

2 Métodos monopaso

2.1 Introducción

En este capítulo se van a considerar diversos métodos monopaso para la aproximación numérica de la solución del problema de valor inicial

$$\begin{cases} x'(t) = f(t, x(t)), & t \in [t_0, T] \\ x(t_0) = \xi_0, \end{cases} \quad (2.1)$$

donde $\xi_0 \in \mathbb{R}^n$ y $t_0, T \in \mathbb{R}$ con $t_0 < T$. Recuérdese que, como se comentó en la Observación 1.16, supondremos que la función f es continua y lipschitziana en la segunda variable con constante de Lipschitz $L > 0$, por lo que el problema (2.1) tiene existencia y unicidad de solución. Además, siguiendo lo visto en la Sección 1.5, dado $N \in \mathbb{N}$ y tomando como paso $h = \frac{T-t_0}{N}$, dividiremos el intervalo $[t_0, T]$ en N partes iguales por medio de puntos igualmente espaciados

$$t_i = t_0 + ih, \quad i = 0, 1, \dots, N. \quad (2.2)$$

2.2 Método de Euler y sus variantes

2.2.1 Método de Euler

El *Método de Euler* (explícito), también conocido como *Método de Euler progresivo* (Forward Euler Method), es el método más simple para resolver el problema de valor inicial (2.1). Teniendo en cuenta que

$$x'(t) = \lim_{h \rightarrow 0} \frac{x(t+h) - x(t)}{h},$$

para h suficientemente pequeño se verifica que

$$x'(t) \simeq \frac{x(t+h) - x(t)}{h} \in \mathbb{R}^n,$$

por lo que podemos aproximar

$$x'(t_i) = f(t_i, x(t_i))$$

por

$$\frac{x(t_{i+1}) - x(t_i)}{h} \simeq f(t_i, x(t_i)).$$

De esta forma, podemos tomar

$$\frac{x_{i+1} - x_i}{h} = f(t_i, x_i), \quad (2.3)$$

donde x_i será una aproximación de $x(t_i)$ (valor de la solución exacta $x = x(t)$ del problema (2.1) en el punto t_i) en principio, tanto mejor cuanto menor sea el paso h . Para calcular la solución de (2.3) basta utilizar el siguiente esquema numérico:

Método de Euler (explícito)	
<u>Paso 1</u>	$x_0 \simeq \xi_0$
<u>Paso 2</u>	Para $i = 0, 1, \dots, N-1$ $x_{i+1} = x_i + hf(t_i, x_i).$

Teniendo en cuenta los conceptos definidos en la Sección 1.6, es claro que este método es monopaso y explícito.

A continuación se presenta un algoritmo para la implementación del Método de Euler:

Algoritmo para resolver (2.1) por el Método de Euler (explícito)	
ENTRADA	f (función $f(t, \xi)$), t_0 (instante inicial), T (instante final), x_0 (valor inicial), N (número de puntos de la malla menos 1)
SALIDA	t (vector con los nodos de la malla $\{t_i\}_{i=0}^N$) x (matriz con los vectores $\{x_i\}_{i=0}^N$)
<u>Paso 1</u>	$h = (T - t_0) / N$ (paso de la discretización)
<u>Paso 2</u>	(Inicialización) $t(0) = t_0$ (inicializa t) $x(0) = x_0$ (inicializa x)
<u>Paso 3</u>	Para $i = 0, 1, \dots, N-1$ $t(i+1) = t(i) + h$ (siguiente valor de t) $x(i+1) = x(i) + hf(t(i), x(i))$ (siguiente valor de x)

Nótese que el algoritmo anterior está escrito en un pseudocódigo general que permite ser implementado, con las adecuadas modificaciones, en la mayoría de lenguajes de programación.

2.2.2 Interpretación geométrica del Método de Euler

Suponiendo que $n = 1$ y $x_0 = x(t_0) = \xi_0$ (nótese que no siempre puede tomarse x_0 como el valor exacto ξ_0), la ecuación de la recta tangente a la curva $x(t)$ en el punto (t_0, x_0) viene dada por

$$y(t) = x_0 + x'(t_0)(t - t_0) = x_0 + (t - t_0)f(t_0, x_0).$$

Al sustituir en $t = t_1$, se obtiene

$$y(t_1) = x_0 + (t_1 - t_0)f(t_0, x_0) = x_0 + hf(t_0, x_0) = x_1.$$

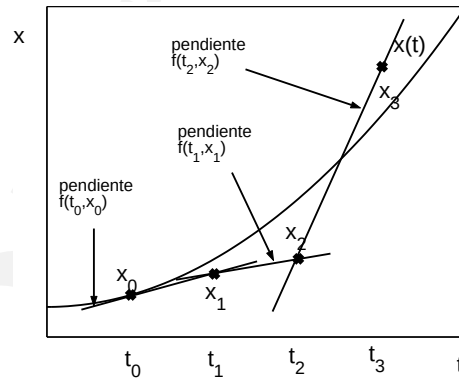


Figura 2.1: Interpretación geométrica del Método de Euler en dimensión 1.

De manera similar, para obtener el siguiente valor x_2 se considera la solución $\tilde{x}(t)$ del problema de valor inicial

$$\begin{cases} \tilde{x}'(t) = f(t, \tilde{x}(t)), & t \in [t_1, T] \\ \tilde{x}(t_1) = x_1 \end{cases}$$

(nótese que $\tilde{x}(t) \neq x(t)$, salvo en el caso de que $x_1 = x(t_1)$). Entonces, si $y(t)$ es la ecuación de la recta tangente a la curva $\tilde{x}(t)$ en el punto (t_1, x_1) , es decir,

$$y(t) = x_1 + \tilde{x}'(t_1)(t - t_1) = x_1 + (t - t_1)f(t_1, x_1),$$

al sustituir en $t = t_2$ se obtiene

$$y(t_2) = x_1 + hf(t_1, x_1) = x_2.$$

Argumentando de forma análoga se obtienen los restantes valores $\{x_3, x_4, \dots, x_N\}$. Uniendo sucesivamente los puntos $(t_0, x_0), (t_1, x_1), \dots, (t_N, x_N)$ mediante segmentos, se obtiene la denominada *poligonal de Euler* (1.14). La Figura 2.1 muestra un gráfico de esta interpretación geométrica.

Ejemplo 2.1 En la Tabla 2.1 se muestran los resultados (y, encima, la representación gráfica) que se obtienen al aplicar el Método de Euler con paso $h = 0.1$ para aproximar la solución del problema de valor inicial

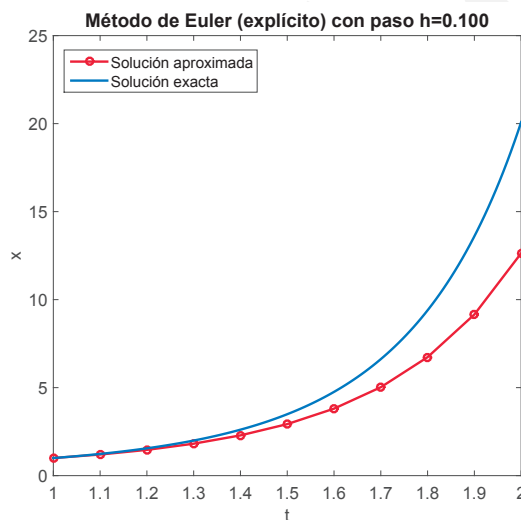
$$\begin{cases} x'(t) = 2tx(t), & t \in [1, 2] \\ x(1) = 1, \end{cases}$$

que tiene por solución exacta a

$$x(t) = e^{t^2-1}.$$

En la columna “error absoluto” se encuentran los valores $|x(t_i) - x_i|$ y en la de “error relativo (%)” las cantidades

$$100 \times \left| \frac{x(t_i) - x_i}{x(t_i)} \right|. \quad \square$$



Punto	Método de Euler	Valor exacto	Error absoluto	Error relativo (%)
1'0000	1'0000	1'0000	0'0000	0'0000
1'1000	1'2000	1'2337	0'0337	2'7299
1'2000	1'4640	1'5527	0'0887	5'7131
1'3000	1'8154	1'9937	0'1784	8'9459
1'4000	2'2874	2'6117	0'3243	12'4189
1'5000	2'9278	3'4903	0'5625	16'1168
1'6000	3'8062	4'7588	0'9527	20'0189
1'7000	5'0241	6'6194	1'5952	24'0996
1'8000	6'7323	9'3933	2'6610	28'3286
1'9000	9'1560	13'5991	4'4431	32'6720
2'0000	12'6352	20'0855	7'4503	37'0929

Tabla 2.1: Método de Euler (explícito) con paso $h = 0.1$.

2.2.3 θ -Método

Otra manera de deducir el Método de Euler consiste en integrar la ecuación diferencial del problema (2.1) entre dos nodos consecutivos t_i y t_{i+1} , con lo que se obtiene

$$x(t_{i+1}) - x(t_i) = \int_{t_i}^{t_{i+1}} x'(s) ds = \int_{t_i}^{t_{i+1}} f(s, x(s)) ds \quad (\in \mathbb{R}^n).$$

Utilizando la *Regla del Rectángulo* de integración numérica

$$\int_{t_i}^{t_{i+1}} g(s) ds \simeq (t_{i+1} - t_i)g(t_i) = hg(t_i)$$

aplicada a la función $g(s) = f(s, x(s))$ se obtiene, de nuevo, el *Método de Euler* (explícito).

Es claro que se podría haber utilizado cualquier otra regla de integración numérica. Por ejemplo, dado $\theta \in [0, 1]$, podemos considerar la regla

$$\int_{t_i}^{t_{i+1}} g(s) ds \simeq h((1-\theta)g(t_i) + \theta g(t_{i+1})) \quad (2.4)$$

que aproxima la integral de la función g entre las abscisas t_i y t_{i+1} por el área de un rectángulo de base h y altura una combinación lineal convexa entre los valores que toma g en los puntos t_i y t_{i+1} (véase la Figura 2.2).

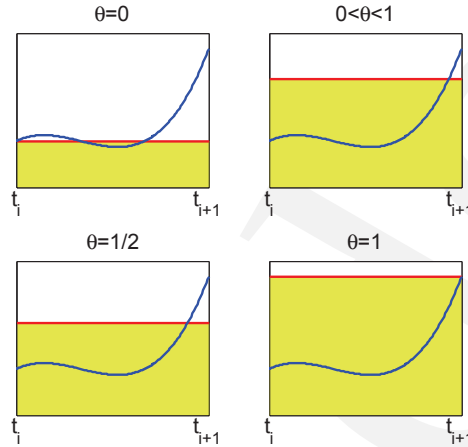


Figura 2.2: Interpretación geométrica de la integración aproximada (2.4) en función del valor del parámetro $\theta \in [0, 1]$.

De esta forma, al aplicar la regla (2.4) a la función $g(s) = f(s, x(s))$, se obtiene que

$$x(t_{i+1}) - x(t_i) \simeq h((1-\theta)f(t_i, x(t_i)) + \theta f(t_{i+1}, x(t_{i+1})))$$

y, por tanto, podemos tomar

$$x_{i+1} - x_i = h((1-\theta)f(t_i, x_i) + \theta f(t_{i+1}, x_{i+1})),$$

donde x_i será una aproximación de $x(t_i)$ (valor de la solución exacta $x = x(t)$ del problema (2.1) en el punto t_i). Se obtiene así la siguiente familia de esquemas numéricos conocidos como θ -Métodos:

θ -Método	
<u>Paso 1</u>	$x_0 \simeq \xi_0$
<u>Paso 2</u>	Para $i = 0, 1, \dots, N-1$ $x_{i+1} = x_i + h((1-\theta)f(t_i, x_i) + \theta f(t_{i+1}, x_{i+1})).$

(2.5)

A continuación se presenta un algoritmo para la implementación del θ -Método:

Algoritmo para resolver (2.1) por el θ -Método	
ENTRADA	f (función $f(t, \xi)$), t_0 (instante inicial), T (instante final), x_0 (valor inicial), N (número de puntos de la malla menos 1), θ (valor del parámetro)
SALIDA	t (vector con los nodos de la malla $\{t_i\}_{i=0}^N$) x (matriz con los vectores $\{x_i\}_{i=0}^N$)
<u>Paso 1</u>	$h = (T - t_0) / N$ (paso de la discretización)
<u>Paso 2</u>	(Inicialización) $t(0) = t_0$ (inicializa t) $x(0) = x_0$ (inicializa x)
<u>Paso 3</u>	Para $i = 0, 1, \dots, N-1$ $t(i+1) = t(i) + h$ (siguiente valor de t) $x(i+1) = x(i) + h((1-\theta)f(t(i), x(i)) + \theta f(t(i+1), x(i+1)))$ (siguiente valor de x)

Casos particulares (véase la Figura 2.2):

- a) Para $\theta = 0$ se recupera el *Método de Euler* (explícito).
b) Para $\theta = 1$ se obtiene *Método de Euler implícito*, también conocido como *Método de Euler regresivo* (Backward Euler Method)

Método de Euler implícito	
<u>Paso 1</u>	$x_0 \simeq \xi_0$
<u>Paso 2</u>	Para $i = 0, 1, \dots, N - 1$ $x_{i+1} = x_i + hf(t_{i+1}, x_{i+1})$.

- c) Para $\theta = \frac{1}{2}$ se está aplicando la *fórmula del Trapecio* de integración numérica (véase [Infante-Rey]), por lo que se obtiene

Método del Trapecio (o Método de Crank–Nicolson trapezoidal)	
<u>Paso 1</u>	$x_0 \simeq \xi_0$
<u>Paso 2</u>	Para $i = 0, 1, \dots, N - 1$ $x_{i+1} = x_i + \frac{h}{2} (f(t_i, x_i) + f(t_{i+1}, x_{i+1}))$.

que es una especie de “promedio” entre los Métodos de Euler explícito e implícito.

Nótese que para $\theta \in (0, 1]$ se obtiene un método monopaso implícito: el valor x_{i+1} se obtiene a partir de x_i resolviendo el sistema de ecuaciones en y (en general, no lineal)

$$y = x_i + h((1 - \theta)f(t_i, x_i) + \theta f(t_{i+1}, y)). \quad (2.6)$$

¿Cómo garantizar la existencia y unicidad de soluciones de la ecuación (2.6)? El siguiente teorema es de gran utilidad a ese respecto.

Teorema 2.2 (Punto Fijo de Banach (o de la aplicación contractiva)) Sea $T : \mathbb{R}^n \rightarrow \mathbb{R}^n$ una aplicación contractiva sobre $\mathbb{R}^n$, es decir, existe un número real $k \in [0, 1)$ tal que

$$\|T(x) - T(y)\| \leq k \|x - y\|, \quad x, y \in \mathbb{R}^n.$$

Entonces T admite un único punto fijo $\tilde{x} \in \mathbb{R}^n$ (es decir, $T(\tilde{x}) = \tilde{x}$). Además, este punto fijo puede encontrarse como límite de la sucesión

$$\begin{cases} x_0 \in \mathbb{R}^n \text{ arbitrario} \\ x_i = T(x_{i-1}), \quad i \in \mathbb{N} \end{cases}$$

y , para todo $i \in \mathbb{N}$, se tiene la siguiente estimación del error cuando se aproxima $\tilde{x}$ por x_i :

$$\|\tilde{x} - x_i\| \leq \frac{k^i}{1 - k} \|x_1 - x_0\|.$$

DEMOSTRACIÓN. Véase [Guzmán, Página 75]. $\square$

Observación 2.3 Consideremos, para cada $i = 0, 1, \dots, N - 1$, la función

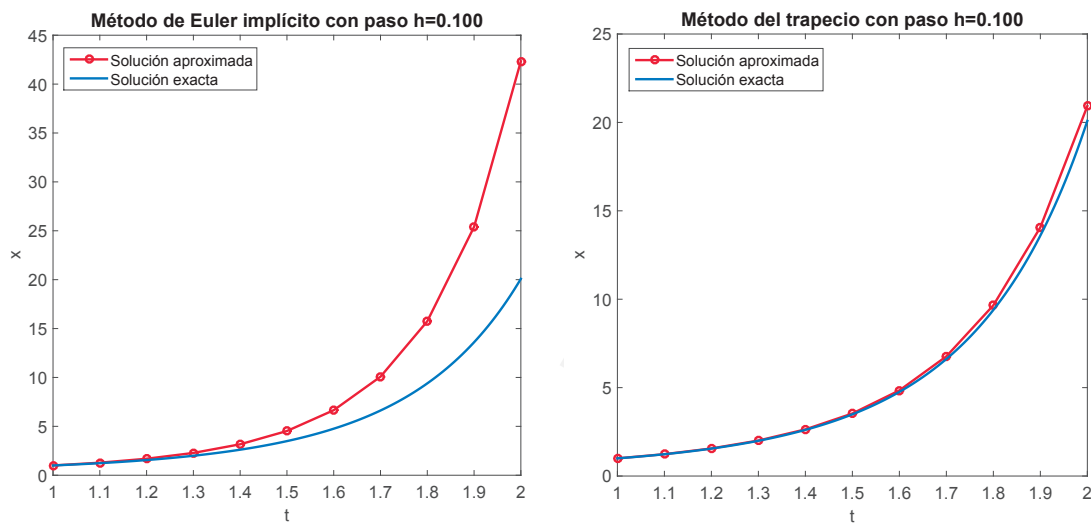
$$T_i(y) = x_i + h((1 - \theta)f(t_i, x_i) + \theta f(t_{i+1}, y))$$

con $\theta \in (0, 1]$. Si la función f es lipschitziana en la segunda variable con constante de Lipschitz $L > 0$, se tiene que

$$\|T_i(y_1) - T_i(y_2)\| = h\theta \|f(t_{i+1}, y_1) - f(t_{i+1}, y_2)\| \leq h\theta L \|y_1 - y_2\|, \quad \forall y_1, y_2 \in \mathbb{R}^n,$$

por lo que eligiendo h suficientemente pequeño (en concreto, basta tomar $0 < h < \frac{1}{\theta L}$), se verifica que cada función T_i es contractiva en $\mathbb{R}^n$ de constante $k = h\theta L$. Por tanto, por el Teorema 2.2 se deduce que el sistema de ecuaciones (2.6) tiene una única solución x_{i+1} . $\square$

Ejemplo 2.4 Regresando al problema de valor inicial considerado en el Ejemplo 2.1, en la Tabla 2.2 se muestran los resultados que se obtienen (y, encima, la representación gráfica) al aplicar, respectivamente, el *Método de Euler implícito* y el *Método del Trapecio* con paso $h = 0.100$. Como se aprecia, el Método del Trapecio aproxima (en este caso) “mucho mejor” la solución del problema de valor inicial que los métodos de Euler (tanto explícito como implícito). □



Punto	Método de Euler implícito	Valor exacto	Error absoluto	Error relativo (%)
1'0000	1'0000	1'0000	0'0000	0'0000
1'1000	1'2821	1'2337	0'0484	3'9211
1'2000	1'6869	1'5527	0'1314	8'6431
1'3000	2'2796	1'9937	0'2859	14'3397
1'4000	3'1661	2'6117	0'5544	21'2285
1'5000	4'5230	3'4903	1'0327	29'5870
1'6000	6'6515	4'7588	1'8927	39'7723
1'7000	10'0781	6'6194	3'4587	52'2510
1'8000	15'7470	9'3933	6'3536	67'6398
1'9000	25'3983	13'5991	11'7993	86'7654
2'0000	42'3305	20'0855	22'2450	110'7513

Punto	Método del Trapecio	Valor exacto	Error absoluto	Error relativo (%)
1'0000	1'0000	1'0000	0'0000	0'0000
1'1000	1'2360	1'2337	0'0023	0'1846
1'2000	1'5590	1'5527	0'0063	0'4046
1'3000	2'0070	1'9937	0'0133	0'6650
1'4000	2'6371	2'6117	0'0254	0'9716
1'5000	3'5368	3'4903	0'0464	1'3304
1'6000	4'8420	4'7588	0'0832	1'7482
1'7000	6'7672	6'6194	0'1478	2'2326
1'8000	9'6556	9'3933	0'2622	2'7918
1'9000	14'0661	13'5991	0'4671	3'4348
2'0000	20'9234	20'0855	0'8379	4'1714

Tabla 2.2: Métodos de Euler implícito y del Trapecio con paso $h = 0.1$.

2.2.4 Análisis del error cometido en el θ -Método

En esta sección vamos a estudiar los errores cometidos en la aproximación numérica de la solución del problema de valor inicial (2.1) mediante el θ -Método. Para ello utilizaremos las nociones sobre errores expuestas en la Sección 1.5.

En el θ -Método, el *error local de truncamiento* $\tau(t, h)$, asociado a una solución x de la ecuación diferencial

$$x'(t) = f(t, x(t)), \quad t \in [t_0, T], \quad (2.7)$$

a partir del instante $t \in [t_0, T)$, viene dado por

$$\begin{aligned}\tau(t, h) &= \frac{x(t+h) - x(t)}{h} - ((1-\theta)f(t, x(t)) + \theta f(t+h, x(t+h))) \\ &= \frac{x(t+h) - x(t)}{h} - ((1-\theta)x'(t) + \theta x'(t+h))\end{aligned}\quad (2.8)$$

para todo $h \in (0, T-t]$. Recordamos que, tal y como se ha explicado en la Sección 1.5, la terminología “error local de truncamiento” se debe a que indica el error local que se comete al aproximar una expresión con un término que incluye un proceso de paso al límite, como es la derivada

$$x'(t) = \lim_{h \rightarrow 0} \frac{x(t+h) - x(t)}{h},$$

por otra expresión sin este tipo de términos, como son los cocientes incrementales

$$x'(t) \simeq \frac{x(t+h) - x(t)}{h}.$$

Por el *Teorema de Taylor*, si la función $x(t) = (x^1(t), x^2(t), \dots, x^n(t))^T$ es de clase $\mathcal{C}^3$ (según se ha visto en la Sección 1.3 basta con que $f \in \mathcal{C}^2$), para todo $t \in [t_0, T-h]$ se verifican las relaciones

$$x(t+h) = x(t) + hx'(t) + \frac{h^2}{2}x''(t) + \frac{h^3}{6} \begin{pmatrix} (x^1)'''(\zeta^1) \\ (x^2)'''(\zeta^2) \\ \dots \\ (x^n)'''(\zeta^n) \end{pmatrix} \quad (2.9)$$

y

$$x'(t+h) = x'(t) + hx''(t) + \frac{h^2}{2} \begin{pmatrix} (x^1)'''(\vartheta^1) \\ (x^2)'''(\vartheta^2) \\ \dots \\ (x^n)'''(\vartheta^n) \end{pmatrix}, \quad (2.10)$$

donde $\zeta^k, \vartheta^k \in (t, t+h)$ para todo $k = 1, 2, \dots, n$. De esta forma, la relación (2.8) determina que

$$\begin{aligned}\tau(t, h) &= x'(t) + \frac{h}{2}x''(t) + \frac{h^2}{6} \begin{pmatrix} (x^1)'''(\zeta^1) \\ (x^2)'''(\zeta^2) \\ \dots \\ (x^n)'''(\zeta^n) \end{pmatrix} - (1-\theta)x'(t) \\ &\quad - \theta \left(x'(t) + hx''(t) + \frac{h^2}{2} \begin{pmatrix} (x^1)'''(\vartheta^1) \\ (x^2)'''(\vartheta^2) \\ \dots \\ (x^n)'''(\vartheta^n) \end{pmatrix} \right) \\ &= \left(\frac{1}{2} - \theta \right) hx''(t) + \frac{1}{6} \left(\begin{pmatrix} (x^1)'''(\zeta^1) \\ (x^2)'''(\zeta^2) \\ \dots \\ (x^n)'''(\zeta^n) \end{pmatrix} - 3\theta \begin{pmatrix} (x^1)'''(\vartheta^1) \\ (x^2)'''(\vartheta^2) \\ \dots \\ (x^n)'''(\vartheta^n) \end{pmatrix} \right) h^2.\end{aligned}\quad (2.11)$$

Así pues, para todo $t \in [t_0, T-h]$, se verifica que

$$\begin{aligned}\|\tau(t, h)\| &\leq c \left| \frac{1}{2} - \theta \right| \|x''(t)\|_\infty h + \frac{c}{6} \left(\left\| \begin{pmatrix} (x^1)'''(\zeta^1) \\ (x^2)'''(\zeta^2) \\ \dots \\ (x^n)'''(\zeta^n) \end{pmatrix} \right\|_\infty + 3\theta \left\| \begin{pmatrix} (x^1)'''(\vartheta^1) \\ (x^2)'''(\vartheta^2) \\ \dots \\ (x^n)'''(\vartheta^n) \end{pmatrix} \right\|_\infty \right) h^2 \\ &\leq c \left| \frac{1}{2} - \theta \right| M_2 h + \frac{(1+3\theta)cM_3}{6} h^2,\end{aligned}$$

siendo

$$M_2 = \max_{t \in [t_0, T]} \|x''(t)\|_\infty, \quad (2.12)$$

$$M_3 = \max_{t \in [t_0, T]} \|x'''(t)\|_\infty \quad (2.13)$$

y $c > 0$ una constante positiva¹ verificando que

$$\|\zeta\| \leq c \|\zeta\|_\infty \quad (2.14)$$

para todo $\zeta \in \mathbb{R}^n$. De esta forma, si definimos el *error local de truncamiento* asociado a la solución $x(t)$ como la función $\bar{\tau} : (0, T - t_0] \rightarrow \mathbb{R}$ dada por

$$\bar{\tau}(h) = \max_{t \in [t_0, T-h]} \|\tau(t, h)\|, \quad (2.15)$$

se verifica que

$$\bar{\tau}(h) \leq c \left(\left| \frac{1}{2} - \theta \right| M_2 h + \frac{(1 + 3\theta)M_3}{6} h^2 \right). \quad (2.16)$$

Observación 2.5 Nótese que si sólo se tiene la regularidad $x \in \mathcal{C}^2$ (para lo cual basta con que $f \in \mathcal{C}^1$), para todo $\theta \in [0, 1]$ se verifica que

$$\tau(t, h) = \left(\frac{1}{2} \begin{pmatrix} (x^1)''(\zeta^1) \\ (x^2)''(\zeta^2) \\ \vdots \\ (x^n)''(\zeta^n) \end{pmatrix} - \theta \begin{pmatrix} (x^1)''(\vartheta^1) \\ (x^2)''(\vartheta^2) \\ \vdots \\ (x^n)''(\vartheta^n) \end{pmatrix} \right) h \quad (2.17)$$

con $\zeta^k, \vartheta^k \in (t, t + h)$ para todo $k = 1, 2, \dots, n$ (ver hojas de problemas) y, por tanto,

$$\bar{\tau}(h) \leq c \left(\frac{1}{2} + \theta \right) M_2 h. \quad \square \quad (2.18)$$

Observación 2.6 En los problemas considerados no se utilizan valores arbitrarios de $h \in (0, T - t_0]$, sino valores de h pertenecientes al conjunto

$$\mathcal{H} = \left\{ \frac{T - t_0}{N} : N \in \mathbb{N} \right\}. \quad \square \quad (2.19)$$

Observación 2.7 La función $\tau(t, h)$ está definida para todo $(t, h) \in \mathcal{U}$ (es decir, $\tau : \mathcal{U} \rightarrow \mathbb{R}^n$), siendo $\mathcal{U} \subset \mathbb{R}^2$ el conjunto (véase la Figura 2.3)

$$\mathcal{U} = \{(t, h) \in \mathbb{R}^2 : h > 0, t, t + h \in [t_0, T]\} = \bigcup_{t \in [t_0, T]} \{t\} \times (0, T - t] = \bigcup_{h \in (0, T - t_0]} [t_0, T - h] \times \{h\} \quad (2.20)$$

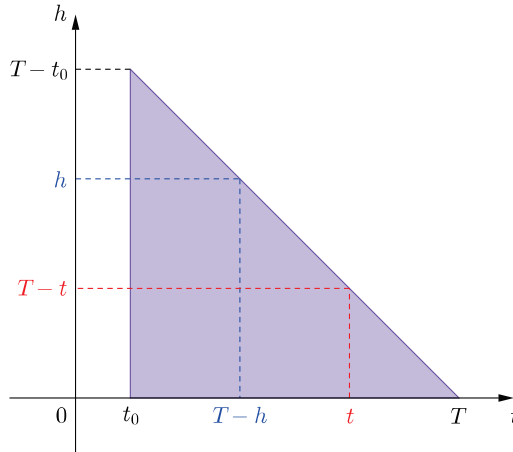


Figura 2.3: Conjunto $\mathcal{U}$.

(nótese que $\mathcal{U} \subset [t_0, T] \times [0, T - t_0]$). Es claro que si $f \in \mathcal{C}([t_0, T] \times \mathbb{R}^n; \mathbb{R}^n)$, la función $\tau \in \mathcal{C}(\mathcal{U})$. Además, definiendo

$$\tau(t, 0) = \lim_{\substack{(s, h) \rightarrow (t, 0) \\ (s, h) \in \mathcal{U}}} \tau(s, h) = x'(t) - f(t, x(t)) = 0$$

para todo $t \in [t_0, T]$, se tiene que la función $\tau \in \mathcal{C}(\bar{\mathcal{U}})$ y, por tanto, τ es uniformemente continua en $\bar{\mathcal{U}}$.

Geométricamente, $\bar{\tau}(h)$ es el máximo de la norma de la función τ sobre el segmento horizontal de $\mathcal{U}$ a altura h (véase la Figura 2.3). $\square$

¹Esta constante existe debido a la equivalencia de las normas $\|\cdot\|$ y $\|\cdot\|_\infty$ en $\mathbb{R}^n$.

Vamos a describir de forma rigurosa varios conceptos (y su correspondiente notación) que sirven para comparar los tamaños de dos funciones cerca de un punto:

Definición 2.8 (O de Landau) Sean $n \in \mathbb{N}$, $a, b \in \mathbb{R}$, $f : (a, b) \rightarrow \mathbb{R}^n$, $g : (a, b) \rightarrow \mathbb{R}$ y $h_0 \in (a, b)$. Se dice que

$$f(h) = O(g(h)) \text{ cuando } h \rightarrow h_0,$$

si existen $M > 0$ y $\delta > 0$ tales que

$$\|f(h)\|_\infty \leq M|g(h)| \text{ si } |h - h_0| < \delta.$$

Cuando $h_0 = 0$ suele escribirse, por simplicidad, $f(h) = O(g(h))$ (omitiendo “cuando $h \rightarrow 0$ ”). $\square$

Observación 2.9 Si existe el límite

$$\lim_{h \rightarrow h_0} \frac{\|f(h)\|_\infty}{|g(h)|}$$

y es finito, entonces se verifica que $f(h) = O(g(h))$ cuando $h \rightarrow h_0$. Nótese que el recíproco de esta propiedad, en general, no es cierto (por ejemplo, si tomamos $f(h) = \sin\left(\frac{1}{h}\right)$ y $g(h) = 1$, se verifica que $f(h) = O(g(h))$; en cambio,

$$\text{no existe } \lim_{h \rightarrow 0} \frac{|f(h)|}{|g(h)|} = \lim_{h \rightarrow 0} \left| \sin\left(\frac{1}{h}\right) \right|. \quad \square$$

Ejemplo 2.10 $4(h+1)^3 + 7(h+1)^4 = O((h+1)^3)$ cuando $h \rightarrow -1$ y $2 - 3h + 4h^2 = O(1) = 2 + O(h) = 2 - 3h + O(h^2)$. $\square$

Observación 2.11 Claramente, se verifican las siguientes propiedades:

- a) Toda función $O(h^r)$ es $O(h^s)$ si $r \geq s$.
- b) $O(h^r) + O(h^s) = O(h^p)$ con $p = \min\{r, s\}$.
- c) $O(h^r)O(h^s) = O(h^{r+s})$. $\square$

Observación 2.12 La Definición 2.8 puede extenderse al caso en el que el intervalo (a, b) se sustituye por el conjunto discreto $\mathcal{H}$ dado en (2.19), con $h_0 = 0$. Así, por ejemplo, si $n \in \mathbb{N}$, $f : \mathcal{H} \rightarrow \mathbb{R}^n$ y $g : \mathcal{H} \rightarrow \mathbb{R}$, diremos que $f(h) = O(g(h))$ si existe $M > 0$ tal que

$$\|f(h)\|_\infty \leq M|g(h)| \text{ para todo } h \in \mathcal{H}. \quad \square$$

Definición 2.13 Un método numérico es *consistente con una ecuación diferencial del tipo (2.7)* si el error local de truncamiento $\bar{\tau}(h)$ asociado a cualquier solución $x(t)$ de dicha ecuación tiene la propiedad de que

$$\lim_{h \rightarrow 0} \bar{\tau}(h) = 0.$$

Además, se dice que el método es de *orden de precisión* $p \in \mathbb{N}$ para una ecuación diferencial del tipo (2.7) si para cualquier solución $x(t)$ de dicha ecuación se verifica que

$$\bar{\tau}(h) = O(h^p). \quad \square$$

Observación 2.14

- a) A grandes rasgos, un método de aproximación numérica es consistente con la ecuación diferencial cuando el esquema numérico converge a dicha ecuación al hacer tender el paso h a cero.
- b) La noción de consistencia que se acaba de definir va ligada únicamente a ecuaciones diferenciales y no a problemas de valor inicial. Puesto que estamos estudiando el problema de valor inicial (2.1), requeriremos la continuidad de la función f y que sea lipschitziana en la segunda variable para, de esta forma, tener garantizada la existencia y unicidad de soluciones de dicho problema (véase la Observación 1.16).
- c) Obviamente, todo método de orden de precisión p para la ecuación diferencial (2.7) es de orden de precisión $q \leq p$ para la ecuación diferencial (2.7) y, en particular, es consistente con dicha ecuación. $\square$

Ejemplo 2.15 Teniendo en cuenta que la solución del problema de valor inicial

$$\begin{cases} x'(t) = \alpha t + \beta, & t \in [t_0, T] \\ x(t_0) = \gamma \end{cases}$$

donde $\alpha, \beta, \gamma \in \mathbb{R}$, es

$$x(t) = \frac{\alpha}{2}(t^2 - t_0^2) + \beta(t - t_0) + \gamma,$$

al aplicar el Método de Euler (explícito) se verifica que (compruébese)

$$\tau(h) = \frac{|\alpha|}{2} h = \begin{cases} O(h^r) & \text{para todo } r > 0 & \text{si } \alpha = 0 \\ O(h) & & \text{si } \alpha \neq 0. \quad \square \end{cases} \quad (2.21)$$

De esta forma, se tiene el siguiente resultado:

Teorema 2.16 Sea $\theta \in [0, 1]$.

- a) Si $f \in C^1([t_0, T] \times \mathbb{R}^n; \mathbb{R}^n)$ se verifica que el θ -Método es de orden de precisión 1 para la ecuación diferencial (2.7).
- b) Si $f \in C^2([t_0, T] \times \mathbb{R}^n; \mathbb{R}^n)$ se verifica que el Método del Trapecio es de orden de precisión 2 para la ecuación diferencial (2.7). $\square$

La consistencia y orden de precisión del θ -Método puede obtenerse con condiciones más débiles que las anteriores. En concreto, se tiene el siguiente resultado:

Teorema 2.17 Sea $\theta \in [0, 1]$.

- a) Si $f \in C([t_0, T] \times \mathbb{R}^n; \mathbb{R}^n)$ se verifica que el θ -Método es consistente con la ecuación diferencial (2.7).
- b) Si f es lipschitziana (en las dos variables), se verifica que el θ -Método es de orden de precisión 1 para la ecuación diferencial (2.7).
- c) Si $\theta = \frac{1}{2}$, $f \in C^1([t_0, T] \times \mathbb{R}^n; \mathbb{R}^n)$ y la función

$$f^{(1)}(t, \xi) = \frac{\partial f}{\partial t}(t, \xi) + D_\xi f(t, \xi)f(t, \xi) \quad (2.22)$$

(véase la Sección 1.3) es lipschitziana (en las dos variables), se verifica que el Método del Trapecio es de orden de precisión 2 para la ecuación diferencial (2.7).

DEMOSTRACIÓN. La equivalencia entre las normas $\|\cdot\|$ y $\|\cdot\|_\infty$ en $\mathbb{R}^n$ determina la existencia de dos constantes $c, C > 0$ tales que

$$C \|\zeta\|_\infty \leq \|\zeta\| \leq c \|\zeta\|_\infty \quad (2.23)$$

para todo $\zeta \in \mathbb{R}^n$. Sea $x(t)$ una solución arbitraria de la ecuación (2.7) (suponiendo que existe).

- a) Puesto que por el Teorema 1.18 se tiene que $x(t)$ que es de clase C^1 (por ser f continua), dado $t \in [t_0, T]$, mediante un desarrollo de Taylor de primer orden, para cada $h \in (0, T - t]$ (nótese, por tanto, que $(t, h) \in \mathcal{U}$), se verifica que

$$x(t+h) = x(t) + h \begin{pmatrix} (x^1)'(\zeta^1) \\ (x^2)'(\zeta^2) \\ \vdots \\ (x^n)'(\zeta^n) \end{pmatrix} = x(t) + h \begin{pmatrix} f^1(\zeta^1, x(\zeta^1)) \\ f^2(\zeta^2, x(\zeta^2)) \\ \vdots \\ f^n(\zeta^n, x(\zeta^n)) \end{pmatrix}$$

donde $\zeta^k \in (t, t+h)$ para todo $k = 1, 2, \dots, n$. Consecuentemente,

$$\begin{aligned}\tau(t, h) &= \frac{x(t+h) - x(t)}{h} - ((1-\theta)f(t, x(t)) + \theta f(t+h, x(t+h))) \\ &= \begin{pmatrix} f^1(\zeta^1, x(\zeta^1)) \\ f^2(\zeta^2, x(\zeta^2)) \\ \dots \\ f^n(\zeta^n, x(\zeta^n)) \end{pmatrix} - (1-\theta) \begin{pmatrix} f^1(t, x(t)) \\ f^2(t, x(t)) \\ \dots \\ f^n(t, x(t)) \end{pmatrix} - \theta \begin{pmatrix} f^1(t+h, x(t+h)) \\ f^2(t+h, x(t+h)) \\ \dots \\ f^n(t+h, x(t+h)) \end{pmatrix} \\ &= (1-\theta) \left[\begin{pmatrix} f^1(\zeta^1, x(\zeta^1)) \\ f^2(\zeta^2, x(\zeta^2)) \\ \dots \\ f^n(\zeta^n, x(\zeta^n)) \end{pmatrix} - \begin{pmatrix} f^1(t, x(t)) \\ f^2(t, x(t)) \\ \dots \\ f^n(t, x(t)) \end{pmatrix} \right] + \theta \left[\begin{pmatrix} f^1(\zeta^1, x(\zeta^1)) \\ f^2(\zeta^2, x(\zeta^2)) \\ \dots \\ f^n(\zeta^n, x(\zeta^n)) \end{pmatrix} - \begin{pmatrix} f^1(t+h, x(t+h)) \\ f^2(t+h, x(t+h)) \\ \dots \\ f^n(t+h, x(t+h)) \end{pmatrix} \right].\end{aligned}$$

Puesto que las funciones $f^k(\cdot, x(\cdot))$, con $k = 1, 2, \dots, n$, son continuas sobre el compacto $[t_0, T]$ y, por tanto, uniformemente continuas en $[t_0, T]$, dado $\varepsilon > 0$ existe $\delta = \delta(\varepsilon) > 0$ tal que

$$|f^k(t, x(t)) - f^k(s, x(s))| < \varepsilon \quad \text{si} \quad |t - s| < \delta. \quad (2.24)$$

Así pues, si se toma $h \in (0, \delta)$, se verifica que

$$\max_{k \in \{1, 2, \dots, n\}} |\zeta^k - t| < \delta \quad \text{y} \quad \max_{k \in \{1, 2, \dots, n\}} |\zeta^k - (t+h)| < \delta,$$

y, por tanto,

$$\begin{aligned}\|\tau(t, h)\| &\leq (1-\theta)c \left\| \begin{pmatrix} f^1(\zeta^1, x(\zeta^1)) - f^1(t, x(t)) \\ f^2(\zeta^2, x(\zeta^2)) - f^2(t, x(t)) \\ \dots \\ f^n(\zeta^n, x(\zeta^n)) - f^n(t, x(t)) \end{pmatrix} \right\|_{\infty} \\ &\quad + \theta c \left\| \begin{pmatrix} f^1(\zeta^1, x(\zeta^1)) - f^1(t+h, x(t+h)) \\ f^2(\zeta^2, x(\zeta^2)) - f^2(t+h, x(t+h)) \\ \dots \\ f^n(\zeta^n, x(\zeta^n)) - f^n(t+h, x(t+h)) \end{pmatrix} \right\|_{\infty} \\ &= (1-\theta)c \max_{k \in \{1, 2, \dots, n\}} |f^k(\zeta^k, x(\zeta^k)) - f^k(t, x(t))| \\ &\quad + \theta c \max_{k \in \{1, 2, \dots, n\}} |f^k(\zeta^k, x(\zeta^k)) - f^k(t+h, x(t+h))| \\ &< (1-\theta)c\varepsilon + \theta c\varepsilon = c\varepsilon\end{aligned} \quad (2.25)$$

para todo $(t, h) \in \mathcal{U}$ con $h \in (0, \delta)$. Por tanto, se verifica que

$$\lim_{h \rightarrow 0} \bar{\tau}(h) = 0,$$

lo que prueba la consistencia del método.

- b) Si, además, la función $f = (f^k)_{k=1}^n$ es lipschitziana en las dos variables con constante de Lipschitz $L > 0$, entonces cada componente f^k es, también, lipschitziana en las dos variables de constante $\frac{L}{C} > 0$ (véase la hoja de problemas). De esta forma, a partir de (2.25), se verifica que

$$\begin{aligned}\|\tau(t, h)\| &\leq (1-\theta)c \max_{k \in \{1, 2, \dots, n\}} |f^k(\zeta^k, x(\zeta^k)) - f^k(t, x(t))| \\ &\quad + \theta c \max_{k \in \{1, 2, \dots, n\}} |f^k(\zeta^k, x(\zeta^k)) - f^k(t+h, x(t+h))| \\ &\leq \frac{(1-\theta)cL}{C} \max_{k \in \{1, 2, \dots, n\}} (|\zeta^k - t| + \|x(\zeta^k) - x(t)\|) \\ &\quad + \frac{\theta cL}{C} \max_{k \in \{1, 2, \dots, n\}} (|\zeta^k - (t+h)| + \|x(\zeta^k) - x(t+h)\|).\end{aligned} \quad (2.26)$$

Puesto que para cada $k = 1, 2, \dots, n$ se verifica que

$$\begin{cases} |\zeta^k - t| < h \\ |\zeta^k - (t+h)| < h \end{cases}$$

y, por el teorema del Valor Medio,

$$\begin{cases} \|x(\zeta^k) - x(t)\| \leq c \|x(\zeta^k) - x(t)\|_\infty \leq cM_1|\zeta^k - t| < cM_1h, \\ \|x(\zeta^k) - x(t+h)\| \leq c \|x(\zeta^k) - x(t+h)\|_\infty \leq cM_1|\zeta^k - (t+h)| < cM_1h, \end{cases}$$

donde

$$M_1 = \max_{t \in [t_0, T]} \|x'(t)\|_\infty < +\infty, \quad (2.27)$$

se verifica que

$$\begin{cases} \max_{k \in \{1, 2, \dots, n\}} (|\zeta^k - t| + \|x(\zeta^k) - x(t)\|) < (1 + cM_1)h \\ \max_{k \in \{1, 2, \dots, n\}} (|\zeta^k - (t+h)| + \|x(\zeta^k) - x(t+h)\|) < (1 + cM_1)h. \end{cases}$$

De esta forma, regresando a la expresión (2.26), se concluye que

$$\|\tau(t, h)\| \leq \frac{cL(1 + cM_1)}{C} h$$

para todo $(t, h) \in \mathcal{U}$, lo que indica que

$$\bar{\tau}(h) = O(h).$$

- c) Puesto que por el Teorema 1.18 se tiene que $x(t)$ que es de clase $\mathcal{C}^2$ (por ser $f \in \mathcal{C}^1$), a partir de (2.17) con $\theta = \frac{1}{2}$, para cada $(t, h) \in \mathcal{U}$, se verifica que existen $\zeta^k, \vartheta^k \in (t, t+h)$ para todo $k = 1, 2, \dots, n$, tales que

$$\tau(t, h) = \frac{1}{2} \begin{pmatrix} (f^{(1)})^1(\zeta^1, x(\zeta^1)) - (f^{(1)})^1(\vartheta^1, x(\vartheta^1)) \\ (f^{(1)})^2(\zeta^2, x(\zeta^2)) - (f^{(1)})^2(\vartheta^2, x(\vartheta^2)) \\ \dots \\ (f^{(1)})^n(\zeta^n, x(\zeta^n)) - (f^{(1)})^n(\vartheta^n, x(\vartheta^n)) \end{pmatrix} h,$$

donde se ha tenido en cuenta que para todo $t \in [t_0, T]$ se verifica que

$$x''(t) = \frac{\partial f}{\partial t}(t, x(t)) + D_\xi f(t, x(t))x'(t) = \frac{\partial f}{\partial t}(t, x(t)) + D_\xi f(t, x(t))f(t, x(t)) = f^{(1)}(t, x(t)).$$

Por tanto,

$$\begin{aligned} \|\tau(t, h)\| &\leq \frac{ch}{2} \left\| \begin{pmatrix} (f^{(1)})^1(\zeta^1, x(\zeta^1)) - (f^{(1)})^1(\vartheta^1, x(\vartheta^1)) \\ (f^{(1)})^2(\zeta^2, x(\zeta^2)) - (f^{(1)})^2(\vartheta^2, x(\vartheta^2)) \\ \dots \\ (f^{(1)})^n(\zeta^n, x(\zeta^n)) - (f^{(1)})^n(\vartheta^n, x(\vartheta^n)) \end{pmatrix} \right\|_\infty \\ &= \frac{ch}{2} \max_{k \in \{1, 2, \dots, n\}} |(f^{(1)})^k(\zeta^k, x(\zeta^k)) - (f^{(1)})^k(\vartheta^k, x(\vartheta^k))| \end{aligned} \quad (2.28)$$

para todo $(t, h) \in \mathcal{U}$. Puesto que la función $f^{(1)} = ((f^{(1)})^k)_{k=1}^n$ es lipschitziana en las dos variables con constante de Lipschitz $L_1 > 0$, entonces cada componente $(f^{(1)})^k$ es, también, lipschitziana en las dos variables con constante de Lipschitz $\frac{L_1}{C} > 0$. De esta forma, a partir de (2.28), se verifica que

$$\|\tau(t, h)\| \leq \frac{cL_1h}{2C} \max_{k \in \{1, 2, \dots, n\}} (|\zeta^k - \vartheta^k| + \|x(\zeta^k) - x(\vartheta^k)\|). \quad (2.29)$$

Puesto que para cada $k = 1, 2, \dots, n$ se verifica que

$$|\zeta^k - \vartheta^k| < h$$

y, por el teorema del Valor Medio,

$$\|x(\zeta^k) - x(\vartheta^k)\| \leq c \|x(\zeta^k) - x(\vartheta^k)\|_\infty \leq cM_1|\zeta^k - \vartheta^k| < cM_1h,$$

donde M_1 está definido en (2.27), se tiene que

$$\max_{k \in \{1, 2, \dots, n\}} (|\zeta^k - \vartheta^k| + \|x(\zeta^k) - x(\vartheta^k)\|) < (1 + cM_1)h.$$

De esta forma, regresando a la expresión (2.29), se concluye que

$$\|\tau(t, h)\| \leq \frac{cL_1(1 + cM_1)}{2C} h^2$$

para todo $(t, h) \in \mathcal{U}$, lo que indica que

$$\bar{\tau}(h) = O(h^2). \quad \square$$

A la vista de los resultados obtenidos en los Teoremas 2.16 y 2.17, y con el objeto de simplificar sus enunciados, vamos a considerar la siguiente definición.

Definición 2.18 Un método numérico es *consistente* si es consistente para cualquier ecuación diferencial del tipo (2.7) con $f \in \mathcal{C}([t_0, T] \times \mathbb{R}^n; \mathbb{R}^n)$. Además, el método es de *orden de precisión* $p \in \mathbb{N}$ si lo es para cualquier ecuación diferencial del tipo (2.7) con $f \in \mathcal{C}^p([t_0, T] \times \mathbb{R}^n; \mathbb{R}^n)$. $\square$

Con la noción anterior se tiene el siguiente resultado:

Corolario 2.19 Sea $\theta \in [0, 1]$. El θ -Método es consistente y de orden de precisión $\begin{cases} 1 & \text{si } \theta \neq \frac{1}{2} \\ 2 & \text{si } \theta = \frac{1}{2}. \end{cases} \square$

Supongamos a continuación que la función f es continua y lipschitziana en la segunda variable, con lo que se tiene garantizada la existencia y unicidad de soluciones del problema de valor inicial (2.1) (véase la Observación 1.16). Tiene sentido, por tanto, preguntarse por el error cometido al aproximar la (única) solución de dicho problema.

Definición 2.20 Dado $h = \frac{T-t_0}{N} \in \mathcal{H}$, definimos el *error de discretización global absoluto* como

$$\varepsilon(h) = \max_{i \in \{0, 1, \dots, N\}} \|\varepsilon_i(h)\|, \quad (2.30)$$

siendo

$$\varepsilon_i(h) = x(t_i) - x_i \in \mathbb{R}^n,$$

para $i = 0, 1, \dots, N$, donde $x(t)$ es la solución del problema de valor inicial (2.1) y $\{x_i\}_{i=0}^N$ es la aproximación numérica obtenida. $\square$

En cada paso del Método de Euler (explícito), para $i = 0, 1, \dots, N-1$, se tiene que

$$x_{i+1} - x_i = hf(t_i, x_i), \quad (2.31)$$

mientras que la solución $x(t)$ verifica

$$x(t_{i+1}) - x(t_i) = hf(t_i, x(t_i)) + h\tau(t_i, h) \quad (2.32)$$

(considerar (2.8) con $\theta = 0$ y $t = t_i$). Por tanto, en el cálculo de cada nuevo valor x_{i+1} se introduce un error de tamaño $h\|\tau(t_i, h)\|$ que, para f suficientemente regular (véanse los Teoremas 2.16 y 2.17), es una $O(h^2)$. Puesto que estos errores se van acumulando en cada etapa se puede pensar, de una manera intuitiva, que al final del proceso de N cálculos, se tendrá un error de orden

$$N \times O(h^2) = \frac{T-t_0}{h} \times O(h^2) = O(h).$$

Veamos que esto es así. En efecto, si a la ecuación (2.32) se le resta la ecuación (2.31), se obtiene

$$x(t_{i+1}) - x_{i+1} = (x(t_i) - x_i) + h(f(t_i, x(t_i)) - f(t_i, x_i)) + h\tau(t_i, h),$$

es decir,

$$\varepsilon_{i+1}(h) = \varepsilon_i(h) + h(f(t_i, x(t_i)) - f(t_i, x_i)) + h\tau(t_i, h) \quad (2.33)$$

para $i = 0, 1, \dots, N-1$. De esta forma, si la función f es lipschitziana en la segunda variable con constante de Lipschitz $L > 0$, entonces

$$\|\varepsilon_{i+1}(h)\| \leq \|\varepsilon_i(h)\| + hL\|\varepsilon_i(h)\| + h\|\tau(t_i, h)\| \leq (1 + hL)\|\varepsilon_i(h)\| + h\bar{\tau}(h) \quad (2.34)$$

para $i = 0, 1, \dots, N-1$.

Observación 2.21 Si se considera la función $R : \mathbb{R} \rightarrow \mathbb{R}$ dada por

$$R(z) = 1 + z = e^z + O(z^2), \quad (2.35)$$

la estimación (2.34) determina que

$$\|\varepsilon_{i+1}(h)\| \leq R(hL)\|\varepsilon_i(h)\| + h\bar{\tau}(h) \simeq (e^{hL} + O(h^2))\|\varepsilon_i(h)\| + h\bar{\tau}(h)$$

para $i = 0, 1, \dots, N-1$. Es decir, la estimación sobre $\|\varepsilon_i(h)\|$ crece en cada paso multiplicada (o amplificada) por una función con cierta similitud a una exponencial. Por esta razón, la función R definida en (2.35) se denomina *función de amplificación* (o *función de estabilidad*) del método. Como se verá más adelante en la Sección 5.3, esta misma función nos aparecerá en el estudio de la estabilidad absoluta del método. $\square$

Para continuar, resulta de gran utilidad el siguiente resultado técnico:

Lema 2.22 (Acotación) Si $\{y_i\}_{i=0}^{\infty}$ es una sucesión de números reales no negativos que verifica

$$y_{i+1} \leq (1 + A)y_i + B$$

para $i = 0, 1, 2, \dots$ con $A > 0$ y $B \geq 0$, entonces

$$y_i \leq (1 + A)^i y_0 + \frac{(1 + A)^i - 1}{A} B \quad (2.36)$$

para $i = 0, 1, 2, \dots$. En particular, se verifica que

$$y_i \leq e^{iA} y_0 + \frac{e^{iA} - 1}{A} B \quad (2.37)$$

para $i = 0, 1, 2, \dots$

DEMOSTRACIÓN. Probamos (2.36) por inducción en el índice i :

a) Para $i = 0$ el resultado es evidente.

b) Supuesto cierto el resultado para $i - 1$ lo probamos para i . Aplicando la hipótesis de inducción se tiene que:

$$\begin{aligned} y_i &\leq (1 + A)y_{i-1} + B \leq (1 + A) \left((1 + A)^{i-1} y_0 + \frac{(1 + A)^{i-1} - 1}{A} B \right) + B \\ &= (1 + A)^i y_0 + \frac{(1 + A)^i - 1 - A}{A} B + B = (1 + A)^i y_0 + \frac{(1 + A)^i - 1}{A} B, \end{aligned}$$

con lo que queda probado (2.36).

Finalmente, la relación (2.37) se deduce, trivialmente, a partir de (2.36) sin más que tener en cuenta que, por ser $A > 0$,

$$e^A = \sum_{k=0}^{\infty} \frac{A^k}{k!} = 1 + A + \frac{A^2}{2} + \frac{A^3}{6} + \dots > 1 + A,$$

lo que determina que $(1 + A)^i \leq e^{iA}$ para $i = 0, 1, 2, \dots$ $\square$

Regresando a la estimación del error de discretización global absoluto en el Método de Euler (explícito), teniendo en cuenta la expresión (2.34), si aplicamos el Lema 2.22 tomando $y_i = \|\varepsilon_i(h)\|$, $A = hL$ y $B = h\bar{\tau}(h)$, para cada $i = 0, 1, \dots, N$ se tiene que

$$\|\varepsilon_i(h)\| \leq e^{ihL} \|\varepsilon_0(h)\| + \frac{e^{ihL} - 1}{L} \bar{\tau}(h),$$

de donde se deduce la estimación

$$\varepsilon(h) \leq e^{NhL} \|\varepsilon_0(h)\| + \frac{e^{NhL} - 1}{L} \bar{\tau}(h) = e^{L(T-t_0)} \|\varepsilon_0(h)\| + \frac{e^{L(T-t_0)} - 1}{L} \bar{\tau}(h). \quad (2.38)$$

En general, para el θ -Método, se tiene la siguiente estimación del error de discretización global absoluto:

Teorema 2.23 Sea $\theta \in [0, 1]$. Si $f \in \mathcal{C}([t_0, T] \times \mathbb{R}^n; \mathbb{R}^n)$ y es lipschitziana en la segunda variable con constante de Lipschitz $L > 0$ y $0 < \tilde{h} < \frac{1}{\theta L}$, el error de discretización global absoluto en el θ -Método verifica que

$$\varepsilon(h) \leq \psi(\tilde{h}) \|\varepsilon_0(h)\| + \frac{\psi(\tilde{h}) - 1}{L} \bar{\tau}(h) \quad (2.39)$$

para todo $h \in \mathcal{H}$ tal que $h \leq \tilde{h}$, estando la función ψ definida como

$$\psi(h) = e^{\frac{L(T-t_0)}{1-\theta hL}}, \quad h \geq 0.$$

DEMOSTRACIÓN. Si a la ecuación (2.8) con $t = t_i$ y multiplicada por $h = \frac{T-t_0}{N} \in \mathcal{H}$ se le resta la ecuación del θ -Método, se obtiene que

$$\begin{aligned} x(t_{i+1}) - x_{i+1} &= (x(t_i) - x_i) + (1 - \theta)h(f(t_i, x(t_i)) - f(t_i, x_i)) \\ &\quad + \theta h(f(t_{i+1}, x(t_{i+1})) - f(t_{i+1}, x_{i+1})) + h\tau(t_i, h), \end{aligned}$$

es decir,

$$\begin{aligned} \varepsilon_{i+1}(h) &= \varepsilon_i(h) + (1 - \theta)h(f(t_i, x(t_i)) - f(t_i, x_i)) \\ &\quad + \theta h(f(t_{i+1}, x(t_{i+1})) - f(t_{i+1}, x_{i+1})) + h\tau(t_i, h) \end{aligned} \quad (2.40)$$

para $i = 0, 1, \dots, N - 1$. De esta forma, como la función f es lipschitziana en la segunda variable con constante de Lipschitz $L > 0$, entonces

$$\begin{aligned} \|\varepsilon_{i+1}(h)\| &\leq \|\varepsilon_i(h)\| + (1 - \theta)hL \|\varepsilon_i(h)\| + \theta hL \|\varepsilon_{i+1}(h)\| + h \|\tau(t_i, h)\| \\ &\leq (1 + (1 - \theta)hL) \|\varepsilon_i(h)\| + \theta hL \|\varepsilon_{i+1}(h)\| + h\bar{\tau}(h) \end{aligned} \quad (2.41)$$

para $i = 0, 1, \dots, N - 1$, con lo que si tomamos $h \in \mathcal{H}$ tal que:

- a) $h < \frac{1}{\theta L}$ en el caso de que $\theta \in (0, 1]$ (hipótesis bajo la que, tal y como se vió en la Observación 2.3, se tiene garantizada la existencia y unicidad de las ecuaciones implícitas que hay que resolver en cada paso).
- b) $h > 0$ cuando $\theta = 0$ (recuérdese que en este caso no hay ninguna restricción sobre el paso h).

se verifica que

$$\|\varepsilon_{i+1}(h)\| \leq \frac{1 + (1 - \theta)hL}{1 - \theta hL} \|\varepsilon_i(h)\| + \frac{h}{1 - \theta hL} \bar{\tau}(h) \quad (2.42)$$

para $i = 0, 1, \dots, N - 1$. Por tanto, aplicando el Lema 2.22 con $y_i = \|\varepsilon_i(h)\|$,

$$A = \frac{1 + (1 - \theta)hL}{1 - \theta hL} - 1 = \frac{hL}{1 - \theta hL} \quad \text{y} \quad B = \frac{h}{1 - \theta hL} \bar{\tau}(h) = \frac{A}{L} \bar{\tau}(h),$$

para cada $i = 0, 1, \dots, N$ se tiene que

$$\|\varepsilon_i(h)\| \leq e^{\frac{ihL}{1 - \theta hL}} \|\varepsilon_0(h)\| + \frac{e^{\frac{ihL}{1 - \theta hL}} - 1}{L} \bar{\tau}(h),$$

de donde, dado que $1 - \theta hL > 0$, se deduce la estimación

$$\begin{aligned} \varepsilon(h) &\leq e^{\frac{NhL}{1 - \theta hL}} \|\varepsilon_0(h)\| + \frac{e^{\frac{NhL}{1 - \theta hL}} - 1}{L} \bar{\tau}(h) \\ &= e^{\frac{L(T-t_0)}{1 - \theta hL}} \|\varepsilon_0(h)\| + \frac{e^{\frac{L(T-t_0)}{1 - \theta hL}} - 1}{L} \bar{\tau}(h) \\ &= \psi(h) \|\varepsilon_0(h)\| + \frac{\psi(h) - 1}{L} \bar{\tau}(h). \end{aligned} \quad (2.43)$$

Puesto que la función $\psi(h)$ verifica que

$$\psi'(h) = \frac{\theta L^2(T - t_0)}{(1 - \theta hL)^2} e^{\frac{L(T-t_0)}{1 - \theta hL}} \geq 0$$

para todo $h \geq 0$ (nótese que $\psi(h)$ es constante cuando $\theta = 0$) y

$$\psi(0) = e^{L(T-t_0)} > 0,$$

se tiene que

$$0 < \psi(h) \leq \psi(\tilde{h}) \quad \text{si} \quad 0 \leq h \leq \tilde{h}.$$

De esta forma, a partir de (2.43), se verifica la relación (2.39) para valores $h \in \mathcal{H}$ tales que $h \leq \tilde{h} < \frac{1}{\theta L}$. $\square$

Observación 2.24 La expresión (2.39) relaciona el error de discretización global absoluto con el error local de truncamiento y el (posible) error en los datos iniciales. $\square$

Definición 2.25

- a) Dada una función $f : [t_0, T] \times \mathbb{R}^n \rightarrow \mathbb{R}^n$, un método monopaso es *convergente para el problema de valor inicial* (2.1) (supuesto que admite solución única) si el error de discretización global absoluto tiene la propiedad de que

$$\lim_{h \rightarrow 0} \varepsilon(h) = 0,$$

supuesto que

$$\lim_{h \rightarrow 0} \varepsilon_0(h) = 0. \quad (2.44)$$

- b) Un método monopaso es *convergente* si para toda función $f \in \mathcal{C}([t_0, T] \times \mathbb{R}^n; \mathbb{R}^n)$ y lipschitziana en la segunda variable se verifica que el método es convergente para el problema de valor inicial (2.1). $\square$

Observación 2.26

- a) Un método numérico es convergente cuando la solución del esquema numérico converge a la (única) solución del problema de valor inicial al hacer tender el paso h a cero.
- b) Obviamente, los métodos convergentes son los únicos que tiene sentido considerar desde el punto de vista de las aplicaciones.
- c) En la práctica, la condición (2.44) juega un papel importante, en el sentido de que es conveniente escoger buenas aproximaciones iniciales. $\square$

Corolario 2.27 El θ -Método es convergente para todo $\theta \in [0, 1]$. Además:

- a) Si $f \in \mathcal{C}^1([t_0, T] \times \mathbb{R}^n; \mathbb{R}^n)$ es lipschitziana en la segunda variable (o f es lipschitziana en las dos variables) y $\|\varepsilon_0(h)\| = O(h)$ se verifica que $\varepsilon(h) = O(h)$.
- b) Si $\theta = \frac{1}{2}$ y $f \in \mathcal{C}^2([t_0, T] \times \mathbb{R}^n; \mathbb{R}^n)$ es lipschitziana en la segunda variable (o $f \in \mathcal{C}^1([t_0, T] \times \mathbb{R}^n; \mathbb{R}^n)$ es lipschitziana en la segunda variable y la función $f^{(1)}$ dada en (2.22) es lipschitziana en las dos variables) y $\|\varepsilon_0(h)\| = O(h^2)$ se tiene que $\varepsilon(h) = O(h^2)$.

DEMOSTRACIÓN. Basta aplicar los Teoremas 2.16, 2.17 y 2.23. $\square$

Observación 2.28 Cuando $f \in \mathcal{C}^1([t_0, T] \times \mathbb{R}^n; \mathbb{R}^n)$ es lipschitziana en la segunda variable y $\|\varepsilon_0(h)\| = O(h)$, los valores $\{x_0, x_1, \dots, x_N\}$ que se obtienen a partir del θ -Método aproximan los valores exactos $\{x(t_0), x(t_1), \dots, x(t_N)\}$ al menos con un orden de error $O(h)$ (nótese que al disminuir h aumenta N y, por tanto, el número de puntos de la malla). Téngase en cuenta que, aún siendo $\|\varepsilon_0(h)\| = O(h^r)$ con $r > 1$, no se mejora el resultado anterior. $\square$

Observación 2.29 Debe resaltarse el hecho de que la estimación (2.39) puede resultar “engañosa”. Así, por ejemplo, cuando $\theta = 0$, $f \in \mathcal{C}^1$ y f es lipschitziana en la segunda variable con constante de Lipschitz $L > 0$, de (2.39) y (2.18) se sigue que

$$\varepsilon(h) \leq e^{L(T-t_0)} \|\varepsilon_0(h)\| + \frac{e^{L(T-t_0)} - 1}{L} \frac{cM_2}{2} h, \quad 0 < h \leq \tilde{h},$$

pudiendo ocurrir que las constantes

$$e^{L(T-t_0)}, \quad \frac{e^{L(T-t_0)} - 1}{L} \quad \text{y} \quad \frac{cM_2}{2}$$

sean muy grandes (por ejemplo, cuando son muy grandes la longitud del intervalo y/o la constante de Lipschitz), lo que puede obligar a utilizar pasos h extremadamente pequeños para obtener resultados de aproximación, en muchos casos, poco satisfactorios. $\square$

Observación 2.30 Dado $\theta \in [0, 1]$, si se considera la *función de amplificación* (o *de estabilidad*) $R : (-\frac{1}{\theta}, \frac{1}{\theta}) \rightarrow \mathbb{R}$ definida como

$$\begin{aligned} R(z) &= \frac{1 + (1 - \theta)z}{1 - \theta z} = (1 + (1 - \theta)z) \sum_{k=0}^{\infty} (\theta z)^k = (1 + (1 - \theta)z) (1 + \theta z + \theta^2 z^2 + \theta^3 z^3 + O(z^4)) \\ &= 1 + \theta z + \theta^2 z^2 + \theta^3 z^3 + (1 - \theta)z + (1 - \theta)\theta z^2 + (1 - \theta)\theta^2 z^3 + O(z^4) \\ &= 1 + z + \theta z^2 + \theta^2 z^3 + O(z^4) = \begin{cases} e^z + O(z^2) & \text{si } \theta \neq \frac{1}{2} \\ e^z + O(z^3) & \text{si } \theta = \frac{1}{2}, \end{cases} \end{aligned} \quad (2.45)$$

la estimación (2.42) determina que

$$\|\varepsilon_{i+1}(h)\| \leq R(hL) \|\varepsilon_i(h)\| + \frac{h}{1-\theta hL} \bar{\tau}(h) \simeq \begin{cases} (e^{hL} + O(h^2)) \|\varepsilon_i(h)\| + \frac{h}{1-\theta hL} \bar{\tau}(h) & \text{si } \theta \neq \frac{1}{2} \\ (e^{hL} + O(h^3)) \|\varepsilon_i(h)\| + \frac{2h}{2-hL} \bar{\tau}(h) & \text{si } \theta = \frac{1}{2} \end{cases}$$

para $i = 0, 1, \dots, N-1$. Es decir, la estimación sobre $\|\varepsilon_i(h)\|$ crece en cada paso multiplicada (o amplificada) por una función con cierta similitud a una exponencial. Nótese que, en este caso, el crecimiento de la estimación para el Método del Trapecio es más rápido que el obtenido en el Método de Euler (explícito) pero menos que en el Método de Euler implícito, ya que

$$1 + hL < \frac{2 + hL}{2 - hL} < \frac{1}{1 - hL}$$

si $0 < h < \frac{1}{L}$. De hecho, cuando $0 < h < \frac{1}{L}$, la función $\Psi : [0, 1] \rightarrow \mathbb{R}$ dada por

$$\Psi(\theta) = \frac{1 + (1 - \theta)hL}{1 - \theta hL}$$

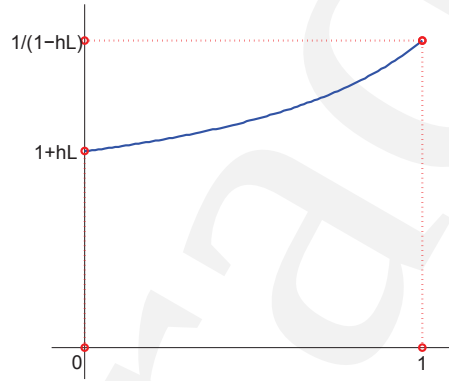


Figura 2.4: Gráfica de la función $\Psi(\theta) = \frac{1 + (1 - \theta)hL}{1 - \theta hL}$, $\theta \in [0, 1]$ cuando $h = 0.6$ y $L = 1$.

(véase la Figura 2.4) tiene las propiedades

$$\Psi(0) = 1 + hL, \Psi(1) = \frac{1}{1 - hL} \text{ y } \Psi'(\theta) = \left(\frac{hL}{1 - \theta hL} \right)^2 > 0, \theta \in (0, 1)$$

(compruébense). Por tanto, siempre que $0 < h < \frac{1}{L}$, se verifica que

$$1 + hL < \frac{1 + (1 - \theta)hL}{1 - \theta hL} < \frac{1}{1 - hL}, \theta \in (0, 1).$$

De esta forma, en un primer análisis, parecería que el Método de Euler explícito es “mejor” y “más barato” que el Método de Euler implícito. Más adelante veremos que esta afirmación es falsa pues, para que se verifique la condición de estabilidad absoluta en los métodos explícitos, pueden requerirse valores del paso h extremadamente pequeños. $\square$

Observación 2.31 Como acabamos de ver, cuando la función f es suficientemente regular, el orden de precisión del Método del Trapecio es dos y, por tanto, mayor que el orden del Método de Euler (explícito) (véanse los Teoremas 2.16 y 2.17). Esto, en principio, no significa que el Método del Trapecio sea “mejor” que el de Euler puesto que, mientras que el Método de Euler es explícito (el vector x_{i+1} se obtiene a partir de x_i de una forma bastante sencilla), el del Trapecio es implícito y requiere mayor coste computacional pues, en cada etapa, debe resolverse un sistema de ecuaciones que, en general, es no lineal. Por eso, en muchas ocasiones es preferible un método de orden bajo pero con un coste computacional pequeño (aunque haya que dar más pasos) que uno de orden más elevado con un coste computacional mayor. $\square$

2.2.5 Otras características y propiedades del Método de Euler y sus variantes

Finalizamos esta sección haciendo algunas consideraciones sobre el Método de Euler explícito que, fácilmente, pueden generalizarse al θ -Método.

Teorema 2.32 Dada $f \in \mathcal{C}^1([t_0, T] \times \mathbb{R}^n; \mathbb{R}^n)$ y lipschitziana en la segunda variable (respectivamente, f lipschitziana en las dos variables), se considera el método de Euler (explícito)

$$\begin{cases} x_0 \simeq \xi_0 \\ x_{i+1} = x_i + hf(t_i, x_i), \quad i = 0, 1, \dots, N-1 \end{cases}$$

y la poligonal de Euler (1.14), es decir,

$$x^h(t) = x_i + \frac{x_{i+1} - x_i}{h} (t - t_i), \quad t \in [t_i, t_{i+1}], \quad (2.46)$$

para $i = 0, 1, \dots, N-1$. Si $\|\xi_0 - x_0(h)\| = O(h)$, se verifica que

$$\|x^h(\cdot) - x(\cdot)\|_{\mathcal{C}([t_0, T]; \mathbb{R}^n)} = O(h),$$

siendo $x(t)$ la única solución del problema de valor inicial (2.1).^a

^aRecuérdese que $\|g\|_{\mathcal{C}([t_0, T]; \mathbb{R}^n)} = \max_{t \in [t_0, T]} \|g(t)\|$ para $g \in \mathcal{C}([t_0, T]; \mathbb{R}^n)$.

DEMOSTRACIÓN. Claramente, la función $x^h \in \mathcal{C}([t_0, T])$ y es derivable en (t_0, T) salvo, quizá, en los puntos de la partición $\{t_0, t_1, \dots, t_N\}$ (véase la Figura 1.1(a)). Fijado $h > 0$ suficientemente pequeño e $i \in \{0, 1, \dots, N-1\}$ arbitrario, para todo punto $t \in [t_i, t_{i+1}]$ se verifica que

$$\begin{aligned} \|x^h(t) - x(t)\| &= \|x^h(t) - x_i + x_i - x(t_i) + x(t_i) - x(t)\| \\ &\leq \|x^h(t) - x_i\| + \|x_i - x(t_i)\| + \|x(t_i) - x(t)\|. \end{aligned} \quad (2.47)$$

Acotemos cada uno de los tres sumandos del miembro de la derecha:

a) Por el Teorema 2.23 se tiene que

$$\|x_i - x(t_i)\| = \|\varepsilon_i(h)\| \leq \varepsilon(h) \leq e^{L(T-t_0)} \|\varepsilon_0(h)\| + \frac{e^{L(T-t_0)} - 1}{L} \bar{\tau}(h) \leq c_1 h \quad (2.48)$$

para cierta constante $c_1 > 0$ ya que, por hipótesis, $\|\xi_0 - x_0\| = O(h)$ y, por el Teorema 2.16 (respectivamente, el Teorema 2.17), se tiene que $\bar{\tau}(h) = O(h)$.

b) Puesto que $t_i \leq t \leq t_{i+1} = t_i + h$, por el teorema del Valor Medio se verifica que

$$\|x(t_i) - x(t)\| \leq c \|x(t_i) - x(t)\|_\infty \leq cM_1 |t_i - t| \leq cM_1 h,$$

estando $c > 0$ y M_1 definidas, respectivamente, en (2.14) y (2.27).

c) De (2.46) se deduce que

$$\|x^h(t) - x_i\| = \left\| \frac{x_{i+1} - x_i}{h} (t - t_i) \right\| = \|f(t_i, x_i)(t - t_i)\| = \|f(t_i, x_i)\| |t - t_i| \leq \|f(t_i, x_i)\| h. \quad (2.49)$$

Puesto que la función f es lipschitziana en la segunda variable con constante de Lipschitz $L > 0$, se verifica que

$$\|f(t_i, x_i) - f(t_i, x(t_i))\| \leq L \|x_i - x(t_i)\|,$$

de donde se obtiene que

$$\|f(t_i, x_i)\| \leq \|f(t_i, x(t_i))\| + L \|x_i - x(t_i)\| = \|x'(t_i)\| + L \|x_i - x(t_i)\| \leq cM_1 + Lc_1 h \quad (2.50)$$

(véase (2.48)). Por tanto, regresando a (2.49), se verifica que

$$\|x^h(t) - x_i\| \leq (cM_1 + Lc_1 h) h.$$

Consecuentemente, sustituyendo las cotas obtenidas en la expresión (2.47), se tiene que existen $c_2 > 0$ y $\tilde{h} > 0$ tales que

$$\|x^h(t) - x(t)\| \leq (cM_1 + Lc_1 h) h + c_1 h + cM_1 h = (2cM_1 + c_1(1 + Lh)) h \leq c_2 h$$

para todo $h \in (0, \tilde{h})$ y para todo $t \in [t_0, T]$, lo cual concluye el resultado. $\square$

Observación 2.33 Nótese que, en particular, el Teorema 2.32 implica que $\{x^h(\cdot)\}_h \rightarrow x(\cdot)$ uniformemente en $[t_0, T]$ cuando $h \rightarrow 0$. $\square$

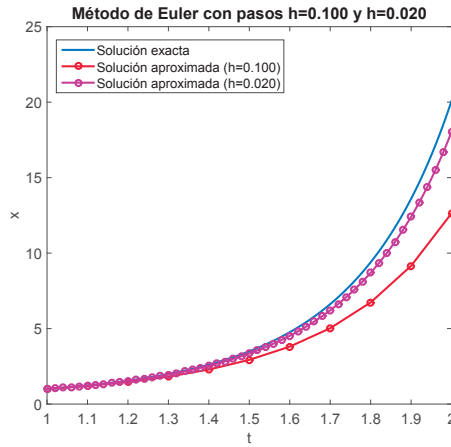


Figura 2.5: Método de Euler con pasos $h = 0.1$ y $h = 0.02$.

Ejemplo 2.34 En la Figura 2.5 se comparan los resultados al aplicar el Método de Euler (explícito) para la resolución del problema de valor inicial del Ejemplo 2.1 cuando se toman como pasos $h = 0.1$ y $h = 0.02$, respectivamente. $\square$

Como se aprecia en las expresiones (2.16) y (2.18), a veces resulta interesante disponer de una estimación de la norma de las derivadas de la solución. ¿Cómo se puede obtener teniendo en cuenta que la solución es, en general, desconocida? En algunas ocasiones puede obtenerse a partir de la función f que aparece en la ecuación diferencial pues, como se vió en la demostración del Teorema 1.18, se verifica que

$$x''(t) = \frac{\partial}{\partial t} f(t, x(t)) + D_{\xi} f(t, x(t)) f(t, x(t))$$

y, a partir de aquí, se pueden obtener las derivadas sucesivas de $x(t)$ expresadas en términos de la función f y sus derivadas.

Comenzamos con un lema en el que aparece una desigualdad diferencial, debido a T. H. Gronwall, que permite obtener una estimación de una función a partir de una inecuación integral.

Lema 2.35 (Gronwall) Sean $\alpha : [a, b] \rightarrow \mathbb{R}$, $\beta : [a, b] \rightarrow [0, +\infty)$ y $u : [a, b] \rightarrow \mathbb{R}$ funciones continuas tales que

$$u(t) \leq \alpha(t) + \int_a^t \beta(s) u(s) ds, \quad \forall t \in [a, b].$$

Entonces se verifica que

$$u(t) \leq \alpha(t) + \int_a^t \alpha(s) \beta(s) \exp \left(\int_s^t \beta(r) dr \right) ds, \quad \forall t \in [a, b]. \quad (2.51)$$

En particular, si la función α es no decreciente, se tiene que

$$u(t) \leq \alpha(t) \exp \left(\int_a^t \beta(s) ds \right), \quad \forall t \in [a, b]. \quad (2.52)$$

DEMOSTRACIÓN. Véase [Guzmán, Lema 4.4.1]. $\square$

Veamos cómo se puede obtener una cota de $\|x(t)\|$ utilizando la siguiente estimación a priori:

Teorema 2.36 (Estimación a priori) Sea $f \in C^1([t_0, T] \times \mathbb{R}^n; \mathbb{R}^n)$ y lipschitziana en la segunda variable con constante de Lipschitz $L > 0$ (o f lipschitziana en las dos variables con constante de Lipschitz $L > 0$). Si $x(t)$ es la única solución del problema de valor inicial (2.1), se verifica que

$$\|x(t)\| \leq e^{L(T-t_0)} \|\xi_0\| + \frac{e^{L(T-t_0)} - 1}{L} M \quad (2.53)$$

para todo $t \in [t_0, T]$, siendo

$$M = \max_{s \in [t_0, T]} \|f(s, 0)\|. \quad (2.54)$$

DEMOSTRACIÓN. Para todo $t \in [t_0, T]$ se verifica que

$$x(t) = \xi_0 + \int_{t_0}^t f(s, x(s)) ds$$

y, por tanto,

$$\|x(t)\| \leq \|\xi_0\| + \int_{t_0}^t \|f(s, x(s))\| ds. \quad (2.55)$$

Por otra parte, como la función f es lipschitziana en la segunda variable con constante de Lipschitz $L > 0$, para todo $s \in [t_0, T]$ se tiene que

$$\|f(s, x(s)) - f(s, 0)\| \leq L\|x(s)\|$$

y, por tanto,

$$\|f(s, x(s))\| \leq L\|x(s)\| + \|f(s, 0)\| \leq L\|x(s)\| + M$$

(véase (2.54)). De esta forma, utilizando esta cota en la expresión (2.55), para todo $t \in [t_0, T]$ se verifica que

$$\|x(t)\| \leq \|\xi_0\| + \int_{t_0}^t (L\|x(s)\| + M) ds = \|\xi_0\| + M(t - t_0) + L \int_{t_0}^t \|x(s)\| ds.$$

Aplicando la desigualdad (2.51) del Lema de Gronwall a las funciones $u(t) = \|x(t)\|$, $\alpha(t) = \|\xi_0\| + M(t - t_0)$ y $\beta(t) \equiv L$, para todo $t \in [t_0, T]$ se obtiene que

$$\begin{aligned} \|x(t)\| &\leq \|\xi_0\| + M(t - t_0) + L \int_{t_0}^t (\|\xi_0\| + M(s - t_0)) e^{L(t-s)} ds \\ &= \|\xi_0\| + M(t - t_0) - \|\xi_0\| e^{L(t-t_0)} \Big|_{t_0}^t + ML \int_{t_0}^t (s - t_0) e^{L(t-s)} ds \\ &= \|\xi_0\| + \cancel{M(t - t_0)} - \|\xi_0\| + \|\xi_0\| e^{L(t-t_0)} - \cancel{M(t - t_0)} + \frac{e^{L(t-t_0)} - 1}{L} M \\ &\leq \|\xi_0\| e^{L(T-t_0)} + \frac{e^{L(T-t_0)} - 1}{L} M, \end{aligned}$$

donde se ha utilizado, integrando por partes, que

$$\int_{t_0}^t (s - t_0) e^{L(t-s)} ds = -\frac{t - t_0}{L} + \frac{e^{L(t-t_0)} - 1}{L^2}. \quad \square$$

Ejemplo 2.37 Consideremos el problema de valor inicial

$$\begin{cases} x'(t) = x(t) + t, & t \in [0, 1] \\ x(0) = 0, \end{cases} \quad (2.56)$$

cuya solución exacta es $x(t) = e^t - t - 1$ y, por tanto,

$$\|x''(t)\|_\infty = \max_{t \in [0, 1]} e^t = e \simeq 2.7183. \quad (2.57)$$

Por otra parte, la cota que se obtiene de la norma de la derivada segunda de la solución del problema (2.56) en función de $f(t, \xi) = t + \xi$ es

$$\|x''(t)\|_\infty = \left\| \frac{\partial}{\partial t} f(t, x(t)) + \frac{\partial}{\partial \xi} f(t, x(t)) f(t, x(t)) \right\|_\infty = \|1 + t + x(t)\|_\infty \leq 2 + \|x(t)\|_\infty. \quad (2.58)$$

Puesto que, en este caso, $L = T - t_0 = 1$ y $\xi_0 = 0$, la estimación (2.53) determina que

$$\|x(t)\|_\infty \leq (e - 1) \max_{s \in [0,1]} \|s\|_\infty = e - 1.$$

Sustituyendo esta cota en (2.58) se obtiene la estimación

$$\|x''(t)\|_\infty \leq 2 + e - 1 = e + 1 \simeq 3'7183$$

(compárese con el valor exacto (2.57)).

En las secciones anteriores se ha considerado que, tanto los números, como las operaciones con números, se realizan de forma exacta. Sin embargo, tal y como se explicaba en la Sección 1.5, esto no se suele cumplir al utilizar ordenadores y se introducen errores, tanto de redondeo como operacionales. De esta forma, cuando se implementa el Método de Euler en un ordenador, en realidad no se obtiene la secuencia $\{x_0, x_1, \dots, x_N\}$, sino otra secuencia $\{x_0^*, x_1^*, \dots, x_N^*\}$ dada por

Método de Euler (explícito) en un ordenador

Paso 1 $x_0^* = \xi_0 + \delta_0$

Paso 2 Para $i = 0, 1, \dots, N - 1$
 $x_{i+1}^* = x_i^* + hf(t_i, x_i^*) + \delta_{i+1}$

siendo δ_i el error de redondeo y operacional (debido al ordenador) en la etapa i -ésima. Si denotamos

$$\varepsilon^*(h) = \max_{i \in \{0, 1, \dots, N\}} \|\varepsilon_i^*(h)\|,$$

siendo

$$\varepsilon_i^*(h) = x(t_i) - x_i^* \in \mathbb{R}^n$$

para $i = 0, 1, \dots, N$, y consideramos el error de redondeo y operacional máximo

$$\delta = \max_{i \in \{0, 1, \dots, N\}} \|\delta_i\|,$$

argumentando como en la obtención de (2.33) se obtiene que

$$\varepsilon_{i+1}^*(h) = \varepsilon_i^*(h) + h(f(t_i, x(t_i)) - f(t_i, x_i^*)) + h\tau(t_i, h) - \delta_{i+1}$$

para $i = 0, 1, \dots, N - 1$. De esta forma, si la función f es lipschitziana en la segunda variable con constante de Lipschitz $L > 0$, se tiene que

$$\|\varepsilon_{i+1}^*(h)\| \leq (1 + hL) \|\varepsilon_i^*(h)\| + h\bar{\tau}(h) + \delta$$

para $i = 0, 1, \dots, N - 1$. Por tanto, aplicando el Lema 2.22 con $y_i = \|\varepsilon_i^*(h)\|$, $A = hL$ y $B = h\bar{\tau}(h) + \delta$, para cada $i = 0, 1, \dots, N$ se tiene que

$$\|\varepsilon_i^*(h)\| \leq e^{ihL} \|\varepsilon_0^*(h)\| + \frac{e^{ihL} - 1}{L} \left(\bar{\tau}(h) + \frac{\delta}{h} \right),$$

de donde se deduce la estimación

$$\varepsilon^*(h) \leq e^{NhL} \|\varepsilon_0^*(h)\| + \frac{e^{NhL} - 1}{L} \left(\bar{\tau}(h) + \frac{\delta}{h} \right) = e^{L(T-t_0)} \|\delta_0\| + \frac{e^{L(T-t_0)} - 1}{L} \left(\bar{\tau}(h) + \frac{\delta}{h} \right). \quad (2.59)$$

Observación 2.38 A diferencia de lo que ocurría con $\varepsilon(h)$ (véase la Observación 2.28), si $\delta > 0$ ya no se tiene que $\varepsilon^*(h) = O(h)$ pues, en general, ni siquiera se cumple que

$$\lim_{h \rightarrow 0} \varepsilon^*(h) = 0$$

debido al error de máquina que se puede producir en cada paso. Así, por ejemplo, si aplicamos el Método de Euler (explícito) a la resolución del problema de valor inicial

$$\begin{cases} x'(t) = 0, & t \in [0, 1] \\ x(0) = \pi, \end{cases}$$

cuya solución exacta es $x(t) = \pi$, se obtiene que

$$x_i^* = r(\pi)$$

para $i = 0, 1, \dots, N$, donde $r(\pi)$ indica el redondeo del número π (número máquina más cercano a él²). Por tanto, para todo $h > 0$ se verifica que

$$\varepsilon^*(h) = \max_{i \in \{0, 1, \dots, N\}} |x(t_i) - x_i^*| = |\pi - r(\pi)| > 0.$$

El hecho de que el número π no sea un número máquina hace que la diferencia, en valor absoluto, entre π y $r(\pi)$ sea siempre una cantidad positiva que no puede tender a cero. $\square$

Observación 2.39 La estimación (2.59) indica que, para $\delta > 0$ fijo conocido, a la hora de elegir el paso h hay que seguir dos estrategias opuestas: por una lado interesa que h sea pequeño (para mejorar la precisión de $\bar{\tau}(h)$) mientras que, por otro lado, no interesa que h sea demasiado pequeño (para no empeorar la estimación gobernada por el término $\frac{\delta}{h}$). Puede ser conveniente hacer un balance entre las dos estrategias. Más concretamente, si $f \in C^1([t_0, T] \times \mathbb{R}^n; \mathbb{R}^n)$, f es lipschitziana en la segunda variable con constante de Lipschitz $L > 0$ y $\delta_0 = 0$, de las relaciones (2.59) y (2.18) tomando $\theta = 0$, se sigue que

$$\varepsilon^*(h) \leq \frac{e^{L(T-t_0)} - 1}{L} \varphi(h), \quad (2.60)$$

siendo

$$\varphi(h) = \frac{cM_2}{2} h + \frac{\delta}{h}.$$

De esta forma, para minimizar la cota de la estimación (2.60) basta minimizar $\varphi(h)$. Para ello, dado que la función $\varphi \in C^\infty(0, +\infty)$ tiene las propiedades

$$\varphi(h) > 0 \quad \forall h > 0, \quad \lim_{h \rightarrow 0^+} \varphi(h) = +\infty, \quad \lim_{h \rightarrow +\infty} \varphi(h) = +\infty \quad \text{y} \quad \varphi'(h) = \frac{cM_2}{2} - \frac{\delta}{h^2} \begin{cases} < 0 & \text{si } h < h_{\min} \\ = 0 & \text{si } h = h_{\min} \\ > 0 & \text{si } h > h_{\min} \end{cases}$$

siendo

$$h_{\min} = \sqrt{\frac{2\delta}{cM_2}},$$

se verifica que la función $\varphi(h)$ alcanza el mínimo absoluto en el punto $h_{\min}$, siendo este valor mínimo

$$\varphi_{\min} = \min_{h>0} \varphi(h) = \varphi(h_{\min}) = \sqrt{2cM_2\delta}$$

(compruébese). En la Figura 2.6 se muestra la gráfica de la función $\varphi(h)$ cuando $c = 1$, $M_2 = 2$ y $\delta = 10^{-6}$, en cuyo caso $h_{\min} = 10^{-3}$ y $\varphi_{\min} = 2 \times 10^{-3}$.

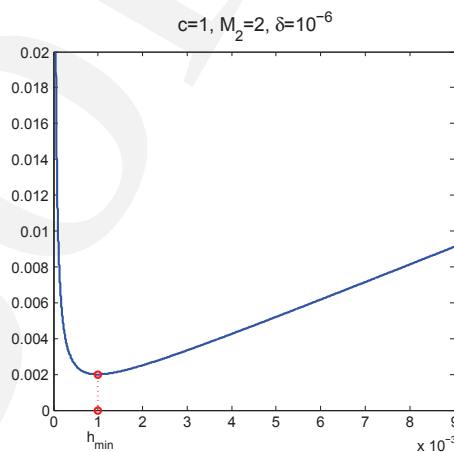


Figura 2.6: Gráfica de la función $\varphi(h)$.

Claramente, a medida que se tomen valores de h más pequeños que $h_{\min}$ la función $\varphi(h)$ va tomando valores cada vez más grandes. También debemos destacar que la optimalidad obtenida es sobre la cota de estimación dada en (2.60) (es decir, sobre su término de la derecha), lo que no impide que en ejemplos concretos se puedan obtener mejores resultados para valores de h más pequeños. $\square$

²En la mayoría de los ordenadores actuales se tiene que $r(\pi) = 3'14159265358979$ en doble precisión.

2.3 Métodos de Taylor

Hasta ahora, el único método que hemos obtenido de orden de precisión 2 ha sido el Método del Trapecio. Tal y como se comentó en la Observación 2.31, este método tiene la desventaja de ser implícito. Por otra parte, el Método de Euler sí que es explícito pero, tan sólo, es de primer orden. Cabe, por tanto, preguntarse si es posible obtener métodos explícitos de orden superior a uno. En esta sección vamos a ver cómo se pueden obtener métodos de este tipo a partir de los desarrollos de Taylor.

2.3.1 Deducción de los métodos

Según se ha visto en la Sección 1.3, si $f \in C^p([t_0, T] \times \mathbb{R}^n; \mathbb{R}^n)$ con $p \in \mathbb{N} \cup \{0\}$, entonces la solución $x(t)$ del problema de valor inicial (2.1) es de clase $C^{p+1}([t_0, T]; \mathbb{R}^n)$ y

$$\begin{aligned} x^{(p+1)}(t) &= \frac{d^p}{dt^p} f(t, x(t)) = \frac{d}{dt} \left(\frac{d^{p-1}}{dt^{p-1}} f(t, x(t)) \right) \\ &= \frac{\partial}{\partial t} \left(\frac{d^{p-1}}{dt^{p-1}} f(t, x(t)) \right) + D_\xi \left(\frac{d^{p-1}}{dt^{p-1}} f(t, x(t)) \right) x'(t) \\ &= \frac{\partial}{\partial t} \left(\frac{d^{p-1}}{dt^{p-1}} f(t, x(t)) \right) + D_\xi \left(\frac{d^{p-1}}{dt^{p-1}} f(t, x(t)) \right) f(t, x(t)) \\ &= f^{(p)}(t, x(t)), \end{aligned}$$

donde, por definición,

$$\begin{cases} f^{(0)}(t, \xi) = f(t, \xi) \\ f^{(k)}(t, \xi) = \frac{\partial f^{(k-1)}}{\partial t}(t, \xi) + D_\xi f^{(k-1)}(t, \xi) f(t, \xi), \quad k = 1, 2, \dots \end{cases} \quad (2.61)$$

Observación 2.40 Antes de continuar veamos cómo son algunas de las funciones consideradas en (2.61) cuando $n = 1$. Para simplificar la notación, vamos a denotar

$$f_t(t, \xi) = \frac{\partial f}{\partial t}(t, \xi) \quad \text{y} \quad f_\xi(t, \xi) = \frac{\partial f}{\partial \xi}(t, \xi),$$

prescindiendo, si no hay lugar a confusión, del argumento (t, ξ) . De esta forma,

$$f^{(1)} = f_t + f_\xi f$$

y

$$f^{(2)} = f_t^{(1)} + f_\xi^{(1)} f = f_{tt} + f_{t\xi} f + f_\xi f_t + (f_{t\xi} + f_{\xi\xi} f + f_\xi^2) f = f_{tt} + 2f_{t\xi} f + f_{\xi\xi} f^2 + (f_t + f_\xi f) f_\xi. \quad (2.62)$$

Así, por ejemplo, si consideramos la función

$$f(t, \xi) = t^2 + \xi^2$$

se tiene que

$$f^{(1)}(t, \xi) = 2t + 2\xi(t^2 + \xi^2)$$

y

$$f^{(2)}(t, \xi) = 2 + 2(t^2 + \xi^2)^2 + (2t + 2\xi(t^2 + \xi^2))2\xi = 2 + 4t\xi + (2t^2 + 6\xi^2)(t^2 + \xi^2). \quad \square$$

De esta forma, por el Teorema de Taylor, para todo $(t, h) \in \mathcal{U}$ se verifica que

$$\begin{aligned} x(t+h) &= x(t) + hx'(t) + \frac{h^2}{2}x''(t) + \dots + \frac{h^p}{p!}x^{(p)}(t) + \frac{h^{p+1}}{(p+1)!} \begin{pmatrix} (x^1)^{p+1}(\zeta^1) \\ (x^2)^{p+1}(\zeta^2) \\ \dots \\ (x^n)^{p+1}(\zeta^n) \end{pmatrix} \\ &= x(t) + hf^{(0)}(t, x(t)) + \frac{h^2}{2}f^{(1)}(t, x(t)) + \dots + \frac{h^p}{p!}f^{(p-1)}(t, x(t)) + \frac{h^{p+1}}{(p+1)!} \begin{pmatrix} (x^1)^{p+1}(\zeta^1) \\ (x^2)^{p+1}(\zeta^2) \\ \dots \\ (x^n)^{p+1}(\zeta^n) \end{pmatrix} \quad (2.63) \\ &= x(t) + hT_p(t, x(t), h) + \frac{h^{p+1}}{(p+1)!} \begin{pmatrix} (x^1)^{p+1}(\zeta^1) \\ (x^2)^{p+1}(\zeta^2) \\ \dots \\ (x^n)^{p+1}(\zeta^n) \end{pmatrix}, \end{aligned}$$

donde $\zeta^k \in (t, t+h)$ para todo $k = 1, 2, \dots, n$ y $T_p(t, \xi, h)$ es el *polinomio de Taylor* de orden p (que es un polinomio de grado $p-1$ en la variable h) dado por

$$T_p(t, \xi, h) = f^{(0)}(t, \xi) + \frac{h}{2} f^{(1)}(t, \xi) + \dots + \frac{h^{p-1}}{p!} f^{(p-1)}(t, \xi) = \sum_{k=0}^{p-1} \frac{h^k}{(k+1)!} f^{(k)}(t, \xi). \quad (2.64)$$

Lo anterior conduce al siguiente esquema:

Método de Taylor de orden p	
<u>Paso 1</u>	$x_0 \simeq \xi_0$
<u>Paso 2</u>	Para $i = 0, 1, \dots, N-1$ $x_{i+1} = x_i + hT_p(t_i, x_i, h).$

(2.65)

Teniendo en cuenta los conceptos definidos en la Sección 1.6, es claro que este método es explícito y monopaso, pero tiene la dificultad práctica de tener que calcular y evaluar todas las funciones $\{f^{(0)}, f^{(1)}, \dots, f^{(p-1)}\}$.

Observación 2.41 Para $p = 1$ se obtiene el Método de Euler (explícito) dado que, en este caso, $T_1(t, \xi, h) = f(t, \xi)$. $\square$

2.3.2 Análisis del error cometido

Según las nociones sobre errores expuestas en la Sección 1.5 y la argumentación seguida en la Sección 2.2.4, definimos el error local de truncamiento $\tau(t, h)$, asociado a una solución exacta $x(t)$ de la ecuación diferencial (2.7) a partir del instante $t \in [t_0, T)$, como

$$\tau(t, h) = \frac{x(t+h) - x(t)}{h} - T_p(t, x(t), h) \quad (2.66)$$

para todo $h \in (0, T-t]$. A la vista de (2.63), se tiene que

$$\tau(t, h) = \frac{h^p}{(p+1)!} \begin{pmatrix} (x^1)^{p+1}(\zeta^1) \\ (x^2)^{p+1}(\zeta^2) \\ \dots \\ (x^n)^{p+1}(\zeta^n) \end{pmatrix},$$

donde $\zeta^k \in (t, t+h)$ para todo $k = 1, 2, \dots, n$. De esta forma, para el error local de truncamiento $\bar{\tau}(h)$ dado en (2.15), se verifica que

$$\bar{\tau}(h) \leq \frac{cM_{p+1}}{(p+1)!} h^p, \quad (2.67)$$

donde

$$M_{p+1} = \max_{t \in [t_0, T]} \|x^{p+1}(t)\|_\infty$$

y $c > 0$ verifica (2.14). Hemos probado, por tanto, el siguiente resultado:

Teorema 2.42 El Método de Taylor (2.65) es de orden de precisión p . $\square$

Respecto al error de discretización global absoluto definido en (2.30), se tiene el siguiente resultado:

Teorema 2.43 Sea $f \in C^p([t_0, T] \times \mathbb{R}^n; \mathbb{R}^n)$, con $p \in \mathbb{N}$, de forma que la función $f^{(k)}$ definida en (2.61) es lipschitziana en la segunda variable con constante de Lipschitz $L_k > 0$ para $k = 0, 1, \dots, p-1$. Dado $\tilde{h} > 0$, el error de discretización global absoluto en el Método de Taylor de orden p verifica que

$$\varepsilon(h) \leq e^{(T-t_0)\psi(\tilde{h})} \|\varepsilon_0(h)\| + \frac{e^{(T-t_0)\psi(\tilde{h})} - 1}{L_0} \bar{\tau}(h) \quad (2.68)$$

para todo $h \in \mathcal{H}$ tal que $0 < h \leq \tilde{h}$, siendo

$$\psi(h) = \sum_{k=0}^{p-1} \frac{L_k}{(k+1)!} h^k = L_0 + \frac{L_1}{2} h + \dots + \frac{L_{p-1}}{p!} h^{p-1}, \quad h \geq 0. \quad (2.69)$$

DEMOSTRACIÓN. Si a la ecuación (2.66) multiplicada por h y particularizada en $t = t_i$ se le resta la ecuación del Método de Taylor (2.65), se obtiene

$$x(t_{i+1}) - x_{i+1} = (x(t_i) - x_i) + h(T_p(t_i, x(t_i), h) - T_p(t_i, x_i, h)) + h\tau(t_i, h),$$

es decir,

$$\begin{aligned}\varepsilon_{i+1}(h) &= \varepsilon_i(h) + h(T_p(t_i, x(t_i), h) - T_p(t_i, x_i, h)) + h\tau(t_i, h) \\ &= \varepsilon_i(h) + h \sum_{k=0}^{p-1} \frac{h^k}{(k+1)!} \left(f^{(k)}(t_i, x(t_i)) - f^{(k)}(t_i, x_i) \right) + h\tau(t_i, h)\end{aligned}\quad (2.70)$$

para $i = 0, 1, \dots, N-1$. De esta forma, puesto que la función $f^{(k)}$ es lipschitziana en la segunda variable con constante de Lipschitz $L_k > 0$ para $k = 0, 1, \dots, p-1$, se verifica que

$$\|\varepsilon_{i+1}(h)\| \leq \left(1 + h \sum_{k=0}^{p-1} \frac{L_k}{(k+1)!} h^k\right) \|\varepsilon_i(h)\| + h\|\tau(t_i, h)\| \leq (1 + h\psi(h)) \|\varepsilon_i(h)\| + h\bar{\tau}(h) \quad (2.71)$$

para $i = 0, 1, \dots, N-1$. Por tanto, aplicando el Lema 2.22 con $y_i = \|\varepsilon_i(h)\|$,

$$A = h\psi(h) \text{ y } B = h\bar{\tau}(h) = A \frac{\bar{\tau}(h)}{\psi(h)},$$

para cada $i = 0, 1, \dots, N$ se tiene que

$$\|\varepsilon_i(h)\| \leq e^{ih\psi(h)} \|\varepsilon_0(h)\| + \frac{e^{ih\psi(h)} - 1}{\psi(h)} \bar{\tau}(h),$$

de donde, utilizando que $\psi(h) \geq 0$ para $h \geq 0$, se deduce la estimación

$$\varepsilon(h) \leq e^{Nh\psi(h)} \|\varepsilon_0(h)\| + \frac{e^{Nh\psi(h)} - 1}{\psi(h)} \bar{\tau}(h) = e^{(T-t_0)\psi(h)} \|\varepsilon_0(h)\| + \frac{e^{(T-t_0)\psi(h)} - 1}{\psi(h)} \bar{\tau}(h). \quad (2.72)$$

Puesto que la función $\psi(h)$ verifica que

$$\psi'(h) = \sum_{k=1}^{p-1} \frac{kL_k}{(k+1)!} h^{k-1} = \frac{L_1}{2} + \frac{L_2}{3} h + \dots + \frac{(p-1)L_{p-1}}{p!} h^{p-2} > 0$$

para $h \geq 0$ y $\psi(0) = L_0$, se tiene que

$$0 < L_0 \leq \psi(h) \leq \psi(\tilde{h}) \text{ si } 0 \leq h \leq \tilde{h}$$

y, por tanto, a partir de (2.72), se obtiene la relación (2.68) para valores $0 < h \leq \tilde{h}$. $\square$

Observación 2.44 La expresión (2.68) relaciona el error de discretización global absoluto con el error local de truncamiento y el (posible) error en los datos iniciales. $\square$

Corolario 2.45 Si la función $f \in C^p([t_0, T] \times \mathbb{R}^n; \mathbb{R}^n)$ con $p \in \mathbb{N}$ es tal que las funciones $\{f^{(0)}, f^{(1)}, \dots, f^{(p-1)}\}$ definidas en (2.61) son lipschitzianas en la segunda variable, se verifica que el Método de Taylor de orden p es convergente para el correspondiente problema de valor inicial (2.1). Además, si $\|\varepsilon_0(h)\| = O(h^p)$ se tiene que $\varepsilon(h) = O(h^p)$.

DEMOSTRACIÓN. Basta aplicar los Teoremas 2.42 y 2.43. $\square$

Ejemplo 2.46 Los métodos de Taylor de órdenes 2 y 3 aplicados al problema de valor inicial

$$\begin{cases} x'(t) = t^2 + (x(t))^2, & t \in [t_0, T] \\ x(t_0) = \xi_0 \end{cases}$$

son, de acuerdo con lo visto en la Observación 2.40, $x_0 \simeq \xi_0$,

$$x_{i+1} = x_i + h \left(f^{(0)}(t_i, x_i) + \frac{h}{2} f^{(1)}(t_i, x_i) \right) = x_i + h(t_i^2 + x_i^2) + h^2(t_i + x_i(t_i^2 + x_i^2))$$

y

$$\begin{aligned} x_{i+1} &= x_i + h \left(f^{(0)}(t_i, x_i) + \frac{h}{2} f^{(1)}(t_i, x_i) + \frac{h^2}{6} f^{(2)}(t_i, x_i) \right) \\ &= x_i + h(t_i^2 + x_i^2) + h^2(t_i + x_i(t_i^2 + x_i^2)) + \frac{h^3}{3}(1 + 2t_i x_i + (t_i^2 + 3x_i^2)(t_i^2 + x_i^2)), \end{aligned}$$

respectivamente. $\square$

2.4 Métodos de Runge–Kutta

2.4.1 Generalidades

Según se ha visto, con los Métodos de Taylor se pueden alcanzar órdenes de convergencia muy altos, pero tienen la desventaja de requerir un análisis previo a su programación, dado que se necesitan calcular y evaluar las funciones $\{f^{(0)}, f^{(1)}, \dots, f^{(p-1)}\}$. Los *Métodos de Runge–Kutta* evitan estas dificultades puesto que sólo necesitan evaluaciones de la propia función f y no de sus derivadas. Para ello se aproxima la función $T_p(t, \xi, h)$ del Método de Taylor de orden p por cierta función $\Phi(t, \xi, h)$, que sólo necesita evaluaciones de la función f , verificando que

$$\Phi(t, \xi, h) = T_p(t, \xi, h) + O(h^p). \quad (2.73)$$

Concretamente, se utiliza el siguiente esquema:

Método de Runge–Kutta de r etapas	
<u>Paso 1</u>	$x_0 \simeq \xi_0$
<u>Paso 2</u>	Para $i = 0, 1, \dots, N-1$ $x_{i+1} = x_i + h\Phi(t_i, x_i, h)$,

(2.74)

siendo

$$\Phi(t, \xi, h) = \sum_{k=1}^r c_k F_k(t, \xi, h) \quad (2.75)$$

y

$$F_k(t, \xi, h) = f \left(t + hb_k, \xi + h \sum_{j=1}^r a_{kj} F_j(t, \xi, h) \right), \quad k = 1, 2, \dots, r, \quad (2.76)$$

donde las constantes c_k , b_k y a_{kj} se eligen adecuadamente en función del orden de precisión que se quiera obtener. Cada una de las posibles elecciones determina un Método de Runge–Kutta (monopaso) de r etapas.

Observación 2.47 En general, el orden p del método no tiene por qué coincidir con el número r de etapas. $\square$

Observación 2.48 Una forma habitual de expresar los Métodos de Runge–Kutta de r etapas es a través de los *tableros de Butcher*

b_1	a_{11}	a_{12}	$\cdots$	a_{1r}
b_2	a_{21}	a_{22}	$\cdots$	a_{2r}
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
b_r	a_{r1}	a_{r2}	$\cdots$	a_{rr}
	c_1	c_2	$\cdots$	c_r

o, en forma matricial,

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

donde

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1r} \\ a_{21} & a_{22} & \cdots & a_{2r} \\ \cdots & \cdots & \cdots & \cdots \\ a_{r1} & a_{r2} & \cdots & a_{rr} \end{pmatrix}, \quad b = (b_1, b_2, \dots, b_r)^T \text{ y } c = (c_1, c_2, \dots, c_r)^T. \quad \square$$

Observación 2.49

a) Si se verifica que

$$a_{kj} = 0 \text{ cuando } j \geq k, \quad (2.77)$$

es decir, cuando la matriz A es triangular inferior con ceros en la diagonal principal, cada función F_k puede obtenerse de forma explícita en términos de las funciones anteriores $\{F_1, F_2, \dots, F_{k-1}\}$ y se dice que el Método de Runge–Kutta (2.74) es *explícito*. Son aquellos para los cuales

$$F_k(t, \xi, h) = f \left(t + hb_k, \xi + h \sum_{j=1}^{k-1} a_{kj} F_j(t, \xi, h) \right), \quad k = 1, 2, \dots, r. \quad (2.78)$$

Los diagramas de Butcher de los métodos explícitos son de la forma

$$\begin{array}{c|cccc} b_1 & 0 & 0 & \cdots & 0 & 0 \\ b_2 & a_{21} & 0 & \cdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ b_r & a_{r1} & a_{r2} & \cdots & a_{r,r-1} & 0 \\ \hline & c_1 & c_2 & \cdots & c_{r-1} & c_r \end{array} \quad (2.79)$$

b) Si no se verifica la condición (2.77) se dice que el Método de Runge–Kutta (2.74) es *implícito*.

c) En el caso particular de que se verifique

$$a_{kj} = 0 \text{ cuando } j > k,$$

es decir, cuando la matriz A es triangular inferior, se tiene que

$$F_k(t, \xi, h) = f \left(t + hb_k, \xi + h \sum_{j=1}^k a_{kj} F_j(t, \xi, h) \right), \quad k = 1, 2, \dots, r$$

y se dice que el Método de Runge–Kutta (2.74) es *semi-implícito*. Los diagramas de Butcher de los métodos semi-implícitos son de la forma

$$\begin{array}{c|cccc} b_1 & a_{11} & 0 & \cdots & 0 & 0 \\ b_2 & a_{21} & a_{22} & \cdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ b_r & a_{r1} & a_{r2} & \cdots & a_{r,r-1} & a_{rr} \\ \hline & c_1 & c_2 & \cdots & c_{r-1} & c_r \end{array}$$

d) Cuando el Método de Runge–Kutta no es explícito, las funciones $F_k(t, \xi, h)$ se obtienen a partir de (2.74), (2.75) y (2.76) (no sólo de (2.76)). En efecto, consideremos, por ejemplo, el Método de Runge–Kutta semi-implícito dado por el diagrama de Butcher

$$\begin{array}{c|c} \frac{1}{2} & \frac{1}{2} \\ \hline & 1 \end{array}$$

es decir,

$$x_{i+1} = x_i + hF_1(t_i, x_i, h), \quad (2.80)$$

donde

$$F_1(t_i, x_i, h) = f \left(t_i + \frac{h}{2}, x_i + \frac{h}{2} F_1(t_i, x_i, h) \right).$$

De (2.80) se tiene que

$$F_1(t_i, x_i, h) = \frac{x_{i+1} - x_i}{h},$$

de donde se deduce que

$$x_{i+1} = x_i + hf \left(t_i + \frac{h}{2}, x_i + \frac{h}{2} \frac{x_{i+1} - x_i}{h} \right).$$

Así pues, el método resultante es

$$x_{i+1} = x_i + hf \left(t_i + \frac{h}{2}, \frac{x_{i+1} + x_i}{2} \right). \quad \square$$

Observación 2.50 El Método de Runge–Kutta más simple es el Método de Euler (explícito), que es un Método de Runge–Kutta de 1 etapa y orden 1, dado por la función

$$\Phi(t, \xi, h) = c_1 f(t + hb_1, \xi)$$

con $c_1 = 1$ y $b_1 = 0$. $\square$

2.4.2 Análisis del error cometido

De manera análoga a lo hecho para los θ -Métodos y los métodos de Taylor, definimos el error local de truncamiento $\tau(t, h)$, asociado a una solución exacta $x(t)$ de la ecuación diferencial (2.7) a partir del instante $t \in [t_0, T]$, como

$$\tau(t, h) = \frac{x(t+h) - x(t)}{h} - \Phi(t, x(t), h), \quad (2.81)$$

para todo $h \in (0, T - t]$, donde Φ está dada en (2.75). Por otra parte, puesto que

$$\tau^{\text{Taylor}}(t, h) = \frac{x(t+h) - x(t)}{h} - T_p(t, x(t), h),$$

restando las dos expresiones anteriores se obtiene, bajo la hipótesis (2.73), que

$$\tau(t, h) = \tau^{\text{Taylor}}(t, h) + O(h^p),$$

para todo $(t, h) \in \mathcal{U}$. De esta forma, el Teorema 2.42 determina que, si $f \in \mathcal{C}^p([t_0, T] \times \mathbb{R}^n; \mathbb{R}^n)$, entonces el error local de truncamiento $\tau(h)$ dado en (2.15) verifica que

$$\tau(h) = \tau^{\text{Taylor}}(h) + O(h^p) = O(h^p),$$

por lo que el Método de Runge–Kutta considerado tiene orden de precisión p .

En todo lo que sigue, únicamente consideraremos Métodos de Runge–Kutta explícitos de r etapas, que vienen dados a partir de la función

$$\Phi(t, \xi, h) = \sum_{k=1}^r c_k F_k(t, \xi, h) = \sum_{k=1}^r c_k f \left(t + hb_k, \xi + h \sum_{j=1}^{k-1} a_{kj} F_j(t, \xi, h) \right) \quad (2.82)$$

(véase (2.78)).

Observación 2.51 Nótese que el número de evaluaciones de la función f que se necesitan aplicar en cada paso del Método de Runge–Kutta definido a partir de (2.82) coincide con el número r de etapas. $\square$

2.4.3 Métodos de Runge–Kutta explícitos de órdenes 1 y 2

Veamos cómo deben elegirse los coeficientes b_k , c_k y a_{kj} para que el método anterior sea, al menos, de orden 1. Para ello, se debe cumplir que

$$\Phi(t, \xi, h) = T_1(t, \xi, h) + O(h) = f(t, \xi) + O(h). \quad (2.83)$$

Si $f \in \mathcal{C}^1([t_0, T] \times \mathbb{R}^n; \mathbb{R}^n)$, un desarrollo de Taylor de primer orden de la función $g(h) = \Phi(t, \xi, h)$ determina que

$$g(h) = g(0) + O(h),$$

es decir,

$$\Phi(t, \xi, h) = \Phi(t, \xi, 0) + O(h) = \left(\sum_{k=1}^r c_k \right) f(t, \xi) + O(h)$$

(véase (2.82)). Por tanto, para que se cumpla (2.83), los coeficientes c_k deben ser tales que

$$\sum_{k=1}^r c_k = 1, \quad (2.84)$$

pudiendo elegirse los coeficientes b_k y a_{kj} de forma arbitraria. Hemos probado así el siguiente resultado:

Teorema 2.52 La condición necesaria y suficiente para que un Método de Runge–Kutta explícito de r etapas tenga, al menos, orden de precisión 1 es que se verifique (2.84). $\square$

Si además $f \in C^2([t_0, T] \times \mathbb{R}^n; \mathbb{R}^n)$ y queremos que el método anterior sea al menos de orden 2, debe verificarse que

$$\Phi(t, \xi, h) = T_2(t, \xi, h) + O(h^2), \quad (2.85)$$

siendo

$$T_2(t, \xi, h) = f^{(0)}(t, \xi) + \frac{h}{2} f^{(1)}(t, \xi) = f(t, \xi) + \frac{h}{2} \left(\frac{\partial f}{\partial t}(t, \xi) + D_\xi f(t, \xi) f(t, \xi) \right). \quad (2.86)$$

Un desarrollo de Taylor de segundo orden de la función $g(h) = \Phi(t, \xi, h)$, hace que

$$\Phi(t, \xi, h) = \Phi(t, \xi, 0) + \frac{\partial \Phi}{\partial h}(t, \xi, 0)h + O(h^2).$$

Puesto que, a partir de (2.82), se verifica que

$$\Phi(t, \xi, 0) = \left(\sum_{k=1}^r c_k \right) f(t, \xi)$$

y

$$\begin{aligned} \frac{\partial \Phi}{\partial h}(t, \xi, 0) &= c_1 b_1 \frac{\partial f}{\partial t}(t, \xi) + \sum_{k=2}^r c_k \left(b_k \frac{\partial f}{\partial t}(t, \xi) + D_\xi f(t, \xi) \left(\sum_{j=1}^{k-1} a_{kj} \overbrace{F_j(t, \xi, 0)}^{=f(t, \xi)} \right) \right) \\ &= \left(\sum_{k=1}^r c_k b_k \right) \frac{\partial f}{\partial t}(t, \xi) + \left(\sum_{k=2}^r c_k \sum_{j=1}^{k-1} a_{kj} \right) D_\xi f(t, \xi) f(t, \xi), \end{aligned}$$

se tiene que

$$\Phi(t, \xi, h) = \left(\sum_{k=1}^r c_k \right) f(t, \xi) + \left(\left(\sum_{k=1}^r c_k b_k \right) \frac{\partial f}{\partial t}(t, \xi) + \left(\sum_{k=2}^r c_k \sum_{j=1}^{k-1} a_{kj} \right) D_\xi f(t, \xi) f(t, \xi) \right) h + O(h^2).$$

Comparando esta expresión con (2.85) (teniendo en cuenta (2.86)), los coeficientes c_k , b_k y a_{kj} deben ser tales que

$$\begin{cases} \sum_{k=1}^r c_k = 1 \\ \sum_{k=1}^r c_k b_k = \frac{1}{2} \\ \sum_{k=2}^r c_k \sum_{j=1}^{k-1} a_{kj} = \frac{1}{2}. \end{cases} \quad (2.87)$$

Se tiene así el siguiente resultado:

Teorema 2.53 La condición necesaria y suficiente para que un Método de Runge–Kutta explícito de r etapas tenga, al menos, orden de precisión 2 es que se verifique (2.87). $\square$

Observación 2.54 Puesto que el sistema (2.87) tiene muchas más incógnitas que ecuaciones y, por tanto, en general no va a tener unicidad de solución, por simplicidad, suelen escogerse los coeficientes de forma que

$$\begin{cases} b_1 = 0 \\ b_k = \sum_{j=1}^{k-1} a_{kj}, \quad k = 2, 3, \dots, r. \end{cases} \quad (2.88)$$

De esta forma, los coeficientes b_k están unívocamente determinados una vez conocidos los coeficientes a_{kj} y las componentes del vector $b = (b_1, b_2, \dots, b_r)^T$ se obtienen sumando, respectivamente, las filas de la matriz $A = (a_{kj})_{k,j=1}^r$ en el diagrama de Butcher (2.79). En esta situación, basta encontrar soluciones del sistema

$$\begin{cases} \sum_{k=1}^r c_k = 1 \\ \sum_{k=2}^r c_k \sum_{j=1}^{k-1} a_{kj} = \frac{1}{2}. \quad \square \end{cases} \quad (2.89)$$

2.4.4 Métodos de Runge–Kutta explícitos de 2 etapas y orden 2

La fórmula general de un Método de Runge–Kutta explícito de 2 etapas es

$$x_{i+1} = x_i + h\Phi(t_i, x_i, h),$$

siendo

$$\Phi(t, \xi, h) = c_1 f(t + hb_1, \xi) + c_2 f(t + hb_2, \xi + ha_{21}f(t + hb_1, \xi)) \quad (2.90)$$

(véase (2.82)). Veamos cómo elegir los coeficientes c_1, c_2, b_1, b_2 y a_{21} para que el orden del método (2.90) sea, al menos, dos. Para $r = 2$ el sistema (2.87) toma la forma

$$\begin{cases} c_1 + c_2 = 1 \\ c_1 b_1 + c_2 b_2 = \frac{1}{2} \\ c_2 a_{21} = \frac{1}{2}, \end{cases}$$

cuya solución general es

$$c_1 = 1 - \alpha, \quad c_2 = \alpha, \quad b_1 = \beta, \quad b_2 = \frac{1 - 2(1 - \alpha)\beta}{2\alpha} \quad \text{y} \quad a_{21} = \frac{1}{2\alpha}.$$

Se obtiene así que la familia de Métodos de Runge–Kutta explícitos de 2 etapas y orden 2 viene definida a partir de la función

$$\Phi(t, \xi, h) = (1 - \alpha)f(t + h\beta, \xi) + \alpha f\left(t + h\frac{1 - 2(1 - \alpha)\beta}{2\alpha}, \xi + \frac{h}{2\alpha}f(t + h\beta, \xi)\right), \quad (2.91)$$

con $\alpha \in \mathbb{R} \setminus \{0\}$ y $\beta \in \mathbb{R}$ arbitrarios.

Observación 2.55 La simplificación considerada en (2.88) equivale a tomar $\beta = 0$. En tal caso, se tiene que

$$b_1 = \beta = 0 \quad \text{y} \quad b_2 = \frac{1 - 2(1 - \alpha)\beta}{2\alpha} = \frac{1}{2\alpha} = a_{21}$$

para $\alpha \neq 0$, lo que conduce a la subfamilia de Métodos de Runge–Kutta

$$\Phi(t, \xi, h) = (1 - \alpha)f(t, \xi) + \alpha f\left(t + \frac{h}{2\alpha}, \xi + \frac{h}{2\alpha}f(t, \xi)\right) \quad (2.92)$$

con $\alpha \in \mathbb{R} \setminus \{0\}$. Los Métodos de Runge–Kutta explícitos de 2 etapas y orden 2 más conocidos³ son los correspondientes a los valores:

a) $\alpha = 1$: Método de Euler modificado

Método de Euler modificado	
<u>Paso 1</u>	$x_0 \simeq \xi_0$
<u>Paso 2</u>	Para $i = 0, 1, \dots, N - 1$ $x_{i+1} = x_i + hf\left(t_i + \frac{h}{2}, x_i + \frac{h}{2}f(t_i, x_i)\right).$

b) $\alpha = \frac{3}{4}$: Método de Heun

³Hay que destacar que esta nomenclatura no es universal y los nombres de estos métodos pueden variar de unos textos a otros.

Método de Heun	
<u>Paso 1</u>	$x_0 \simeq \xi_0$
<u>Paso 2</u>	Para $i = 0, 1, \dots, N - 1$ $x_{i+1} = x_i + \frac{h}{4} \left(f(t_i, x_i) + 3f\left(t_i + \frac{2h}{3}, x_i + \frac{2h}{3}f(t_i, x_i)\right) \right).$

c) $\alpha = \frac{1}{2}$: Método de Euler mejorado

Método de Euler mejorado	
<u>Paso 1</u>	$x_0 \simeq \xi_0$
<u>Paso 2</u>	Para $i = 0, 1, \dots, N - 1$ $x_{i+1} = x_i + \frac{h}{2} (f(t_i, x_i) + f(t_{i+1}, x_i + hf(t_i, x_i))).$

(2.93)

Observación 2.56 Los tableros de Butcher asociados a los métodos anteriores son:

Método de Euler modificado

Método de Heun

Método de Euler mejorado

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

0	0	0
$\frac{2}{3}$	$\frac{2}{3}$	0
	$\frac{1}{4}$	$\frac{3}{4}$

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

Observación 2.57 Nótese que en los Métodos de Runge–Kutta de la familia anterior sólo se necesitan, en cada etapa, dos evaluaciones de la función f frente a las tres que se precisan, de f o sus derivadas, en el Método de Taylor de orden 2 cuando $n = 1$ (que son $f(t_i, x_i)$, $f_t(t_i, x_i)$ y $f_x(t_i, x_i)$). □

Observación 2.58 (Interpretación del Método de Euler mejorado) En el Método de Euler se utiliza la pendiente en el extremo izquierdo del intervalo $[t_i, t_{i+1}]$ para aproximar el valor de la solución $x(t)$ en el nodo t_{i+1} , a partir de su valor en el nodo t_i :

$$x(t_{i+1}) \simeq x(t_i) + hx'(t_i) = x(t_i) + hf(t_i, x(t_i)),$$

de donde se obtiene el Método de Euler (explícito)

$$x_{i+1} = x_i + hf(t_i, x_i). \quad (2.94)$$

Podemos mejorar esta técnica si, en lugar de considerar la pendiente en el extremo izquierdo del intervalo $[t_i, t_{i+1}]$, consideramos el promedio de las pendientes en los extremos del mismo:

$$x(t_{i+1}) \simeq x(t_i) + h \frac{x'(t_i) + x'(t_{i+1})}{2} = x(t_i) + h \frac{f(t_i, x(t_i)) + f(t_{i+1}, x(t_{i+1}))}{2},$$

de donde se obtiene el Método del Trapecio

$$x_{i+1} = x_i + h \frac{f(t_i, x_i) + f(t_{i+1}, x_{i+1})}{2}, \quad (2.95)$$

que proporciona, tal y como se ha visto en la Sección 2.2.4, una mejor estimación de x en el nodo t_{i+1} que la obtenida con el Método de Euler (explícito). Sin embargo, no es posible aplicar directamente la fórmula (2.95) ya que, por tratarse de un método implícito, se desconoce el valor de x_{i+1} . El Método de Euler mejorado da respuesta a este problema al estimar o “predecir” un valor de x_{i+1} por medio del Método de Euler (2.94) para luego usar este valor en el Método del Trapecio (2.95), obteniendo, así, una estimación mejorada o “corregida” de x_{i+1} . Por este motivo, al Método de Euler mejorado también se le conoce como *Método de predicción–corrección de Euler* (en el Capítulo 4 se estudiarán este tipo de métodos). La manera habitual de implementar (2.93) es

Método de Euler mejorado	
<u>Paso 1</u>	$x_0 \simeq \xi_0$
<u>Paso 2</u>	Para $i = 0, 1, \dots, N - 1$ $\begin{cases} F_1 = f(t_i, x_i) \\ F_2 = f(t_{i+1}, x_i + hF_1) \end{cases}$ $x_{i+1} = x_i + \frac{h}{2} (F_1 + F_2).$

Nótese que el orden de precisión de este método, al igual que el del Trapecio, es dos y tiene la ventaja de ser explícito. $\square$

Observación 2.59 Obviamente, existen Métodos de Runge–Kutta explícitos de 2 etapas y orden 2 que no forman parte de la subfamilia relativa a la función (2.92). Así ocurre, por ejemplo, cuando se toma $\alpha = \frac{3}{4}$ y $\beta = 2$, que da lugar a la función

$$\Phi(t, \xi, h) = \frac{1}{4} f(t + 2h, \xi) + \frac{3}{4} f\left(t, \xi + \frac{2h}{3} f(t + 2h, \xi)\right),$$

a partir de la cual se obtiene el correspondiente Método de Runge–Kutta, cuyo tablero de Butcher es

$$\begin{array}{c|cc} 2 & 0 & 0 \\ 0 & \frac{2}{3} & 0 \\ \hline & \frac{1}{4} & \frac{3}{4} \end{array}$$

Nótese que, al igual que sucedía con la familia de métodos dada por (2.92), en el método anterior sólo se necesitan, en cada etapa, dos evaluaciones de la función f (véanse la Observaciones 2.57 y 2.51). $\square$

Respecto al error global de discretización, se tiene el siguiente resultado:

Teorema 2.60 Sea $f \in \mathcal{C}([t_0, T] \times \mathbb{R}^n; \mathbb{R}^n)$ y lipschitziana en la segunda variable con constante de Lipschitz $L > 0$. Dado $\tilde{h} > 0$, el error de discretización global absoluto en la familia de Métodos de Runge–Kutta explícitos de 2 etapas definidos a partir de (2.91) verifica que

$$\varepsilon(h) \leq e^{(T-t_0)\psi(\tilde{h})} \|\varepsilon_0(h)\| + \frac{e^{(T-t_0)\psi(\tilde{h})} - 1}{(|1 - \alpha| + |\alpha|)L} \bar{\tau}(h) \quad (2.96)$$

para todo $h \in \mathcal{H}$ tal que $0 < h \leq \tilde{h}$, siendo

$$\psi(h) = (|1 - \alpha| + |\alpha|)L + \frac{L^2}{2} h, \quad h \geq 0.$$

DEMOSTRACIÓN. Si a la ecuación (2.81) multiplicada por h y particularizada en $t = t_i$ se le resta la ecuación del Método de Runge–Kutta de orden 2, se obtiene

$$x(t_{i+1}) - x_{i+1} = (x(t_i) - x_i) + h(\Phi(t_i, x(t_i), h) - \Phi(t_i, x_i, h)) + h\tau(t_i, h),$$

es decir,

$$\begin{aligned} \varepsilon_{i+1}(h) &= \varepsilon_i(h) + h(\Phi(t_i, x(t_i), h) - \Phi(t_i, x_i, h)) + h\tau(t_i, h) \\ &= \varepsilon_i(h) + h(1 - \alpha) \left[f(t_i + h\beta, x(t_i)) - f(t_i + h\beta, x_i) \right] \\ &\quad + h\alpha \left[f\left(t_i + h\frac{1 - 2(1 - \alpha)\beta}{2\alpha}, x(t_i) + \frac{h}{2\alpha} f(t_i + h\beta, x(t_i))\right) \right. \\ &\quad \left. - f\left(t_i + h\frac{1 - 2(1 - \alpha)\beta}{2\alpha}, x_i + \frac{h}{2\alpha} f(t_i + h\beta, x_i)\right) \right] + h\tau(t_i, h) \end{aligned} \quad (2.97)$$

para $i = 0, 1, \dots, N - 1$. De esta forma, puesto que f es lipschitziana en la segunda variable con constante de Lipschitz $L > 0$, entonces

$$\begin{aligned} \|\varepsilon_{i+1}(h)\| &\leq \|\varepsilon_i(h)\| + h|1 - \alpha|L \|\varepsilon_i(h)\| + h|\alpha|L \left(\|\varepsilon_i(h)\| + \frac{hL}{2|\alpha|} \|\varepsilon_i(h)\| \right) + h\|\tau(t_i, h)\| \\ &\leq \left(1 + h|1 - \alpha|L + h|\alpha|L \left(1 + \frac{hL}{2|\alpha|} \right) \right) \|\varepsilon_i(h)\| + h\bar{\tau}(h) \\ &= (1 + h\psi(h)) \|\varepsilon_i(h)\| + h\bar{\tau}(h) \end{aligned} \quad (2.98)$$

para $i = 0, 1, \dots, N - 1$. Por tanto, aplicando el Lema 2.22 con $y_i = \|\varepsilon_i(h)\|$,

$$A = h\psi(h) \text{ y } B = h\bar{\tau}(h) = A \frac{\bar{\tau}(h)}{\psi(h)},$$

para cada $i = 0, 1, \dots, N$ se obtiene que

$$\|\varepsilon_i(h)\| \leq e^{ih\psi(h)} \|\varepsilon_0(h)\| + \frac{e^{ih\psi(h)} - 1}{\psi(h)} \bar{\tau}(h),$$

de donde, utilizando que $\psi(h) \geq 0$ para $h \geq 0$, se deduce la estimación

$$\varepsilon(h) \leq e^{Nh\psi(h)} \|\varepsilon_0(h)\| + \frac{e^{Nh\psi(h)} - 1}{\psi(h)} \bar{\tau}(h) = e^{(T-t_0)\psi(h)} \|\varepsilon_0(h)\| + \frac{e^{(T-t_0)\psi(h)} - 1}{\psi(h)} \bar{\tau}(h). \quad (2.99)$$

Puesto que la función $\psi(h)$ verifica que

$$\psi'(h) = \frac{L^2}{2} > 0$$

para $h \geq 0$ y $\psi(0) = (|1 - \alpha| + |\alpha|)L > 0$, se tiene que

$$0 < (|1 - \alpha| + |\alpha|)L \leq \psi(h) \leq \psi(\tilde{h}) \text{ si } 0 \leq h \leq \tilde{h}$$

y, por tanto, a partir de (2.99), se obtiene la relación (2.96) para valores $0 < h \leq \tilde{h}$. $\square$

Observación 2.61 La expresión (2.96) relaciona el error de discretización global absoluto con el error local de truncamiento y el (posible) error en los datos iniciales. $\square$

Observación 2.62 Bajo las hipótesis del Teorema 2.60, puesto que

$$|1 - \alpha| + |\alpha| \begin{cases} = 1 & \text{si } \alpha \in (0, 1] \\ > 1 & \text{si } \alpha \notin [0, 1], \end{cases}$$

a partir de la relación (2.96) se verifica que

$$\varepsilon(h) \leq e^{(T-t_0)\psi(\tilde{h})} \|\varepsilon_0(h)\| + \frac{e^{(T-t_0)\psi(\tilde{h})} - 1}{L} \tau(h)$$

para todo $0 < h \leq \tilde{h}$. $\square$

Corolario 2.63 Si la función $f \in C^2([t_0, T] \times \mathbb{R}^n; \mathbb{R}^n)$ y es lipschitziana en la segunda variable, se verifica que la familia de Métodos de Runge–Kutta explícitos de dos etapas definidos a partir de (2.91) es convergente. Además, si $\|\varepsilon_0(h)\| = O(h^2)$, se tiene que $\varepsilon(h) = O(h^2)$.

DEMOSTRACIÓN. Basta aplicar el Teorema 2.60. $\square$

2.4.5 Otros métodos de Runge–Kutta

La fórmula general de un Método de Runge–Kutta explícito de 3 etapas es

$$x_{i+1} = x_i + h\Phi(t_i, x_i, h),$$

siendo

$$\begin{aligned} \Phi(t, \xi, h) = & c_1 f(t + hb_1, \xi) + c_2 f(t + hb_2, \xi + ha_{21}f(t + hb_1, \xi)) \\ & + c_3 f\left(t + hb_3, \xi + ha_{31}f(t + hb_1, \xi) + ha_{32}f(t + hb_2, \xi + ha_{21}f(t + hb_1, \xi))\right). \end{aligned}$$

En este caso, nos planteamos encontrar los coeficientes $c_1, c_2, c_3, b_1, b_2, b_3, a_{21}, a_{31}$ y a_{32} para que el orden del método anterior sea, al menos, tres. Argumentando de forma análoga a como se ha hecho en la Sección 2.4.4, bajo la hipótesis de simplificación de suma por filas

$$\begin{cases} b_1 = 0 \\ b_2 = a_{21} \\ b_3 = a_{31} + a_{32} \end{cases}$$

(véase (2.88)), se deduce (tras cálculos algo tediosos) que los restantes 6 coeficientes son solución del sistema

$$\begin{cases} c_1 + c_2 + c_3 = 1 \\ c_2 b_2 + c_3 b_3 = \frac{1}{2} \\ c_2 b_2^2 + c_3 b_3^2 = \frac{1}{3} \\ c_3 b_2 a_{32} = \frac{1}{6}. \end{cases}$$

En particular, una solución de los sistemas anteriores viene dada por

$$c_1 = \frac{2}{9}, \quad c_2 = \frac{1}{3}, \quad c_3 = \frac{4}{9}, \quad b_1 = 0, \quad b_2 = \frac{1}{2}, \quad b_3 = \frac{3}{4}, \quad a_{21} = \frac{1}{2}, \quad a_{31} = 0 \quad \text{y} \quad a_{32} = \frac{3}{4},$$

que dá lugar al siguiente método:

Método de Runge–Kutta de 3 etapas y orden 3	
<u>Paso 1</u>	$x_0 \simeq \xi_0$
<u>Paso 2</u>	Para $i = 0, 1, \dots, N - 1$
	$\begin{cases} F_1 = f(t_i, x_i) \\ F_2 = f\left(t_i + \frac{h}{2}, x_i + \frac{h}{2} F_1\right) \\ F_3 = f\left(t_i + \frac{3h}{4}, x_i + \frac{3h}{4} F_2\right) \end{cases}$
	$x_{i+1} = x_i + \frac{h}{9} (2F_1 + 3F_2 + 4F_3),$

(2.100)

cuyo tablero de Butcher es

0	0	0	0
$\frac{1}{2}$	$\frac{1}{2}$	0	0
$\frac{3}{4}$	0	$\frac{3}{4}$	0
$\frac{2}{9}$	$\frac{1}{3}$	$\frac{4}{9}$	

Aunque los Métodos de Runge–Kutta de cuatro etapas y orden cuarto se deducen de manera semejante a los de segundo y tercer orden, la complejidad de las operaciones aumenta considerablemente dado que, al tener que comparar términos hasta h^4 , se obtiene un sistema de 10 ecuaciones con 14 incógnitas (hay, por tanto, cuatro grados de libertad). Con el fin de evitar muchas operaciones de cálculo que pueden distraer la atención del lector sin aportar ningún aspecto nuevo relevante, no deduciremos estos métodos, sino que presentamos, a continuación, el método de orden cuatro más utilizado (puede demostrarse que, efectivamente, es de orden 4 y, además, convergente):

Método de Runge–Kutta de 4 etapas y orden 4 (clásico)	
<u>Paso 1</u>	$x_0 \simeq \xi_0$
<u>Paso 2</u>	Para $i = 0, 1, \dots, N - 1$
	$\begin{cases} F_1 = f(t_i, x_i) \\ F_2 = f\left(t_i + \frac{h}{2}, x_i + \frac{h}{2} F_1\right) \\ F_3 = f\left(t_i + \frac{h}{2}, x_i + \frac{h}{2} F_2\right) \\ F_4 = f(t_i + h, x_i + h F_3) \end{cases}$
	$x_{i+1} = x_i + \frac{h}{6} (F_1 + 2F_2 + 2F_3 + F_4),$

cuyo tablero de Butcher es

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

Observación 2.64 Como ya se vio en la derivación de los Métodos de Runge–Kutta explícitos de dos etapas y orden dos, no hay un método de Runge–Kutta de un orden dado, sino una familia de métodos de Runge–Kutta de ese orden (debido a que hay ciertos parámetros que deben seleccionarse). En la siguiente tabla se observa cómo el número de etapas requeridas aumenta más rápidamente que el correspondiente orden máximo de los métodos de Runge–Kutta explícitos:

Número de etapas	1	2	3	4	5	6	7	8	9
Orden máximo del Método de Runge–Kutta explícito	1	2	3	4	4	5	6	6	7

(véase [Kincaid–Cheney]). En general, puede demostrarse que un Método de Runge–Kutta explícito de r etapas no puede tener un orden de precisión superior a r (véase [Butcher] ó [Hairer–Norsett–Wanner]). Esto hace que el método de Runge–Kutta que se use con mayor frecuencia sea el método clásico de Runge–Kutta de orden 4: se trata de un método muy utilizado que da excelentes resultados. Los métodos de Runge–Kutta de orden superior como, por ejemplo,

Método de Runge–Kutta de 6 etapas y orden 5	
Paso 1	$x_0 \simeq \xi_0$
Paso 2	Para $i = 0, 1, \dots, N - 1$
	$\left\{ \begin{array}{l} F_1 = f(t_i, x_i) \\ F_2 = f\left(t_i + \frac{h}{2}, x_i + \frac{h}{2} F_1\right) \\ F_3 = f\left(t_i + \frac{h}{2}, x_i + \frac{h}{4} (F_1 + F_2)\right) \\ F_4 = f\left(t_i + h, x_i + h(-F_2 + 2F_3)\right) \\ F_5 = f\left(t_i + \frac{2h}{3}, x_i + \frac{h}{27} (7F_1 + 10F_2 + F_4)\right) \\ F_6 = f\left(t_i + \frac{h}{5}, x_i + \frac{h}{625} (28F_1 - 125F_2 + 546F_3 + 54F_4 - 378F_5)\right) \end{array} \right.$
	$x_{i+1} = x_i + \frac{h}{336} (14F_1 + 35F_4 + 162F_5 + 125F_6)$

son menos atractivos debido al aumento de su coste computacional con respecto a otros métodos, como el anterior de orden 4. $\square$