Análisis Numérico

Universidad de Málaga

Grado en Matemáticas

Febrero de 2024

Contenidos

TEMA 1

Introducción

1.1. Problemas de Cauchy

El principal propósito de esta asignatura es el estudio de métodos numéricos que permitan aproximar soluciones de ecuaciones diferenciales ordinarias (EDO). Concretamente, se tratarán problemas del estilo

$$(P)\begin{cases} y'(t) = f(t, y(t)), \ t \in [t_0, t_0 + T], \\ y(t_0) = y^0, \end{cases}$$

siendo $(t_0, y^0) \in \mathbb{R} \times \mathbb{R}^n$ y $f : [t_0, t_0 + T] \times \mathbb{R}^n \to \mathbb{R}^n$. Problemas de este tipo, habitualmente denominados *problemas de Cauchy*, ya han sido estudiados en asignaturas precedentes. Es por esto que en absoluto nos adentraremos en asuntos relacionados con existencia y unicidad de soluciones; simplemente se recuerdan algunas nociones básicas necesarias para asegurar que el problema (P) posee una única solución.

Definición 1. Fijada una norma vectorial $||\cdot||$ en $\mathbb{R}^n$, una función $f : [t_0, t_0 + T] \times \mathbb{R}^n \to \mathbb{R}^n$ se dice que es **de Lipschitz en la variable y** si existe una constante $L > 0$ tal que

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

para todo $t \in [t_0, t_0 + T]$ y todos $y_1, y_2 \in \mathbb{R}^n$.

Teorema 1. *Si una función $f : [t_0, t_0 + T] \times \mathbb{R}^n \to \mathbb{R}^n$ es continua y de Lipschitz en la variable y, entonces el problema (P) tiene solución única en $[t_0, t_0 + T]$.*

Demostración. Corresponde a la asignatura *Ecuaciones Diferenciales II*. □

De aquí en adelante, la función f que figura en el problema (P) se supondrá continua y de Lipschitz en la variable y. Ya se puede asegurar la existencia y unicidad de una única solución del problema (P). El asunto que ahora nos ocupa es el de aproximar dicha solución.

1.2. Algunos modelos matemáticos basados en EDO

Ejemplo. Se tiene un recipiente con 20 litros de agua pura en un instante inicial. Supóngase que

- ***(i)*** entra agua salada en el recipiente a razón de 2 litros por segundo con una concentración de 3 gramos por litro;
- ***(ii)*** sale agua salada del recipiente a razón de 2 litros por segundo.

Se trata de encontrar la función $s(t)$ que proporcione la cantidad de sal que hay en el depósito en cada instante t. En primer lugar, la velocidad a la que entra sal en el recipiente es de 6 gramos por segundo, mientras que la velocidad a la que sale sal del recipiente es de

$$\underbrace{\frac{s(t)}{20}}_{g/l} \cdot \underbrace{2}_{l/s} = \frac{s(t)}{10}$$

gramos por segundo. Por tanto, $s'(t) = 6 - \frac{s(t)}{10}$ es la velocidad a la que varía la sal que hay en el recipiente. Así, el problema a resolver es

$$(P)\begin{cases} s'(t) = 6 - \dfrac{s(t)}{10}, \\ s(0) = 0, \end{cases}$$

y su solución, como se comprueba fácilmente, sería $s(t) = 60(1 - e^{-\frac{1}{10}t})$.

Ejemplo. Sea $x(t)$ el número de individuos de una cierta población, y sean k_n y k_m las respectivas tasas de natalidad y mortalidad de la población. En ausencia de depredadores, los modelos para esta población aislada son ecuaciones de la forma

$$x'(t) = k_n x - k_m x,$$

así que, de ser k_n y k_m constantes, fijando un dato inicial $(t_0, x_0) \in \mathbb{R}^2$ y llamando $k = k_n - k_m$, la solución sería $x(t) = x_0 e^{kt}$.

Ejemplo. La casuística es la siguiente:

- ***(i)*** Hay dos especies que interactúan: las presas, $x(t)$, y los depredadores, $y(t)$.
- ***(ii)*** La tasa de natalidad de las presas, k_n^p, es constante, y la tasa de mortalidad es $k_m^p = c + \beta y$, con $c, \beta > 0$.
- ***(iii)*** La tasa de natalidad de los depredadores es $k_n^d = d + \delta x$, con $d, \delta > 0$, mientras que la tasa de mortalidad, k_m^d, es constante.

Las ecuaciones serían, siguiendo el modelo anterior, $x' = k_n^p x - k_m^p x$ e $y' = k_n^d y - k_m^d y$. Sustituyendo y llamando $x(0) = x_0$ e $y(0) = y_0$, el problema a resolver no es más que

$$(LV)\begin{cases} x' = \alpha x - \beta x y, & x(0) = x_0, \\ y' = -\gamma y + \delta x y, & y(0) = y_0, \end{cases}$$

donde $\alpha = k_n^p - c$ y $\gamma = k_m^d - d$. Este modelo se conoce como *modelo presa-depredador de Lotka-Volterra*.

1.3. Método de Euler

El proceso habitual para aproximar $y\colon [t_0, t_0 + T] \to \mathbb{R}$, la única solución del problema de partida (P) en el caso unidimensional, consistirá en tomar una partición uniforme $t_0 < t_1 < \ldots < t_{n-1} < t_n = t_0 + T$ del intervalo $[t_0, t_0 + T]$ y, para cada $k \in \{0, 1, \ldots, n\}$, obtener una aproximación de $y(t_k)$, que será denotada por y_k.

La longitud de cada subintervalo de la partición se denomina *paso de malla*, y habitualmente se denota por h. Así, para cada $k \in \{0, 1, \ldots, n\}$, se tiene

$$t_k = t_0 + kh$$

A bote pronto, la aproximación más sencilla de la gráfica de y en el intervalo $[t_0, t_1]$ sería la recta tangente a la gráfica de la solución en (t_0, y_0). La ordenada del punto de dicha recta con abscisa t_1 es

$$y_1 = y_0 + f(t_0, y_0)(t_1 - t_0) = y_0 + hf(t_0, y_0)$$

Al repetir esta interpolación en los demás subintervalos de la partición, se obtiene una aproximación razonable de la solución de (P) en el intervalo $[t_0, t_0 + T]$.

Definición 2. El **método de Euler** es aquel definido por

$$y_{k+1} = y_k + hf(t_k, y_k), \quad k \in \{0, 1, \ldots, n-1\}$$

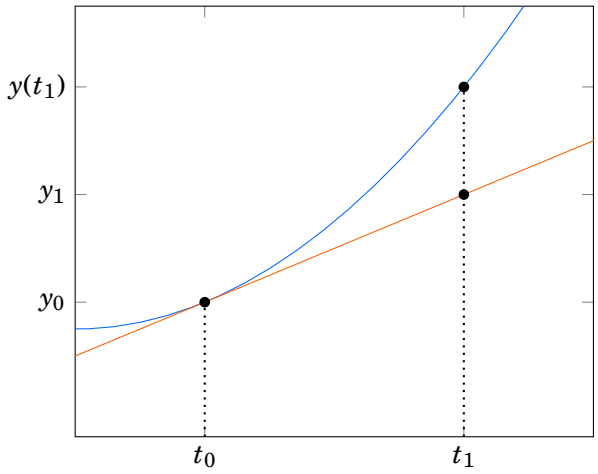

Figura 1. Interpretación gráfica del método de Euler.

A continuación, se tratará de estudiar cómo de buenas son las aproximaciones proporcionadas por el método de Euler. Para ello, es conveniente recordar nociones básicas sobre polinomios de Taylor.

Definición 3. Sea $I \subset \mathbb{R}$ un intervalo y sea $g : I \to \mathbb{R}$ una función n veces derivable en un punto $t_0 \in I$. Se define el **polinomio de Taylor de la función g de grado $n \in \mathbb{N}$ centrado en t_0** como

$$P_{g,n,t_0}(t) = \sum_{k=0}^{n} \frac{g^{(k)}(t_0)}{k!}(t-t_0)^k$$

Teorema 2 (Fórmula del resto de Lagrange). *Sea $I \subset \mathbb{R}$ un intervalo abierto y sea $g \in \mathcal{C}^{n+1}(I)$. Dados $t, t_0 \in I$, existe ξ en el intervalo abierto de extremos t_0 y t tal que*

$$g(t) - P_{g,n,t_0}(t) = \frac{g^{(n+1)}(\xi)}{(n+1)!}(t-t_0)^{n+1}$$

Demostración. Corresponde a la asignatura *Análisis Matemático II*. □

Teorema 3. *Sean $\{y_k\}_{k=0}^n$ las aproximaciones obtenidas mediante el método de Euler. Si la función f del problema (P) es de clase 1, entonces existe una constante $K > 0$, independiente de h, tal que*

$$e(h) := \max_{k=0,1,\dots,n} |y(t_k) - y_k| \leq Kh$$

Demostración. Para cada $k \in \{0, 1, \dots, n\}$, definiremos

$$e_k := |y(t_k) - y_k|$$

Evidentemente, $e_0 = 0$. En $k = 1$, se tiene

$$e_1 = |y(t_1) - y_1| = |y(t_1) - y_0 - hf(t_0, y_0)| = |y(t_1) - y(t_0) - hy'(t_0)| = |y(t_1) - P_{y,1,t_0}(t_1)|$$

Por la fórmula del resto de Lagrange, existe $\xi_0 \in (t_0, t_1)$ tal que

$$e_1 = \left| \frac{y''(\xi_0)}{2} h^2 \right| = \frac{|y''(\xi_0)|}{2} h^2$$

Por ser f de clase 1 y tenerse $y' = f(t, y)$, puede afirmarse que y' tiene derivada continua en $[t_0, t_0 + T]$, luego $|y''|$ es continua en el compacto $[t_0, t_0 + T]$, y por tanto alcanza el máximo. Por tanto, $e_1 \leq Ch^2$, donde

$$C := \frac{1}{2} \max_{t \in [t_0, t_0+T]} |y''(t)|$$

En $k=2$, se verifica

$$\begin{aligned}
e_2 &= |y(t_2)-y_2| \\
&= |y(t_2)-y_1-hf(t_1,y_1)| \\
&= |y(t_2)+y(t_1)-y(t_1)+hf(t_1,y(t_1))-hf(t_1,y(t_1))-y_1-hf(t_1,y_1)| \\
&\leq |y(t_2)-y(t_1)-hf(t_1,y(t_1))|+|y(t_1)-y_1|+h|f(t_1,y(t_1))-f(t_1,y_1)| \\
&= \underbrace{|y(t_2)-P_{y,1,t_1}(t_2)|}_{I} + e_1 + \underbrace{h|f(t_1,y(t_1))-f(t_1,y_1)|}_{II}
\end{aligned}$$

Para acotar I, se razona como antes: por la fórmula del resto de Lagrange, existe $\xi_1 \in (t_1,t_2)$ tal que

$$|y(t_2)-P_{y,1,t_1}(t_2)| = \frac{|y''(\xi_1)|}{2}h^2 \leq Ch^2$$

Para acotar II, se recuerda que f es de Lipschitz en la variable y, así que existe $L>0$ verificando

$$h|f(t_1,y(t_1))-f(t_1,y_1)| \leq hL|y(t_1)-y_1| = hLe_1,$$

y en consecuencia,

$$e_2 \leq Ch^2+(1+hL)e_1$$

Así, es fácil probar que para cada $k \in \{1,2,\ldots,n\}$ se verifica

$$e_k \leq Ch^2+(1+hL)e_{k-1},$$

luego

$$\begin{aligned}
e_k &\leq Ch^2+(1+hL)e_{k-1} \\
&\leq Ch^2+(1+hL)(Ch^2+(1+hL)e_{k-2}) = Ch^2+(1+hL)Ch^2+(1+hL)^2e_{k-2} \\
&\leq Ch^2+(1+hL)Ch^2+(1+hL)^2(Ch^2+(1+hL)e_{k-3}) \\
&\leq \ldots \\
&\leq Ch^2\sum_{j=0}^{k-1}(1+hL)^j = Ch^2\frac{1-(1+hL)^k}{1-1-hL} = Ch\frac{(1+hL)^k-1}{L} \\
&\overset{(*)}{\leq} Ch\frac{(1+hL)^n-1}{L} \\
&\overset{(**)}{\leq} Ch\frac{e^{nhL}-1}{L} = Ch\frac{e^{TL}-1}{L} = Kh,
\end{aligned}$$

donde

$$K = C\frac{e^{TL}-1}{L}$$

es una constante positiva que no depende de h. Un par de aclaraciones:

- $(*)$ Como $h>0$ y $L>0$, entonces $1+hL>1$ y por tanto $(1+hL)^k \leq (1+hL)^n$ al ser $k \leq n$.
- $(**)$ Se ha usado la desigualdad $1+x \leq e^x$ para todo $x \in \mathbb{R}$, de donde se deduce, elevando a $n \in \mathbb{N}$, que $(1+x)^n \leq e^{nx}$. Para $x \leq 0$, es evidente que $1+x \leq e^x$; si $x>0$, no hay más que recordar el desarrollo en serie de Taylor de la función exponencial:

$$e^x = \sum_{n=0}^{\infty}\frac{x^n}{n!} > \sum_{n=0}^{1}\frac{x^n}{n!} = 1+x$$

Así, como se ha probado que $e_k \leq Kh$ para todo $k \in \{0,1,\ldots,n\}$, el teorema está demostrado. □

1.4. Métodos de Taylor

Partiendo de las mismas condiciones de la sección anterior, como para todo $k = 0, 1, \ldots, n-1$ se tiene $y'(t_k) = f(t_k, y(t_k))$, puede escribirse

$$y_{k+1} = y_k + hf(t_k, y_k) = y_k + y'(t_k)(t_{k+1} - t_k) = P_{y,1,t_k}(t_{k+1})$$

Naturalmente, si en lugar de polinomios de Taylor de primer grado se usan polinomios de mayor grado, las aproximaciones obtenidas deberían ser más precisas. Queremos definir un método mediante

$$y_{k+1} = y_k + \sum_{q=1}^{p} \frac{y^{(q)}(t_k)}{q!} h^q,$$

pero antes es necesario conocer las derivadas de orden superior de y. Siempre que la función f sea lo suficientemente regular, se puede derivar en la ecuación $y'(t) = f(t, y(t))$, obteniéndose

$$y''(t) = \frac{d}{dt} f(t, y(t)) = f_t(t, y(t)) + y'(t) \cdot f_y(t, y(t)) = f_t(t, y(t)) + f(t, y(t)) \cdot f_y(t, y(t)) \tag{1}$$

Derivando otra vez y omitiendo los puntos en los que se evalúan las funciones por motivos de comodidad, se llega a

$$y''' = f_{tt} + f \cdot f_{ty} + (f_t + f \cdot f_y) \cdot f_y + f \cdot (f_{yt} + f \cdot f_{yy}), \tag{2}$$

lo que sugiere lo siguiente:

Notación. Si $f : [t_0, t_0 + T] \times \mathbb{R} \to \mathbb{R}$ tiene derivadas parciales de orden q para todo $q \in \{0, 1, \ldots, p\}$, se define

$$f^{(0)}(t, y) := f(t, y),$$

y para $q \in \{0, 1, \ldots, p-1\}$, se define

$$f^{(q+1)}(t, y) := f_t^{(q)}(t, y) + f(t, y) f_y^{(q)}(t, y)$$

Así, la expresión (1) afirma que $y''(t) = f^{(1)}(t, y(t))$, mientras que (2) se traduce en $y'''(t) = f^{(2)}(t, y(t))$. En general, si f es p veces derivable, es claro que para todo $q \in \{0, 1, \ldots, p\}$ se verifica

$$y^{(q+1)}(t) = f^{(q)}(t, y(t))$$

Ya estamos en condiciones de definir el método deseado:

Definición 4. Si $p \in \mathbb{N}$ y f tiene derivadas parciales de hasta orden p, el **método de Taylor de orden *p*** es aquel definido por

$$y_{k+1} = y_k + \sum_{q=1}^{p} \frac{f^{(q-1)}(t_k, y_k)}{q!} h^q, \quad k \in \{0, 1, \ldots, n-1\}$$

Ejemplo. Se trata de aproximar la solución del problema

$$(P) \begin{cases} y' = \frac{1}{2}(t^2 - y) \\ y(0) = 1 \end{cases}$$

mediante el método de Taylor de grado 2. Va a ser necesario conocer la segunda derivada de y, la cual puede obtenerse derivando en la ecuación del problema: para todo $t \in \mathbb{R}$ se tiene

$$y''(t) = t - \frac{y'(t)}{2} = t - \frac{1}{4}(t^2 - y(t))$$

Por tanto,

$$\begin{aligned} P_{y,2,t_k}(t_{k+1}) &= y(t_k) + y'(t_k)h + \frac{y''(t_k)}{2}h^2 \\ &= y(t_k) + \frac{1}{2}(t_k^2 - y(t_k))h + \frac{1}{2}(t_k - \frac{1}{2}y'(t_k))h^2 \\ &= y(t_k) + \frac{1}{2}(t_k^2 - y(t_k))h + \frac{1}{2}(t_k - \frac{1}{4}t_k^2 + \frac{1}{4}y(t_k))h^2 \end{aligned}$$

En consecuencia, el método de Taylor de segundo orden viene dado por

$$y_{k+1} = y_k + \frac{1}{2}(t_k^2 - y_k)h + \frac{1}{2}(t_k - \frac{1}{4}t_k^2 + \frac{1}{4}y_k)h^2$$

para cada $k \in \{0, 1, \ldots, n-1\}$.

Ejemplo. Considerando el problema del ejemplo anterior y tomando $h = 0.1$, se trata de aproximar $y(0.1)$ haciendo uso de

(i) el método de Euler con $h = 0.1$:

$$y(0.1) \approx y_0 + \frac{1}{2}(t_0^2 - y_0)h = 1 + \frac{1}{2}(0^2 - 1)0.1 = 0.95$$

(ii) el método de Taylor de segundo orden:

$$y(0.1) \approx y_0 + \frac{1}{2}(t_0^2 - y_0)h + \frac{1}{2}(t_0 - \frac{1}{4}t_0^2 + \frac{1}{4}y_0)h^2 = 0.95125$$

(iii) el método de Taylor de tercer orden:

$$y(0.1) \approx y_0 + \frac{1}{2}(t_0^2 - y_0)h + \frac{1}{2}(t_0 - \frac{1}{4}t_0^2 + \frac{1}{4}y_0)h^2 + \frac{1}{6}(1 - \frac{1}{2}t_k + \frac{1}{8}t_k^2 + \frac{1}{8}y_k)h^3 = \ldots$$

1.5. Método del punto medio y método de Heun

Partiendo de las condiciones de la sección anterior y fijando $k = 0, 1, \ldots, n-1$, como se está suponiendo que f es continua, entonces, para cualquier $t^* \in [t_0, t_0 + T]$, el problema (P) y la ecuación

$$y(t) = y(t^*) + \int_{t^*}^{t} f(s, y(s))\, ds$$

son totalmente equivalentes, suponiendo que $y(t^*)$ es conocido. En particular,

$$y(t_{k+1}) = y(t_k) + \int_{t_k}^{t_{k+1}} f(t, y(t))\, dt, \tag{3}$$

así que aproximar $y(t_{k+1})$ es equivalente a aproximar la integral de la derecha. Para este propósito, se va a utilizar, por ejemplo, la *fórmula del rectángulo a la izquierda*, que consiste en

$$\int_a^b g(x)\, dx \approx g(a)(b-a)$$

Si se aplica esta fórmula en (3), la aproximación obtenida es precisamente la del método de Euler, o sea, $y_{k+1} = y_k + hf(t_k, y_k)$. Esto sugiere un nuevo procedimiento de obtención de aproximaciones de y: usar fórmulas de integración más precisas que la del rectángulo a la izquierda. Repitamos el razonamiento de antes pero echando mano de la *fórmula del punto medio*, es decir,

$$\int_a^b g(x)\, dx \approx g\big(\frac{b+a}{2}\big)(b-a),$$

que aplicándose en (3) quedaría

$$y_{k+1} = y_k + hf\big(t_k + \frac{h}{2}, y\big(t_k + \frac{h}{2}\big)\big)$$

El valor de y en el punto $t_k + \frac{h}{2}$ tampoco se conoce, pero puede estimarse mediante el método de Euler. Se tendría entonces

$$y_{k+\frac{1}{2}} = y_k + \frac{h}{2} f(t_k, y_k),$$

y sustituyendo arriba, tenemos un método más:

Definición 5. El **método del punto medio** es aquel definido por

$$y_{k+1} = y_k + hf\big(t_k + \frac{h}{2}, y_k + \frac{h}{2} f\big(t_k, y_k\big)\big), \quad k \in \{0, 1, \ldots, n-1\},$$

o, alternativamente,

$$\begin{cases} y_{k+\frac{1}{2}} = y_k + \frac{h}{2} f(t_k, y_k) \\ y_{k+1} = y_k + hf\big(t_k + \frac{h}{2}, y_{k+\frac{1}{2}}\big) \end{cases}$$

Ahora, en lugar de utilizar la fórmula del punto medio, probemos a aproximar la integral que aparece en (3) mediante la *fórmula del trapecio*, es decir,

$$\int_a^b g(x)\,dx \approx (b-a)\frac{g(a)+g(b)}{2},$$

lo que proporcionaría la expresión

$$y_{k+1} = y_k + \frac{h}{2}(f(t_k, y_k) + f(t_{k+1}, y(t_{k+1})))$$

El mismo inconveniente de antes: hay que realizar una nueva aproximación porque no se sabe quién es $y(t_{k+1})$. Vuélvase a recurrir al método de Euler para aproximar $y(t_{k+1})$, y hállese así un nuevo método:

Definición 6. El **método de Heun** es aquel definido por

$$y_{k+1} = y_k + \frac{h}{2}(f(t_k, y_k) + f(t_k + h, y_k + hf(t_k, y_k))), \qquad k \in \{0, 1, \ldots, n-1\},$$

o, alternativamente,

$$\begin{cases} y^*_{k+1} = y_k + hf(t_k, y_k) \\ y_{k+1} = y_k + \frac{h}{2}(f(t_k, y_k) + f(t_{k+1}, y^*_{k+1})) \end{cases}$$

1.6. Métodos de Runge-Kutta

Al igual que antes, el procedimiento de obtención de los métodos de Runge-Kutta tratará de aproximar la integral de (3) mediante una cierta fórmula de integración. Concretamente, se va a emplear una *fórmula de cuadratura de p puntos*, o sea, una fórmula de la forma

$$\int_a^b g(x)\,dx \approx (b-a)\sum_{i=1}^{p} b_i g(a + c_i(b-a)),$$

donde $c_i \in [0,1]$ y $b_i \in \mathbb{R}$. Se dice que $a + c_i(b-a)$ son los *nodos* y b_i los *pesos*. Al sustituir en la integral de (3) quedaría

$$\int_{t_k}^{t_{k+1}} f(t, y(t))\,dt \approx h\sum_{i=1}^{p} b_i f(t_k + c_i h, y(t_k + c_i h)) \tag{4}$$

Esta aproximación no sirve de mucho si no se conocen los valores $y(t_k + c_i h)$, con $i \in \{0,1,\dots,p\}$. Para aproximarlos, se procede de la misma forma pero en el intervalo $[t_k, t_k + c_i h]$. Se tiene que

$$y(t_k + c_i h) = y(t_k) + \int_{t_k}^{t_k + c_i h} f(t, y(t))\, dt$$

Ahora se vuelve a aplicar una fórmula de cuadratura de p puntos con los mismos nodos de antes y pesos nuevos, quedando

$$\int_{t_k}^{t_k + c_i h} f(t, y(t))\, dt \approx h \sum_{j=1}^{p} a_{i,j} f(t_k + c_j h, y(t_k + c_j h)) \tag{5}$$

Así, se obtienen las aproximaciones

$$y(t_k + c_i h) \approx y(t_k) + h \sum_{j=1}^{p} a_{i,j} f(t_k + c_j h, y(t_k + c_j h))$$

Al sustituir en (4) y volver a (3), se obtiene el método siguiente:

Definición 7. Sea $p \in \mathbb{N}$, sean $a_{i,j}, b_i, c_i \in \mathbb{R}$ con $c_i \in [0,1]$ para $i,j \in \{1,2,\dots,p\}$ y llamemos

$$t_k^{(i)} := t_k + c_i h, \quad i \in \{1,2,\dots,p\}, k \in \{0,1,\dots,n\}$$

El **método de Runge-Kutta de p etapas** es aquel definido por

$$\begin{cases} y_k^{(1)} = y_k + h(a_{1,1} f(t_k^{(1)}, y_k^{(1)}) + a_{1,2} f(t_k^{(2)}, y_k^{(2)}) + \dots + a_{1,p} f(t_k^{(p)}, y_k^{(p)})) \\ y_k^{(2)} = y_k + h(a_{2,1} f(t_k^{(1)}, y_k^{(1)}) + a_{2,2} f(t_k^{(2)}, y_k^{(2)}) + \dots + a_{2,p} f(t_k^{(p)}, y_k^{(p)})) \\ \quad \vdots \\ y_k^{(p)} = y_k + h(a_{p,1} f(t_k^{(1)}, y_k^{(1)}) + a_{p,2} f(t_k^{(2)}, y_k^{(2)}) + \dots + a_{p,p} f(t_k^{(p)}, y_k^{(p)})) \\ y_{k+1} = y_k + h(b_1 f(t_k^{(1)}, y_k^{(1)}) + \dots + b_p f(t_k^{(p)}, y_k^{(p)})), \end{cases}$$

o lo que es lo mismo,

$$y_{k+1} = y_k + h(b_1 k_1 + \dots + b_p k_p),$$

donde

$$\begin{aligned} k_1 &= f(t_k^{(1)}, y_k + h(a_{11} k_1 + \dots + a_{1p} k_p)) \\ k_2 &= f(t_k^{(2)}, y_k + h(a_{21} k_1 + \dots + a_{2p} k_p)) \\ &\vdots \\ k_p &= f(t_k^{(p)}, y_k + h(a_{p1} k_1 + \dots + a_{pp} k_p)) \end{aligned}$$

El método más célebre de la familia de los métodos de Runge-Kutta es uno de 4 etapas definido de la siguiente manera:

Definición 8. El **método RK4** es aquel definido por

$$\begin{cases} y_k^{(1)} = y_k \\ y_k^{(2)} = y_k + \dfrac{h}{2} f(t_k, y_k^{(1)}) \\ y_k^{(3)} = y_k + \dfrac{h}{2} f(t_k + \dfrac{h}{2}, y_k^{(2)}) \\ y_k^{(4)} = y_k + h f(t_k + \dfrac{h}{2}, y_k^{(3)}) \\ y_{k+1} = y_k + \dfrac{h}{6}(f(t_k, y_k^{(1)}) + 2f(t_k + \dfrac{h}{2}, y_k^{(2)}) + 2f(t_k + \dfrac{h}{2}, y_k^{(3)}) + f(t_{k+1}, y_k^{(4)})), \end{cases}$$

o lo que es lo mismo,

$$y_{k+1} = y_k + \frac{h}{6}(k_1 + 2k_2 + 2k_3 + k_4), \qquad k \in \{0, 1, \ldots, n-1\},$$

donde

$$k_1 = f(t_k, y_k), \qquad k_2 = f(t_k + \frac{h}{2}, y_k + \frac{h}{2}k_1), \qquad k_3 = f(t_k + \frac{h}{2}, y_k + \frac{h}{2}k_2), \qquad k_4 = f(t_k + h, y_k + hk_3)$$

Regresando al caso general, cabe remarcar que a los coeficientes $a_{i,j}$, b_i y c_i se les va a pedir que verifiquen

$$\sum_{i=1}^{p} b_i = 1, \qquad c_i = \sum_{j=1}^{p} a_{i,j}$$

Esto se debe a que las fórmulas de integración (4) y (5) deberían ser exactas, por lo menos, para la función $f \equiv 1$. De poco serviría una fórmula de integración que no es capaz de aproximar ni la longitud de un segmento.

Por otra parte, los números $a_{i,j}, b_i, c_i$ suman, en total, $p^2 + 2p$ coeficientes, y suelen ser dispuestos en lo que se conoce como un *tablero de Butcher*, es decir, un diagrama del estilo

$$\begin{array}{c|ccc} c_1 & a_{1,1} & \cdots & a_{1,p} \\ \vdots & \vdots & \ddots & \vdots \\ c_p & a_{p,1} & \cdots & a_{p,p} \\ \hline & b_1 & \cdots & b_p \end{array}$$

Veamos ahora que estos métodos generalizan al de Euler, al del punto medio, al de Heun y al RK4.

(*i*) El método de Euler es más simple que el mecanismo de un botijo, pues se escribe como

$$y_{k+1} = y_k + hk_1,$$

donde

$$k_1 = f(t_k^{(1)}, y_k), \qquad t_k^{(1)} = t_k,$$

así que el tablero de Butcher sería

$$\begin{array}{c|c} 0 & 0 \\ \hline & 1 \end{array}$$

(*ii*) El método del punto medio se puede escribir como

$$y_{k+1} = y_k + hk_2,$$

donde

$$k_1 = f(t_k^{(1)}, y_k), \qquad k_2 = f(t_k^{(2)}, y_k + \frac{h}{2}k_1), \qquad t_k^{(1)} = t_k, \qquad t_k^{(2)} = t_k + \frac{h}{2},$$

así que el tablero de Butcher sería

$$\begin{array}{c|cc} 0 & 0 & 0 \\ 1/2 & 1/2 & 0 \\ \hline & 0 & 1 \end{array}$$

(*iii*) El método de Heun se puede escribir como

$$y_{k+1} = y_k + \frac{h}{2}(k_1 + k_2),$$

donde

$$k_1 = f(t_k^{(1)}, y_k), \qquad k_2 = f(t_k^{(2)}, y_k + hk_1), \qquad t_k^{(1)} = t_k, \qquad t_k^{(2)} = t_k + h,$$

así que el tablero de Butcher sería

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

(iv) Mirando cara a cara la definición del método RK4, el tablero de Butcher sería

0	0	0	0	0
1/2	1/2	0	0	0
1/2	0	1/2	0	0
1	0	0	1	0
	1/6	1/3	1/3	1/6

Ejemplo. Considérese el problema

$$(P)\ \begin{cases} y' = \frac{1}{2}(t^2 - y) \\ y(0) = 1 \end{cases}$$

Aproximemos $y(0.1)$ con un paso del *método de Euler implícito*, o sea, el método cuyo tablero es

1	1
	1

Este método sería

$$\begin{cases} y_k^{(1)} = y_k + hf(t_k + h, y_k^{(1)}) \\ y_{k+1} = y_k + hf(t_k + h, y_k^{(1)}) \end{cases}$$

Poniendo $k = 0$, se obtiene

$$y_0^{(1)} = 1 + 0.1f(0.1, y_0^{(1)}) = 1 + \frac{0.1}{2}(0.1^2 - y_0^{(1)}) = 1 + 0.0005 - 0.05y_0^{(1)} = 1.0005 - 0.05y_0^{(1)}$$

Despejando,

$$y_0^{(1)} = \frac{1.0005}{1.05} \approx 0.95285$$

Por tanto, la aproximación buscada es

$$y(0.1) \approx y_1 = 1 + 0.1f(0.1, 0.9528) = 1 + \frac{0.1}{2}(0.1^2 - 0.9528) = 0.95286$$

Ejemplo. Vamos a dar la expresión general del *método del trapecio*, o sea, el método con tablero

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

En primer lugar,

$$t_k^{(1)} = t_k, \qquad t_k^{(2)} = t_k + h = t_{k+1}$$

Por tanto, el método es

$$\begin{cases} y_k^{(1)} = y_k \\ y_k^{(2)} = y_k + h\big(\frac{1}{2}f(t_k, y_k + \frac{1}{2}f(t_k^{(2)}, y_k^{(2)})\big) \\ y_{k+1} = y_k + h\big(\frac{1}{2}f(t_k, y_k) + \frac{1}{2}f(t_k^{(2)}, y_k^{(2)})\big), \end{cases}$$

También puede escribirse

$$y_{k+1} = y_k + \frac{h}{2}(f(t_k, y_k) + f(t_{k+1}, y_{k+1}))$$

Definición 9. Un método de Runge-Kutta de orden p es **explícito** si su tablero de Butcher es de la forma

c_1	0	0	$\dots$	0	0
c_2	$a_{2,1}$	0	$\dots$	0	0
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
c_{p-1}	$a_{p-1,1}$	$a_{p-1,2}$	$\dots$	0	0
c_p	$a_{p,1}$	$a_{p,2}$	$\dots$	$a_{p,p-1}$	0
	b_1	b_2	$\dots$	b_{p-1}	b_p

o si se pueden reordenar las etapas para que sea de dicha forma. En caso contrario, se dice que el método es **implícito**, y si además $a_{i,j} = 0$ para $j > i$, se dirá que es **diagonalmente implícito**.

Obsérvese que, en un método explícito, para calcular y_{k+1} basta con evaluar f y hacer unas cuantas operaciones, mientras que en un método implícito habría que resolver sistemas de ecuaciones. En el caso diagonal implícito, sería necesario resolver entre 1 y p ecuaciones para hallar y_{k+1}.

Respecto a los métodos conocidos, es claro que el método de Euler, el del punto medio y el de Heun son explícitos, mientras que el de Euler implícito y el del trapecio son diagonalmente implícitos.

TEMA 2

Métodos unipaso

En este tema se estudiarán ciertos métodos numéricos que generalizan a los que se estudiaron en el tema anterior. Una vez más, si $(t_0, y^0) \in \mathbb{R} \times \mathbb{R}^n$ y $f : [t_0, t_0 + T] \times \mathbb{R}^n \to \mathbb{R}^n$ es continua y de Lipschitz en la variable y, partimos de un problema del estilo

$$(P) \begin{cases} y'(t) = f(t, y(t)), \ t \in [t_0, t_0 + T], \\ y(t_0) = y^0, \end{cases}$$

y de una partición uniforme $t_0 < t_1 < \ldots < t_{n-1} < t_n = t_0 + T$. En estas circunstancias, el objetivo es aproximar $y(t_k)$ para cada $k \in \{0, 1, \ldots, n\}$, denotándose por y_k a estas aproximaciones. Para ello, se emplearán métodos como los siguientes:

Definición 10. Un **método unipaso** es aquel definido por

$$y_{k+1} = y_k + h\,\Phi(t_k, y_k, h),$$

siendo $\Phi : [t_0, t_0 + T] \times \mathbb{R} \times [0, T] \to \mathbb{R}$ una función continua, denominada **función incremento**.

Evidentemente, todos los métodos vistos hasta ahora son métodos unipaso, pues en cada uno de ellos puede hallarse y_{k+1} a partir de t_k, y_k y h. Por ejemplo, en el método de Euler, la función incremento sería $\Phi(t, y, h) = f(t, y)$.

2.1. Consistencia, estabilidad y convergencia

Definición 11. Sea Φ la función incremento de un método unipaso. El **error de truncamiento local en t_k** o **error de discretización local en t_k** se define como

$$\varepsilon_k := y(t_{k+1}) - y(t_k) - h\,\Phi(t_k, y(t_k), h)$$

para cada $k \in \{0, 1, \ldots, n-1\}$.

Definición 12. Se dice que un método unipaso es **consistente** si

$$\lim_{h \to 0} \sum_{k=0}^{n-1} |\varepsilon_k| = 0$$

Definición 13. Un método unipaso se dice que es **estable** si existe una constante M positiva e independiente de h verificando lo siguiente: si $\{y_k\}_{k=0}^n$, $\{z_k\}_{k=0}^n$, $\{\delta_k\}_{k=0}^{n-1}$ son tales que

$$y_{k+1} = y_k + h\,\Phi(t_k, y_k, h), \qquad z_{k+1} = z_k + h\,\Phi(t_k, z_k, h) + \delta_k$$

para cada $k \in \{0, 1, \ldots, n-1\}$, entonces se tiene

$$\max_{k=0,1,\ldots,n} |y_k - z_k| \leq M \left(|y_0 - z_0| + \sum_{k=0}^{n-1} |\delta_k| \right)$$

Se puede pensar en los números $\{y_k\}_{k=0}^n$ como las aproximaciones teóricas del método, $\{z_k\}_{k=0}^n$ como los

resultados de calcular las aproximaciones en un ordenador y $\{\delta_k\}_{k=0}^{n-1}$ como los errores de redondeo que comete el ordenador.

Definición 14. Se dice que un método unipaso es **convergente** si

$$\lim_{h\to 0} e(h) = 0,$$

donde

$$e(h) := \max_{k=0,1,\dots,n} |y_k - y(t_k)|$$

Teorema 4. *Consistencia y estabilidad implican convergencia.*

Demostración. Considérese un método unipaso consistente y estable, y sean $\{z_k\}_{k=0}^{n}$ y $\{\delta_k\}_{k=0}^{n-1}$ tales que $z_k = y(t_k)$ y $\delta_k = \varepsilon_k$. Entonces se verifica trivialmente

$$z_{k+1} = z_k + h\,\Phi(t_k, z_k, h) + \delta_k$$

Sean $\{y_k\}_{k=0}^{n}$ las aproximaciones del método, es decir,

$$y_{k+1} = y_k + h\,\Phi(t_k, y_k, h)$$

Entonces

$$e(h) = \max_{k=0,1,\dots,n} |y_k - y(t_k)| = \max_{k=0,1,\dots,n} |y_k - z_k| \overset{(*)}{\leq} M(|y_0 - z_0| + \sum_{k=0}^{n-1} |\delta_k|) = M(|\cancel{y_0 - y(t_0)}| + \sum_{k=0}^{n-1} |\varepsilon_k|) = M \sum_{k=0}^{n-1} |\varepsilon_k|,$$

donde en $(*)$ se ha utilizado la estabilidad. Como el método es consistente, entonces

$$\lim_{h\to 0} M \sum_{k=0}^{n-1} |\varepsilon_k| = 0,$$

deduciéndose que

$$\lim_{h\to 0} e(h) = 0,$$

luego el método converge. □

Estudiar la consistencia y estabilidad de un método unipaso utilizando las definiciones puede llegar a ser bastante penoso. Los dos resultados siguientes acuden al rescate:

Teorema 5 (Caracterización de la consistencia). *Un método unipaso con función incremento Φ es consistente si y solo si para todo $t \in [t_0, t_0 + T]$ y todo $y \in \mathbb{R}$ se tiene*

$$\Phi(t, y, 0) = f(t, y)$$

Demostración. La igualdad siguiente va a ser de extrema importancia en la demostración, pero, por desgracia, no va a probarse:

$$\lim_{h\to 0} \sum_{k=0}^{n-1} |\varepsilon_k| = \int_{t_0}^{t_0+T} |f(t, y(t)) - \Phi(t, y(t), 0)|\, dt \tag{6}$$

Supóngase primero que se verifica $\Phi(t, y, 0) = f(t, y)$ para todo $t \in [t_0, t_0 + T]$ y todo $y \in \mathbb{R}$. Entonces, por (6),

$$\lim_{h\to 0} \sum_{k=0}^{n-1} |\varepsilon_k| = \int_{t_0}^{t_0+T} 0\, dt = 0,$$

así que el método es consistente.

Recíprocamente, supóngase que el método unipaso es consistente. Entonces la igualdad (6) permite

afirmar que

$$\int_{t_0}^{t_0+T} |f(t,y(t)) - \Phi(t,y(t),0)|\,dt = 0$$

Como el integrando es una función no negativa y se trata de una integral de Riemann, entonces $|f(t,y(t)) - \Phi(t,y(t),0)| = 0$, o sea,

$$f(t,y(t)) = \Phi(t,y(t),0) \tag{7}$$

para todo $t \in [t_0, t_0+T]$. Todavía no está todo el pescado vendido: fijemos $(t^*, y^*) \in [t_0, t_0+T] \times \mathbb{R}$ y veamos que, como propone el enunciado, $\Phi(t^*, y^*, 0) = f(t^*, y^*)$. Sea $y\colon [t_0, t_0+T] \times \mathbb{R} \to \mathbb{R}$ la única solución de

$$(P) \begin{cases} y'(t) = f(t, y(t)), t \in [t_0, t_0+T], \\ y(t^*) = y^* \end{cases}$$

Si se aplica la igualdad (7) en t^* se obtiene $f(t^*, y(t^*)) = \Phi(t^*, y(t^*), 0)$, o sea, $f(t^*, y^*) = \Phi(t^*, y^*, 0)$. □

Teorema 6 (Condición suficiente para la estabilidad). *Si la función incremento de un método unipaso es de Lipschitz en la variable y, entonces el método es estable.*

Demostración. Considérense $\{y_k\}_{k=0}^n$, $\{z_k\}_{k=0}^n$, $\{\delta_k\}_{k=0}^{n-1}$ tales que, para cada $k \in \{0, 1, \ldots, n-1\}$, se tiene

$$y_{k+1} = y_k + h\,\Phi(t_k, y_k, h), \qquad z_{k+1} = z_k + h\,\Phi(t_k, z_k, h) + \delta_k,$$

y sea $L > 0$ la constante de Lipschitz de Φ. Entonces

$$\begin{aligned}
|y_{k+1} - z_{k+1}| &\le |y_k - z_k| + h|\Phi(t_k, y_k, h) - \Phi(t_k, z_k, h)| + |\delta_k| \\
&\overset{(i)}{\le} |y_k - z_k| + hL|y_k - z_k| + |\delta_k| = (1+hL)|y_k - z_k| + |\delta_k| \\
&\overset{(ii)}{\le} e^{hL}|y_k - z_k| + |\delta_k| \\
&\le e^{hL}(e^{hL}|y_{k-1} - z_{k-1}| + |\delta_{k-1}|) + |\delta_k| = e^{2hL}|y_{k-1} - z_{k-1}| + e^{hL}|\delta_{k-1}| + |\delta_k| \\
&\le e^{2hL}(e^{hL}|y_{k-2} - z_{k-2}| + |\delta_{k-2}|) + e^{hL}|\delta_{k-1}| + |\delta_k| \\
&\le \ldots \\
&\le e^{(k+1)hL}|y_0 - z_0| + e^{khL}|\delta_0| + \ldots + e^{2hL}|\delta_{k-2}| + e^{hL}|\delta_{k-1}| + |\delta_k| \\
&\overset{(iii)}{\le} e^{TL}|y_0 - z_0| + e^{TL}|\delta_0| + \ldots + e^{TL}|\delta_{k-2}| + e^{TL}|\delta_{k-1}| + |\delta_k| \\
&\overset{(iv)}{\le} e^{TL}|y_0 - z_0| + e^{TL}|\delta_0| + \ldots + e^{TL}|\delta_{k-2}| + e^{TL}|\delta_{k-1}| + e^{TL}|\delta_k| = e^{TL}(|y_0 - z_0| + \sum_{j=0}^{k} |\delta_j|) \\
&\le e^{TL}(|y_0 - z_0| + \sum_{j=0}^{n-1} |\delta_j|)
\end{aligned}$$

Como este último miembro no depende de k y e^{TL} es una constante positiva e independiente de h, entonces el método es estable. Algunas aclaraciones:

- ***(i)*** Se ha utilizado que Φ es de Lipschitz en la variable y.
- ***(ii)*** Se ha utilizado, de nuevo, que $1 + x \le e^x$ para todo $x \in \mathbb{R}$.
- ***(iii)*** Se ha utilizado que $h = \frac{T}{n}$ y por tanto $kh \le nh = T$ para todo $k \le n$.
- ***(iv)*** Se ha utilizado que $TL > 0$ y por tanto $e^{TL} > 1$, luego $e^{TL}|\delta_k| \ge |\delta_k|$. □

Corolario 1. *Sea $\Phi\colon [t_0, t_0+T] \times \mathbb{R} \times [0,T] \to \mathbb{R}$ la función incremento de un método unipaso. Si $\frac{\partial \Phi}{\partial y}$ existe y es acotada, entonces el método es estable.*

Demostración. No hay más que aplicar el teorema anterior teniendo en cuenta que la existencia y acotación de $\frac{\partial \Phi}{\partial y}$ implican que Φ es de Lipschitz en la variable y. □

Corolario 2. *Si la función incremento $\Phi\colon [t_0,t_0+T]\times\mathbb{R}\times[0,T]\to\mathbb{R}$ de un método unipaso verifica $\Phi(t,y,0)=f(t,y)$ y es de Lipschitz en la variable y, entonces el método converge.*

Demostración. Consecuencia directa de los últimos tres teoremas. □

Ejemplo. Si f es de Lipschitz en la variable y, entonces el método de Heun es convergente. En efecto, la función incremento es

$$\Phi(t,y,h)=\frac{1}{2}(f(t,y)+f(t+h,y+hf(t,y))),$$

que verifica

(i) $\Phi(t,y,0)=\dfrac{1}{2}(f(t,y)+f(t,y))=f(t,y)$ para todos $t,y\in[t_0,t_0+T]\times\mathbb{R}$;

(ii)
$$\begin{aligned}|\Phi(t,y,h)-\Phi(t,z,h)| &= \Big|\frac{1}{2}(f(t,y)+f(t+h,y+hf(t,y)))-\frac{1}{2}(f(t,z)+f(t+h,z+hf(t,z)))\Big|\\ &\le \frac{1}{2}\big|f(t,y)-f(t,z)\big|+\frac{1}{2}\big|f(t+h,y+hf(t,y))-f(t+h,z+hf(t,z))\big|\\ &\le \frac{L}{2}|y-z|+\frac{L}{2}\big|y+hf(t,y)-z-hf(t,z)\big|\\ &\le L|y-z|+\frac{Lh}{2}|f(t,y)-f(t,z)|\\ &\le \Big(L+\frac{L^2T}{2}\Big)|y-z|\end{aligned}$$

para todos $(t,y,h),(t,z,h)\in[t_0,t_0+T]\times\mathbb{R}\times[0,T]\to\mathbb{R}$.

Por tanto, el método es consistente y estable, luego convergente.

2.2. Orden de un método unipaso

Notación. Sean $f,g\colon[0,T]\to\mathbb{R}$ dos funciones con $g(h)\neq 0$ para todo $h\in[0,T]$.

(i) Se dice que $f=o(g)$ si

$$\lim_{h\to 0}\frac{f(h)}{g(h)}=0,$$

es decir, si para todo $\varepsilon>0$ existe $h^*>0$ tal que para todo $h\in(0,h^*)$ se tiene que

$$\Big|\frac{f(h)}{g(h)}\Big|<\varepsilon,$$

o, equivalentemente,

$$|f(h)|<\varepsilon|g(h)|$$

(ii) Se dice que $f=O(g)$ si existen $C,h^*>0$ tales que para todo $h\in(0,h^*)$ se tiene que

$$\Big|\frac{f(h)}{g(h)}\Big|\le C,$$

o, equivalentemente,

$$|f(h)|\le C|g(h)|$$

Obsérvese que si f y g son continuas y se encuentran en las condiciones de la definición anterior, entonces $f=o(g)$ implica $f=O(g)$, pero el recíproco no es cierto.

Proposición 1. *Si $f=O(h^p)$ (lo que quiere decir que $f=O(g)$, con g definida por $g(h)=h^p$), entonces $f=O(h^q)$ para todo $q\le p$.*

Demostración. En efecto, por hipótesis, existen $C, h^* > 0$ tales que, para todo $h \in (0, h^*)$,

$$|f(h)| \leq Ch^p = Ch^{p-q}h^q \leq C(h^*)^{p-q}h^q = \tilde{C}h^q,$$

donde $\tilde{C} = C(h^*)^{p-q}$ es una constante positiva. □

Proposición 2. *Si $f = O(h^p)$ y $g = O(h^q)$, entonces*

$$f + g = O\left(h^{\max\{p,q\}}\right) \qquad \text{y} \qquad fg = O(h^{p+q})$$

Demostración. Ejercicio. □

Proposición 3. *Si $y \in \mathcal{C}^{p+1}([a,b],\mathbb{R})$ y $t_0 \in [a,b], h > 0$ son tales que $t_1 = t_0 + h \in [a,b]$, entonces*

$$y(t_1) - P_{y,p,t_0}(t_1) = O(h^{p+1}),$$

lo que también suele escribirse como

$$y(t_1) = P_{y,p,t_0}(t_1) + O(h^{p+1})$$

Demostración. Obsérvese que lo que dice el enunciado es que $f = O(g)$, siendo f la función dada por $f(h) = y(t_0 + h) - P_{y,p,t_0}(t_0 + h)$, y g la función definida por $g(h) = h^{p+1}$. La demostración en sí: por la fórmula del resto de Lagrange, existe $\xi \in (t_0, t_1)$ tal que

$$y(t_1) - P_{y,p,t_0}(t_1) = \frac{y^{(p+1)}(\xi)}{(p+1)!}h^{p+1} \leq \max_{t\in[a,b]} |y^{(p+1)}(t)| \frac{h^{p+1}}{(p+1)!} = Ch^{p+1},$$

donde

$$C = \frac{1}{(p+1)!} \max_{t\in[a,b]} |y^{(p+1)}(t)|$$

es una constante positiva. □

Definición 15. Si $y \in \mathcal{C}^{p+1}([t_0, t_0 + T],\mathbb{R})$ con $p \in \mathbb{N}$, se dice que un método unipaso es **de orden p** si

$$\sum_{k=0}^{n-1} |\varepsilon_k| = O(h^p)$$

Nótese que la suma anterior es una función que depende de h (ya que depende de n), y por tanto la definición tiene perfecto sentido.

Teorema 7. *Si $f \in \mathcal{C}^p([t_0, t_0 + T] \times \mathbb{R}, \mathbb{R})$ y se tiene un método unipaso estable y de orden p, entonces $e(h) = O(h^p)$.*

Demostración. Nótese que $f \in \mathcal{C}^p([t_0, t_0 + T] \times \mathbb{R}, \mathbb{R})$ implica $y \in \mathcal{C}^{p+1}([t_0, t_0 + T],\mathbb{R})$, luego tiene sentido preguntarse si el método es de orden p. Sean $\{z_k\}_{k=0}^n$ y $\{\delta_k\}_{k=0}^{n-1}$ con $z_k = y(t_k)$ y $\delta_k = \varepsilon_k$. Es claro que

$$z_{k+1} = z_k + h\,\Phi(t_k, z_k, h) + \delta_k$$

Sean $\{y_k\}_{k=0}^n$ las aproximaciones del método, es decir,

$$y_{k+1} = y_k + h\,\Phi(t_k, y_k, h)$$

Entonces

$$e(h) = \max_{k=0,1,\ldots,n} |y(t_k) - y_k| \leq M(|\cancel{y(t_0) - y_0}| + \sum_{j=0}^{n-1} |\varepsilon_j|) = M \sum_{j=0}^{n-1} |\varepsilon_j| \leq MCh^p,$$

donde se ha utilizado que el método es estable y de orden p. □

Determinar el orden de un método unipaso a través de la definición puede ser una tarea dura y pesarosa.

En la práctica, lo más frecuente será recurrir al resultado que sigue.

Teorema 8 (Caracterización del orden). *Sea $f \in C^p([t_0, t_0+T]\times\mathbb{R},\mathbb{R})$ y sea Φ la función incremento de un método unipaso. Supóngase que para cada $i \in \{1,\dots,p\}$ existe $\frac{\partial^i \Phi}{\partial h^i}$ y es continua. Entonces el método unipaso es de orden p si y solo si para todo $(t,y) \in [t_0, t_0+T]\times\mathbb{R}$ se verifica*

$$\frac{\partial^i \Phi}{\partial h^i}(t,y,0) = \frac{1}{i+1} f^{(i)}(t,y), \quad i \in \{0,1,\dots,p-1\},$$

entendiéndose que para $i=0$ la igualdad anterior dice que

$$\Phi(t,y,0) = f(t,y)$$

Demostración. Solo va a probarse una implicación. Supóngase que la igualdad del enunciado es cierta y veamos que el método es de orden p. Se tiene que

$$|\varepsilon_k| = |y(t_{k+1}) - y(t_k) - h\,\Phi(t_k, y(t_k), h)|$$

Por un lado,

$$y(t_{k+1}) - y(t_k) = \sum_{i=1}^{p} y^{(i)}(t_k)\frac{h^i}{i!} + O(h^{p+1}) = \sum_{i=1}^{p} f^{(i-1)}(t_k, y(t_k))\frac{h^i}{i!} + O(h^{p+1})$$

Por otra parte,

$$\Phi(t_k, y(t_k), h) = \sum_{i=0}^{p-1} \frac{\partial^i \Phi}{\partial h^i}(t_k, y(t_k), 0)\frac{h^i}{i!} + O(h^p) = \sum_{i=0}^{p-1} f^{(i)}(t_k, y(t_k))\frac{h^i}{(i+1)!} + O(h^p),$$

y en consecuencia,

$$h\,\Phi(t_k, y(t_k), h) = \sum_{i=0}^{p-1} f^{(i)}(t_k, y(t_k))\frac{h^{i+1}}{(i+1)!} + O(h^{p+1}) = \sum_{i=1}^{p} f^{(i-1)}(t_k, y(t_k))\frac{h^i}{i!} + O(h^{p+1}),$$

De aquí se deduce que $\varepsilon_k = O(h^{p+1})$ para todo $k \in \{0,\dots,n-1\}$, o sea, que existen $C, h^* > 0$ tales que para todo $h \in (0, h^*)$ se tiene $|\varepsilon_k| \le Ch^{p+1}$ y, por tanto,

$$\sum_{k=0}^{n-1} |\varepsilon_k| \le \sum_{k=0}^{n-1} Ch^{p-1} = Cnh^{p+1} = Cnhh^p = CTh^p,$$

donde CT es una constante positiva. Se concluye que el método es de orden p. □

Corolario 3. *Un método unipaso es consistente si y solo si es de orden 1.*

Demostración. No hay más que leer el teorema anterior y el Teorema 5. □

Nótese que *ser de orden 1* no impide *ser de orden p* para $p > 1$. De forma natural, se introduce la definición siguiente:

Definición 16. Dado $p \in \mathbb{N}$, un método unipaso se dice que es **de orden exactamente *p*** si es de orden p pero no es de orden $p+1$.

Corolario 4. *Sea $f \in C^p([t_0, t_0+T]\times\mathbb{R},\mathbb{R})$ y sea Φ la función incremento de un método unipaso. Si para cada $i \in \{1,\dots,p\}$ existe $\frac{\partial^i \Phi}{\partial h^i}$ y es continua, entonces el método unipaso es de orden exactamente p si y solo si se verifica lo siguiente:*

(i) *Para todo* $(t,y)\in[t_0,t_0+T]\times\mathbb{R}$ *se tiene*

$$\frac{\partial^i\Phi}{\partial h^i}(t,y,0)=\frac{1}{i+1}f^{(i)}(t,y),\quad i\in\{0,1,\dots,p-1\},$$

entendiéndose que para $i=0$ *la igualdad anterior dice que* $\Phi(t,y,0)=f(t,y)$.

(ii) *Existe* $(t,y)\in[t_0,t_0+T]\times\mathbb{R}$ *tal que*

$$\frac{\partial^p\Phi}{\partial h^p}(t,y,0)\neq\frac{1}{p+1}f^{(p)}(t,y)$$

Demostración. Véase el teorema anterior. □

Ejemplo. Como ya se había anticipado, el método de Euler es de orden 1. ¿Será de orden 2? Pues

$$\frac{1}{2}f^{(1)}(t,y)=\frac{1}{2}\Big(\frac{\partial f}{\partial t}(t,y)+f(t,y)\frac{\partial f}{\partial y}(t,y)\Big)$$

En general, no parece que se pueda asegurar que esta expresión coincida con $\frac{\partial\Phi}{\partial h}(t,y,0)=0$, pero si f es constante, sí que se verifica. Es más, el método de Euler en este caso es de orden p para todo $p\in\mathbb{N}$ (y con razón, pues aproxima la solución del problema de forma exacta).

Ejemplo. Estudiemos el orden del método de Heun:

$$y_{k+1}=y_k+\frac{h}{2}(f(t_k,y_k)+f(t_k+h,y_k+hf(t_k,y_k)))$$

La función incremento sería $\Phi(t,y,h)=\frac{1}{2}(f(t,y)+f(t+h,y+hf(t,y)))$. Para poder aplicar el teorema anterior, se le va a pedir a f toda la regularidad que sea necesaria. Se tiene que

$$\Phi(t,y,0)=\frac{1}{2}2f(t,y)=f(t,y),$$

luego el método es de orden 1. Además,

$$\frac{\partial\Phi}{\partial h}(t,y,h)=\frac{1}{2}\Big(\frac{\partial f}{\partial t}(t+h,y+hf(t,y))+f(t,y)\frac{\partial f}{\partial y}(t+h,y+hf(t,y))\Big),$$

así que

$$\frac{\partial\Phi}{\partial h}(t,y,0)=\frac{1}{2}\Big(\frac{\partial f}{\partial t}(t,y)+f(t,y)\frac{\partial f}{\partial y}(t,y)\Big)=\frac{1}{2}f^{(1)}(t,y),$$

luego el método es de orden 2. Seguimos:

$$\begin{aligned}\frac{\partial^2\Phi}{\partial h^2}(t,y,h)=\frac{1}{2}\Big(&\frac{\partial^2 f}{\partial t^2}(t+h,y+hf(t,y))+f(t,y)\frac{\partial^2 f}{\partial t\partial y}(t+h,y+hf(t,y))+\\&f(t,y)\frac{\partial^2 f}{\partial y\partial t}(t+h,y+hf(t,y))+f(t,y)^2\frac{\partial^2 f}{\partial y^2}(t+h,y+hf(t,y))\Big)\end{aligned}$$

En consecuencia,

$$\frac{\partial^2\Phi}{\partial h^2}(t,y,0)=\frac{1}{2}\Big(\frac{\partial^2 f}{\partial t^2}(t,y)+f(t,y)\frac{\partial^2 f}{\partial t\partial y}(t,y)+f(t,y)\frac{\partial^2 f}{\partial y\partial t}(t,y)+f(t,y)^2\frac{\partial^2 f}{\partial y^2}(t,y)\Big)$$

Pero

$$\begin{aligned}\frac{1}{3}f^{(2)}(t,y)&=\frac{1}{3}\Big(\frac{\partial f^{(1)}}{\partial t}(t,y)+f(t,y)\frac{\partial f^{(1)}}{\partial y}(t,y)\Big)\\&=\frac{1}{3}\Big(\frac{\partial^2 f}{\partial t^2}(t,y)+2f(t,y)\frac{\partial^2 f}{\partial y\partial t}(t,y)+\frac{\partial f}{\partial t}(t,y)\frac{\partial f}{\partial y}(t,y)+f(t,y)^2\frac{\partial^2 f}{\partial y^2}(t,y)+f(t,y)\Big(\frac{\partial f}{\partial y}(t,y)\Big)^2\Big)\end{aligned}$$

Sin entrar en detalles, parece que el método de Heun, en general, es de orden exactamente 2.

Ejemplo. ¿Cuál será el orden del método de Taylor de orden p? La función incremento es

$$\Phi(t,y,h)=\sum_{q=1}^{p}\frac{f^{(q-1)}(t,y)}{q!}h^{q-1}$$

A hacer cuentas: en primer lugar,

$$\Phi(t,y,0)=f^{(0)}(t,y)=f(t,y),$$

así que el método es de orden 1. Ahora se deriva:

$$\frac{\partial\Phi}{\partial h}(t,y,h)=\sum_{q=2}^{p}\frac{q-1}{q!}f^{(q-1)}(t,y)h^{q-2}$$

Por tanto,

$$\frac{\partial\Phi}{\partial h}(t,y,0)=\frac{1}{2}f^{(1)}(t,y),$$

luego el método es de orden 2. Otra más:

$$\frac{\partial^2\Phi}{\partial h^2}(t,y,h)=\sum_{q=3}^{p}\frac{(q-1)(q-2)}{q!}f^{(q-1)}(t,y)h^{q-3}$$

En consecuencia,

$$\frac{\partial^2\Phi}{\partial h^2}(t,y,0)=\frac{1}{3}f^{(2)}(t,y),$$

y el método es de orden 3. Ya se ven por dónde van los tiros: la derivada de orden $p-1$ sería

$$\frac{\partial^{p-1}\Phi}{\partial h^{p-1}}(t,y,h)=\frac{(p-1)!}{p!}f^{(p-1)}(t,y)=\frac{1}{p}f^{(p-1)}(t,y),$$

y al evaluar en $h=0$ obtenemos que el método es de orden p. Si se deriva una vez más, se llega a

$$\frac{\partial^p\Phi}{\partial h^p}=0,$$

pero, en general, seguramente se cumpla

$$\frac{1}{p+1}f^{(p)}(t,y)\neq 0$$

La conclusión obtenida es verdaderamente sorprendente: el método de Taylor de orden p es de orden exactamente p.

2.3. Análisis de los métodos de Runge-Kutta

En la Sección 1.6 tuvo lugar un acto de valentía y descaro al definirse un método numérico mediante las ecuaciones

$$(S)\begin{cases} y_k^{(1)}=y_k+h(a_{1,1}f(t_k^{(1)},y_k^{(1)})+a_{1,2}f(t_k^{(2)},y_k^{(2)})+\ldots+a_{1,p}f(t_k^{(p)},y_k^{(p)}))\\ y_k^{(2)}=y_k+h(a_{2,1}f(t_k^{(1)},y_k^{(1)})+a_{2,2}f(t_k^{(2)},y_k^{(2)})+\ldots+a_{2,p}f(t_k^{(p)},y_k^{(p)}))\\ \quad\vdots\\ y_k^{(p)}=y_k+h(a_{p,1}f(t_k^{(1)},y_k^{(1)})+a_{p,2}f(t_k^{(2)},y_k^{(2)})+\ldots+a_{p,p}f(t_k^{(p)},y_k^{(p)}))\\ y_{k+1}=y_k+h(b_1f(t_k^{(1)},y_k^{(1)})+\ldots+b_pf(t_k^{(p)},y_k^{(p)})) \end{cases}$$

Cabría preguntarse si este sistema presenta algún problema de existencia de soluciones, en cuyo caso la definición de los métodos de Runge-Kutta correría serio peligro. Supóngase primero que el método es explícito, quedando el sistema anterior como sigue:

$$\begin{cases} y_k^{(1)} = y_k \\ y_k^{(2)} = y_k + ha_{2,1} f(t_k^{(1)}, y_k^{(1)}) \\ \quad \vdots \\ y_k^{(p)} = y_k + h(a_{p,1} f(t_k^{(1)}, y_k^{(1)}) + a_{p,2} f(t_k^{(2)}, y_k^{(2)}) + \ldots + a_{p,p} f(t_k^{(p)}, y_k^{(p)})) \\ y_{k+1} = y_k + h(b_1 f(t_k^{(1)}, y_k^{(1)}) + \ldots + b_p f(t_k^{(p)}, y_k^{(p)})) \end{cases}$$

En este caso, los $y_k^{(i)}$ se pueden calcular sin problema alguno a partir del anterior; el sistema posee solución y es única. Si el método fuese diagonalmente implícito, el sistema adquiere la forma

$$\begin{cases} y_k^{(1)} = y_k + ha_{1,1} f(t_k^{(1)}, y_k^{(1)}) \\ y_k^{(2)} = y_k + ha_{2,1} f(t_k^{(1)}, y_k^{(1)}) + a_{2,2} f(t_k^{(2)}, y_k^{(2)}) \\ \quad \vdots \\ y_k^{(p)} = y_k + h(a_{p,1} f(t_k^{(1)}, y_k^{(1)}) + a_{p,2} f(t_k^{(2)}, y_k^{(2)}) + \ldots + a_{p,p} f(t_k^{(p)}, y_k^{(p)})) \\ y_{k+1} = y_k + h(b_1 f(t_k^{(1)}, y_k^{(1)}) + \ldots + b_p f(t_k^{(p)}, y_k^{(p)})) \end{cases}$$

Para la hallar $y_k^{(1)}$, habría que resolver una ecuación del estilo

$$y = y_k + ha_{1,1} f(t_k^{(1)}, y)$$

Considérese la funcion $g\colon \mathbb{R} \to \mathbb{R}$ dada por

$$g(y) = y_k + ha_{1,1} f(t_k^{(1)}, y)$$

Se ha transformado la ecuación a resolver en una ecuación de punto fijo ($g(y) = y$), así que convendría que g tuviese un punto fijo y, ya que estamos, que este sea único. De esta manera, el sistema anterior tendría solución única y el método estaría bien definido.

Teorema 9 (Teorema del punto fijo de Banach). *Sea (X,d) un espacio métrico completo y supóngase que $T\colon X \to X$ es una aplicación contractiva, es decir, existe $\alpha \in (0,1)$ con*

$$d(T(x), T(y)) \leq \alpha d(x,y)$$

para cualesquiera $x, y \in X$. Entonces T existe un único $x^* \in X$ tal que $T(x^*) = x^*$. Es más, para cada $x_0 \in X$, la sucesión $\{T^k(x_0)\}_{k=1}^{\infty}$ converge a x^*.

Demostración. Corresponde a la asignatura *Ecuaciones Diferenciales II*. □

Estudiemos entonces la contractividad de g. Si $y_1, y_2 \in \mathbb{R}$,

$$|g(y_2) - g(y_1)| = |ha_{1,1}(f(t_k^{(1)}, y_1) - f(t_k^{(1)}, y_2))| \leq h|a_{1,1}|L|y_1 - y_2|,$$

donde L es la constante de Lipschitz de f. Así, siempre que se tenga

$$h < \frac{1}{L|a_{1,1}|}$$

la función g será contractiva e $y_k^{(1)}$ quedará determinado de forma única. Que h sea tan pequeño como se quiera no es una condición muy restrictiva, pues no hay más que dividir la partición en subintervalos

pequeños. Ya se conoce $y_k^{(1)}$; vamos con $y_k^{(2)}$. Habría que resolver la ecuación

$$y = y_k + ha_{2,1}f(t_k^{(1)}, y_k^{(1)}) + a_{2,2}f(t_k^{(2)}, y)$$

La dinámica va a ser exactamente la misma: considérese la función $g\colon \mathbb{R} \to \mathbb{R}$ dada por

$$g(y) = y_k + ha_{2,1}f(t_k^{(1)}, y_k^{(1)}) + a_{2,2}f(t_k^{(2)}, y)$$

Razonando como antes, se demuestra fácilmente que basta tener

$$h < \frac{1}{L|a_{2,2}|}$$

para que g posea un único punto fijo y, así, que $y_k^{(2)}$ quede determinado de manera única. Continuando con este procedimiento se demuestra que una condición suficiente para que el sistema que define un método diagonalmente implícito tenga solución única es

$$h < \min_{i=1,\dots,n}\left\{\frac{1}{L|a_{i,i}|}\right\}$$

Siguiendo esta idea, el próximo teorema va a recoger una condición suficiente para asegurar la buena definición de los métodos de Runge-Kutta, y ya de paso se estudiarán la estabilidad y la consistencia. Previamente, será necesario recordar algunas nociones de asignaturas pasadas:

Definición 17. Dada una matriz $A \in \mathcal{M}_n(\mathbb{R})$, se define el **radio espectral de A** como

$$\rho(A) := \max\{|\lambda| \colon \lambda \text{ es autovalor de } A\}$$

Proposición 4. *$(\mathbb{R}^n, ||\cdot||_\infty)$ es un espacio métrico completo, donde $||\cdot||_\infty \colon \mathbb{R}^n \to \mathbb{R}$ está definida por*

$$||Y||_\infty = \max_{i=1,\dots,n} |Y_i|,$$

Además, dada $A \in \mathcal{M}_n(\mathbb{R})$, la norma matricial subordinada a $||\cdot||_\infty$ viene dada por

$$||A||_\infty = \max_{i=1,\dots,n} \sum_{j=1}^{n} |a_{i,j}|$$

Demostración. Corresponde a la asignatura *Métodos Numéricos II*. □

Proposición 5. *Dada $A \in \mathcal{M}_n(\mathbb{R})$, se verifican las siguientes propiedades:*

(i) *Si $||\cdot||$ es una norma matricial cualquiera,*

$$\rho(A) = \lim_{k\to\infty} ||A^k||^{1/k}$$

(ii) *$\rho(A) < 1$ si y solo si $\lim_{k\to\infty} A^k = 0$.*

(iii) *Si $||\cdot||$ es una norma matricial subordinada y $||A|| < 1$, entonces $I + A$ es inversible.*

Demostración. Corresponde a la asignatura *Métodos Numéricos II*. □

Se advierte que el valor absoluto de una matriz se entenderá como la matriz formada por el valor absoluto de cada componente, y lo mismo con el valor absoluto de un vector.

Teorema 10. *Sea $A \in \mathcal{M}_p(\mathbb{R})$ la matriz asociada a un método de Runge-Kutta de p etapas.*

(i) *Supóngase que $h \in [0,T]$ es tal que*

$$h < \frac{1}{L\rho(|A|)}$$

Entonces el sistema (S) tiene solución única y, en consecuencia, el método está bien definido.

(ii) *Supóngase además que*

$$T < \frac{1}{L\rho(|A|)}$$

Entonces el método es estable.

(iii) *Más aún, el método es consistente si y solo si*

$$\sum_{i=1}^{p} b_i = 1,$$

(iv) *Bajo las hipótesis de todos los apartados anteriores, el método es convergente.*

Demostración.

(i) Se va a intentar aplicar el teorema del punto fijo a la función $G\colon \mathbb{R}^p \to \mathbb{R}^p$ definida por

$$G(Y) = \begin{pmatrix} y_k + h(a_{1,1}f(t_k^{(1)},Y_1) + a_{1,2}f(t_k^{(2)},Y_2) + \ldots + a_{1,p}f(t_k^{(p)},Y_p)) \\ y_k + h(a_{2,1}f(t_k^{(1)},Y_1) + a_{2,2}f(t_k^{(2)},Y_2) + \ldots + a_{2,p}f(t_k^{(p)},Y_p)) \\ \vdots \\ y_k + h(a_{p,1}f(t_k^{(1)},Y_1) + a_{p,2}f(t_k^{(2)},Y_2) + \ldots + a_{p,p}f(t_k^{(p)},Y_p)) \end{pmatrix} = y_k E + hAF(t,Y),$$

donde

$$E = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \qquad A = \begin{pmatrix} a_{1,1} & a_{1,2} & \ldots & a_{1,p} \\ a_{2,1} & a_{2,2} & \ldots & a_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ a_{p,1} & a_{p,2} & \ldots & a_{p,p} \end{pmatrix} \qquad F(t,Y) = \begin{pmatrix} f(t_k^{(1)},Y_1) \\ f(t_k^{(2)},Y_2) \\ \vdots \\ f(t_k^{(p)},Y_p) \end{pmatrix}$$

Si $i \in \{1,\ldots,n\}$ e $Y, Z \in \mathbb{R}^p$,

$$|G_i(Y) - G_i(Z)| = \left| h \sum_{j=1}^{p} a_{i,j}(f(t_k^{(j)},Y_j) - f(t_k^{(j)},Z_j)) \right| \leq hL \sum_{j=1}^{p} |a_{i,j}|\,|Y_j - Z_j|$$

Por tanto,

$$|G(Y) - G(Z)| \leq hL\,|A|\,|Y - Z|, \tag{8}$$

donde los vectores se están comparando componente a componente. Además, como

$$||A||_\infty = \max_{i=1,\ldots,p} \sum_{j=1}^{p} |a_{i,j}|, \qquad ||Y - Z||_\infty = \max_{j=1,\ldots,p} |Y_j - Z_j|,$$

entonces

$$||G(Y) - G(Z)||_\infty \leq hL\,||A||_\infty\,||Y - Z||_\infty$$

En consecuencia, siempre que se tenga

$$h < \frac{1}{L||A||_\infty}$$

el teorema del punto fijo nos dará una única solución para el sistema (S), y por tanto el método de Runge-Kutta de p etapas estará bien definido. Pero esto no es lo que se pide demostrar, pues hay que probar una desigualdad más fuerte (ya que $\rho(A) \leq ||A||$). Para ello, se tiene en cuenta que, en

realidad, no es necesario que G sea contractiva; basta que lo sea una iterada suya. Si $n \in \mathbb{N}$, se demuestra fácilmente a partir de (8) que

$$|G^n(Y) - G^n(Z)| \leq h^n L^n |A|^n |Y - Z|,$$

luego

$$||G^n(Y) - G^n(Z)||_\infty \leq h^n L^n |||A|^n||_\infty ||Y - Z||_\infty$$

Ahora bien, por hipótesis se tiene

$$hL\rho(|A|) < 1,$$

y como

$$\rho(|A|) = \lim_{n\to\infty} |||A|^n||_\infty^{1/n},$$

entonces

$$\lim_{n\to\infty} hL|||A|^n||_\infty^{1/n} = hL\rho(|A|) < 1,$$

así que existe $n_0 \in \mathbb{N}$ tal que

$$hL|||A|^{n_0}||_\infty^{1/n_0} < 1,$$

luego

$$h^{n_0} L^{n_0} |||A|^{n_0}||_\infty < 1,$$

concluyéndose que G^{n_0} es contractiva y tiene un único punto fijo, que es la única solución de (S).

(ii) En lugar de probar la estabilidad por definición, se va a echar mano del Teorema 6. La función incremento es

$$\Phi(t, y, h) = \sum_{i=1}^{p} b_i f(t^{(i)}, y^{(i)}),$$

donde, para cada $i \in \{1, \ldots, p\}$,

$$t^{(i)} = t + c_i h, \qquad y^{(i)} = y + h \sum_{j=1}^{p} a_{i,j} f(t^{(j)}, y^{(j)})$$

Si $y, z \in \mathbb{R}$, se tiene

$$\left|y^{(i)} - z^{(i)}\right| = \left|y - z + h \sum_{j=1}^{p} a_{i,j}(f(t^{(j)}, y^{(j)}) - f(t^{(j)}, z^{(j)}))\right| \leq |y - z| + hL \sum_{j=1}^{p} |a_{i,j}||y^{(j)} - z^{(j)}|$$

Sean

$$Y = \begin{pmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(p)} \end{pmatrix} \qquad Z = \begin{pmatrix} z^{(1)} \\ z^{(2)} \\ \vdots \\ z^{(p)} \end{pmatrix}$$

Entonces

$$\begin{aligned} |Y - Z| &\leq |y - z|E + hL|A||Y - Z| \\ &\leq |y - z|E + TL|A||Y - Z| \\ &\leq |y - z|E + TL|A|\big(|y - z|E + TL|A||Y - Z|\big) = |y - z|\big(I + TL|A|\big)E + (TL|A|)^2|Y - Z| \end{aligned}$$

Por recurrencia, para todo $n \in \mathbb{N}$ se tiene

$$\begin{aligned} |Y - Z| &\leq |y - z|\big(I + TL|A| + (TL|A|)^2 + \ldots + (TL|A|)^{n-1}\big)E + (TL|A|)^n|Y - Z| \\ &= |y - z|(I - TL|A|)^{-1}(I - (TL|A|)^n)E + (TL|A|)^n|Y - Z|, \end{aligned}$$

donde en la igualdad se ha utilizado la fórmula para la sucesión de sumas parciales de una serie geométrica de matrices. Nótese que, por ser $TL\rho(|A|) < 1$, se tiene que $||TL|A|||_\infty < 1$ y por tanto

$I - TL|A|$ es inversible. Ahora, utilizando que $||E||_\infty = 1$,

$$||Y - Z||_\infty \leq |y - z|||(I - TL|A|)^{-1}||_\infty ||I - (TL|A|)^n||_\infty + ||(TL|A|)^n||_\infty ||Y - Z||_\infty \tag{9}$$

Si se verifica la condición

$$T < \frac{1}{L\rho(|A|)},$$

es decir,

$$TL\rho(|A|) = \rho(TL|A|) < 1,$$

entonces

$$\lim_{n\to\infty} (TL|A|)^n = 0,$$

luego

$$\lim_{n\to\infty} ||I - (TL|A|)^n||_\infty = 1 \qquad \text{y} \qquad \lim_{n\to\infty} ||(TL|A|)^n||_\infty = 0$$

Tomando límite en (9) se deduce que

$$||Y - Z||_\infty \leq |y - z|||(I - hL|A|)^{-1}||_\infty$$

Consecuentemente,

$$\begin{aligned}
|\Phi(t,y,h) - \Phi(t,z,h)| &\leq \sum_{i=1}^{p} |b_i||f(t_k^{(i)}, y^{(i)}) - f(t_k^{(i)}, z^{(i)})| \\
&\leq L \sum_{i=1}^{p} |b_i||y^{(i)} - z^{(i)}| \\
&\leq L \sum_{i=1}^{p} |b_i|||Y - Z||_\infty \\
&\leq \left(L||(I - hL|A|)^{-1}||_\infty \sum_{i=1}^{p} |b_i| \right) |y - z|,
\end{aligned}$$

deduciéndose que Φ es de Lipschitz en la variable y, así que el método es estable.

(iii) Como

$$\Phi(t,y,0) = f(t,y) \sum_{i=1}^{p} b_i,$$

el Teorema 5 permite afirmar que el método es consistente si y solo si

$$\sum_{i=1}^{p} b_i = 1$$

(iv) Consecuencia inmediata de los apartados anteriores y del Teorema 4. □

Una vez estudiada la convergencia de los métodos de Runge-Kutta, la próxima parada es el estudio del orden, empleándose para ello el Teorema 8.

Teorema 11. *Considérese un método de Runge-Kutta dado por las matrices*

$$b = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix} \qquad C = \begin{pmatrix} c_1 & 0 & \dots & 0 \\ 0 & c_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & c_n \end{pmatrix} \qquad A = \begin{pmatrix} a_{1,1} & a_{1,2} & \dots & a_{1,n} \\ a_{2,1} & a_{2,2} & \dots & a_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n,1} & a_{n,2} & \dots & a_{n,n} \end{pmatrix}$$

Entonces

(i) *El método es de orden* 1 *si y solo si*

$$B^tE = 1$$

(ii) *El método es de orden* 2 *si y solo si se verifica la condición de orden* 1*, y además,*

$$B^tAE = \frac{1}{2}$$

(iii) *El método es de orden* 3 *si y solo si se verifican las condiciones de orden* 2*, y además,*

$$\begin{cases} B^tC^2E = \dfrac{1}{3} \\ B^tACE = \dfrac{1}{6} \end{cases}$$

(iv) *El método es de orden* 4 *si y solo si se verifican las condiciones de orden* 3*, y además,*

$$\begin{cases} B^tC^3E = \dfrac{1}{4} \\ B^tAC^2E = \dfrac{1}{12} \\ B^tA^2CE = \dfrac{1}{24} \\ B^tCACE = \dfrac{1}{8} \end{cases}$$

Demostración. Que no cunda el pánico: solo se van a probar los dos primeros apartados. En primer lugar, el teorema anterior permite afirmar que el método es consistente si y solo si $B^tE = 1$, luego, por el Corolario 3, el método es de orden 1 si y solo si $B^tE = 1$. Para el apartado segundo, se recuerda que

$$\Phi(t,y,h) = \sum_{i=1}^{p} b_i f(t^{(i)}, y^{(i)}), \qquad t^{(i)} = t + c_i h, \qquad y^{(i)} = y + h\sum_{j=1}^{p} a_{i,j} f(t^{(j)}, y^{(j)}),$$

donde $t^{(i)}$ e $y^{(i)}$ pueden verse como funciones de (t,y,h). En consecuencia,

$$\frac{\partial \Phi}{\partial h}(t,y,h) = \sum_{i=1}^{p} b_i \left(\frac{\partial f}{\partial t}(t^{(i)}, y^{(i)}) \frac{\partial t^{(i)}}{\partial h}(t,y,h) + \frac{\partial f}{\partial y}(t^{(i)}, y^{(i)}) \frac{\partial y^{(i)}}{\partial h}(t,y,h) \right)$$

Por un lado,

$$\frac{\partial t^{(i)}}{\partial h}(t,y,h) = 0$$

Por otro,

$$\frac{\partial y^{(i)}}{\partial h}(t,y,h) = \sum_{j=1}^{p} a_{i,j} f(t^{(j)}, y^{(j)}) + h \frac{\partial}{\partial h} \left(\sum_{j=1}^{p} a_{i,j} f(t^{(j)}, y^{(j)}) \right)$$

Poniendo $h = 0$,

$$\frac{\partial y^{(i)}}{\partial h}(t,y,0) = f(t,y) \sum_{j=1}^{p} a_{i,j}$$

Volviendo arriba,

$$\begin{aligned} \frac{\partial \Phi}{\partial h}(t,y,0) &= \sum_{i=1}^{p} b_i \left(c_i \frac{\partial f}{\partial t}(t,y) + f(t,y) \frac{\partial f}{\partial y}(t,y) \sum_{j=1}^{p} a_{i,j} \right) \\ &= \frac{\partial f}{\partial t}(t,y) \sum_{i=1}^{p} b_i c_i + f(t,y) \frac{\partial f}{\partial y}(t,y) \sum_{i=1}^{p} b_i \sum_{j=1}^{p} a_{i,j} \end{aligned}$$

Recuérdese que lo primero que se hizo después de definir los métodos de Runge-Kutta es suponer

$$\sum_{j=1}^{p} a_{i,j} = c_i$$

para cada $i \in \{1,\dots,p\}$, o, equivalentemente, $AE = CE$. Así,

$$\begin{aligned}\frac{\partial \Phi}{\partial h}(t,y,0) &= \frac{\partial f}{\partial t}(t,y)\sum_{i=1}^{p} b_i c_i + f(t,y)\frac{\partial f}{\partial y}(t,y)\sum_{i=1}^{p} b_i c_i \\ &= \left(\frac{\partial f}{\partial t}(t,y) + f(t,y)\frac{\partial f}{\partial y}(t,y)\right)\sum_{i=1}^{p} b_i c_i \\ &= f^{(1)}(t,y)\sum_{i=1}^{p} b_i c_i\end{aligned}$$

Por tanto, el método es de orden 2 si y solo si

$$\sum_{i=1}^{p} b_i = 1 \qquad \text{y} \qquad \sum_{i=1}^{p} b_i c_i = \frac{1}{2},$$

o sea, si y solo si $B^t E = 1$ y $B^t CE = B^t AE = \frac{1}{2}$, como quería probarse. □

Teorema 12. *Considérese un método de Runge-Kutta de p etapas. Entonces*

(i) *El orden del método es a lo sumo $2p$.*

(ii) *Si el método es explícito, su orden es a lo sumo p.*

(iii) *Si el método es explícito y $q \geq 5$, su orden es estrictamente menor que p.*

Demostración. Escapa a los propósitos de la asignatura. □

Ejemplo. Se recuerda que el tablero de Butcher del método de Euler es

0	0
	1

Se observa que $b_1 = 1$ y $b_1 c_1 = 0 \neq \frac{1}{2}$, luego el método es de orden exactamente 1.

Ejemplo. Se recuerda que el tablero de Butcher del método de Euler implícito es

1	1
	1

Se observa que $b_1 = 1$ y $b_1 c_1 = 1 \neq \frac{1}{2}$, luego el método es de orden exactamente 1.

Ejemplo. El tablero de Butcher de un método de Runge-Kutta de una etapa es

a	a
	1

Por tanto, si $a = \frac{1}{2}$, el método es de orden 2, y si no, el método es de orden exactamente 1.

Ejemplo. Dados $b_1, b_2 \in \mathbb{R}$ con $b_1 + b_2 = 1$, el tablero de Butcher de un método de Runge-Kutta de dos etapas explícito es

0	0	0
a	a	0
	b_1	b_2

Por tanto, si $ab_2 = \frac{1}{2}$, el método es de orden 2, y si no, el método es de orden exactamente 1.

Ejemplo. Los métodos de Runge-Kutta dados por los tableros siguientes son de orden 2:

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

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

En efecto, no hay más que volver al ejemplo anterior observando que en ambos casos se tiene

$$b_1 = 1 - \frac{1}{2a} \qquad \text{y} \qquad b_2 = \frac{1}{2a}$$

TEMA 3

Métodos multipaso

3.1. Motivación

Como se ha estudiado en el tema anterior, los datos que utiliza un método unipaso para calcular la aproximación y_{k+1} son h, t_k e y_k. Con objeto de mejorar las aproximaciones obtenidas, quizá sería conveniente utilizar todas las aproximaciones realizadas (es decir, $y_0, y_1, \ldots, y_k$) para hallar y_{k+1}. De nuevo, consideramos la ecuación

$$y(t_{k+1}) = y(t_k) + \int_{t_k}^{t_{k+1}} f(t, y(t))\, dt, \tag{10}$$

y tenemos que aproximar $y(t_{k+1})$ es equivalente a aproximar la integral. Si ya se conocen y_{k-1} e y_k, podemos estimar la gráfica de f en $[t_k, t_{k+1}]$ mediante la recta r que pasa por $f(t_{k-1}, y_{k-1})$ y $f(t_k, y_k)$. De esta manera, la integral a calcular no es más que el área encerrada por dicha recta y el eje x. Si llamamos $f_k = f(t_k, y_k)$, una parametrización de la recta r sería

$$r(t) = f_k + \frac{f_k - f_{k-1}}{h}(t - t_k)$$

y aproximamos la integral como sigue:

$$\int_{t_k}^{t_{k+1}} f(t, y(t))\, dt \approx \int_{t_k}^{t_{k+1}} r(t)\, dt = f_k h + \frac{f_k - f_{k-1}}{h}\frac{h^2}{2} = f_k h + h\frac{f_k - f_{k-1}}{2} = \frac{h}{2}(3f_k - f_{k-1})$$

En consecuencia,

$$y_{k+1} = y_k + \frac{h}{2} + \frac{h}{2}(3f_k - f_{k-1}), \quad k \in \{1, \ldots, n-1\}$$

Esta expresión no es válida para $k = 0$, pero y_1 se puede aproximar usando un método unipaso. Como se ha conseguido aproximar la integral utilizando dos aproximaciones calculadas previamente, se dice que el método obtenido es *de dos pasos*. Si también se quiere usar y_{k-2} para aproximar la integral, se puede hallar el polinomio de interpolación de los puntos (t_{k-2}, y_{k-2}), (t_{k-1}, y_{k-1}) y (t_k, y_k) y estimar la integral en (10) mediante la integral de dicho polinomio en el intervalo $[t_k, t_{k+1}]$.

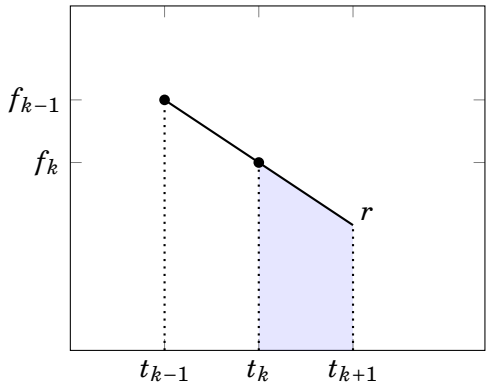

Figura 2. Interpretación gráfica de un método de dos pasos.

3.2. Interpolación polinómica

La situación es la siguiente: dados $k+1$ puntos del plano $(t_0, f_0), (t_1, f_1), \ldots, (t_k, f_k)$ tales que $t_i \neq t_j$ si $i \neq j$, se trata de encontrar un polinomio P de grado menor o igual que k cuya gráfica pase por los $k+1$ puntos, es decir, tal que $P(t_i) = f_i$ para todo $i \in \{0, 1, \ldots, k\}$. Es bien sabido que este polinomio existe y es único, y se denomina *polinomio de interpolación de los $k+1$ puntos*. Se proporcionan a continuación un par de posibles construcciones del polinomio de interpolación:

Definición 18. Los polinomios

$$l_i(t) = \frac{(t-t_0)\ldots\widehat{(t-t_i)}\ldots(t-t_k)}{(t_i-t_0)\ldots\widehat{(t_i-t_i)}\ldots(t_i-t_k)}, \qquad i \in \{0, 1, \ldots, k\}$$

se denominan **polinomios de base de interpolación de Lagrange**. La **forma de Lagrange del polinomio de interpolación** no es más que

$$P(t) = \sum_{i=0}^{k} f_i l_i(t)$$

Definición 19. La **forma de Newton del polinomio de interpolación** es

$$P(t) = f[t_k] + f[t_{k-1}, t_k](t-t_k) + \ldots + f[t_0, t_1, \ldots, t_k](t-t_k)\ldots(t-t_1),$$

donde, para cualquier $i \in \{0, 1, \ldots, l\}$,

$$f[t_i] := f_i$$

es una **diferencia dividida de orden 0**, y para cualquier subconjunto de índices distintos dos a dos $\{i_0, \ldots, i_m\} \subset \{0, 1, \ldots, k\}$,

$$f[t_{i_0}, t_{i_1}, \ldots, t_{i_m}] := \frac{f[t_{i_1}, t_{i_2}, \ldots, t_{i_m}] - f[t_{i_0}, t_{i_1}, \ldots, t_{i_{m-1}}]}{t_{i_m} - t_{i_0}}$$

es una **diferencia dividida de orden *m***.

La comprobación de que las formas de Lagrange y Newton son válidas (es decir, que el polinomio que definen es el que interpola los $k+1$ puntos) no corresponde a esta asignatura.

Por otra parte, para calcular de las diferencias divididas que aparecen en la forma de Newton del polinomio de interpolación puede resultar de utilidad la tabla siguiente:

Puntos	Orden 0	Orden 1	Orden 2	…	Orden $k-1$	Orden k
t_0	$f[t_0]$					
		$f[t_0, t_1]$				
t_1	$f[t_1]$		$f[t_0, t_1, t_2]$			
		$f[t_1, t_2]$		$\ddots$		
t_2	$f[t_2]$					
					$f[t_0, \ldots, t_{k-1}]$	
$\vdots$	$\vdots$	$\vdots$	$\vdots$			$f[t_0, \ldots, t_k]$
					$f[t_1, \ldots, t_k]$	
t_{k-2}	$f[t_{k-2}]$					
		$f[t_{k-2}, t_{k-1}]$		$\iddots$		
t_{k-1}	$f[t_{k-1}]$		$f[t_{k-2}, t_{k-1}, t_k]$			
		$f[t_{k-1}, t_k]$				
t_k	$f[t_k]$					

Ahora bien, como los puntos nos interesa tomarlos de forma que $t_k - t_{k-1} = h$ para cada $k \in \{0,1,\dots,k\}$, entonces el denominador de cualquier diferencia dividida de orden m es $m!h^m$. ¿Qué sucede con el numerador? La definición siguiente resultará bastante conveniente:

Definición 20. Dados $k+1$ puntos $f_0,\dots,f_k$, se definen las **diferencias regresivas de orden 0** como

$$\nabla^0 f_i = f_i, \qquad i \in \{0,1,\dots,k\},$$

y para $m \in \mathbb{N}$ con $m \leq k$, se definen las **diferencias regresivas de orden m** como

$$\nabla^m f_i = \nabla^{m-1} f_i - \nabla^{m-1} f_{i-1}, \qquad i \in \{m,\dots,k\}$$

De esta manera, las diferencias divididas de orden m que aparecen en la tabla son

$$f[t_l,\dots,t_{l+m}] = \frac{\nabla^m f_{l+m}}{m!h^m}$$

para cualquier $l \in \mathbb{N}\cup\{0\}$ con $0 \leq l \leq k-m$. Como los denominadores de una misma columna son comunes, basta con escribir la tabla de las diferencias divididas atendiendo solo a los numeradores:

Puntos	Orden 0	Orden 1	Orden 2	…	Orden $k-1$	Orden k
t_0	$\nabla^0 f_0$					
		$\nabla^1 f_1$				
t_1	$\nabla^0 f_1$		$\nabla^2 f_2$			
		$\nabla^1 f_2$		$\ddots$		
t_2	$\nabla^0 f_2$				$\nabla^{k-1} f_{k-1}$	
$\vdots$	$\vdots$	$\vdots$	$\vdots$			$\nabla^k f_k$
					$\nabla^{k-1} f_k$	
t_{k-2}	$\nabla^0 f_{k-2}$					
		$\nabla^1 f_{k-1}$		$\iddots$		
t_{k-1}	$\nabla^0 f_{k-1}$		$\nabla^2 f_k$			
		$\nabla^1 f_k$				
t_k	$\nabla^0 f_k$					

Además, el polinomio de interpolación en su forma de Newton sería

$$P(t) = \nabla^0 f_k + \frac{\nabla^1 f_k}{h}(t-t_k) + \frac{\nabla^2 f_k}{2h^2}(t-t_k)(t-t_{k-1}) + \ldots + \frac{\nabla^k f_k}{k!h^k}(t-t_k)\ldots(t-t_1)$$

Si llamamos $s = \frac{t-t_k}{h}$, se tiene que

$$P(t) = \nabla^0 f_k + \nabla^1 f_k s + \frac{\nabla^2 f_k}{2}s(s+1) + \frac{\nabla^3 f_k}{3!}s(s+1)(s+2) + \ldots + \frac{\nabla^k f_k}{k!}s(s+1)\ldots(s+k-1)$$

Ahora, para cualquier $s \in \mathbb{R}$ definimos

$$\binom{s}{0} := 1,$$

y para cualquier $j \in \mathbb{N}$,

$$\binom{s+j-1}{j} := \frac{s(s+1)\ldots(s+j-1)}{j!}$$

En consecuencia, puede introducirse la definición siguiente:

Definición 21. Si se define

$$\widetilde{P}(s) := \sum_{j=0}^{k} \nabla^j f_k \binom{s+j-1}{j},$$

entonces la expresión

$$P(t) = \widetilde{P}\left(\frac{t-t_k}{h}\right)$$

para el polinomio de interpolación de los puntos

$$(t_0, f_0), (t_1, f_1), \ldots, (t_k, f_k)$$

se conoce como **forma regresiva de Gregory-Newton**.

De esta manera, si $q \in \mathbb{N}$ es tal que $q \leq k$ y se define

$$\widetilde{P}(s) := \sum_{j=0}^{q} \nabla^j f_k \binom{s+j-1}{j},$$

entonces el polinomio

$$P_q(t) = \widetilde{P}_q\left(\frac{t-t_k}{h}\right)$$

es el polinomio que interpola los datos $(t_k, f_k), (t_{k-1}, f_{k-1}), \ldots, (t_{k-q}, f_{k-q})$. Téngase en cuenta que el polinomio de interpolación también puede escribirse como

$$P(t) = f[t_0] + f[t_0, t_1](t-t_0) + \ldots + f[t_0, t_1, \ldots, t_k](t-t_0)\ldots(t-t_{k-1}),$$

y entonces se pueden introducir de forma totalmente análoga las *diferencias progresivas* y la *forma progresiva* del polinomio de interpolación:

Definición 22. Dados $k+1$ puntos $f_0, \ldots, f_k$, se definen las **diferencias progresivas de orden 0** como

$$\Delta^0 f_i = f_i, \qquad i \in \{0, 1, \ldots, k\},$$

y para $m \in \mathbb{N}$ con $m \leq k$, se definen las **diferencias progresivas de orden *m*** como

$$\Delta^m f_i = \Delta^{m-1} f_{i+1} - \Delta^{m-1} f_i, \qquad i \in \{0, 1, \ldots, k-m\}$$

Definición 23. Si se define

$$\widetilde{P}(s) := \sum_{j=0}^{k} \Delta^j f_0 \binom{s}{j},$$

entonces la expresión

$$P(t) = \widetilde{P}\left(\frac{t-t_0}{h}\right)$$

para el polinomio de interpolación de los puntos

$$(t_0, f_0), (t_1, f_1), \ldots, (t_k, f_k)$$

se conoce como **forma progresiva de Gregory-Newton**.

3.3. Métodos basados en integración numérica

Como ya se adelantó al comienzo del tema, nos interesamos en aproximar la integral que aparece en la ecuación

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

utilizando alguna de las aproximaciones $y_0, y_1, \ldots, y_k$ calculadas previamente. Para ello, se hará uso de la forma regresiva de Gregory-Newton del polinomio de interpolación de ciertos puntos.

3.3.1. Métodos de Adams-Bashforth

Supóngase que quiere hallarse y_{k+1} usando q de las aproximaciones calculadas anteriormente ($q \in \mathbb{N}$, $q \leq k$), a saber, $y_k, y_{k-1}, \ldots, y_{k-q+1}$. El procedimiento a seguir consistirá en aproximar la gráfica de la función $t \mapsto f(t, y(t))$, $t \in [t_k, t_{k+1}]$ mediante el polinomio de interpolación de los puntos

$$(t_k, f_k), (t_{k-1}, f_{k-1}), \ldots, (t_{k-q+1}, f_{k-q+1}),$$

donde $f_j = f(t_j, y_j)$, $j \in \{0, 1, \ldots, k\}$ son valores ya conocidos. De esta manera, la aproximación de $y(t_{k+1})$ sería

$$y_{k+1} = y_k + \int_{t_k}^{t_{k+1}} P_{q-1}(t)\, dt,$$

donde P_{q-1} es el único polinomio de grado menor o igual que $q-1$ que interpola los q puntos

$$(t_k, f_k), (t_{k-1}, f_{k-1}), \ldots, (t_{k-q+1}, f_{k-q+1})$$

Ahora se halla la integral haciendo uso la forma regresiva de Gregory-Newton para el polinomio de interpolación, quedando

$$\int_{t_k}^{t_{k+1}} P_{q-1}(t)\, dt = \int_{t_k}^{t_{k+1}} \widetilde{P}_{q-1}\left(\frac{t - t_k}{h}\right) dt = h \int_0^1 \widetilde{P}_{q-1}(s)\, ds = h \int_0^1 \sum_{j=0}^{q-1} \nabla^j f_k \binom{s+j-1}{j} ds$$

Surge de aquí la definición del método siguiente:

Definición 24. Se define el **método de Adams-Bashforth de q pasos** o **método AB de q pasos** como

$$y_{k+1} = y_k + h \sum_{j=0}^{q-1} \gamma_j \nabla^j f_k,$$

donde, para cada $j \in \{0, 1, \ldots, q-1\}$,

$$\gamma_j = \int_0^1 \binom{s+j-1}{j} ds$$

A continuación, se hallará la expresión del método de Adams-Bashforth de 1, 2 y 3 pasos.

(i) El método AB de 1 paso viene dado por

$$y_{k+1} = y_k + h\gamma_0 \nabla^0 f_k,$$

siendo

$$\gamma_0 = \int_0^1 \binom{s-1}{0} ds = 1, \qquad \nabla^0 f_k = f_k$$

Así, obtenemos

$$y_{k+1} = y_k + h f(t_k, y_k)$$

Una grata sorpresa: el método de Euler.

(ii) El método AB de 2 pasos viene dado por

$$y_{k+1} = y_k + h\left(\gamma_0 \nabla^0 f_k + \gamma_1 \nabla^1 f_k\right),$$

siendo

$$\gamma_1 = \int_0^1 \binom{s}{1} ds = \frac{1}{2}, \qquad \nabla^1 f_k = f_k - f_{k-1}$$

Así, obtenemos

$$y_{k+1} = y_k + h\left(f_k + \frac{1}{2}f_k - \frac{1}{2}f_{k-1}\right) = y_k + \frac{h}{2}(3f_k - f_{k-1})$$

(iii) El método AB de 3 pasos viene dado por

$$y_{k+1} = y_k + h\left(\gamma_0\nabla^0 f_k + \gamma_1\nabla^1 f_k + \gamma_2\nabla^2 f_k\right),$$

siendo

$$\gamma_2 = \int_0^1 \binom{s+1}{2} ds = \int_0^1 \frac{s(s+1)}{2} ds = \frac{5}{12}, \qquad \nabla^2 f_k = f_k - 2f_{k-1} + f_{k-2}$$

Así, obtenemos

$$y_{k+1} = y_k + \frac{h}{2}(3f_k - f_{k-1}) + \frac{5h}{12}(f_k - 2f_{k-1} + f_{k-2}) = y_k + \frac{h}{12}(23f_k - 16f_{k-1} + 5f_{k-2})$$

En general, adoptando un pequeño cambio de notación, el método AB de q pasos se puede escribir en la forma

$$y_{k+q} = y_{k+q-1} + h\sum_{j=0}^{q-1} f_{k+j}\beta_j = y_{k+q-1} + h\left(\beta_{q-1}f_{k+q-1} + \beta_{q-2}f_{k+q-2} + \ldots + \beta_0 f_k\right), \quad k \in \{0,1,\ldots,n-q\}$$

3.3.2. Métodos de Adams-Moulton

Procediendo de forma totalmente análoga, se trata de aproximar la gráfica de la función $t \mapsto f(t, y(t))$, $t \in [t_k, t_{k+1}]$ mediante el polinomio de interpolación de los puntos

$$(t_{k+1}, f_{k+1}), (t_k, f_k), \ldots, (t_{k-q+1}, f_{k-q+1}),$$

donde ahora f_{k+1} es un valor desconocido a priori (lo que va a dar lugar a métodos implícitos) y $q \in \mathbb{N}\cup\{0\}$ es tal que $q - 1 \leq k$. De esta manera, la aproximación de $y(t_{k+1})$ sería

$$y_{k+1} = y_k + \int_{t_k}^{t_{k+1}} Q_q(t)\,dt,$$

donde Q_q es el único polinomio de grado menor o igual que q que interpola los $q+1$ puntos

$$(t_{k+1}, f_{k+1}), (t_k, q_k), \ldots, (t_{k-q+1}, f_{k-q+1}),$$

Ahora se halla la integral haciendo uso la forma regresiva de Gregory-Newton para el polinomio de interpolación, quedando

$$\int_{t_k}^{t_{k+1}} Q_q(t)\,dt = \int_{t_k}^{t_{k+1}} \widetilde{Q}_q\left(\frac{t - t_{k+1}}{h}\right) dt = h\int_{-1}^{0} \widetilde{Q}_q(s)\,ds = h\int_{-1}^{0} \sum_{j=0}^{q} \nabla^j f_{k+1} \binom{s+j-1}{j} ds$$

Surge de aquí la definición del método siguiente:

Definición 25. Se define el **método de Adams-Moulton de q pasos** o **método AM de q pasos** como

$$y_{k+1} = y_k + h\sum_{j=0}^{q} \widetilde{\gamma}_j \nabla^j f_{k+1},$$

donde, para cada $j \in \{0,1,\ldots,q-1\}$,

$$\widetilde{\gamma}_j = \int_{-1}^{0} \binom{s+j-1}{j} ds$$

A continuación, se hallará la expresión del método de Adams-Moulton de 0, 1 y 2 pasos.

(*i*) El método AM de 0 pasos viene dado por

$$y_{k+1} = y_k + h\widetilde{\gamma}_0 \nabla^0 f_{k+1},$$

siendo

$$\widetilde{\gamma}_0 = \int_{-1}^{0} \binom{s-1}{0} ds = 1, \qquad \nabla^0 f_{k+1} = f_{k+1}$$

Así, obtenemos

$$y_{k+1} = y_k + hf(t_{k+1}, y_{k+1}),$$

es decir, el método de Euler implícito.

(*ii*) El método AM de 1 paso viene dado por

$$y_{k+1} = y_k + h\left(\widetilde{\gamma}_0 \nabla^0 f_{k+1} + \widetilde{\gamma}_1 \nabla^1 f_{k+1}\right),$$

siendo

$$\widetilde{\gamma}_1 = \int_{-1}^{0} \binom{s}{1} ds = -\frac{1}{2}, \qquad \nabla^1 f_{k+1} = f_{k+1} - f_k$$

Así, obtenemos

$$y_{k+1} = y_k + h\left(f_{k+1} - \frac{1}{2} f_{k+1} + \frac{1}{2} f_k\right) = y_k + \frac{h}{2}(f_{k+1} + f_k),$$

es decir, el método del trapecio.

(*iii*) El método AM de 2 pasos viene dado por

$$y_{k+1} = y_k + h\left(\widetilde{\gamma}_0 \nabla^0 f_{k+1} + \gamma_1 \nabla^1 f_{k+1} + \widetilde{\gamma}_2 \nabla^2 f_{k+1}\right),$$

siendo

$$\widetilde{\gamma}_2 = \int_{-1}^{0} \binom{s+1}{2} ds = \int_{-1}^{0} \frac{s(s+1)}{2} ds = -\frac{1}{12}, \qquad \nabla^2 f_{k+1} = f_{k+1} - 2f_k + f_{k-1}$$

Así, obtenemos

$$y_{k+1} = y_k + \frac{h}{2}(f_{k+1} + f_k) - \frac{h}{12}(f_{k+1} - 2f_k + f_{k-1}) = y_k + \frac{h}{12}(5f_{k+1} + 8f_k - f_{k-1})$$

En general, todas las aproximaciones del método de q pasos pueden escribirse en la forma

$$y_{k+q} = y_{k+q-1} + h\sum_{j=0}^{q} f_{k+j}\widetilde{\beta}_j = y_{k+q-1} + h\left(\widetilde{\beta}_q f_{k+q} + \widetilde{\beta}_{q-1} f_{k+q-1} + \ldots + \widetilde{\beta}_0 f_k\right), \quad k \in \{0, 1, \ldots, n-q\}$$

Esta es la expresión de un método implícito; para hallar y_{k+q} es necesario resolver una ecuación.

3.4. Métodos basados en diferenciación numérica

El problema a resolver es, una vez más, el cálculo de y_{k+1} una vez conocidas las aproximaciones $y_0, \ldots, y_k$. Se sabe que

$$y'(t_{k+1}) = f(t_{k+1}, y(t_{k+1})), \tag{11}$$

lo que sugiere de manera clamorosa un nuevo procedimiento de obtención de y_{k+1}: emplear ténicas de diferenciación numérica. La aproximación más natural de $y'(t_{k+1})$ surge de la propia definición de la derivada:

$$y'(t_{k+1}) \approx \frac{y_{k+1} - y_k}{h}$$

También resulta natural realizar la aproximación

$$f(t_{k+1}, y(t_{k+1})) \approx f(t_{k+1}, y_{k+1}) = f_{k+1}$$

Sustituyendo en (11), se obtiene

$$y_{k+1} = y_k + h f_{k+1},$$

que no es más que el método de Euler implícito.

En general, si quieren usarse q aproximaciones ya calculadas para aproximar $y'(t_{k+1})$, por ejemplo, $y_{k+1}, y_k, \ldots, y_{k-q+1}$, el procedimiento a seguir tratará de hallar el polinomio de interpolación de los puntos

$$(t_{k+1}, y_{k+1}), (t_k, y_k), \ldots, (t_{k-q+1}, y_{k-q+1}),$$

y calcular su derivada en el punto t_{k+1}. Llamando R_q a este polinomio, la ecuación (11) se aproximaría mediante el método siguiente:

$$R_q'(t_{k+1}) = f_{k+1} \tag{12}$$

Ahora bien, se recuerda que

$$R_q(t) = \widetilde{R}_q\left(\frac{t - t_{k+1}}{h}\right),$$

donde

$$\widetilde{R}_q(s) = \sum_{j=0}^{q} \nabla^j y_{k+1} \binom{s+j-1}{j}$$

Por tanto, por la regla de la cadena,

$$R_q'(t) = \frac{1}{h} \widetilde{R}_q'\left(\frac{t - t_{k+1}}{h}\right)$$

En consecuencia,

$$R_q'(t_{k+1}) = \frac{1}{h} \widetilde{R}_q'(0) = \frac{1}{h} \sum_{j=0}^{q} \nabla^j y_{k+1} \frac{d}{ds} \binom{s+j-1}{j}\bigg|_{s=0}$$

Al sustituir en (12), se obtiene la siguiente familia de métodos:

Definición 26. El **método BDF de q pasos** es aquel dado por

$$\sum_{j=0}^{q} \nabla^j y_{k+1} \delta_j = h f_{k+1},$$

donde

$$\delta_j = \frac{d}{ds} \binom{s+j-1}{j}\bigg|_{s=0}$$

Las siglas BDF proceden de *Backward Differentiation Formula*. Hallemos la expresión de estos métodos en los casos más sencillos:

(*i*) El método BDF de 1 paso viene dado por

$$\delta_1 \nabla^1 y_{k+1} = h f_{k+1},$$

donde

$$\delta_1 = \frac{ds}{ds}\bigg|_{s=0} = 1, \qquad \nabla^1 y_{k+1} = y_{k+1} - y_k$$

Así,

$$y_{k+1} = y_k + h f_{k+1}$$

Otra vez: el método de Euler implícito.

(ii) El método BDF de 2 pasos viene dado por

$$\delta_1 \nabla^1 y_{k+1} + \delta_2 \nabla^2 y_{k+1} = h f_{k+1},$$

donde

$$\delta_2 = \frac{d}{ds}\frac{s(s+1)}{2}\bigg|_{s=0} = \frac{1}{2}, \qquad \nabla^2 y_{k+1} = y_{k+1} - 2y_k + y_{k-1}$$

Así, sustituyendo arriba,

$$y_{k+1} - y_k + \frac{1}{2}y_{k+1} - y_k + \frac{1}{2}y_{k-1} = \frac{3}{2}y_{k+1} - 2y_k + \frac{1}{2}y_{k-1} = h f_{k+1}$$

En general, el método BDF de q pasos adopta la forma

$$\sum_{j=0}^{q} \alpha_j y_{k+j} = h\beta_q f_{k+q}, \quad k \in \{0, 1, \ldots, n-q\}$$

para unos ciertos coeficientes $\alpha_j, \beta \in \mathbb{R}$.

3.5. Expresión general de un método multipaso

Recapitulando, hemos visto que los métodos basados en integración numérica de q pasos se escriben como

$$y_{k+q} = y_{k+q-1} + h\left(\widetilde{\beta}_q f_{k+q} + \widetilde{\beta}_{q-1} f_{k+q-1} + \ldots + \widetilde{\beta}_0 f_k\right), \quad k \in \{0, 1, \ldots, n-q\}$$

Dependiendo de si $\widetilde{\beta}_q = 0$ o $\widetilde{\beta}_q \neq 0$, estaremos ante un método AM o un método AB. Por otra parte, los métodos basados en diferenciación numérica admiten una expresión del tipo

$$\alpha_0 y_k + \alpha_1 y_{k+1} + \ldots + \alpha_q y_{k+q} = h\beta_q f_{k+q}, \quad k \in \{0, 1, \ldots, n-q\},$$

con $\alpha_q \neq 0$ para poder despejar y_{k+q}. Las dos expresiones anteriores pueden aunarse en una sola para proporcionar la definición que sigue.

Definición 27. Un método numérico de la forma

$$\sum_{j=0}^{q} \alpha_j y_{k+j} = h \sum_{j=0}^{q} \beta_j f_{k+j}, \quad k \in \{0, 1, \ldots, n-q\}$$

se denomina **método multipaso lineal de q pasos**, o, simplemente, **método de q pasos**, donde

$$|\alpha_0| + |\beta_0| > 0, \qquad \alpha_q \neq 0$$

Si $\beta_q = 0$, se dice que el método es **explícito**, y en caso contrario, que es **implícito**.

En primer lugar, si se considera un método de q pasos, téngase en cuenta que el cálculo de las q primeras aproximaciones debe realizarse mediante un método unipaso, y, a partir de ahí, el método multipaso arranca sin problema.

En segundo lugar, la condición $\alpha_q \neq 0$ se pide para que aparezca y_{k+q} en el método, mientras que la condición $|\alpha_0| + |\beta_0| > 0$ se exige para que salga y_k. Usualmente, por motivos de comodidad, se va a dividir por α_q en dicha expresión para que quede otra totalmente equivalente en la que el coeficiente de y_{k+q} es 1 En caso de que $\beta_q \neq 0$, puesto que f_{k+q} es un dato desconocido, será necesario resolver la ecuación

$$y_{k+q} = h\beta_q f(t_{k+q}, y_{k+q}) + C_{k+q}, \tag{13}$$

siendo

$$C_{k+q} = -\sum_{j=0}^{q-1} \alpha_j y_{k+j} + h \sum_{j=0}^{q-1} \beta_j f_{k+j}$$

un número conocido. Evidentemente, para la buena definición de los métodos implícitos es menester que la ecuación (13) tenga solución única. Para ello, como se hizo en el capítulo anterior, se trata de probar la contractividad de la función $g\colon \mathbb{R} \to \mathbb{R}$ dada por

$$g(y) = h\beta_q f(t_{k+q}, y) + C_{k+q}$$

Si $y_1, y_2 \in \mathbb{R}$, usando que f es de Lipschitz en la variable y,

$$|g(y_1) - g(y_2)| \leq hL|\beta_q|\,|y_1 - y_2|$$

Esta desigualdad y el teorema del punto fijo permiten afirmar que una condición suficiente para que el método implícito esté bien definido es

$$h < \frac{1}{L|\beta_q|}$$

3.6. Orden de un método multipaso

Una vez asegurada la buena definición de toda clase de métodos multipaso lineales, continuamos con el estudio del error, el orden, la convergencia y ese tipo de características. Las definiciones que van a introducirse serán rescatadas del tema anterior, y cuando no se puedan copiar y pegar de forma literal, serán adaptadas convenientemente.

Definición 28. Considérese un método de q pasos.

(*i*) Dado $k \in \{0, 1, \ldots, n\}$, se define el **error en la etapa *k*-ésima** como

$$e_k = |y(t_k) - y_k|$$

(*ii*) Se denomina **error global** al número real

$$e(h) = \max_{k=0,1,\ldots,n} e_k$$

(*iii*) Se dice que el método es **convergente** si

$$\lim_{h \to 0} e(h) = 0$$

(*iv*) Dado $k \in \{0, 1, \ldots, n\}$, se define el **error de discretización local en la etapa $k+q$-ésima** como

$$\varepsilon_{k+q} = y(t_{k+q}) - \widetilde{y}_{k+q},$$

donde $\widetilde{y}_{k+q}$ viene dado por

$$\alpha_q \widetilde{y}_{k+q} + \sum_{j=0}^{q-1} \alpha_j y(t_{k+j}) = h\beta_q f(t_{k+q}, \widetilde{y}_{k+q}) + h \sum_{j=0}^{q-1} \beta_j f(t_{k+j}, y(t_{k+j})),$$

o sea, $\widetilde{y}_{k+q}$ es la aproximación de $y(t_{k+q})$ que daría el método si se conociesen de forma exacta las aproximaciones $y_k, \ldots, y_{k+q-1}$. Si $l \in \{1, 2, \ldots, q-1\}$, el **error de discretización local en la etapa *l*-ésima** se define como

$$\varepsilon_l = y(t_{l+1}) - y(t_l) - h\Phi(t_l, y(t_l), h)$$

donde Φ es la función incremento del método unipaso usado para calcular $y_0, y_1, \ldots, y_{q-1}$.

(v) Se dice que el método es **consistente** si

$$\lim_{h\to 0} \sum_{k=1}^{n} |\varepsilon_k| = 0$$

(vi) Si $y \in \mathcal{C}^{p+1}([t_0, t_0+T], \mathbb{R})$, se dice que el método es **de orden p** si para todo $k \in \{0,1,\ldots,n-q\}$ es

$$\varepsilon_{k+q} = O(h^{p+1})$$

En principio, esta definición de orden parece tener poco que ver con la que se proporcionó para métodos unipaso. Ahora bien, recuérdese que en la demostración del Teorema 8 se probó que si un método es de orden p (y además se verificaban ciertas condiciones de regularidad), entonces $\varepsilon_k = O(h^{p+1})$ para todo $k \in \{0,\ldots,n-1\}$. Resulta que el recíproco de esta proposición es cierto (aunque no vaya a probarse), lo que justifica que el orden de un método multipaso se defina de esta manera.

Teorema 13. *Considérese un método de q pasos dado por*

$$\sum_{j=0}^{q} \alpha_j y_{k+j} = h \sum_{j=0}^{q} \beta_j f_{k+j}$$

Si se verifica

$$\sum_{j=0}^{q} \alpha_j j^l = l \sum_{j=0}^{q} \beta_j j^{l-1} \text{ para todo } l \in \{1,2,\ldots,p\}, \qquad \sum_{j=0}^{q} \alpha_j = 0,$$

entonces el método es orden p.

Demostración. Probemos que $\varepsilon_{k+q} = O(h^{p+1})$, con

$$\varepsilon_{k+q} = y(t_{k+q}) - \widetilde{y}_{k+q},$$

siendo $\widetilde{y}_{k+q}$ tal que

$$\widetilde{y}_{k+q} + \sum_{j=0}^{q-1} \alpha_j y(t_{k+j}) = h\beta_q f(t_{k+q}, \widetilde{y}_{k+q}) + h \sum_{j=0}^{q-1} \beta_j f(t_{k+j}, y(t_{k+j})),$$

donde se ha tomado $\alpha_q = 1$ en la expresión del método multipaso. Se tiene entonces

$$\begin{aligned}
\varepsilon_{k+q} &= y(t_{k+q}) + \sum_{j=0}^{q-1} \alpha_j y(t_{k+j}) - h\beta_q f(t_{k+q}, \widetilde{y}_{k+q}) - h \sum_{j=0}^{q-1} \beta_j f(t_{k+j}, y(t_{k+j})) \\
&= \sum_{j=0}^{q} \alpha_j y(t_{k+j}) - h \sum_{j=0}^{q} \beta_j f(t_{k+j}, y(t_{k+j})) + h\beta_q \left(f(t_{k+q}, y(t_{k+q})) - f(t_{k+q}, \widetilde{y}_{k+q}) \right) \\
&= \sum_{j=0}^{q} \alpha_j y(t_{k+j}) - h \sum_{j=0}^{q} \beta_j y'(t_{k+j}) + h\beta_q \left(f(t_{k+q}, y(t_{k+q})) - f(t_{k+q}, \widetilde{y}_{k+q}) \right) \\
&= L(y, t_k, h) + h\beta_q \left(f(t_{k+q}, y(t_{k+q})) - f(t_{k+q}, \widetilde{y}_{k+q}) \right),
\end{aligned}$$

donde

$$L(y, t_k, h) := \sum_{j=0}^{q} \alpha_j y(t_{k+j}) - h \sum_{j=0}^{q} \beta_j y'(t_{k+j})$$

Así,

$$L(y, t_k, h) = \varepsilon_{k+q} - h\beta_q \left(f(t_{k+q}, y(t_{k+q})) - f(t_{k+q}, \widetilde{y}_{k+q}) \right)$$

Vamos a probar primero que $L(y,t_k,h)=O(h^{p+1})$ si y solo si

$$\sum_{j=0}^{q}\alpha_j j^l = l\sum_{j=0}^{q}\beta_j j^{l-1} \text{ para todo } l\in\{1,2,\dots,p\}, \qquad \sum_{j=0}^{q}\alpha_j = 0,$$

Como y es de clase $p+1$, por la fórmula del resto de Lagrange, para cada $j\in\{0,1,\dots,q\}$ se tiene

$$\begin{aligned} y(t_{k+j}) &= y(t_k+jh)\\ &= y(t_k)+y'(t_k)jh+\frac{y''(t_k)}{2}(jh)^2+\ldots+\frac{y^{(p)}(t_k)}{p!}(jh)^p+O(h^{p+1})\\ &= \sum_{l=0}^{p}\frac{y^{(l)}(t_k)}{l!}(jh)^l+O(h^{p+1}) \end{aligned}$$

Análogamente,

$$\begin{aligned} y'(t_{k+j}) &= y'(t_k+jh)\\ &= y'(t_k)+y''(t_k)jh+\frac{y'''(t_k)}{2}(jh)^2+\ldots+\frac{y^{(p)}(t_k)}{(p-1)!}(jh)^{p-1}+O(h^p)\\ &= \sum_{l=1}^{p}\frac{y^{(l)}(t_k)}{(l-1)!}(jh)^{l-1}+O(h^p) \end{aligned}$$

Por tanto,

$$\begin{aligned} L(y,t_k,h) &= \sum_{j=0}^{q}\alpha_j\left(\sum_{l=0}^{p}\frac{y^{(l)}(t_k)}{l!}(jh)^l+O(h^{p+1})\right)-h\sum_{j=0}^{q}\beta_j\left(\sum_{l=1}^{p}\frac{y^{(l)}(t_k)}{(l-1)!}(jh)^{l-1}+O(h^p)\right)\\ &= \sum_{j=0}^{q}\alpha_j\sum_{l=0}^{p}\frac{y^{(l)}(t_k)}{l!}(jh)^l-h\sum_{j=0}^{q}\beta_j\sum_{l=1}^{p}\frac{y^{(l)}(t_k)}{(l-1)!}(jh)^{l-1}+O(h^{p+1})\\ &= \sum_{j=0}^{q}\left(\alpha_j\sum_{l=0}^{p}\frac{y^{(l)}(t_k)}{l!}(jh)^l-h\beta_j\sum_{l=1}^{p}\frac{y^{(l)}(t_k)}{(l-1)!}(jh)^{l-1}\right)+O(h^{p+1})\\ &= \sum_{j=0}^{q}\left(\alpha_j y(t_k)+\alpha_j\sum_{l=1}^{p}\frac{y^{(l)}(t_k)}{l!}(jh)^l-h\beta_j\sum_{l=1}^{p}\frac{y^{(l)}(t_k)}{(l-1)!}(jh)^{l-1}\right)+O(h^{p+1})\\ &= \sum_{j=0}^{q}\left(\alpha_j y(t_k)+\sum_{l=1}^{p}\frac{y^{(l)}(t_k)}{l!}h^l\left(\alpha_j j^l-l\beta_j j^{l-1}\right)\right)+O(h^{p+1})\\ &= y(t_k)\sum_{j=0}^{q}\alpha_j+\sum_{l=1}^{p}\frac{y^{(l)}(t_k)}{l!}h^l\sum_{j=0}^{q}\left(\alpha_j j^l-l\beta_j j^{l-1}\right)+O(h^{p+1}) \end{aligned}$$

De aquí se deduce que $L(y,t_k,h)=O(h^{p+1})$ si y solo si

$$\sum_{j=0}^{q}\alpha_j = 0, \qquad \sum_{j=0}^{q}\alpha_j j^l = l\sum_{j=0}^{q}\beta_j j^{l-1} \text{ para todo } l\in\{1,2,\dots,p\}$$

Ahora se va a probar que $L(y,t_k,h)=O(h^{p+1})$ si y solo si $\varepsilon_{k+q}=O(h^{p+1})$, lo que concluirá la prueba. Recuérdese que

$$L(y,t_k,h)=\varepsilon_{k+q}-h\beta_q\left(f(t_{k+q},y(t_{k+q}))-f(t_{k+q},\widetilde{y}_{k+q})\right)$$

y por tanto si el método es explícito ($\beta_q=0$) hemos terminado. Supóngase entonces que $\beta_q\neq 0$, y también que $\varepsilon_{k+q}=O(h^{p+1})$. Esto significa que existen $C,h^*>0$ tales que para todo $h\in(0,h^*)$ se verifica

$$|\varepsilon_{k+q}|\leq Ch^{p+1}$$

Por tanto, usando la desigualdad triangular y la condición de Lipschitz en la variable y para f,

$$|L(y,t_k,h)| \leq |\varepsilon_{k+q}| + h^* L|\beta_q||\varepsilon_{k+q}| = (1+h^* L|\beta_q|)|\varepsilon_{k+q}| \leq C(1+h^* L|\beta_q|)h^{p+1} = \widetilde{C}h^{p+1},$$

siendo $\widetilde{C} = C(1+h^* L|\beta_q|)$ una constante positiva e independiente de h. Así, se puede afirmar que $L(y,t_k,h) = O(h^{p+1})$.

Recíprocamente, supóngase que $L(y,t_k,h) = O(h^{p+1})$. En primer lugar, tómese $h_1^* > 0$ tal que

$$h_1^* < \frac{1}{L|\beta_q|}$$

Entonces

$$\left|h\beta_q\left(f(t_{k+q},y(t_{k+q})) - f(t_{k+q},\widetilde{y}_{k+q})\right)\right| \leq h^* L|\beta_q||\varepsilon_{k+q}| < |\varepsilon_{k+q}|,$$

es decir,

$$-|\varepsilon_{k+q}| < h\beta_q\left(f(t_{k+q},y(t_{k+q})) - f(t_{k+q},\widetilde{y}_{k+q})\right) < |\varepsilon_{k+q}|$$

De estas desigualdades y de la definición de $L(y,t_k,h)$ se deduce que $L(y,t_k,h)$ y ε_{k+q} tienen el mismo signo. Supóngase primero que son ambos positivos. Al ser $L(y,t_k,h) = O(h^{p+1})$, existen $C, h_2^* > 0$ tales que para todo $h \in (0,h_2^*)$ se tiene

$$|L(y,t_k,h)| = L(y,t_k,h) \leq Ch^{p+1}$$

Sea $h^* = \min\{h_1^*, h_2^*\}$. Entonces, para todo $h \in (0,h^*)$ se verifica

$$L(y,t_k,h) = \varepsilon_{k+q} - h\beta_q\left(f(t_{k+q},y(t_{k+q})) - f(t_{k+q},\widetilde{y}_{k+q})\right) \geq \varepsilon_{k+q} - h^* L|\beta_q|\varepsilon_{k+q} = \varepsilon_{k+q}(1-h^* L|\beta_q|)$$

Por tanto,

$$Ch^{p+1} \geq L(y,t_k,h) \geq \varepsilon_{k+q}(1-h^* L|\beta_q|),$$

y como $h^* L|\beta_q| < 1$, entonces $1-h^* L|\beta_q| > 0$, así que

$$\varepsilon_{k+q} = |\varepsilon_{k+q}| \leq \frac{C}{1-h^* L|\beta_q|}h^{p+1} = \widetilde{C}h^{p+1},$$

siendo

$$\widetilde{C} = \frac{C}{1-h^* L|\beta_q|}$$

una constante positiva e independiente de h. Así, tenemos que $\varepsilon_{k+q} = O(h^{p+1})$. Si el signo de $L(y,t_k,h)$ y ε_{k+q} fuese negativo, se razona análogamente. □

Ejemplo. Estudiemos el orden del método de 2 pasos dado por

$$y_{k+2} - \frac{4}{3}y_{k+1} + \frac{1}{3}y_k = \frac{2h}{3}f_{k+2}$$

En primer lugar,

$$\sum_{j=0}^{2}\alpha_j = \frac{1}{3} - \frac{4}{3} + 1 = 0$$

Además,

$$\sum_{j=0}^{2}\alpha_j j = -\frac{4}{3} + 2 = \frac{2}{3}, \qquad \sum_{j=0}^{2}\beta_j = \frac{2}{3},$$

luego el método es de orden 1. Más aún,

$$\sum_{j=0}^{2}\alpha_j j^2 = -\frac{4}{3} + 4 = \frac{8}{3}, \qquad 2\sum_{j=0}^{2}\beta_j j = 2\frac{4}{3} = \frac{8}{3},$$

así que el método es de orden 2. Sin embargo,

$$\sum_{j=0}^{2} \alpha_j j^3 = -\frac{4}{3} + 8 = \frac{20}{3}, \qquad 3\sum_{j=0}^{2} \beta_j j^2 = 3\frac{8}{3} = 8,$$

luego el método es de orden exactamente 2.

Ejemplo. Estudiemos el orden del método AB de 3 pasos:

$$y_{k+3} = y_{k+2} + \frac{h}{12}(23f_{k+2} - 16f_{k+1} + 5f_k)$$

En primer lugar,

$$\sum_{j=0}^{3} \alpha_j = -1 + 1 = 0$$

Además,

$$\sum_{j=0}^{3} \alpha_j j = -2 + 3 = 1, \qquad \sum_{j=0}^{3} \beta_j = 1,$$

luego el método es de orden 1. Más aún,

$$\sum_{j=0}^{3} \alpha_j j^2 = -4 + 9 = 5, \qquad 2\sum_{j=0}^{3} \beta_j j = \frac{2}{12}(-16 + 46) = 5$$

así que el método es de orden 2. Pero es que

$$\sum_{j=0}^{3} \alpha_j j^3 = -8 + 27 = 19, \qquad 3\sum_{j=0}^{3} \beta_j j^2 = \frac{3}{12}(-16 + 23 \cdot 4) = 19$$

El método es incluso de orden 3. Una más:

$$\sum_{j=0}^{3} \alpha_j j^4 = -16 + 81 = 65, \qquad 4\sum_{j=0}^{3} \beta_j j^3 = \frac{4}{12}(-16 + 23 \cdot 8) = 56$$

Una lástima: el método es orden exactamente 3.

Ejemplo. Se trata de encontrar un método explícito de 2 pasos que tenga el mayor orden posible. La expresión del método es

$$y_{k+2} + \alpha_1 y_{k+1} + \alpha_0 y_k = h\left(\beta_0 f_k + \beta_1 f_{k+1}\right)$$

Se tiene que

$$\sum_{j=0}^{2} \alpha_j = 0 \iff 1 + \alpha_1 + \alpha_0 = 0$$

$$\sum_{j=0}^{2} \alpha_j j = \sum_{j=0}^{2} \beta_j \iff \alpha_1 + 2 = \beta_0 + \beta_1$$

$$\sum_{j=0}^{2} \alpha_j j^2 = 2\sum_{j=0}^{2} \beta_j j \iff \alpha_1 + 4 = 2\beta_1$$

$$\sum_{j=0}^{2} \alpha_j j^3 = 3\sum_{j=0}^{2} \beta_j j^2 \iff \alpha_1 + 8 = 3\beta_1$$

Ya se dispone de un sistema de 4 ecuaciones con 4 incógnitas. Al hacer los cálculos se obtiene

$$\alpha_0 = -5, \qquad \alpha_1 = 4, \qquad \beta_0 = 2, \qquad \beta_1 = 4,$$

luego el método explícito de 2 pasos de orden máximo es de orden al menos 3, y adopta la expresión

$$y_{k+2} + 4y_{k+1} - 5y_k = h(2f_k + 4f_{k+1})$$

¿Será este método de orden 4? Pues no, porque se tiene

$$\sum_{j=0}^{2} \alpha_j j^4 = 4 + 16 = 20, \qquad 4\sum_{j=0}^{2} \beta_j j^3 = 16$$

Teorema 14. *Respecto a los métodos multipaso lineales estudiados, se verifica*

(i) *el método AB de q pasos es de orden q;*

(ii) *el método AM de q pasos es de orden $q+1$;*

(iii) *el método BDF de q pasos es de orden q.*

Demostración. Solo se va a demostrar el apartado primero. El método AB de q pasos se escribe como

$$y_{k+q} = y_{k+q-1} + h\left(\beta_{q-1} f_{k+q-1} + \beta_{q-2} f_{k+q-2} + \ldots + \beta_0 f_k\right)$$

Es claro que

$$\sum_{j=0}^{q} \alpha_j = 1 - 1 = 0$$

Se trata de probar que para todo $l \in \{1, \ldots, q\}$ se verifica

$$\sum_{j=0}^{q} \alpha_j j^l = l \sum_{j=0}^{q} \beta_j j^{l-1}$$

Fijemos $l \in \{1, \ldots, q\}$, y considérese el problema

$$\begin{cases} y'(t) = g(t, y(t)) \\ y(0) = 0, \end{cases}$$

con $g(t,y) = lt^{l-1}$. La única solución de este problema es $y(t) = t^l$, y el error de discretización local es

$$\varepsilon_{k+q} = y(t_{k+q}) - \widetilde{y}_{k+q},$$

donde

$$\begin{aligned} \widetilde{y}_{k+q} &= -\sum_{j=0}^{q-1} \alpha_j y(t_{k+j}) + h\beta_q g(t_{k+q}, \widetilde{y}_{k+q}) + h\sum_{j=0}^{q-1} \beta_j g(t_{k+j}, y(t_{k+j})) \\ &= y(t_{k+q-1}) + h\sum_{j=0}^{q-1} \beta_j l t_{k+j}^{l-1}, \end{aligned}$$

Ahora bien, por la propia definición de los métodos de Adams-Bashforth,

$$h\sum_{j=0}^{q-1} \beta_j l t_{k+j}^{l-1} = \int_{t_{k+q-1}}^{t_{k+q}} P_{q-1}(t)\,dt,$$

siendo P_{q-1} el único polinomio de grado menor o igual que q que interpola los datos

$$(t_{k+q-1}, l t_{k+q-1}^{l-1}), (t_{k+q-2}, l t_{k+q-2}^{l-1}), \ldots, (t_k, l t_k^{l-1}),$$

Pero $t \mapsto lt^{l-1}$ es un polinomio de grado menor o igual que $q-1$ que interpola los puntos anteriores, así

que $P_{q-1}(t) = lt^{l-1}$, y en consecuencia,

$$\varepsilon_{k+q} = y(t_{k+q}) - y(t_{k+q-1}) - \int_{t_{k+q-1}}^{t_{k+q}} lt^{l-1}\,dt = t_{k+q}^l - t_{k+q-1}^l - t_{k+q}^l + t_{k+q-1}^l = 0$$

En la demostración del teorema anterior se definió

$$L(y,t_k,h) := \sum_{j=0}^{q} \alpha_j y(t_{k+j}) - h\sum_{j=0}^{q} \beta_j y'(t_{k+j}),$$

y se razonó que

$$L(y,t_k,h) = \varepsilon_{k+q} - h\beta_q\left(g(t_{k+q},y(t_{k+q})) - g(t_{k+q},\widetilde{y}_{k+q})\right),$$

En este caso, por ser $\beta_q = 0$, se tiene

$$L(y,t_k,h) = \varepsilon_{k+q} = 0$$

En particular, para $k = 0$,

$$0 = L(y,t_0,h) = \sum_{j=0}^{q} \alpha_j y(t_j) - h\sum_{j=0}^{q} \beta_j y'(t_j) = \sum_{j=0}^{q} \alpha_j t_j^l - hl\sum_{j=0}^{q} \beta_j t_j^{l-1}$$

Ahora bien, como $t_0 = 0$ y $t_j = t_0 + jh$ para todo $j \in \{0,1,\ldots,q\}$, entonces

$$0 = \sum_{j=0}^{q} \alpha_j j^l h^l - hl\sum_{j=0}^{q} \beta_j j^{l-1} h^{l-1} = h^l\left(\sum_{j=0}^{q} \alpha_j j^l - l\sum_{j=0}^{q} \beta_j j^{l-1}\right)$$

Como $h^l > 0$, entonces debe ser

$$\sum_{j=0}^{q} \alpha_j j^l - l\sum_{j=0}^{q} \beta_j j^{l-1} = 0,$$

o sea,

$$\sum_{j=0}^{q} \alpha_j j^l = l\sum_{j=0}^{q} \beta_j j^{l-1},$$

que es lo que se quería demostrar. □

3.7. Estabilidad de un método multipaso

Una vez concluido el estudio del orden, la estabilidad de un método multipaso se va a definir de forma totalmente análoga al caso de los métodos unipaso.

Definición 29. Considérese un método de q pasos,

$$\sum_{j=0}^{q} \alpha_j y_{k+j} = h\sum_{j=0}^{q} \beta_j f_{k+j}, \quad k \in \{0,1,\ldots,n-q\}$$

Se dice que el método es **estable** si existe una constante M positiva e independiente de h verificando lo siguiente: si $\{y_k\}_{k=0}^n$, $\{z_k\}_{k=0}^n$, $\{\delta_k\}_{k=q}^n$ son tales que

$$\sum_{j=0}^{q} \alpha_j y_{k+j} = h\sum_{j=0}^{q} \beta_j f(t_{k+j},y_{k+j}), \qquad \sum_{j=0}^{q} \alpha_j z_{k+j} = h\sum_{j=0}^{q} \beta_j f(t_{k+j},z_{k+j}) + \delta_{k+q}$$

para cada $k \in \{0,1,\ldots,n-q\}$, entonces se tiene

$$\max_{k=q,\ldots,n} |y_k - z_k| \le M\left(\max_{k=0,1,\ldots,q-1} |y_k - z_k| + \sum_{k=q}^{n} |\delta_k|\right)$$

En la práctica, comprobar si un método multipaso es estable empleando esta definición parece que va a ser una auténtica pesadilla. Se va a tratar de obtener alguna condición necesaria para la estabilidad de un método multipaso. Para ello, se va a considerar un problema absolutamente trivial, como por ejemplo

$$\begin{cases} y'(t) = 0, \\ y(0) = 0, \end{cases}$$

cuya única solución en $\mathbb{R}$ es, sin ninguna duda, la función idénticamente nula. Sea

$$\sum_{j=0}^{q} \alpha_j y_{k+j} = h \sum_{j=0}^{q} \beta_j f_{k+j}, \quad k \in \{0,1,\dots,n-q\}$$

la expresión de un método multipaso cualquiera, y sean $\{y_k\}_{k=0}^n$, $\{z_k\}_{k=0}^n$, $\{\delta_k\}_{k=q}^n$ tales que, para cada $k \in \{0,1,\dots,n-q\}$,

$$\sum_{j=0}^{q} \alpha_j y_{k+j} = h \sum_{j=0}^{q} \beta_j f(t_{k+j}, y_{k+j}), \qquad \sum_{j=0}^{q} \alpha_j z_{k+j} = h \sum_{j=0}^{q} \beta_j f(t_{k+j}, z_{k+j}) + \delta_{k+q}$$

Como $f \equiv 0$, entonces

$$y_{k+q} = -\sum_{j=0}^{q-1} \alpha_j y_{k+j}, \qquad z_{k+q} = -\sum_{j=0}^{q-1} \alpha_j z_{k+j} + \delta_{k+q}$$

Supóngase que el método es estable, y elíjanse valores concretos para y_k, z_k y δ_k: si $k \in \{0,1,\dots,n\}$, escogemos $y_k = 0$, y si $k \in \{q, q+1, \dots, n\}$, tomamos $\delta_k = 0$. Asimismo, sean $\{z_k\}_{k=0}^n$ dados por

$$\begin{cases} z_0, z_1, \dots, z_{q-1} \in \mathbb{R}, \\ z_{k+q} = -\displaystyle\sum_{j=0}^{q-1} \alpha_j z_{k+j}, \quad k \in \{0,1,\dots,n-q\} \end{cases}$$

La condición de estabilidad para estos datos sería

$$\max_{k=q,\dots,n} |z_k| \leq M \left(\max_{k=0,1,\dots,q-1} |z_k| \right)$$

para cierta constante M positiva e independiente de h (es decir, independiente de n, y por tanto la desigualdad anterior debe tenerse para cualquier $n \in \mathbb{N}$). De esto se deduce que una condición necesaria para que el método sea estable es que la sucesión $\{z_k\}_{k=0}^\infty$ sea acotada.

Para estudiar sucesiones de este tipo, será conveniente recordar el resultado siguiente:

Proposición 6. *Sea $\{a_n\}_{n=0}^\infty$ una sucesión que satisface una relación de recurrencia lineal homogénea de orden $k \in \mathbb{N}$, es decir, existen $\lambda_1, \dots, \lambda_k \in \mathbb{R}$ tales que*

$$a_n = \lambda_1 a_{n-1} + \lambda_2 a_{n-2} + \dots + \lambda_k a_{n-k}$$

para todo $n \geq k$. Consideremos el polinomio característico de la ecuación,

$$p(X) = X^k - \lambda_1 X^{k-1} - \lambda_2 X^{k-2} - \dots - \lambda_{k-1} X - \lambda_k$$

Sean $\mu_1, \mu_2, \dots, \mu_s$ las raíces de este polinomio y sean $m_1, m_2, \dots, m_s$ sus respectivas multiplicidades. Entonces existen polinomios $p_1, p_2, \dots, p_s$ tales que $\deg(p_i(X)) < m_i$ para todo $i \in \{1,2,\dots,s\}$ y

$$a_n = p_1(n)\mu_1^n + p_2(n)\mu_2^n + \dots + p_s(n)\mu_s^n$$

para todo $n \in \mathbb{N} \cup \{0\}$.

Demostración. Corresponde a la asignatura *Matemática Discreta*. □

Ejemplo. Considérese el método de dos pasos siguiente:

$$y_{k+2} + 4y_{k+1} - 5y_k = h(4f_{k+1} + 2f_k),$$

y sea $\{z_k\}_{k=0}^{\infty}$ la sucesión dada por

$$\begin{cases} z_0, z_1 \in \mathbb{R}, \\ z_{k+2} + 4z_{k+1} - 5z_k = 0, \quad k \in \mathbb{N} \cup \{0\} \end{cases}$$

Se puede comprobar por inducción que esta sucesión es acotada, pero en su lugar, se va a tratar de dar una fórmula explícita para z_k. Considérese el polinomio

$$p(\lambda) = \lambda^2 + 4\lambda - 5,$$

y hállense sus raíces:

$$p(\lambda) = 0 \iff \lambda = \frac{-4 \pm \sqrt{16+20}}{2} \iff \lambda \in \{-5, 1\}$$

Por la proposición anterior, puede afirmarse que la sucesión $\{z_k\}_{k=0}^{\infty}$ es de la forma

$$z_k = c_1 \cdot 1^k + c_2 \cdot (-5)^k = c_1 + (-5)^k c_2, \quad k \in \mathbb{N} \cup \{0\},$$

donde $c_1, c_2 \in \mathbb{R}$. En particular, para $k = 0$ y $k = 1$,

$$\begin{cases} z_0 = c_1 + \ c_2 \\ z_1 = c_1 - 5c_2 \end{cases}$$

Resolviendo el sistema se obtiene

$$c_1 = \frac{5z_0 + z_1}{6}, \qquad c_2 = \frac{z_0 - z_1}{6}$$

En consecuencia,

$$z_k = \frac{5z_0 + z_1}{6} + (-5)^k \frac{z_0 - z_1}{6}$$

De esto puede deducirse fácilmente que la sucesión $\{z_k\}_{k=0}^{\infty}$ es acotada.

Ejemplo. Considérese el método multipaso de 2 pasos siguiente:

$$y_{k+2} - 2y_{k+1} + y_k = h(f_{k+1} - f_k),$$

y sea $\{z_k\}_{k=0}^{\infty}$ la sucesión dada por

$$\begin{cases} z_0, z_1 \in \mathbb{R}, \\ z_{k+2} - 2z_{k+1} + z_k = 0, \quad k \in \mathbb{N} \cup \{0\} \end{cases}$$

Estudiemos si esta sucesión es acotada. Considérese el polinomio

$$p(\lambda) = \lambda^2 - 2\lambda + 1,$$

y hállense sus raíces:

$$p(\lambda) = 0 \iff \lambda = \frac{2 \pm \sqrt{4-4}}{2} \iff \lambda = 1$$

Por tanto, la sucesión $\{z_k\}_{k=0}^{\infty}$ es de la forma

$$z_k = p(k), \quad k \in \mathbb{N} \cup \{0\},$$

donde $p(X) = c_1 + c_2 X$ es un polinomio de grado menor o igual que 1. Independientemente de $c_1, c_2 \in \mathbb{R}$, puede afirmarse que la sucesión $\{z_k\}_{k=0}^{\infty}$ no es acotada, luego el método no es estable.

Ahora que se le ha dado tanto bombo a la acotación de sucesiones dadas por relaciones de recurrencia, sería una auténtica lástima que esta condición necesaria para la estabilidad de un método multipaso no fuese suficiente.

Teorema 15. *Considérese un método multipaso lineal dado por*

$$y_{k+q} + \sum_{j=0}^{q-1} \alpha_j y_{k+j} = h \sum_{j=0}^{q} \beta_j f_{k+j}, \quad k \in \{0, 1, \ldots, n-q\},$$

y considérese el polinomio

$$p(\lambda) = \lambda^q + \alpha_{q-1}\lambda^{q-1} + \ldots + \alpha_0$$

Entonces el método es estable si y solo si todas las raíces de $p(\lambda)$ son de módulo menor o igual que 1 y las que tienen módulo 1 (en caso de que las haya) son simples.

Demostración. La omitimos. □

Definición 30. Dado un método multipaso lineal, el polinomio que figura en el teorema anterior se dice que es el **polinomio característico del método**.

Corolario 5. *Los métodos de Adams-Bashforth y Adams-Moulton son estables.*

Demostración. El polinomio característico de ambos métodos es $p(\lambda) = \lambda^q - \lambda^{q-1}$, que tiene como raíces a 1, con multiplicidad 1, y 0, con multiplicidad $q-1$. □

Corolario 6. *Un método BDF de q pasos es estable si y solo si $q \leq 6$.*

Demostración. Sería consecuencia inmediata del teorema anterior si se hubiesen estudiado con más profundidad los coeficientes de los métodos BDF. □

Teorema 16. *Un método multipaso es consistente si y solo si es de orden 1, es decir, si y solo si*

$$\sum_{j=0}^{q} \alpha_j j = \sum_{j=0}^{q} \beta_j, \qquad \sum_{i=1}^{q} \alpha_j = 0$$

Demostración. También se puede omitir. □

Corolario 7. *Un método de q pasos es convergente si y solo si es estable y consistente, o sea, si y solo si se satisface lo siguiente:*

(i) $\sum_{j=0}^{q} \alpha_j j = \sum_{j=0}^{q} \beta_j$.

(ii) $\sum_{i=1}^{q} \alpha_j = 0$.

(iii) *Todas las raíces del polinomio característico son de módulo menor o igual que 1 y las que tienen módulo 1 (en caso de que las haya) son simples.*

Demostración. Sería trivial si se probasen los dos últimos teoremas. □

Teorema 17. *Si $y \in C^{p+1}([t_0, t_0+T], \mathbb{R})$ y se tiene un método multipaso estable y de orden p, entonces*

$$e(h) = O(h^p)$$

Demostración. Se omite. □

Teorema 18 (Primera barrera de Dahlquist). *Un método estable de q pasos es de orden a lo sumo $q+1$ si q es impar, y de orden a lo sumo $q+2$ si q es par. Si además el método es explícito, entonces es de orden a lo sumo q.*

Demostración. Esta también. □

3.8. Métodos predictor-corrector

El objetivo de esta sección es desarollar un método numérico que combine métodos multipaso explícitos con métodos multipaso implícitos. Considérese de un método implícito de q pasos,

$$\sum_{j=0}^{q} \alpha_j y_{k+j} = h \sum_{j=0}^{q} \beta_j f_{k+j}, \quad k \in \{0,1,\dots,n-q\},$$

siendo $\beta_q \neq 0$ y, por comodidad, $\alpha_q = 1$. El método puede ser escrito como

$$y_{k+q} = h\beta_q f(t_{k+q}, y_{k+q}) + C_{k+q}, \tag{14}$$

siendo

$$C_{k+q} = -\sum_{j=0}^{q-1} \alpha_j y_{k+j} + h \sum_{j=0}^{q-1} \beta_j f_{k+j}$$

Supongamos conocidas las aproximaciones $y_0, y_1, \dots, y_{k+q-1}$. Como ya se razonó en secciones anteriores, la ecuación (14) puede verse como una ecuación de punto fijo, que tiene solución única si se toma h lo suficientemente pequeño; concretamente, si

$$h < \frac{1}{L|\beta_q|}$$

De esta manera, el valor y_{k+q} se puede aproximar mediante un método iterativo de punto fijo, esto es, mediante un método iterativo de la forma

$$\begin{cases} x_0 \in \mathbb{R}, \\ x_{n+1} = g(x_n), \quad n \in \mathbb{N} \cup \{0\} \end{cases}$$

En el caso que nos ocupa, lo más natural es tomar y_{k+q-1} como semilla y

$$g(y) = h\beta_q f(t_{k+q}, y) + C_{k+q}$$

como función de iteración, de forma que el método de punto fijo quedaría

$$\begin{cases} y_{k+q}^{(0)} = y_{k+q-1}, \\ y_{k+q}^{(l)} = h\beta_q f(t_{k+q}, y_{k+q}^{(l-1)}) + C_{k+q}, \quad l \in \mathbb{N} \end{cases}$$

Este es el procedimiento que se suele seguir en la práctica para hallar las aproximaciones $\{y_k\}_{k=0}^n$ de un método multipaso implícito. Si se quieren mejorar las aproximaciones de este método de punto fijo, lo primero que hay que optimizar es la elección de la semilla: en lugar de escoger y_{k+q-1}, se va a considerar un método explícito de q pasos cualquiera,

$$y_{k+q} + \sum_{j=0}^{q-1} \alpha_j^* y_{k+j} + h \sum_{j=0}^{q-1} \beta_j^* f_{k+j}, \quad k \in \{0,1,\dots,n-q\},$$

y se va a tomar como semilla la aproximación de y_{k+q} que daría este método explícito, es decir,

$$y_{k+q}^{(0)} = -\sum_{j=0}^{q-1} \alpha_j^* y_{k+j} + h \sum_{j=0}^{q-1} \beta_j^* f_{k+j}$$

Tenemos entonces el método de punto fijo siguiente:

$$\begin{cases} y_{k+q}^{(0)} = -\sum_{j=0}^{q-1} \alpha_j^* y_{k+j} + h \sum_{j=0}^{q-1} \beta_j^* f_{k+j}, \\ y_{k+q}^{(l)} = h\beta_q f(t_{k+q}, y_{k+q}^{(l-1)}) + C_{k+q}, \quad l \in \mathbb{N}, \end{cases}$$

y la aproximación y_{k+q} sería $y_{k+q}^{(m)}$ para algún $m \in \mathbb{N}$, que en la práctica se escoge atendiendo a algún criterio de parada establecido previamente. Dado $l \in \{1, \dots, m\}$, los valores

$$f_{k+q}^{(l)} = f(t_{k+q}, y_{k+q}^{(l)})$$

se conocen como *evaluaciones*, mientras que los números

$$y_{k+q}^{(l)} = h\beta_q f_{k+q}^{(l)} + C_{k+q}$$

se dice que son las *predicciones*. Se recoge el método numérico obtenido en la definición siguiente.

Definición 31. Dados $m, q \in \mathbb{N}$, Un **método predictor-corrector** es aquel definido mediante

$$\begin{cases} y_{k+q}^{(0)} = -\sum_{j=0}^{q-1} \alpha_j^* y_{k+j} + h \sum_{j=0}^{q-1} \beta_j^* f_{k+j}, \\ f_{k+q}^{(l)} = f(t_{k+q}, y_{k+q}^{(l)}), \qquad l \in \{0, 1, \dots, m-1\}, \\ y_{k+q}^{(l+1)} = h\beta_q f_{k+q}^{(l)} + C_{k+q}, \quad l \in \{0, 1, \dots, m-1\}, \\ y_{k+q} = y_{k+q}^{(m)} \end{cases}$$

para cada $k \in \{0, 1, \dots, n-q\}$.

Estos métodos habitualmente se denotan por P(EC)^{m}E, haciendo referencia al número de iteraciones de punto fijo que se realizan en cada paso. En particular, en el caso $m = 1$, se escribe simplemente PECE (*Predict–Evaluate–Correct–Evaluate*), y el método quedaría como sigue: si $k \in \{0, 1, \dots, n-q\}$,

$$\begin{cases} y_{k+q}^* = -\sum_{j=0}^{q-1} \alpha_j^* y_{k+j} + h \sum_{j=0}^{q-1} \beta_j^* f_{k+j}, \\ f_{k+q}^* = f(t_{k+q}, y_{k+q}^*), \\ y_{k+q} = h\beta_q f_{k+q}^* + C_{k+q} \end{cases}$$

TEMA 4

Estabilidad absoluta

4.1. Introducción

El objeto de estudio de este tema es el comportamiento en el infinito de las soluciones numéricas de un problema de Cauchy. El problema que protagoniza la definición siguiente actuará como sujeto de pruebas, y se utilizará constantemente a lo largo del tema.

Definición 32. Dados $y_0, \lambda \in \mathbb{C}$, un problema de Cauchy del tipo

$$(D)\ \begin{cases} y'(t) = \lambda y(t),\ t \in [0,\infty), \\ y(0) = y_0, \end{cases}$$

se conoce como **problema test de Dahlquist**.

La única solución del problema (D) es la función $y\colon [0,\infty) \to \mathbb{C}$ dada por $y(t) = y_0 e^{\lambda t}$, que verifica

$$\lim_{t\to\infty} y(t) = 0 \iff \mathrm{Re}(\lambda) < 0,$$

siendo esta convergencia más rápida cuanto mayor es $|\mathrm{Re}(\lambda)|$. Tratemos de aplicar el método de Euler a este problema: dado $k \in \mathbb{N}\cup\{0\}$,

$$y_{k+1} = y_k + h\lambda y_k = (1+h\lambda)y_k = (1+h\lambda)^2 y_{k-1} = \ldots = (1+h\lambda)^{k+1} y_0,$$

o sea,

$$y_k = (1+h\lambda)^k y_0$$

Cabría esperar que el comportamiento en el infinito de las aproximaciones del método de Euler imitase al de la solución exacta del problema. Nos centramos en el caso en que $\mathrm{Re}(\lambda) < 0$. Obsérvese que

$$\lim_{k\to\infty} y_k = 0 \iff |1+h\lambda| < 1 \iff h\lambda \in \Delta(-1,1) = \{z \in \mathbb{C}\colon |z+1| < 1\},$$

En particular, si $\lambda \in (-\infty, 0)$, entonces

$$\lim_{k\to\infty} y_k = 0 \iff |1+h\lambda| < 1 \iff -2 < h\lambda < 0 \iff h\lambda \in (-2,0)$$

Tratemos de generalizar todo esto a otro tipo de métodos numéricos:

Definición 33. Considérese un método unipaso que adopta una expresión del tipo

$$y_k = R(\hat{h})^k y_0, \quad k \in \mathbb{N},$$

cuando se aplica a un problema test de Dahlquist, siendo $\hat{h} = h\lambda$ y $h > 0$.

(i) La función $R\colon \mathbb{C}^- = \{\hat{h} \in \mathbb{C}\colon \mathrm{Re}(\hat{h}) < 0\} \to \mathbb{C}$ se conoce como **función de estabilidad absoluta**.

(ii) El conjunto

$$D_A = \{\hat{h} \in \mathbb{C}\colon |R(\hat{h})| < 1\}$$

se denomina **región de estabilidad absoluta** o **dominio de estabilidad absoluta**.

(iii) El intervalo

$$I_A = D_A \cap (-\infty, 0)$$

se conoce como **intervalo de estabilidad absoluta**.

(iv) Se dice que el método es ***A*-estable** si $\mathbb{C}^- \subset D_A$.

El interés de los métodos A-estables es que, para cualquier $\mathrm{Re}(\lambda) < 0$ y cualquier paso de malla $h > 0$, el comportamiento de las aproximaciones en el infinito es el mismo que el de la solución del problema (D).

Ejemplo. Para el método de Euler, según lo visto anteriormente, la función de estabilidad absoluta viene dada por

$$R(\hat{h}) = 1 + \hat{h},$$

mientras que el dominio de estabilidad absoluta es

$$D_A = \Delta(-1, 1),$$

y el intervalo de estabilidad absoluta,

$$I_A = D_A \cap (-\infty, 0) = (-2, 0)$$

Ejemplo. Repitamos este procedimiento con el método de Euler implícito, que para el problema de Dahlquist adopta la expresión

$$y_{k+1} = y_k + h\lambda y_{k+1}, \quad k \in \mathbb{N} \cup \{0\},$$

es decir,

$$y_{k+1} = \frac{1}{1 - h\lambda} y_k = \frac{1}{1 - \hat{h}} y_k = \frac{1}{(1 - \hat{h})^2} y_{k-1} = \ldots = \frac{1}{(1 - \hat{h})^{k+1}} y_0, \quad k \in \mathbb{N} \cup \{0\},$$

de donde

$$y_k = \frac{1}{(1 - \hat{h})^k} y_0, \quad k \in \mathbb{N}$$

Ahora se tendría

$$R(\hat{h}) = \frac{1}{1 - \hat{h}}, \qquad D_A = \mathbb{C} \setminus \overline{\Delta(1, 1)}, \qquad I_A = (-\infty, 0)$$

Ejemplo. Se trata de hallar la función de estabilidad absoluta del método de Heun:

$$\begin{cases} y^*_{k+1} = y_k + h\lambda y_k \\ y_{k+1} = y_k + \dfrac{h}{2}(\lambda y_k + \lambda y^*_{k+1}) \end{cases}$$

Equivalentemente,

$$y_{k+1} = y_k + \frac{h}{2}(2\lambda y_k + h\lambda^2 y_k) = \left(1 + h\lambda + \frac{(h\lambda)^2}{2}\right) y_k$$

En consecuencia,

$$R(\hat{h}) = 1 + \hat{h} + \frac{\hat{h}^2}{2}$$

es la función de estabilidad absoltua del método.

Proposición 7. *La función de estabilidad absoluta R de un método no contiene al eje real positivo en un entorno del origen. En otras palabras, existe $h^* > 0$ tal que $R(h) > 1$ para todo $h \in (0, h^*)$.*

Demostración. Hay que creérselo. □

Corolario 8. *La región de estabilidad absoluta de un método verifica*

$$D_A \subsetneq \mathbb{C}$$

Demostración. Inmediata a partir de la proposición anterior. □

El objetivo próximo consiste en generalizar todo esto al caso multidimensional. El problema test de Dahlquist para sistemas sería de la forma

$$(\widetilde{D})\ \begin{cases} Y'(t) = MY(t),\ t \in [0,\infty), \\ Y(0) = Y_0, \end{cases}$$

donde $Y_0 \in \mathbb{R}^n$ y $M \in \mathcal{M}_n(\mathbb{C})$ una matriz a la que más adelante se le pedirán ciertos requisitos para que la única solución del problema verifique

$$\lim_{t\to\infty} Y(t) = 0$$

Para poder resolver $(\widetilde{D})$ cómodamente, se supondrá que M es diagonalizable, o sea, existen matrices $P, \Lambda \in \mathcal{M}_n(\mathbb{C})$ con P inversible y Λ diagonal tales que

$$\Lambda = P^{-1}MP$$

Pongamos cara y ojos a estas matrices:

$$\Lambda = \begin{pmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_n \end{pmatrix} \qquad P = \left(\begin{array}{c|c|c|c} P_1 & P_2 & \dots & P_n \end{array} \right)$$

Obsérvese que, en estas circunstancias, para cada $i \in \{1,\dots,n\}$, λ_i es un autovalor de la matriz M con autovector asociado P_i. Tratemos de escribir el problema $(\widetilde{D})$ en términos de estas matrices: como P es inversible, entonces sus columnas constituyen una base de $\mathbb{R}^n$, y dado $Y \in \mathbb{R}^n$, su expresión en la base formada por las columnas de P sería $Z = P^{-1}Y$. De esta manera, el problema $(\widetilde{D})$ es equivalente a

$$(\overline{D})\ \begin{cases} Z'(t) = \Lambda Z(t),\ t \in [0,\infty), \\ Z(0) = Z_0, \end{cases}$$

donde

$$Z_0 = P^{-1}Y_0 = \begin{pmatrix} z_{0,1} \\ z_{0,2} \\ \vdots \\ z_{0,n} \end{pmatrix}$$

Vectorialmente,

$$(\overline{D})\ \begin{cases} z_1'(t) = \lambda_1 z_1(t), & z_1(0) = z_{0,1} \\ z_2'(t) = \lambda_2 z_2(t), & z_2(0) = z_{0,2} \\ \qquad \vdots & \\ z_n'(t) = \lambda_n z_n(t), & z_n(0) = z_{0,n} \end{cases}$$

La única solución de este problema es la función $Z\colon [0,\infty) \to \mathbb{R}^n$ dada por

$$Z(t) = \begin{pmatrix} z_{0,1}e^{\lambda_1 t} \\ z_{0,2}e^{\lambda_2 t} \\ \vdots \\ z_{0,2}e^{\lambda_n t} \end{pmatrix}$$

Por tanto, la única solución de $(\widetilde{D})$ es la función $Y:[0,\infty)\to\mathbb{R}^n$ dada por $Y(t)=PZ(t)$. Se tiene que

$$\lim_{t\to\infty} Y(t)=0 \iff \lim_{t\to\infty} Z(t)=0 \iff \mathrm{Re}(\lambda_i)<0 \text{ para todo } i\in\{1,\dots,n\}$$

En consecuencia, el caso que capta nuestro interés es aquel en que la matriz M posee autovalores de parte real negativa. Pasamos a estudiar el comportamiento en el infinito de las aproximaciones del método de Euler para el problema de Dahlquist multidimensional.

Ejemplo. Se trata de aplicar el método de Euler al problema $(\widetilde{D})$ y estudiar bajo qué condiciones puede asegurarse que

$$\lim_{k\to\infty} Y_k=0$$

Dado $k\in\mathbb{N}\cup\{0\}$, las aproximaciones del método son

$$Y_{k+1}=Y_k+hMY_k=(1+hM)Y_k,$$

o lo que es lo mismo, llamando $Z=P^{-1}Y$,

$$Z_{k+1}=P^{-1}(1+hM)PZ_k=(I+h\Lambda)Z_k=(I+h\Lambda)^2Z_{k-1}=\ldots=(I+h\Lambda)^{k+1}Z_0$$

Por tanto, el método de Euler se puede expresar como

$$Z_k=(I+h\Lambda)^kZ_0$$

Pero

$$Z_k=(I+h\Lambda)^kZ_0 \iff \begin{pmatrix} z_{k,1}\\ z_{k,2}\\ \vdots\\ z_{k,n}\end{pmatrix}=\begin{pmatrix} (1+h\lambda_1)^k & 0 & \dots & 0\\ 0 & (1+h\lambda_2)^k & \dots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \dots & (1+h\lambda_n)^k\end{pmatrix}\begin{pmatrix} z_{0,1}\\ z_{0,2}\\ \vdots\\ z_{0,n}\end{pmatrix}$$

$$\iff \begin{cases} z_{k,1}=(1+h\lambda_1)^kz_{0,1}\\ z_{k,2}=(1+h\lambda_2)^kz_{0,2}\\ \quad\vdots\\ z_{k,n}=(1+h\lambda_n)^kz_{0,n}\end{cases}$$

En consecuencia,

$$\lim_{k\to\infty} Y_k=0 \iff \lim_{k\to\infty} Z_k=0 \iff |1+h\lambda_i|<1\ \forall\ i\in\{1,\dots,n\} \iff h\lambda_i\in\Delta(-1,1)\ \forall\ i\in\{1,\dots,n\}$$

A continuación, fijado $\lambda=x+iy\in\mathbb{C}$ con $x<0$, trataremos de hallar los valores de $h^*>0$ tales que $|1+h^*\lambda|=1$. Se tiene que

$$|1+h^*\lambda|^2=1 \iff (1+h^*x)^2+h^{*2}y^2=1 \iff h^{*2}x^2+2h^*x+h^{*2}y^2=0 \iff h^*\left(h^*x^2+2x+h^*y^2\right)=0$$
$$\iff h^*|\lambda|^2+2\mathrm{Re}(\lambda)=0 \iff h^*=-\frac{2\mathrm{Re}(\lambda)}{|\lambda|^2}$$

De esto se deduce que

$$|1+h^*\lambda|<1 \iff |1+h^*\lambda|^2<1 \iff 0<h^*<-\frac{2\mathrm{Re}(\lambda)}{|\lambda|^2}$$

Concluimos que

$$\lim_{k\to\infty} Y_k=0 \iff 0<h^*<\min_{i=1,\dots,n}-\frac{2\mathrm{Re}(\lambda_i)}{|\lambda_i|^2}$$

4.2. Estabilidad absoluta de los métodos de Runge-Kutta

Como el propio título sugiere, en esta sección se estudiarán los acontecimientos resultantes de aplicar el método

$$\begin{cases} y_k^{(1)} = y_k + h(a_{1,1}f(t_k^{(1)}, y_k^{(1)}) + a_{1,2}f(t_k^{(2)}, y_k^{(2)}) + \ldots + a_{1,p}f(t_k^{(p)}, y_k^{(p)})) \\ y_k^{(2)} = y_k + h(a_{2,1}f(t_k^{(1)}, y_k^{(1)}) + a_{2,2}f(t_k^{(2)}, y_k^{(2)}) + \ldots + a_{2,p}f(t_k^{(p)}, y_k^{(p)})) \\ \quad \vdots \\ y_k^{(p)} = y_k + h(a_{p,1}f(t_k^{(1)}, y_k^{(1)}) + a_{p,2}f(t_k^{(2)}, y_k^{(2)}) + \ldots + a_{p,p}f(t_k^{(p)}, y_k^{(p)})) \\ y_{k+1} = y_k + h(b_1 f(t_k^{(1)}, y_k^{(1)}) + \ldots + b_p f(t_k^{(p)}, y_k^{(p)})) \end{cases}$$

al problema test de Dahlquist (caso unidimensional, afortunadamente). Poniendo $f(t,y) = \lambda y$, se obtiene

$$\begin{cases} y_k^{(1)} = y_k + \hat{h}(a_{1,1}y_k^{(1)} + a_{1,2}y_k^{(2)} + \ldots + a_{1,p}y_k^{(p)}) \\ y_k^{(2)} = y_k + \hat{h}(a_{2,1}y_k^{(1)} + a_{2,2}y_k^{(2)} + \ldots + a_{2,p}y_k^{(p)}) \\ \quad \vdots \\ y_k^{(p)} = y_k + \hat{h}(a_{p,1}y_k^{(1)} + a_{p,2}y_k^{(2)} + \ldots + a_{p,p}y_k^{(p)}) \\ y_{k+1} = y_k + \hat{h}(b_1 y_k^{(1)} + \ldots + b_p y_k^{(p)}) \end{cases}$$

Equivalentemente,

$$\begin{cases} y_k = (1 - \hat{h}a_{1,1})y_k^{(1)} - \hat{h}a_{1,2}y_k^{(2)} - \ldots - \hat{h}a_{1,p}y_k^{(p)} \\ y_k = -\hat{h}a_{2,1}y_k^{(1)} + (1 - \hat{h}a_{2,2})\hat{h}a_{2,2}y_k^{(2)} - \ldots - \hat{h}a_{2,p}y_k^{(p)} \\ \quad \vdots \\ y_k = \hat{h}a_{p,1}y_k^{(1)} - \hat{h}a_{p,2}y_k^{(2)} - \ldots + (1 - \hat{h}a_{p,p})y_k^{(p)} \\ y_{k+1} = y_k + \hat{h}(b_1 y_k^{(1)} + \ldots + b_p y_k^{(p)}) \end{cases}$$

También puede escribirse

$$\begin{cases} y_k E = (I - \hat{h}A)Y, \\ y_{k+1} = y_k + \hat{h}B^t Y \end{cases}$$

donde

$$A = \begin{pmatrix} a_{1,1} & a_{1,2} & \ldots & a_{1,p} \\ a_{2,1} & a_{2,2} & \ldots & a_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ a_{p,1} & a_{p,2} & \ldots & a_{p,p} \end{pmatrix} \qquad B = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_p \end{pmatrix} \qquad E = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \qquad Y = \begin{pmatrix} y_k^{(1)} \\ y_k^{(2)} \\ \vdots \\ y_k^{(p)} \end{pmatrix}$$

Por la Proposición 5, si se verificase $\rho(\hat{h}A) = h\rho(\lambda A) < 1$, o lo que es lo mismo,

$$h < \frac{1}{\rho(\lambda A)},$$

entonces $\rho(\hat{h}A) \leq ||\hat{h}A||$ para cualquier norma matricial subordinada $||\cdot||$ y la matriz $I - \hat{h}A$ tendría inversa. Así, tendríamos $Y = y_k(I - \hat{h}A)^{-1}E$ y por tanto

$$y_{k+1} = y_k + \hat{h}B^t Y = y_k + y_k \hat{h}B^t(I - \hat{h}A)^{-1}E = \left(1 + \hat{h}B^t(I - \hat{h}A)^{-1}E\right) y_k,$$

Concluimos que la función de estabilidad absoluta de un método de Runge-Kutta no es más que

$$R(\hat{h}) = 1 + \hat{h}B^t(I - \hat{h}A)^{-1}E$$

En la práctica, el cálculo de matrices inversas es indeseable, así que se tratará de dar una expresión alternativa para $R(\hat{h})$. Un método de Runge-Kutta se puede escribir como un sistema de $p+1$ ecuaciones con $p+1$ incógnitas:

$$\begin{pmatrix} 1-\hat{h}a_{1,1} & -\hat{h}a_{1,2} & \dots & -\hat{h}a_{1,p-1} & 0 \\ -\hat{h}a_{2,1} & 1-\hat{h}a_{2,2} & \dots & -\hat{h}a_{2,p-1} & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ -\hat{h}a_{p,1} & -\hat{h}a_{p,2} & \dots & 1-\hat{h}a_{p,p} & 0 \\ -\hat{h}b_1 & -\hat{h}b_2 & \dots & -\hat{h}b_p & 1 \end{pmatrix} \begin{pmatrix} y_k^{(1)} \\ y_k^{(2)} \\ \vdots \\ y_k^{(p)} \\ y_{k+1} \end{pmatrix} = \begin{pmatrix} y_k \\ y_k \\ \vdots \\ y_k \\ y_k \end{pmatrix}$$

En consecuencia, por la regla de Cramer,

$$y_{k+1} = \frac{\begin{vmatrix} 1-\hat{h}a_{1,1} & -\hat{h}a_{1,2} & \dots & -\hat{h}a_{1,p-1} & y_k \\ -\hat{h}a_{2,1} & 1-\hat{h}a_{2,2} & \dots & -\hat{h}a_{2,p-1} & y_k \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ -\hat{h}a_{p,1} & -\hat{h}a_{p,2} & \dots & 1-\hat{h}a_{p,p} & y_k \\ -\hat{h}b_1 & -\hat{h}b_2 & \dots & -\hat{h}b_p & y_k \end{vmatrix}}{\begin{vmatrix} 1-\hat{h}a_{1,1} & -\hat{h}a_{1,2} & \dots & -\hat{h}a_{1,p-1} & 0 \\ -\hat{h}a_{2,1} & 1-\hat{h}a_{2,2} & \dots & -\hat{h}a_{2,p-1} & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ -\hat{h}a_{p,1} & -\hat{h}a_{p,2} & \dots & 1-\hat{h}a_{p,p} & 0 \\ -\hat{h}b_1 & -\hat{h}b_2 & \dots & -\hat{h}b_p & 1 \end{vmatrix}}$$

Tras hacer un par de cálculos sencillos en el numerador se demuestra que

$$y_{k+1} = \frac{\left|I-\hat{h}A+\hat{h}EB^t\right|}{\left|I-\hat{h}A\right|} y_k,$$

luego

$$R(\hat{h}) = \frac{\left|I-\hat{h}A+\hat{h}EB^t\right|}{\left|I-\hat{h}A\right|}$$

es una expresión alternativa para la función de estabilidad absoluta de los métodos de Runge-Kutta. Se exponen a continuación otras propiedades sobre la estabilidad absoluta de los métodos de Runge-Kutta.

Proposición 8. *La función de estabilidad absoluta de un método de Runge-Kutta de p etapas es*

$$R(\hat{h}) = \frac{P(\hat{h})}{Q(\hat{h})},$$

donde P y Q son polinomios de grado menor o igual que p.

Demostración. Trivial a partir de la última expresión dada para $R(\hat{h})$. □

Proposición 9. *La función de estabilidad absoluta de un método de Runge-Kutta explícito de p etapas es un polinomio de grado menor o igual que p.*

Demostración. Solo hay que observar que en un método explícito, la matriz $I-\hat{h}A$ es triangular y con diagonal llena de unos. □

Proposición 10. *Un método de Runge-Kutta de p etapas no puede ser explícito y A-estable.*

Demostración. Si fuese $|R(\hat{h})| < 1$ para todo $\hat{h} \in \mathbb{C}^-$, entonces R tendría que ser un polinomio constante de módulo menor que 1. Esto es imposible porque si tomamos $\lambda = 0$, el método de Runge-Kutta aplicado al correspondiente problema de Dahlquist sería

$$\begin{cases} y_k^{(1)} = y_k \\ y_k^{(2)} = y_k \\ \vdots \\ y_k^{(p)} = y_k \\ y_{k+1} = y_k \end{cases}$$

Como la expresión del método es $y_{k+1} = y_k$, entonces $R(h \cdot 0) = R(0) = 1$, luego R no puede ser una constante de módulo menor que 1, y por tanto ningún método explícito es A-estable. □

Proposición 11. *Si un método de Runge-Kutta de p etapas es de orden s, entonces*

$$R(h) = e^h + O(h^s)$$

Demostración. El problema de Dahlquist para $\lambda = 1$ tiene como solución a $y(t) = y_0 e^t$. Además, por el Teorema 7, existen $C, h^* > 0$ tales que para todo $h \in (0, h^*)$ es

$$|y(t_1) - y_1| \leq e(h) \leq Ch^s$$

Pero

$$|y(t_1) - y_1| = |y(h) - R(h)y_0| = |y_0||e^h - R(h)|,$$

y por tanto $R(h) - e^h = O(h^s)$. □

Corolario 9. *Si un método de Runge-Kutta de p etapas es de orden s, entonces*

$$R(h) = 1 + h + \frac{h^2}{2} + \ldots + \frac{h^s}{s!} + O(h^{s+1})$$

Demostración. Consecuencia directa de la proposición anterior. □

Proposición 12. *Si un método de Runge-Kutta explícito de p etapas es de orden p, entonces*

$$R(\hat{h}) = 1 + \hat{h} + \frac{\hat{h}^2}{2} + \ldots + \frac{\hat{h}^p}{p!}$$

Demostración. Por una proposición anterior,

$$R(\hat{h}) = a_0 + a_1\hat{h} + \ldots + a_p\hat{h}^p$$

En particular, si $h > 0$,

$$R(h) = a_0 + a_1 h + \ldots + a_p h^p$$

Por otra proposición anterior,

$$R(h) = 1 + h + \frac{h^2}{2} + \ldots + \frac{1}{p!}h^p + O(h^{p+1})$$

Por tanto,

$$a_0 - 1 + h(a_1 - 1) + h^2\left(a_2 - \frac{1}{2}\right) + \ldots + h^p\left(a_p - \frac{1}{p!}\right) = O(h^{p+1})$$

Esto significa que existen $C, h^* > 0$ tales que para todo $h \in (0, h^*)$ es

$$\left| a_0 - 1 + h(a_1 - 1) + h^2\left(a_2 - \frac{1}{2}\right) + \ldots + h^p\left(a_p - \frac{1}{p!}\right)\right| \leq Ch^{p+1},$$

es decir,

$$\left| \frac{a_0 - 1}{h^{p+1}} + \frac{a_1 - 1}{h^p} + \frac{a_2 - \frac{1}{2}}{h^{p-1}} + \ldots + \frac{a_p - \frac{1}{p!}}{h} \right| \leq C,$$

De ser alguno de los $a_i - \frac{1}{i!}$ no nulo, el cociente $\frac{a_i - \frac{1}{i!}}{h^{p+1-i}}$ tendría límite ∞ cuando $h \to 0^+$ y por tanto la suma anterior no puede estar acotada por C. La única posibilidad es que para todo $i \in \{1, \ldots, p\}$ se verifique

$$a_i - \frac{1}{i!} = 0,$$

o sea,

$$a_i = \frac{1}{i!}$$

de donde se deduce inmediatamente la igualdad del enunciado. □

Ejemplo. Se trata de dar la función de estabilidad absoluta del método del punto medio. Ya se sabe que este método es un método de Runge-Kutta, y su tablero de Butcher es

$$\begin{array}{c|cc} 0 & 0 & 0 \\ 1/2 & 1/2 & 0 \\ \hline & 0 & 1 \end{array}$$

Se tiene que

$$I - \hat{h}A + \hat{h}EB^t = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} - \hat{h}\begin{pmatrix} 0 & 0 \\ 1/2 & 0 \end{pmatrix} + \hat{h}\begin{pmatrix} 1 \\ 1 \end{pmatrix}\begin{pmatrix} 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ -\hat{h}/2 & 1 \end{pmatrix} + \begin{pmatrix} 0 & \hat{h} \\ 0 & \hat{h} \end{pmatrix} = \begin{pmatrix} 1 & \hat{h} \\ -\hat{h}/2 & 1+\hat{h} \end{pmatrix}$$

Por tanto,

$$R(\hat{h}) = \frac{\begin{vmatrix} 1 & \hat{h} \\ -\hat{h}/2 & 1+\hat{h} \end{vmatrix}}{\begin{vmatrix} 1 & 0 \\ -\hat{h}/2 & 1 \end{vmatrix}} = 1 + \hat{h} + \frac{\hat{h}^2}{2}$$

Ejemplo. Veamos que el método del trapecio es A-estable. Su tablero de Butcher es

$$\begin{array}{c|cc} 0 & 0 & 0 \\ 1 & 1/2 & 1/2 \\ \hline & 1/2 & 1/2 \end{array}$$

Se tiene que

$$I - \hat{h}A + \hat{h}EB^t = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} - \hat{h}\begin{pmatrix} 0 & 0 \\ 1/2 & 1/2 \end{pmatrix} + \hat{h}\begin{pmatrix} 1 \\ 1 \end{pmatrix}\begin{pmatrix} 1/2 & 1/2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ -\hat{h}/2 & 1-\hat{h}/2 \end{pmatrix} + \begin{pmatrix} \hat{h}/2 & \hat{h}/2 \\ \hat{h}/2 & \hat{h}/2 \end{pmatrix}$$

Por tanto,

$$R(\hat{h}) = \frac{\begin{vmatrix} 1+\hat{h}/2 & \hat{h}/2 \\ 0 & 1 \end{vmatrix}}{\begin{vmatrix} 1 & 0 \\ -\hat{h}/2 & 1-\hat{h}/2 \end{vmatrix}} = \frac{1+\hat{h}/2}{1-\hat{h}/2} = \frac{2+\hat{h}}{2-\hat{h}}$$

Además,

$$\left|\frac{2+\hat{h}}{2-\hat{h}}\right|^2 < 1 \iff |2+\hat{h}|^2 < |2-\hat{h}|^2 \iff (2+x)^2 + y^2 < (2-x)^2 + y^2 \iff x < 0,$$

donde $\hat{h} = x + iy$. Concluimos que

$$D_A = \{z \in \mathbb{C} : \mathrm{Re}(z) < 0\} = \mathbb{C}^-,$$

luego el método es A-estable.

4.3. Estabilidad absoluta de los métodos multipaso

El panorama es el siguiente: se trata de aplicar un método de q pasos dado por

$$\sum_{j=0}^{q} \alpha_j y_{k+j} = h \sum_{j=0}^{q} \beta_j f_{k+j}, \quad k \in \mathbb{N} \cup \{0\},$$

al problema test de Dahlquist, y estudiar bajo qué condiciones se tiene

$$\lim_{k \to \infty} y_k = 0$$

Nótese que este límite es independiente de las primeras q aproximaciones, $y_0, \ldots, y_{q-1}$. Al sustituir en el método la función $f(t, y) = \lambda y$, se obtiene

$$\sum_{j=0}^{q} \alpha_j y_{k+j} = \hat{h} \sum_{j=0}^{q} \beta_j y_{k+j}, \quad k \in \mathbb{N} \cup \{0\},$$

es decir,

$$\sum_{j=0}^{q} \left(\alpha_j - \hat{h}\beta_j\right) y_{k+j} = 0, \quad k \in \mathbb{N} \cup \{0\}$$

Obsérvese que la sucesión $\{y_k\}_{k=0}^{\infty}$ está definida por recurrencia como sigue:

$$\begin{cases} y_0, y_1, \ldots, y_{q-1} \in \mathbb{R}, \\ \displaystyle\sum_{j=0}^{q} \left(\alpha_j - \hat{h}\beta_j\right) y_{k+j} = 0, \quad k \in \mathbb{N} \cup \{0\}, \end{cases}$$

Por la Proposición 6, puede escribirse

$$y_k = p_1(k)\mu_1^n + p_2(k)\mu_2^k + \ldots + p_s(k)\mu_s^k, \quad k \in \mathbb{N} \cup \{0\},$$

donde $\mu_1, \mu_2, \ldots, \mu_s$ son las raíces del polinomio

$$p(z) = \sum_{j=0}^{q} \left(\alpha_j - \hat{h}\beta_j\right) z^j,$$

de multiplicidades $m_1, m_2, \ldots, m_s$, y p_i es un polinomio de grado menor o igual que $m_i - 1$ para cada $i \in \{1, \ldots, s\}$. Puede probarse que

$$\lim_{k \to \infty} y_k = 0 \iff |\mu_i| < 1 \ \forall \ i \in \{1, \ldots, s\}$$

Definición 34. Considérese un método de q pasos dado por

$$\sum_{j=0}^{q} \alpha_j y_{k+j} = h \sum_{j=0}^{q} \beta_j f_{k+j}, \quad k \in \mathbb{N} \cup \{0\}$$

(i) Se define el **primer polinomio característico del método** como

$$\rho(z) = \sum_{i=0}^{q} \alpha_i z^i$$

(ii) Se define el **segundo polinomio característico del método** como

$$\sigma(z) = \sum_{i=0}^{q} \beta_i z^i$$

(iii) Fijado $\hat{h} \in \mathbb{C}$, se define el **polinomio de estabilidad asboluta del método** como

$$\pi_{\hat{h}}(z) = \rho(z) - \hat{h}\sigma(z) = \sum_{j=0}^{q} \left(\alpha_j - \hat{h}\beta_j\right) z^j$$

(iv) La **región de estabilidad absoluta del método** o **dominio de estabilidad absoluta del método** no es más que

$$D_A = \left\{\hat{h} \in \mathbb{C} \colon \pi_{\hat{h}} \text{ tiene todas sus raíces de módulo menor que } 1\right\}$$

(v) Se conoce como **intervalo de estabilidad absoluta del método** a

$$I_A = D_A \cap (-\infty, 0)$$

(vi) Se dice que el método es ***A*-estable** si $\mathbb{C}^- \subset D_A$.

Ejemplo. Considérese el método del trapecio:

$$y_{k+1} = y_k + \frac{h}{2}(f_k + f_{k+1}), \quad k \in \mathbb{N}$$

Este método puede ser visto como uno de la familia RK (método unipaso) o uno de la familia AM (método multipaso con $q = 1$). Estaría bien que el dominio de estabilidad absoluta según la definición para métodos unipaso fuese el mismo que el que proporciona la definición para métodos multipaso. En el último ejemplo ya se vio que

$$R(\hat{h}) = \frac{2 + \hat{h}}{2 - \hat{h}}$$

Por otra parte,

$$\rho(z) = z - 1, \qquad \sigma(z) = \frac{1}{2}(z + 1)$$

Por tanto,

$$\pi_{\hat{h}}(z) = z - 1 - \frac{\hat{h}}{2}(z + 1) = \left(1 - \frac{\hat{h}}{2}\right) z - 1 - \frac{\hat{h}}{2}$$

Se tiene que

$$\pi_{\hat{h}}(z) = 0 \iff z = \frac{1 + \frac{\hat{h}}{2}}{1 - \frac{\hat{h}}{2}} = \frac{2 + \hat{h}}{2 - \hat{h}} = R(\hat{h}),$$

luego el dominio de estabilidad absoluta es el mismo.

Teorema 19 (Segunda barrera de Dahlquist). *Ningún métodos multipaso es explícito y A-estable. Es más, un método multipaso y A-estable es a lo sumo de orden 2.*

Demostración. Lo mismo que la otra barrera: se queda sin demostrar. □

4.4. Método de localización de la frontera

En la práctica, como haya que enfrentarse a un método multipaso de expresión enrevesada, calcular la región de estabilidad del método empleando la definición no va a ser para nada trivial. Es por ello que conviene desarrollar un procedimiento alternativo para el cálculo de D_A.

Definición 35. Dado un método multipaso con región de estabilidad absoluta D_A, se define la **frontera de D_A** como

$$\partial D_A := \{\hat{h} \in \mathbb{C} : \pi_{\hat{h}} \text{ tiene al menos una raíz de módulo 1}\}$$

Lo primero que debe observarse es que esta frontera es distinta de la frontera de toda la vida, la frontera topológica. Por otra parte, nótese que $\hat{h} \in \partial D_A$ si y solo si existe $\theta \in \mathbb{R}$ tal que $e^{i\theta} = \cos\theta + i\,\mathrm{sen}\,\theta$ es raíz de $\pi_{\hat{h}}$. Pero

$$\pi_{\hat{h}}(e^{i\theta}) = 0 \iff \rho(e^{i\theta}) - \hat{h}\sigma(e^{i\theta}) = 0 \iff \hat{h} = \frac{\rho(e^{i\theta})}{\sigma(e^{i\theta})},$$

siempre que $\sigma(e^{i\theta}) \neq 0$. Así, D_A puede verse como una curva en $\mathbb{C}$ parametrizada por la función

$$\theta \mapsto \frac{\rho(e^{i\theta})}{\sigma(e^{i\theta})}, \quad \theta \in \mathbb{R},$$

y a partir de aquí se puede representar gráficamente ∂D_A de forma más o menos fácil. Si se tuviese $\sigma(e^{i\theta^*}) = 0$ para un cierto $\theta^* \in \mathbb{R}$, se distinguen dos casos:

- ***(i)*** Si $\rho(e^{i\theta^*}) = 0$, entonces $\pi_{\hat{h}}(e^{i\theta^*}) = 0$ para todo $\hat{h} \in \mathbb{C}$ y por tanto $D_A = \emptyset$.
- ***(ii)*** Si $\rho(e^{i\theta^*}) \neq 0$, entonces $\pi_{\hat{h}}(e^{i\theta^*}) \neq 0$ para todo $\hat{h} \in \mathbb{C}$ y no se puede decir mucho más.

Con esta información y las propiedades de la proposición que sigue, se puede obtener información acerca de D_A.

Proposición 13. *Sea D_A la región de estabilidad absoluta de un método multipaso. Se verifican las siguientes propiedades:*

- ***(i)*** *D_A es conexo (pudiendo ser vacío).*
- ***(ii)*** *D_A no contiene al eje real positivo en un entorno del origen.*
- ***(iii)*** *$D_A \subsetneq \mathbb{C}$.*

Demostración. Nos lo creemos. □

Ejemplo. Se trata de aplicar el método de localización de la frontera al método del trapecio. Se tiene

$$\rho(z) = z - 1, \qquad \sigma(z) = \frac{1}{2}(z+1),$$

luego

$$\frac{\rho(e^{i\theta})}{\sigma(e^{i\theta})} = \frac{2\left(e^{i\theta}-1\right)}{e^{i\theta}+1} = \frac{2\left(1+e^{i\theta}-e^{-i\theta}-1\right)}{1+e^{i\theta}+e^{-i\theta}+1} = \frac{2\,\mathrm{sen}\,\theta}{1+\cos\theta}\,i$$

La función $\theta \mapsto \frac{2\,\mathrm{sen}\,\theta}{1+\cos\theta}$, $\theta \in (-\pi, \pi)$ está bien definida y puede comprobarse que su imagen es todo $\mathbb{R}$, y de aquí se deduce que ∂D_A es el eje imaginario:

$$\partial D_A = \{\hat{h} \in \mathbb{C} : \mathrm{Re}(\hat{h}) = 0\}$$

Hay dos posibilidades: $D_A = \mathbb{C}^-$, $D_A = \mathbb{C}^+$ y $D_A = \emptyset$. La segunda no puede darse porque D_A contendría al eje real positivo; la tercera, tampoco, pues se comprueba fácilmente que $-1 \in D_A$. La conclusión es que $D_A = \mathbb{C}^-$.

Ejemplo. Se trata de aplicar el método de localización de la frontera al método

$$y_{k+2} = y_k + \frac{h}{3}(f_{k+2} + 4f_{k+1} + f_k)$$

Se tiene

$$\rho(z) = z^2 - 1, \qquad \sigma(z) = \frac{1}{3}\left(z^2 + 4z + 1\right),$$

Además, se comprueba fácilmente que

$$\frac{\rho(e^{i\theta})}{\sigma(e^{i\theta})} = \frac{3\,\mathrm{sen}\,\theta}{2 + \cos\theta} i$$

La función $\theta \mapsto \frac{2\,\mathrm{sen}\,\theta}{1+\cos\theta}$, $\theta \in \mathbb{R}$ está bien definida y puede comprobarse que su imagen es $[-\sqrt{3}, \sqrt{3}]$, y de aquí puede deducirse que

$$\partial D_A = \left\{\hat{h} \in \mathbb{C} : \mathrm{Re}(\hat{h}) = 0, \mathrm{Im}(\hat{h}) \in [-\sqrt{3}, \sqrt{3}]\right\}$$

Hay dos posibilidades: $D_A = \mathbb{C}$ y $D_A = \emptyset$. La primera no puede darse porque D_A no puede contener al eje real positivo, concluyéndose que $D_A = \emptyset$.

TEMA 5

Problemas de contorno

5.1. Introducción

Hasta ahora, todos los esfuerzos se han centrado en el estudio de problemas del estilo

$$\begin{cases} y'(t) = f(t, y(t)), \ t \in [t_0, t_0 + T], \\ y(t_0) = y^0 \end{cases}$$

Los problemas que van a protagonizar este tema son ligeramente distintos:

Definición 36. Un **problema de contorno** es un problema de la forma

$$(P) \begin{cases} y''(x) = f(x, y(x), y'(x)), \ x \in [a, b], \\ y(a) = \alpha, \\ y(b) = \beta, \end{cases}$$

donde $\alpha, \beta \in \mathbb{R}$ y $f : [a, b] \times \mathbb{R} \times \mathbb{R} \to \mathbb{R}$.

Para obtener aproximaciones numéricas de soluciones de problemas de contorno interesa trabajar bajo condiciones que aseguren que el problema tiene solución y es única. Se reúnen dichas condiciones en el teorema que sigue.

Teorema 20. *Sea $f : [a, b] \times \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ una función verificando*

(i) *f, $\frac{\partial f}{\partial y}$, $\frac{\partial f}{\partial z}$ son continuas.*

(ii) *$\frac{\partial f}{\partial y}(t, y, z) > 0$ para todo $(t, y, z) \in [a, b] \times \mathbb{R} \times \mathbb{R}$.*

(iii) *Existen $m, M, L > 0$ tales que para todo $(t, y, z) \in [a, b] \times \mathbb{R} \times \mathbb{R}$ se tiene*

$$m \leq \left| \frac{\partial f}{\partial y}(t, y, z) \right| \leq M, \qquad \left| \frac{\partial f}{\partial y}(t, y, z) \right| \leq L$$

Entonces el problema (P) tiene solución única.

Demostración. No corresponde a esta asignatura. □

5.2. Método del tiro

Una primera estrategia para resolver un problema del tipo

$$(P) \begin{cases} y''(x) = f(x, y(x), y'(x)), \ x \in [a, b], \\ y(a) = \alpha, \\ y(b) = \beta, \end{cases}$$

consiste en trabajar con un problema que resulte más familiar, por ejemplo

$$(Q_v)\begin{cases} y''(x) = f(x, y(x), y'(x)),\ x \in [a,b], \\ y(a) = \alpha, \\ y'(a) = v, \end{cases}$$

para cualquier $v \in \mathbb{R}$. Si denotamos y_v a la solución del problema, el problema se reduce a encontrar $v \in \mathbb{R}$ tal que $y_v(b) = \beta$, lo que proporcionará una solución (la única) del problema (P). Para poder emplear los métodos estudiados, hay que traducir el problema (Q_v) al caso conocido:

$$(\widetilde{Q}_v)\begin{cases} y'(x) = z, \\ z'(x) = f(x, y(x), z(x)), \\ y(a) = \alpha, \\ z(a) = v \end{cases}$$

Esta estrategia de resolución de problemas de contorno se conoce como *método del tiro*, y no se va a profundizar en ella.

Antes de desarrollar otro tipo de métodos que sí sean de nuestro interés, conviene recordar las nociones básicas de derivación numérica.

5.3. Derivación numérica

El problema es el siguiente: considérese una partición uniforme de $\mathbb{R}$ dada por $x_i = x_0 + ih$, $i \in \mathbb{Z}$, y supóngase que se conocen los valores de una función u lo suficientemente regular en los puntos de la partición. El objetivo es aproximar $u^{(k)}(c)$ para $k \in \mathbb{N}$, $c \in \mathbb{R}$ cualesquiera. Si $k = 1$, la aproximación más natural de $u'(c)$ nace de la propia definición de derivada, que sugiere lo siguiente: si $c \in (x_i, x_{i+1})$ para cierto $i \in \mathbb{Z}$,

$$u'(c) \approx \frac{u(x_{i+1}) - u(x_i)}{h}$$

Si fuese $c = x_i$, se pueden realizar las aproximaciones

$$u'(x_i) \approx \frac{u(x_{i+1}) - u(x_i)}{h}, \qquad u'(x_i) \approx \frac{u(x_i) - u(x_{i-1})}{h} \qquad \text{o} \qquad u'(x_i) \approx \frac{u(x_{i+1}) - u(x_{i-1})}{2h}$$

A partir de aquí, también se puede intentar aproximar $u''(x_i)$ como sigue:

$$u''(x_i) \approx \frac{u'(x_i + \frac{h}{2}) - u'(x_i - \frac{h}{2})}{h} \approx \frac{\frac{u(x_{i+1}) - u(x_i)}{h} - \frac{u(x_i) - u(x_{i-1})}{h}}{h} = \frac{u(x_{i+1}) - 2u(x_i) + u(x_{i-1})}{h^2}$$

Las fórmulas empleadas van a ser como las que se recogen en la definición siguiente.

Definición 37. Una **fórmula de derivación numérica** es una expresión del tipo

$$u^{(k)}(c) \approx \mathcal{D}^k_{n+1}(u) = \frac{1}{h^k} \sum_{i=l}^{r} \alpha_i u(x_i),$$

donde $\alpha_l, \ldots, \alpha_r \in \mathbb{R}$ y $l, r \in \mathbb{N}$ son tales que $r - l = n$. Si $c = \frac{x_l + x_r}{2}$, se dirá que la fórmula es **centrada**.

Definición 38. Una fórmula de derivación numérica $\mathcal{D}^k_{n+1}(u)$ es dice que es **de orden p** si para toda función $u \in \mathcal{C}^{k+p}([\alpha, \beta])$ se tiene que

$$u^{(k)}(c) = \mathcal{D}^k_{n+1}(u) + O(h^p),$$

donde $[\alpha, \beta]$ es un intervalo que contiene al intervalo $[x_l, x_r]$.

Ejemplo. Estudiemos el error de la fórmula

$$\mathcal{D}_2^1(u) = \frac{u(x_{i+1}) - u(x_i)}{h}$$

Por la fórmula del resto de Lagrange, existe $\xi \in (x_i, x_{i+1})$ tal que

$$u(x_{i+1}) = u(x_i) + u'(x_i)h + \frac{u''(\xi)}{2}h^2,$$

luego

$$u'(x_i) \approx \mathcal{D}_2^1(u) = \frac{\cancel{u(x_i)} + u'(x_i)h + \frac{u''(\xi)}{2}h^2 - \cancel{u(x_i)}}{h} = u'(x_i) + \frac{u''(\xi)}{2}h$$

Por tanto,

$$|\mathcal{D}_2^1(u) - u'(x_i)| \leq Mh,$$

donde

$$M = \max_{x\in[a,b]} |u''(x)|$$

y se está suponiendo que u es de clase 2 en un intervalo $[\alpha, \beta]$ que contiene a $[x_i, x_{i+1}]$. La fórmula $\mathcal{D}_1^2(u)$ es de primer orden.

Ejemplo. Considérese ahora la fórmula

$$\mathcal{D}_2^1(u) = \frac{u(x_{i+1}) - u(x_{i-1})}{2h}$$

Por la fórmula del resto de Lagrange, existen $\xi_1 \in (x_i, x_{i+1})$, $\xi_2 \in (x_{i-1}, x_i)$ tales que

$$u(x_{i+1}) = u(x_i) + u'(x_i)h + \frac{u''(x_i)}{2}h^2 + \frac{u'''(\xi_1)}{6}h^3, \qquad u(x_{i+1}) = u(x_i) - u'(x_i)h + \frac{u''(x_i)}{2}h^2 - \frac{u'''(\xi_2)}{6}h^3,$$

luego

$$u'(x_i) \approx \mathcal{D}_2^1(u) = \frac{2u'(x_i)h + \frac{u'''(\xi_1)}{6}h^3 - \frac{u'''(\xi_2)}{6}h^3}{2h} = u'(x_i) + \left(u'''(\xi_1) - u'''(\xi_2)\right)\frac{h^2}{12}$$

Por tanto,

$$|\mathcal{D}_2^1(u) - u'(x_i)| \leq M\frac{h^2}{6},$$

donde

$$M = \max_{x\in[a,b]} |u'''(x)|$$

y se está suponiendo que u es de clase 3 en un intervalo $[\alpha, \beta]$ que contiene a $[x_{i-1}, x_{i+1}]$. Tenemos entonces que la fórmula $\mathcal{D}_1^2(u)$ es de segundo orden.

En general, se pueden seguir ciertos procedimientos para encontrar fórmulas del mayor orden posible. Los más comunes son los expuestos en las subsecciones siguientes.

5.3.1. Método de Taylor

Para ahorrarnos una explicación teórica tediosa e insulsa, se expondrá el método en cuestión a través de un ejemplo.

Ejemplo. Consideramos una fórmula del tipo

$$\mathcal{D}_3^1(u) = \frac{\alpha_0 u(x_i) + \alpha_1 u(x_{i+1}) + \alpha_2 u(x_{i+2})}{h}$$

Se trata de escoger $\alpha_0, \alpha_1, \alpha_2 \in \mathbb{R}$ adecuadamente para que el orden sea lo más grande posible. Por un

lado,

$$\alpha_1 u(x_{i+1}) = \alpha_1 u(x_i) + \alpha_1 h u'(x_i) + \alpha_1 \frac{h^2}{2} u''(x_i) + \alpha_1 \frac{h^3}{6} u'''(x_i) + \dots$$

Por otro,

$$\alpha_2 u(x_{i+2}) = \alpha_2 u(x_i) + 2\alpha_2 h u'(x_i) + 4\alpha_2 \frac{h^2}{2} u''(x_i) + 8\alpha_2 \frac{h^3}{6} u'''(x_i) + \dots$$

Por tanto,

$$\mathcal{D}_3^1(u) = \frac{1}{h}\left((\alpha_0 + \alpha_1 + \alpha_2)u(x_i) + (\alpha_1 + 2\alpha_2)h u'(x_i) + (\alpha_1 + 4\alpha_2)\frac{h^2}{2} u''(x_i) + \dots\right)$$

Para que el método tenga orden 2, interesa que se verifiquen las igualdades

$$\begin{cases} \alpha_0 + \alpha_1 + \alpha_2 = 0 \\ \alpha_1 + 2\alpha_2 = 1 \\ \alpha_1 + 4\alpha_2 = 0 \end{cases}$$

La solución del sistema es

$$\alpha_0 = -\frac{3}{2}, \qquad \alpha_1 = 2, \qquad \alpha_2 = -\frac{1}{2}$$

De esta manera, tenemos asegurado que la fórmula

$$\mathcal{D}_3^1(u) = \frac{-3u(x_i) + 4u(x_{i+1}) - u(x_{i+2})}{2h}$$

es de orden 2. Para comprobar si es de orden 3, se añade un término más:

$$\mathcal{D}_3^1(u) = \frac{1}{h}\left(h u'(x_i) + (\alpha_1 + 8\alpha_2)\frac{h^3}{6} u'''(x_i) + \dots\right)$$

Como $\alpha_1 + 8\alpha_2 = -2 \neq 0$, la fórmula no es de orden 3.

Proposición 14. *El orden máximo de una fórmula de derivación numérica $\mathcal{D}_{n+1}^k(u)$ es*

(i) *$n+1-k$ si la fórmula es centrada.*

(ii) *$n+1-k+1 = n+2-k$ si la fórmula es centrada o si $n \equiv k$ mód 2.*

Demostración. Nos lo creemos y seguimos adelante. □

Ejemplo. Los órdenes máximos de las fórmulas

$$u'(x_i) \approx \frac{u(x_{i+1}) - u(x_i)}{h}, \qquad u'(x_i) \approx \frac{u(x_{i+1}) - u(x_{i-1})}{2h} \qquad \text{y} \qquad u''(x_i) \approx \frac{u(x_{i+1}) - 2u(x_i) + u(x_{i-1})}{h^2}$$

son 2, 3 y 2, respectivamente.

5.3.2. Método de interpolación

Una forma natural de encontrar fórmulas del tipo

$$\mathcal{D}_{n+1}^k(u) = \frac{1}{h^k} \sum_{i=l}^{r} \alpha_i u(x_i),$$

consiste en tomar el polinomio P que interpola los puntos

$$(x_l, u(x_l)), \dots, (x_r, u(x_r))$$

y aproximar $u^{(k)}(c) \approx P^{(k)}$. Se sabe que el polinomio P adopta una expresión de la forma

$$P(x) = \sum_{i=l}^{r} u(x_i) l_i(x),$$

donde, para $i \in \{l, l+1, \ldots, r\}$,

$$l_i(x) = \frac{(x-x_l)\ldots\widehat{(x-x_i)}\ldots(x-x_r)}{(x_i-x_l)\ldots\widehat{(x_i-x_i)}\ldots(x_i-x_r)}$$

Para aprovechar el hecho de que se está trabajando con particiones uniformes, dado $x \in \mathbb{R}$, se va a denotar $t = \frac{x-x_0}{h}$ y entonces

$$l_i(x) = \frac{(t-l)h\ldots\widehat{(t-i)h}\ldots(t-r)h}{(i-l)h\ldots\widehat{(i-i)h}\ldots(i-r)h} = \frac{(t-l)\ldots\widehat{(t-i)}\ldots(t-r)}{(i-l)\ldots\widehat{(i-i)}\ldots(i-r)}$$

Así,

$$l_i(x) = \widetilde{l}_i\left(\frac{x-x_0}{h}\right),$$

donde, para $t \in \mathbb{R}$,

$$\widetilde{l}_i(t) = \frac{(t-l)\ldots\widehat{(t-i)}\ldots(t-r)}{(i-l)\ldots\widehat{(i-i)}\ldots(i-r)}$$

Por la regla de la cadena,

$$l_i^{(k)}(c) = \frac{1}{h^k}\widetilde{l}_i^{(k)}\left(\frac{c-x_0}{h}\right) = \frac{1}{h^k}\widetilde{l}_i^{(k)}(t_c),$$

donde $t_c = \frac{c-x_0}{h}$. Así,

$$u^{(k)}(c) \approx P^{(k)}(c) = \sum_{i=l}^{r} u(x_i) l_i^{(k)}(c) = \frac{1}{h^k}\sum_{i=l}^{r} u(x_i)\widetilde{l}_i^{(k)}(t_c)$$

Obtenemos así la fórmula de derivación numérica

$$\mathcal{D}_{n+1}^{k}(u) = \frac{1}{h^k}\sum_{i=l}^{r} \alpha_i u(x_i),$$

donde $\alpha_i = \widetilde{l}_i^{(k)}(t_c)$ para cada $i \in \{l, l+1, \ldots, r\}$. Alternativamente, si a_k es el coeficiente de grado k del polinomio P, se puede usar que

$$P^{(k)}(c) = k!a_k = k!u[x_l, \ldots, x_r]$$

Ejemplo. Tratemos de encontrar la fórmula proporcionada por el método anterior para $k = 2$ y $n = 3$, o sea, queremos aproximar $u''(c)$ usando $u(x_{i-1})$, $u(x_i)$ y $u(x_{i+1})$. Hay que hallar la diferencia dividida $u[x_{i-1}, x_i, x_{i+1}]$.

Puntos	Orden 0	Orden 1	Orden 2
x_{i-1}	$u(x_{i-1})$		
		$\frac{u(x_i)-u(x_{i-1})}{h}$	
x_i	$u(x_i)$		$\frac{u(x_{i+1})-2u(x_i)+u(x_{i-1})}{2h^2}$
		$\frac{u(x_{i+1})-u(x_i)}{h}$	
x_{i+1}	$u(x_{i+1})$		

Si se quisiera aproximar $u'''(c)$ usando $u(x_{i-1})$, $u(x_i)$, $u(x_{i+1})$ y $u(x_{i+2})$, habría que hallar la diferencia dividida $u[x_{i-1}, x_i, x_{i+1}, x_{i+2}]$.

Puntos	Orden 0	Orden 1	Orden 2	Orden 3
x_{i-1}	$u(x_{i-1})$	$\frac{u(x_i)-u(x_{i-1})}{h}$		
x_i	$u(x_i)$		$\frac{u(x_{i+1})-2u(x_i)+u(x_{i-1})}{2h^2}$	$\frac{u(x_{i+2})-3u(x_{i+1})+3u(x_i)-u(x_{i-1})}{6h^3}$
x_{i+1}	$u(x_{i+1})$	$\frac{u(x_{i+1})-u(x_i)}{h}$	$\frac{u(x_{i+2})-2u(x_{i+1})+u(x_i)}{2h^2}$	
x_{i+2}	$u(x_{i+2})$	$\frac{u(x_{i+2})-u(x_{i+1})}{h}$		

Por tanto,

$$u'''(c) \approx \frac{u(x_{i+2}) - 3u(x_{i+1}) + 3u(x_i) - u(x_{i-1})}{h^3}$$

Puede comprobarse que la fórmula es de orden 1 si $c \neq \frac{x_i + x_{i+2}}{2}$ y de orden 2 en otro caso.

5.4. Método de diferencias finitas para el caso lineal

Se trata de aproximar la solución (en caso de haberla) de un problema de contorno lineal:

$$(L) \begin{cases} y''(x) = p(x)y' + q(x)y + r(x), \ x \in [a,b], \\ y(a) = \alpha, \\ y(b) = \beta, \end{cases}$$

donde $p, q, r \colon [a,b] \to \mathbb{R}$, $\alpha, \beta \in \mathbb{R}$. Si p, q y r son continuas y $q(x) > 0$ para todo $x \in [a,b]$, se dan las hipótesis del Teorema 20 y el problema (L) tiene solución única. Aproximemos entonces dicha solución. Se considera una partición uniforme $x_0 = a < x_1 < \ldots < x_n < x_{n+1} = b$, de forma que $x_i = x_0 + ih$ para cada $i \in \{0, 1, \ldots, n+1\}$, con

$$h = \frac{b-a}{n+1}$$

Para cada $i \in \{1, 2, \ldots, n\}$, se realizan las aproximaciones

$$y''(x_i) \approx \frac{y(x_{i+i}) - 2y(x_i) + y(x_{i-1})}{h^2}$$

Por otro lado,

$$y'(x_i) \approx \frac{y(x_{i+1}) - y(x_{i-1})}{2h},$$

luego

$$y''(x_i) = p(x_i)y'(x_i) + q(x_i)y(x_i) + r(x_i) \approx p(x_i)\frac{y(x_{i+1}) - y(x_{i-1})}{2h} + q(x_i)y(x_i) + r(x_i)$$

y el problema se reduce a encontrar $u_0, u_1, \ldots, u_{n+1}$ tales que para cada $i \in \{1, 2, \ldots, n\}$ se verifique

$$\frac{u_{i+i} - 2u_i + u_{i-1}}{h^2} = p(x_i)\frac{u_{i+1} - u_{i-1}}{2h} + q(x_i)u_i + r(x_i),$$

o lo que es lo mismo,

$$\left(\frac{1}{h^2} + \frac{p(x_i)}{2h}\right)u_{i-1} - \left(\frac{2}{h^2} + q(x_i)\right)u_i + \left(\frac{1}{h^2} - \frac{p(x_i)}{2h}\right)u_{i+1} = r(x_i) \tag{15}$$

Multiplicando por $-\frac{h^2}{2}$,

$$-\frac{1}{2}\left(1 + \frac{p(x_i)h}{2}\right)u_{i-1} + \left(1 + \frac{q(x_i)h^2}{2}\right)u_i - \frac{1}{2}\left(1 - \frac{p(x_i)h}{2}\right)u_{i+1} = -\frac{h^2}{2}r(x_i),$$

es decir,

$$-b_i u_{i-1} + a_i u_i - c_i u_{i+1} = -\frac{h^2}{2} r(x_i), \tag{16}$$

donde

$$a_i = 1 + \frac{q(x_i)h^2}{2}, \qquad b_i = \frac{1}{2}\left(1 + \frac{p(x_i)h}{2}\right), \qquad c_i = \frac{1}{2}\left(1 - \frac{p(x_i)h}{2}\right)$$

Así, $u_1, u_2, \ldots, u_n$ satisfacen un sistema lineal de n ecuaciones y n incógnitas. Tratemos de poner cara y ojos a este sistema. Para $i = 1$, la ecuación (16) no es más que

$$-b_1 u_0 + a_1 u_1 - c_1 u_2 = -\frac{h^2}{2} r(x_1)$$

Escogiendo $u_0 = \alpha$,

$$a_1 u_1 - c_1 u_2 = -\frac{h^2}{2} r(x_1) + b_1 \alpha$$

Para $i = n$, la ecuación (16) sería

$$-b_n u_{n-1} + a_n u_n - c_n u_{n+1} = -\frac{h^2}{2} r(x_i)$$

Tomando $u_{n+1} = \beta$,

$$-b_n u_{n-1} + a_n u_n = -\frac{h^2}{2} r(x_i) + c_n \beta$$

En resumen, el sistema a resolver sería

$$(S) \begin{cases} a_1 u_1 - c_1 u_2 = -\dfrac{h^2}{2} r(x_1) + b_1 \alpha, \\ -b_i u_{i-1} + a_i u_i - c_i u_{i+1} = -\dfrac{h^2}{2} r(x_i), & i \in \{2, 3, \ldots, n-1\}, \\ -b_n u_{n-1} + a_n u_n = -\dfrac{h^2}{2} r(x_n) + c_n \beta, \end{cases}$$

o lo que es lo mismo, en forma matricial,

$$AU = F,$$

donde

$$A = \begin{pmatrix} a_1 & -c_1 & 0 & \ldots & 0 & 0 & 0 \\ -b_2 & a_2 & -c_2 & \ldots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & -b_{n-1} & a_{n-1} & -c_{n-1} \\ 0 & 0 & 0 & \ldots & 0 & -b_n & a_n \end{pmatrix} \qquad U = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_{n-1} \\ u_n \end{pmatrix} \qquad F = \begin{pmatrix} -\frac{h^2}{2} r(x_1) + b_1 \alpha \\ -\frac{h^2}{2} r(x_2) \\ \vdots \\ -\frac{h^2}{2} r(x_{n-1}) \\ -\frac{h^2}{2} r(x_n) + c_n \beta \end{pmatrix}$$

De la ecuación (15) y de las condiciones iniciales se obtiene inmediatamente una expresión equivalente para el sistema (S) que resulta un poco más agradable a la vista:

Definición 39. Un método numérico para la resolución del problema (L) del tipo

$$(MDF) \begin{cases} u_0 = \alpha \\ \dfrac{u_{i-1} - 2u_i + u_{i+1}}{h^2} = p(x_i)\dfrac{u_{i+1} - u_{i-1}}{2h} + q(x_i)u_i + r(x_i), & i \in \{1, \ldots, n\} \\ u_{n+1} = \beta \end{cases}$$

se conoce como **método de diferencias finitas**.

Es conveniente averiguar cuándo el sistema anterior posee una única solución (o sea, $\det(A) \neq 0$).

Definición 40. Una matriz $M = (m_{i,j})_{i,j=1,\dots,n}$ es **de diagonal estrictamente dominante por filas** si para todo $i \in \{1,2,\dots,n\}$ se verifica

$$|m_{i,i}| > \sum_{j \neq i} |m_{i,j}|$$

Proposición 15. *Una matriz de diagonal estrictamente dominante por filas tiene determinante no nulo.*

Demostración. Corresponde a la asignatura *Métodos Numéricos II*. □

Corolario 10. *Si $p \equiv 0$, el sistema $AU = F$ tiene solución única.*

Demostración. En efecto, si $p \equiv 0$, entonces $b_i = c_i = \frac{1}{2}$ para todo $i \in \{1,2,\dots,n\}$, y usando que $q(x) > 0$ para todo $x \in [a,b]$,

$$|a_i| = \left|1 + \frac{q(x_i)h^2}{2}\right| = 1 + \frac{q(x_i)h^2}{2} > 1 = |-b_i| + |-c_i|,$$

luego A es de diagonal estrictamente dominante por filas y por tanto $\det(A) \neq 0$, concluyéndose que el sistema $AU = F$ posee solución única. □

Corolario 11. *Si*

$$h||p||_\infty = h \max_{x \in [a,b]} |p(x)| < 2,$$

entonces el sistema $AU = F$ tiene solución única.

Demostración. Por un lado,

$$b_i = \frac{1}{2}\left(1 + \frac{p(x_i)}{2}h\right) \geq \frac{1}{2}\left(1 - \frac{|p(x_i)|}{2}h\right) \geq \frac{1}{2}\left(1 - \frac{||p||_\infty}{2}h\right) > \frac{1}{2}\left(1 - \frac{2}{2}\right) = 0$$

Por otro lado,

$$c_i = \frac{1}{2}\left(1 - \frac{p(x_i)}{2}h\right) \geq \frac{1}{2}\left(1 - \frac{|p(x_i)|}{2}h\right) \geq \frac{1}{2}\left(1 - \frac{||p||_\infty}{2}h\right) > \frac{1}{2}\left(1 - \frac{2}{2}\right) = 0$$

En consecuencia, como $q(x_i) > 0$,

$$|-b_i| + |-c_i| = b_i + c_i = 1 < 1 + \frac{q(x_i)h^2}{2} = a_i = |a_i|$$

y por tanto A es de diagonal estrictamente dominante por filas, así que $\det(A) \neq 0$ y el sistema $AU = F$ posee solución única. □

Ahora que disponemos de un método numérico para aproximar el problema (L), el próximo paso es definir el error, el orden, la estabilidad, la consistencia... y ese tipo de cosas.

Definición 41. Sean $\{u_0, u_1, \dots, u_{n+1}\}$ las aproximaciones de un método de diferencias finitas. Sean

$$Y = \begin{pmatrix} y(x_0) \\ y(x_1) \\ \vdots \\ y(x_{n+1}) \end{pmatrix} \qquad R = \begin{pmatrix} r(x_1) \\ r(x_2) \\ \vdots \\ r(x_n) \end{pmatrix}$$

y sea $\mathcal{L}_h : \mathbb{R}^{n+2} \to \mathbb{R}^n$ la función definida por

$$(\mathcal{L}_h Z)_i = \frac{z_{i-1} - 2z_i + z_{i+1}}{h^2} - p(x_i)\frac{z_{i+1} - z_{i-1}}{2h} - q(x_i)z_i, \quad i \in \{1,2,\dots,n\},$$

donde $(\mathcal{L}_h Z)_i$ es la i-ésima componente de $\mathcal{L}_h Z \in \mathbb{R}^n$ para cualquier $Z \in \mathbb{R}^{n+2}$.

(i) Dado $i \in \{0,1,\ldots,n+1\}$, se define el **error en la etapa *i*-ésima** como

$$e_i = y(x_i) - u_i$$

(ii) Se define el **error global** como

$$e(h) := \max_{i=0,\ldots,n+1} |e_i|$$

(iii) Se dice que el método es **convergente** si

$$\lim_{h\to 0} e(h) = 0$$

(iv) Se define el **error de discretización local** como

$$\tau(h) := \mathcal{L}_h Y - R,$$

(v) Se dice que el método es **consistente** si

$$\lim_{h\to 0} ||\tau(h)||_\infty = 0$$

(vi) Se dice que el método es **estable** si existen $M, h^* > 0$ verificando lo siguiente: si $Z \in \mathbb{R}^{n+2}$, $\delta \in \mathbb{R}^n$ son tales que

$$\mathcal{L}_h Z = \delta,$$

entonces

$$\max_{i=0,\ldots,n+1} |z_i| \leq M\left(\max\{|z_0|, |z_{n+1}|\} + ||\delta||_\infty\right)$$

(vii) Se dice que el método es **de orden *p*** si

$$||\tau(h)||_\infty = O(h^p)$$

Teorema 21. *Consistencia y estabilidad implica convergencia.*

Demostración. La expresión del método de diferencias finitas quiere decir que $\mathcal{L}_h U = R$. Además, por definición, $\mathcal{L}_h Y = R + \tau(h)$. Si llamamos

$$E = Y - U = \begin{pmatrix} e_0 \\ e_1 \\ \vdots \\ e_{n+1} \end{pmatrix}$$

entonces $\mathcal{L}_h E = R + \tau(h) - R = \tau(h)$. Tomando $Z = E \in \mathbb{R}^{n+2}$ y $\delta = \tau(h) \in \mathbb{R}^n$ en la definición de estabilidad, se tiene

$$\max_{i=0,\ldots,n+1} |e_i| \leq M\left(\max\{|e_0|, |e_{n+1}|\} + ||\tau(h)||_\infty\right)$$

Ahora bien, como $e_0 = y(x_0) - u_0 = \alpha - \alpha = 0$ y $e_{n+1} = y(x_{n+1}) - u_{n+1} = \beta - \beta = 0$, entonces

$$0 \leq \max_{i=0,\ldots,n+1} |e_i| \leq M||\tau(h)||_\infty$$

Como el método es consistente, al tomar límite en estas desigualdades se obtiene

$$\lim_{h\to 0} e(h) = 0,$$

concluyéndose que el método converge. □

Teorema 22. *Si un método es estable y de orden p, entonces*

$$e(h) = O(h^p)$$

Demostración. Ya va tocando saltarse una demostración. □

5.5. Otras condiciones de contorno

Considérese el problema

$$(\widetilde{L}) \begin{cases} y''(x) = p(x)y' + q(x)y + r(x), \ x \in [a,b], \\ y'(a) = \alpha, \\ y(b) = \beta, \end{cases}$$

que es idéntico al problema (L) salvo por una de las condiciones iniciales. Sean $\{u_0, u_1, \ldots, u_{n+1}\}$ las aproximaciones de un método de diferencias finitas para el problema (L). Casi todas las aproximaciones siguen siendo válidas para el problema $(\widetilde{L})$; la única que no sirve es u_0, pues ya no se pide $y(a) = \alpha$. Para cada $i \in \{1, 2, \ldots, n\}$ se verifica

$$\frac{u_{i-1} - 2u_i + u_{i+1}}{h^2} = p(x_i)\frac{u_{i+1} - u_{i-1}}{2h} + q(x_i)u_i + r(x_i)$$

Poniendo $i = 1$,

$$\frac{u_0 - 2u_1 + u_2}{h^2} = p(x_1)\frac{u_2 - u_0}{2h} + q(x_1)u_1 + r(x_1),$$

Resulta natural realizar la aproximación

$$y'(a) = y'(x_0) \approx \frac{y(x_1) - y(x_0)}{h} \approx \frac{u_1 - u_0}{h},$$

así que puede considerarse el método dado por

$$(M_1) \begin{cases} \dfrac{u_1 - u_0}{h} = \alpha \\ \dfrac{u_{i-1} - 2u_i + u_{i+1}}{h^2} = p(x_i)\dfrac{u_{i+1} - u_{i-1}}{2h} + q(x_i)u_i + r(x_i), & i \in \{1, \ldots, n\} \\ u_{n+1} = \beta \end{cases}$$

Si se quieren obtener aproximaciones más precisas, basta aproximar $y'(x_0)$ con mejores fórmulas de derivación numérica. Rescatamos la fórmula de un ejemplo anterior:

$$y'(x_0) \approx \frac{-3y(x_i) + 4y(x_{i+1}) - y(x_{i+2})}{2h}$$

El método obtenido para la resolución de $(\widetilde{L})$ sería

$$(M_2) \begin{cases} \dfrac{-3u_0 + 4u_1 - u_2}{2h} = \alpha \\ \dfrac{u_{i-1} - 2u_i + u_{i+1}}{h^2} = p(x_i)\dfrac{u_{i+1} - u_{i-1}}{2h} + q(x_i)u_i + r(x_i), & i \in \{1, \ldots, n\} \\ u_{n+1} = \beta \end{cases}$$

Otra manera de aproximar $y'(x_0)$ consiste en añadir un punto más a la izquierda de x_0, llámese x_{-1}, y hacer

$$y'(x_0) \approx \frac{y(x_1) - y(x_{-1})}{2h},$$

lo que daría lugar al método

$$(M_3)\begin{cases} \dfrac{u_1-u_{-1}}{2h}=\alpha \\ \dfrac{u_{i-1}-2u_i+u_{i+1}}{h^2}=p(x_i)\dfrac{u_{i+1}-u_{i-1}}{2h}+q(x_i)u_i+r(x_i), & i\in\{0,\dots,n\} \\ u_{n+1}=\beta \end{cases}$$

A priori, no se sabe quién es u_{-1}, pero de la primera ecuación se deduce $u_{-1}=u_1-2\alpha h$, así que la segunda ecuación del método para $i=0$ quedaría

$$\frac{u_1-2\alpha h-2u_0+u_1}{h^2}=p(x_0)\frac{u_1-u_1-2\alpha h}{2h}+q(x_0)u_0+r(x_0),$$

es decir,

$$\frac{2(u_1-u_0)}{h^2}=p(x_0)\alpha+q(x_0)u_0+r(x_0)+\frac{2\alpha}{h}$$

Ya nos hemos desentendido de u_{-1}: el método que queda es

$$(M_3)\begin{cases} \dfrac{2(u_1-u_0)}{h^2}=p(x_0)\alpha+q(x_0)u_0+r(x_0)+\dfrac{2\alpha}{h} \\ \dfrac{u_{i-1}-2u_i+u_{i+1}}{h^2}=p(x_i)\dfrac{u_{i+1}-u_{i-1}}{2h}+q(x_i)u_i+r(x_i), & i\in\{1,\dots,n\} \\ u_{n+1}=\beta \end{cases}$$

5.6. Método de diferencias finitas para el caso general

Lo que se ha hecho hasta ahora es tratar de aproximar la solución del problema

$$(L)\begin{cases} y''(x)=p(x)y'+q(x)y+r(x),\ x\in[a,b], \\ y(a)=\alpha, \\ y(b)=\beta, \end{cases}$$

mediante un método de la forma

$$(MDF)\begin{cases} u_0=\alpha \\ \dfrac{u_{i-1}-2u_i+u_{i+1}}{h^2}=p(x_i)\dfrac{u_{i+1}-u_{i-1}}{2h}+q(x_i)u_i+r(x_i), & i\in\{1,\dots,n\} \\ u_{n+1}=\beta \end{cases}$$

Si consideramos un problema de contorno arbitrario, no necesariamente lineal,

$$(P)\begin{cases} y''(x)=f(x,y(x),y'(x)),\ x\in[a,b], \\ y(a)=\alpha, \\ y(b)=\beta, \end{cases}$$

entonces, por analogía al caso lineal, es tentador considerar un método del estilo

$$(MDF)\begin{cases} u_0=\alpha \\ \dfrac{u_{i-1}-2u_i+u_{i+1}}{h^2}=f\left(x_i,u_i,\dfrac{u_{i+1}-u_{i-1}}{2h}\right), & i\in\{1,\dots,n\} \\ u_{n+1}=\beta \end{cases}$$

Para hallar $\{u_1,u_2,\dots,u_n\}$ hay que resolver el sistema no lineal de n ecuaciones y n incógnitas dado por

$$G(U)=0,$$

donde $G\colon \mathbb{R}^n \to \mathbb{R}^n$ es la función definida por

$$G(Z)_i = \frac{z_{i-1} - 2z_i + z_{i+1}}{h^2} - f\left(x_i, z_i, \frac{z_{i+1} - z_{i-1}}{2h}\right), \qquad i \in \{1, \ldots, n\}$$

En analogía al método de Newton (método de punto fijo dado por $z_{n+1} = z_n - \frac{g(u_n)}{g'(u_n)}$ para resolver la ecuación $g(u) = 0$, con $g\colon \mathbb{R} \to \mathbb{R}$), pueden aproximarse las soluciones de $G(U) = 0$ mediante la sucesión

$$U_{n+1} = U_n - J(U_n)^{-1} G(U_n), \qquad n \in \mathbb{N} \cup \{0\},$$

donde, para $U \in \mathbb{R}^n$ cualquiera,

$$J(U) = \begin{pmatrix} \frac{\partial G_1}{\partial u_1} & \cdots & \frac{\partial G_1}{\partial u_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial G_n}{\partial u_1} & \cdots & \frac{\partial G_n}{\partial u_n} \end{pmatrix}$$

Generalmente, el cálculo de matrices inversas debe ser evitado a toda costa. Se tiene que

$$U_{n+1} = U_n - J(U_n)^{-1} G(U_n) \iff J(U_n)(U_n - U_{n+1}) = G(U_n) \iff \begin{cases} J(U_n) V_n = G(U_n) \\ U_{n+1} = U_n - V_n \end{cases}$$

Encontrar $V_n \in \mathbb{R}^n$ tal que $J(U_n)V_n = G(U_n)$ no es más que resolver un sistema de ecuaciones con matriz de coeficientes $J(U_n)$ y vector de términos independientes $G(U_n)$, que se calculan fácilmente. Nótese que, por cómo está definida la función G, la i-ésima fila de la matriz $J(U)$ solo tiene tres términos no nulos: el de la diagonal y los dos colindantes.