10}$$

La equivalencia se sigue de tomar

$$S^1 = P^T(Q-q) - S,$$

y de que

$$d(P^T Q) = P^T dQ + Q^T dP.$$

La transformación simpléctica definida por S^1 se deduce comparando los coeficietes de dq y dP en la ecuación (2.4.10)

$$p = P + \frac{\partial S^1}{\partial q}(P,q), \qquad Q = q + \frac{\partial S^1}{\partial P}(P,q). \tag{2.4.11}$$

Puesto que hemos visto que toda transformación simpléctica se puede escribir en términos de una función generatriz, demostramos a continuación un resultado que determina una función generatriz para los métodos Runge-Kutta simplécticos.

Teorema 2.4.3. *Supongamos que los coeficientes de un método Runge-Kutta (1.1.14) satisfacen*

$$b_i a_{ij} + b_j a_{ji} = b_i b_j, \qquad 1 \le i,j \le s. \tag{2.4.12}$$

Entonces, denotando en (2.3.4)–(2.3.5) $p = p_n$, $q = q_n$, $P = p_{n+1}$ *y* $Q = q_{n+1}$, *las ecuaciones del método Runge-Kutta*

$$\begin{aligned} P &= p - h\sum_{i=1}^{s} b_i H_q(P_i, Q_i), \qquad & P_i &= p - h\sum_{j=1}^{s} a_{ij} H_q(P_j, Q_j), \\ Q &= q + h\sum_{i=1}^{s} b_i H_p(P_i, Q_i), \qquad & Q_i &= q + h\sum_{j=1}^{s} a_{ij} H_p(P_j, Q_j) \end{aligned} \tag{2.4.13}$$

pueden escribirse como (2.4.11) con

$$S^1(P, q, h) = h\sum_{i=1}^{s} b_i H(P_i, Q_i) - h^2 \sum_{i,j=1}^{s} b_i a_{ij} H_q(P_i, Q_i)^T H_p(P_j, Q_j). \tag{2.4.14}$$

Demostración. Usamos las abreviaturas $H[i] = H(P_i, Q_i), H_p[i] = H_p(P_i, Q_i)$, etc. Derivamos en primer lugar $S^1(P, q, h)$ respecto a q.

$$\begin{aligned} \frac{\partial}{\partial q}\left(\sum_{i=1} b_i H[i]\right) &= \sum_i b_i H_p[i]^T \left(\frac{\partial p}{\partial q} - h\sum_j a_{ij}\frac{\partial}{\partial q} H_q[j]\right) \\ &+ \sum_i b_i H_q[i]^T \left(I + h\sum_j a_{ij}\frac{\partial}{\partial q} H_p[j]\right). \end{aligned}$$

Diferenciando la primera relación de (2.4.13)

$$0 = \frac{\partial p}{\partial q} - h\sum_j b_j \frac{\partial}{\partial q} H_q[j].$$

Utilizando la regla de Leibniz,

$$\frac{\partial}{\partial q}(H_q[i]^T H_p[j]) = H_q[i]^T \frac{\partial}{\partial q} H_p[j] + H_p[j]^T \frac{\partial}{\partial q} H_q[i]. \tag{2.4.15}$$

Y teniendo en cuenta (2.4.12), se obtiene

$$\frac{\partial S^1(P, q, h)}{\partial q} = h\sum_{i=1}^{s} b_i H_q[i], \qquad \frac{\partial S^1(P, q, h)}{\partial P} = h\sum_{i=1}^{s} b_i H_p[i].$$

La segunda relación se prueba de manera análoga. Por tanto las fórmulas Runge-Kutta (2.4.13) son equivalentes a (2.4.11) con S^1 dada por (2.4.14) . □

Por último, introducimos sin demostración un resultado que va a proporcionar una función generatriz para el flujo exacto de un sistema Hamiltoniano. Supongamos ahora fijo (p, q) y que el punto $(P(t), Q(t))$ se mueve en el flujo de un sistema Hamiltoniano. Queremos determinar una función generatriz $S(q, Q, t)$, ahora también dependiente de t, que genere via (2.4.9) la transformación simpléctica $(p, q) \to (P(t), Q(t))$ del flujo exacto del sistema Hamiltoniano. Se debe satisfacer

$$P_i(t) = \frac{\partial S}{\partial Q_i}(q, Q(t), t), \qquad p_i = -\frac{\partial S}{\partial q_i}(q, Q(t), t). \tag{2.4.16}$$

Teorema 2.4.4. *Si $S(q, Q, t)$ es una solución de*

$$\frac{\partial S}{\partial t} + H\left(\frac{\partial S}{\partial Q_1}, \ldots, \frac{\partial S}{\partial Q_d}, Q_1, \ldots, Q_d\right) = 0, \tag{2.4.17}$$

y si la matriz $\left(\frac{\partial^2 S}{\partial q_i \partial Q_j}\right)$ es invertible, existe una aplicación $(p, q) \to (P(t), Q(t))$ definida por (2.4.16) que es una solución del sistema Hamiltoniano. La ecuación de (2.4.17) se llama ecuación de Hamilton-Jacobi.

Ya hemos visto que todo método simpléctico de un paso, Ψ_h, puede expresarse localmente en términos de una función generatriz. Esta propiedad nos permite mostrar que la ecuación modificada (1.1.5) es Hamiltoniana con $H_h(p, q)$ definido en el mismo dominio que la función generatriz.

Teorema 2.4.5. ***Existencia de un Hamiltoniano modificado global*** *Supongamos que el método simpléctico Ψ_h tiene como función generatriz*

$$S(P, q, h) = hS_1(P, q) + h^2 S_2(P, q) + h^3 S_3(P, q) + \cdots \tag{2.4.18}$$

con S_j regulares y definidas en un conjunto abierto D. Se aplica el método a un sistema Hamiltoniano con un Hamiltoniano suficientemente regular, $H : \mathbb{R}^{2d} \to \mathbb{R}$. Entonces la ecuación diferencial modificada es Hamiltoniana con

$$H_h(p, q) = H(p, q) + hH_2(p, q) + h^2 H_3(p, q) + \cdots, \tag{2.4.19}$$

donde las funciones $H_j(p, q)$ están definidas en todo D.

Demostración. La solución exacta de (P, Q)=$(\tilde{p}(t), \tilde{q}(t))$ del sistema Hamiltoniano correspondiente a $H_h(p, q)$ está dado por

$$p = P + \frac{\partial \tilde{S}}{\partial q}(P, q, t), \qquad Q = q + \frac{\partial \tilde{S}}{\partial P}(P, q, t),$$

donde $\tilde{S}$ es la solución de la ecuación diferencial de Hamilton-Jacobi

$$\frac{\partial \tilde{S}}{\partial t}(P,q,t) = H_h\left(P, q + \frac{\partial \tilde{S}}{\partial P}(P,q,t)\right), \qquad \tilde{S}(P,q,0) = 0. \tag{2.4.20}$$

Como H_h depende del parámetro h, también lo hará $\tilde{S}$. Nuestro objetivo es determinar las funciones $H_j(p,q)$ tales que la solución $\tilde{S}(P,q,t)$ de (2.4.20) coincida en $t = h$ con (2.4.18).
Expresamos primero $\tilde{S}(P,q,t)$ como una serie

$$\tilde{S}(P,q,t) = t\tilde{S}_1(P,q,h) + t^2\tilde{S}_2(P,q,h) + t^3\tilde{S}_3(P,q,h) + \cdots .$$

Insertándolo en (2.4.20) y comparando potencias de t, podemos obtener las funciones $\tilde{S}_j(P,q,h)$ recursivamente en función de las derivadas de H_h.

$$\begin{aligned} \tilde{S}_1(p,q,h) &= H_h(p,q) \\ 2\tilde{S}_2(p,q,h) &= \left(\frac{\partial H_h}{\partial q}\cdot\frac{\partial \tilde{S}_1}{\partial P}\right)(p,q,h) \\ 3\tilde{S}_3(p,q,h) &= \left(\frac{\partial H_h}{\partial q}\cdot\frac{\partial \tilde{S}_2}{\partial P}\right)(p,q,h) + \frac{1}{2}\left(\frac{\partial^2 H_h}{\partial q^2}\left(\frac{\partial \tilde{S}_1}{\partial P}, \frac{\partial \tilde{S}_1}{\partial P}\right)\right)(p,q,h). \end{aligned} \tag{2.4.21}$$

Escribimos $\tilde{S}_j$ como una serie

$$\tilde{S}_j(p,q,h) = \tilde{S}_{j1}(p,q) + h\tilde{S}_{j2}(p,q) + h^2\tilde{S}_{j3}(p,q) + \cdots ,$$

y lo insertamos junto con el desarrollo (2.4.19) en (2.4.21), y comparamos potencias de h. Esto arroja $\tilde{S}_{1k}(p,q) = H_k(p,q)$ y para $j > 1$ $\tilde{S}_{jk}(p,q)$ es una función de las derivadas de H_l con $l < k$.
Imponiendo $S(p,q,h) = \tilde{S}(p,q,h)$ se llega a que

$$\begin{aligned} S_1(p,q) &= \tilde{S}_{11}(p,q), \\ S_2(p,q) &= \tilde{S}_{12}(p,q) + \tilde{S}_{21}(p,q), etc, \end{aligned}$$

luego,

$$S_j(p,q) = H_j(p,q) + \text{ "función de las derivadas de } H_k(p,q) \text{ con } k < j\text{".}$$

Para una función generatriz $S(P,q,h)$ estas relaciones de recurrencia permiten determinar sucesivamente las $H_j(p,q)$. De las fórmulas explícitas se sigue que H_j está definida en el mismo dominio que las S_j. □
Las funciones $H_{N,h}$, N=2,3..., son truncaciones de la serie H_h en potencias de h. Si esta serie de potencias converge y además $\Psi_{h,H} = \Phi_{h,H_h}$, entonces los puntos calculados se situan exactamente en las soluciones del sistema con Hamiltoniano

H_h.
En general, para problemas no lineales, la serie H_h no converge. Sin embargo, si H se comporta suficientemente bien, se puede mostrar que reteniendo para cada $h > 0$ un número finito de términos $N = N(h)$ de la serie de H_h, se puede obtener un Hamiltoniano $H_{N,h}$ para el cual el flujo correspondiente difiera de $\Psi_{h,H}$ en términos que tienden a cero exponencialmente cuando $h \to 0$.
Esto hace posible una interpretación regresiva de los errores de los resultados numéricos: la solución calculada resulta de calcular exactamente (o muy aproximadamente) un problema Hamiltoniano próximo.

Supongamos que $f(y)$, el lado derecho en (1.1.4), es analítica en la bola $B_{2R}(y_0)$ y se cumple

$$\|f(y)\| \leq M, \text{ para } \|y - y_0\| \leq 2R. \tag{2.4.22}$$

Entonces, se tiene la siguiente acotación para los coeficientes del desarrollo de Taylor (1.1.7) de un método Runge-Kutta [7]

$$\|d_j(y)\| \leq C_0\, C_j^{j-1}, \text{ para } \|y - y_0\| \leq R, \tag{2.4.23}$$

donde C_0 y C_j son constantes que dependen de M y de los coeficientes del método RK, y C_j depende además de R.
Con estas hipótesis introducimos sin demostración un resultado sobre la acotación de las funciones f_j de la ecuación modificada y la proximidad entre la solución numérica y la solución de la ecuación modificada correspondiente (ver [7]).

Teorema 2.4.6. *Sea $f(y)$ analítica en $B_{2R}(y_0)$ y sean los coeficientes d_j de (1.1.7) analíticos en $B_R(y_0)$. Supongamos que se dan las acotaciones (2.4.22) y (2.4.23). Entonces tenemos para los coeficientes de la ecuación diferencial modificada*

$$\|f_j(y)\| \leq B_1\, (B_2\, j)^{j-1}, \text{ para } \|y - y_0\| \leq R/2, \tag{2.4.24}$$

donde B_1 y B_2 son constantes que dependen de M y de los coeficientes del método RK, y B_2 además depende de R.
Además, si $h \leq \frac{h_0}{4}$ con h_0 dependiendo de M, R y los coeficientes del método RK, entonces existe $N = N(h)$ (N igual al entero más grande satisfaciendo $hN \leq h_0$) tal que la diferencia entre la solución numérica $y_1 = \Psi_{h,f}(y_0)$ y la solución exacta $\Phi_{t,f_{N,h}}(y_0)$ de la ecuación modificada truncada

$$\dot{\tilde{y}} = f_{N,h}(\tilde{y}), \qquad f_{N,h}(\tilde{y}) = f(\tilde{y}) + hf_2(\tilde{y}) + \cdots + h^{N-1} f_N(\tilde{y}),$$

con valor inicial $\tilde{y}(0) = y_0$, satisface

$$\|\Psi_{h,f}(y_0) - \Phi_{h,f_{N,h}}(y_0)\| \leq hCMe^{-h_0/h}, \tag{2.4.25}$$

donde C depende solamente del método.

Una primera aplicación del teorema anterior es el estudio de la conservación de la energía en largos periodos de tiempo usando métodos simplécticos para resolver numéricamente sistemas Hamiltonianos. Sabemos ya que la ecuación diferencial modificada también es Hamiltoniana (Teorema 2.4.1). Tras truncar, obtenemos un Hamiltoniano modificado

$$H_{N,h}(y) = H(y) + h^r H_{r+1}(y) + \cdots + h^{N-1} H_N(y), \tag{2.4.26}$$

que asumimos está definido en el mismo conjunto abierto que el Hamiltoniano original H (Teorema 2.4.5). Supongamos que se dan las condiciones para aplicar el Teorema 2.4.6.

Teorema 2.4.7. *Consideremos un sistema Hamiltoniano con $H : D \to \mathbb{R}$ analítico (donde $D \subset \mathbb{R}^{2d}$) y aplicamos un método numérico simpléctico Ψ_h con longitud de paso h. Si la solución numérica permanece en un conjunto compacto $K \subset D$, entonces existe h_0 y $N = N(h)$ tal que*

$$\begin{aligned} H_{N,h}(y_n) &= H_{N,h}(y_0) + O(e^{-h_0/2h}), \\ H(y_n) &= H(y_0) + O(h^r), \end{aligned}$$

sobre intervalos de tiempo exponencialmente largos $nh \leq e^{h_0/2h}$.

Demostración. Sea $\Phi_{t,H_{N,h}}$ el flujo de la ecuación diferencial modificada. Como esta ecuación diferencial es Hamiltoniana con $H_{N,h}$ como en (2.4.26),

$$H_{N,h}(\Phi_{t,H_{N,h}}(y_0)) = H_{N,h}(y_0)$$

se cumple para todo tiempo t. Del Teorema 2.4.6 conocemos la acotación (2.4.25) y usando una constante global de Lipschitz que no depende de h para $H_{N,h}$, tenemos también $H_{N,h}(y_{n+1}) - H_{N,h}(\Phi_{h,H_{N,h}}(y_n)) = O(he^{-h_0/h})$. De la identidad

$$\begin{aligned} H_{N,h}(y_n) - H_{N,h}(y_0) &= \sum_{j=1}^{n} (H_{N,h}(y_j) - H_{N,h}(y_{j-1})) \\ &= \sum_{j=1}^{n} (H_{N,h}(y_j) - H_{N,h}(\Phi_{N,h}(y_{j-1}))) \end{aligned}$$

obtenemos entonces $H_{N,h}(y_n) - H_{N,h}(y_0) = O(nhe^{-h_0/h})$, con lo que la conservación de $H_{N,h}$ es una consecuencia inmediata.
Para probar la segunda afirmación del teorema, despejamos $H(y_n) - H(y_0)$ de (2.4.26),

$$\begin{aligned} H(y_n) - H(y_0) &= H_{N,h}(y_n) - H_{N,h}(y_0) - h^r(H_{r+1}(y_n) - H_{r+1}(y_0)) - \\ &\quad h^{r+1}(H_{r+2}(y_n) - H_{r+2}(y_0)) - \cdots - h^{N-1}(H_N(y_n) - H_N(y_0)). \end{aligned}$$

El resultado se sigue de la primera afirmación del teorema y del hecho de que $H_{r+1}(y) + hH_{r+2}(y) + \cdots + h^{N-r-1}H_N(y)$ está uniformemente acotado en K independientemente de h y de N (debido al Lema 2.4.1, a que $f_j(y) = J^{-1}\nabla H_j(y)$ y usando la acotación (2.4.24)). □

Finalizamos este capítulo recalcando en el contexto de los integradores simplécticos para sistemas Hamiltonianos, que la interpretación regresiva de los errores solo se sostiene si la solución numérica en un tiempo t_n se calcula iterando n veces la misma transformación simpléctica. Si alternativamente, componemos n aplicaciones simplécticas (una de t_0 a t_1, otra distinta de t_1 a t_2, etc.) entonces se pierde esta interpretación. Veámos cómo afecta este hecho al caso de los integradores con paso variable.

Sea $\Psi_{h,H}$ un integrador simpléctico y lo aplicamos a un sistema Hamiltoniano. Dado un entero N, se construye un $H_{N,h}$ tal que $\Psi_{h,H} = \Phi_{h,H_{N,h}} + O(h^{N+1})$. Si h se mantiene constante durante la integración desde $t = 0$ hasta $t = t_n$, con $t = t_n$ en un intervalo acotado,

$$\overbrace{\Psi_{h,H}\Psi_{h,H}\cdots\Psi_{h,H}}^{n} = \Psi_{t_n,H}$$

diferirá en términos exponencialmente pequeños de la composición

$$\overbrace{\Phi_{h,H_{N,h}}\Phi_{h,H_{N,h}}\cdots\Phi_{h,H_{N,h}}}^{n} = \Phi_{t_n,H_{N,h}}$$

Sin embargo, si avanzamos la solución con paso variable,

$$\Psi_{h_n,H}\Psi_{h_{n-1},H}\cdots\Psi_{h_1,H}$$

como aproximación de

$$\Phi_{h_n,H_{N,h_n}}\Phi_{h_{n-1},H_{N,h_{n-1}}}\cdots\Phi_{h_1,H_{N,h_1}},$$

esta última expresión ya no es el flujo hasta t_n de un problema Hamiltoniano, ya que las funciones Hamiltonianas usadas en cada paso son diferentes. Por tanto la interpretación regresiva del error del integrador simpléctico ya no se mantiene en el caso de paso variable.

Capítulo 3

Comparación entre integradores simplécticos y convencionales.

El objetivo de este capítulo es averiguar si verdaderamente, para problemas Hamiltonianos, los métodos simplécticos son ventajosos en la práctica. Para ello se han realizado test numéricos y se han justificado teóricamente los resultados obtenidos, haciendo uso del análisis regresivo de los errores estudiado en el Capítulo 2. El tipo de experimento llevado a cabo consiste en la utilización de un método simpléctico con paso fijo h para la integración de problemas Hamiltonianos (2.1.1) en intervalos de tiempo largos, y su comparación con un integrador convencional y eficiente de paso variable. Concretamente, usamos el método Runge-Kutta-Nyström (RKN) simpléctico propuesto en [4], ya que presenta la ventaja de ser simpléctico y explícito, y como integrador no simpléctico tomamos el par encajado de métodos Runge-Kutta explícitos implementado en la función ode45 de Matlab.

El experimento llevado a cabo permite ilustrar que la integración de un problema Hamiltoniano con un método simpléctico de paso fijo puede llegar a ser más eficiente que un código comercial de paso variable. Es en una integración larga donde las ventajas de la simplecticidad se aprecian mejor. Para intervalos temporales cortos, el error local de la fórmula es de gran importancia, mientras que a medida que el intervalo temporal se alarga, las ventajas que ofrece el método simpléctico son, por un lado, un mejor comportamiento cualitativo de la solución numérica, y por otro, un mecanismo de propagación del error más favorable.

El problema test considerado es el problema de los dos cuerpos, más conocido como ***problema de Kepler***, cuya solución describe el movimiento en un plano de un punto material atraído hacia el origen con una fuerza inversamente propor-

cional al cuadrado de la distancia. El Hamiltoniano que describe este movimiento es

$$H([p_1, p_2, q_1, q_2]) = T(p_1, p_2) + V(q_1, q_2) = \frac{1}{2}(p_1^2 + p_2^2) - \frac{1}{\sqrt{q_1^2 + q_2^2}}, \qquad (3.0.1)$$

que claramente es separable.
Las ecuaciones del movimiento escritas como sistema diferencial de primer orden son

$$\dot{p}_i = \frac{q_i}{(q_1^2 + q_2^2)^{3/2}}, \qquad \dot{q}_i = p_i, \qquad i = 1, 2, \qquad (3.0.2)$$

y en los experimentos numéricos vamos a considerar condiciones iniciales de la forma

$$q_1(0) = 1 - e, \quad q_2(0) = 0, \quad p_1(0) = 0, \quad p_2(0) = \sqrt{(1+e)(1-e)}, \qquad (3.0.3)$$

con $0 < e < 1$ un parámetro real. La solución de (3.0.2)-(3.0.3) es 2π periódica y su proyección en el plano (q_1, q_2) es una elipse con excentricidad e y semieje mayor 1. Además, como el problema es autónomo, la energía total H es una cantidad conservada.

3.1. Experimentos numéricos

Se efectuará la integración con las condiciones iniciales dadas en (3.0.3) y excentricidad $e = 0.5$ hasta $t = 21870 \times 2\pi$, obteniendo aproximaciones a la solución en los tiempos intermedios $3^k \times 10 \times 2\pi$, para $0 \leq k \leq 7$. El mismo problema se ha integrado con la función ode45 de Matlab (método Runge-Kutta de orden 5 implementado con paso variable). Se elegirán para el método simpléctico valores de h de la forma $2\pi/\nu$, con ν un número entero positivo, y para el integrador de paso variable tolerancias de la forma 10^{-m} con valor de m adecuado, para que o bien el costo computacional (medido en evaluaciones de función) requerido por ambos procedimientos, o bien los errores generados, sean comparables.

Para las dos implementaciones consideradas se medirán errores tanto en la solución (con la norma euclídea de $\mathbb{R}^4$) como en la energía, en los tiempos intermedios considerados, comparando el valor de las variables (p, q) y de la energía en dichos tiempos, con los correspondientes valores en la condición inicial dada, aprovechando el carácter 2π-periódico de la solución y que los tiempos de integración que se van a utilizar son múltiplos enteros de 2π. Se realizarán en primer lugar gráficas que muestren la evolución de estos errores con el tiempo de integración t. Se harán también gráficas de eficiencia (error en la solución frente al número de evaluaciones de función del lado derecho del sistema diferencial) en

las integraciones hasta 10, 90, 810 y 7290 periodos. Las longitudes de paso h y las tolerancias concretas consideradas para cada método se especificarán más adelante. Con estos datos se realizarán cuatro gráficas de eficiencia. Análogamente se realizarán dos gráficas de eficiencia para el error en la energía, correspondientes a integraciones realizadas hasta 90 y 7290 periodos.

Empezamos describiendo la familia de métodos que se va a utilizar en los experimentos numéricos.

3.1.1. Métodos Runge-Kutta-Nyström simplécticos

Los sistemas Hamiltonianos de ecuaciones diferenciales con la forma particular

$$\frac{dp}{dt} = f(q), \qquad \frac{dq}{dt} = p, \tag{3.1.1}$$

con f igual al gradiente de cierta función escalar $-V$, se pueden reescribir como sistemas diferenciales de segundo orden

$$\frac{d^2q}{dt^2} = f(q), \tag{3.1.2}$$

y pueden integrarse con un método Runge-Kutta-Nyström, cuyas ecuaciones para avanzar un paso de longitud h son

$$\begin{aligned} Q_i &= q_n + h\gamma_i p_n + h^2 \sum_{j=1}^{s} \alpha_{ij} f(Q_j) \qquad , 1 \leq i \leq s, \\ p_{n+1} &= p_n + h \sum_{i=1}^{s} b_i f(Q_i), \\ q_{n+1} &= q_n + h p_n + h^2 \sum_{i=1}^{s} \beta_i f(Q_i), \end{aligned}$$

donde los coeficientes del método son los del tablero

$$\begin{array}{c|ccc} \gamma_1 & \alpha_{11} & \cdots & \alpha_{1s} \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_s & \alpha_{s1} & \cdots & \alpha_{ss} \\ \hline & b_1 & \cdots & b_s \\ \hline & \beta_1 & \cdots & \beta_s \end{array} .$$

Tabla 3.1: Tablero de Butcher de un método Runge-Kutta-Nyström.

Enunciamos sin demostración [13] un resultado que establece bajo qué condiciones sobre sus coeficientes un método RKN es simpléctico.

Teorema 3.1.1. *Supongamos que los coeficientes del método RKN son como en la Tabla 3.1 y satisfacen las condiciones*

$$\begin{aligned} \beta_i &= b_i(1-\gamma_i), \qquad 1 \le i \le s, \\ b_i(\beta_j - \alpha_{ij}) &= b_j(\beta_i - \alpha_{ji}), \qquad 1 \le i, j \le s. \end{aligned} \tag{3.1.3}$$

Entonces el método es simpléctico cuando se aplica a problemas Hamiltonianos con función Hamiltoniana de la forma

$$H(p,q,t) = \frac{1}{2} p^T M^{-1} p + V(q,t), \tag{3.1.4}$$

donde M es normalmente una matriz diagonal, y sus elementos diagonales representan las masas del sistema.

Los métodos RKN que son a la vez explícitos y simplécticos, tienen $\alpha_{ij} = 0$ para $i \le j$. Esto, junto con la segunda condición de (3.1.3) implica que para $i > j$

$$b_i(\beta_j - \alpha_{ij}) = b_j \beta_i.$$

Utilizando ahora la primera condición de (3.1.3) se tiene

$$b_i[b_j(1-\gamma_j) - \alpha_{ij}] = b_j b_i (1-\gamma_i),$$

de donde se deduce que si $b_i \neq 0$

$$\alpha_{ij} = b_j(\gamma_i - \gamma_j), \quad i > j.$$

El tablero de Butcher de un método RKN explícito y simpléctico tiene la forma

γ_1	0	0	$\cdots$	0
γ_2	$b_1(\gamma_2 - \gamma_1)$	0	$\cdots$	0
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
γ_s	$b_1(\gamma_s - \gamma_1)$	$b_2(\gamma_s - \gamma_2)$	$\ldots$	0
	b_1	b_2	$\ldots$	b_s
	$b_1(1-\gamma_1)$	$b_2(1-\gamma_2)$	$\ldots$	$b_s(1-\gamma_s)$

Tabla 3.2: Tablero de un método RKN simpléctico y explícito.

Implementación de un método RKN simpléctico y explícito

A la vista de la Tabla 3.2, las ecuaciones del método se pueden escribir como

$$\begin{aligned} Q_i &= q_n + h\gamma_i p_n + h^2 \sum_{j=1}^{i-1} b_j(\gamma_i - \gamma_j) f(Q_j), \quad 1 \le i \le s, \\ p_{n+1} &= p_n + h \sum_{i=1}^{s} b_i f(Q_i), \\ q_{n+1} &= q_n + h p_n + h^2 \sum_{i=1}^{s} b_i(1-\gamma_i) f(Q_i). \end{aligned}$$

Se puede ver fácilmente que para $1 \le i \le s-1$,

$$Q_{i+1} - Q_i = h(\gamma_{i+1} - \gamma_i)p_n + h^2(\gamma_{i+1} - \gamma_i) \sum_{j=1}^{i} b_j f(Q_j) = h(\gamma_{i+1} - \gamma_i)P_i,$$

donde

$$P_i = p_n + h \sum_{j=1}^{i} b_j f(Q_j), \quad 1 \le i \le s-1.$$

Además, la definición dada para P_i se puede extender también a $i = s$ y se tiene

$$\begin{aligned} P_{i+1} - P_i &= h b_{i+1} f(Q_{i+1}), \quad 1 \le i \le s-1, \\ q_{n+1} - Q_s &= h(1-\gamma_s)P_s. \end{aligned}$$

El método se implementa entonces, para avanzar un paso de (p_n, q_n) a (p_{n+1}, q_{n+1}), como

$Q_0 = q_n,$
$P_0 = p_n,$

Para $\quad i = 1, ..., s$

$$\begin{aligned} Q_i &= Q_{i-1} + h(\gamma_{i+1} - \gamma_i)P_{i-1}, \\ P_i &= P_{i-1} + h b_i f(Q_i), \end{aligned}$$

$q_{n+1} = Q_s + h(1-\gamma_s)P_s,$
$p_{n+1} = P_s.$

Notemos que si $\gamma_1 = 0$ y $\gamma_s = 1$, el esquema RKN simpléctico explícito tiene la propiedad FSAL (First Same As Last), es decir, los pesos $b_i(1-\gamma_i)$ de la Tabla

3.2 coinciden con la última fila de la matriz $\mathcal{A}$ del tablero, y se puede ahorrar una evaluación de función por paso.
En los experimentos numéricos se ha utilizado el método RKN simpléctico y explícito construido en [4] cuyos coeficientes se recogen en la Tabla 3.3.

$$\begin{array}{ll} \gamma_1 = 0, & b_1 = 0.0617588581356263250, \\ \gamma_2 = 0.2051776615422863869, & b_2 = 0.3389780265536433551, \\ \gamma_3 = 0.6081989431465009739, & b_3 = 0.6147913071755775662, \\ \gamma_4 = 0.4872780668075869657, & b_4 = 0.1405480146593733802, \\ \gamma_5 = 1, & b_5 = 0.1250198227945261338. \end{array}$$

Tabla 3.3: Coeficientes del método RKN simpléctico y explícito propuesto en [4].

En general, los integradores simplécticos requieren, para la misma precisión, más trabajo que los no simplécticos dado que, al imponer las condiciones (3.1.3), para el mismo número de etapas se dispone de menos parámetros libres para satisfacer las condiciones de orden.

3.1.2. Resultados numéricos obtenidos

Para las gráficas del error en la solución y en la energía frente al tiempo de integración se utilizó ode45 con tolerancia 10^{-11} (tanto para el error absoluto como para el error relativo) y el método RKN con un paso $h = 2\pi/512$. Para las gráficas de eficiencia en las que se ha representado el error frente al número de evaluaciones del lado derecho de (3.0.2) se emplearon diferentes valores de la tolerancia y del tamaño de paso h. La elección de tales valores se hizo para obtener errores finales próximos con ambos integradores, siempre inferiores a 0.1, por lo que dichos valores fueron haciéndose más restrictivos a medida que el tiempo de integración era más largo. Se indican en la Tabla 3.4 las elecciones de TOL y h hechas para cada tiempo final de integración.

t_F	TOL	h
$10\times 2\pi$	$10^{-6}, \ldots, 10^{-11}$	$2\pi/64, \ldots, 2\pi/1024$
$90 \times 2\pi$	$10^{-7}, \ldots, 10^{-12}$	$2\pi/64, \ldots, 2\pi/1024$
$810\times 2\pi$	$10^{-8}, \ldots, 10^{-13}$	$2\pi/128, \ldots, 2\pi/1024$
$7290\times 2\pi$	$10^{-10}, \ldots, 10^{-14}$	$2\pi/256, \ldots, 2\pi/2048$

Tabla 3.4: Valores de TOL y h empleados en las gráficas de eficiencia.

La línea que representa el crecimiento del error con el tiempo observado en la Figura 3.1 presenta una pendiente aproximadamente igual a 2 para ode45 y muy próxima a 1 para RKN.

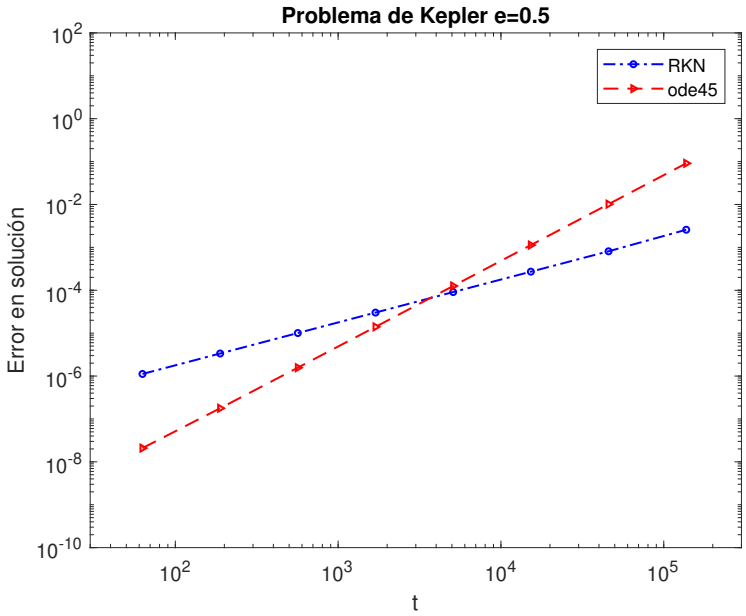


Figura 3.1: Error en la solución como función de t

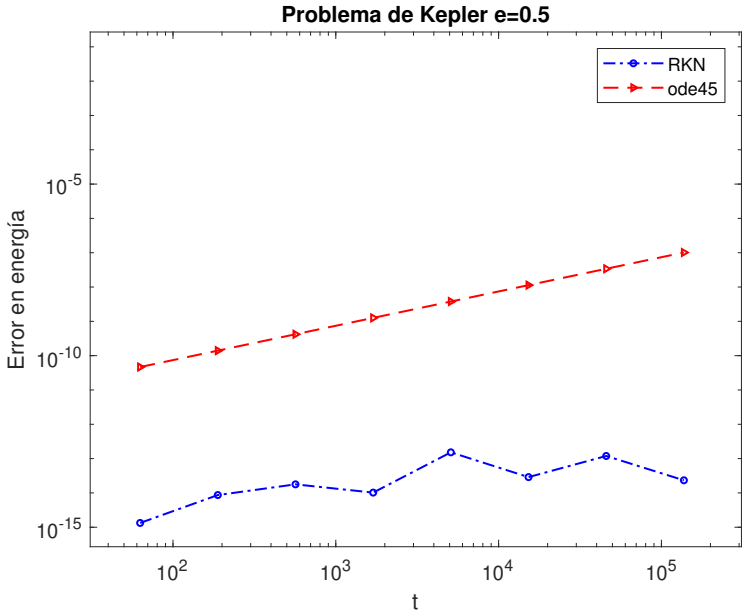


Figura 3.2: Error en la energía como función de t

Veremos que esto es lo esperado en base al análisis del error que se afectuará

en la Sección 3.2 (Teoremas 3.2.1 y 3.2.2).
En la Figura 3.2 se ilustra el error en la energía

$$|H(p_n, q_n) - H(p(t_n), q(t_n))| = |H(p_n, q_n) - H(p(0), q(0))|$$

frente al tiempo de integración. Se puede observar que aunque la energía no se conserva exactamente, el error es mucho más pequeño que el correspondiente al error en la solución $\|(p_n, q_n) - p(0), q(0)\|$, llegando a alcanzar los errores valores de 10^{-15} en el caso simpléctico, muy próximos a los errores de redondeo. En el caso del método ode45, la gráfica muestra una recta de pendiente aproximadamente 1.

A continuación procedemos a ilustrar el error en la solución tras 10, 90, 810 y 7290 periodos frente al costo computacional, medido como el número de evaluaciones del lado derecho del sistema de ecuaciones diferenciales. El resultado se puede ver en las Figuras 3.3, 3.4, 3.5 y 3.6, respectivamente.

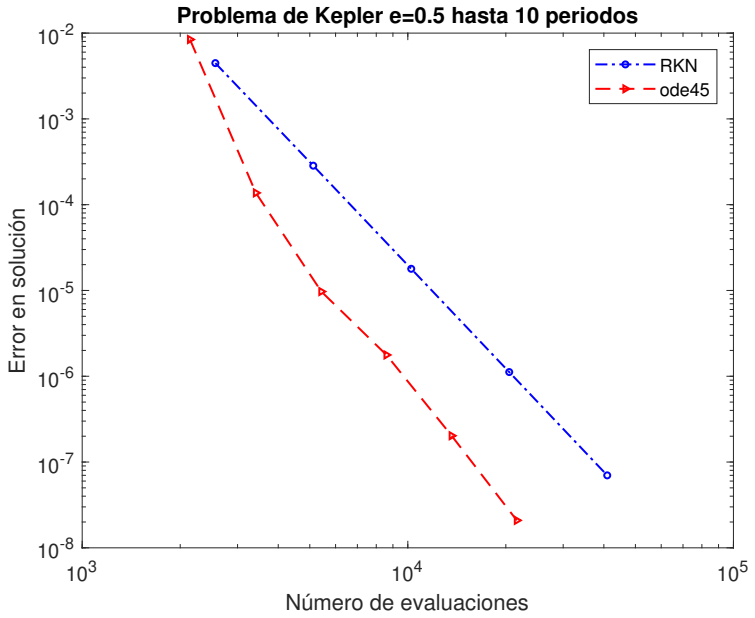


Figura 3.3: Error en la solución frente al número de evaluaciones tras 10 periodos

Para el método RKN, el error es proporcional a h^4, y puesto que h a su vez es inversamente proporcional al número de pasos N, y a su vez N es proporcional al número de evaluaciones de función (4 evaluaciones por paso), teóricamente la pendiente del error, en escala doblemente logarítmica, es -4. Por su parte, el método ode45 es de orden 5, y por el mismo razonamiento esperamos encontrar

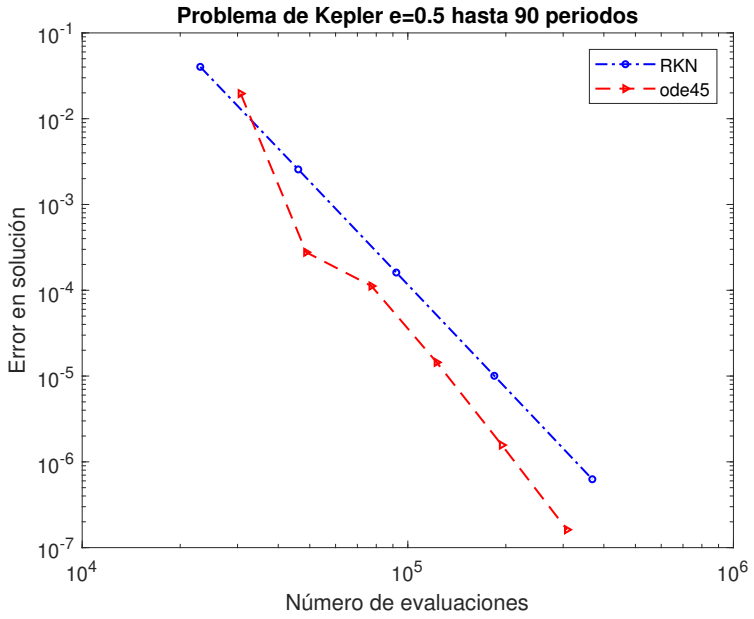


Figura 3.4: Error en la solución frente al número de evaluaciones tras 90 periodos

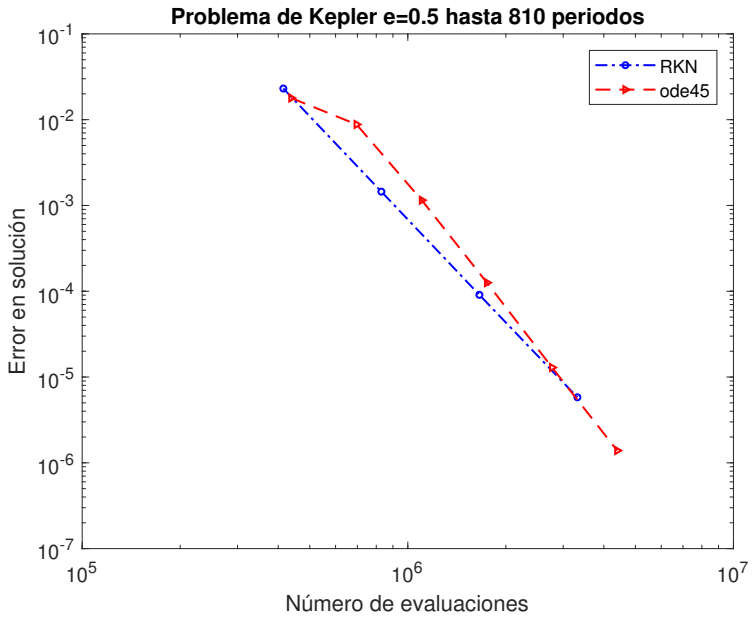


Figura 3.5: Error frente al número de evaluaciones tras 810 periodos

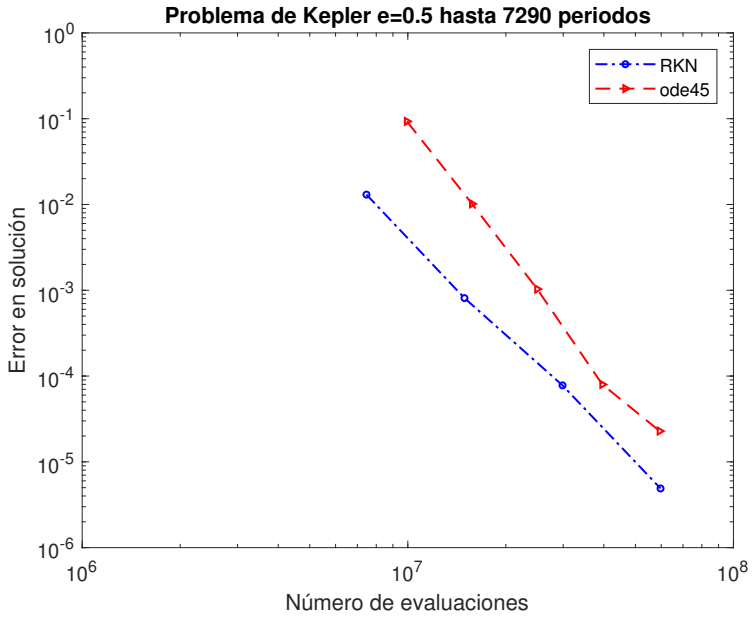


Figura 3.6: Error frente al número de evaluaciones tras 7290 periodos

una pendiente de la línea que representa el error que genera igual a -5 en escala doblemente logarítmica.

Procedemos ahora a comparar la eficiencia entre los dos métodos en vista de las gráficas obtenidas. Para un tiempo final de integración $t_F = 10 \times 2\pi$ (Figura 3.3), un error dado requiere un mayor costo computacional para el método RKN simpléctico, puesto que la línea azul se mantiene siempre a la derecha de la línea roja (para el mismo error necesita más evaluaciones). A medida que aumenta el tiempo final de integración, las líneas que representan a los dos métodos se van acercando hasta que, pasados 7290 periodos (Figura 3.6), se ve que claramente el método RKN simpléctico es más eficiente que el integrador ode45 para todas las longitudes de paso y tolerancias utilizadas. Por ejemplo, para obtener un error de tamaño 10^{-3}, ode45 necesita casi el doble de evaluaciones de función que el método RKN. Esto parece indicar que el mecanismo de propagación de los errores da ventaja al algoritmo simpléctico sobre el no simpléctico. El Teorema 3.2.2 confirma el resultado obtenido.

Por tanto se concluye que para este problema, el método simpléctico mejora al método no simpléctico cuando el tiempo final de integración supera los 810 periodos, y la desventaja que suponía tener que utilizar paso fijo, es compensada por un buen mecanismo de propagación de los errores. Respecto a las pendientes

del error observadas en las gráficas de eficiencia, se confirma lo predicho teóricamente, obteniendo pendientes muy próximas a -4 y -5 para el método RKN y ode45 respectivamente.

Por último, se muestra en las Figuras 3.7 y 3.8 el error en la energía tras 90 y 7290 periodos de integración, frente al número de evaluaciones del lado derecho del sistema diferencial. Para realizar el estudio de las pendientes de las líneas

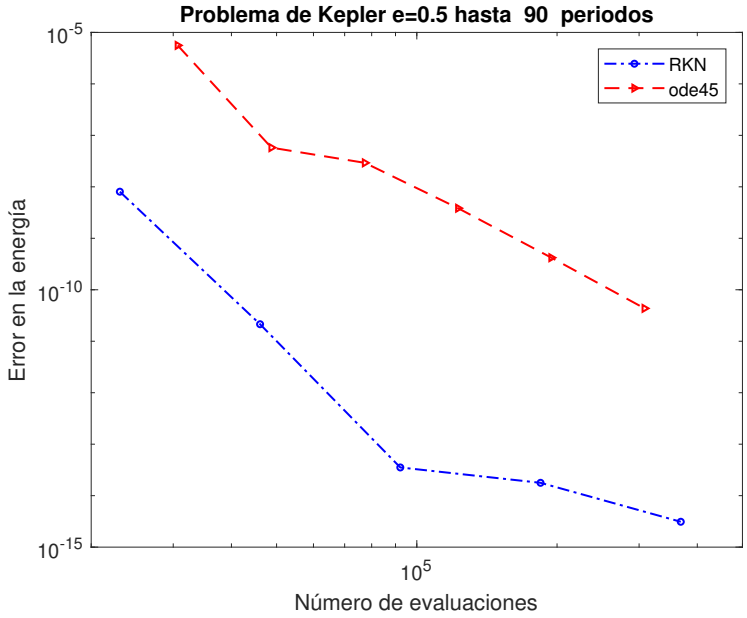


Figura 3.7: Error en la energía frente al número de evaluaciones tras 90 periodos

representadas, nos centramos en el rango de valores de los errores en la energía en los que los errores de redondeo aún no son tan significativos como para dominar el comportamiento de la gráfica. Si consideramos que la precisión de la máquina es de 1e-16, en una integración hasta tiempo final de 90 periodos, solo podemos esperar que el error cometido por el método numérico sea representativo para errores superiores a 1e-14. Para ese rango de valores, se observan pendientes de -8 y -5 para los métodos RKN y ode45, respectivamente. En el caso de la integración hasta un tiempo final de 7290 periodos, nos fijamos en los errores superiores a 1e-12, donde la línea correspondiente a ode45 muestra una pendiente de -5 y los errores de RKN se mantienen casi constantes y del tamaño de los errores de redondeo. Análogamente a lo que sucedía con las gráficas de eficiencia para los errores en la solución, encontraremos respuesta al comportamiento descrito en los Teoremas 3.2.2 y 3.2.1.

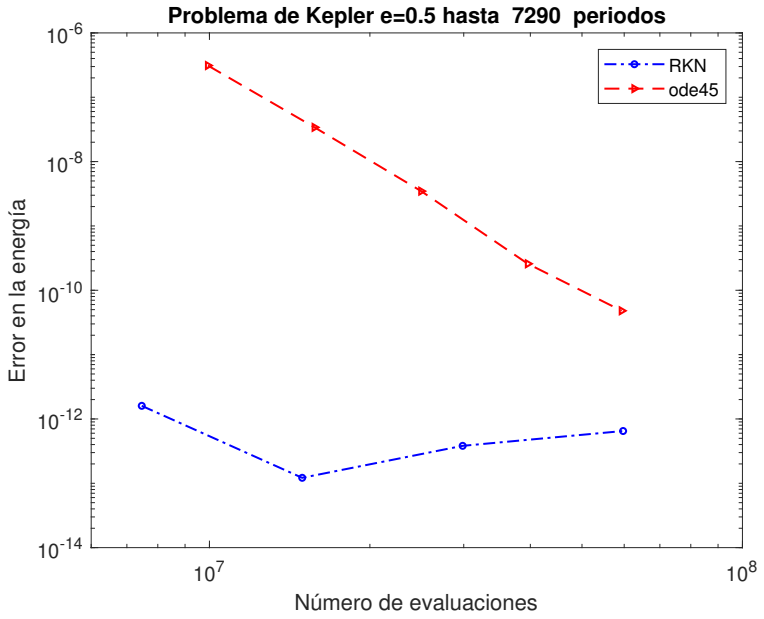


Figura 3.8: Error en la energía frente al número de evaluaciones tras 7290 periodos

3.2. Análisis del error

Reescribiendo el problema de Kepler (3.0.2) como

$$\dot{Y} = F(Y)$$

donde $Y = [p^1, p^2, q^1, q^2]$ y $F = [f^T, p^T]$ siendo $f = f(q)$ la fuerza. Usaremos la notación $G = G(Y)$ para referirnos al gradiente ∇H del Hamiltoniano H respecto a Y. Notemos que F y G son ortogonales en cada punto Y.
Consideramos la región del espacio de fases cubierta por movimientos elípticos, es decir, la región donde la energía H es menor que 0, donde no es posible un escape a infinito. Todas las soluciones son periódicas con periodo

$$T = T(H) = 2\pi/\sqrt{(2|H|)^3}.$$

Fijamos una condición inicial Y_0. Llamamos $F_0 = F(Y_0)$, $G_0 = G(Y_0)$. Denotamos por Φ a la transformación que avanza el flujo un periodo, φ_{T_0}, $T_0 = T(H(Y_0))$.

Lema 3.2.1. *La diferencial Φ_0' es una modificación de rango uno de la identidad dada por*

$$\Phi_0' = I + W_0 G_0^T, \tag{3.2.1}$$

con $W_0 = T'(H(Y_0))F_0$ un vector no nulo en $\mathbb{R}^4$ tangente en Y_0 a la solución del problema de Kepler.

Demostración. Sea

$$\Phi(Y) = \varphi_\tau(\varphi_{T(H(Y))}(Y)) = \varphi_\tau(Y),$$

donde $\tau = T_0 - T(H(Y))$ es una función de Y. Entonces

$$\begin{aligned}\Phi'(Y) &= \varphi'_\tau(Y) + \left(\frac{d}{d\tau}\varphi_\tau(Y)\right)(\nabla\tau)^T \\ &= \varphi'_\tau(Y) - \left(\frac{d}{d\tau}\varphi_\tau(Y)\right)T'(H(Y))G(Y)^T\end{aligned}$$

y como φ_τ satisface, como función de t, $\dot{Y} = F(Y)$,

$$\Phi'(Y) = \varphi'_\tau(Y) - F(\varphi_\tau(Y))T'(H(Y))G(Y)^T.$$

Ahora evaluando $Y = Y_0$ conduce a $\tau = 0$ y por lo tanto φ_τ es igual a la aplicación identidad, luego $\varphi'_\tau\ (Y_0) = I$. $\square$
Calculamos la potencia N-ésima de Φ'_0,

$$(\Phi'_0)^N = \sum_{k=0}^{N}\binom{N}{k} I^{N-k}(W_0G_0^T)^k.$$

Notemos que $(W_0G_0^T)^2 = W_0G_0^TW_0G_0^T = W_0(G_0^TW_0)G_0^T = 0$ puesto que G_0 y W_0 son ortogonales. Por tanto, si $k \geq 2$, todas las potencias $(W_0G_0^T)^k$ se anulan. La potencia N-ésima de Φ'_0 resulta ser entonces

$$(\Phi'_0)^N = I + NW_0G_0^T. \tag{3.2.2}$$

Consideremos ahora un método de un paso ψ_h de orden r, es decir,

$$\psi_h^n(Y_0) - \varphi_h^n(Y_0) = O(h^r) \text{ cuando } h \to 0.$$

Suponemos además que las matrices jacobianas $(\psi_h^n)'(Y)$ también convergen con orden r a la matriz jacobiana del flujo,

$$(\psi_h^n)'(Y) - (\varphi_h^n)'(Y) = O(h^r), h \to 0.$$

Denotaremos por Ψ_h la aplicación que avanza la solución numérica T_0 unidades de tiempo, y se quiere estudiar la diferencia E_N entre la solución numérica $\Psi_h^N(Y_0)$ y la teórica $\Phi^N(Y_0) = Y_0$ tras N periodos.
Escribimos

$$\begin{aligned}E_N &= \Psi_h^N(Y_0) - Y_0 = \Psi_h(\Psi_h^{N-1}(Y_0)) - \Psi_h(Y_0) + E_1 \\ &= \Psi'_hE_{N-1} + O(\|E_{N-1}\|^2) + E_1.\end{aligned}$$

Teniendo en cuenta que $E_{N-1} = O(h^r)$ y que $\Psi_h - \Phi_h = O(h^r)$, entonces

$$E_N = \Psi'_h E_{N-1} + E_1 + O(h^{2r}) = \Phi'_0 E_{N-1} + E_1 + O(h^{2r}).$$

Por inducción, se llega a

$$E_N = [I + \Phi'_0 + \cdots + {\Phi'_0}^{N-1}]E_1 + O(h^{2r}). \tag{3.2.3}$$

Aplicando 3.2.2 a lo anterior obtenemos

Teorema 3.2.1. *Bajo las hipótesis anteriores, el error tras N periodos viene dado por*

$$E_N = NE_1 + \frac{1}{2}(N^2 - N)(G_0^T E_1)W_0 + O(h^{2r}). \tag{3.2.4}$$

El error en la energía tras N periodos satisface,

$$E_N^{energia} = H(\Psi_h^N) - H(Y_0) = NE_1^{energia} + O(h^{2r}). \tag{3.2.5}$$

Es decir, excepto por los términos $O(h^{2r})$, el error E_N tras N periodos, crece cuadráticamente con N. El término dominante del crecimiento N^2, es en la dirección tangente a la solución en Y_0 (error de fase). El error en la energía tras N periodos es, salvo por los términos $O(h^{2r})$, N veces el error en la energía tras el primer periodo.

Demostración. Insertando la expresión (3.2.2) en (3.2.3),

$$E_N = \left[I + (I + W_0 G_0^T) + (I + 2W_0 G_0^T) + \cdots (I + (N-1)W_0 G_0^T)\right] E_1 + O(h^{2r}) \tag{3.2.6}$$

y teniendo en cuenta que la suma de los $N-1$ primeros números naturales es $N(N-1)/2$, se llega al resultando

$$E_N = NE_1 + \frac{N(N-1)}{2} W_0 G_0^T E_1 + O(h^{2r}).$$

En cuanto al error en la energía,

$$\begin{aligned} E_N^{energia} &= H(\Psi_h^N) - H(Y_0) = H(E_N + Y_0) - H(Y_0) = G_0^T E_N + O(\|E_N\|^2) \\ &= NG_0^T E_1 + G_0^T \frac{N(N-1)}{2}(G_0^T E_1)W_0 + O(h^{2r}), \end{aligned}$$

y debido a la ortogonalidad de G_0 y W_0,

$$E_N^{energia} = NG_0^T E_1 + O(h^{2r}). \quad \square$$

En el caso simpléctico es posible mejorar el comportamiento del error en el Teorema 3.2.1 para métodos de un paso generales.

Teorema 3.2.2. *Para un método simpléctico de paso fijo,*

$$E_N = NE_1 + O(h^{2r}). \tag{3.2.7}$$

El error en la energía satisface

$$E_N^{energia} = H(\Psi_h^N(Y_0)) - H(Y_0) = O(h^{2r}).$$

Demostración. En el caso simpléctico, dado un entero q, ya vimos en el Teorema 2.4.1 que es posible construir un Hamiltoniano modificado $H_h = H + O(h^r)$ tal que Ψ_h es consistente de orden q con el problema Hamiltoniano asociado a H_h, es decir, $\Psi_h - \varphi_{h,H_h} = O(h^{q+1})$ donde φ_{h,H_h} es el flujo del problema con Hamiltoniano H_h. Aquí tomamos $q = 2r$. Los puntos calculados Y_n distan solo en $O(h^{2r})$ de la solución exacta del problema modificado con condición inicial Y_0. Por tanto

$$H_h(\Psi_h(Y_0)) - H_h(\varphi_{T_0,H_h}(Y_0)) = H_h(\Psi_h(Y_0)) - H_h(Y_0) = O(h^{2r}), \tag{3.2.8}$$

ya que H_h es una cantidad conservada para el flujo φ_{T_0,H_h}. Al mismo tiempo, si hacemos el desarrollo en serie de Taylor de (3.2.8),

$$H_h(\Psi_h(Y_0)) - H_h(\varphi_{T_0,H_h}(Y_0)) = (G_0^h)^T E_1 + O(\|E_1\|^2) = (G_0^h)^T E_1 + O(h^{2r}),$$

donde G_0^h es el gradiente de H_h en Y_0. Comparando las dos ecuaciones anteriores, se llega a que $(G_0^h)^T E_1 = O(h^{2r})$, es decir, E_1 y el gradiente G_0^h son casi ortogonales. Por último,

$$|G_0^T E_1| = |(G_0 - G_0^h)^T E_1 + G_0^{h^T} E_1| \leq \|G_0 - G_0^h\|\|E_1\| + |G_0^{h^T} E_1| = O(h^{2r}),$$

donde se ha usado que las derivadas de H_h aproximan a las derivadas de H con el mismo orden, $O(h^r)$, con el que H_h aproxima a H. □

Bibliografía

[1] V.I. Arnold, *Mecánica clásica: métodos matemáticos,* Paraninfo. (1983)

[2] J. C. Butcher, *Numerical Methods for Ordinary Differential equations*, John Wiley & Sons Inc. (2016).

[3] M.P. Calvo, A. Murua & J.M. Sanz-Serna, *Modified equations for ODEs,* Contemporary Mathematics 172 (1994), 63-74.

[4] M.P. Calvo & J.M. Sanz-Serna, *The development of variable-step symplectic integrators with application to the two-body problem,* SIAM J. Sci. Comput. 14 (1993), 936-952.

[5] M. P. Calvo & J. M. Sanz-Serna, *Canonical B-series,* Numer. Math. 67 (1994), 161-175.

[6] E. Hairer, *Backward analysis of numerical integrators and symplectic methods,* Annals of Numerical Mathematics 1 (1994), 107-132.

[7] E. Hairer, Ch. Lubich & G. Wanner, *Geometric Numerical Integration. Structure-Preserving Algorithms for Ordinary Differential Equations.* Springer Series in Computational Mathematics 31. Springer-Verlag, Berlin (2002).

[8] E. Hairer, S.P. Norsett & G. Wanner, *Solving Ordinary Differential Equations I. Nonstiff Problems. Second edition.* Springer Series in Computational Mathematics 8. Springer-Verlag, Berlin (1993).

[9] E. Hairer & G. Wanner, *On the Butcher Group and General Multi-Value Methods*, Computing 13 (1974), 1-15.

[10] B. Leimkuhler & S. Reich, *Simulating Hamiltonian Dynamics*, Cambridge Monographs on Applied and Computational Mathematics (2004), 50-55.

[11] J.M. Sanz-Serna, *Symplectic Integrators for Hamiltonian Problems: an Overview,* Acta Numérica 1 (1991), 243-286.

[12] J.M. Sanz-Serna, L. Abia, *Order conditions for canonical Runge-Kutta schemes*, SIAM J.Numer. Anal. 28, 1081-1096.

[13] J.M. Sanz-Serna & M.P. Calvo, *Numerical Hamiltonian Problems,* Dover Publications Inc., Mineola, New York (2018).

Apéndice A

Programas de Matlab

En este apéndice se incluyen los códigos de las funciones de Matlab utilizadas para generar las gráficas de las Figuras 1.1 y 1.2, y los experimentos numéricos del Capítulo 3.

A.1. Gráficas del Capítulo 1

La función **ecmod.m** genera una gráfica que representa la evolución con el tiempo $y(t)$ de la solución exacta del problema de valores iniciales (1.1.9), la solución numérica generada con el método de Euler explícito y la solución *exacta* de la ecuación modificada truncada tras 2 y 3 términos, respectivamente . La longitud de paso utilizada en la integración numérica está definida en el programa.

```
function [] = ecmod()
global h
tF=0.99;
Nsteps=20;
h=tF/Nsteps;
y0=1;

y1=zeros(Nsteps+1,1);
y1(1)=1;
t1=[0];
options=odeset('RelTol',1e-14,'AbsTol',1e-14);

for i=1:Nsteps
    y1(i+1,1)=y1(i,1)+h*odefun1(t1(i),y1(i,1));
    t1=[t1;t1(end)+h];
end

t2=linspace(0,0.99,100);
yex=1./(1-t2);
sol2=ode45(@odefun2,[0 0.99],y0,options);
sol3=ode45(@odefun3,[0 0.99],y0,options);
```

```
y2=deval(sol2,t2);
y3=deval(sol3,t2);

figure(1)
clf
plot(t2,yex,'-b','LineWidth',1,'MarkerSize',3)
hold on
plot(t1,y1,'or','LineWidth',1,'MarkerSize',3)
hold on
plot(t2,y2,'-y','LineWidth',1,'MarkerSize',3)
hold on
plot(t2,y3,'-g','LineWidth',1,'MarkerSize',3)
hold on
legend('Sol exacta','Euler h=0.02','Sol exacta ec mod + O(h^{2})','
    Sol exacta ec mod + O(h^{3})');
axis([0 1.1 0 20])
xlabel('t');
ylabel('y(t)');
hold off
end

function f =odefun1(t,y)
f=y^2;
end

function f = odefun2(t,y)
global h
f =y^2-h*y^3;
end

function f = odefun3(t,y)
global h
f =y^2*(1-h*y*(1-1.5*h*y));
end
```

La función **ecmod2** genera las gráficas de la Figura 1.2, con las soluciones *exactas* y numéricas del problema de Lotka-Volterra, utilizando los métodos de Euler explícito y Euler simpléctico, ambos con longitud de paso $h = 0{,}1$, y las soluciones *exactas* de las correspondientes ecuaciones modificadas truncadas tras dos términos.

```
function [] = ecmod2()
global h
h=0.1;
options=odeset('RelTol',1e-14,'AbsTol',1e-14);
t=linspace(0,12,1000);
teuler=linspace(0,14.5,1500);

%solucion exacta problema original con condicion inicial (1,3)
solvolterra=ode45(@odefunvolterra, [0 14.5], [1 3],...
```

```
    options);
yex=deval(solvolterra,t);

%solucion exacta problema original con condicion inicial (1,6)
solvolterra2=ode45(@odefunvolterra,[0 12],[1 6], options);
yex2=deval(solvolterra2,t);

%solucion exacta ecuacion modificada para Euler explicito,
solvolterramod=ode45(@odefunvolterramod, [0 14.5], [1 3], options);
yexmod=deval(solvolterramod,teuler);

%solucion exacta ecuacion modificada para Euler simplectico, y
    condicion inicial (1,3)
solvolterramodsimpl=ode45(@odefunvolterramodsimpl, [0 12],...
    [1 3], options);
yexmodsimpl=deval(solvolterramodsimpl,t);

%solucion exacta ecuacion modificada para Euler simplectico, y
    condicion inicial (1,6)
solvolterramodsimpl2=ode45(@odefunvolterramodsimpl, [0 12], [1 6],
    options);
yexmodsimpl2=deval(solvolterramodsimpl2,t);


Npasos=12/h;
Npasoseuler=floor(14.5/h);

yE=zeros(Npasoseuler+1,2); %solucion numerica Euler
yEsim=zeros(Npasos+1,2); %solucion numerica Euler simplectico CI
    (1,3)
yEsim2=zeros(Npasos+1,2); %solucion numerica Euler simplectico CI
    (1,6)

%condiciones iniciales
yE(1,:)=[1,3];
yEsim(1,:)=[1,3];
yEsim2(1,:)=[1,6];

%Implementacion de Euler y Euler simplectico
for i=1:Npasos
yEsim(i+1,:)=Eulersimpl(t,yEsim(i,:))';
yEsim2(i+1,:)=Eulersimpl(t,yEsim2(i,:))';
end

for i=1:Npasoseuler
yE(i+1,:)=yE(i,:)+h*odefunvolterra(t,yE(i,:))';
end

figure(1)
clf
plot(yex(2,:),yex(1,:),'-b','LineWidth',1,'MarkerSize',3)
```

```
hold on
plot(yE(:,2),yE(:,1),'or', 'LineWidth',1,'MarkerSize',3)
hold on
plot(yexmod(2,:),yexmod(1,:),'-y','LineWidth',1,'MarkerSize',3)
hold on
plot(3,1,'.k','MarkerSize',15)
hold on
legend('Sol exacta','Euler h=0.1','Sol exacta ec mod + O(h^{2})')
xlabel('q')
ylabel(['p'])
axis([0 7.5 0 6]);
hold off

figure(2)
clf
plot(yex(2,:),yex(1,:),'-b','LineWidth',1,'MarkerSize',3)
hold on
plot(yEsim(:,2),yEsim(:,1),'or', 'LineWidth',1,'MarkerSize',3)
hold on
plot(yexmodsimpl(2,:),yexmodsimpl(1,:),'-y','LineWidth',1,'
    MarkerSize',3)
hold on
plot(3,1,'.k','MarkerSize',15)
hold on
plot(yex2(2,:),yex2(1,:),'-b','LineWidth',1,'MarkerSize',3)
hold on
plot(yEsim2(:,2),yEsim2(:,1),'or', 'LineWidth',1,'MarkerSize',3)
hold on
plot(yexmodsimpl2(2,:),yexmodsimpl2(1,:),'-y','LineWidth',1,'
    MarkerSize',3)
hold on
plot(6,1,'.k','MarkerSize',15)
hold on
legend('Sol exacta','Euler simpl ctico h=0.1','Sol exacta ec mod +
     O(h^{2})')
axis([0 7.5 0 6]);
xlabel('q')
ylabel(['p'])
hold off

end

function f1 = odefunvolterra(t,y)
f1=[y(1)*(2-y(2)); y(2)*(y(1)-1)];
end

function f2= Eulersimpl(t,y)
global h
f2=[y(1)/(1-h*(2-y(2)));y(2)+h*y(2)*((y(1))/(1-h*(2-y(2))) -1)];
end
```

```
function f3 = odefunvolterramod(t,y)
global h
f3=[-y(1)*(y(2)-2)-(h/2)*y(1)*(y(2)^2-y(1)*y(2)-3*y(2)+4);y(2)*(y
    (1)-1)-(h/2)*y(2)*(y(1)^2-y(1)*y(2)+1)];
end

function f4= odefunvolterramodsimpl(t,y)
global h
f4=[-y(1)*(y(2)-2)+(h/2)*y(1)*(y(2)^2+y(1)*y(2)-5*y(2)+4);y(2)*(y
    (1)-1)-(h/2)*y(2)*(y(1)^2+y(1)*y(2)-4*y(1)+1)];
end
```

A.2. Test numérico del Capítulo 3.

La función **expnumKEPLER.m** toma como parámetro de entrada *exc*, la excentricidad de la órbita. Devuelve las gráficas del error en la solución numérica y en la energía frente al tiempo, y las gráficas de eficiencia, en las que se comparan las integraciones realizadas con el método RKN simpléctico con coeficientes definidos en la Tabla 3.3 con el integrador de Matlab ode45.
Se utilizan como funciones auxiliares:

- **Kepler.m**: función que evalúa el lado derecho del sistema diferencial $dq^2/dt^2 = f(q)$ y lo guarda en un vector columna f. Se utiliza como función auxiliar de **RKN.m**.
- **Keplercompleta**: función que evalúa el lado derecho del sistema diferencial (3.0.2) del problema de Kepler y lo guarda en un vector columna g.
- **ode45**
- **RKN.m**: función que integra un problema de 2^{o} orden de la forma (3.1.2) con un método RKN simpléctico y devuelve la solución numérica en dos vectores columna P y Q, además del número de llamadas a la función **Kepler.m**.
- **graficerrortiempo.m**: función que calcula la norma del error en la solución y el error en la energía proporcionados por **RKN.m** y **ode.45** y los representa frente al tiempo.
- **graficef.m**: con los datos de las soluciones y del número de evaluaciones del lado derecho del sistema diferencial en la integración del problema de Kepler proporcionados por ambos integradores hasta el tiempo final de integración indicado como argumento de entrada, para unos parámetros de tolerancias y longitudes de paso h establecidos en el programa principal, realiza las gráficas error en la solución y en la energía frente al número de evaluaciones realizadas.

```
function [] = expnumKEPLER(exc)
format long

%Condiciones iniciales para RKN
q0=[1-exc;0];
p0=[0;sqrt((1+exc)/(1-exc))];

%CI para ode45
```

```
y0=[q0;p0];

%Cantidad conservada
Iconserv0=0.5*(p0(1)*p0(1)+p0(2)*p0(2))-1/sqrt(q0(1)*q0(1)+q0(2)*q0
    (2));

%Graficas del error en solucion y en energia
graficerrortiempo(q0,p0,Iconserv0)
%Graficas de eficiencia
TOL10=[10^-6,10^-7,10^-8,10^-9,10^-10,10^-11];
h10=2*pi*[1/64;1/128;1/256;1/512;1/1024];
graficef(10,TOL10,h10,q0,p0)

TOL90=[10^-7,10^-8,10^-9,10^-10,10^-11,10^-12];
h90=2*pi*[1/64;1/128;1/256;1/512;1/1024];
graficef(90,TOL90,h90,q0,p0)

TOL810=[10^-8,10^-9,10^-10,10^-11,10^-12,10^-13];
h810=2*pi*[1/128;1/256;1/512;1/1024];
graficef(810,TOL810,h810,q0,p0)

TOL7290=[10^-10,10^-11,10^-12,10^-13,10^-14];
h7290=2*pi*[1/256;1/512;1/1024;1/2048];
graficef(7290,TOL7290,h7290,q0,p0)
end
```

A.2.1. Programación del método RKN simpléctico

La función **RKN** presenta los siguientes parámetros de entrada:

- t_F: corresponde al tiempo final, que es un número entero de periodos, hasta el que se realiza la integración numérica.
- h: tamaño del paso temporal de integración.
- q_0: vector bidimensional columna con los valores iniciales de las coordenadas generalizadas.
- p_0: vector bidimensional columna con los valores iniciales de los momentos generalizados.

Los coeficientes del método en concreto se definen dentro del propio código como dos vectores columna, b y B.
Los valores de salida son los siguientes:

- Q: vector columna bidimensional con la aproximación numérica de las coordenadas generalizadas.
- P: vector columna bidimensional con la aproximación numérica de los momentos generalizadas.
- *neval*: número de evaluaciones del lado derecho del sistema diferencial $d^2q/dt^2 = f(q)$.

```
function [Q,P,neval] = RKN(tF,h,q0,p0)
gamma1=0;
gamma2= 0.2051776615422863869;
gamma3=0.6081989431465009739;
gamma4=0.4872780668075869657;
gamma5=1;

b
    =[0.0617588581356263250;0.3389780265536433551;0.6147913071755775662;...

-0.1405480146593733802;0.1250198227945261338];
B=[gamma1;gamma2-gamma1;gamma3-gamma2;gamma4-gamma3;gamma5-gamma4];

Q=zeros(2,1);
P=zeros(2,1);
neval=0;
Npasos=floor((tF)/h);

Q=q0;
P=p0;
tt=0;
```

```
for n=1:Npasos
for i=1:5
  Q=Q+h*B(i)*P ;
  P=P+h*b(i)*Kepler(0,Q);
end
tt=tt+h;
neval=neval+4;
end
haux=(tF-tt);
if (haux~=0)
for i=1:5
  Q=Q+haux*B(i)*P ;
  P=P+haux*b(i)*Kepler(0,Q);
end
neval=neval+4;
end
end
```

A.2.2. Problema de Kepler

La función **Kepler.m** tiene como valores de entrada el tiempo t y un vector de coordenadas generalizadas $y(t) = [y_1(t), y_2(t)]$. Como salida devuelve la evaluación del lado derecho del sistema diferencial $d^2q/dt^2 = f(q)$.
La función **Keplercompleta.m** tiene como valores de entrada el tiempo t y un vector con las coordenadas generalizadas y los momentos $y = [y(1), y(2), y(3), y(4)] = [q^1, q^2, p^1, p^2]$. Como salida devuelve la evaluación del lado derecho de (3.0.2).

```
function f = Kepler(t,y)
f =[-y(1)/((y(1)*y(1)+y(2)*y(2))^(3/2));-y(2)/((y(1)*y(1)+y(2)*y(2)
    )^(3/2))];
end
```

```
function g = Keplercompleta(t,y);
g =[y(3);y(4);-y(1)/((y(1)*y(1)+y(2)*y(2))^(3/2));-y(2)/((y(1)*y(1)
    +y(2)*y(2))^(3/2))];
end
```

A.2.3. Gráficas

La función **graficef.m**, descrita anteriormente recibe como argumentos de entrada el número de periodos correspondiente al tiempo final de integración, los valores de h y TOL que se van a utlizar en la integración con **RKN.m** y **ode45** respectivamente, las condiciones iniciales que van a necesitar en la integración con ambos métodos numéricos y utiliza como funciones auxiliares **RKN.m** y **ode45**. Como salida se obtienen las gráficas de eficiencia.

```
function []=graficef(nperiodos,TOL,h,q0,p0)
```

```
l=length(TOL);
m=length(h);
error=zeros(m,1);
errorode=zeros(l,1);
errorenergia=zeros(m,1);
errorenergiaode=zeros(l,1);
y0=[q0;p0];

tF=2*pi*nperiodos;
neval=zeros(m,1);
nevalode=zeros(l,1);

Iconserv0=0.5*(p0(1)*p0(1)+p0(2)*p0(2))-1/sqrt(q0(1)*q0(1)+q0(2)*q0
    (2));
for j=1:l   %para cada tolerancia
options=odeset('RelTol',TOL(j),'AbsTol',TOL(j));
sol=ode45(@Keplercompleta,[0 tF],y0,options);
aux=sol.stats;
nevalode(j,1)=aux.nfevals;
yode = deval(sol,tF);
errorode(j,1)=norm(yode-y0);
errorenergiaode(j,1)= abs(Iconserv0-0.5*(yode(3)*yode(3)+yode(4)*
    yode(4))+1/sqrt(yode(1)*yode(1)+yode(2)*yode(2)));

end

for j=1:m %para cada longitud de paso
[q,p,neval(j,1)]=RKN(tF,h(j),q0,p0);
error(j,1)=norm([q;p]-[q0;p0]);
errorenergia(j,1)= abs(Iconserv0-0.5*(p(1)*p(1)+p(2)*p(2))+1/sqrt(q
    (1)*q(1)+q(2)*q(2)));

end

figure(3)
clf
loglog(neval(:,1),error(:,1),'-.ob','LineWidth',1,'MarkerSize',3)
hold on
loglog(nevalode(:,1),errorode(:,1),'-->r','LineWidth',1,'MarkerSize
    ',3)
hold on
xlabel('N mero de evaluaciones')
ylabel(['Error en soluci n'])
title(['Problema de Kepler e=0.5 hasta ' num2str(nperiodos) '
    periodos'])
legend('RKN', 'ode45')
hold off

figure(4)
clf
```

```
loglog(neval(:,1),errorenergia(:,1),'-.ob','LineWidth',1,'
    MarkerSize',3)
hold on
loglog(nevalode(:,1),errorenergiaode(:,1),'-->r','LineWidth',1,'
    MarkerSize',3)
hold on
xlabel('N mero de evaluaciones')
ylabel(['Error en la energ a'])
title(['Problema de Kepler e=0.5 hasta  ' num2str(nperiodos) '
    periodos'])
legend('RKN', 'ode45')
hold off

end
```