Resolución numérica de PVI

M. Palacios

Dpto. Matemática Aplicada (CPS)
Universidad de Zaragoza, Spain. *

1. Introducción a la resolución de P.V.I.
   (a) Planteamiento del problema.
   (b) El método de Euler.
   (c) El método de Heun.
   (d) Métodos de un paso
2. Métodos Runge-Kutta
   (a) Definición de métodos R-K con paso fijo.
   (b) Métodos R-K con paso variable. Estimación del error.
      i. Estimación por extrapolación de Richardson.
      ii. Estimación por pares encajados.
   (c) Extrapolación de Richardson.
3. Resolución de problemas "stiff"
   (a) Planteamiento del problema.
   (b) A-estabilidad. Función de amplificación.
   (c) Métodos implícitos

*e-mail: mpala@posta.unizar.es

Introducción a la resolución numérica de P.V.I.

1 Planteamiento del problema.

En este capítulo se estudiará la resolución numérica de problemas de valor inicial (PVI) que tengan una solución única. Para describir un PVI se utilizará la notación siguiente:

$$\begin{array}{l} y' = f(t,y), \quad t \in [t_0, T] \\ y(t_0) = \eta \end{array} \tag{1}$$

donde $y \in \mathbb{R}^s$ and $f : [t_0, T] \times \mathbb{R}^s \longrightarrow \mathbb{R}^s$. Aunque si no se dice nada en contra se supondrá que $s = 1$.

Antes de intentar buscar una aproximación a la solución del PVI, interesa saber si existe una única solución y, caso de que exista, si es estable, es decir, si pequeños cambios en las condiciones iniciales o en la ecuación diferencial originarán cambios también pequeños en la solución. Habitualmente esta va a ser la situación, ya que al menos se cometerán errores de redondeo.

A continuación se recordarán algunos resultados sobre ecuaciones diferenciales.

Definición 1 *Se dice que una función $f(t,y)$ satisface una* **condición de Lipschitz** *para la variable y en un conjunto $D \in \mathbb{R}^{s+1}$, si existe una constante $L \in \mathbb{R}^+$ tal que*

$$|f(t,y_1) - f(t,y_2)| \leq L\,|y_1 - y_2|$$

$\forall (t,y_1), (t,y_2) \in D$.

Se puede demostrar fácilmente que "si D es convexo y f tiene derivadas parciales con respecto a y acotadas en D, entonces f es lipschitziana con respecto a y". Esta condición suele ser mucho más fácil de comprobar que la definición.

En el teorema siguiente se presenta una versión del teorema de existencia y unicidad de solución para un PVI (la demostración puede consultarse en libros de ecuaciones diferenciales).

Teorema 2 *Sea $D = [t_0, T] \times \mathbb{R}^s$. Si f es continua y lipschitziana para y en D, entonces el problema de valor inicial (1) tiene una única solución*

Ejemplo 3 *Probar que existe una única solución del PVI:*

$$y' = 1 + t\,sen(t\,y), \quad , t \in [0,2]\,, \quad y(0) = 0$$

Sol.: Como f es continua y derivable con respecto a y en D, el teorema del valor medio permite escribir:

$$|f(t,y_1) - f(t,y_2)| = |y_1 - y_2|\,|t^2\,cos\,(t\xi)| \leq 4\,|y_1 - y_2|$$

Luego, f es lipschitziana con respecto a y con constante $L = 4$. (Por supuesto, la constante de Lipschitz no es única). ■

Seguidamente se aborda la estabilidad del PVI, es decir, el efecto que ciertas perturbaciones sobre el PVI producen sobre la solución.

Suele llamarse **problema perturbado** del PVI original al siguiente PVI:

$$y' = f(t,y) + \delta(t), \quad , t \in [t_0, T], \quad y(t_0) = \eta + \epsilon_0$$

Definición 4 *Se dice que el PVI original es* **estable** *(o que está bien planteado) si, teniendo solución única $y(t)$, existen constantes ϵ, $k \in \mathbb{R}^+$ tales que el problema perturbado también tiene solución única $z(t)$ que satisface:*

$$|z(t) - y(t)| < k\,\epsilon$$

suponiendo que $|\epsilon_0| < \epsilon$ y $|\delta(t)| < \epsilon$.

El siguiente teorema presenta condiciones suficientes para la estabilidad del PVI.

Teorema 5 *Si f es continua y lipschitziana con respecto a y en D, entonces el PVI (1) es estable.*

Ejemplo 6 *Comprobar que el PVI:*

$$y' = 1 + t - y, \quad t \in [0,1], \quad y(0) = 1$$

es estable

Sol.: El PVI original tiene como solución: $y(t) = t + e^{-t}$ y el problema perturbado

$$z' = 1 + t - z + \delta, \quad t \in [0,1], \quad z(0) = 1 + \epsilon_0$$

donde δ y ϵ_0 son constantes, tiene como solución: $z(t) = (1 + \epsilon_0 - \delta)\,e^{-t} + t + \delta$. En consecuencia,

$$|z(t) - y(t)| = |(\delta - \epsilon_0)\,e^{-t} - \delta| \leq |\epsilon_0| + |\delta||1 - e^{-t}| \leq 2\,\epsilon$$

$\forall\, t \in [0,1]$.

Es inmediato comprobar que se cumplen las condiciones suficientes. ∎

2 Método de Euler.

Se presenta en esta sección el método de Euler que, aunque raramente se utiliza en la práctica, sin embargo, su construcción es un ejemplo ilustrativo de las técnicas más avanzadas que se usan en otros métodos más potentes.

El objetivo de cualquier método numérico de integración de un PVI es la construcción aproximada de la solución en una red de N puntos o nodos del intervalo $[t_0, T]$ conociendo el valor de la solución y la pendiente de la solución en cualquier punto. En otros puntos diferentes de los nodos la solución puede construirse utilizando las técnicas de interpolación ya conocidas.

Los puntos de la red serán denotados por $\{t_0, t_1, ..., t_N\}$ y si están uniformente distribuidos (lo que supondremos si no se dice nada en contra) cumplirán:

$$t_k = t_0 + k\,h, \quad k = 0, 1, ..., N$$

siendo $h = \dfrac{T - t_0}{N}$ lo que se denomina **paso de integración**.

Suponiendo que la solución del PVI (1) es de clase $C^{(2}\,[t_0, T]$, para cada nodo la solución puede desarrollarse en serie de Taylor en la forma siguiente:

$$y(t_{k+1}) = y(t_k) + h\,y'(t_k) + \frac{h^2}{2}y''(\xi_k)$$

para algún $\xi_i \in [t_0, T]$.

Puesto que $y(t)$ es solución de la ecuación diferencial (1) se cumple:

$$y(t_{k+1}) = y(t_k) + h\,f(t_k, y(t_k)) + \frac{h^2}{2}y''(\xi_k)$$

y, suponiendo que h es suficientemente pequeño, se pueden construir, a partir del valor inicial $y(t_0) = \eta$, soluciones aproximadas $y_k \approx y(t_k)$ de la solución en los nodos mediante el algoritmo:

$$\begin{array}{l} y_0 = \eta_h \\ y_{n+1} = y_n + h\,f(t_n, y_n; h), \quad 0 \le n \le N-1 \end{array} \tag{2}$$

que es la ecuación en diferencias asociada al método de Euler.

Geométricamente, el método de Euler puede ser interpretado en la forma que se muestra en la figura 2; para ello es preciso tener en cuenta que si y_k es una aproximación suficientemente buena de la solución y el PVI es estable, se tiene:

$$f(t_k, y_k) \approx y'(t_k) = f(t_k, y(t_k))$$

El error al aproximar la solución exacta en los nodos por los valores proporcionados por el método viene recogido en el siguiente

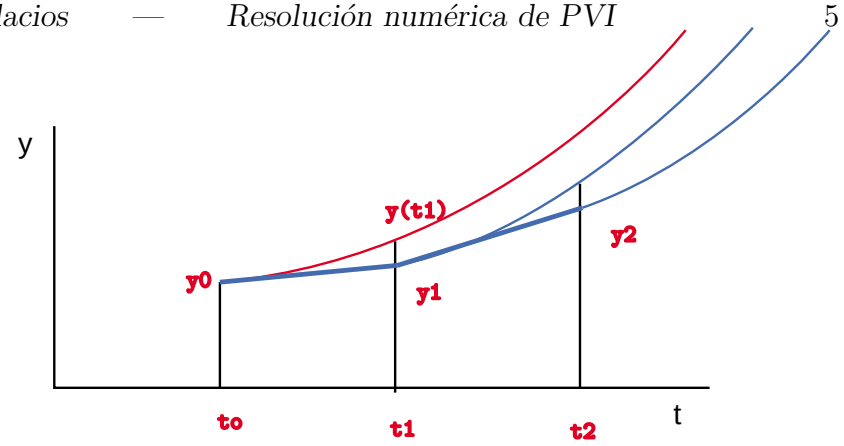

Figure 1: Interpretación geométrica del método de Euler.

Teorema 7 *Sea $y(t)$ la solución del PVI (1) y sean $y_0, y_1, ..., y_N$ las aproximaciones generadas por el método de Euler. Si f es lipschitziana con respecto a y con constante L y la derivada segunda $y''(t)$ existe y está acotada en $[t_0, T]$ por una constante M, entonces se verifica:*

$$|y(t_k) - y_k| \leq \frac{h\,M}{2\,L}\,\left[e^{L\,(t_k - t_0)} - 1\right]$$

$\forall\, k = 0, 1, 2, ..., N$

Dem.: El resultado es cierto para $k = 0$, ya que $y(t_0) = y_0$.

Por otro lado, para cualquier valor $k = 0, 1, 2, ..., N-1$, desarrollando en serie de Taylor se tiene:

$$y(t_{k+1}) = y(t_k) + h\,f(t_k, y(t_k)) + \frac{h^2}{2} y''(\xi_k)$$

Por consiguiente,

$$|y(t_{k+1}) - y_{k+1}| \leq |y(t_k) - y_k| + h\,|f(t_k, y(t_k)) - f(t_k, y_k))| + \frac{h^2}{2}|y''(\xi_k)|$$

Como, por hipótesis, f es lipschitziana y $|y''(t)| \leq M$, resulta

$$|y(t_{k+1}) - y_{k+1}| \leq |y(t_k) - y_k|\,(1 + h\,L) + \frac{h^2\,M}{2}$$

Sustituyendo cada diferencia $|y(t_k) - y_k|$ por su expresión en función de la anterior, se llega a:

$$|y(t_{k+1}) - y_{k+1}| \leq |y(t_0) - y_0|\,(1 + h\,L)^{k+1} +$$
$$\frac{h^2\,M}{2}[\left[1 + (1 + h\,L) + (1 + h\,L)^2 + ... + (1 + h\,L)^k\right]$$

La suma de los términos de la progresión geométrica de razón $(1 + h\,L)$ resulta:

$$\frac{1 - (1 + h\,L)^{k+1}}{1 - (1 + h\,L)} = \frac{(1 + h\,L)^{k+1} - 1}{h\,L}$$

Como el término $(1+h\,L)^{k+1} \leq e^{(k+1)\,h\,L}$, resulta finalmente:

$$|y(t_k) - y_k| \leq \frac{h\,M}{2\,L}\,\left[e^{L\,(t_k - t_0)} - 1\right]$$

$\forall\, k = 0, 1, 2, ..., N$. ■

Evidentemente, la utilidad de este resultado radica en el conocimiento de una cota de la derivada segunda; cuanto mejor sea ésta, más realista será la acotación. Para algunos problemas es posible conocer la $y''(t)$ sin conocer $y(t)$, ya que

$$\begin{aligned} y''(t) &= \frac{d\,y'(t)}{d\,t} = \frac{d\,f(t,t(t))}{d\,t} \\ &= \frac{\partial f(t,t(t))}{\partial t} + \frac{\partial f(t,t(t))}{\partial y}\,f(t,t(t)) \end{aligned}$$

y en consecuencia obtener una cota para $y''(t)$.

Lo más interesante de la acotación anterior es el hecho de que **la cota del error depende linealmente del paso**. El error se puede desarrollar asintóticamente [0] y encontrar que

$$|y(t_k) - y_k| = \mathcal{O}(h) = C\,h + ...$$

en donde la contante positiva C es denominada **constante del error**; para este método, $C = 0.256$.

Ejemplo 8 *Comparar el error "global" al utilizar el método de Euler con diferentes pasos para la integración del PVI siguiente:*

$$y' = \tfrac{1}{2}\,(y - t), \quad t \in [0, 3]\,, \quad y(0) = 1$$

Sol.: En la tabla 1 se da el error (comparando con la solución exacta) y el valor $C\,h$ para diferentes valores del paso de integración en el nodo $t_N = 3$. Se puede comprobar que el error se reduce aproximadamente a la mitad cuando el paso se divide por 2.

Paso de int., h	Número de pasos, N	y_N	$\lvert y(t_N) - y_N \rvert$	$\mathcal{O}(h) \approx Ch$
1	3	1.375000	0.294390	0.256
1/2	6	1.533936	0.135454	0.128
1/4	12	1.604252	0.065138	0.064
1/8	24	1.637429	0.031961	0.032
1/16	48	1.653557	0.01583	0.016
1/32	96	1.661510	0.0007880	0.008
1/64	192	1.665459	0.0003931	0.004

Table 1: El error en el método de Euler.

Obsérvese que en el teorema de acotación del error no se ha considerado el efecto que el error de redondeo ejerce sobre la elección del paso. Evidentemente,

al disminuir el tamaño del paso aumenta el número de iteraciones, por lo que si se utiliza aritmética finita el error de redondeo aumentará en la misma medida. En la práctica, en lugar del algoritmo del método expresado en (2), se utiliza el siguiente:

$$\begin{array}{l} y_0 = \eta_h + \delta_0 \\ \tilde{y}_{n+1} = \tilde{y}_n + h\, f(t_n, \tilde{y}_n; h) + \delta_n, \quad 0 \leq n \leq N-1 \end{array} \tag{3}$$

donde δ_0 es el error de redondeo en el valor inicial y δ_n es el error de redondeo asociado al cálculo del segundo miembro.

De forma parecida a la utilizada en la obtención de la acotación del error se deduce el siguiente resultado.

Teorema 9 *Bajo las hipótesis del teorema anterior, si $\tilde{y}_k$ son las aproximaciones obtenidas con el algoritmo (3) y si $|\delta_k| < \delta$, $k = 0, 1, ..., N$, entonces:*

$$|y(t_k) - \tilde{y}_k| \leq \frac{1}{L}\left(\frac{h\,M}{2} + \frac{\delta}{h}\right)\left[e^{L\,(t_k - t_0)} - 1\right] + |\delta_0|\; e^{L\,(t_k - t_0)}$$

$\forall\, k = 0, 1, 2, ..., N$.

Nótese que la cota del error ya no es lineal y, puesto que el término $\phi(h) = \dfrac{h\,M}{2} + \dfrac{\delta}{h}$ ntiende hacia 0 cuando $h \to 0$, el error se hará grande para valores suficientemente pequeños de h. El valor mínimo de ϕ se alcanza cuando $h = \sqrt{2\delta/M}$, por lo que reducir el paso más de este valor no mejorará la aproximación. Normalmente, este valor mínimo suele ser mucho menor que los valores del paso que se utilizan.

3 El método de Heun.

Una idea diferente a la de Euler [1] es la utilizada al construir el método de Heun para la resolución numérica del PVI (1). En esta ocasión se integra la ecuación diferencial en el intervalo $[t_k, t_{k+1}]$, suponiendo que $y(t_k)$ es conocido, para obtener:

$$y(t_{k+1}) = y(t_k) + \int_{t_k}^{t_{k+1}} f(t, y(t))\, dt$$

Para calcular la integral puede utilizarse una fórmula de cuadratura, por ejemplo, la fórmula trapezoidal con paso $h = t_{k+1} - t_k$, resultando:

$$y(t_{k+1}) = y(t_k) + \frac{h}{2}\left[f(t_k, y(t_k)) + f(t_{k+1}, y(t_{k+1}))\right]$$

Como el valor de la solución en t_{k+1} no es conocido, se puede aproximar a partir de $y(t_k)$ por el método de Euler, obteniendose el siguiente **algoritmo de Heun**:

$$\begin{array}{l} y_0 = \eta = y(t_0) \\ y_{n+1} = y_n + \dfrac{h}{2}\left[f(t_n, y_n) + f(t_n + h, y_n + h\, f(t_n, y_n))\right] \dfrac{h}{2} \end{array} \tag{4}$$

Obsérvese el papel que juegan la diferenciación y la integración en este método. En primer lugar, siguiendo la recta tangente a la solución $y(t)$ en el punto y_0, se determina el punto (t_1, y_1); en segundo lugar, se toma la fórmula de cuadratura trapezoidal con vértices (t_0, f_0), (t_1, f_1) para aproximar la integral.

Más adelante será probado que el método de Heun es de orden 2, es decir,

$$|y(t_k) - y_k| = \mathcal{O}(h^2) \approx C\, h^2$$

donde $C \approx -0.0432$

Ejemplo 10 *Comparar el error "global" al utilizar el método de Heun con diferentes pasos para la integración del PVI siguiente:*

$$y' = \tfrac{1}{2}(y - t), \quad t \in [0, 3], \quad y(0) = 1$$

Sol.: En la tabla 2 se da el error global (comparando con la solución exacta) y el valor $C\, h^2$ para diferentes valores del paso de integración en el nodo $t_N = 3$. Se puede comprobar que el error se reduce aproximadamente a la cuarta parte cuando el paso se divide por 2. Puede observarse su mejoría cuando se compara con la misma tabla para el método de Euler.

[1] Idea de Runge al estudiar el movimiento de partículas cargadas en una aurora boreal

Paso de int., h	Número de pasos, N	y_N	$\lvert y(t_N) - y_N \rvert$	$O(h^2) \approx Ch^2$
1	3	1.732422	-0.063032	-0.043200
1/2	6	1.682121	-0.012310	-0.010800
1/4	12	1.672269	-0.002879	-0.002700
1/8	24	1.670076	-0.000686	-0.000675
1/16	48	1.669558	-0.000168	-0.000169
1/32	96	1.669432	-0.000042	-0.000042
1/64	192	1.669401	-0.000011	-0.000011

Table 2: El error en el método de Heun.

4 Métodos de un paso.

En este sección estudiaremos las propiedades básicas de los métodos de un paso para la resolución numérica del problema de valor inicial (1)

Los métodos de un paso pueden formularse, como generalización del método de Euler, en la forma siguiente:

$$\begin{aligned} &y_0 = \eta_h \\ &y_{n+1} = y_n + h\,\phi_f(t_n, y_n; h), \quad 0 \le n \le N-1 \end{aligned} \tag{5}$$

donde la función ϕ, que depende de la función f y del paso h considerado, se denomina **función incremento**.

Observemos que el valor inicial η_h no tiene por qué coincidir con el valor de la solución en el instante inicial η, debido a que no se conozca su valor exacto o, aún conociendose, a que al ser implementado se produzca un error de redondeo.

4.1 Convergencia

La primera condición que debe exigirse a todo método numérico es que los valores calculados y_n se aproximen a los valores de la solución exacta $y(t_n)$ del PVI original cuando $h \to 0$. Para plantear con más precisión este concepto se introducen los **errores de discretización** del método (5) respecto al PVI (1), definidos por:

$$d_n = y_n - y(t_n), \quad 0 \le n \le N$$

Definición 11 *Se dice que el* **método** *es* **convergente** *si para todo PVI del tipo (1) y para condiciones iniciales tales que $\eta_h \to \eta$ cuando $h \to 0$ se verifica:*

$$\lim_{h\to 0} \max_{0\le n\le N} |d_n| = 0$$

En otros términos, esto significa que la solución aproximada proporcionada por el algoritmo converge uniformemente a la solución exacta del PVI.

Para matizar más el grado de aproximación se dice que el **método** es **convergente de orden p** si

$$\max_{0\le n\le N} |d_n| = O(h^p), \quad h \to 0,$$

para lo que se debe exigir que las condicones iniciales η_h cumplan: $|\eta_h - \eta| = O(h^p)$.

Ya que la convergencia es una propiedad cualitativa, su comprobación suele ser difícil, por ello es conveniente dar alguna condición más sencilla de analizar que garantice la convergencia.

4.2 Consistencia

Sea $(t^*, y^*) \in [t_0, T) \times \mathbb{R}$, se considera el siguiente PVI:

$$\begin{aligned} z' &= f(t, z), \quad t \in [t^*, T] \\ z(t^*) &= y^* \end{aligned} \tag{6}$$

cuya solución $z = z(t)$ define la trayectoria solución de la ecuación diferencial que pasa por el punto (t^*, y^*). Si se aplica el método de un paso (5) para calcular una aproximación $\bar{y}$ de la solución en el instante $t^* + h \in [t_0, T]$, resulta

$$\bar{y} = y^* + h\,\phi(t^*, y^*; h)$$

Con este planteamiento, se define el **error local** en el punto (t^*, y^*) como

$$e(t^*, y^*; h) = z(t^* + h) - z(t^*) - h\,\phi(t^*, z(t^*); h)$$

Por la propia definición de $\bar{y}$, el error local en (t^*, y^*) resulta ser la diferencia $e(t^*, y^*; h) = z(t^* + h) - \bar{y}$, es decir, el error que se cometería al aplicar el algoritmo una vez partiendo de dicho punto con paso h.

En la figura puede verse una interpretación geométrica.

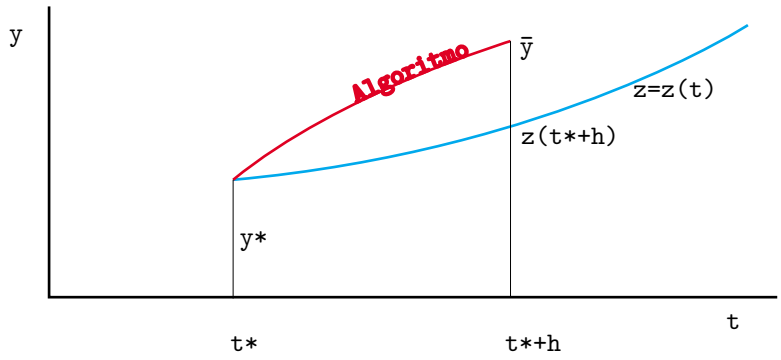

Figure 2: El error local

Dado un PVI (1) y una red de nodos $\{t_n \,|\, n = 0(1)N\}$, se pueden definir los errores locales en los puntos $(t_n, y(t_n)), 0 \le n \le N-1$ dados por $e_n = e(t_n, y(t_n); h)$. Entonces, se puede dar la siguiente

Definición 12 *El método (5) se dice* **consistente** *con el PVI (1) si se cumple que*

$$\lim_{h \to 0} \max_{0 \le n \le N-1} h^{-1} |e_n| = 0$$

En general, se cumple que si un método es consistente la condición anterior se cumple respecto de todo PVI del tipo considerado.

También se dice que el **método** es **consistente de orden q** si éste es el máximo número entero tal que

$$\max_{0\le n\le N-1} h^{-1}\,|e_n| = O(h^q), \quad h \to 0$$

para todo PVI del tipo considerado suficientemente diferenciable.

Ejemplo 13 *Analizar la consistencia del método de Heun*

Sol.: La función incremento en el método de Heun es

$$\phi[t_n, y_n; h) = \frac{1}{2}[f(t_n, y_n) + f(t_n + h, y_n + hf(t_n, y_n))],$$

por lo tanto, el error local en $(t_n, y(t_n))$ será:

$$e_n = y(t_n + h) - y(t_n) - \frac{h}{2}[f(t_n, y(t_n)) + f(t_n + h, y(t_n) + hf(t_n, y(t_n)))];$$

en consecuencia, si la solución es $y(t) \in C^3[t_0, T]$ y desarrollando $y(t)$ y $f(t, y(t))$ en serie de Taylor en torno de t_n y $(t_n, y(t_n))$, respectivamente, se tiene

$$\begin{aligned} e_n &= y(t_n) + h\,y'(t_n) + \frac{1}{2}h^2 y''(t_n) + O(h^3) \\ &- y(t_n) - \frac{h}{2}[f(t_n, y(t_n)) + f(t_n, y(t_n)) + \\ &+ h\,f_t(t_n, y(t_n)) + h\,f(t_n, y(t_n))\,f_y(t_n, y(t_n)) + O(h^2)] \\ &= O(h^3) \end{aligned}$$

Así que:

$$\max_{0\le n\le N-1} h^{-1}\,|e_n| = O(h^2), \quad h \to 0$$

por lo que el método es consistente de orden 2. ■

Ejemplo 14 *Estudiar la consistencia de los siguientes métodos:*

a) Trapecio: $y_{n+1} = y_n + \dfrac{h}{2}[f(y_n) + f(y_{n+1})]$

b) Heun: $y_{n+1} = y_n + \dfrac{h}{2}[f(t_n, y_n) + f(t_n + h, y_n + hf(t_n, y_n))]$

c) Euler modificado: $y_{n+1} = y_n + h\,f(t_n + \dfrac{h}{2}, y_n + \dfrac{h}{2}f(t_n, y_n))$

d) El método de un paso: $y_{n+1} = y_n + h\,\phi(t_n, y_n; h)$,

siendo $\phi(t, y; h) = (1 - \alpha)\,f(t, y) + \alpha\,f(t + \dfrac{h}{2\alpha}, y + \dfrac{h}{2\alpha}f(t, y))$

Véase a continuación una caracterización de la consistencia.

Teorema 15 *Suponiendo que la función incremento es continua y que* $(t^*, y^*) \in [t_0, T] \times \mathbb{R}$*, una condición necesaria y suficiente para la consistencia de un método de un paso es*

$$\phi_f(t, y; 0) \equiv f(t, y)$$

Demostr.: Como se sabe, existe un valor inicial η tal que la solución del PVI (1) verifica $y(t^*) = y^*$. Además, por la continuidad:

$$\begin{array}{l} y(t_n + h) = y(t_n) + h\, f(t_n, y(t_n)) + h\, \rho_1(h) \\ \phi(t_n, y(t_n); h) = \phi(t_n, y(t_n); 0) + \rho_2(h) \end{array} \tag{7}$$

en donde tanto ρ_1 como ρ_2 tienden a cero cuando $h \to 0$. Por lo tanto,

$$\begin{array}{rcl} h^{-1}\, e_n & = & h^{-1}\, [y(t_n + h) - y(t_n) - h\, \phi(t_n, y(t_n); h)] = \\ & = & f(t_n, y(t_n)) - \phi(t_n, y(t_n); 0) + \rho(h) \end{array}$$

donde $\rho(h) \to 0$ cuando $h \to 0$. En consecuencia, si el método es consistente,

$$f(t, y(t) - \phi(t, y(t); 0) = 0, \quad \forall t \in [t_0, T]$$

El recíproco es inmediato. ∎

4.3 Estabilidad

En esta sección se estudiará el comportamiento de un método de un paso frente a perturbaciones, es decir, se compararán las soluciones del problema discretizado original (que recordamos)

$$\begin{array}{l} y_0 = \eta \\ y_{n+1} = y_n + h\, \phi(t_n, y_n; h), \quad 0 \le n \le N - 1 \end{array} \tag{8}$$

y del **problema perturbado**

$$\begin{array}{l} \bar{y}_0 = \eta + \omega_0 \\ \bar{y}_{n+1} = \bar{y}_n + h\, \phi(t_n, \bar{y}_n; h) + \omega_{n+1}, \quad 0 \le n \le N - 1 \end{array} \tag{9}$$

en donde w_i son las perturbaciones, que pueden entenderse como errores de redondeo.

La diferencia entre las soluciones del problema original y el perturbado está acotada como se muestra en el siguiente teorema (cuya demostración puede consultarse en [0]).

Teorema 16 *(de* **estabilidad***) Si* $\phi(t, y; h)$ *es uniformemente lipschitziana con respecto a* y $\forall t \in [t_0, T]$*,* $h \in (0, h_0]$*, con constante de Lipschitz* Λ*, entonces existen constantes positivas* C_1 *y* C_2*, independientes de* h *y de las perturbaciones, tales que*

$$C_1 \max_{0 \le n \le N} \left| \sum_{j=0}^{n} \omega_j \right| \le \max_{0 \le n \le N} |\bar{y}_n - y_n| \le C_2 \max_{0 \le n \le N} \left| \sum_{j=0}^{n} \omega_j \right|$$

Una consecuencia de este teorema de estabilidad da un resultado básico de convergencia.

Teorema 17 *Si la función incremento es continua con respecto a sus argumentos y lipschitziana con respeto a y, una condición necesaria y suficiente para la convergencia del método es la consistencia (es decir, $\phi(t, y; 0) \equiv f(t, y)$*

Teorema 18 *Bajo las hipótesis del teorema anterior, si un método de un paso es consistente de orden q, entonces es convergente de orden q.*

Resumiendo, respecto de los tres conceptos anteriores podemos decir lo siguiente:

a) convergencia, indica que los valores exactos y calculados de la solución tienden a confundirse cuando $h \to 0$.

b) consistencia, mide la separación de la ecuación (en diferencias) del método y de la ecuación diferencial.

c) estabilidad, muestra el comportamiento de la ecuación del método frente a perturbaciones (por ejemplo, frente a errores de cálculo).

2. Métodos Runge-Kutta

5 Definición de métodos R-K con paso fijo.

Habida cuenta de las propiedades de los métodos de un paso, es evidente que el uso de métodos de orden más elevado que el de Euler evitaría tener que tomar paso pequeño y, consecuentemente, realizar un número elevado de evaluaciones, lo que posiblemente produciría errores de redondeo intolerables.

Una forma de elevar el orden es actuar como en el método de Euler, es decir, considerar como función incremento un desarrollo limitado de Taylor, lo que lleva consigo la dificultad de calcular las derivadas de la solución a través de la ecuación diferencial.

Otra forma de proceder es actuar como en el método de Heun, es decir, aproximar la integral más adecuadamente, lo que conduce a construir la función incremento como una combinación lineal de valores de la función f en puntos adecuados. En este capítulo sólo estudiaremos los métodos de este tipo denomidados de Runge-Kutta explícitos.

Definición 19 *Se denomina* **método Runge-Kutta (R-K)** *de m etapas para la resolución numérica del PVI (1) al representado por el algoritmo*

$$\begin{aligned} y_0 &= \eta \\ y_{n+1} &= y_n + h \sum_{j=1}^{m} b_j \, g_{nj}, \quad 0 \leq n \leq N-1 \end{aligned} \tag{10}$$

donde las funciones g_{nj} están definidas por

$$\begin{aligned} &g_{n1} = f(t_n, y_n) \\ &g_{n2} = f(t_n + c_2\, h, y_n + h\, a_{21}\, g_{n1}) \\ &g_{n3} = f(t_n + c_3\, h, y_n + h\, (a_{31}\, g_{n1} + a_{32}\, g_{n2})) \\ &\cdots \\ &g_{nm} = f(t_n + c_m\, h, y_n + h\, (a_{m1}\, g_{n1} + a_{m2}\, g_{n2} + \cdots + a_{mm-1}\, g_{nm-1})) \end{aligned} \tag{11}$$

siendo las constantes a_{jk}, b_j, c_j dependientes del método, pero independientes del PVI, y tales que $c_j = \sum_{k=1}^{j-1} a_{jk}$

Una forma cómoda de representar un método R-K es la forma matricial de Butcher que se observa en la tabla 3 en forma explícita y abreviada.

$$\begin{array}{c|ccccc} 0 & 0 & & & & \\ c_2 & a_{21} & & & & \\ c_3 & a_{31} & a_{32} & & & \\ \vdots & \vdots & & \ddots & & \\ c_m & a_{m1} & a_{m2} & \cdots & a_{mm-1} & \\ \hline & b_1 & b_2 & \cdots & b_{m-1} & b_m \end{array} \qquad \begin{array}{c|c} c & \mathcal{A} \\ \hline & b \end{array}$$

Table 3: R-K explícito en la forma matricial de Butcher.

Como una condición necesaria y suficiente para consistencia de los métodos de un paso es : $\phi_f(t, y(t); 0) = f(t, y(t))$, el método R-K será consistente solamente si $\sum_{j=1}^{m} b_j = 1$.

En adelante, supondremos que los métodos R-K considerados serán, al menos, consistentes, por lo que se cumplirán siempre las relaciones siguientes:

$$\begin{aligned} \sum_{j=1}^{m} b_j &= 1 \\ \sum_{k=1}^{j-1} a_{jk} &= c_j \end{aligned} \tag{12}$$

Ejemplo 20 *El método de Heun es un método R-K de dos etapas y orden 2.*

En efecto, puede escribirse en la forma

$$y_{n+1} = y_n + h \left[\tfrac{1}{2}\, g_{n1} + \tfrac{1}{2}\, g_{n2}\right]$$

definiendo

$$\begin{aligned} g_{n1} &= f(t_n, y_n) \\ g_{n2} &= f(t_n + h, y_n + h\, g_{n1}) \end{aligned} \tag{13}$$

En forma matricial de Butcher se puede representar mediante la tabla 4 adjunta.

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

Table 4: Forma matricial del método de Heun como un R-K explícito.

Ejemplo 21 *El método del trapecio*

$$y_{n+1} = y_n + \frac{h}{2}\, [f(t_n, y_n) + f(t_{n+1}, y_{n+1})]$$

es un método R-K de dos etapas y orden 2, pero es implícito.

5.1 Convergencia

Como los métodos R-K son métodos de un paso, para estudiar su convergencia es suficiente analizar el cumplimiento de la condición de Lipschitz para la función incremento.

Propiedad 22 *La función incremento de un método R-K,*

$$\phi(t, y; h) = \sum_{j=1}^{m} b_j\, g_j\,,$$

es uniformemente lipschitziana con respeto a y con constante de Lipschitz

$$\Lambda = \beta_b\, L\, (1 + h_0\, L\, \beta_a)^m$$

donde L es la constante de Lipschitz de la función f, $\beta_a = max\,|a_{jk}|$, $\beta_b = \sum_{j=1}^{m} |b_j|$ y $h \in (0, h_0]$

Dem.: Por recurrencia, se puede probar que:

$$|g_j(t,\tilde{y}) - g_j(t,y)| \leq L\,\nu_j\;|\tilde{y} - y|\,, \quad j = 1,2,3,\cdots,m$$

donde los números positivos ν_j están definidos por

$$\begin{aligned} \nu_1 &= 1 \\ \nu_i &= 1 + h\beta_a\,L\sum_{k=1}^{i-1}\nu_k, \quad i = 2,3,\cdots,m. \end{aligned}$$

Teniendo en cuenta que

$$\nu_j = (1 + h\,L\,\beta_a)^{j-1} \leq (1 + h\,L\,\beta_a)^m,$$

se verifica:

$$\begin{aligned} |\phi(t,\tilde{y};h) - \phi(t,y;h)| &\leq \left(\sum_{j=1}^{m}|b_j|\right)\max_{1\leq j\leq m}|g_j(t,\tilde{y}) - g_j(t,y)| \\ &\leq \beta_b\,L\,\max_{1\leq j\leq m}\nu_j\;|\tilde{y} - y| \leq \beta_b\,L\,(1 + h\,L\,\beta_a)^m \end{aligned}$$

En consecuencia, se puede enunciar el siguiente resultado:

Corolario.- Un método R-K es **convergente** si se verifican las condiciones de consistencia (12). Además, si el método es consistente de orden q, también es convergente del mismo orden.

5.2 Consistencia

En este párrafo se van a determinar los parámetros de un R-K para que éste alcance orden máximo. Para ello, en el caso de que f sea suficientemente diferenciable, se desarrolla en serie de potencias de h tanto la función incremento ϕ como la función incremento exacto $\Delta(t,y;h) = h^{-1}\,(y(t+h) - y(t))$ y se les hace coincidir hasta el orden h^{p-1}, según la definición de consistencia de orden p. No supone ninguna restricción el considerar que el sistema diferencial es autónomo. Así que, en esta situación,

$$\Delta(y;h) = f(y) + \frac{h}{2}\,f^{(1)}(y) + \frac{h^2}{3}\,f^{(2)}(y) + \cdots$$

donde, con la notación de Butcher,

$$\begin{aligned} f^{(1)}(y) &= f_y\,f = f'\,f = \{f\} \\ f^{(2)}(y) &= f''\,f\,f + f'\,f'\,f = \{f,f\} + \{\{f\}\} \\ f^{(3)}(y) &= f'''\,f^3 + 3f''\,f'\,f^2 + f''\,f'\,f^2 + f'^3\,f \end{aligned}$$

En el caso de que f sea una función vectorial, las expresiones anteriores se complican enormemente y se hace preciso utilizar una notación especial: se introduce el concepto de **diferenciales elementales** a las que se les asocia un **orden** (de derivación) y un **"arbol"**. De esta forma se puede obtener "*la derivada n-ésima de $f(y)$ como combinación lineal de diferenciales elementales de orden $n+1$ con*

coeficientes positivos"; estos coeficientes son fácilmente determinados en función del arbol considerado. No será desarrollada aquí la teoría, puede consultarse en [0], simplemente se aplicará a algunos casos sencillos, aunque pueden comprobarse los resultados por el camino clásico.

A continuación se verá el orden máximo alcanzable para algunos casos particulares de métodos R-K.

Ejemplo 23 *Métodos R-K de dos etapas:*

$$\begin{array}{c|cc} 0 & 0 & \\ c_2 & a_{21} & \\ \hline & b_1 & b_2 \end{array}$$

Condiciones a satisfacer:

$$\begin{array}{ll} \text{definición:} & a_{21} = c_2 \\ \text{orden 1:} & b_1 + b_2 = 1 \\ \text{orden 2:} & b_2\, c_2 = \frac{1}{2} \\ \text{orden 3:} & b^T\ (c \cdot\ c) = \dfrac{1}{3} \\ & b^T\, \mathcal{A}\, c = 0 = \dfrac{1}{6} \end{array}$$

Como la última de las condiciones, la asociada a la diferencial elemental $\{\{f\}\}$, no puede verificarse nunca, el método solo puede ser de orden dos como máximo; en este caso, hay 3 ecuaciones con cuatro incognitas, por lo que habrá un parámetro libre y resulta así una familia uniparamétrica de métodos R-K de orden 2, representada en la tabla 5.

$$\begin{array}{c|cc} 0 & 0 & \\ \frac{1}{2\,\alpha} & \frac{1}{2\,\alpha} & \\ \hline & 1-\alpha & \alpha \end{array}$$

Table 5: R-K explícito de 2 etapas y orden 2.

Ejemplo 24 *Métodos R-K de tres etapas:*

$$\begin{array}{c|ccc} 0 & 0 & & \\ c_2 & a_{21} & & \\ c_3 & a_{31} & a_{32} & \\ \hline & b_1 & b_2 & b_3 \end{array}$$

Condiciones a satisfacer:

$$\begin{array}{ll} \text{definición:} & a_{21} = c_2 \\ & a_{31} + a_{32} = c_3 \\ \text{orden 1:} & b_1 + b_2 + b_3 = 1 \\ \text{orden 2:} & b_2\, c_2 + b_3\, c_3 = \frac{1}{2} \\ \text{orden 3:} & b_2\, c_2^2 + b_3\, c_3^2 = \dfrac{1}{3} \\ & b_3\, a_{32}\, c_2 = \dfrac{1}{6} \end{array}$$

Como una de las condiciones de orden 4, la asociada a la diferencial elemental $\{\{\{f\}\}\}$, no puede verificarse nunca, el método solo puede ser de orden tres como máximo; en este caso, hay 6 ecuaciones no lineales con ocho incognitas, por lo que habrá dos parámetros libres y resulta así una familia biparamétrica de métodos R-K de orden 3. Un miembro de esta familia es la fórmula de Kutta que se muestra en la tabla 6.

0	0		
1/2	1/2		
1	−1	2	
	1/6	4/6	1/6

Table 6: R-K explícito de 3 etapas. Fórmula de Kutta

Ejemplo 25 *Métodos R-K de cuatro etapas:*

0	0			
c_2	a_{21}			
c_3	a_{31}	a_{32}		
c_4	a_{41}	a_{42}	a_{43}	
	b_1	b_2	b_3	b_4

Table 7: R-K explícito de 4 etapas.

Condiciones a satisfacer:

$$
\begin{aligned}
&\text{definición:} && a_{21} = c_2 \\
& && a_{31} + a_{32} = c_3 \\
& && a_{41} + a_{42} + a_{43} = c_4 \\
&\text{orden 1:} && b_1 + b_2 + b_3 + b_4 = 1 \\
&\text{orden 2:} && b_2\, c_2 + b_3\, c_3 + b_4\, c_4 = \tfrac{1}{2} \\
&\text{orden 3:} && b_2\, c_2^2 + b_3\, c_3^2 + b_4\, c_4^2 = \frac{1}{3} \\
& && b_3\, a_{32}\, c_2 + b_4\, a_{42}\, c_2 + b_4\, a_{43}\, c_3 = \frac{1}{6} \\
&\text{orden 4:} && b_2\, c_2^3 + b_3\, c_3^3 + b_4\, c_4^3 = \frac{1}{4} \\
& && b_3\, c_3\, a_{32}\, c_2 + b_4\, c_4\, a_{42}\, c_2 + b_4\, c_4\, a_{43}\, c_3 = \frac{1}{8} \\
& && b_3\, a_{32}\, c_2^2 + b_4\, a_{42}\, c_2^2 + b_4\, a_{43}\, c_3^2 = \frac{1}{12} \\
& && b_4\, a_{43}\, a_{32}\, c_2 = \frac{1}{24}
\end{aligned}
$$

Como una de las condiciones de orden 5, la asociada a la diferencial elemental $\{\{\{\{f\}\}\}\}$, no puede verificarse nunca, el método solo puede ser de orden cuatro como máximo; en este caso, hay 11 ecuaciones no lineales con 13 incógnitas, por lo que habrá dos parámetros libres y resulta así una familia biparamétrica de métodos R-K de orden 4. Un miembro de esta familia, el RK4 de la tabla 8, es el R-K más conocido.

0	0			
1/2	1/2			
1/2	0	1/2		
1	0	0	1	
	1/6	2/6	2/6	1/6

Table 8: RK4. La fórmula de Runge-Kutta más popular

Un análisis del orden de los métodos R-K resulta muy complicado debido a la no linealidad y al número creciente de las condiciones que van apareciendo. En los trabajos de Butcher se encuentra la relación entre el número de etapas, m, y el orden máximo alcanzable, que aparecen en la tabla 9. Para $m \geq 9$, el orden máximo es menor o igual que $m - 2$.

Núm. etapas	2	3	4	5	6	7	9
Orden máximo	2	3	4	4	5	6	7

Table 9: Orden máximo alcanzable frente a número de etapas en los métodos Runge-Kutta

Obsérvese que el principal esfuerzo computacional al aplicar los métodos R-K reside en la evaluación de la función f. Como se observa en la tabla 9, parece más acansejable utilizar métodos R-K con orden menor que cinco y paso pequeño que métodos de orden mayor y paso más grande.

A la hora de comparar métodos R-K de orden bajo, debe tomarse el paso h adecuado para alcanzar el nuevo nodo con el mismo número de evaluaciones de función; por ejemplo, si se compara el método R-K de orden 4 con uno de orden 2, debe tomarse en el segundo paso la mitad que en el primero.

Ejemplo 26 *Comparar los métodos de Euler, Heun y R-K popular para el PVI:*

$$y' = -y + 1, \quad t \in [0, 1], \quad y(0) = 0$$

Sol.: Se toma, por ejemplo, R-K popular con paso $h = 0.1$, Heun con paso $h/2 = 0.05$ y Euler con paso $h/4 = 0.025$. Los resultados pueden verse en la tabla 10

t	Euler, $h = 0.025$	Heun, $h = 0.05$	R-K (4), $h = 0.1$	Valor exacto
0.1	0.096312	0.095123	0.09516250	0.095162582
0.2	0.183348	0.181198	0.18126910	0.181269470
0.3	0.262001	0.259085	0.25918158	0.259181779
0.4	0.333079	0.329563	0.32967971	0.329679954
0.5	0.397312	0.393337	0.39346906	0.393469340

Table 10: Comparación de métodos Runge-Kutta

6 Motivación de métodos R-K con paso variable.

Podría decirse que un método ideal de un paso sería aquel que diera la solución con un error global menor que la tolerancia prefijada utilizando el mínimo número de nodos. Sin embargo, esta posibilidad es incompatible con que los nodos estén igualmente espaciados. Por ello se buscan maneras de estimar el error local (la parte principal del mismo) y, en consecuencia, elegir en cada iteración el paso adecuado para que el error sea menor que una tolerancia dada; una de éstas está basada en la extrapolación al límite de Richardson; otra utiliza pares de métodos R-K, lo que permite disminuir el número de evaluaciones de función requeridas; ambos se describen sucintamente a continuación.

6.1 Estimación del error mediante extrapolación de Richardson.

Suponiendo que se utiliza un método Runge-Kutta de orden p y que la solución numérica ha sido encontrada en el nodo t_n, se calculará la solución en el nodo t_{n+2} de dos formas: a) realizando dos iteraciones con paso h; b) realizando una iteración con paso $2\,h$.

Llamando y^h_{n+2} y y^{2h}_{n+2} a las soluciones numéricas respectivas y desarrollando en serie de Taylor y teniendo en cuenta que el mtodo RK es de orden p, se obtiene (ver [0]):

$$y(t_n+2h) = y^h_{n+2} + 2\,h^{p+1}\,\phi(t_n,y_n) + \mathcal{O}(h^{p+2}) \tag{14}$$

$$y(t_n+2h) = y^{2h}_{n+2} + 2\,h(2\,h)^p\,\phi(t_n,y_n) + \mathcal{O}(h^{p+2}) \tag{15}$$

Restando ambas ecuaciones se deduce:

$$y^h_{n+2} - y^{2h}_{n+2} = h^{p+1}\,\phi(t_n,y_n)\,(2-2^{p+1}) + \mathcal{O}(h^{p+2})$$

Esta última expresión permite estimar la parte principal del error local en t_{n+2} cuando se toma paso h (obsérvese (14)), en la forma:

$$y(t_n+2h) - y^h_{n+2} \approx 2\,h^{p+1}\,\phi(t_n,y_n) \approx \frac{y^{2h}_n - y^{h+2}_{n+2}}{2^p-1}$$

Aunque esta estimación del error es fácilmente calculable y programable, y por ello ampliamente usada, requiere un costo computacional más alto que si no se utilizase control del error. EN la siguiente sección se obtiene otro tipo de estimación del error que requiere menos costo computacional.

6.2 Estimación del error mediante pares encajados.

Se considera un método R-K de m etapas y orden p, que se denotará RK(p), del que se desea obtener una estimación del error local; para ello se considera también otro R-K de orden $p+1$, que se denotará RK(p+1); se tiene entonces:

$$RK(p), \quad y_{n+1} = y(t_{n+1}) - e(t_n,y_n;h)$$

$$RK(p+1), \quad \tilde{y}_{n+1} = y(t_{n+1}) - \tilde{e}(t_n,y_n;h)$$

siendo

$$\begin{aligned} RK(p), \quad e(t_n, y_n; h) &= h^{p+1}\,\phi(t_n, y_n) + \mathcal{O}(h^{p+2}) \\ RK(p+1), \quad \tilde{e}(t_n, y_n; h) &= \mathcal{O}(h^{p+2}) \end{aligned}$$

Restando miembro a miembro, resulta:

$$\tilde{y}_{n+1} - y_{n+1} = h^{p+1}\,\phi(t_n, y_n) + \mathcal{O}(h^{p+2}) = e(t_n, y_n; h) + \mathcal{O}(h^{p+2}) \tag{16}$$

Por lo que la diferencia de valores calculados da una estimación del error local para la fórmula de orden p. En consecuencia, la parte principal del error local (relativo al paso) del método de orden p verifica:

$$h^{-1}\,e(t_n, y_n; h) \approx C\,h^p = \frac{1}{h}\,(\tilde{y}_{n+1} - y_{n+1}) \tag{17}$$

donde C es la constante del error.

Esta estimación permite controlar el error eligiendo el paso adecuado en cada iteración. En efecto, sea $\tilde{h} = q\,h$, $q > 0$, el nuevo paso de integración. La igualdad (17) permite escribir ahora:

$$\tilde{h}^{-1}\,e(t_n, y_n; \tilde{h}) \approx C\,\tilde{h}^p = \frac{q^p}{h}\,(\tilde{y}_{n+1} - y_{n+1})$$

Entonces, el error local relativo será menor que una tolerancia $TOL = \epsilon$ prefijada, si se cumple

$$q \leq \left(\frac{\epsilon\,h}{|\tilde{y}_{n+1} - y_{n+1}|}\right)^{\frac{1}{p}} \tag{18}$$

En general, los coeficientes a_{jk} de ambos métodos no tienen por qué ser iguales, por lo que para obtener la estimación (16) se hace necesario realizar todas las evaluaciones del método $RK(p)$ y todas las del $RK(p+1)$.

Para reducir el número de evaluaciones se ha ideado la técnica de la **estimación mediante pares encajados**, que consiste en lo siguiente:

Para un número de etapas m prefijado construir dos métodos R-K que tengan los mismos c_j y a_{jk} y que sean de orden p y $p+1$, respectivamente. En consecuencia, los dos métodos

$$y_{n+1} = y_n + h \sum_{j=1}^{m} b_j\,g_j, \quad 0 \leq n \leq N-1$$

$$\tilde{y}_{n+1} = y_n + h \sum_{j=1}^{m} \tilde{b}_j\,g_j, \quad 0 \leq n \leq N-1$$

pueden ser representados en una única matriz de Butcher como se observa en la tabla 11 expícita y abreviadamente. cumpliendose, además, las relaciones siguientes:

$$\sum_{j=1}^{m} b_j = 1, \quad \sum_{j=1}^{m} \tilde{b}_j = 1 \tag{19}$$

$$\sum_{k=1}^{j-1} a_{jk} = c_j \tag{20}$$

La obtención de pares encajados es un poco más difícil que la de un solo método de orden máximo.

c_2	a_{21}				
c_3	a_{31}	a_{32}			
$\vdots$	$\vdots$		$\ddots$		
c_m	a_{m1}	a_{m2}	$\cdots$	a_{mm-1}	
	b_1	b_2	$\cdots$	b_{m-1}	b_m
	$\tilde{b}_1$	$\tilde{b}_2$	$\cdots$	$\tilde{b}_{m-1}$	$\tilde{b}_m$

c	$\mathcal{A}$
	b
	$\tilde{b}$

Table 11: R-K encajado de m etapas.

Ejemplo 27 *Obtención de un par encajado RK2(3), con $m = 3$ etapas.*

Sol.: Como se sabe, el orden máximo alcanzable es 3; luego, a la fuerza, será un RK(2) encajado en un RK(3). Las condiciones de orden que deben cumplirse son las siguientes:

$$\begin{array}{lll}
\text{definición:} & a_{21} = c_2 & \\
 & a_{31} + a_{32} = c_3 & \\
\text{orden 1:} & \tilde{b}_1 + \tilde{b}_2 + \tilde{b}_3 = 1, & \text{para el RK(3)} \\
\text{orden 2:} & \tilde{b}_2\, c_2 + \tilde{b}_3\, c_3 = \frac{1}{2} & \\
\text{orden 3:} & \tilde{b}_2\, c_2^2 + \tilde{b}_3\, c_3^2 = \dfrac{1}{3} & \\
 & \tilde{b}_3\, a_{32}\, c_2 = \dfrac{1}{6} &
\end{array}$$

$$\begin{array}{lll}
\text{orden 1:} & b_1 + b_2 + b_3 = 1, & \text{para el RK(2)} \\
\text{orden 2:} & b_2\, c_2 + b_3\, c_3 = \frac{1}{2} &
\end{array}$$

La resolución llevaría el siguiente proceso: en primer lugar, se resuelve el sistema para el $RK(3)$ en función de 2 parámetros, a continuación, se resuelve el sistema para el $RK(2)$ en función de los valores anteriores. De esta forma se obtiene, por ejemplo, el par $RK2(3)$ de la tabla 12

$\frac{1}{3}$	$\frac{1}{3}$		
$\frac{2}{3}$	0	$\frac{2}{3}$	
$RK(2)$	0	$\frac{1}{2}$	$\frac{1}{2}$
$RK(3)$	$\frac{1}{4}$	0	$\frac{1}{4}$

Table 12: Un R-K encajado de 3 etapas.

Teniendo en cuenta (17), la estimación del error local puede escribirse ahora en la forma

$$\tilde{y}_{n+1} - y_{n+1} = h \sum_{j=1}^{m} (\tilde{b}_j - b_j)\, g_j. \tag{21}$$

Nótese que no es preciso calcular explícitamente $\tilde{y}_{n+1}$, ni restar dos cantidades próximas para estimar el error.

Nótese también que la estimación anterior puede ser sustituida en la expresión (18) que define el nuevo paso.

Así, pues, la forma más apropiada de expresar los pares encajados es la que se muestra en la tabla 13.

c	$\mathcal{A}$
$RK(p)$	b
Estim.	$\tilde{b}-b$

Table 13: R-K encajado de m etapas.

Uno de los métodos encajados más habituales, denominado $RKF(4,5)$, es el obtenido por Fehlberg (1970) imponiendo algunas condiciones sobre los coeficientes y que sean lo más pequeños posible para aliviar en lo posible los pesados cálculos; puede verse en la tabla 14

0						
$\frac{1}{4}$	$\frac{1}{4}$					
$\frac{3}{8}$	$\frac{3}{32}$	$\frac{9}{32}$				
$\frac{12}{13}$	$\frac{1932}{2197}$	$\frac{-7200}{2197}$	$\frac{7296}{2197}$			
1	$\frac{439}{216}$	-8	$\frac{3680}{513}$	$\frac{-845}{4104}$		
$\frac{1}{2}$	$\frac{-8}{27}$	2	$\frac{-3544}{2565}$	$\frac{1859}{4104}$	$\frac{-11}{40}$	
$RK(4)$	$\frac{25}{216}$	0	$\frac{1408}{2565}$	$\frac{2197}{4104}$	$\frac{-1}{5}$	0
$RK(5)$	$\frac{16}{135}$	0	$\frac{6656}{12825}$	$\frac{28561}{56430}$	$\frac{-9}{50}$	$\frac{2}{55}$
Estimac.	$\frac{1}{360}$	0	$\frac{-128}{4275}$	$\frac{-2187}{75240}$	$\frac{1}{50}$	$\frac{2}{55}$

Table 14: Un método R-K-Fehlberg, el RKF(4,5).

Conviene observar que una ventaja clara de este método es que sólo se requieren 6 evaluaciones de la función f por paso frente a las 10 que necesitaría un par no encajado.

Cuando se realiza el control del error mediante estimación del nuevo paso (ver (18)), el valor de q se utiliza con dos propósitos: a) rechazar la elección del paso y b) elegir adecuadamente el nuevo paso. Para reducir el costo computacional (en términos de evaluaciones de función) se tiende a escoger q de manera conservadora, como en la fórmula siguiente para el RKF(4,5):

$$q=\left(\frac{TOL\,h}{2\,|\tilde{y}_{n+1}-y_{n+1}|}\right)^{\frac{1}{4}}=0.84\,\left(\frac{TOL\,h}{|\tilde{y}_{n+1}-y_{n+1}|}\right)^{\frac{1}{4}} \tag{22}$$

y, además, no se permiten grandes modificaciones del paso.

Ejemplo 28 *Resolver mediante el RKF(4,5) el PVI siguiente:*

$$y'=-y+t+1,\quad t\in[0,1]\,,\,y(0)=1$$

tomando una tolerancia $TOL=5*10^{-5}$ *y paso* $0.02\leq h\leq 0.1$

Sol.: Tomando un paso inicial $h = TOL^{1/4}$, se obtienen los resultados, con 6 dígitos, que aparecen en la tabla 15. Se puede obervar que sólo se modifica el paso en la primera iteración.

n	t_n	h	estim. error loc. rel.	error global
1	0.0840896	0.0840896	$9.674 * 10^{-8}$	0.
2	0.1840896	0.1	$2.398 * 10^{-7}$	$2. * 10^{-6}$
3	0.2840896	0.1	$1.420 * 10^{-7}$	$2. * 10^{-6}$
4	0.3840896	0.1	$1.863 * 10^{-7}$	$4. * 10^{-6}$
5	0.4840896	0.1	$1.257 * 10^{-7}$	$6. * 10^{-6}$
6	0.5840896	0.1	$1.490 * 10^{-7}$	$6. * 10^{-6}$
7	0.6840896	0.1	$2.002 * 10^{-7}$	$6. * 10^{-6}$
8	0.7840896	0.1	$1.839 * 10^{-7}$	$8. * 10^{-6}$
9	0.8840896	0.1	$1.350 * 10^{-7}$	$9. * 10^{-6}$
10	0.9840896	0.1	$8.615 * 10^{-8}$	$1. * 10^{-5}$

Table 15: Un ejemplo de aplicación del RKF(4,5).

3. Resolución de problemas "stiff"

7 Planteamiento del problema.

Según es conocido, un método numérico aplicado a un problema de valor inicial no es más que una discretización del mismo. Sería deseable, por lo tanto, que las propiedades de estabilidad de la ecuación diferencial se trasladaran a la ecuación en diferencias del método numérico. La formulación de esta cuestión para problemas generales sería muy difícil, por ello, se suele restringir a cierto tipo de problemas que surgen muy a menudo en aplicaciones (cinética de las reacciones químicas, teoría de control, etc.)

Ejemplo 29 *Sea el sistema diferencial (Grigorieff, 1977) [0]*

$$y_1' = \frac{\lambda_1 + \lambda_2}{2}\, y_1 + \frac{\lambda_1 - \lambda_2}{2}\, y_2$$
$$y_2' = \frac{\lambda_1 - \lambda_2}{2}\, y_1 + \frac{\lambda_1 + \lambda_2}{2}\, y_2$$

donde $\lambda_1, \lambda_2 \in \mathbb{R}^-$.

La solución general es

$$y_1(x) = C_1\, e^{\lambda_1 x} + C_2\, e^{\lambda_2}\, x$$
$$y_2(x) = C_1\, e^{\lambda_1 x} - C_2\, e^{\lambda_2}\, x$$

$\forall x \geq 0$. Si la ecuación anterior es integrada numéricamente, por ejemplo, con el método de Euler, la solución numérica se puede expresar en la forma:

$$y_{1\,n} = C_1\,(1 + h\lambda_1)^n + C_2\,(1 + h\lambda_2)^n$$
$$y_{2\,n} = C_1\,(1 + h\lambda_1)^n - C_2\,(1 + h\lambda_2)^n$$

Evidentemente, la aproximación converge a cero cuando $h \to 0$ solamente si se toma el paso h suficientemente pequeño para que

$$|1 + h\lambda_1| < 1 \quad y \quad |1 + h\lambda_2| < 1 \tag{23}$$

Supóngase que $|\lambda_2|$ es grande comparado con $|\lambda_1|$. Como ambos son negativos, la influencia de la componente $e^{\lambda_2 x}$ es despreciable frente a la otra. Sin embargo, esto es falso para la solución numérica. Según (23), el paso de integración debería ser escogido tan pequeño como

$$0 < h < \frac{2}{|\lambda_2|}.$$

Por ejemplo, si $\lambda_1 = -1, \lambda_2 = -1000$, se debe tomar $h < 0.002$, es decir muy pequeño. Este comportamiento de la solución numérica suele ser denominado "stiffness" (rigidez) y se habla de un "problema stiff". Es evidente que el método de Euler no es adecuado para tratar problemas stiff.

8 A-estabilidad de los métodos R-K.

Para el estudio de los problemas stiff suele considerarse una **ecuación de prueba** lineal y homogenea como la siguiente:

$$\begin{aligned} y'(t) &= \lambda\, y(t) \\ y(0) &= y_0 \end{aligned} \tag{24}$$

$\lambda \in \mathbb{C}$.

Si se aplica, ahora, un método R-K de m etapas con paso fijo h a la ecuación de prueba (24), resulta la sucesión recurrente:

$$Y_j = y_n + h\,\lambda \sum_{k=1}^{m} a_{jk} Y_k$$
$$y_{n+1} = y_n + h\,\lambda \sum_{j=1}^{m} b_j Y_j$$

que se puede escribir abreviadamente, en forma matricial, como sigue:

$$\begin{aligned} Y &= y_n e + h\,\lambda\,\mathcal{A}\,Y \\ y_{n+1} &= y_n + h\,\lambda\, b^T Y \end{aligned} \tag{25}$$

llamando $Y = (Y_1, Y_2, \ldots, Y_m)^T$ y $e = (1, 1, \ldots, 1)$; finalmente, suponiendo que la matriz $(I - h\lambda\mathcal{A})$ es regular, se obtiene:

$$y_{n+1} = (1 + h\,\lambda\, b^T (I - h\lambda\mathcal{A})^{-1} e) y_n = R(h\lambda) y_n \tag{26}$$

Definición 30 *Se denomina* **función o factor de amplificación** *a la función*

$$R(z) = 1 + z\, b^T (I - z\mathcal{A})^{-1} e \tag{27}$$

Puesto que la solución del PVI está acotada $\forall \lambda \in \mathbb{C}$, la solución numérica estará también acotada si y solo si $|R(z)| \leq 1$, $\forall z \in \mathbb{C}$ con $\Re(z) \leq 0$; naturalmente, esto no se cumple en general, por eso se dan las siguientes definiciones.

Definición 31 *Se denomina* **dominio de estabilidad absoluta** *o A-estabilidad del método RK (3) al conjunto*

$$D = \{z \in \mathbb{C} \,|\, |R(z)| \leq 1\}$$

Definición 32 *Se dice que un* **método RK** *es* **A-estable** *si su dominio de estabilidad absoluta contiene al semiplano complejo negativo. A la inteserción de D con el eje real se le denomina* **intervalo de estabilidad absoluta** *(IEA)*

Nótese que, para los métodos RK explícitos, la matriz $I - z\mathcal{A}$ es triangular inferior con unos en la diagonal, por lo que $R(z)$ es un polinomio de grado $\leq m$; sin embargo, en general, será una función racional, cociente de dos polinomios de grado $\leq m$.

Se puede probar rápidamente que la elección de h tal que $h\lambda = z \in D$ implica también la estabilidad en la propagación de errores, tal como se estudió en un párrafo anterior, cuando el método se aplica a la ecuación de prueba. Una propiedad interesante de la función de amplificación (ver demostración en [0]) es la siguiente:

Teorema 33 *Si el método RK tiene orden p, $R(z)$ es una aproximación de orden $p+1$ de la exponencial, es decir,*

$$R(z) = e^z + \mathcal{O}(z^{p+1}), \quad z \to 0$$

A continuación se estudiará el dominio de estabilidad para los RK explícitos. Su función de amplificación $R(z)$ es un polinomio, por lo que $|R(z)| \to +\infty$ cuando $|z| \to 0$, en consecuencia, el dominio de estabilidad es finito y

"ningún método RK explícito puede ser A-estable."

Por otro lado, si el método es de orden p, $R(z)$ debe coincidir con la exponencial en sus primeros términos, por lo que se presentan dos casos:

a) $m = p$. En este caso,

$$R(z) = 1 + z + \frac{z^2}{2} + ... + \frac{z^p}{p!}$$

es decir, $R(z)$ está determinada univocamente por el número de etapas. Así, para $m = 2$, resulta

$$D = \{z \in \mathbb{C} \,|\, |1 + z + \frac{z^2}{2}| \leq 1\},$$

cuyo IEA es $[-2, 0]$.

Análogamente, para los RK de orden 3 y tres etapas se encuentra el IEA $= [-2.51, 0]$ y para los RK de orden 4 y cuatro etapas se encuentra el IEA $= [-2.78, 0]$. Normalmente, la localización de la frontera de D suele ser complicada y se hace frecuentemente por procedimientos numéricos.

b) $m > p$. En este caso,

$$R(z) = 1 + z + \frac{z^2}{2} + ... + \frac{z^p}{p} + \beta_{p+1} z^{p+1} + ... + \beta_m z^m$$

donde los coeficientes β dependen de los coeficientes del método RK que se escogen de modo que el IEA sea lo mayor posible.

Burden, R. and Faires, J. D. : *Análisis numérico*. PWS-Kent Publishing Co., 1993.

Calvo, M.; Montijano, J. y Rández, L.: *Análisis numérico: Métodos Runge-Kutta*. Universidad de Zaragoza,1990.

Mathews, J.: *Numerical Methods for Mathematics, Science and Engineering*. Prentice-Hall Intern., 1992.

Stoer, J. and Bulirsch, R.: *Introduction to Numerical Analysis*. Springer-Verlag, Berlin, 1980.

Scheid, F. and DiConstanzo, R.E.: *Métodos numéricos*. Colec. Schaum, McGraw-Hill, 1991.