Apuntes de Métodos Numéricos para EDO

Fernando Quirós Gracián

Versión del 8 de abril de 2013

Índice

Capítulo 1

El problema de valor inicial

1.1. Problemas de valor inicial y problemas de contorno

Al modelizar problemas de la ciencia, la ingeniería y la economía aparecen con frecuencia ecuaciones diferenciales ordinarias[1]. Una ecuación diferencial ordinaria (en adelante una EDO) es una relación entre una variable independiente y una función de dicha variable y sus derivadas. Nosotros nos centraremos en ecuaciones de *primer orden* (la derivada de mayor orden que aparece es la de orden uno) escritas en la *forma estándar*,

$$y'(x) = f(x, y(x)), \qquad a \leq x \leq b, \tag{1.1}$$

donde $f : [a, b] \times \mathbb{R}^d \mapsto \mathbb{R}^d$ es continua. Una solución de (1.1) en $[a, b]$ es una función $y : [a, b] \mapsto \mathbb{R}^d$, $y \in C^1([a, b])$ que satisface (1.1). En el caso vectorial[2],

[1] Las ecuaciones diferenciales son tan antiguas como el cálculo diferencial. Aparecen por primera vez en el tratado de Isaac Newton (1642–1727) "Methodus Fluxionum et Serierum Infinitorum", publicado por primera vez en 1671.

[2] Se entiende que los vectores se derivan componente a componente.

$d > 1$, que se puede interpretar como un sistema de ecuaciones escalares, y y f tienen d componentes cada una,

$$y = (y^1, y^2, \ldots, y^d)^T, \qquad f = (f^1, f^2, \ldots, f^d)^T.$$

Una EDO en general no define por sí sola una solución única, y se hace necesario añadir a la formulación del problema un cierto número de condiciones adicionales. Éstas son o bien "condiciones de frontera", si la información adicional se da en dos o más valores de x, o "condiciones iniciales", si se especifican todas las condiciones sobre y en un único valor de x. En los próximos capítulos nos centraremos en el caso en que se dan condiciones iniciales. Así pues, dado $\eta = (\eta^1, \eta^2, \ldots, \eta^d)^T \in \mathbb{R}^d$, buscamos una solución del *problema de valor inicial* en $[a, b]$ con dato inicial η, esto es, una función $y \in C^1([a, b])$ que satisfaga

$$y'(x) = f(x, y(x)) \qquad \text{para } a \leq x \leq b, \qquad y(a) = \eta. \tag{PVI}$$

Problemas

1. La restricción a ecuaciones de primer orden no supone pérdida de generalidad, pues siempre es posible transformar un problema de mayor orden en uno de primer orden a costa de aumentar la dimensión del sistema. Para ello simplemente hay que añadir como nuevas variables dependientes a cada una de las derivadas de las variables dependientes originales de orden estrictamente menor que las de mayor orden que aparecieran en las ecuaciones de partida.

   (a) Escribir la ecuación lineal $y^{(n)} + a_{n-1}y^{(n-1)} + \cdots + a_1 y' + a_0 y = 0$ en la forma $Y' = AY$ con $Y = (y, y', \ldots, y^{(n-1)})^T$ y A una matriz $n \times n$.

(b) Reducir el problema de valor inicial

$$u''' = u'' + v', \qquad v'' = u^2 + e^x u' v' + \operatorname{sen} v,$$
$$u(0) = u'(0) = u''(0) = v'(0) = 0, \qquad v(0) = 1,$$

a un problema de valor inicial para una EDO de primer orden escrita en la forma estándar.

2. Una EDO de primer orden escrita en la forma estándar se dice *autónoma* si la función f del lado derecho no depende explícitamente de la variable independiente x. Cualquier EDO se puede transformar en autónoma introduciendo una nueva variable dependiente dada precisamente por x.

   Transformar el problema de valor inicial

$$\begin{aligned} u' &= xu + x^2 v, & u(0) &= 3, \\ v' &= u - v + 2xw, & v(0) &= 2, \\ w' &= u + \frac{v}{1+x}, & w(0) &= 5, \end{aligned}$$

   en un problema de valor inicial para una EDO autónoma.

3. Al restringirnos a ecuaciones que se puedan escribir en forma estándar, sí dejamos fuera algunos problemas: la EDO de primer orden más general, $F(x, y(x), y'(x)) = 0$, donde y, y' y F tienen d componentes cada una, no siempre equivale a una única EDO escrita en forma estándar.

   Encontrar todas las soluciones de la EDO de primer orden en forma no estándar $(y')^2 - (2x + y)y' + 2xy = 0$.

1.2. Unicidad

Como se puede ver en el siguiente ejemplo[3], algunos problemas de valor inicial tienen más de una solución.

[3]Debido a Giuseppe Peano (1858–1932), matemático italiano conocido sobre todo por sus contribuciones a la teoría de conjuntos y por haber fundado la lógica simbólica.

Ejemplo. Consideramos el problema de valor inicial $y' = |y|^\alpha$ en $[a, b]$, $y(a) = 0$, siendo α un número real fijo, $\alpha \in (0, 1)$. Es facil comprobar que para cualquier número real no negativo $c \in [a, b]$,

$$y_c(x) = \begin{cases} 0, & a \le x \le c, \\ (1-\alpha)^{1/(1-\alpha)}(x-c)^{1/(1-\alpha)}, & c \le x \le b, \end{cases}$$

es una solución del problema. Así pues, si bien el problema tiene solución, ésta no es única. Sin embargo, en contraste con el caso $\alpha \in (0, 1)$, si $\alpha \ge 1$ el problema de valor inicial tiene una única solución, $y(x) \equiv 0$. ♣

Este ejemplo muestra que hay que pedir a la función f algo más que continuidad para asegurar que la solución del problema (PVI) sea única. Nosotros pediremos una condición de crecimiento con respecto al segundo argumento de la función.

Definición 1.1. *La función $f : D \subset \mathbb{R} \times \mathbb{R}^d \mapsto \mathbb{R}^d$ satisface una* condición de Lipschitz[4] en D con respecto a su segunda variable *si existe una constante L, conocida como* constante de Lipschitz*, tal que*

$$\|f(x, y) - f(x, \hat{y})\| \le L\|y - \hat{y}\| \qquad \forall (x, y), (x, \hat{y}) \in D.$$

Observación. Dado que en un espacio vectorial de dimensión finita todas las normas son equivalentes, la propiedad de ser Lipschitz no depende de qué norma tomemos. ♠

Si f, además de ser continua en $D = [a, b] \times \mathbb{R}^d$, satisface una condición de Lipschitz con respecto a su segunda variable, la solución, de existir, será única. Esto es una consecuencia inmediata del siguiente lema.

Lema 1.2. *Sea $D = [a, b] \times \mathbb{R}^d$ y sea f continua en D y Lipschitz con respecto a su segunda variable en D con constante L. Sean $y, \hat{y}$ dos soluciones de la*

[4]Rudolf Otto Sigismund Lipschitz (1832–1903), matemático alemán. Trabajó en una amplia gama de áreas: teoría de números, álgebras con involución, análisis matemático, geometría diferencial y mecánica clásica.

ecuación (1.1) *en* $[a,b]$. *Entonces, para todo* $a \le x \le b$,

$$\|y(x) - \hat{y}(x)\| \le \|y(a) - \hat{y}(a)\| \exp(L(x-a)). \tag{1.2}$$

Demostración. Tenemos que[5]

$$y(x) = y(a) + \int_a^x f(s, y(s))\, ds,$$
$$\hat{y}(x) = \hat{y}(a) + \int_a^x f(s, \hat{y}(s))\, ds.$$

Restando y tomando normas, y usando que f es Lipschitz en su segunda variable, se tiene que

$$\begin{aligned} \|y(x) - \hat{y}(x)\| &\le \|y(a) - \hat{y}(a)\| + \int_a^x \|f(s, y(s)) - f(s, \hat{y}(s))\|\, ds \\ &\le \underbrace{\|y(a) - \hat{y}(a)\| + L\int_a^x \|y(s) - \hat{y}(s)\|\, ds}_{g(x)}. \end{aligned} \tag{1.3}$$

Dado que $g'(x) = L\|y(x) - \hat{y}(x)\|$, llegamos a $g'(x) \le Lg(x)$. Multiplicando esta desigualdad por el factor integrante $\exp(-L(x-a))$ se tiene que

$$\frac{d}{dx}(g(x)\exp(-L(x-a))) \le 0,$$

de donde $g(x) \le g(a)\exp(L(x-a))$. Esto, combinado con (1.3), produce el resultado. □

Observación. El paso de la estimación integral (1.3) a la estimación puntual (1.2) es lo que se conoce en la literatura como Lema de Grönwall[6]. ♠

[5]Se entiende que los vectores se integran componente a componente.

[6]Thomas Hakon Grönwall (1877–1932), matemático sueco, desarrolló una gran parte de su carrera en los Estados Unidos de América. Sus intereses oscilaron entre las matemáticas, en las que hizo contribuciones de primer nivel en varios campos, y la investigación aplicada, especialmente en el área de la física-química.

La desigualdad (1.2) no sólo demuestra la unicidad de la solución del problema de valor inicial (PVI); también muestra que las soluciones de dicho problema dependen de manera continua del dato inicial.

La cota (1.2) es óptima en algunos casos, por ejemplo para la ecuación escalar lineal $y' = \lambda y$, $\lambda > 0$. Sin embargo, puede ser demasiado pesimista para otros.

Ejemplo. Consideramos la ecuación $y' = -\lambda y$, $\lambda > 0$, cuyo lado derecho tiene constante de Lipschitz $L = \lambda$. La diferencia entre dos soluciones verifica

$$y(x) - \hat{y}(x) = (y(a) - \hat{y}(a))e^{-\lambda(x-a)}.$$

Es decir, no sólo no crece con el tiempo, sino que decae exponencialmente. ♣

Problemas

1. Comprobar que las siguientes funciones satisfacen una condición de Lipschitz con respecto a su segunda variable para $x \in [1, \infty)$:

   (a) $f(x, y) = 2yx^{-4}$, (b) $f(x, y) = e^{-x^2} \arctan y$.

2. Consideramos el problema de valor inicial

   $$y'(x) = Ay(x) \qquad \text{para } 0 \le x \le 2, \qquad y(0) = \eta,$$

   donde

   $$A = \begin{pmatrix} -\sqrt{7} & \sqrt{2} \\ 0 & -2 \end{pmatrix}.$$

   (a) Encontrar la mejor constante de Lipschitz para la función del lado derecho usando las normas $\|\cdot\|_1$, $\|\cdot\|_2$ y $\|\cdot\|_\infty$[7].

[7]Recordemos que, para $x \in \mathbb{C}^d$ se define $\|x\|_p = \left(\sum_{j=1}^d |x_j|^p\right)^{1/p}$ si $p \in [1, \infty)$, $\|x\|_\infty = \max_{1\le j\le d} |x_j|$. Por otra parte, dada una matriz A de entradas $a_{ij} \in \mathbb{C}$, $1 \le i, j \le d$, se define

(b) Sean y e $\hat{y}$ las soluciones del problema correspondientes a datos iniciales $(0,1)^T$ y $(1,0)^T$ respectivamente. Acotar $\max_{x\in[0,2]} \|y(x)-\hat{y}(x)\|$ para cada una de las normas del apartado anterior.

3. Consideramos la función $f : \mathbb{R} \times \mathbb{R}^2 \mapsto \mathbb{R}^2$ dada por

$$f(x,y) = (x + \operatorname{sen} y^2, \frac{x^2}{2} + \cos y^1)^T.$$

(a) Encontrar, usando las normas $\|\cdot\|_1$, $\|\cdot\|_2$ y $\|\cdot\|_\infty$, una constante de Lipschitz para f con respecto a $y = (y^1, y^2)^T$.

(b) Sean y e $\hat{y}$ las soluciones de $y'(x) = f(x, y(x))$ en el intervalo $[0,1]$ con datos iniciales $y(0) = (0,1)^T$ e $\hat{y(0)} = (1,0)^T$ respectivamente. Acotar $\max_{x\in[0,1]} \|y(x)-\hat{y}(x)\|$ para cada una de las normas del apartado anterior.

4. Demostrar que cualquier problema de valor inicial para las ecuaciones diferenciales

(a) $y'(x) = 2y(1+y^2)^{-1}(1+e^{-|x|})$, (b) $y'(x) = 3|y(x)+1| + \cos x$,

tiene a lo sumo una solución.

5. Consideramos la función $f : \mathbb{R} \mapsto \mathbb{R}$ dada por $f(y) = |y|^\alpha$, $\alpha > 0$.

(a) Si $\alpha \in (0,1)$, demostrar que f no es Lipschitz en ningún intervalo que contenga el origen.

(b) Si $\alpha > 1$, demostrar que f es Lipschitz en cualquier intervalo de la forma $[-K, K]$, pero no en todo $\mathbb{R}$.

$\|A\|_p = \sup_{x\neq 0} \frac{\|Ax\|_p}{\|x\|_p}$, $1 \le p \le \infty$, y se demuestra que

$$\|A\|_1 = \max_{1\le j\le d} \sum_{i=1}^{d} |a_{ij}|, \qquad \|A\|_2 = \sqrt{r_\sigma(A^*A)}, \qquad \|A\|_\infty = \max_{1\le i\le d} \sum_{j=1}^{d} |a_{ij}|,$$

donde $r_\sigma(B)$ denota el *radio espectral* de B, $r_\sigma(B) = \max_{\lambda\in\sigma(B)} |\lambda|$, siendo $\sigma(B)$ el *espectro* de B, es decir, el conjunto de todos sus autovalores.

6. La función f satisface una *condición de Lipschitz unilateral* en $D = [a,b] \times \mathbb{R}^d$ con *constante de Lipschitz unilateral* l si

$$\langle f(x,y) - f(x,\hat{y}), y - \hat{y}\rangle \leq l\|y - \hat{y}\|_2^2 \qquad \forall (x,y),(x,\hat{y}) \in D.$$

   (a) Si $l < 0$ se dice que el problema de valor inicial PVI es *disipativo*. Demostrar que el problema de valor inicial del problema 2 es disipativo.

   (b) Demostrar que si f satisface una condición de Lipschitz unilateral con constante l y las funciones y e $\hat{y}$ son soluciones de $y'(x) = f(x, y(x))$ en $[a,b]$, entonces

$$\|y(x) - \hat{y}(x)\|_2 \leq \exp(l(x-a))\|y(a) - \hat{y}(a)\|_2 \qquad \text{si } a \leq x \leq b.$$

   (c) Dar una cota para la norma euclídea, $\|y(x)\|_2$, de la solución del problema 2 con dato inicial $y(0) = (0,1)^T$.

7. Sea y solución de (1.1) e $\hat{y}$ solución de la ecuación perturbada

$$\hat{y}'(x) = f(x, \hat{y}(x)) + r(x, \hat{y}(x)), \qquad x \in [a,b],$$

   con f continua y Lipschitz (con constante L) con respecto a su segunda variable en $D = [a,b] \times \mathbb{R}$ y r acotada en D, $\|r\| \leq M$. Demostrar que

$$\|y(x) - \hat{y}(x)\| \leq e^{L(x-a)}\|y(a) - \hat{y}(a)\| + \frac{M}{L}(e^{L(x-a)} - 1), \qquad x \in [a,b].$$

1.3. Existencia

La condición de Lipschitz junto con la continuidad resultan ser también suficientes para demostrar la existencia de una solución[8].

[8]Para que exista una solución (en un intervalo que contenga al punto donde se da el dato inicial) basta con que la función del lado derecho sea continua. Giuseppe Peano publicó por

Teorema 1.3 (Picard (1890))**.** *Sea $D = [a, b] \times \mathbb{R}^d$. Si f es continua en D y Lipschitz con respecto a su segunda variable en D, entonces existe una única solución del problema* (PVI) *en $[a, b]$.*

Demostración. Sólo falta por probar la existencia. La clave es que, por ser f continua, el Teorema Fundamental del Cálculo nos permite demostrar que el problema (PVI) es equivalente a encontrar una función $y \in C([a, b])$ tal que

$$y(x) = \eta + \int_a^x f(s, y(s))\, ds. \tag{1.4}$$

Las soluciones de esta ecuación integral son puntos fijos del operador integral $T : C([a, b]) \mapsto C([a, b])$ definido por

$$(Ty)(x) = \eta + \int_a^x f(s, y(s))\, ds.$$

La idea para hallar un punto fijo para este operador es la misma que se utiliza para hallar puntos fijos de aplicaciones en espacios de dimensión finita: la iteración. Definimos una sucesión de funciones $\{y_n\}_{n=0}^{\infty}$ por medio de la recurrencia

$$\begin{cases} y_n(x) = (Ty_{n-1})(x) = \eta + \displaystyle\int_a^x f(s, y_{n-1}(s))\, ds, \quad n = 1, 2, \ldots, \\ y_0(x) = \eta. \end{cases} \tag{1.5}$$

Si la sucesión $\{y_n\}$ converge *uniformemente* a una función y, por un lado ésta será continua; por otro lado, por ser f Lipschitz con respecto a su segunda

primera vez este resultado en 1886 con una demostración incorrecta. En 1890 dio una nueva demostración, esta vez correcta, usando aproximaciones sucesivas. Charles-Émile Picard (1856–1941) descubrió este método de forma independiente. Aunque el método se conoce hoy en día como método de Picard, el propio Picard se lo atribuye a Hermann Amandus Schwarz (1843–1921). Entre las muchas contribuciones de Picard a diversos campos de las matemáticas, destacan sus teoremas de variable compleja. Era también famoso por ser un magnífico profesor. Schwarz, nacido en Alemania, es conocido por sus aportaciones a la teorías de superficies mínimas y de aplicaciones conformes, y sobre todo por la desigualdad de Cauchy-Schwarz, de la cual probó un caso particular.

variable,

$$\|f(x, y_n(x)) - f(x, y(x))\|_\infty \leq L\|y_n(x) - y(x)\|_\infty$$

y, por tanto, $\{f(\cdot, y_n(\cdot))\}$ converge uniformemente a $f(\cdot, y(\cdot))$. Por consiguiente, al tomar límites en la relación de recurrencia (1.5) podemos intercambiar el límite con la integral, para obtener que y es solución de (1.4).

Dado que

$$y_n(x) = y_0(x) + \sum_{k=1}^{n}(y_k(x) - y_{k-1}(x)),$$

la convergencia uniforme de la sucesión de funciones $\{y_n\}$ se seguirá inmediatamente de la prueba M[9] de Weierstrass[10] si encontramos una sucesión numérica $\{a_k\}$ tal que $\|y_k(x) - y_{k-1}(x)\|_\infty \leq a_k$ y $\sum_{k=1}^{\infty} a_k < \infty$.

La acotación

$$\|y_k(x) - y_{k-1}(x)\|_\infty \leq \frac{K_\eta}{L}\frac{(L(x-a))^k}{k!},$$

donde L es la constante de Lipschitz de f respecto de la segunda variable y $K_\eta = \max_{x\in[a,b]} \|f(x,\eta)\|_\infty$, nos dice que podemos tomar $a_k = \frac{K_\eta}{L}\frac{(L(b-a))^k}{k!}$. Esta acotación se demuestra por inducción: es cierta para $k = 1$, pues

$$\|y_1(x) - y_0(x)\|_\infty \leq \int_a^x \|f(s,\eta)\|_\infty \, ds \leq K_\eta(x-a),$$

y, si es cierta para $k-1$, entonces

$$\begin{aligned}
\|y_k(x) - y_{k-1}(x)\|_\infty &\leq \int_a^x \|f(s, y_{k-1}(s)) - f(s, y_{k-2}(s))\|_\infty \, ds \\
&\leq L\int_a^x \|y_{k-1}(s) - y_{k-2}(s)\|_\infty \, ds \\
&\leq L\int_a^x \frac{K_\eta}{L}\frac{(L(s-a))^{k-1}}{(k-1)!}\, ds = \frac{K_\eta}{L}\frac{(L(x-a))^k}{k!},
\end{aligned}$$

[9] Prueba M de Weierstrass: Sea $\{f_n\}$ una sucesión de funciones definidas sobre A y con valores en $\mathbb{R}$. Si existe una sucesión de números $\{M_n\}$ tal que: (i) $|f_n(x)| \leq M_n$ para todo $x \in A$; y (ii) $\sum_{n=1}^{\infty} M_n < \infty$, entonces $\sum_{n=1}^{\infty} f_n$ converge uniformemente en A.

[10] Karl Theodor Wilhelm Weierstrass (1815–1897), matemático alemán, uno de los fundadores de la moderna teoría de funciones. Según él, "un verdadero matemático que no tenga también algo de poeta, nunca será un matemático perfecto".

es decir, es cierta para k. □

En resumen, las condiciones:

$$\left.\begin{array}{ll} \text{(i)} & f \text{ continua en } D = [a,b] \times \mathbb{R}^d, \\ \text{(ii)} & f \text{ Lipschitz en } D \text{ respecto de su segunda variable,} \end{array}\right\} \qquad (\mathrm{H}_f)$$

garantizan que el problema (PVI) está *bien planteado*[11]: (i) tiene solución; (ii) es única y (iii) depende de manera continua de los datos iniciales. En lo sucesivo supondremos siempre que f cumple las hipótesis (H_f).

En ocasiones necesitaremos que la solución sea más regular. Para conseguirlo bastará con imponer mayor regularidad a la f. En efecto, es fácil ver que si, además de ser Lipschitz con respecto a su segunda variable, $f \in C^p([a,b]\times\mathbb{R}^d)$, entonces $y \in C^{p+1}([a,b])$.

Ejemplo. Si $f \in C^1([a,b])$, usando la ecuación se obtiene que

$$y''(x) = \frac{\partial f}{\partial x}(x, y(x)) + \sum_{J=1}^{d} \frac{\partial f}{\partial y^J}(x, y(x))(y^J)'(x).$$

Por consiguiente, puesto que $y' \in C([a,b])$, se concluye que $y \in C^2([a,b])$. ♣

Problemas

1. Consideramos la ecuación integral

$$y(x) = 10 + \int_1^x y(s)\,ds, \qquad 1 \le x \le 2.$$

   (a) Demostrar que tiene una única solución continua.

   (b) Calcularla.

[11] Esta noción fue introducida por Jacques Salomon Hadamard (1865–1963), matemático francés conocido principalmente por haber demostrado el teorema del número primo en 1896.

2. Demostrar que existe un único par de funciones $u, v \in C([0,10])$ que resuelve el sistema de ecuaciones integrales

$$\begin{cases} u(x) = 1 + \displaystyle\int_0^x \left(u(s) + 2v(s) + s + \int_0^{v(s)} \mathrm{e}^{-t^2}\, dt \right) ds, \\ v(x) = 1 + \displaystyle\int_0^x \left(-u(s) + 3v(s) + \frac{s^2}{2} + \arctan(\exp(u(s))) \right) ds, \end{cases}$$

para todo $x \in [0,10]$.

3. Sean c y d funciones continuas en $[a,b]$. Consideramos el problema de valor inicial PVI con $f : [a,b] \times \mathbb{R} \mapsto \mathbb{R}$, $f(x,y) = c(x)y + d(x)$.

   (a) Demostrar que el problema tiene solución única.

   (b) Si además $c, d \in C^\infty([a,b])$, demostrar que $y \in C^\infty([a,b])$.

   Sugerencia. Usar la fórmula de Leibniz[12] para la derivada k-ésima de un producto,

$$(fg)^{(k)} = \sum_{j=0}^{k} \binom{k}{j} f^{(j)} g^{(k-j)}.$$

   (c) Si $b = a + 3$ y $c(x) = \operatorname{sen}(\exp(\arctan x))$, determinar un valor de δ que garantice que dos soluciones que inicialmente distan menos que δ no se separan más de 10^{-1} en todo el intervalo $[a,b]$.

4. Sea f continua en $D = [a,b] \times \{y \in \mathbb{R}^d : \ \|y - \eta\|_\infty \leq C\}$ y Lipschitz en D, con constante L, con respecto a su segunda variable. Sea $K > 0$ una constante tal que $\max_{(x,y)\in D} \|f(x,y)\|_\infty \leq K$. Si

$$C \geq \frac{K}{L} \left(e^{L(b-a)} - 1 \right), \tag{CEL}$$

probar que existe una única solución del problema de valor inicial PVI en $[a,b]$ tal que $\|y(x) - \eta\| \leq C$ en $[a,b]$.

[12] Gottfried Wilhelm von Leibniz (1646–1716), matemático y filósofo alemán, inventó el cálculo infinitesimal independientemente de Isaac Newton. La notación que introdujo para las derivadas e integrales es la de uso común hoy en día.

5. Consideramos el problema

$$y'(x) = y^2(x) \arctan y(x), \quad x \in [0, b], \qquad y(0) = 1. \tag{1.6}$$

   (a) Aplicar el resultado del problema anterior para encontrar un valor $b > 0$ tal que el problema (1.6) tenga solución.

   (b) Sea $T = \sup\{b : \text{ (1.6) tiene solución}\}$. Probar que $\lim_{x \to T^-} y(x) = \infty$.

   (c) Demostrar que $T < \infty$. Este fenómeno se denomina *explosión*.

6. Se considera el PVI

$$y'(x) = -y^2(x) + y(x) + 2y(x)x^2 - x^2 - x^4, \quad y(0) = 1/2.$$

   (a) Demostrar que mientras la solución existe se verifica que

$$x^2 < y(x) < x^2 + 1.$$

   (b) Utilizar el apartado anterior para demostrar que la solución existe para todo $x \geq 0$.

7. El Teorema de Picard, además de la continuidad Lipschitz de $f = f(x, y)$ con respecto a y, tiene también como hipótesis la continuidad de f con respecto a sus dos variables. Demuéstrese mediante un contraejemplo que no se puede eliminar esta segunda hipótesis.

Capítulo 2

Métodos numéricos. Convergencia

2.1. Métodos numéricos

Aunque el problema (PVI) tenga solución, en la mayor parte de los casos no es posible encontrar una forma cerrada para la misma por métodos analíticos.

Ejemplo. Consideramos el problema de valor inicial

$$y'(x) = e^{-y^2(x)} + \frac{1}{1+x^2}, \qquad y(0) = 0. \tag{2.1}$$

Usando el teorema del valor medio se tiene que

$$|f(x,y) - f(x,\hat{y})| = |e^{-y^2} - e^{-\hat{y}^2}| = |-2\xi e^{-\xi^2}||y - \hat{y}| \leq L|y - \hat{y}|,$$

pero no sabemos encontrar una fórmula para la solución. ♣

En casos como éste habrá que contentarse con obtener una aproximación numérica de la solución. La demostración del teorema de Picard sugiere una manera de construir una. En efecto, las funciones y_n obtenidas mediante la relación de recurrencia (1.5) aproximan uniformemente a la solución.

Ejemplo. Si aplicamos el método de Picard al problema lineal escalar con coeficientes constantes

$$y' = y, \qquad y(0) = 1,$$

obtenemos la recurrencia

$$y_0(x) = 1, \qquad y_n(x) = 1 + \int_0^x y_{n-1}(s)\, ds, \quad n = 1, 2, \ldots,$$

cuya solución es

$$y_n(x) = \sum_{k=0}^{n} \frac{x^k}{k!},$$

es decir, el polinomio de Taylor[1] de grado n de la exponencial en $x = 0$. ♣

Pero, ¡cuidado!, este ejemplo es muy especial. En general será difícil, o incluso imposible, evaluar las integrales involucradas en forma cerrada. Así que buscamos una idea distinta, la *discretización*.

Sustituimos el intervalo continuo $[a, b]$ por un conjunto discreto de puntos, $a = x_0 < x_1 < \cdots < x_N = b$, al que llamaremos *malla*. Nuestro objetivo es encontrar una forma de producir una sucesión de valores $\{y_n\}_{n=0}^N$ que aproxime a la solución y de (PVI) en los puntos $\{x_n\}_{n=0}^N$,

$$y_n \approx y(x_n).$$

Un *método numérico* para la integración[2] de (PVI) será un procedimiento para producir la sucesión de valores aproximados $\{y_n\}$. Los valores y_n se pueden interpolar para obtener una aproximación de $y(x)$ en puntos x que no pertenezcan a la malla.

Los métodos numéricos para resolver el PVI que vamos a considerar producen los valores $\{y_n\}_{n=0}^N$ de forma iterativa: el valor de y_{n+1} se calcula a partir de

[1] Brook Taylor (1685–1731), matemático inglés. Además del famoso desarrollo en serie que lleva su nombre, creó el cálculo de diferencias finitas e inventó la integración por partes.

[2] Históricamente se ha reservado el término *integración* para la resolución de EDOs, empleándose el término *cuadratura* para el cálculo de integrales.

los $k \geq 1$ valores $\{y_j\}_{j=n-k+1}^{n}$ obtenidos previamente. Para empezar el procedimiento será por tanto necesario disponer de k *valores de arranque*, $y_0, \ldots, y_{k-1}$, obtenidos de forma independiente al método. El número k se conoce como el *número de pasos* del método.

Es costumbre describir el cálculo de y_{n+1} como "dar un paso" de longitud $h_n = x_{n+1} - x_n$. Para facilitar la exposición, en la mayor parte de lo que sigue nos restringiremos a mallas equiespaciadas,

$$x_n = a + nh, \qquad n = 0, 1, 2, \ldots, N, \qquad h = (b-a)/N,$$

en las que la longitud del paso es constante, $h_n = h$. Pero hay que tener presente que gran parte de la potencia de los algoritmos modernos proviene de su capacidad para cambiar h_n automáticamente a medida que se realizan los cálculos. Veremos cómo hacer esto en un capítulo posterior.

Problemas

1. Utilizar los iterantes de Picard para construir una aproximación numérica de la solución del problema de valor inicial

$$y'' + y = 0, \quad 0 \leq x \leq 2\pi, \qquad y(0) = 0,\ y'(0) = 1. \tag{2.2}$$

2. Si $f \in C^p([a,b] \times \mathbb{R}^d)$, podemos encontrar una solución aproximada del problema de valor inicial (PVI) utilizando la fórmula de Taylor de orden p alrededor de $x = a$, $y(x) \approx \sum_{j=0}^{p} y^{(j)}(a)\frac{(x-a)^j}{j!}$. Los valores de las derivadas $y^{(j)}(a)$ se pueden calcular derivando la EDO y usando la condición inicial. Utilizar este procedimiento para aproximar la solución del problema

$$y'(x) = -y^2(x),\ 0 \leq x \leq 1/2, \qquad y(0) = 1,$$

por un polinomio de orden p y estimar el error cometido.

3. *Método de las series de potencias*[3]. Si la solución y de una EDO se puede escribir como una serie de potencias, $y(x) = \sum_{n=0}^{\infty} a_n x^n$, se puede intentar usar la ecuación para obtener cada valor a_n a partir de los anteriores, $a_0, \ldots, a_{n-1}$. Los primeros valores, necesarios para arrancar el procedimiento, se extraen de las condiciones iniciales. Usar este método para resolver el problema (2.2).

4. Resolver las siguientes relaciones de recurrencia[4], definidas todas ellas para $n \geq 0$:

   (a) $y_{n+1} - y_n = 2n + 3$, $y_0 = 1$;

   (b) $y_{n+1} - y_n = 3n^2 - n$, $y_0 = 3$;

   (c) $y_{n+1} - 2y_n = 5$, $y_0 = 1$;

   (d) $y_{n+1} - 2y_n = 2^n$, $y_0 = 1$;

   (e) $y_{n+2} + 3y_{n+1} + 2y_n = 3^n$, $y_0 = 0$, $y_1 = 1$;

   (f) $y_{n+2} + 4y_{n+1} + 4y_n = 7$, $y_0 = 1$, $y_1 = 2$;

   (g) $y_{n+2} - 6y_{n+1} + 9y_n = 3 \cdot 2^n + 7 \cdot 3^n$, $y_0 = 1$, $y_1 = 4$.

5. El número a_n de euros de activo de una compañía se incrementa cada año cinco veces lo que se incrementó el año anterior. Si $a_0 = 3$ y $a_1 = 7$, calcular a_n.

[3]Este método, introducido por Newton en su "Methodus Fluxionum et Serierum Infinitorum" (1671), es el primero que se usó para resolver ecuaciones diferenciales ordinarias.

[4]Para repasar todo lo que tiene que ver con este tema recomiendo leer el capítulo 8 del libro "Diez lecciones de cálculo numérico", J. M. Sanz-Serna, Universidad de Valladolid. Secretariado de publicaciones e intercambio científico, 2010 (segunda edición).

2.2. El método de Euler

Nuestro primer método, padre de alguna manera de todos los demás que vamos a estudiar, es el *método de Euler*[5]. Se define por medio de la recurrencia

$$y_{n+1} = y_n + hf(x_n, y_n), \qquad n = 0, \ldots, N-1. \tag{2.3}$$

Esta recurrencia se debe complementar con un *valor de arranque* $y_0 \approx y(x_0)$[6].

Vamos a ver cómo llegar al método de Euler desde cuatro puntos de vista diferentes. Cada uno de ellos se podrá generalizar, como veremos en en el siguiente capítulo, para obtener nuevos métodos numéricos.

• Fórmula de Taylor. Desarrollamos la solución alrededor de x_n y usamos la ecuación diferencial para obtener

$$\begin{aligned} y(x_{n+1}) &= y(x_n) + hy'(x_n) + \frac{h^2}{2}\, y''(\overline{\xi}_n) \\ &= y(x_n) + hf(x_n, y(x_n)) + \frac{h^2}{2}\, y''(\overline{\xi}_n). \end{aligned} \tag{2.4}$$

La notación $\overline{\xi}_n$ indica que el punto intermedio $\overline{\xi}_n \in [x_n, x_{n+1}]$ puede variar de componente a componente.

Así pues, la solución teórica resuelve la recurrencia

$$y(x_{n+1}) = y(x_n) + hf(x_n, y(x_n)) + R_n, \qquad n = 0, \ldots, N-1, \tag{2.5}$$

con $R_n = \frac{h^2}{2}\, y''(\overline{\xi}_n)$. Esta última cantidad es desconocida, lo que impide resolver la recurrencia. No obstante, sabemos que R_n es pequeña si h es pequeña. Si la recurrencia (2.3) es estable ante pequeñas perturbaciones, esperamos que

[5] Leonhard Euler (1707-1783): "Institutionum Calculi Integralis" (1768), volumen I, sección II, capítulo VII. *Opera Omnia*, vol. XI. Para aproximarse un poco a la ingente obra de Euler recomiendo el magnífico libro "Euler. El maestro de todos los matemáticos", William Dunham, Nivola, 2006 (segunda edición).

[6] Lo ideal sería tomar $y_0 = y(x_0)$, pero en general esto no es posible, debido a la precisión finita del ordenador.

$\{y(x_n)\}$ se parezca a la solución $\{y_n\}$ de la recurrencia sin perturbar (2.3). La cantidad R_n, que es lo que le sobra a la solución teórica para ser solución de la ecuación del método numérico, recibe el nombre de *residuo*.

Geométricamente (en el caso $d = 1$), al despreciar el residuo lo que estamos haciendo es aproximar a la solución por su tangente.

• Cuadratura numérica. Integramos la EDO en (x_n, x_{n+1}), y obtenemos

$$y(x_{n+1}) = y(x_n) + \int_{x_n}^{x_{n+1}} f(x, y(x))\, dx.$$

Para calcular la integral utilizamos la fórmula de cuadratura numérica

$$\int_c^{c+h} g = hg(c) + \mathcal{E}_h(g), \qquad \mathcal{E}_h(g) = \frac{h^2}{2} g'(\overline{\xi}), \quad \overline{\xi} \in [c, c+h],$$

llamada regla de los rectángulos, válida si $g \in C^1([c, c+h])$, y llegamos de nuevo a (2.5).

• Diferenciación numérica. Si aplicamos la fórmula de diferenciación progresiva

$$g'(x) = \frac{g(x+h) - g(x)}{h} + \mathcal{E}_h(g), \qquad \mathcal{E}_n(g) = \frac{h}{2} g''(\overline{\xi}),$$

válida para funciones $g \in C^2([x, x+h))$, a la función y, obtenemos que

$$f(x_n, y(x_n)) = y'(x_n) = \frac{y(x_{n+1}) - y(x_n)}{h} + \frac{h}{2} y''(\overline{\xi}_n),$$

de donde, despejando, obtenemos una vez más (2.5).

• Colocación. Consideramos el polinomio u de grado menor o igual que 1 que verifica

$$u(x_n) = y(x_n), \qquad u'(x_n) = f(x_n, u(x_n)).$$

Es decir, pedimos que u satisfaga la condición inicial y también la ecuación en el punto $x = x_n$. Es inmediato ver que $u(x) = y(x_n) + f(x_n, y(x_n))(x - x_n)$. Si tomamos $y(x_{n+1}) \approx u(x_{n+1})$ llegamos otra vez a (2.5).

Problemas

1. Aplicamos el método de Euler con paso fijo $h = 1/N$ al PVI $y' = y$ en $[0, 1]$, $y(0) = 1$. Como valor de arranque tomamos $y_0 = 1$. Calcular la solución numérica, $\{y_n\}_{n=0}^N$ y demostrar que existe una constante C tal que $|y(x_n) - y_n| \leq Ch$.
2. Consideramos el PVI $y'' + y = 0$, $x \in [0, 1]$, $y(0) = 1$, $y'(0) = 0$. Escribirlo como un problema de primer orden, y aplicarle entonces el método de Euler con paso fijo $h = 1/N$, suponiendo que no se comete ningún error en los valores de arranque. Demostrar que el máximo error cometido, $\max\limits_{n=0,\ldots,N} |y(x_n) - y_n|$, es menor o igual que Ch para alguna constante C.
3. Consideramos el problema de valor inicial $y'(x) = f(x, y(x))$, $x \in [a, b]$, $y(a) = \eta$.
   (a) Supongamos que existe un vector columna c tal que $c^T f(x, y) = 0$. Demostrar que la solución del problema satisface la *ley de conservación lineal* $c^T y(x) = c^T \eta$ para todo $x \in [a, b]$.
   (b) Demostrar que las soluciones obtenidas al aplicar el método de Euler a este PVI también satisfacen la misma ley de conservación lineal, $c^T y_n = c^T y_0$, $n = 0, \ldots, N$.

2.3. Convergencia

¿Cuándo podemos decir que un método numérico aproxima bien a la verdadera solución del problema (PVI)? Una medida posible de la bondad de la aproximación es el mayor error cometido,

$$\max_{0 \leq n \leq N} \|y(x_n) - y_n\|.$$

¿Se puede hacer el error tan pequeño como se desee tomando un número suficientemente grande de puntos en la malla? Si esto es así diremos que el método es convergente.

Definición 2.1. *Un método numérico de k pasos para resolver el* (PVI) *se dice* convergente[7] *si para todo* PVI *se tiene que*

$$\lim_{N\to\infty} \max_{k\le n\le N} \|y(x_n)-y_n\| = 0 \qquad si \lim_{N\to\infty} \max_{0\le n\le k-1} \|y(x_n)-y_n\| = 0.$$

Observación. La condición $\lim_{N\to\infty} \max_{0\le n\le k-1} \|y(x_n)-y_n\| = 0$ sobre los valores de arranque es equivalente a pedir que $\lim_{h\to 0^+} y_n = y(x_0)$ para $n = 0,\dots,k-1$. Estamos pidiendo que los valores de arranque, $\{y_n\}_{n=0}^{k-1}$, aproximen bien al dato inicial $y(x_0)$; si esto no es así no hay por qué esperar que la solución numérica se parezca a la teórica. ♠

Nuestro primer ejemplo de método convergente es el método de Euler.

Teorema 2.2. *El método de Euler es convergente.*

Demostración. Consta de dos pasos.

Paso 1: Estabilidad. Restando la ecuación del método (2.3) de la recurrencia perturbada (2.5) satisfecha por la verdadera solución, tomando normas y usando la condición de Lipschitz, se tiene que

$$\begin{aligned}\|y(x_{n+1})-y_{n+1}\| &\le \|y(x_n)-y_n\| + h\|f(x_n,y(x_n))-f(x_n,y_n)\| + \|R_n\| \\ &\le (1+Lh)\|y(x_n)-y_n\| + \max_{0\le n\le N-1}\|R_n\|.\end{aligned}$$

A partir de aquí se demuestra fácilmente por inducción que

$$\|y(x_n)-y_n\| \le e^{L(x_n-x_0)}\|y(x_0)-y_0\| + e^{L(x_n-x_0)}(x_n-x_0)\max_{0\le n\le N-1}\frac{\|R_n\|}{h}; \quad (2.6)$$

es decir, podemos controlar la diferencia entre la solución de la ecuación perturbada y la solución de la ecuación sin perturbar si controlamos la perturbación: la recurrencia es estable.

[7] En el caso de métodos de paso variable hay convergencia si el máximo error tiende a 0 cuando el *diámetro de la malla*, $\max_{0\le n\le N-1}(x_{n+1}-x_n)$, tiende a 0.

Paso 2: Control del residuo (consistencia). Por el Teorema del Valor Medio,

$$\frac{R_n}{h} = \frac{y(x_{n+1}) - y(x_n)}{h} - f(x_n, y(x_n)) = y'(\overline{\xi}_n) - y'(x_n).$$

Usando la continuidad uniforme[8] de $y'(x)$ en $[a, b]$ se concluye que

$$\max_{0 \le n \le N-1} \frac{\|R_n\|}{h} \to 0,$$

y por tanto la convergencia. □

Problemas

1. Sea y la solución del problema (PVI) y sea y^I la función continua lineal a trozos que pasa por los puntos (x_n, y_n), $n = 0, \ldots, N$, donde $\{y_n\}$ es una solución numérica del problema. Una medida alternativa del error cometido viene dada por $\max_{a \le x \le b} \|y(x) - y^I(x)\|_\infty$. Hallar una cota para esta cantidad en términos de $\max_{0 \le n \le N} \|y(x_n) - y_n\|$ y de h, suponiendo que $f \in C^1([a, b] \times \mathbb{R}^d)$.
2. Demostrar la fórmula (2.6).
3. Demostrar que la convergencia del método de Euler, aplicado a un problema "de cuadratura", $y'(x) = f(x)$, implica la convergencia de las sumas de Riemann $\sum_{n=0}^{N-1} hf(a + nh)$ cuando h tiende a cero al valor de la integral $\int_a^b f(x)\, dx$.

2.4. Convergencia de orden p

Si miramos con cuidado la demostración de la convergencia del método de Euler, veremos que cuando $f \in C^1$ se tiene algo mejor. En este caso $R_n =$

[8] Una función $f(x)$ es uniformemente continua si pequeños cambios en el valor de x producen pequeños cambios en el valor de la función (continuidad) y el tamaño de los cambios en $f(x)$ depende sólo del tamaño de los cambios en x pero no del valor de x (uniformidad). Las funciones continuas son uniformemente sobre conjuntos compactos.

$y''(\overline{\xi}_n)h^2/2$ y, utilizando (2.6), se llega a la cota para el error

$$\|y(x_n) - y_n\| \leq e^{L(b-a)}\|y(x_0) - y_0\| + e^{L(b-a)}(b-a)Ch, \tag{2.7}$$

donde $C = \frac{1}{2}\,\text{máx}_{x\in[a,b]}\,\|y''(x)\|$ y $L > 0$ es una constante de Lipschitz para f con respecto a su segunda variable. No es necesario conocer la solución para estimar C; se puede hacer utilizando la ecuación diferencial.

Ejemplo. Sea y la solución de $y' = \operatorname{sen}(\exp(y))$ en $[0,1]$, $y(0) = 0$. Queremos estimar y'' sin conocer y. Para ello derivamos la EDO,

$$y'' = \cos(\exp(y))\exp(y)y' = \cos(\exp(y))\exp(y)\operatorname{sen}(\exp(y)).$$

Para acotar el factor exponencial tenemos que tener una idea del tamaño de y. Dado que $|y'| \leq 1$ tenemos que $|y(x)| \leq x \leq 1$. Concluimos que $|y''| \leq e$. ♣

La cota (2.7) en general sobreestima el error en muchos órdenes de magnitud, y no se debe por tanto utilizar en la práctica. Sin embargo, da una información importante: el error del método de Euler es una $O(h)$[9], siempre y cuando $\|y(x_0) - y_0\| = O(h)$. Eso motiva la siguiente definición.

Definición 2.3. *Un método numérico convergente para resolver el* (PVI) *es* convergente de orden p *si este es el mayor entero tal que para todo* PVI *con* $f \in C^q(D)$, $q \leq p$, *se tiene que*

$$\max_{k\leq n\leq N}\|y(x_n) - y_n\| = O(h^q) \qquad si \quad \max_{0\leq n\leq k-1}\|y(x_n) - y_n\| = O(h^q)$$

cuando $h \to 0^+$.

[9] Notaciones O y o de Landau. Escribimos $f = O(g)$, $x \to x_0$ (se lee "f es O grande de g cuando x tiende a x_0"), si existen constantes $K, r > 0$ tales que para cada x que verifica $0 < |x - x_0| < r$ se tiene que f y g están definidas y satisfacen $|f(x)| \leq K|g(x)|$. Escribimos $f = o(g)$, $x \to x_0$ (se lee "f es o pequeña de g cuando x tiende a x_0"), si existe $r > 0$ tal que para cada x que verifica $0 < |x - x_0| < r$ se tiene que f y g están definidas, g no se anula y $\lim_{x\to x_0} f(x)/g(x) = 0$. Edmund Georg Hermann Landau (1877–1938) fue un matemático alemán especializado en teoría analítica de números.

Con esta definición, el método de Euler es convergente de orden al menos uno. Para ver que no es convergente de orden 2 lo aplicamos al problema

$$y'(x) = x \qquad \text{para } x \in [0, b], \qquad y(0) = 0,$$

cuya solución exacta es $y(x) = x^2/2$. Si tomamos valor de arranque $y_0 = 0$ tenemos que

$$\begin{aligned} y_1 &= y_0 + hx_0 = 0, \\ y_2 &= y_1 + hx_1 = h^2, \\ y_3 &= y_2 + hx_2 = h^2(1+2) \\ &\vdots \\ y_n &= h^2(1 + \cdots + (n-1)). \end{aligned}$$

Es decir,

$$y_n = \frac{x_n^2}{2} - \frac{x_n h}{2},$$

y por tanto

$$\max_{1 \le n \le N} \|y(x_n) - y_n\| = \max_{1 \le n \le N} \|x_n h/2\| = bh/2.$$

El error tiende a cero cuando $h \to 0^+$, pero lo hace como h y no como h^2.

Que un método sea de orden p significa que para *todos* los problemas razonables el error que se comete con el método es de ese orden. Pero puede suceder que para algún problema concreto el método sea mejor.

Ejemplo. Consideramos el problema

$$y'(x) = 1 \quad \text{en } [0, 1], \qquad y(0) = 0,$$

cuya solución es $y(x) = x$. Al aplicar Euler obtenemos la recurrencia $y_{n+1} = y_n + h$, cuya solución es $y_n = nh + y_0$. Por consiguiente, dado cualquier entero positivo p, si $|y(x_0) - y_0| = O(h^p)$, es decir, si $y_0 = O(h^p)$, se tiene que $|y(x_n) - y_n| = O(h^p)$ y el método es por tanto al menos de orden p. ♣

Problemas

1. Consideramos la ecuación escalar $y' = \arctan\, y$. Encontrar una cota para y'' e y''' en el intervalo $[0, 1]$ sin hallar y explícitamente.
2. Demostrar las siguientes identidades:
   (a) $C \cdot O(h^k) = O(h^k)$ para cualquier constante C;
   (b) $O(h^k) \pm O(h^k) = O(h^k)$;
   (c) $O(h^k) + O(h^m) = O(h^{\min(k,m)})$;
   (d) $O(h^k) \cdot O(h^m) = O(h^{k+m})$;
   (e) $\int_0^h O(t^k)dt = O(h^{k+1})$;
   (f) $\dfrac{1}{1+O(h^k)} = 1 + O(h^k)$.
3. Sean $g_1(h) = 1 - h^2 + O(h^4)$ y $g_2(h) = 3 + h + O(h^2)$. Encontrar el mayor entero k tal que $g_1(h) \cdot g_2(h) = 3 + O(h^k)$.
4. Calcular la solución teórica del problema de valor inicial

$$y'(x) = \min(y(x), 2), \quad 0 \le x \le 2, \qquad y(0) = 1,$$

   y aproximarla mediante el método de Euler. Comprobar que la solución numérica converge a la teórica. ¿Es la convergencia de orden 1?
5. Consideramos el problema de valor inicial $y'(x) = x(\operatorname{sen} y(x))^2$, $x \in [0, 1]$, $y(0) = 1$. Si se resuelve por el método de Euler con $y_0 = 1$, ¿qué valor de h habrá que tomar para garantizar un error menor que 10^{-3}?

2.5. Programación del método de Euler

Programar el método de Euler para sistemas es realmente fácil usando Matlab. Basta con "copiar" la fórmula (2.3) (la hemos adaptado para permitir el uso de mallas no equiespaciadas).

```
1  function y=euler(ld,x,y0);
2
3  % ld: Nombre del fichero que contiene la función que calcula el lado derecho de la
4  %     EDO. Tendrá siempre dos argumentos de entrada, x (escalar) e y (vector
5  %     columna). La salida será un vector columna.
6  % x:  Malla (vector fila de longitud N).
7  % y0: Dato inicial (vector columna de tamaño d, la dimensión del sistema de edos).
8  % y:  Solución numérica en los puntos de la malla (matriz de tamaño d x N).
9
10 N=length(x)-1;
11 y(:,1)=y0;
12 for n=1:N,
13     y(:,n+1)=y(:,n)+(x(n+1)-x(n))*feval(ld,x(n),y(:,n));
14 end
```

Programa 2.1: Método de Euler.

Veamos al método en acción. Se lo aplicaremos al PVI

$$y'(x) = -y(x) - 5e^{-x}\operatorname{sen} 5x, \quad x \in [0,3], \qquad y(0) = 1. \tag{2.8}$$

Podemos calcular la solución teórica por medio de la instrucción

```
dsolve('Dy=-y-5*exp(-x)*sin(5*x)','y(0)=1','x')
```

que nos devuelve

```
ans =
exp(-x)*cos(5*x).
```

Observación. Por defecto, la función `dsolve` supone que la variable independiente es `t`. Si queremos que sea cualquier otra variable, tenemos que dar el nombre de la misma como último argumento de la función. ♠

Para calcular la solución numérica con nuestro programa `euler` lo primero que tenemos que hacer es escribir una función que evalúe el lado derecho. Para ello creamos en el fichero `ejescalar.m` la función

```
function yprima=ejescalar(x,y)
yprima=-y-5*exp(-x)*sin(5*x);
```

Si ahora ejecutamos el conjunto de instrucciones

```
fplot(@(x) exp(-x)*cos(5*x), [0 3]), hold

x=0:.2:3;
y=euler(@ejescalar,x,1);
plot(x,y,'r*')

x=0:.1:3;
y=euler(@ejescalar,x,1);
plot(x,y,'gd')

legend('Solucion','h=0.2','h=0.1')
xlabel('x'), ylabel('y','Rotation',0)
```


obtenemos el dibujo de la Figura 2.1, en el que se superponen la solución exacta y las aproximaciones numéricas obtenidas con longitudes de paso $h = 0{,}1$ y $h = 0{,}2$.

Se observa a simple vista que la aproximación con $h = 0{,}1$ es mejor que con $h = 0{,}2$. Para confirmar esta observación elaboraremos un *diagrama de eficiencia*. En este tipo de diagramas se representa en escalas logarítmicas la precisión conseguida al aplicar un método a un cierto problema frente al coste computacional. A continuación se da un ejemplo de programa para elaborar un diagrama de eficiciencia para el método de Euler aplicado al problema (2.8) en el que se mide el coste computacional mediante el número de evaluaciones de función[10], en este caso N (evaluar funciones suele ser lo más costoso a la hora de aplicar un método).

[10]Otra posible medida del coste computacional, quizá más justa, es el tiempo de cálculo.

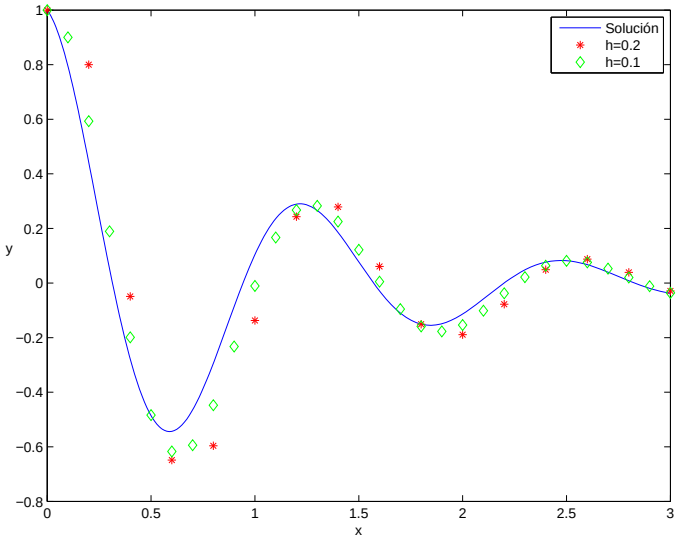

Figura 2.1: Solución numérica del PVI (2.8) por el método de Euler.

```
1  function eficienciaeuler
2
3  % Diagrama de eficiencia para Euler aplicado al ejemplo escalar
4  eeuler=[];
5  feuler=[];
6  N=30;
7  for i=1:10
8      x=linspace(0,3,N+1);
9      y=euler(@ejescalar,x,1);
10     eeuler=[eeuler,max(abs(solee(x)-y))];
11     feuler=[feuler,N];
12     N=2*N;
13 end
14 figure(1)
15 loglog(feuler,eeuler,'-o')
16 grid
17 xlabel('evaluaciones de funcion'), ylabel('error')
18
19 function sol = solee(x)
20 sol = exp(-x).*cos(5*x);
```

Programa 2.2: Elaboración de un diagrama de eficiencia para el problema (2.8).

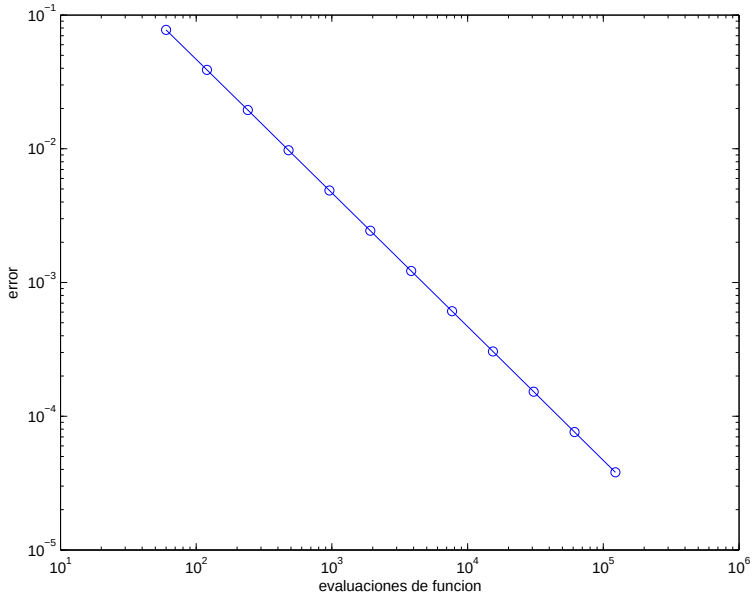

Figura 2.2: Diagrama de eficiencia para el problema (2.8).

El diagrama resultante de ejecutar este programa se muestra en la figura 2.2.

Como era de esperar, a mayor trabajo, menor error. Por otra parte, puesto que el método es de orden 1 y $f \in C^1$, sabemos que existe una constante C tal que $\log_{10} \texttt{error} \leq -\log_{10} \texttt{N} + C$. En el diagrama observamos que se tiene algo mejor que un menor o igual: una igualdad aproximada. La razón es que si, como sucede en nuestro caso, $f \in C^2$, el error global admite un desarrollo asintótico de la forma $y(x_n) - y_n = d(x_n)h + O(h^2)$, para una cierta función $d \in C^2([a, b])$.

Problemas

1. La ecuación de (2.8) da para cada punto (x, y) la derivada de la solución que pasa por ese punto. Define de esta forma un *campo de pendientes*. Las soluciones tienen que ser tangentes en cada punto a este campo.

Podemos representar el campo de pendientes dibujando en cada pundo (x, y) una flecha cuya pendiente sea $f(x, y)$.

(a) Usar la función `quiver` de Matlab para dibujar el campo de pendientes de la ecuación que aparece en (2.8).

(b) Superponer en el dibujo anterior la verdadera solución.

(c) Superponer también la solución numérica obtenida con el método de Euler y longitud de paso $h = 0{,}5$, uniendo los puntos obtenidos por medio de líneas rectas.

2. La ecuación del péndulo simple, $\theta'' + \operatorname{sen}\theta = 0$, se puede escribir como un sistema de primer orden,

$$(y^1)' = y^2, \qquad (y^2)' = -\operatorname{sen}(y^1),$$

para las variables $y^1 = \theta$, $y^2 = \theta'$.

(a) Utilizar el método de Euler con longitud de paso $h = 0{,}005$ para aproximar numéricamente en el intervalo $[0, 10]$ a las soluciones con datos iniciales $(1, 1)^T$, $(-5, 2)^T$ y $(5, -2)^T$.

(b) Dibujar en el plano y^1–y^2 las curvas correspondientes a las soluciones calculadas en el apartado anterior, y superponer el campo de vectores asociado a la EDO. Obtendremos de esta forma un esbozo del *plano de fases* asociado a la EDO.

(c) Demostrar que las soluciones de la ecuación del péndulo conservan la energía: la cantidad $E(t) = \frac{1}{2}\left(y^2(t)\right)^2 - \cos y^1(t)$ es constante para todo t.

(d) Comprobar que la energía sólo se conserva de forma aproximada para las soluciones numéricas obtenidas por medio del método de Euler.

3. La función error, $\operatorname{erf}(x)$, se define usualmente por medio de una integral,

$$\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-s^2}\, ds,$$

pero también se puede definir como la solución de la ecuación diferencial

$$y'(x) = \frac{2}{\sqrt{\pi}}\, e^{-x^2}, \qquad y(0) = 0.$$

Usar el método de Euler con longitud de paso $h = 0,01$ para resolver esta ecuación diferencial en el intervalo $[0, 2]$. Comparar los resultados con los valores proporcionados por la función de Matlab `erf`.

4. Estimar, a partir del diagrama de eficiencia de la figura 2.2, qué longitud de paso hay que tomar para conseguir que el máximo error cometido al aplicar el método de Euler al problema (2.8) sea 0,002. Comprobar que la cota (2.7) sobreestima el error para esta longitud de paso en varios órdenes de magnitud.

5. Consideramos un método de orden p que, para funciones f suficientemente regulares, admite un desarrollo asintótico para el error global,

$$y(x_n) - y_n = d(x_n)h^p + O(h^{p+1}), \qquad d \in C^2([a, b]).$$

Lo aplicamos a un problema con dos longitudes de paso distintas, h y h'. Demostrar que

$$p \approx \frac{\log(\texttt{error}(h)) - \log(\texttt{error}(h'))}{\log h - \log h'}.$$

Esto da una forma de determinar empíricamente el orden de un método. Como aplicación, determinar el orden empírico de un método que aplicado con longitudes de paso $h = 0,001$ y $h = 0,003$ produce errores $1,011 \cdot 10^{-8}$ y $2,699 \cdot 10^{-7}$ respectivamente.

6. Elaborar un diagrama de eficiencia para el método de Euler aplicado al PVI

$$y'(x) = -\frac{xy(x)}{1 - x^2}, \quad x \in [0, 1], \qquad y(0) = 1,$$

cuya solución exacta es $y(x) = \sqrt{1 - x^2}$. ¿Qué orden empírico se observa?

Capítulo 3

Construcción de métodos

Nuestro objetivo en este capítulo es ver, a través de algunos ejemplos significativos, formas razonables de construir métodos numéricos. Supondremos siempre que f es tan regular como sea necesario para justificar los cálculos que se hagan.

3.1. Métodos de Taylor

En lugar de sólo uno, como hicimos al obtener el método de Euler, se pueden tomar más términos en la fórmula de Taylor para $y(x_{n+1})$ alrededor de $x = x_n$. Esto da lugar a los llamados *métodos de Taylor*[1]. Como ejemplo construimos el método de Taylor de orden 2 en el caso $d = 1$.

El desarrollo de Taylor produce

$$y(x_{n+1}) = y(x_n) + hy'(x_n) + \frac{h^2}{2}y''(x_n) + \frac{h^3}{3!}y'''(\xi_n).$$

[1]Fueron propuestos por Euler como un ejercicio en su "Institutionum Calculi Integralis" (1768).

Para calcular $y'(x_n)$ e $y''(x_n)$ usamos la ecuación diferencial,

$$y'(x) = f(x, y(x)),$$
$$y''(x) = f_x(x, y(x)) + f_y(x, y(x))y'(x) = f_x(x, y(x)) + f_y(x, y(x))f(x, y(x)).$$

Si despreciamos el residuo, $R_n = \frac{h^3}{3!}y'''(\xi_n)$, obtenemos el método

$$y_{n+1} = y_n + hf(x_n, y_n) + \frac{h^2}{2}(f_x(x_n, y_n) + f_y(x_n, y_n)f(x_n, y_n)).$$

Puesto que en este caso el residuo es más pequeño que para el método de Euler (si h es pequeño), esperamos que la aproximación sea mejor. A cambio hay que hacer tres evaluaciones de función por paso, mientras que con el método de Euler sólo hay que hacer una.

Por el mismo procedimiento podemos generar métodos de Taylor del orden que queramos. El inconveniente es que involucran cada vez más derivadas, que hay que calcular y evaluar.

Problemas

1. Construir el método de Taylor de orden 3 para $d = 1$ y comprobar que al aplicarlo al problema $y' = 2xy$, $y(0) = 1$, lleva a la iteración

$$y_{n+1} = y_n + h\big(2x_n y_n + y_n(1 + 2x_n^2)h + 2x_n y_n(3 + 2x_n^2)\frac{h^2}{3}\big).$$

2. Construir el método de Taylor de orden 2 para problemas de dimensión d arbitraria y elaborar un programa de Matlab que resuelva el problema (PVI) por medio de dicho método.

   Sintaxis: `y=taylororden2(ld,dxld,jyld,x,y0)`, siendo `dxld` el nombre del fichero que contiene la derivada con respecto a x de la función del lado derecho de la EDO en el punto (x, y) y `jyld` el nombre del fichero que contiene el jacobiano con respecto a y de la función del lado derecho de la EDO en el punto (x, y).

3.2. Métodos basados en fórmulas de cuadratura

Como vimos en el capítulo anterior, el método de Euler se obtiene a partir de la igualdad

$$y(x_{n+1}) = y(x_n) + \int_{x_n}^{x_{n+1}} f(x, y(x))\, dx, \tag{3.1}$$

usando la reglar rectangular izquierda para aproximar la integral. Si usamos otras fórmulas de cuadratura obtendremos nuevos métodos de integración. Veamos algunos ejemplos.

• *Regla del trapecio.* Si calculamos la integral (3.1) por medio de la regla del trapecio, que es de orden 2, en lugar de con la regla rectangular derecha, que es sólo de orden 1, se tiene que

$$y(x_{n+1}) - y(x_n) = \frac{h}{2}\big(f(x_n, y(x_n)) + f(x_{n+1}, y(x_{n+1}))\big) + R_n.$$

Despreciando el residuo R_n, se llega a la llamada *regla del trapecio*,

$$y_{n+1} = y_n + \frac{h}{2}\big(f(x_n, y_n) + f(x_{n+1}, y_{n+1})\big). \tag{3.2}$$

El método de Euler es *explícito*; el valor y_{n+1} viene dado explícitamente en términos del valor anterior y_n y se puede calcular fácilmente mediante una evaluación de f y dos operaciones aritméticas. Por el contrario, la regla del trapecio es un método *implícito*; para calcular y_{n+1} hay que resolver un sistema de ecuaciones no lineales, lo que en general es computacionalmente costoso.

Los valores y_{n+1} que resuelven el sistema (3.2) son puntos fijos de la aplicación

$$G(y) = y_n + \frac{h}{2}(f(x_n, y_n) + f(x_{n+1}, y)).$$

Para ver que hay uno (y sólo uno), al menos para h pequeño, usaremos el

Teorema de la Aplicación Contractiva[2] de Banach[3]. En efecto, la función de iteración satisface

$$\|G(y) - G(\hat{y})\| = \frac{h}{2}\|f(x_{n+1}, y) - f(x_{n+1}, \hat{y})\| \leq \frac{Lh}{2}\|y - \hat{y}\|,$$

que es contractiva si $h < 2/L$.

• *Predictor-corrector Euler/Trapecio.* Para evitar el coste computacional derivado del carácter implícito de la regla del trapecio, podemos sustituir el valor y_{n+1} en el lado derecho de (3.2) por una aproximación de ese valor dada por el método de Euler,

$$y^*_{n+1} = y_n + hf(x_n, y_n), \tag{3.3}$$

de manera que

$$y_{n+1} = y_n + \frac{h}{2}\left(f(x_n, y_n) + f(x_{n+1}, y^*_{n+1})\right). \tag{3.4}$$

Así pues, primero *predecimos* un valor de y_{n+1} mediante Euler y luego lo *corregimos* mediante la fórmula (3.4): se trata de un *par predictor-corrector*.

• *Euler modificado.* Si calculamos la integral de (3.1) por medio de la regla del punto medio, se tiene que

$$y(x_{n+1}) - y(x_n) \approx hf\left(x_n + \frac{h}{2}, y\left(x_n + \frac{h}{2}\right)\right).$$

Como no conocemos el valor $y(x_n + \frac{h}{2})$, lo aproximamos por el método de Euler. Llegamos finalmente a

$$y(x_{n+1}) \approx y(x_n) + hf\left(x_n + \frac{h}{2}, y(x_n) + \frac{h}{2}f(x_n, y(x_n))\right).$$

[2]Sea M un espacio métrico con métrica d. Una aplicación $f : M \to M$ es una *contracción* si existe una constante $K < 1$ tal que $d(f(x), f(y)) \leq Kd(x, y)$ para todo x, y en M.
Teorema de la Aplicación Contractiva: Sea M un espacio métrico completo. Si $f : M \to M$ es una contracción, entonces f tiene un único punto fijo en M.

[3]Stefan Banach (1892-1945), matemático polaco, fundador del análisis funcional moderno.

El método de Euler modificado se obtiene imponiendo que esta relación se satisfaga de forma exacta,

$$y_{n+1} = y_n + hf\left(x_n + \frac{h}{2}, y_n + \frac{h}{2}f(x_n, y_n)\right). \tag{3.5}$$

Se puede hacer más patente que este método requiere dos evaluaciones de función por paso escribiéndolo como

$$k_1 = f(x_n, y_n), \qquad k_2 = f\left(x_n + \frac{h}{2}, y_n + h\frac{k_1}{2}\right), \qquad y_{n+1} = y_n + hk_2. \tag{3.6}$$

Pertenece por tanto a la familia de *métodos de Runge*[4]*-Kutta*[5], que son aquellos que se pueden escribir en la forma

$$\begin{cases} k_i = f(x_n + c_i h, y_n + h\sum_{j=1}^{s} a_{ij}k_j), & i = 1, 2, \ldots, s, \\ y_{n+1} = y_n + h\sum_{i=1}^{s} b_i k_i. \end{cases}$$

Cada una de las evaluaciones de función k_i es una *etapa*. Estos métodos se representan de forma resumida por medio de su *tablero de Butcher*[6]:

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

[4]Carl David Tolmé Runge (1856–1927) matemático alemán. Comenzó su carrera dedicándose a la matemática pura. Sin embargo, tras conseguir una catedra en la Technische Hochschule de Hannover, centró sus investigaciones en la física, haciendo importantes contribuciones en el campo de la espectroscopía. Este cambio de intereses fue considerado por muchos de sus antiguos profesores y compañeros como una traición.

[5]Martin Wilhelm Kutta (1867–1944). Nacido en la actual Polonia, es conocido sobre todo por el llamado método de Runge-Kutta para la integración de EDO (que introdujo en su tesis doctoral) y por su contribuciones a la aerodinámica.

[6]John Charles Butcher (1933–), matemático australiano especializado en métodos numéricos para la resolución de EDOs.

• *Método leap-frog (regla del punto medio).* La "dificultad" en el método de Euler modificado reside en que hay que encontrar una aproximación a la solución en el punto $x_n+\frac{h}{2}$, que no pertenece a la malla, lo que exige una evaluación de función extra. Para evitarlo podemos integrar la EDO en (x_n, x_{n+2}). Al usar la regla del punto medio, se tiene ahora que

$$y(x_{n+2}) - y(x_n) = \int_{x_n}^{x_{n+2}} f(x, y(x))\, dx \approx 2hf(x_{n+1}, y(x_{n+1})),$$

lo que da lugar al método *leap-frog*, un método de dos pasos también conocido como regla del punto medio,

$$y_{n+2} = y_n + 2hf(x_{n+1}, y_{n+1}), \tag{3.7}$$

que sólo requiere una evaluación de función por paso.

Para calcular y_{n+2} la regla del punto medio utiliza *dos* valores anteriores, y_n e y_{n+1}, y no uno sólo, como los métodos vistos hasta ahora. Diremos que es un *método de dos pasos*. Necesitaremos por tanto dos valores de arranque, y_0 e y_1. Para y_0 tomamos un valor próximo al dato inicial, $y_0 \approx y(x_0)$. El valor de y_1 se tendrá que obtener por otro procedimiento, por ejemplo por el método de Euler.

• *Adams*[7]*-Bashforth*[8] *de dos pasos.* Si integramos la EDO en (x_{n+1}, x_{n+2}), se tiene que

$$y(x_{n+2}) - y(x_{n+1}) = \int_{x_{n+1}}^{x_{n+2}} f(x, y(x))\, dx. \tag{3.8}$$

Aproximamos la integral sustituyendo el integrando por el polinomio de grado menor o igual que uno que interpola a $f(x, y(x))$ en los nodos x_n y x_{n+1},

$$f(x, y(x)) \approx f(x_n, y(x_n))\frac{x - x_{n+1}}{x_n - x_{n+1}} + f(x_{n+1}, y(x_{n+1}))\frac{x - x_n}{x_{n+1} - x_n},$$

[7]John Couch Adams (1819–1892), matemático y astrónomo breitánico, una de las dos personas que de forma independiente descubrieron el planeta Neptuno, a partir de las irregularidades observadas en el movimiento de Urano. El otro descubridor fue el francés Urbain Jean Joseph Le Verrier (1811-1877).

[8]Francis Bashforth (1819–1912), matemático inglés experto en balística.

de forma que

$$y(x_{n+2}) - y(x_{n+1}) \approx \frac{h}{2}(3f(x_{n+1}, y(x_{n+1})) - f(x_n, y(x_n))).$$

Llegamos al método

$$y_{n+2} - y_{n+1} = \frac{h}{2}(3f(x_{n+1}, y_{n+1}) - f(x_n, y_n)),$$

conocido como *método de Adams-Bashforth de dos pasos*. Sólo requiere una evaluación de función por paso.

• *Adams-Moulton*[9] *de dos pasos.* Si sustituimos el integrando en (3.8) por el polinomio de grado menor o igual que dos que interpola a $f(x, y(x))$ en los nodos x_n, x_{n+1} y x_{n+2},

$$\begin{aligned} f(x, y(x)) \approx\ & f(x_n, y(x_n))\frac{(x - x_{n+1})(x - x_{n+2})}{(x_n - x_{n+1})(x_n - x_{n+2})} \\ & + f(x_{n+1}, y(x_{n+1}))\frac{(x - x_n)(x - x_{n+2})}{(x_{n+1} - x_n)(x_{n+1} - x_{n+2})} \\ & + f(x_{n+2}, y(x_{n+2}))\frac{(x - x_n)(x - x_{n+1})}{(x_{n+2} - x_n)(x_{n+2} - x_{n+1})}, \end{aligned}$$

llegamos a

$$y(x_{n+2}) - y(x_{n+1}) \approx \frac{h}{12}(5f(x_{n+2}, y(x_{n+2})) + 8f(x_{n+1}, y(x_{n+1})) - f(x_n, y(x_n))),$$

y finalmente al método implícito

$$y_{n+2} - y_{n+1} = \frac{h}{12}(5f(x_{n+2}, y_{n+2}) + 8f(x_{n+1}, y_{n+1}) - f(x_n, y_n)), \qquad (3.9)$$

conocido como *método de Adams-Moulton de dos pasos.*

Los métodos que, como la regla del punto medio, el método de Adams-Bashforth de dos pasos o el método de Adams-Moulton de dos pasos se pueden escribir en la forma

$$\sum_{j=0}^{k} \alpha_j y_{n+j} = h \sum_{j=0}^{k} \beta_j f(x_{n+j}, y_{n+j}), \qquad n = 0, \ldots, N - k,$$

[9]Forest Ray Moulton (1872-1952), astrónomo estadounidense especialmente interesado en las aplicaciones de las matemáticas a la astronomía. Tuvo un programa de divulgación científica en la radio que le hizo muy popular.

se conocen como métodos *lineales* de k-pasos, porque y_{n+k} es una combinación lineal de y_{n+j} y $f(x_{n+j}, y_{n+j})$, $j = 0, \ldots, k$, involucrando su cálculo a las aproximaciones en los k nodos anteriores. Estos métodos requieren k valores de arranque.

• *Predictor-corrector AB2/AM2.* Podemos combinar los dos métodos anteriores para obtener un nuevo método: en primer lugar *predecimos* un valor de y_{n+2} mediante el método de Adams-Bashforth de dos pasos y a continuación lo *corregimos* usando esta cantidad para calcular el lado derecho de (3.9),

$$y^*_{n+2} = y_{n+1} + \tfrac{h}{2}(3f(x_{n+1}, y_{n+1}) - f(x_n, y_n)),$$
$$y_{n+2} = y_{n+1} + \tfrac{h}{12}(5f(x_{n+2}, y^*_{n+2}) + 8f(x_{n+1}, y_{n+1}) - f(x_n, y_n)).$$

El método resultante, explícito y de dos pasos, es un *par predictor-corrector*. Requiere dos evaluaciones de función por paso.

Problemas

1. Escribir el par predictor-corrector Euler/Trapecio, (3.3)–(3.4), como un método de Runge-Kutta. Visto de esta forma se le conoce como método de Euler mejorado.
2. Consideramos el problema (PVI) con una f tal que $c^T f(x, y) = 0$ para un cierto vector columna c. Según vimos en un problema del capítulo anterior, en estas condiciones la solución del problema satisface la ley de conservación lineal $c^T y(x) = c^T \eta$ para todo $x \in [a, b]$. Estudiar si las soluciones obtenidas al aplicar al problema un método de Runge-Kutta satisfacen la misma ley de conservación lineal, $c^T y_n = c^T y_0$, $n = 0, \ldots, N$.
3. Si integramos la EDO en (x_n, x_{n+1}) y aplicamos la regla de Simpson, obtenemos que

   $$y(x_{n+1}) - y(x_n) \approx \frac{h}{6}(f(x_n, y(x_n)) + 4f(x_{n+\frac{1}{2}}, y(x_{n+\frac{1}{2}})) + f(x_{n+1}, y(x_{n+1}))),$$

   donde $x_{n+\frac{1}{2}} = x_n + \frac{h}{2}$. Como no conocemos el valor de $y(x_{n+\frac{1}{2}})$, aproximamos esta cantidad por Euler, $y(x_{n+\frac{1}{2}}) \approx y(x_n) + \frac{h}{2} f(x_n, y(x_n))$. Tampoco

conocemos el valor de $y(x_{n+1})$. Para estimarlo podemos avanzar a partir de x_n utilizando una media ponderada de las dos pendientes calculadas hasta ahora, $k_1 = f(x_n, y(x_n))$, $k_2 = f(x_{n+\frac{1}{2}}, y(x_n) + \frac{h}{2}k_1)$; es decir, tomamos $y(x_{n+1}) \approx y(x_n) + h(\alpha k_1 + \beta k_2)$. Llegamos así al método

$$\begin{cases} k_1 = f(x_n, y_n), \\ k_2 = f(x_n + \frac{h}{2}, y_n + \frac{h}{2}k_1), \\ k_3 = f(x_n + h, y_n + h(\alpha k_1 + \beta k_2)) \\ y_{n+1} = y_n + \frac{h}{6}(k_1 + 4k_2 + k_3). \end{cases}$$

Determinar α y β para que $\|R_n\| \leq Ch^4$ para alguna constante C.

4. Consideramos la igualdad

$$y(x_{n+k}) - y(x_{n+k-1}) = \int_{x_{n+k-1}}^{x_{n+k}} f(t, y(t))\, dt. \tag{3.10}$$

Si aproximamos la integral sustituyendo el integrando por el polinomio $p_{n,k}(t)$ de grado menor o igual que $k-1$ que interpola a $f(t, y(t))$ en los nodos $x_n, \ldots, x_{n+k-1}$, y pedimos que la aproximación y_n satisfaga la relación resultante exactamente, se obtiene el método de Adams-Bashforth de k pasos.

(a) Demostrar que el método de Adams-Bashforth de k pasos para mallas equiespaciadas viene dado por la recurrencia

$$y_{n+k} - y_{n+k-1} = h \sum_{j=0}^{k-1} \beta_{kj} f(x_{n+j}, y_{n+j}),$$

donde los coeficiente β_{kj} satisfacen

$$\beta_{kj} = \int_{k-1}^{k} \prod_{l=0,\ l\neq j}^{k-1} \frac{s-l}{j-l}\, ds.$$

(b) Usando la fórmula para el error de interpolación, obtener una expresión para el residuo del método,

$$\begin{aligned} R_n &= y(x_{n+k}) - y(x_{n+k-1}) - \int_{x_{n+k-1}}^{x_{n+k}} p_{n,k}(t)\, dt \\ &= \int_{x_{n+k-1}}^{x_{n+k}} (y'(t) - p_{n,k}(t))\, dt, \end{aligned}$$

que demuestre que $\|R_n\| \leq Ch^{k+1}$.

(c) Obtener el método de Adams-Bashforth de 2 pasos para mallas generales, no necesariamente equiespaciadas, y una expresión para su residuo.

5. Si aproximamos la integral de la igualdad (3.10) sustituyendo el integrando por el polinomio $\hat{p}_{n,k}(t)$ de grado menor o igual que k que interpola a $f(t, y(t))$ en los nodos $x_n, \ldots, x_{n+k}$, y pedimos que la aproximación y_n satisfaga la relación resultante exactamente, se obtiene el método de Adams-Moulton de k pasos.

(a) Demostrar que el método de Adams-Moulton de k pasos para mallas equiespaciadas viene dado por la recurrencia

$$y_{n+k} - y_{n+k-1} = h \sum_{j=0}^{k} \beta^*_{kj} f(x_{n+j}, y_{n+j}), \tag{3.11}$$

donde los coeficiente β_{kj} satisfacen

$$\beta^*_{kj} = \int_{k-1}^{k} \prod_{l=0,\ l\neq j}^{k} \frac{s-l}{j-l}\, ds.$$

(b) Usando la fórmula para el error de interpolación, obtener una expresión para el residuo del método,

$$R_n = \int_{x_{n+k-1}}^{x_{n+k}} (y'(t) - \hat{p}_{n,k}(t))\, dt,$$

que demuestre que $\|R_n\| \leq Ch^{k+2}$.

(c) Dar una condición sobre h que garantice que el sistema (3.11) tiene una única solución y_{n+k}.

6. Consideramos la igualdad

$$y(x_{n+k}) - y(x_{n+k-2}) = \int_{x_{n+k-2}}^{x_{n+k}} f(t, y(t))\, dt.$$

Si aproximamos la integral sustituyendo el integrando por el polinomio de grado menor o igual que $k-1$ que interpola a $f(t, y(t))$ en los nodos $x_n, \ldots, x_{n+k-1}$, y pedimos que la aproximación y_n satisfaga la relación resultante exactamente, se obtiene el método de Nyström[10] de k pasos.

(a) Obtener una fórmula para el método de Nyström de k pasos.

(b) Usar la fórmula para el error de interpolación para demostrar que existe una constante C tal que el residuo del método de Nyström de k pasos satisface $\|R_n\| \leq Ch^k$.

(c) Demostrar, usando desarrollos de Taylor, que se tiene algo mejor: $\|R_n\| \leq Ch^{k+1}$.

7. Programar funciones de Matlab que calculen aproximaciones numéricas para el problema (PVI) por medio de los siguientes métodos:

(a) El método de Euler mejorado.

(b) El método de Euler modificado.

(c) El método de Adams-Bashforth de 2 pasos (para mallas uniformes). Hay que programar el método de manera que en cada paso (salvo el primero) sólo se haga una evaluación de función. El segundo valor de arranque, y_1, se calculará mediante el método de Euler.

La estructura de las variables de entrada y salida será la misma que la del programa `euler` presentado en el capítulo 2.

[10]Evert Johannes Nyström (1895–1960), matemático finés conocido por sus métodos numéricos para ecuaciones integrales y para problemas de valor inicial para ecuaciones de segundo orden.

3.3. Métodos basados en fórmulas de diferenciación numérica

Como hemos visto, al aproximar la derivada $y'(x)$ mediante la *diferencia progresiva*

$$y'(x) \approx \frac{y(x+h) - y(x)}{h},$$

se obtiene el método de Euler. Al usar otras fórmulas para aproximar la derivada obtenemos otros métodos. Veamos algunos ejemplos.

• *Método de Euler implícito.* Si aproximamos $y'(x)$ mediante la *diferencia regresiva*

$$y'(x) \approx \frac{y(x) - y(x-h)}{h},$$

obtenemos que

$$f(x_{n+1}, y(x_{n+1})) = y'(x_{n+1}) \approx \frac{y(x_{n+1}) - y(x_n)}{h},$$

lo que da lugar a la recurrencia

$$y_{n+1} = y_n + hf(x_{n+1}, y_{n+1}), \tag{3.12}$$

un método conocido como *método de Euler implícito.*

Es fácil ver, usando la fórmula de Taylor, que tanto para la diferencia progresiva como para la regresiva el error es del mismo orden,

$$y'(x) - \frac{y(x+h) - y(x)}{h} = -\frac{h}{2}y''(\bar{\xi}), \qquad \xi \in [x, x+h],$$
$$y'(x) - \frac{y(x) - y(x-h)}{h} = \frac{h}{2}y''(\bar{\eta}), \qquad \eta \in [x-h, x].$$

Sin embargo, el método de Euler implícito es, como su nombre indica, *implícito.* Por consiguiente, dar un paso con este método es mucho más costoso que darlo con el método de Euler. A pesar de ello, como veremos más adelante, en ocasiones el método de Euler implícito puede ser más ventajoso, pues aunque cada paso sea más costoso, quizá haya que dar muchos menos pasos.

• *Regla del punto medio (método leap-frog).* Si aproximamos la derivada $y'(x)$ mediante la diferencia central

$$y'(x) \approx \frac{y(x+h) - y(x-h)}{2h},$$

obtenemos que

$$f(x_{n+1}, y(x_{n+1})) = y'(x_{n+1}) \approx \frac{y(x_{n+2}) - y(x_n)}{2h},$$

lo que da lugar a la regla del punto medio (3.7), que ya hemos visto anteriormente. Puesto que este método proviene de una regla de derivación numérica de orden 2, podríamos pensar que es superior al método de Euler. Sin embargo, como veremos, este método no funciona correctamente.

Problemas

1. Obtener, por medio del desarrollo de Taylor, una aproximación de orden 2 de la derivada $y'(x)$ mediante una combinación lineal de $y(x-2h)$, $y(x-h)$ e $y(x)$ (fórmula de diferenciación regresiva de tres puntos). Utilizarla para construir un método de dos pasos para resolver el PVI. Obtener una expresión para el residuo. El método construido se conoce como fórmula BDF ("backward differentiation formulae") de dos pasos.
2. Obtener una aproximación de orden 4 de la derivada $y'(x)$ a partir de los valores $y(x+2h)$, $y(x+h)$, $y(x)$, $y(x-h)$ e $y(x-2h)$ (fórmula de diferenciación centrada de cinco puntos). Utilizarla para construir un método de cuatro pasos para resolver problemas de valor inicial para EDOs. Obtener una expresión para el residuo.
3. Consideramos el polinomio $p_1(x)$ de grado menor o igual que uno que pasa por los puntos $(x_n, y(x_n))$, $(x_{n+1}, y(x_{n+1}))$. Se sabe que

   $$y'(x) = p_1'(x) + O(h), \qquad x \in [x_n, x_{n+1}].$$

   Utilizar esta idea para construir un método numérico de un paso para resolver el PVI.

3.4. Métodos de colocación

La idea de los métodos de colocación es aproximar la solución y del PVI por una función u perteneciente a un subespacio de dimensión finita de $C^1([a,b])$. La función u se determina imponiendo que pase por ciertos puntos y que satisfaga la ecuación en otros (los llamados *puntos de colocación*). Veamos un par de ejemplos.

• *Regla implícita del punto medio.* Buscamos un polinomio u de grado 1 tal que

$$u(x_n) = y_n, \qquad u'\left(x_n + \frac{h}{2}\right) = f\left(x_n + \frac{h}{2}, u\left(x_n + \frac{h}{2}\right)\right);$$

es decir, pedimos que u tome un cierto dato inicial y que sea solución de la EDO en el *punto de colocación* $x = x_n + \frac{h}{2}$. Puesto que u' es constante, tenemos que $u'(x) = k_1 \equiv f\left(x_n + \frac{h}{2}, u\left(x_n + \frac{h}{2}\right)\right)$, que integrado en (x_n, x) produce

$$u(x) = y_n + k_1(x - x_n).$$

Si definimos $y_{n+1} = u(x_{n+1})$, se tiene que

$$y_{n+1} = y_n + hk_1, \qquad k_1 = f\left(x_n + \frac{h}{2}, y_n + h\frac{k_1}{2}\right). \tag{3.13}$$

Se trata por tanto de un método de Runge-Kutta de tablero

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

Es fácil comprobar que el método se puede escribir en la forma

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

expresión de la que procede el nombre del método.

• *Fórmula BDF de 2 pasos.* Dadas y_n e y_{n+1} queremos obtener y_{n+2}. Consideramos el polinomio $Q_{n,2}$ de grado menor o igual que 2 que interpola a y_n,

y_{n+1} e y_{n+2}, es decir

$$Q_{n,2}(x) = y_n \frac{(x-x_{n+1})(x-x_{n+2})}{(x_n-x_{n+1})(x_n-x_{n+2})} + y_{n+1} \frac{(x-x_n)(x-x_{n+2})}{(x_{n+1}-x_n)(x_{n+1}-x_{n+2})} + y_{n+2} \frac{(x-x_n)(x-x_{n+1})}{(x_{n+2}-x_n)(x_{n+2}-x_{n+1})}.$$

Este polinomio no se puede construir, porque desconocemos y_{n+2}. Se impone ahora que $Q_{n,2}$ satisfaga la ecuación diferencial en x_{n+2}, esto es

$$Q'_{n,2}(x_{n+2}) = f(x_{n+2}, Q_{n,2}(x_{n+2})) = f(x_{n+2}, y_{n+2}). \tag{3.14}$$

El método resultante,

$$y_{n+2} - \frac{4}{3}y_{n+1} + \frac{1}{3}y_n = \frac{2}{3}hf(x_{n+2}, y_{n+2}), \tag{3.15}$$

es un método lineal implícito y de dos pasos que se conoce como *fórmula BDF de 2 pasos*.

Problemas

1. Dados ν *parámetros de colocación* $c_1, \ldots, c_\nu$ (preferiblemente en $[0,1]$, aunque no es imprescindible), buscamos un polinomio u de grado ν (con coeficientes vectoriales) tal que

   $$u(x_n) = y_n, \qquad u'(x_n + c_j h) = f(x_n + c_j h, u(x_n + c_j h)),\ j = 1, \ldots, \nu.$$

   Al calcular u y tomar $y_{n+1} = u(x_{n+1})$ se obtiene un método de colocación que pertenece a la familia de los métodos de Runge-Kutta.
   Obtener el método de Runge-Kutta de colocación que corresponde a los parámetros de colocación $c_1 = 1/3$, $c_2 = 1$ (este método se conoce como método de Radau IIA de orden 3).
2. Comprobar que la condición (3.14) conduce a la recurrencia (3.15).
3. Dadas las aproximaciones $y_n, \ldots, y_{n+k-1}$, consideramos el polinomio $Q_{n,k}$ de grado menor o igual que k que interpola a $(x_n, y_n), \ldots, (x_{n+k}, y_{n+k})$.

Obviamente este polinomio no se puede construir, porque desconocemos y_{n+k}. Se impone ahora que $Q_{n,k}$ satisfaga la ecuación diferencial en x_{n+k}, esto es

$$Q'_{n,k}(x_{n+k}) = f(x_{n+k}, Q_{n,k}(x_{n+k})) = f(x_{n+k}, y_{n+k}).$$

El método resultante se conoce como *fórmula BDF de k pasos* (del inglés "backward differentiation formulae").

(a) Obtener la recurrencia para la fórmula BDF de k pasos.

(b) Obtener, usando la fórmula para el error de interpolación, una fórmula para el residuo de la fórmula BDF de k pasos que demuestre que $\|R_n\| \leq Ch^{k+1}$.

3.5. Programación de un método implícito: la regla del trapecio

Por ser la regla del trapecio un método implícito, para obtener la sucesión de valores aproximados $\{y_n\}_{n=0}^N$ en cada paso tendremos que resolver un sistema no lineal de ecuaciones. La solución del sistema, y_{n+1}, es, como ya dijimos, un punto fijo de la aplicación

$$G(y) = y_n + \frac{h}{2}(f(x_n, y_n) + f(x_{n+1}, y)),$$

aplicación que es contractiva si $h < 2/L$. Así pues, si h satisface esa restricción, no sólo tendremos garantizado que la solución del sistema es única, sino que además se podrá obtener como límite de la sucesión $\{y_{n+1}^{[k]}\}_{k=0}^{\infty}$ definida por medio iteración

$$y_{n+1}^{[k+1]} = G(y_{n+1}^{[k]}).$$

Como valor de arranque podemos tomar $y_{n+1}^{[0]} = y_n + hf(x_n, y_n)$, valor que estará próximo a y_{n+1} si h es pequeño. Nótese que esto no supone ninguna

evaluación de función extra, pues $f(x_n, y_n)$ se debe calcular en cualquier caso para hacer la iteración de punto fijo.

Estas ideas se pueden poner en práctica de forma bastante sencilla, como se puede ver en el programa 3.1.

```
function [y,evf]=trapeciopf(ld,x,y0,itm,tol);

% ld: Nombre del fichero que contiene la función que calcula el lado derecho de la
%     EDO. Tendrá siempre dos argumentos de entrada, x (escalar) e y (vector
%     columna). La salida será un vector columna.
% x:  Malla (vector fila de longitud N).
% y0: Dato inicial (vector columna de tamaño d, la dimensión del sistema de edos).
% itm: Número máximo de iteraciones en la iteración de punto fijo.
% tol: Tolerancia para la iteración de punto fijo.
% y:   Solución numérica en los puntos de la malla (matriz de tamaño d x N).
% evf: Número de evaluaciones de función llevadas a cabo por el método.

N=length(x)-1;
evf=0;
y(:,1)=y0;
for n=1:N, % Bucle que recorre la discretizacion
    numit=0;
    errorpf=tol+1;
    hn=(x(n+1)-x(n));
    fn=feval(ld,x(n),y(:,n));
    evf=evf+1;
    yk=y(:,n)+hn*fn;
    while (numit<itm & errorpf>tol) % Iteración de punto fijo
        numit=numit+1;
        ykmas1=y(:,n)+.5*hn*(fn+feval(ld,x(n+1),yk));
        evf=evf+1;
        errorpf=norm(ykmas1-yk,inf);
        yk=ykmas1;
    end
    if errorpf<=tol
        y(:,n+1)=yk;
    else
        error('la iteracion de punto fijo no converge')
    end
end
```

Programa 3.1: Regla del trapecio con iteración de punto fijo.

Para estimar el error en la iteración de punto fijo hemos usado la diferencia con la iteración anterior. Como es bien sabido, ésta no es necesariamente una buena estimación del error. En los problemas del final de la sección sugerimos como obtener una estimación mejor.

Es importante recalcar que no se han hecho evaluaciones de función innecesarias. En particular, el valor de $f(x_n, y_n)$ sólo se calcula una vez por paso. Por otra parte, con el fin de poder analizar la eficiencia del método, se ha incluido como variable de salida el número de evaluaciones de función realizadas.

Si $L \gg 1$, la restricción $h < 2/L$ obliga a dar pasos de longitud h muy pequeña. Para evitarlo lo que se hace es determinar y_{n+1} como un cero de la función

$$F(y) = y - \{y_n + \frac{h}{2}(f(x_n, y_n) + f(x_{n+1}, y))\}$$

mediante el método de Newton. El cero vendrá dado como límite de la iteración

$$y_{n+1}^{[k+1]} = y_{n+1}^{[k]} - \left(DF(y_{n+1}^{[k]})\right)^{-1} F(y_{n+1}^{[k]}). \tag{3.16}$$

Invertir matrices es computacionalmente muy costoso. Así que, en lugar de usar directamente la fórmula (3.16), es mucho mejor resolver el sistema

$$DF(y_{n+1}^{[k]})\texttt{correccion} = -F(y_{n+1}^{[k]}),$$

que producirá $\texttt{correcion} = y_{n+1}^{[k+1]} - y_{n+1}^{[k]}$. Aparte del ahorro computacional, matamos dos pajaros de un tiro: además de $y_{n+1}^{[k+1]}$, que se obtiene mediante la fórmula $y_{n+1}^{[k+1]} = y_{n+1}^{[k]} + \texttt{correccion}$, tenemos un criterio de parada para la iteración del método de Newton, ya que `correccion` es una buena estimación para el error cometido, $y_{n+1} - y_{n+1}^{[k]} \approx \texttt{correccion}$.

El jacobiano de la función $F(y)$ se obtiene a partir del jacobiano con respecto a la variable y de la función f del lado derecho por medio de la fórmula

$$DF(y) = I - \frac{h}{2} D_y f(x_{n+1}, y).$$

Como iterante inicial tomaremos una aproximación de y_{n+1} obtenida por el método de Euler, $y_{n+1}^{[0]} = y_n + hf(x_n, y_n)$, que no supone ninguna evaluación de función extra, pues $f(x_n, y_n)$ se tiene que calcular en cualquier caso para hacer las iteraciones de Newton.

El programa 3.2 recoge las observaciones anteriores. Como siempre, tendremos buen cuidado de no hacer evaluaciones de función innecesarias.

```
function [y,evf,evj]=trapecionw(ld,jyld,x,y0,itm,tol);

 % ld: Nombre del fichero que contiene la función que calcula el lado derecho de la
 %     EDO. Tendrá siempre dos argumentos de entrada, x (escalar) e y (vector
 %     columna). La salida será un vector columna.
 % jyld: Nombre del fichero que contiene el jacobiano con respecto a y de la función
 %     del lado derecho de la EDO. Tendrá siempre dos argumentos de entrada, x
 %     (escalar) e y (vector columna). La salida será una matriz de tamaño d x d.
 % x:  Malla (vector fila de longitud N).
 % y0: Dato inicial (vector columna de tamaño d, la dimensión del sistema de edos).
 % itm: Número máximo de iteraciones en la iteración de Newton.
 % tol: Tolerancia para la iteración de Newton.
 % y:   Solución numérica en los puntos de la malla (matriz de tamaño d x N).
 % evf: Número de evaluaciones de la función ld llevadas a cabo por el método.
 % evj: Número de evaluaciones del jacobiano jyld llevadas a cabo por el método.

 N=length(x)-1; evf=0; evj=0; I=eye(length(y0));
 y(:,1)=y0;
 for n=1:N, % Bucle para recorrer la discretizacion
     numit=0;
     enewton=tol+1;
     fn=feval(ld,x(n),y(:,n));
     evf=evf+1;
     yk=y(:,n)+(x(n+1)-x(n))*fn;
     while (numit<itm & enewton>tol) % Bucle del Newton
         numit=numit+1;
         % g es la funcion de la que hay que encontrar el cero
         g=yk-y(:,n)-.5*(x(n+1)-x(n))*(fn+feval(ld,x(n+1),yk));
         evf=evf+1;
         jacg=I-.5*(x(n+1)-x(n))*feval(jyld,x(n+1),yk);
         evj=evj+1;
         correccion=-jacg\g;
         enewton=norm(correccion,inf);
         yk=yk+correccion;
     end
     if enewton<=tol
         y(:,n+1)=yk; %incluyo la ultima correccion
     else
         error('newton no converge')
     end
 end
```

Programa 3.2: Regla del trapecio con Newton.

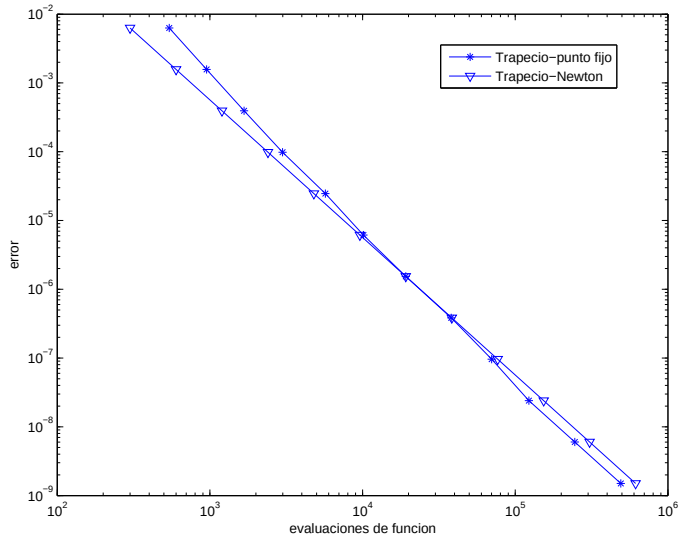

Figura 3.1: Diagrama de eficiencia de dos versiones de la regla del trapecio aplicadas al Problema 1.

En la figura 3.1 comparamos los diagramas de eficiencia correspondientes a aplicar la regla del trapecio al problema (2.8) empleando los programas 3.1 y 3.2. Para evaluar el coste sumaremos las evaluaciones de la función y del jacobiano.

Cuando el número de pasos es pequeño (y por tanto h es grande), aplicar Newton resulta ser más conveniente que aplicar iteración de punto fijo, aunque en ambos casos haya convergencia. La razón reside en que el punto fijo converge con orden 1, con un factor de contracción para el error dado por $Lh/2$. Así que si h es grande, la convergencia va a ser bastante lenta. A esto hay que añadir que el iterante inicial no va a ser demasiado bueno. Por el contrario, para h pequeño la iteración de punto fijo resulta ser superior: la velocidad de convergencia aumenta y además el iterante inicial es una aproximación mucho mejor.

Problemas

1. El uso del método de Newton en la Regla del Trapecio requiere evaluar la matriz jacobiana $D_y f$. Esta función debe ser suministrada al integrador. Sin embargo, en las aplicaciones prácticas especificar estas derivadas parciales analíticamente es con frecuencia una tarea difícil o engorrosa. Una alternativa es usar diferencias aproximadas: en $y = y_n^{[k]}$ se evalúan $\hat{f} = f(x_n, \hat{y})$ y $\tilde{f} = f(x_n, \tilde{y})$, donde $\hat{y}$ y $\tilde{y}$ son perturbaciones de y en una coordenada, $\hat{y}_j = y_j + \epsilon$, $\tilde{y}_j = y_j - \epsilon$, $\hat{y}_l = \tilde{y}_l = y_l$, $l \neq j$. Entonces, la j-ésima columna de $D_y f$ se puede aproximar por

$$D_y f \approx \frac{1}{2\epsilon}(\hat{f} - \tilde{f}),$$

siendo ϵ un parámetro positivo pequeño.

Este truco, fácil de programar, suele funcionar bien con la elección $\epsilon = 10^{-\gamma}$, si se trabaja con aritmética de coma flotante con aproximadamente 2γ dígitos significativos (en el caso de Matlab, $\gamma = 7$).

Observación. Esta aproximación de la matriz jacobiana puede resultar costosa en algunos casos y no siempre funciona bien.

   (a) Modificar el programa `trapecionw` que se os ha proporcionado de forma que use una aproximación del jacobiano en lugar de la función jacobiana suministrada por el usuario.

   (b) Otra alternativa para aproximar el jacobiano numéricamente es usar la función de Matlab `numjac`. Modificar el programa `trapecionw` de manera que aproxime el jacobiano por medio de esta función.

2. El programa `trapeciopf` que se os ha suministrado estima el error en la iteración de punto fijo como la diferencia entre las dos últimas iteraciones. Esta estimación con frecuencia sobreestima o subestima el error. Un criterio mejor es tomar

$$\frac{\lambda}{1-\lambda}\|y_n^{[k+1]} - y_n^{[k]}\| \leq \texttt{TOL},$$

donde `TOL` es la tolerancia para la iteración de punto fijo proporcionada por el usuario y λ es una medida de la velocidad de convergencia de la

iteración, que se puede estimar por

$$\lambda = \left(\frac{\|y_n^{[k+1]} - y_n^{[k]}\|}{\|y_n^{[1]} - y_n^{[0]}\|} \right)^{1/k}.$$

(a) Modificar el programa `trapeciopf` de manera que use esta estimación mejorada del error de punto fijo.

(b) Elaborar un diagrama que mida la eficiencia de ambos métodos (el original y el modificado) al aplicarlos al problema del péndulo.

3. Las evaluaciones del jacobiano en el método de Newton pueden ser costosas, y también lo es la resolución de los sistemas lineales resultantes. Por ello, se suele usar un método de Newton modificado, que mantiene fijo el jacobiano, $D_y f \approx D_y f(x_{n+1}, y_{n+1}^{[0]})$. De esta manera sólo se evalúa el jacobiano una vez. Y lo que es aún mejor, se hace una (única) descomposición LU (lo que conlleva $\frac{d^3}{3} + O(d^2)$ operaciones), que se emplea en todos los pasos de la iteración para resolver los sistemas lineales (el coste es $O(d^2)$).

   - Modificar el programa `trapecionw` de forma que recoja las ideas que acabamos de explicar.
   - Elaborar un diagrama que mida la eficiencia de ambos métodos (el original y el modificado) al aplicarlos al problema del péndulo.

4. Si integramos la EDO $y'(x) = f(x, y(x))$ en el intervalo (x_n, x_{n+1}) se tiene que $y(x_{n+1}) = y(x_n) + \int_{x_n}^{x_{n+1}} f(s, y(s))\, ds$. Si aproximamos el integrando por medio de la regla rectangular derecha, llegamos al llamado *método de Euler implícito*,

$$y_{n+1} = y_n + h f(x_{n+1}, y_{n+1}).$$

Programar una función de Matlab que calcule soluciones numéricas para sistemas por este método. Los sistemas no lineales resultantes se resolverán, bien por punto fijo (nombre de la función: `eulerimplicitopf`),

bien por el método de Newton (nombre de la función: `eulerimplicitonw`). La estructura de las variables de entrada y salida será la misma que la de las funciones `trapeciopf` y `trapecionw` que se os han suministrado.

5. Repetir el ejercicio anterior para la regla implícita del punto medio. Los nombres de los programas serán `rkipf` y `rkinw`.

6. Para dibujar un círculo de radio r en una pantalla gráfica, uno puede evaluar pares $x = r\cos\theta$, $y = r\sin\theta$ para una sucesión de valores de θ. Pero esto es costoso. Un método alternativo más barato consiste en considerar el PVI

$$x'(\theta) = -y(\theta), \quad y'(\theta) = x(\theta), \qquad x(0) = r, \quad y(0) = 0,$$

y aproximarlo utilizando algún método numérico. Sin embargo, hay que asegurarse de que la solución obtenida tiene el aspecto deseado.

Aplicar el método de Euler, el método de Euler implícito y la Regla del trapecio al problema, con paso $h = 0{,}02$, para $0 \le \theta \le 120$. Determinar si la solución forma una espiral hacia fuera, una espiral hacia dentro o, como se desea, un círculo aproximado. Explicar los resultados observados.

Sugerencia. Esto tiene que ver con una cierta función invariante de x e y, más que con el orden de los métodos.

7. El sistema de EDOs

$$(y^1)' = \alpha - y^1 - \frac{4y^1y^2}{1+(y^1)^2}, \quad (y^2)' = \beta y^1\left(1 - \frac{y^2}{1+(y^1)^2}\right),$$

donde α y β son parámetros, representa de forma simplificada a cierta reacción química. Para cada valor del parámetro α hay un valor crítico del parámetro β, $\beta_c = \frac{3\alpha}{5} - \frac{25}{\alpha}$, tal que para $\beta > \beta_c$ las trayectorias de las soluciones decaen en amplitud y se acercan en el plano de fases en forma espiral a un punto fijo estable, mientras que para $\beta < \beta_c$ las trayectorias oscilan sin amortiguarse y se ven atraídas por un ciclo límite estable (esto es lo que se conoce como una bifurcación de Hopf).

(a) Fijamos $\alpha = 10$. Usar cualquiera de los métodos que se han programado hasta el momento con longitud de paso fija $h = 0{,}01$ para aproximar la solución del problema que empieza en $y^1(0) = 0$, $y^2(0) = 2$, para $t \in [0, 20]$, con $\beta = 2$ y $\beta = 4$. En cada caso dibujar y^1 e y^2 frente a t. Describir lo observado.

(b) Investigar cuál es la situación cerca del valor crítico $\beta_c = 3{,}5$ (quizá convenga incrementar la longitud del intervalo de integración para ver mejor lo que sucede).

Capítulo 4

El Teorema de Equivalencia

La demostración de la convergencia del método de Euler que hemos visto en el capítulo 2 tiene dos ingredientes: la estabilidad del método ante perturbaciones y un control del residuo (consistencia). A lo largo de este capítulo veremos que estos son los elementos necesarios y suficientes para la convergencia de la mayoría de los métodos.

4.1. Función de incremento

Todos los métodos del capítulo anterior, y prácticamente todos los que se usan habitualmente, se pueden escribir en la forma

$$\sum_{j=0}^{k} \alpha_j y_{n+j} = h\phi_f(x_n, y_n, \ldots, y_{n+k-1}; h), \qquad n = 0, \ldots, N-k, \qquad \text{(MN)}$$

donde ϕ_f, la *función de incremento*, depende de sus argumentos a través de la función f (y a veces, como en los métodos de Taylor, también de sus derivadas).

Ejemplos. (a) Para el método de Euler $\phi_f(x_n, y_n; h) = f(x_n, y_n)$.

(b) Para la regla del trapecio $\phi_f(x_n, y_n; h)$ viene definida implícitamente por

$$\phi_f(x_n, y_n; h) = \frac{f(x_n, y_n) + f(x_n + h, y_n + h\phi_f(x_n, y_n; h))}{2}. \tag{4.1}$$

Veamos que esta ecuación tiene una única solución para h suficientemente pequeño. El valor de la función de incremento es un punto fijo de la función

$$F(\phi) = \frac{f(x_n, y_n) + f(x_n + h, y_n + h\phi)}{2}. \tag{4.2}$$

Usando que f es Lipschitz con respecto a su segunda variable, con constante L, se tiene que $\|F(\phi) - F(\hat{\phi})\| \leq \frac{Lh}{2}\|\phi - \hat{\phi}\|$. Así pues, si $\frac{Lh}{2} < 1$, la aplicación F es contractiva, y tiene por consiguiente un único punto fijo. En resumen, $\phi_f(x_n, y_n; h)$ está bien definida para todo $h < 2/L$. ♣

Si la función de incremento ϕ_f se puede expresar explícitamente en términos de f (y tal vez de sus derivadas), el método es *explícito*. En caso contrario, el método es *implícito*.

El método (MN) se dice de k *pasos*, pues se necesitan k valores anteriores, $y_n, \ldots, y_{n+k-1}$, para calcular y_{n+k}. Por consiguiente, es necesario disponer de k *valores de arranque*, $y_0, \ldots, y_{k-1}$; sin embargo, el problema de valor inicial sólo proporciona y_0. Si $k \geq 2$ habrá que obtener los $k - 1$ valores de arranque restantes por algún otro procedimiento, como puede ser por desarrollo de Taylor, utilizando un método de un sólo paso, etc.

Supondremos que $\alpha_k = 1$, eliminando así la arbitrariedad que surge del hecho de que podemos multiplicar ambos lados de (MN) por un mismo número distinto de cero sin cambiar el método. Supondremos también que, o bien $\alpha_0 \neq 0$, o bien ϕ_f depende de y_n de forma no trivial. Excluimos así métodos como por ejemplo

$$y_{n+2} - y_{n+1} = hf(x_{n+1}, y_{n+1}),$$

que es esencialmente de 1 paso, y no de 2, y que se puede escribir como

$$y_{n+1} - y_n = hf(x_n, y_n).$$

Para todos los ejemplos del capítulo anterior ϕ_f satisface las siguientes propiedades:

$$\left.\begin{array}{ll} \text{(i)} & \text{es continua;} \\ \text{(ii)} & \text{existen constantes } h_0 \text{ y } L \text{ tales que} \\ & \|\phi_f(x_n, y_n, \ldots, y_{n+k-1}; h) - \phi_f(x_n, \hat{y}_n, \ldots, \hat{y}_{n+k-1}; h)\| \\ & \qquad \leq L \sum_{j=0}^{k-1} \|y_{n+j} - \hat{y}_{n+j}\| \quad \text{si } 0 < h < h_0; \\ \text{(iii)} & \text{si } f = 0, \text{ entonces } \phi_f = 0. \end{array}\right\} \qquad (\mathrm{H_{MN}})$$

En lo sucesivo nos restringiremos a métodos de la forma (MN) que verifiquen las hipótesis ($\mathrm{H_{MN}}$).

Ejemplo. Veamos que la función de incremento de la regla del trapecio satisface las hipótesis ($\mathrm{H_{MN}}$). Sea L_f una constante de Lipschitz para f con respecto a su segunda variable.

(i) Como ya hemos explicado, el valor de la función de incremento se puede obtener por iteración de punto fijo,

$$\phi_f(x_n, y_n; h) = \lim_{k\to\infty} \phi_f^{[k]}(x_n, y_n; h), \qquad \phi_f^{[k]}(x_n, y_n; h) = F(\phi_f^{[k-1]}(x_n, y_n; h)),$$

con la función de iteración dada en (4.2). Si $h < 2/L_f$, condición que supondremos en adelante, hay un único punto fijo y la iteración converge a él sea cual sea el iterante inicial.

Por ser f una función continua, si $\phi_f^{[k-1]}$ es una función continua, la función $\phi_f^{[k]} = F(\phi_f^{[k-1]})$ también lo será, pues es composición de funciones continuas.

Tomamos como iterante inicial $\phi_f^{[0]} = 0$. Todas las funciones $\phi_f^{[k]}$ son entonces continuas, pero el límite no tiene por qué serlo. Sí lo será si la convergencia es uniforme.

Observamos que $\phi_f^{[k]} = \sum_{j=1}^{k} (\phi_f^{[j]} - \phi_f^{[j-1]})$. Si la serie converge uniformemente sobre compactos, la sucesión $\{\phi_f^{[k]}\}_{k=1}^{\infty}$ también lo hará, y el límite será una función continua. Ahora bien,

$$\|\phi_f^{[j]} - \phi_f^{[j-1]}\| = \|F(\phi_f^{[j-1]}) - F(\phi_f^{[j-2]})\| \leq \frac{L_f h}{2} \|\phi_f^{[j-1]} - \phi_f^{[j-2]}\|.$$

Iterando está relación, concluimos que $\|\phi_f^{[j]} - \phi_f^{[j-1]}\| \leq \left(\frac{L_f h}{2}\right)^{j-1} \|\phi_f^{[1]}\|$. Pero

$$\phi_f^{[1]}(x_n, y_n; h) = \frac{f(x_n, y_n) + f(x_n + h, y_n)}{2}.$$

Por ser f continua, está acotada sobre compactos. Así pues, para cada subconjunto compacto K de $[a, b] \times \mathbb{R}^d \times (0, \infty)$, existe una constante M_K tal que $\|\phi_f^{[1]}(x_n, y_n; h)| \leq M_K$ si $(x_n, y_n; h) \in K$. Concluimos que sobre cada compacto $\|\phi_f^{[j]} - \phi_f^{[j-1]}\| \leq M_K \left(\frac{L_f h}{2}\right)^{j-1}$. Como la serie $\sum_{j=1}^{\infty} \left(\frac{L_f h}{2}\right)^j$ es convergente, tenemos, por el criterio M de Weierstrass, que la convergencia es uniforme sobre K. Esto implica que ϕ_f es continua sobre K. Como el compacto K es arbitrario, concluimos que ϕ_f es una función continua de sus argumentos si $h < 2/L_f$.

(ii) Aplicando la desigualdad triangular y la condición de Lipschitz para f,

$$\begin{aligned}&\|\phi_f(x_n, y_n; h) - \phi_f(x_n, \hat{y}_n; h)\| \\ &\quad \leq \tfrac{L_f}{2}\|y_n - \hat{y}_n\| + \tfrac{L_f}{2}(\|y_n - \hat{y}_n\| + h\|\phi_f(x_n, y_n; h) - \phi_f(x_n, \hat{y}_n; h)\|).\end{aligned}$$

Por consiguiente, para cualquier $h_0 < 2/L_f$ se tiene que

$$\|\phi_f(x_n, y_n; h) - \phi(x_n, \hat{y}_n; h)\| \leq \frac{L_f}{1 - \frac{L_f h_0}{2}}\|y_n - \hat{y}_n\|, \qquad 0 < h < h_0.$$

Así pues, se tiene la hipótesis ($\mathrm{H_{MN}}$)-(ii) con $L = L_f/(1 - \frac{L_f h_0}{2})$.

(iii) La hipótesis ($\mathrm{H_{MN}}$)-(iii) se obtiene inmediatamente de la ecuación (4.1) que define ϕ_f implícitamente. ♣

Problemas

1. Para cada uno de los métodos explícitos siguientes, demostrar que si f es continua en D y Lipschitz con respecto a su segunda variable en D, entonces la correspondiente función de incremento satisface las hipótesis ($\mathrm{H_{MN}}$).

(a) Predictor-corrector Euler/Trapecio, (3.3)–(3.4).

(b) Euler modificado, (3.5).

(c) Regla del punto medio, (3.7).

2. Para cada uno de los métodos implícitos siguientes, demostrar que si f es continua en D y Lipschitz con respecto a su segunda variable en D, entonces la correspondiente función de incremento está bien definida para h pequeño y satisface las hipótesis ($\mathrm{H_{MN}}$).

(a) Método de Euler implícito, (3.12).

(b) Método RK de colocación de parámetro $c_1 = 1/2$, (3.13).

4.2. 0-estabilidad

Salvo en casos excepcionales, la solución teórica de un PVI no satisface la recurrencia (MN) que define el método numérico, sino la recurrencia más una pequeña perturbación,

$$\sum_{j=0}^{k} \alpha_j y(x_{n+j}) = h\phi_f(x_n, y(x_n), \ldots, y(x_{n+k-1}); h) + R_n, \qquad n = 0, \ldots, N-k. \tag{4.3}$$

La cantidad R_n, el *residuo*, es lo que le falta a la solución teórica para ser solución del método numérico.

La solución producida por el ordenador tampoco es solución de la ecuación del método, pues los errores de redondeo introducen pequeñas perturbaciones. Por otra parte, ϕ_f en general no se calcula exactamente, sino sólo de forma aproximada. Así que en realidad el ordenador produce una solución $\{\hat{y}_n\}$ de una recurrencia perturbada

$$\begin{aligned} &\sum_{j=0}^{k} \alpha_j \hat{y}_{n+j} = h\big(\phi_f(x_n, \hat{y}_n, \ldots, \hat{y}_{n+k-1}; h) + \varepsilon_n\big) + \mu_n, && n = 0, \ldots, N-k, \\ &\hat{y}_n = y_n + \mu_n, && n = 0, \ldots, k-1, \end{aligned}$$

donde ε_n es el error cometido en el cálculo de $\phi_f(x_n, \hat{y}_n, \ldots, \hat{y}_{n+k-1}; h)$ y μ_n es el error de redondeo.

Para que el método sirva para algo, necesitamos que $y(x_n)$ e $\hat{y}_n$ permanezcan próximas si las perturbaciones R_n, ε_n y μ_n son pequeñas. Esto es, nos gustaría que el método numérico fuese *estable*. Esto motiva la siguiente definición.

Definición 4.1. *Un método* (MN) *se dice* 0-estable *si para cada* PVI *existe una constante* $C > 0$ *tal que para cada dos sucesiones* $\{u_n\}_{n=0}^N$ *y* $\{v_n\}_{n=0}^N$ *satisfaciendo*

$$\begin{aligned} &\sum_{j=0}^{k} \alpha_j u_{n+j} - h\phi_f(x_n, u_n, \ldots, u_{n+k-1}; h) = h\delta_n, \\ &\sum_{j=0}^{k} \alpha_j v_{n+j} - h\phi_f(x_n, v_n, \ldots, v_{n+k-1}; h) = h\gamma_n, \end{aligned} \qquad 0 \le n \le N-k,$$

se verifica que

$$\max_{k\le n\le N} \|u_n - v_n\| \le C \left(\max_{0\le n\le k-1} \|u_n - v_n\| + \max_{0\le n\le N-k} \|\delta_n - \gamma_n\| \right). \tag{4.4}$$

Ejemplo. Todos los métodos de un paso con $\alpha_0 = -1$ son 0-estables. En efecto, sean $\{u_n\}_{n=0}^N$ y $\{v_n\}_{n=0}^N$ satisfaciendo

$$\begin{aligned} u_{n+1} &= u_n + h\phi_f(x_n, u_n; h) + h\delta_n, \\ v_{n+1} &= v_n + h\phi_f(x_n, v_n; h) + h\gamma_n, \end{aligned} \qquad n = 0, \ldots N-1. \tag{4.5}$$

Restando ambas ecuaciones, tomando normas y usando la condición de Lipschitz ($\mathrm{H_{MN}}$)-(ii), se tiene que $e_n = \|u_n - v_n\|$ verifica

$$\begin{aligned} e_{n+1} &\le e_n + h\|\phi_f(x_n, u_n; h) - \phi_f(x_n, v_n; h)\| + h\|\delta_n - \gamma_n\| \\ &\le (1 + Lh)e_n + h\|\delta_n - \gamma_n\|. \end{aligned}$$

A partir de aquí se demuestra fácilmente por inducción que

$$\|u_n - v_n\| \le e^{L(x_n - x_0)}\|u_0 - v_0\| + e^{L(x_n - x_0)}(x_n - x_0) \max_{0\le n\le N-1} \|\delta_n - \gamma_n\|, \tag{4.6}$$

de donde se sigue inmediatamente que el método es 0-estable. ♣

Como corolario de la definición de 0-estabilidad se tiene el siguiente resultado de convergencia.

Corolario 4.2. *Si un método* (MN) *es* 0*-estable y*

$$\tau := \max_{0\le n\le N-k} \|R_n\|/h \to 0 \qquad \textit{cuando } h \to 0^+,$$

entonces el método es convergente. Si además para todo $q \le p$ *se tiene que* $\tau = O(h^q)$ *para todo* PVI *con* $f \in C^q$*, entonces el método es convergente de orden al menos* p*.*

Demostración. Basta con tomar $u_n = y(x_n)$, $\delta_n = R_n/h$, $v_n = y_n$ y $\gamma_n = 0$ en la definición de 0-estabilidad. □

Ejemplo. Gracias al ejemplo anterior sabemos que la regla del trapecio es 0-estable. Por otra parte, por el Teorema del Valor Medio,

$$\begin{aligned}\frac{R_n}{h} &= \frac{y(x_{n+1})-y(x_n)}{h} - \frac{f(x_n,y(x_n))+f(x_{n+1},y(x_{n+1}))}{2}\\ &= y'(\overline{\xi}_n) - \frac{y'(x_n)+y'(x_{n+1})}{2} = y'(\overline{\xi}_n) - y'(x_n) + \frac{y'(x_n)-y'(x_{n+1})}{2}.\end{aligned}$$

Usando la continuidad uniforme de $y'(x)$ en $[a,b]$ se concluye que $\tau \to 0$ cuando $h \to 0^+$, y por tanto la convergencia.

Si además $f \in C^2$, desarrollando por Taylor alrededor de x_n obtenemos que

$$\begin{aligned}R_n &= y(x_{n+1}) - \{y(x_n) + \tfrac{h}{2}[f(x_n,y(x_n)) + f(x_{n+1},y(x_{n+1}))]\}\\ &= y(x_n) + hy'(x_n) + \tfrac{h^2}{2}y''(x_n) + \tfrac{h^3}{3!}y'''(\overline{\xi}_n)\\ &\quad -\{y(x_n) + \tfrac{h}{2}[y'(x_n) + y'(x_n) + hy''(x_n) + \tfrac{h^2}{2}y'''(\overline{\eta}_n)]\}\\ &= \left(\tfrac{y'''(\overline{\xi}_n)}{3!} - \tfrac{y'''(\overline{\eta}_n)}{4}\right)h^3,\end{aligned}$$

de donde se deduce que

$$\tau \le Ch^2, \qquad C = \frac{5}{12}\max_{x\in[a,b]}\|y'''(x)\|.$$

Así pues, el método es convergente de orden al menos 2. ♣

Problemas

1. Consideramos un método de un sólo paso con $|\alpha_0| > 1$. Demostrar que no es 0-estable. *Indicación:* Aplicárselo a un problema con $f = 0$ y valores de arranque $y_0 = 0$ e $\hat{y}_0 = h$.
2. Probar que la regla del trapecio no es convergente de orden 3.
3. Consideramos el problema de valor inicial $y'(x) = \int_0^{y(x)} e^{-s^2}\, ds$, $x \in [0,1]$, $y(0) = 1$. Suponiendo que no se comete ningún error en el valor de arranque, $y_0 = y(x_0)$, determinar un valor h_0 que garantice un error menor que 10^{-3} al resolver el problema con $0 < h \leq h_0$ por medio de:
   (a) el método de Euler implícito;
   (b) la regla del trapecio.
4. Consideramos un método de un paso con $\alpha_0 = -1$. Sean $\{u_n\}_{n=0}^N$ y $\{v_n\}_{n=0}^N$ satisfaciendo (4.5), con ϕ_f verificando las hipótesis ($\mathrm{H_{MN}}$). Demostrar que existe una constante $K > 0$ tal que
$$\max_{0\leq n\leq N} \|u_n - v_n\| \geq K \max_{0\leq n\leq N} \left\| u_0 - v_0 + h\sum_{j=0}^{n-1}(\delta_j - \gamma_j) \right\|.$$
   Concluir de lo anterior que si existe un problema de valor inicial tal que $R_n = Ch^{p+1}$, con una constante C independiente de n, entonces el método no puede tener orden de convergencia mayor que p. Como aplicación, demostrar que el método de Euler modificado no es convergente de orden 3.
5. Intentar obtener un resultado análogo al del problema anterior para métodos de más de una paso.

4.3. Consistencia

La 0-estabilidad garantiza que si $\lim_{h\to 0^+} \tau = 0$, el método converge. Si además $\tau = O(h^p)$ cuando $h \to 0^+$ para todo PVI con $f \in C^p$, converge con orden al menos p. Esto motiva la siguiente definición.

Definición 4.3. *Un método de la forma* (MN) *es* consistente *si para todo* PVI *se tiene que* $\lim_{h\to 0^+} \tau = 0$*; es* consistente de orden p *si, siendo consistente,* p *es el mayor entero tal que para todo* PVI *con* $f \in C^q$*,* $q \le p$*, se tiene que* $\tau = O(h^q)$ *cuando* $h \to 0^+$*.*

Ejemplo. Consideramos la regla del punto medio (3.7). Por el Teorema del Valor Medio,

$$\frac{R_n}{h} = \frac{y(x_{n+2}) - y(x_n)}{h} - 2f(x_{n+1}, y(x_{n+1})) = 2y'(\overline{\xi}_n) - 2y'(x_{n+1}),$$

donde $x_n \le \overline{\xi}_n \le x_{n+2}$. Usando la continuidad uniforme de y' en $[a,b]$ se tiene que para todo $\varepsilon > 0$ existe un δ tal que $\frac{\|R_n\|}{h} \le \varepsilon$ si $|\overline{\xi}_n - x_n| \le \delta$. Esta última condición está garantizada si $h < \delta$. Concluimos por tanto que

$$\max_{0\le n\le N-2} \frac{\|R_n\|}{h} \to 0 \quad \text{cuando } h \to 0^+;$$

es decir, el método es consistente.

Si $f \in C^2$, la regla del punto medio tiene residuo

$$\begin{aligned} R_n &= y(x_{n+2}) - y(x_n) - 2hf(x_{n+1}, y(x_{n+1})) \\ &= y(x_{n+2}) - y(x_n) - 2hy'(x_{n+1}) \\ &= \left(y(x_n) + y'(x_n)(2h) + \tfrac{1}{2!}y''(x_n)(2h)^2 + \tfrac{1}{3!}y'''(\xi_n)(2h)^3\right) \\ &\quad -y(x_n) - 2h\left(y'(x_n) + y''(x_n)h + \tfrac{1}{2!}y'''(\eta_n)h^2\right) \\ &= \left(\tfrac{4}{3}y'''(\xi_n) - y'''(\eta_n)\right)h^3, \end{aligned}$$

con $x_n \le \xi_n \le x_{n+2}$ y $x_n \le \eta_n \le x_{n+1}$. Así pues, $\tau = O(h^2)$, y el método es consistente de orden al menos 2.

Si aplicamos el método al problema

$$y'(x) = \frac{x^2}{2}, \quad 0 \le x \le b, \qquad y(0) = 0, \tag{4.7}$$

cuya solución, $y(x) = x^3/3!$, verifica que $y'''(x) = 1$, se tiene que $R_n = h^3/3$. Así pues, $\tau = h^2/3$, y el método no es consistente de orden 3. ♣

¿Qué tiene que satisfacer un método para ser consistente? El siguiente resultado da una caracterización completa.

Teorema 4.4. *Un método* (MN) *que satisface las hipótesis* ($\mathrm{H_{MN}}$) *es consistente si y sólo si*

$$\sum_{j=0}^{k} \alpha_j = 0 \tag{C1}$$

y además

$$\phi_f(x, y(x), \ldots, y(x); 0) = \left(\sum_{j=0}^{k} j\alpha_j \right) f(x, y(x)). \tag{C2}$$

Demostración. Aplicamos el método (MN) al problema

$$y'(x) = 0, \qquad y(0) = 1,$$

cuya solución es $y(x) = 1$. Puesto que $f = 0$, entonces $\phi_f = 0$, y se tiene que

$$\frac{R_n}{h} = \frac{1}{h} \sum_{j=0}^{k} \alpha_j.$$

Si el método es consistente, el límite de este cociente es cero, algo que sólo es posible si se da la condición (C1).

Consideramos ahora el problema general (PVI) y lo integramos numérica-

mente en el intervalo $[a,x]$, con $x \in (a,b]$. Si se cumple (C1), entonces

$$\begin{aligned}\frac{R_n}{h} &= \frac{1}{h}\sum_{j=0}^{k}\alpha_j y(x_{n+j}) - \phi_f(x_n, y(x_n), \ldots, y(x_{n+k-1}); h)\\ &= \frac{1}{h}\sum_{j=0}^{k}\alpha_j\big(y(x_n) + jhy'(\overline{\xi}_{n,j})\big) - \phi_f(x_n, y(x_n), \ldots, y(x_{n+k-1}); h)\\ &= \sum_{j=0}^{k} j\alpha_j y'(\overline{\xi}_{n,j}) - \phi_f(x_n, y(x_n), \ldots, y(x_{n+k-1}); h)\\ &= \sum_{j=0}^{k} j\alpha_j f(\overline{\xi}_{n,j}, y(\overline{\xi}_{n,j})) - \phi_f(x_n, y(x_n), \ldots, y(x_{n+k-1}); h),\end{aligned} \tag{4.8}$$

donde $x_n \le \xi_{n,j} \le x_{n+j}$. Si el problema es consistente, pasando al límite $h \to 0^+$ para $n = 0$ y usando la continuidad de f y ϕ_f, se obtiene (C2) con $x = a$; haciendo lo mismo para $n = N - k$,

$$\begin{aligned}\frac{R_{N-k}}{h} &= \sum_{j=0}^{k} j\alpha_j f(\overline{\xi}_{N-k,j}, y(\overline{\xi}_{N-k,j})) - \phi_f(x_{N-k}, y(x_{N-k}), \ldots, y(x_{N-1}); h)\\ &\to \sum_{j=0}^{k} j\alpha_j f(x, y(x)) - \phi_f(x, y(x), \ldots, y(x); 0),\end{aligned}$$

de donde, si el método es consistente, se obtiene (C2) para $x \in (a,b]$.

Veamos el recíproco. Supongamos que se cumplen las condiciones (C1) y (C2). Entonces, usando (4.8) se tiene que

$$\frac{R_n}{h} = \sum_{j=0}^{k} j\alpha_j \frac{\phi_f(\overline{\xi}_{n,j}, y(\overline{\xi}_{n,j}), \ldots, y(\overline{\xi}_{n,j}); 0) - \phi_f(x_n, y(x_n), \ldots, y(x_{n+k-1}); h)}{\sum_{j=0}^{k} j\alpha_j}.$$

Utilizando la continuidad uniforme de y y ϕ_f sobre compactos, se concluye que $\tau \to 0$. □

Ejemplo. Los métodos de un paso son consistentes si y sólo si

$$\alpha_0 = -1 \quad \text{y} \quad \phi_f(x, y(x); 0) = f(x, y(x)).$$

Usando (4.1) es fácil ver que la regla del trapecio cumple estas dos condiciones; por tanto es consistente. ♣

Problemas

1. Decidir si los siguientes métodos son consistentes y en caso afirmativo determinar el orden de consistencia.

   (a) Predictor-corrector Euler/Trapecio, (3.3)–(3.4).

   (b) Regla de Simpson,

   $$y_{n+2} - y_n = \frac{h}{3}(f(x_n, y_n) + 4f(x_{n+1}, y_{n+1}) + f(x_{n+2}, y_{n+2})). \quad (4.9)$$

   (c) Fórmula BDF de dos pasos,

   $$y_{n+2} - \frac{4}{3}y_{n+1} + \frac{1}{3}y_n = \frac{2}{3}hf(x_{n+2}, y_{n+2}). \quad (4.10)$$

   (d) Método de colocación de parámetro $c_1 = 1/2$, (3.13).

2. Determinar qué condiciones deben cumplir las constantes α y β para que el método

   $$y_{n+1} = y_n + \alpha hf(x_n, y_n) + \beta hf(x_{n+1}, y_{n+1})$$

   sea: (i) consistente; (ii) consistente de orden al menos uno; y (iii) consistente de orden al menos dos. ¿Es consistente de orden 3 para alguna elección de α y β?

3. Considérese el método

   $$y_{n+1} = y_n - \frac{5}{2}hf(x_n, y_n) + \frac{7}{2}hf(x_n + \frac{h}{7}, y_n + \frac{h}{7}f(x_n, y_n)).$$

   (a) Escribir la función de incremento ϕ_f y comprobar que cumple las hipótesis ($\mathrm{H_{MN}}$).

   (b) Decidir si es consistente y, en caso afirmativo, hallar el orden de consistencia.

4. Repetir el ejercicio anterior para el método

   $$y_{n+1} = y_n + hf(x_n + (1-\theta)h, \theta y_n + (1-\theta)y_{n+1}),$$

   donde $\theta \in [0, 1]$.

4.4. Criterio de la raíz

Si un método de un paso es consistente, entonces $\alpha_0 = -1$ y es 0-estable, y por tanto convergente. ¿Sucede lo mismo con los métodos de más de un paso? La respuesta es que no.

Ejemplo. Consideramos el método

$$y_{n+2} + y_{n+1} - 2y_n = h\big(5f(x_{n+1}, y_{n+1}) - 2f(x_n, y_n)\big). \tag{4.11}$$

Se tiene $\phi_f(x_n, y_n, y_{n+1}; h) = 5f(x_n + h, y_{n+1}) - 2f(x_n, y_n)$, $\sum_{j=0}^{2} \alpha_j = 0$ y $\sum_{j=0}^{2} j\alpha_j = 3$. Por consiguiente, $\phi_f(x, y(x), y(x); 0) = 3f(x, y(x))$. Así pues, el método es consistente.

Aplicamos este método al problema más simple que uno pueda imaginar,

$$y'(x) = 0, \quad a \le x \le b, \qquad y(a) = 0,$$

cuya solución es $y(x) = 0$. El método en este caso queda

$$y_{n+2} + y_{n+1} - 2y_n = 0.$$

Para que $y_n = \lambda^n$ sea solución de esta recurrencia, λ tiene que ser raíz del polinomio $\rho(\zeta) = \zeta^2 + \zeta - 2$. Así, λ debe ser 1 o -2, lo que da lugar a las soluciones $y_n = 1$, $y_n = (-2)^n$. La solución general de la recurrencia será una combinación lineal de estas dos soluciones,

$$y_n = c_1 + c_2(-2)^n.$$

Consideramos las soluciones $u_n = 0$ y $v_n = (-2)^n$. Se tiene que $\delta_n = 0 = \gamma_n$ y que

$$\max_{0 \le n \le 1} \|u_n - v_n\| = 2, \qquad \max_{2 \le n \le N} \|u_n - v_n\| = 2^N.$$

Si el método fuera 0-estable tendría que existir una constante C tal que $2^N \le C2$, lo que es evidentemente falso.

Por otra parte, $y_n = h(-2)^n$ es solucion de la recurrencia y verifica que

$$\max_{0\le n\le 1} |y(x_n) - y_n| = 2h \to 0 \quad \text{cuando } h \to 0^+.$$

Sin embargo

$$\max_{2\le n\le N} |y(x_n) - y_n| = h2^N \to \infty \quad \text{cuando } h \to 0^+.$$

Por tanto, el método no es convergente. ♣

Como se ve en este ejemplo, la 0-estabilidad está relacionada con el tamaño de las raíces del polinomio

$$\rho(\zeta) = \sum_{j=0}^{k} \alpha_j \zeta^j,$$

conocido como *primer polinomio característico* del método, que es el polinomio asociado a la recurrencia que se obtiene al aplicar el método a problemas con $f = 0$. Esto motiva la siguiente definición.

Definición 4.5. *Se dice que un método* (MN) *satisface la* condición de la raíz *si todas las raíces del primer polinomio característico tienen módulo menor o igual que 1, y aquellas que tienen módulo 1 son simples.*

Observación. Si un método es consistente, una de las raíces de $\rho(\zeta)$ debe ser 1. Ésta es la *raíz principal*, a la que se etiqueta como ζ_1. Las restantes raíces, ζ_i, $i = 2, \ldots, k$, son las llamadas *raíces espurias*; surgen porque decidimos representar un sistema diferencial de primer orden por un sistema en diferencias de orden k. Obviamente, los métodos consistentes con $k = 1$ no tienen raíces espurias, y satisfacen por tanto la condición de la raíz. ♠

La 0-estabilidad resulta ser equivalente a que se satisfaga la condición de la raíz.

Teorema 4.6. *Un método* (MN) *es* 0*-estable si y sólo si satisface la condición de la raíz.*

Demostración. Consideramos el problema de valor inicial con $f = 0$ y tomamos dos soluciones u_n y v_n de la ecuación en diferencias sin perturbar

$$\sum_{j=0}^{k} \alpha_j y_{n+j} = 0. \tag{4.12}$$

(i) Supongamos que existe ξ tal que $|\xi| > 1$ y $\rho(\xi) = 0$. Las sucesiones $u_n = 0$ y $v_n = \xi^n$ son solución de la ecuación en diferencias (4.12). Tenemos que

$$\max_{0\le n\le k-1} \|u_n - v_n\| = |\xi|^{k-1}, \qquad \max_{k\le n\le N} \|u_n - v_n\| = |\xi|^N.$$

Si el método fuera 0-estable, existiría una constante C tal que $|\xi|^N \le C|\xi|^{k-1}$, una contradicción.

(ii) Supongamos que existe ξ tal que $|\xi| = 1$, $\rho(\xi) = 0 = \rho'(\xi)$. Las sucesiones $u_n = 0$ y $v_n = n\xi^n$ son solución de la ecuación en diferencias (4.12) (nótese que ξ es raíz doble). Entonces

$$\max_{0\le n\le k-1} \|u_n - v_n\| = k - 1, \qquad \max_{k\le n\le N} \|u_n - v_n\| = N.$$

Si suponemos que el método es 0-estable, de nuevo llegamos a una contradicción.

Veamos ahora que si el método cumple la condición de la raíz, entonces es 0-estable. Damos la demostración (que se puede omitir en una primera lectura) en el caso escalar. El caso vectorial es análogo.

La idea es escribir el método como método de un paso. Para ello denotamos

$$\begin{aligned} &Y_n = (y_{n+k-1}, \dots, y_n)^T, && n = 0, \dots, N-k+1, \\ &F(Y_n) = (\phi_f(x_n, y_n, \dots, y_{n+k-1}; h), 0, \dots, 0)^T, && n = 0, \dots, N-k+1. \end{aligned}$$

Con esta notación podemos escribir el método como

$$Y_{n+1} = AY_n + hF(Y_n), \qquad n = 0, \dots, N-k,$$

donde A es la *matriz compañera* del primer polinomio característico,

$$A = \begin{bmatrix} -\alpha_{k-1} & -\alpha_{k-2} & \dots & -\alpha_1 & -\alpha_0 \\ 1 & 0 & \dots & & 0 \\ 0 & 1 & \ddots & & \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & 1 & 0 \end{bmatrix}.$$

Iterando el proceso tenemos que

$$Y_n = A^n Y_0 + h \sum_{l=0}^{n-1} A^{n-1-l} F(Y_l).$$

Para las u_n y v_n de la definición de estabilidad, llamando

$$R_n = (\delta_n - \gamma_n, 0, \dots, 0)^T, \qquad n = 0, \dots, N-k,$$

y restando tendremos que

$$U_n - V_n = A^n(U_0 - V_0) + h \sum_{l=0}^{n-1} A^{n-1-l}(F(U_l) - F(V_l)) + h \sum_{l=0}^{n-1} A^{n-1-l} R_l.$$

Si las potencias de A están todas ellas acotadas, esto es, si

$$\|A^n\| \le M, \qquad n = 0, 1, \dots,$$

tomando normas y utilizando que ϕ_f satisface la condición de Lipschitz ($\mathrm{H_{MN}}$)-(ii), tenemos que

$$\|U_n - V_n\| \le M\|U_0 - V_0\| + hLM \sum_{l=0}^{n-1} \|U_l - V_l\| + (b-a)M \max_{0 \le l \le N-k} \|\delta_l - \gamma_l\|,$$

desigualdad que se puede escribir como

$$\psi_n \le N + hLM \sum_{l=0}^{n-1} \psi_l,$$

definiendo ψ_n y N de la manera obvia.

Sea $\chi_n = N + hLM \sum_{l=0}^{n-1} \psi_l$. Se tiene que

$$\chi_{n+1} - \chi_n = hLM\psi_n \leq hLM\chi_n,$$

de donde concluimos que

$$\psi_n \leq \chi_n \leq (1 + hLM)^{n-1}\chi_1 \leq e^{(n-1)hLM}(N + hLM\|U_0 - V_0\|),$$

y de aquí la 0-estabilidad.

Para que las potencias de A estén todas acotadas es necesario y suficiente que las raíces del polinomio ρ sean de módulo menor o igual que la unidad, y aquellas que sean de módulo 1 sean simples, es decir, que se cumpla el criterio de la raíz. Para ver esto no hay más que observar que los autovalores de A son las raíces de ρ. □

Ejemplo. El primer polinomio característico de la regla del punto medio (3.7) es

$$\rho(\zeta) = \zeta^2 - 1.$$

Sus raíces son $\zeta_1 = 1$ y $\zeta_2 = -1$, luego el método satisface la condición de la raíz. Por consiguiente, el método es 0-estable. Como también es consistente de orden 2, es convergente de orden al menos 2.

Por otra parte, si aplicamos el método al problema (4.7), se obtiene la recurrencia

$$y_{n+2} - y_n = h^3(n+1)^2, \qquad n = 0, \ldots, N-2,$$

cuya solución general es

$$y_n = c_1 + c_2(-1)^n + \frac{h^3}{3!}(n^3 - n).$$

Si tomamos valores de arranque $y_0 = 0 = y_1$, entonces

$$y_n = \frac{h^3}{3!}(n^3 - n).$$

Así pues, $\max\limits_{0\le n\le 1} |y(x_n)-y_n| = \frac{h^3}{3!}$ y sin embargo

$$\max_{2\le n\le N} |y(x_n)-y_n| = \frac{h^2}{3!} \max_{2\le n\le N} |x_n| = \frac{bh^2}{3!};$$

el método no es por tanto convergente de orden 3. ♣

Problemas

1. Comprobar que los autovalores de la matriz compañera del primer polinomio característico son precisamente las raíces de dicho polinomio.
2. Comprobar que la condición necesaria y suficiente para que todas las potencias de una matriz tengan norma uniformemente acotada es que todos sus autovalores tengan módulo menor o igual que uno y aquellos que tengan módulo uno sean simples.
3. Determinar si son 0-estables la regla de Simpson, (4.9), y la fórmula BDF de 2 pasos, (4.10).
4. Consideramos la familia de métodos

   $$y_{n+3}+(2\alpha-3)y_{n+2}-(2\alpha-3)y_{n+1}-y_n = h\alpha\big(f(x_{n+2},y_{n+2})+f(x_{n+1},y_{n+1})\big),$$

   donde α es un parámetro real. Estudiar para qué valores de α es 0-estable.
5. Extender la demostración del teorema 4.6 al caso vectorial, $d > 1$.

4.5. Teorema de equivalencia

Hemos visto que "0-estabilidad + consistencia ⇒ convergencia". ¿Es cierto el recíproco?

Teorema 4.7 (Teorema de equivalencia)**.** *Un método* (MN) *que satisface las condiciones* ($\mathrm{H_{MN}}$) *es convergente si y sólo si es* 0*-estable y consistente.*

Demostración. En primer lugar probamos que "convergencia ⇒ criterio de la raíz". Consideramos el problema

$$y'(x) = 0, \qquad y(0) = 0,$$

cuya solución es $y(x) = 0$.

(i) Supongamos que existe ξ tal que $|\xi| > 1$ y $\rho(\xi) = 0$. La sucesión $y_n = h\xi^n$, $n = 0, 1, \ldots$, es solución de la ecuación en diferencias y tiene condiciones de arranque $y_n = h\xi^n$, $n = 0, \ldots, k-1$; por tanto

$$\max_{0\le n\le k-1} \|y(x_n) - y_n\| = h|\xi|^{k-1} \to 0.$$

Sin embargo,

$$\max_{k\le n\le N} \|y(x_n) - y_n\| = h|\xi|^N = \frac{b-a}{N}|\xi|^N \to \infty \quad \text{cuando } N \to \infty,$$

y por tanto el método no es convergente.

(ii) Supongamos que existe ξ tal que $|\xi| = 1$, $\rho(\xi) = 0 = \rho'(\xi)$. La sucesión $y_n = \sqrt{h}n\xi^n$, $n = 0, 1, \ldots$, es solución de la ecuación en diferencias y tiene condiciones de arranque $y_n = \sqrt{h}n\xi^n$, $n = 0, \ldots, k-1$; por tanto

$$\max_{0\le n\le k-1} \|y(x_n) - y_n\| = \sqrt{h}(k-1)|\xi|^{k-1} \to 0.$$

Sin embargo,

$$\max_{k\le n\le N} \|y(x_n) - y_n\| = \sqrt{h}N|\xi|^N = \sqrt{b-a}\sqrt{N} \to \infty \quad \text{cuando } N \to \infty,$$

y el método no es convergente.

Veamos ahora que "convergencia ⇒ consistencia". Si un método es convergente, en particular lo es para el problema $y'(x) = 0$, $0 \le x \le b$, $y(0) = 1$, cuya solución es $y(x) = 1$. La aplicación del método numérico a este problema produce la recurrencia

$$\sum_{j=0}^{k} \alpha_j y_{n+j} = 0, \qquad n = 0, \ldots, N-k. \tag{4.13}$$

Si tomamos valores de arranque $y_n = 1$, $n = 0, \ldots, k-1$, entonces

$$\max_{0\le n\le k-1} \|y(x_n) - y_n\| = 0.$$

Por consiguiente, la convergencia implica que

$$\max_{k\le n\le N} \|1 - y_n\| \to 0,$$

y por tanto que $\lim_{h\to 0^+} y_n = 1$. Así, pasando al límite en (4.13) se llega a (C1).

Supongamos ahora que no se cumple la condición (C2),

$$\phi_f(x, y(x), \dots, y(x); 0) - \left(\sum_{j=0}^{k} j\alpha_j\right) f(x, y(x)) \not\equiv 0.$$

Nótese que $\rho'(1) = \sum_{j=0}^{k} j\alpha_j \neq 0$; en otro caso $\zeta_1 = 1$ sería raíz doble, y no se cumpliría la condición de la raíz.

Consideramos el PVI

$$\hat{y}'(x) = \underbrace{\phi_f(x, \hat{y}(x), \dots, \hat{y}(x); 0)/\sum_{j=0}^{k} j\alpha_j}_{\hat{f}(x,\hat{y}(x))}, \qquad \hat{y}(a) = y(a) = \eta. \tag{4.14}$$

Se obtiene fácilmente (ver fórmula (4.8)) que para este problema y con este método

$$\frac{R_n}{h} = \frac{1}{h}\sum_{j=0}^{k} \alpha_j \hat{y}(x_{n+j}) - \phi_f(x_n, \hat{y}(x_n), \dots, \hat{y}(x_{n+j})) \to 0 \quad \text{cuando } h \to 0^+.$$

Como el método es 0-estable, se concluye que y_n converge a la solución $\hat{y} \neq y$ de (4.14), ¡una ecuación equivocada! □

Observación. Las condiciones de consistencia garantizan que estamos resolviendo la ecuación diferencial correcta, que estamos siendo *consistentes* con ella. ♠

Ejemplo. La función de incremento del método

$$y_{n+1} = y_n + h(2f(x_n, y_n) + 3f(x_{n+1}, y_{n+1})) \tag{4.15}$$

viene dada implícitamente por

$$\phi_f(x, y; h) = 2f(x, y) + 3f(x + h, y + h\phi_f(x, y; h)).$$

Así, $\phi_f(x, y(x); 0) = 5f(x, y(x))$, y el método no es convergente, ya que no es consistente. Las soluciones proporcionadas por el método numérico convergerán a una solucion de la EDO con función de lado derecho $\hat{f}(x, y(x)) = \phi_f(x, y(x); 0)/\sum_{j=0}^{1} j\alpha_j = 5f(x, y(x))$. ♣

Problemas

1. Estudiar si es convergente la *regla implícita del punto medio*,

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

2. Repetir el problema anterior para el par predictor-corrector AB2/AM2,

$$\begin{aligned} y^*_{n+2} &= y_{n+1} + \tfrac{h}{2}(3f(x_{n+1}, y_{n+1}) - f(x_n, y_n)), \\ y_{n+2} &= y_{n+1} + \tfrac{h}{12}(5f(x_{n+2}, y^*_{n+2}) + 8f(x_{n+1}, y_{n+1}) - f(x_n, y_n)). \end{aligned}$$

3. Consideramos la familia de métodos

$$y_{n+3}+(2\alpha-3)y_{n+2}-(2\alpha-3)y_{n+1}-y_n = h\alpha\big(f(x_{n+2}, y_{n+2})+f(x_{n+1}, y_{n+1})\big),$$

donde α es un parámetro real. Estudiar para que valores de α es convergente.

4. Construir todos los métodos convergentes de orden 2 de la forma

$$y_{n+1} = y_n + hf(x_n + \alpha h, y_n + \alpha h f(x_n + \beta h, y_n + \beta h f(x_n, y_n))).$$

Comprobar que ninguno tiene orden de convergencia 3. ¿Tiene alguno orden de convergencia 3 para el problema $y' = y$ en $[0, b]$, $y(0) = 1$?

5. Al aplicar el método numérico

$$y_{n+2} + \alpha y_{n+1} + \beta y_n = h\phi(x_n, y_n, y_{n+1}; h)$$

al problema $y'(x) = f(x, y(x))$, $x \in [a, b]$, $y(a) = \eta$, se obtiene una aproximación numérica de la solución del problema $y'(x) = 2f(x, y(x))$, $x \in [a, b]$, $y(a) = \eta$. Si $\phi(x, y(x), y(x); 0) = f(x, y(x))$, ¿cuánto valen α y β?

4.6. Experimentos numéricos

Consideramos el problema lineal

$$\begin{aligned} &y'(x) = Ay(x) + B(x) \qquad \text{para } 0 \le x \le 10, \qquad y(0) = (2,3)^T, \\ &A = \begin{pmatrix} -2 & 1 \\ 1 & -2 \end{pmatrix}, \qquad B(x) = \begin{pmatrix} 2\operatorname{sen} x \\ 2(\cos x - \operatorname{sen} x) \end{pmatrix}, \end{aligned} \tag{4.16}$$

cuya solución es

$$y = 2e^{-x}\begin{pmatrix} 1 \\ 1 \end{pmatrix} + \begin{pmatrix} \operatorname{sen} x \\ \cos x \end{pmatrix}. \tag{4.17}$$

Para empezar le aplicamos el método de Runge-Kutta clásico,

$$\begin{cases} k_1 = f(x_n, y_n), \\ k_2 = f(x_n + \frac{h}{2}, y_n + \frac{hk_1}{2}), \\ k_3 = f(x_n + \frac{h}{2}, y_n + \frac{hk_2}{2}), \\ k_4 = f(x_n + h, y_n + hk_3), \\ y_{n+1} = y_n + \frac{h}{6}(k_1 + 2k_2 + 2k_3 + k_4). \end{cases}$$

En la Figura 4.1 se muestra el correspondiente diagrama de eficiencia. El diagrama tiene pendiente -4, lo que se corresponde con el orden de convergencia del método, hasta $N \approx 15000$. A partir de ahí, aunque trabajemos más no conseguimos una mayor precisión: ¡el error incluso empeora! Es fácil entender lo que está sucediendo. Por muy bueno que sea un método y por muy pequeño que sea el paso, lo que no podemos es conseguir una precisión mayor que la que sea capaz de proporcionar el ordenador. Si nos fijamos en el menor error conseguido, vemos que es del orden de 10^{-15}. Dado que la solución es de tamaño 1, estamos consiguiendo unas 15 cifras significativas, que es aproximadamente la precisión que proporciona el Matlab.

A continuación le aplicamos el método (4.11) que, recordemos, es consistente pero no 0-estable, con $h = 10^{-3}$. En la Figura 4.2 se representa el error en escala logarítmica frente a x en el intervalo $[0, 1]$ (poco después de $x = 1$ el

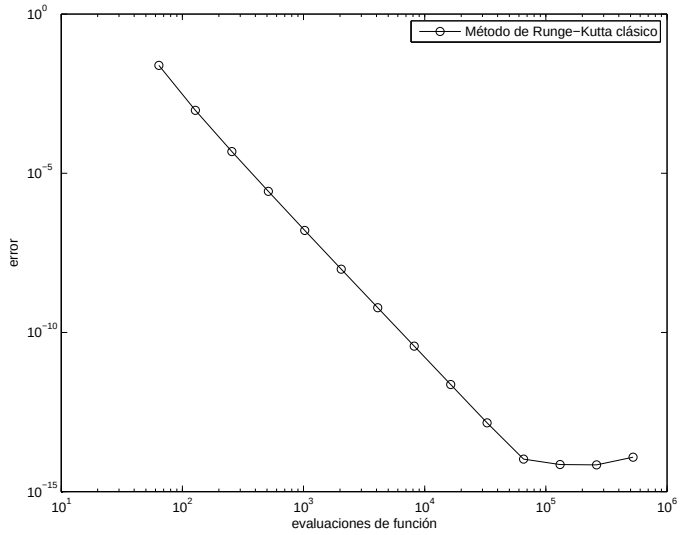

Figura 4.1: Diagrama de eficiencia para el método de Runge-Kutta clásico aplicado al Problema 1.

error excede a las capacidades del Matlab). El error crece exponencialmente; más concretamente, se tiene que $\texttt{error}(x_n) \approx c2^n$. No es casual que crezca como una potencia de 2. Esto es típico de los métodos que no son 0-estables: el crecimiento de la solución numérica viene dado al cabo de unos pocos pasos por aquella de las soluciones de la recurrencia que se obtiene al poner $\phi_f = 0$ en la ecuación del método que tiene un crecimiento más rápido. En nuestro caso, esta solución es un múltiplo de $(-2)^n$, ya que -2 es la raíz de mayor módulo del primer polinomio característico del método.

Para finalizar, aplicamos al problema el método (4.15), que no es consistente, aunque sí 0-estable. El correspondiente diagrama de eficiencia se muestra en la Figura 4.3. A partir de un cierto N el error no parece depender demasiado de h, y se aproxima cada vez más a 1,0060.

Como ya se explicó en la sección anterior, lo que sucede es que la solución numérica converge a la solución del problema con el mismo dato inicial y lado

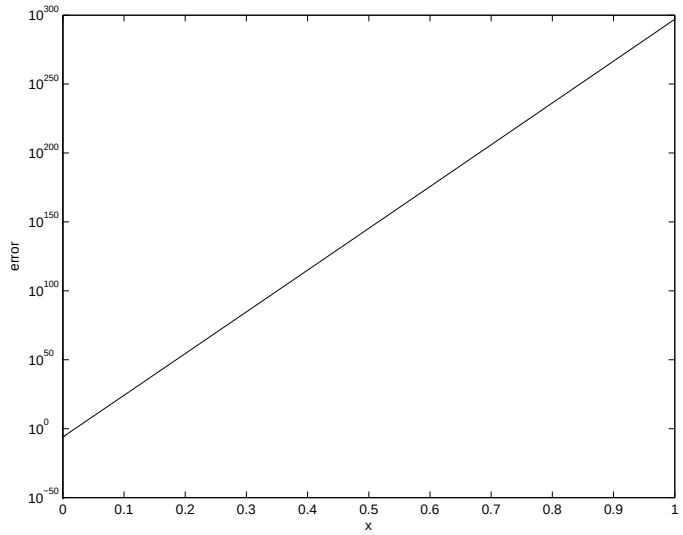

Figura 4.2: Error cometido con el método (4.11) aplicado al Problema 1 con $h = 10^{-3}$.

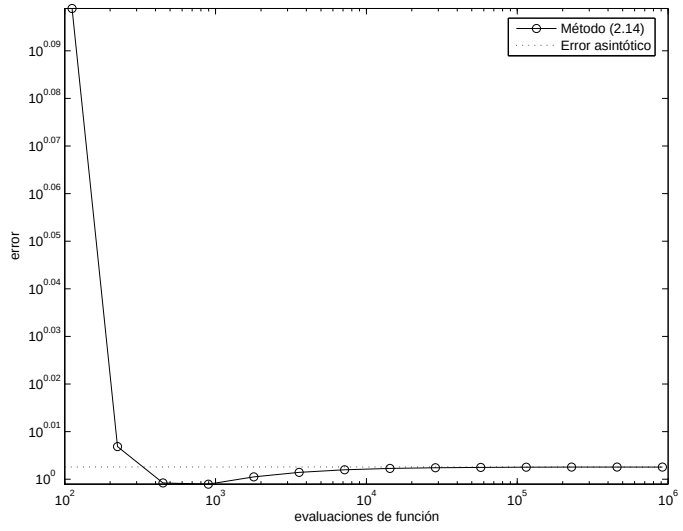

Figura 4.3: Diagrama de eficiencia para el método (4.15) aplicado al Problema 1.

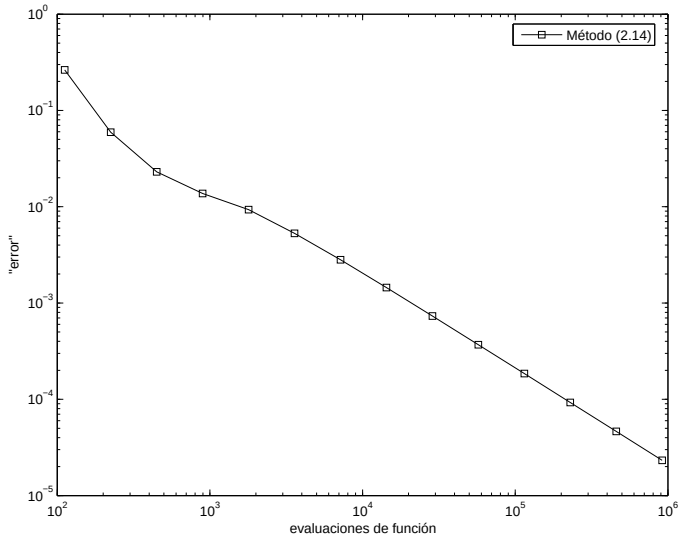

Figura 4.4: Diferencia entre la solución numérica obtenida por el método (4.15) aplicado al Problema 1 y la solución de la EDO con lado derecho $\hat{f}$ vs. el número de evaluaciones de función.

derecho $\hat{f}(x, y(x)) = 5(Ay(x) + B(x))$. Para comprobarlo, en la Figura 4.4 se muestra para varios valores de N, en escalas logarítmicas, la diferencia entre la solución numérica y la solución del problema que tiene por lado derecho a $\hat{f}$. Se ve claramente que la solución numérica converge, con orden 1, a la solución del problema "equivocado". Por consiguiente, el error cometido por el método debería aproximarse cada vez más, al aumentar N, a la máxima diferencia, en norma infinito, entre la solución de la EDO con lado derecho f y la solución de la EDO con lado derecho $\hat{f}$. Esa diferencia resulta ser precisamente 1,0060.

Problemas

1. Supongamos que los errores de evaluación de ϕ_f están acotados por ε y los de redondeo por μ. Consideramos un método 0-estable con orden de consistencia p. Sea C una constante tal que $\tau \leq Ch^p$ y sea $\{\hat{y}_n\}$ la

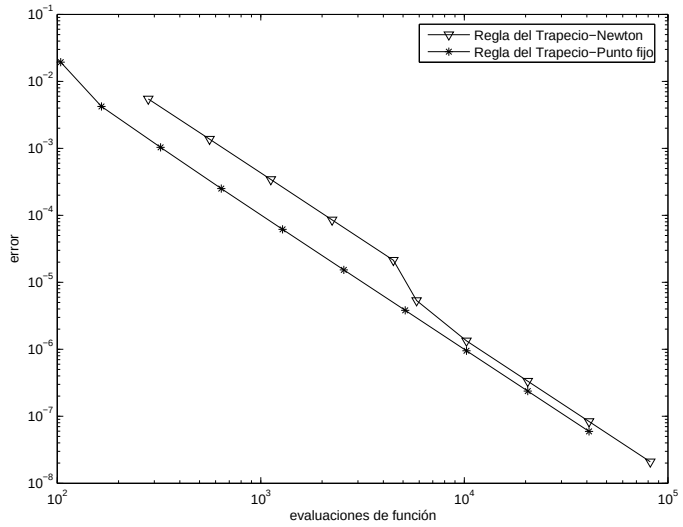

Figura 4.5: Diagramas de eficiencia para la Regla del Trapecio (con Newton y punto fijo) aplicada al Problema 1, con una tolerancia para las iteraciones igual a 10^{-2}.

solución numérica producida por el ordenador. Dar una cota para el error $\max_{k\leq n\leq N} \|y(x_n) - \hat{y}_n\|$ y obtener el valor de h que minimiza dicha cota.

2. Consideramos el método $y_{n+1} - \frac{1}{2}y_n = hf(x_n, y_n)$, que es 0-estable y satisface la condición de consistencia (C2), pero no la condición de consistencia (C1). Estudiar cómo se comporta el error.

3. Aplicamos la Regla del Trapecio (resolviendo los problemas implícitos tanto mediante el método de Newton como mediante iteración de punto fijo) al problema (4.16), pidiendo una tolerancia poco exigente, concretamente 10^{-2}, en las iteraciones de Newton/punto fijo. Se obtienen los diagramas de eficiencia que aparecen representados en la Figura 4.5.

   ¿Puedes explicar el brusco cambio que experimenta el diagrama de la Regla del Trapecio-Newton cuando el número de evaluaciones de función es aproximadamente 2500? ¿Y por qué no se da una situación parecida para la Regla del Trapecio-punto fijo?

4. Sea $f \in C^2$. Demostrar que el error global del método de Euler admite

un desarrollo asintótico de la forma

$$y(x_n) - y_n = d(x_n)h + O(h^2),$$

siendo $d \in C^2([a,b])$ la solución del problema

$$d'(x) = \sum_{J=1}^{d} \frac{\partial f}{\partial y^J}(x, y(x)) d^J(x) + \frac{1}{2} y''(x), \qquad d^J(0) = 0,\ J = 1, \ldots, d.$$

Indicación. Si definimos $y_n^* := y_n + d(x_n)h$, la sucesión $\{y_n^*\}$ es la solución numérica producida por el método de un paso cuya función de incremento es

$$\phi_f^*(x_n, y_n^*; h) = f(x_n, y_n^* - d(x_n)h) + d(x_n + h) - d(x_n).$$

El residuo de dicho método es una $O(h^3)$.

Capítulo 5

Dos familias importantes

En este capítulo aplicaremos la teoría desarrollada en el capítulo 4 a dos familias importantes: los métodos de Runge-Kutta y los métodos lineales multipaso.

5.1. Métodos de Runge-Kutta: definición

Los métodos de Runge-Kutta son aquellos que se pueden escribir en la forma

$$\begin{cases} k_i = f(x_n + c_i h, y_n + h \sum_{j=1}^{s} a_{ij} k_j), & i = 1, 2, \ldots, s, \\ y_{n+1} = y_n + h \sum_{i=1}^{s} b_i k_i. \end{cases} \tag{RK}$$

Se avanza con una pendiente promedio obtenida a partir de evaluaciones de la f en diversos puntos próximos a (x_n, y_n). Cada una de las evaluaciones de función k_i es una *etapa*.

El método (RK) se representa por medio de su *tablero de Butcher*:

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

Si escribimos $c = (c_1, c_2, \dots, c_s)^T$, $b = (b_1, b_2, \dots, b_s)^T$, $A = (a_{ij})$, el tablero se puede resumir como

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

Si $a_{ij} = 0$ para $j \geq i$, $i = 1, 2, \dots, s$, es decir, si la matriz A es triangular inferior estricta, entonces cada uno de los k_i viene dado explícitamente en términos de los anteriormente calculados, k_j, $j = 1, 2, \dots, i-1$. En este caso el método es *explícito*. Al escribir su tablero se suelen omitir los ceros sobre y por encima de la diagonal principal.

Ejemplo. El método de Euler modificado, (3.6), es explícito. Su tablero es

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

♣

Si el método no es explícito, es *implícito*. En general es necesario entonces resolver en cada paso un sistema no lineal para calcular los k_i. Este sistema tiene dimensión ds.

Ejemplo. El método RK-Radau IA de dos etapas, de tablero

$$\begin{array}{c|cc} 0 & 1/4 & -1/4 \\ 2/3 & 1/4 & 5/12 \\ \hline & 1/4 & 3/4, \end{array}$$

es un método de Runge-Kutta implícito. ♣

Hay una situación intermedia; si $a_{ij} = 0$ para $j > i$, $i = 1, 2, \ldots, s$ (matriz triangular inferior), entonces cada k_i está definido individualmente por

$$k_i = f(x_n + c_i h, y_n + h \sum_{j=1}^{i} a_{ij} k_j), \qquad i = 1, 2, \ldots, s.$$

En vez de resolver en cada paso un sistema no lineal de dimensión ds, tendremos que resolver s sistemas no acoplados de dimensión d. Estos métodos se llaman *semi-implícitos*. Al escribir sus tableros se omiten los ceros por encima de la diagonal principal.

Ejemplo. El método RK-Lobatto IIIA de dos etapas, de tablero

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

es un método de Runge-Kutta semi-implícito. ♣

Para que un método de Runge-Kutta implícito esté bien determinado el sistema que define las etapas debe tener solución única. Se puede probar que esto es cierto si h es pequeño.

Ejemplo. El método de colocación de parametro $c_1 = 1/2$ es un método de Runge-Kutta implícito de tablero

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

La única etapa, k_1, es, si existe, un punto fijo de la aplicación

$$G(k_1) = f(x_n + \frac{h}{2}, y_n + \frac{h}{2} k_1).$$

Si f es Lipschitz con respecto de su segunda variable, con constante de Lipschitz L, se tiene que

$$\|G(k_1) - G(\hat{k}_1)\| \leq \frac{Lh}{2} \|k_1 - \hat{k}_1\|.$$

Así pues, si $h < 2/L$ la aplicación G es contractiva, y tiene por tanto un único punto fijo. ♣

Los métodos de Runge-Kutta son métodos de un paso con función de incremento

$$\phi_f(x_n, y_n; h) = \sum_{i=1}^{s} b_i k_i(x_n, y_n; h). \tag{5.1}$$

Puesto que las etapas k_i son evaluaciones de la función f, no es difícil convencerse de que, si f satisface las hipótesis (H_f), entonces ϕ_f verifica las condiciones $(\mathrm{H}_{\mathrm{MN}})$.

Ejemplo. Para el método de Euler modificado,

$$\begin{aligned}
&\|k_1(x_n, y_n; h) - k_1(x_n, \hat{y}_n; h)\| \le L\|y_n - \hat{y}_n\|,\\
&\|k_2(x_n, y_n; h) - k_2(x_n, \hat{y}_n; h)\|\\
&\qquad\qquad \le L(\|y_n - \hat{y}_n\| + \tfrac{h}{2}\|k_1(x_n, y_n; h) - k_1(x_n, \hat{y}_n; h)\|)\\
&\qquad\qquad \le L(1 + \tfrac{Lh}{2})\|y_n - \hat{y}_n\|,
\end{aligned}$$

de donde se obtiene inmediatamente la condición de Lipschitz deseada observando que en este caso $\phi_f(x_n, y_n; h) = k_2(x_n, y_n; h)$. ♣

Ejemplo. En el caso del método RK-Radau I de dos etapas,

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

se tiene que

$$\begin{aligned}
&\|k_1(x_n, y_n; h) - k_1(x_n, \hat{y}_n; h)\| \le L\|y_n - \hat{y}_n\|,\\
&\|k_2(x_n, y_n; h) - k_2(x_n, \hat{y}_n; h)\| \le L(\|y_n - \hat{y}_n\|\\
&\quad + \tfrac{h}{3}\|k_1(x_n, y_n; h) - k_1(x_n, \hat{y}_n; h)\| + \tfrac{h}{3}\|k_2(x_n, y_n; h) - k_2(x_n, \hat{y}_n; h)\|),\\
&\quad \le L(1 + \tfrac{Lh}{3})\|y_n - \hat{y}_n\| + \tfrac{Lh}{3}\|k_2(x_n, y_n; h) - k_2(x_n, \hat{y}_n; h)\|,
\end{aligned}$$

de forma que, si $h < 3/L$,

$$\|k_2(x_n, y_n; h) - k_2(x_n, \hat{y}_n; h)\| \leq \frac{L(1 + \frac{Lh}{3})}{1 - \frac{Lh}{3}} \|y_n - \hat{y}_n\|.$$

De aquí se concluye la condición de Lipschitz deseada utilizando que en este caso $\phi_f(x_n, y_n; h) = \frac{1}{4}k_1(x_n, y_n; h) + \frac{3}{4}k_2(x_n, y_n; h)$. ♣

Problemas

1. Sea $q(x) = \prod_{i=1}^{\nu}(x - c_i)$, $q_l(x) = q(x)/(x - c_l)$, $l = 1, \ldots, \nu$. Demostrar que el método de colocación de parámetros $c_1, \ldots, c_\nu$ es un método de Runge-Kutta implícito de tablero

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

con

$$a_{ij} = \int_0^{c_i} \frac{q_j(x)}{q_j(c_j)}\, dx, \quad b_i = \int_0^1 \frac{q_i(x)}{q_i(c_i)}\, dx, \qquad i, j = 1, \ldots, \nu.$$

2. Dado un método de Runge-Kutta (RK), comprobar que admite la formulación equivalente

$$Y_i = y_n + h \sum_{j=1}^{s} a_{ij} f(x_n + c_j h, Y_j), \qquad i = 1, \ldots, s, \tag{5.2}$$

$$y_{n+1} = y_n + h \sum_{i=1}^{s} b_i f(x_n + c_i h, Y_i), \qquad n = 0, \ldots, N - 1. \tag{5.3}$$

Demostrar, utilizando el Teorema de la Aplicación Contractiva, que para todo h pequeño el sistema de ecuaciones (5.2) tiene una única solución. *Sugerencia.* Demostrar el resultado usando la norma infinito.

3. Demostrar que la función de incremento ϕ_f de un método (RK) es continua en sus tres variables para todo h suficientemente pequeño.

4. Dado un método de Runge-Kutta explícito de s etapas, definimos $\alpha = \max\limits_{1\le i,j\le s} |a_{ij}|$, $\beta = \sum\limits_{i=1}^{s} |b_i|$. Sea L una constante de Lipschitz para f con respecto a su segunda variable.

   (i) Demostrar que

   $$\|k_i(x_n, y_n; h) - k_i(x_n, \hat{y}_n; h)\| \le L\nu_i \|y_n - \hat{y}_n\|,$$

   donde ν_i satisface la recurrencia

   $$\nu_1 = 1, \qquad \nu_i = 1 + h\alpha L \sum_{j=1}^{i-1} \nu_j, \quad i = 2, \dots, s.$$

   (ii) Demostrar que la solución de dicha recurrencia es $\nu_i = (1+h\alpha L)^{i-1}$.

   (iii) Concluir que la función de incremento del método $\phi_f(x_n, y_n; h)$ es Lipschitz respecto a su segunda variable con constante de Lipschitz $\beta L(1 + h_0\alpha L)^{s-1}$ para $h \in [0, h_0]$.

5. Consideramos el método de Runge-Kutta

0	1/4	−1/4
2/3	1/4	5/12
	1/4	3/4.

   (i) Comprobar que la correspondiente función de incremento satisface una condición de Lipschitz con respecto a su segunda variable.

   (ii) Comprobar que no es un método de colocación.

5.2. Métodos de Runge-Kutta: condiciones de orden

Los métodos de Runge-Kutta (RK) verifican la condición de la raíz, y son por tanto 0-estables. Por otra parte, cumplen obviamente la primera de las

condiciones de consistencia, (C1). Así pues, para que sean consistente, y por tanto convergentes, basta con comprobar si cumplen la segunda condición de consistencia, (C2).

Si hacemos $h = 0$ en la primera fórmula de (RK), obtenemos que $k_i(x_n, y_n; 0) = f(x_n, y_n)$ para todo $i = 1, 2, \ldots, s$. Por consiguiente, usando (5.1), se tiene

$$\phi_f(x, y; 0) = \left(\sum_{i=1}^{s} b_i \right) f(x, y).$$

Así, un método de Runge-Kutta es consistente (y por tanto convergente) si y sólo si

$$\sum_{i=1}^{s} b_i = 1. \tag{5.4}$$

¿Qué condiciones sobre los coeficientes garantizan que el método es consistente de orden p?

Tenemos que desarrollar el residuo

$$R_n = y(x_{n+1}) - (y(x_n) + h \sum_{i=1}^{s} b_i k_i(x_n, y(x_n); h))$$

alrededor de x_n en potencias de h. Así pues, lo que hay que hacer es obtener desarrollos en potencias de h para las funciones

$$F(h) := y(x_n + h), \qquad G(h) := y(x_n) + h \sum_{i=1}^{s} b_i k_i(x_n, y(x_n); h).$$

Si estos desarrollos coinciden hasta orden p, existe una constante C independiente de n tal que $\|R_n\| \leq Ch^{p+1}$, y se tendrá consistencia de orden al menos p.

Para que los desarrollos coincidan hasta un cierto orden, los coeficientes del método tienen que satisfacer ciertas condiciones. Como ejemplo obtendremos las condiciones de consistencia de orden 1 y de orden 2. Las condiciones para tener un orden de consistencia más alto se obtienen de manera similar[1].

[1]La obtención de las condiciones de orden se puede sistematizar – y simplificar – usando teoría de árboles. Quien primero se dio cuenta de esto fue el ya mencionado J. Butcher.

En primer lugar desarrollamos $F(h) = y(x_n + h)$. Se tiene que

$$F(h) = y(x_n) + hy'(x_n) + \frac{h^2}{2}y''(x_n) + \frac{h^3}{3!}y'''(\overline{\xi}_n),$$

donde $\overline{\xi}_n \in [x_n, x_{n+1}]$. Usando la EDO para expresar y' e y'' en términos de f,

$$\begin{aligned} F(h) &= y(x_n) + hf(x_n, y(x_n)) + \tfrac{h^2}{2}\tfrac{\partial f}{\partial x}(x_n, y(x_n)) \\ &\quad + \tfrac{h^2}{2}\sum_{J=1}^{d}\tfrac{\partial f}{\partial y^J}(x_n, y(x_n))f^J(x_n, y(x_n)) + \tfrac{h^3}{3!}y'''(\overline{\xi}_n). \end{aligned}$$

Por otra parte, $k_i(x_n, y(x_n); 0) = f(x_n, y(x_n))$ y

$$\begin{aligned} &\tfrac{dk_i}{dh}(x_n, y(x_n); h) = c_i\tfrac{\partial f}{\partial x}\big(x_n + c_i h, y(x_n) + h\sum_{j=1}^{s} a_{ij}k_j(x_n, y(x_n); h)\big) \\ &+ \sum_{J=1}^{d}\tfrac{\partial f}{\partial y^J}\big(x_n + c_i h, y(x_n) + h\sum_{j=1}^{s} a_{ij}k_j(x_n, y(x_n); h)\big)\big(\sum_{j=1}^{s} a_{ij}k_j^J(x_n, y(x_n); h) \\ &+ h\sum_{j=1}^{s} a_{ij}\tfrac{dk_j^J}{dh}(x_n, y(x_n); h)\big), \end{aligned}$$

expresión que evaluada en $h = 0$ produce

$$\frac{dk_i}{dh}(x_n, y(x_n); 0) = c_i\frac{\partial f}{\partial x}(x_n, y(x_n)) + \sum_{j=1}^{s} a_{ij}\sum_{J=1}^{d}\frac{\partial f}{\partial y^J}(x_n, y(x_n))f^J(x_n, y(x_n)).$$

Así, tenemos el siguiente desarrollo para $G(h)$,

$$\begin{aligned} G(h) &= y(x_n) + h\sum_{i=1}^{s} b_i k_i(x_n, y(x_n); h) \\ &= y(x_n) + h\sum_{i=1}^{s} b_i f(x_n, y(x_n)) + h^2\sum_{i=1}^{s} b_i c_i\tfrac{\partial f}{\partial x}(x_n, y(x_n)) \\ &\quad + h^2\sum_{i,j=1}^{s} b_i a_{ij}\sum_{J=1}^{d}\tfrac{\partial f}{\partial y^J}(x_n, y(x_n))f^J(x_n, y(x_n)) \\ &\quad + \tfrac{h^3}{2}\textstyle\sum_{i=1}^{s}\tfrac{d^2k_i}{dh^2}(x_n, y(x_n); \overline{\eta}), \end{aligned}$$

donde $\overline{\eta} \in [0, h]$.

Si $f \in C^2$, es fácil (aunque tedioso) probar que existe una constante C independiente de n tal que $\|y'''(\overline{\xi}_n)\|, \|\frac{d^2k_i}{dh^2}(x_n, y(x_n); \overline{\eta})\| \leq C$. Por tanto, comparando ambos desarrollos concluimos que la condición

$$\sum_{i=1}^{s} b_i = 1 \tag{5.5}$$

es suficiente para tener orden de consistencia al menos 1 y que si además

$$\sum_{i=1}^{s} b_i c_i = \frac{1}{2}, \qquad \sum_{i,j=1}^{s} b_i a_{ij} = \frac{1}{2}, \tag{5.6}$$

entonces el orden de consistencia es al menos 2.

Por otra parte, si consideramos el problema $y'(x) = 1$, $x \in [0, b]$, $y(0) = 0$, cuya solución es $y(x) = x$, se tiene que $R_n = h(1 - \sum_{i=1}^{s} b_i)$. Por consiguiente, si $\sum_{i=1}^{s} b_i \neq 1$ el método no tiene orden de consistencia 1. De manera análoga se demuestra que las condiciones (5.6) son necesarias para tener orden de consistencia 2.

Observación. La condición de consistencia coincide con la condición de consistencia de orden 1. ♠

Observación. En el caso problema autónomos la primera de las condiciones en (5.6) no es necesaria para tener orden 2. Basta con la segunda (además de, por supuesto, la condición de orden uno (5.5)).

El ejemplo anterior debería bastar para convencerse de que calcular condiciones de orden para un método de Runge-Kutta general es una tarea ardua. Así que intentamos simplificar un poco las cosas.

Cualquier problema (PVI) se puede escribir en forma autónoma a costa de aumentar la dimensión del espacio. En efecto, si definimos

$$\bar{y}(x) = \begin{pmatrix} x \\ y(x) \end{pmatrix}, \qquad \bar{\eta} = \begin{pmatrix} a \\ \eta \end{pmatrix}, \qquad \bar{f}(\bar{y}(x)) = \begin{pmatrix} 1 \\ f(x, y(x)) \end{pmatrix},$$

el problema (PVI) es equivalente al problema autónomo

$$\bar{y}'(x) = \bar{f}(\bar{y}(x)), \qquad \bar{y}(a) = \bar{\eta}.$$

Podríamos entonces pensar que basta con estudiar las condiciones de orden para problemas autónomos, lo que es un poco más fácil. Hay una dificultad:

en general un método de Runge-Kutta no produce la misma solución para un problema que para su equivalente autónomo. Sin embargo, se puede comprobar fácilmente que ambas soluciones numéricas coinciden si el método es consistente y satisface la *condición de suma por filas*

$$c_i = \sum_{j=1}^{s} a_{ij}, \qquad i = 1, 2 \ldots, s. \tag{5.7}$$

Por consiguiente, si un método cumple esta condición, podemos calcular el orden de consistencia del método sin pérdida de generalidad suponiendo que el problema es autónomo. En adelante, salvo que se mencione expresamente lo contrario, nos restringiremos a métodos que cumplan la condición de suma por filas.

Observación. Si se cumple la condición de suma por filas, las dos condiciones (5.6) de orden 2 coinciden. ♠

Problemas

1. Demostrar que las condiciones (5.6) son necesarias para tener orden de consistencia 2.
2. Demostrar que para que un método de Runge-Kutta que cumple la condición de suma por filas tenga orden de consistencia 3 es necesario y suficiente que, además de (5.5) y (5.6), cumpla las cinco condiciones[2]

$$\sum_{i=1}^{s} b_i c_i^2 = \frac{1}{3}, \qquad \sum_{i,j=1}^{s} b_i c_i a_{ij} = \frac{1}{3}, \qquad \sum_{i,j,l=1}^{s} b_i a_{ij} a_{il} = \frac{1}{3},$$
$$\sum_{i,j=1}^{s} b_i a_{ij} c_j = \frac{1}{3!}, \qquad \sum_{i,j,l=1}^{s} b_i a_{ij} a_{jl} = \frac{1}{3!}.$$

[2]En el caso de métodos que cumplen la condición de suma por filas (5.7) estas cinco condiciones se reducen a dos,

$$\sum_{i=1}^{s} b_i c_i^2 = \frac{1}{3}, \qquad b^T A c = \sum_{i,j=1}^{s} b_i a_{ij} c_j = \frac{1}{3!}.$$

Determinar cuáles de ellas son necesarias en el caso de problemas autónomos.

3. Restringiéndonos a ecuaciones escalares autónomas, demostrar que el método de Runge-Kutta explícito de tablero

$$\begin{array}{c|cccc} 0 & & & & \\ 1/2 & 1/2 & & & \\ 1/2 & 0 & 1/2 & & \\ 1 & 0 & 0 & 1 & \\ \hline & 1/6 & 1/3 & 1/3 & 1/6 \end{array}$$

tiene orden 4.

4. Comprobar que las soluciones numéricas por un método de Runge-Kutta del problema de valor inicial (PVI) y de su equivalente autónomo coinciden si el método es consistente y satisface la condición de suma por filas (5.7).

5. Consideramos el método de Runge-Kutta de tablero

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

Este método *no cumple la condición de suma por filas*. Determinar su orden de consistencia. ¿Tiene un orden de consistencia mayor cuando nos restringimos a problemas autónomos?

6. Considérese el método cuyo tablero es

$$\begin{array}{c|ccc} 0 & & & \\ 1 & 1 & & \\ 1 & 1/2 & 1/2 & \\ \hline & 3/6 & 1/6 & 2/6 \end{array}$$

(a) Probar que es de orden 2 y no es de orden 3 en general.

(b) Probar que sí es de orden 3 para todas las ecuaciones lineales de la forma $y' = My$ donde M es una matriz de constantes.

7. ¿Es posible encontrar un método que tenga orden tres (en general) para problemas *escalares* autónomos y que tenga orden menor (en general) cuando sean vectoriales?

5.3. Métodos de Runge-Kutta: limitaciones sobre el orden obtenible

¿Cuál es el mejor orden que podemos conseguir para un método de Runge-Kutta con un número de etapas s dado? Las condiciones de orden son relaciones no lineales sobre los coeficientes del método, por lo que la respuesta no es en absoluto trivial. No obstante, tenemos una cota superior para el orden que se puede conseguir si el método es explícito.

Teorema 5.1. *Un método de Runge-Kutta explícito de s etapas no puede tener orden mayor que s.*

Demostración. Aplicamos el método al problema escalar $y' = y$ en $[0, b]$, $y(0) = 1$, cuya solución es $y(x) = e^x$. Así pues,

$$\frac{d^p F}{dh^p}(0) = e^{x_n+h}\Big|_{h=0} = e^{x_n} = y(x_n).$$

En cuanto a G, aplicando la regla de Leibniz para la derivada de un producto obtenemos que

$$\frac{d^p G}{dh^p}(0) = p \sum_{j_1=1}^{s} b_{j_1}\, k_{j_1}^{(p-1)}\Big|_{h=0}.$$

Por otra parte, aplicando de nuevo la regla de Leibniz,

$$\begin{aligned} k_{j_1}^{(p-1)}\Big|_{h=0} &= (p-1)\sum_{j_2=1}^{s} a_{j_1 j_2}\, k_{j_2}^{(p-2)}\Big|_{h=0} \\ &= (p-1)(p-2)\sum_{j_2,j_3=1}^{s} a_{j_1 j_2} a_{j_2 j_3}\, k_{j_3}^{(p-3)}\Big|_{h=0} \\ &= \cdots = (p-1)!\sum_{j_2,j_3,\ldots,j_p=1}^{s} a_{j_1 j_2} a_{j_2 j_3}\ldots a_{j_{p-1} j_p}\, k_{j_p}^{(0)}\Big|_{h=0}. \end{aligned}$$

Para la EDO que estamos considerando $k_{j_p}^{(0)}\Big|_{h=0} = y(x_n)$, por lo que

$$\frac{d^p G}{dh^p}(0) = p!\sum_{j_1,\ldots,j_p=1}^{s} b_{j_1} a_{j_1 j_2} a_{j_2 j_3}\ldots a_{j_{p-1} j_p} y(x_n).$$

Concluimos que la condición

$$p!\sum_{j_1,\ldots,j_p=1}^{s} b_{j_1} a_{j_1 j_2} a_{j_2 j_3}\ldots a_{j_{p-1} j_p} = 1 \tag{5.8}$$

es necesaria para que el método sea de orden p.

Si el método es explícito, $a_{ij} = 0$ para $j \geq i$. Así, a menos que exista una sucesión $j_1, j_2, \ldots, j_p$ de enteros $1, 2, \ldots, s$ tal que $j_1 > j_2 > j_3 > \cdots > j_p$, se tendrá que $\sum_{j_1,\ldots,j_p=1}^{s} b_{j_1} a_{j_1 j_2} a_{j_2 j_3}\ldots a_{j_{p-1} j_p} = 0$, y la condición (5.8) no se podrá cumplir. Dicha sucesión sólo puede existir si $p \leq s$. □

¿Se puede conseguir orden $p = s$? Para $s = 1, 2, 3, 4$ se puede ver que sí, y es relativamente fácil encontrar familias de soluciones a las correspondientes condiciones de orden.

Ejemplo. Sea $s = 1$. La condición de orden 1 es en este caso $b_1 = 1$. El único método explícito de una etapa y orden 1 que cumple la condición de suma por filas es por tanto el método de Euler. ♣

Ejemplo. Sea $s = 2$. Las condiciones de orden 1 y 2 son en este caso $b_1 + b_2 = 1$ y $b_1 c_1 + b_2 c_2 = 1/2$ respectivamente (estamos suponiendo que se cumple la condición de suma por filas). Si el método es explícito, entonces $c_1 = 0$, y la condición de orden 2 queda que $b_2 c_2 = 1/2$. Hay pues una familia uniparamétrica

de métodos explícitos de 2 etapas y orden 2,

$$c_2 = \lambda, \qquad b_1 = 1 - \frac{1}{2\lambda}, \qquad b_2 = \frac{1}{2\lambda}, \qquad \lambda \in \mathbb{R},\ \lambda \neq 0.$$

♣

¿Qué pasa con $s = 5$? Para que un método sea de orden 5 se tienen que satisfacer 17 condiciones de orden. Si $s = 5$ tenemos sólo 15 parámetros libres. Sin embargo, el sistema es no lineal, así que podría haber solución. Pero no es así.

Teorema 5.2. *Para $p \geq 5$ no existe ningún método de Runge-Kutta explícito de orden p con $s = p$ etapas.*

Estas limitaciones sobre el orden obtenible que hemos mencionado se conocen como *barreras de Butcher*. Hay más barreras. Por ejemplo, se tiene el siguiente teorema.

Teorema 5.3. *(i) Para $p \geq 7$ no existe ningún método de Runge-Kutta explícito de orden p con $s = p + 1$ etapas.*

(ii) Para $p \geq 8$ no hay ningún método de Runge-Kutta explícito con $s = p + 2$ etapas.

Problemas

1. Consideramos el siguiente método creado por Heun:

$$y_{n+1} = y_n + \frac{h}{4}\Big(f(x_n, y_n) + 3f\big(x_n + \frac{2h}{3}, y_n + \frac{2h}{3} f(x_n + \frac{h}{3}, y_n + \frac{h}{3} f(x_n, y_n))\big)\Big).$$

   (i) Escribir su tablero.

   (ii) Determinar el orden de convergencia.

2. Consideramos el método de Runge-Kutta implícito

$$\begin{array}{c|cc} 0 & 1/4 & -1/4 \\ 2/3 & 1/4 & 5/12 \\ \hline & 1/4 & 3/4. \end{array}$$

(i) Comprobar que la correspondiente función de incremento satisface una condición de Lipschitz con respecto a su segunda variable.

(ii) Demostrar que es consistente de orden 3. Esto demuestra que puede haber métodos de Runge-Kutta con orden mayor que el número de etapas. Eso sí, tienen que ser implícitos.

3. (a) Demostrar que un método de Runge-Kutta de una etapa no puede tener orden mayor que 2.

(b) Demostrar que los métodos de Runge-Kutta de s etapas y orden p satisfacen

$$\sum_{i=1}^{s} b_i c_i^{l-1} = \frac{1}{l}, \qquad l = 1, \cdots, p.$$

Sugerencia. Aplicar el método a los PVI $y'(x) = x^{l-1}$, $y(0) = 0$, $l = 1, \ldots, p$, y estudiar el valor del residuo R_n para $n = 0$.

(c) Demostrar que un método de Runge-Kutta de dos etapas no puede tener orden mayor que cuatro.

5.4. Métodos lineales multipaso: definición y condiciones de orden

Un método es *lineal* de k pasos si se puede escribir en la forma

$$\sum_{j=0}^{k} \alpha_j y_{n+j} = h \sum_{j=0}^{k} \beta_j f(x_{n+j}, y_{n+j}), \qquad n = 0, \ldots, N-k, \qquad \text{(MLM)}$$

con $\alpha_k = 1$, $|\alpha_0| + |\beta_0| \neq 0$.

Notación. Para abreviar se suele escribir $f_n = f(x_n, y_n)$.

Si $\beta_k = 0$, el método es explícito. En este caso se puede despejar explícitamente cada y_{n+k} en términos de los k valores anteriores $y_n, \ldots, y_{n+k-1}$. Si se han almacenado los valores $f_n, \ldots, f_{n+k-2}$, una única evaluación de función, f_{n+k-1}, producirá y_{n+k}.

Si $\beta_k \neq 0$, el método es implícito. Para obtener y_{n+k}, en cada paso tendremos que resolver la ecuación no lineal

$$y_{n+k} = h\beta_k f(x_{n+k}, y_{n+k}) + g, \tag{5.9}$$

donde g es una función conocida de los valores ya calculados,

$$g = \sum_{j=0}^{k-1} (h\beta_j f_{n+j} - \alpha_j y_{n+j}).$$

Si el lado derecho de la igualdad (5.9) tiene una constante de Lipschitz M con respecto a y_{n+k} tal que $M < 1$, entonces este sistema de ecuaciones tiene una única solución y_{n+k}, que se puede aproximar tanto como se desee por medio de la iteración

$$y_{n+k}^{[\nu+1]} = h\beta_k f(x_{n+k}, y_{n+k}^{[\nu]}) + g, \qquad y_{n+k}^{[0]} \text{ arbitrario.} \tag{5.10}$$

Por el Teorema de la Aplicación Contractiva, se cumple que

$$\lim_{\nu\to\infty} y_{n+k}^{[\nu]} = y_{n+k}.$$

Si f tiene una constante de Lipschitz L con respecto a su segunda variable, podemos forzar que se cumpla la condición $M < 1$ tomando $h < 1/(\beta_k L)$. Si este valor de h fuera demasiado pequeño, se debería optar por otro método iterativo, como el de Newton.

Ya vimos que los coeficientes α_j de un método general (MN) se pueden especificar por medio de ρ, el primer polinomio característico del método. En

el caso de los métodos lineales multipaso se define un *segundo polinomio característico*,

$$\sigma(\zeta) = \sum_{j=0}^{k} \beta_j \zeta^j,$$

con el fin de caracterizar los coeficientes β_j.

¿Qué condiciones deben satisfacer los coeficientes de un método lineal multipaso para que sea consistente? La función de incremento del método (MLM) viene dada por

$$\begin{aligned}\phi_f(x_n, y_n, \ldots, y_{n+k-1}; h) = &\sum_{j=0}^{k-1} \beta_j f(x_n + jh, y_{n+j}) \\ &+\beta_k f(x_n + kh, -\sum_{j=0}^{k-1} \alpha_j y_{n+j} + h\phi_f(x_n, y_n, \ldots, y_{n+k-1}; h)).\end{aligned}$$

Si el método cumple la condición (C1), entonces

$$\begin{aligned}\phi_f(x, y(x), \ldots, y(x); 0) &= \sum_{j=0}^{k-1} \beta_j f(x, y(x)) + \beta_k f\left(x, -\sum_{j=0}^{k-1} \alpha_j y(x)\right) \\ &= \sum_{j=0}^{k} \beta_j f(x, y(x)).\end{aligned}$$

Así pues, el método será consistente si y sólo si

$$\sum_{j=0}^{k} \alpha_j = 0 \quad \text{y} \quad \sum_{j=0}^{k} \beta_j = \sum_{j=0}^{k} j\alpha_j.$$

En términos del primer y segundo polinomio característico estas condiciones se escriben como

$$\rho(1) = 0, \qquad \rho'(1) = \sigma(1).$$

¿Qué condiciones se deben cumplir para que el método sea consistente de orden p? En este caso el residuo viene dado por

$$\begin{aligned}R_n &= \sum_{j=0}^{k} (\alpha_j y(x_{n+j}) - h\beta_j f(x_{n+j}, y(x_{n+j})) \\ &= \sum_{j=0}^{k} (\alpha_j y(x_{n+j}) - h\beta_j y'(x_{n+j})), \qquad n = 0, \ldots, N-k.\end{aligned}$$

Si $f \in C^p$, tomando desarrollos de Taylor para $y(x_{n+j}) = y(x_n + jh)$ e $y'(x_{n+j}) = y'(x_n + jh)$ alrededor del punto $x = x_n$ se tiene que

$$y(x_{n+j}) = y(x_n) + y'(x_n)jh + \cdots + \frac{y^{(p)}(x_n)(jh)^p}{p!} + \frac{y^{(p+1)}(\overline{\xi}_{n,j})(jh)^{p+1}}{(p+1)!},$$
$$y'(x_{n+j}) = y'(x_n) + y''(x_n)jh + \cdots + \frac{y^{(p)}(x_n)(jh)^{p-1}}{(p-1)!} + \frac{y^{(p+1)}(\overline{\eta}_{n,j})(jh)^p}{p!},$$

para algunos puntos intermedios $\overline{\xi}_{n,j}, \overline{\eta}_{n,j} \in (x_n, x_n + jh)$, de donde

$$R_n = C_0 y(x_n) + C_1 y'(x_n)h + \cdots + C_p y^{(p)}(x_n)h^p + O(h^{p+1}),$$

donde

$$\begin{aligned} C_0 &= \sum_{j=0}^{k} \alpha_j, \\ C_1 &= \sum_{j=0}^{k} \alpha_j j - \sum_{j=0}^{k} \beta_j, \\ &\vdots \\ C_q &= \frac{1}{q!}\left(\sum_{j=0}^{k} \alpha_j j^q - q\sum_{j=0}^{k} \beta_j j^{q-1}\right), \qquad q > 1. \end{aligned}$$

Así pues, si $C_q = 0$, $q = 0, \ldots, p$, el método será consistente de orden al menos p. Por otra parte, si $C_{p+1} \neq 0$, el método no puede tener orden de consistencia $p+1$. En efecto, la solución del problema

$$y'(x) = \frac{x^p}{p!}, \quad 0 \leq x \leq b, \qquad y(0) = 0,$$

verifica $y^{(p+1)}(x) = 1$, y por consiguiente $R_n = C_{p+1}h^{p+1} \neq O(h^{p+2})$. El mismo ejemplo sirve para demostrar que si $C_{p+1} \neq 0$ tampoco hay convergencia de orden p.

Si un método lineal multipaso tiene orden de consistencia p, entonces el valor $C_{p+1} \neq 0$ recibe el nombre de *constante de error*.

Ejemplo. Consideramos el *método de de los tres octavos*,

$$y_{n+3} - y_n = \frac{h}{8}\left(3f_{n+3} + 9f_{n+2} + 9f_{n+1} + 3f_n\right).$$

Un sencillo cálculo muestra que $C_0 = C_1 = C_2 = C_3 = C_4 = 0$ y que $C_5 = -3/80 \neq 0$. Así, el método es consistente de orden 3. ♣

Observación. Las condiciones de consistencia del método lineal multipaso (MLM) coinciden con las condiciones de consistencia de orden 1. ♠

Problemas

1. Comprobar que si la función f cumple las condiciones (H_f), entonces la función de incremento ϕ_f del método lineal multipaso (MLM) satisface las hipótesis $(\mathrm{H}_{\mathrm{MN}})$.
2. Según hemos visto, los métodos de Runge-Kutta y los métodos lineales multipaso que son consistentes son también consistentes de orden 1. Dar un ejemplo de un método consistente que no sea consistente de orden 1.
3. Demostrar que si un método lineal multipaso es consistente de orden p, con constante de error $C_{p+1} \neq 0$, entonces

$$y_n = \frac{h^{p+1} n^{p+1}}{(p+1)!} - \frac{C_{p+1} h^{p+1} n}{\sum_{j=1}^{k} j\alpha_j}$$

es solución particular de la recurrencia que se obtiene al aplicar el método al PVI

$$y(x) = x^p/p!, \quad x \in [0, b], \qquad y(0) = 0.$$

Concluir que las condiciones de consistencia de orden p son también necesarias para la convergencia de orden p.

4. Probar que un MLM con polinomios característicos ρ y σ es de orden $p \geq 1$ si y sólo si existe una constante $c \neq 0$ tal que cuando $\zeta \to 1$

$$\rho(\zeta) - \sigma(\zeta) \log \zeta = c(\zeta - 1)^{p+1} + O(|\zeta - 1|^{p+2}).$$

5. Un método lineal de k pasos con orden de consistencia k se dice que es una BDF si $\sigma(\zeta) = \gamma\zeta^k$ para algún $\gamma \in \mathbb{R} \setminus \{0\}$. Determinar ρ y el valor

de γ. Comprobar que las BDF con $k = 2$ y $k = 3$ son convergentes[3].

6. Sea un MLM tal que $\rho(\zeta) = (\zeta - 1)(\zeta^2 + 1)$. Determinar $\sigma(\zeta)$ para que alcance el orden máximo posible. Calcular su constante de error.

7. Cuando un MLM es implícito, para ahorrar trabajo computacional podemos sustituir el valor y_{n+k} en el lado derecho de (MLM) por una aproximación de ese valor dada por un método explícito,

$$y^*_{n+k} = -\sum_{j=0}^{k-1} \hat{\alpha}_j y_{n+j} + h \sum_{j=0}^{k-1} \hat{\beta}_j f(x_{n+j}, y_{n+j}),$$

de manera que

$$y_{n+k} = -\sum_{j=0}^{k-1} \alpha_j y_{n+j} + h \left(\beta_k f(x_{n+k}, y^*_{n+k}) + \sum_{j=0}^{k-1} \beta_j f(x_{n+j}, y_{n+j}) \right).$$

Así pues, primero *predecimos* un valor de y_{n+k} y luego lo *corregimos*. El par predictor-corrector resultante es explícito, pero *no es* un MLM.

   (i) Demostrar que un par predictor-corrector obtenido a partir de dos MLM no puede tener mayor orden de consistencia que el método corrector del par.

   (ii) Determinar cuál es el mínimo orden de consistencia que debe tener el método predictor para garantizar que el orden de consistencia del par predictor-corrector coincide con el del método corrector.

8. Consideramos el par predictor-corrector

$$\begin{cases} y^*_{n+2} - 3\, y_{n+1} + 2\, y_n = \dfrac{h}{2} \big(f(x_{n+1}, y_{n+1}) - 3\, f(x_n, y_n) \big), \\ y_{n+2} - y_n = h \big(f(x_{n+2}, y^*_{n+2}) + f(x_n, y_n) \big). \end{cases}$$

   (i) Determinar el orden de consistencia.

   (ii) Estudiar si es convergente.

[3]Las BDF sólo son 0-estables si $k \leq 6$.

5.5. Métodos lineales multipaso: limitaciones sobre el orden obtenible

Los métodos lineales multipaso que verifican

$$\alpha_j = -\alpha_{k-j}, \qquad \beta_j = \beta_{k-j}$$

se conocen como métodos simétricos Es fácil comprobar, desarrollando el residuo alrededor del punto medio, que los métodos simétricos tienen orden par.

Ejemplo. Si integramos la EDO en (x_n, x_{n+2}) y aproximamos la integral mediante la regla de Simpson, llegamos al método lineal de dos pasos

$$y_{n+2} - y_n = \frac{h}{3}(f_n + 4f_{n+1} + f_{n+2}),$$

conocido como regla de Simpson. Si desarrollamos el residuo alrededor del punto medio, $x = x_{n+1}$, queda un desarrollo en potencias impares de h,

$$\begin{aligned} R_n &= y(x_{n+2}) - y(x_n) - \tfrac{h}{3}(y'(x_n) + 4y'(x_{n+1}) + y'(x_{n+2})) = \\ &= (2y'(x_{n+1})h + \tfrac{2}{3!}y'''(x_{n+1})h^3 + \tfrac{2}{5!}y^{(v)}(x_{n+1})h^5 + \cdots) \\ &\quad -\tfrac{h}{3}(6y'(x_{n+1}) + y'''(x_{n+1})h^2 + \tfrac{2}{4!}y^{(v)}(x_{n+1})h^4 + \cdots). \end{aligned}$$

Concluimos que $R_n = O(h^5)$; por consiguiente, el método es consistente de orden 4. ♣

Ejemplo. Consideramos la regla del trapecio

$$y_{n+1} - y_n = \frac{h}{2}(f_{n+1} + f_n).$$

Si desarrollamos alrededor del punto medio $x = x_{n+\frac{1}{2}} := x_n + \frac{h}{2}$, de nuevo queda un desarrollo en potencias impares de h,

$$\begin{aligned} R_n &= y(x_{n+1}) - y(x_n) - \tfrac{h}{2}(y'(x_n) + y'(x_{n+1})) = \\ &= (y'(x_{n+\frac{1}{2}})h + \tfrac{1}{24}y'''(x_{n+\frac{1}{2}})h^3 + \cdots) \\ &\quad -\tfrac{h}{2}(2y'(x_{n+\frac{1}{2}}) + \tfrac{1}{4}y'''(x_{n+\frac{1}{2}})h^2 + \cdots). \end{aligned}$$

Concluimos que $R_n = O(h^3)$; por consiguiente, el método es consistente de orden 2. ♣

Si un método es simétrico sólo será necesario considerar las condiciones de orden impar, pues las de orden par se cumplen necesariamente.

Ejemplo. El método

$$y_{n+2} - y_n = \frac{h}{3}\left(f(x_{n+2}, y_{n+2}) + 4f(x_{n+1}, y_{n+1}) + f(x_n, y_n)\right).$$

es un método lineal de dos pasos con

$$\alpha_0 = -1, \quad \alpha_1 = 0, \quad \alpha_2 = 1, \qquad \beta_0 = 1/3, \quad \beta_1 = 4/3, \quad \beta_2 = 1/3,$$

y es por tanto simétrico. Basta por tanto con comprobar las condiciones de orden impar. Tenemos que

$$\begin{aligned}
C_1 &= \textstyle\sum_{j=1}^{2} j\alpha_j - \sum_{j=0}^{2}\beta_j = 2 - \left(\frac{1}{3} + \frac{4}{3} + \frac{1}{3}\right) = 0 && \Rightarrow \text{orden al menos 2,}\\
C_3 &= \textstyle\frac{1}{3!}\left(\sum_{j=1}^{2} j^3\alpha_j - 3\sum_{j=1}^{2} j^2\beta_j\right) &&\\
&= \textstyle\frac{1}{3!}\left(8 - 3\left(\frac{4}{3} + \frac{4}{3}\right)\right) = 0 && \Rightarrow \text{orden al menos 4,}\\
C_5 &= \textstyle\frac{1}{5!}\left(\sum_{j=1}^{2} j^5\alpha_j - 5\sum_{j=1}^{2} j^4\beta_j\right) &&\\
&= \textstyle\frac{1}{5!}\left(32 - 5\left(\frac{16}{3} + \frac{4}{3}\right)\right) = -\frac{1}{90} \neq 0 && \Rightarrow \text{orden 4.}
\end{aligned}$$

♣

¿Cuál es el mejor orden que podemos conseguir para un método, para un número de pasos k dado? Puesto que un método lineal de k pasos viene determinado por $2k+1$ coeficientes, y las condiciones de orden de consistencia son relaciones lineales de estos coeficientes, existen métodos lineales de k pasos con orden de consistencia $p = 2k$; reciben el nombre de *maximales*. Sin embargo no sirven para nada, puesto que, salvo para $k = 1$ y $k = 2$, no son convergentes.

Teorema 5.4 (Primera barrera de Dahlquist (1956))**.** *El orden de convergencia p de un método lineal de k pasos 0-estable satisface:*

- $p \leq k+2$ *si* k *es par;*
- $p \leq k+1$ *si* k *es impar;*
- $p \leq k$ *si* $\beta_k \leq 0$ *(en particular si es explícito).*

Problemas

1. Demostrar que existen métodos lineales de k pasos con orden de consistencia $p = 2k$.
2. Sea ρ un polinomio mónico de grado k verificando $\rho(1) = 0$. Demostrar que:
   (a) Existe un único polinomio σ de grado menor o igual que k tal que el MLM dado por ρ y σ tiene orden de consistencia mayor o igual que $k+1$.
   (b) Existe un único polinomio σ de grado menor estricto que k tal que el MLM dado por ρ y σ tiene orden de consistencia mayor o igual que k.
3. Demostrar que si un método lineal de 3 pasos explícito es de orden 4, entonces $\alpha_0+\alpha_2 = 8$ y $\alpha_1 = -9$. Deducir de esto que el método no puede ser convergente.
4. Encontrar el orden de consistencia de los métodos
   (a) $y_{n+2} + y_{n+1} - 2y_n = \frac{h}{4}(f_{n+2} + 8f_{n+1} + 3f_n)$,
   (b) $y_{n+2} - y_n = \frac{2}{3}h(f_{n+2} + f_{n+1} + f_n)$,
   (c) $y_{n+4} - \frac{8}{19}(y_{n+3} - y_{n+1}) - y_n = \frac{6h}{19}(f_{n+4} + 4f_{n+3} + 4f_{n+1} + f_n)$.
   ¿Son convergentes? (El método (c) se conoce como método de Quade.)
5. Repetir el problema anterior para el método lineal multipaso cuyos polinomios característicos son

$$\rho(\zeta) = \zeta^4 - 1 \qquad \text{y} \qquad \sigma(\zeta) = \frac{14}{45}(\zeta^4+1) + \frac{64}{45}(\zeta^3+\zeta) + \frac{24}{45}\zeta^2.$$

6. De entre los métodos lineales multipaso de la forma

$$y_{n+3} + \alpha_2 y_{n+2} + \alpha_1 y_{n+1} + \alpha_0 y_n = h(\beta_2 f_{n+2} + \beta_1 f_{n+1} + \beta_0 f_n),$$

determinar los simétricos convergentes de mayor orden.

5.6. Experimentos numéricos

Consideramos una vez más el problema lineal

$$\begin{gathered} y'(x) = Ay(x) + B(x) \qquad \text{para } 0 \le x \le 10, \qquad y(0) = (2,3)^T, \\ A = \begin{pmatrix} -2 & 1 \\ 1 & -2 \end{pmatrix}, \qquad B(x) = \begin{pmatrix} 2\operatorname{sen} x \\ 2(\cos x - \operatorname{sen} x) \end{pmatrix}, \end{gathered} \tag{5.11}$$

cuya solución es

$$y = 2e^{-x}\begin{pmatrix} 1 \\ 1 \end{pmatrix} + \begin{pmatrix} \operatorname{sen} x \\ \cos x \end{pmatrix}. \tag{5.12}$$

Para empezar le aplicamos la regla implícita del punto medio, que es un método de Runge-Kutta que tiene por tablero

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

El correspondiente diagrama de eficiencia se muestra en la Figura 5.1. Vemos que, a pesar de tener sólo una etapa, el método tiene orden 2. Esto es cierto en general, y no sólo para este problema: los métodos implícitos de s etapas pueden alcanzar orden $2s$.

A continuación consideramos el método de Runge-Kutta de tablero

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

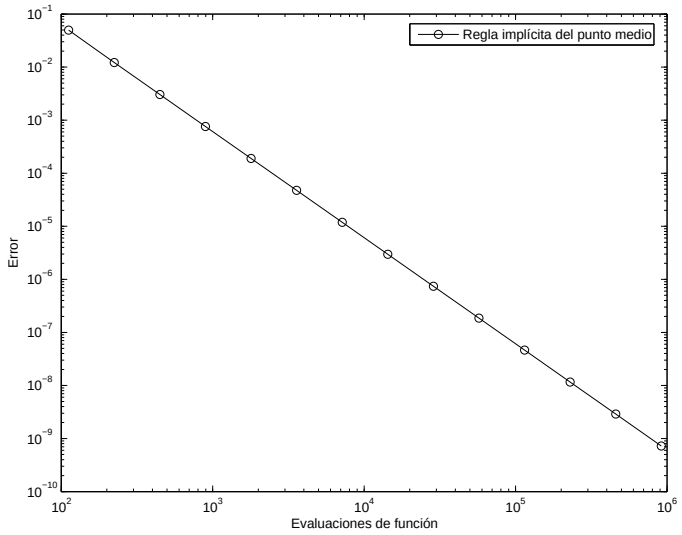

Figura 5.1: Diagrama de eficiencia para la regla implícita del punto medio aplicada al problema (5.11).

que no cumple la condición de suma por filas. Se lo aplicamos al problema (5.11) y a su equivalente autónomo. Los correspondientes diagramas de eficiencia se muestran en la Figura 5.2.

Se observa que el orden que se obtiene al aplicar el método al equivalente autónomo es 2, siendo sólo 1 al aplicárselo al problema original: para obtener condiciones de orden en métodos que no cumplan la condición de suma por filas no basta con considerar tan sólo problemas autónomos.

Seguidamente consideramos el método cuyo tablero es

$$\begin{array}{c|ccc} 0 & & & \\ 1 & 1 & & \\ 1 & 1/2 & 1/2 & \\ \hline & 3/6 & 1/6 & 2/6. \end{array} \tag{5.13}$$

Se puede probar sin mucha dificultad que es de orden 2 y no es de orden 3 en general. Sin embargo, sí es de orden 3 para todas las ecuaciones lineales de

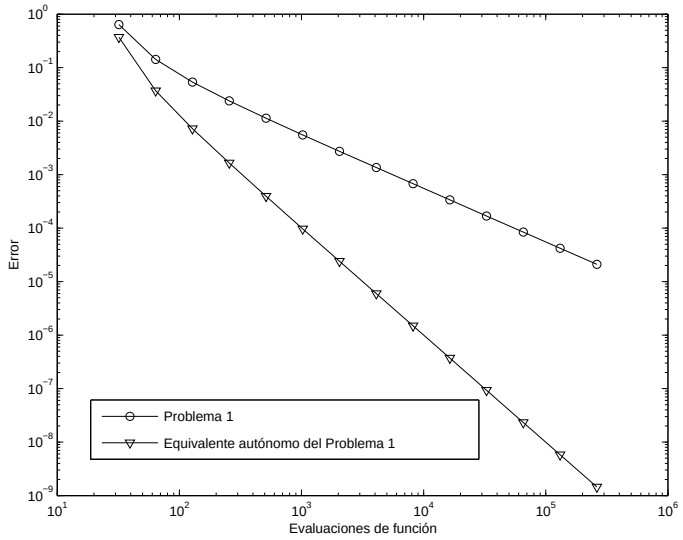

Figura 5.2: Diagrama de eficiencia para un método que no cumple la condición de suma por filas aplicado al problema (5.11) y a su equivalente autónomo.

la forma $y' = Ay$ donde A es una matriz de constantes. Para comprobar estas afirmaciones, aplicamos el método, por un lado al problema (5.11), y por otro al problema

$$y'(x) = \begin{pmatrix} -2 & 1 \\ 1 & -2 \end{pmatrix} y(x) \qquad \text{para } 0 \leq x \leq 10, \qquad y(0) = (2,3)^T,$$

cuya solución es

$$y = \frac{5}{2} e^{-x} \begin{pmatrix} 1 \\ 1 \end{pmatrix} + \frac{1}{2} e^{-3x} \begin{pmatrix} -1 \\ 1 \end{pmatrix}.$$

Los diagramas de eficiencia correspondientes se muestran en la Figura 5.3. Se comprueba que el orden es mejor cuando el método se aplica a un problema lineal homogéneo.

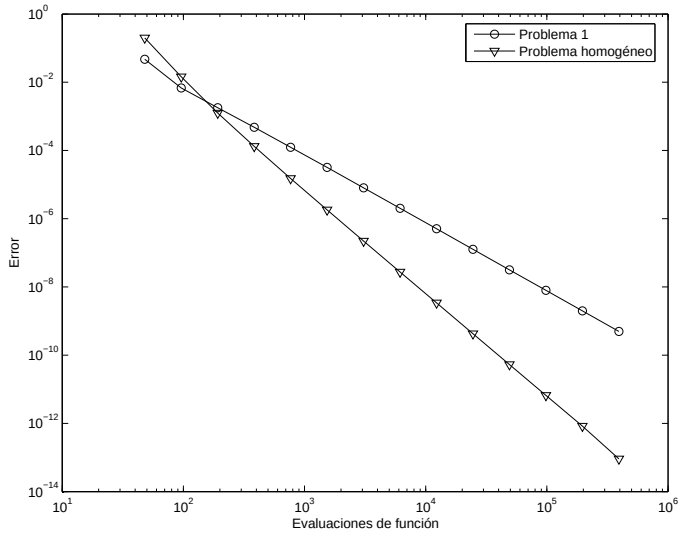

Figura 5.3: Diagrama de eficiencia para el método (5.13) aplicado al problema (5.11) y a un problema lineal homogéneo.

Problemas

1. Programar el método de Runge-Kutta semi-implícito Lobatto IIIA de dos etapas, de tablero

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

   resolviendo los sistemas no lineales que aparecen, por un lado por iteración de punto fijo, y por otro usando el método de Newton.

2. Programar el método de Runge-Kutta implícito Radau IA de dos etapas, de tablero

$$\begin{array}{c|cc} 0 & 1/4 & -1/4 \\ 2/3 & 1/4 & 5/12 \\ \hline & 1/4 & 3/4, \end{array}$$

resolviendo los sistemas no lineales que aparecen tanto por iteración de punto fijo como por el método de Newton.

3. Programar el método de Milne-Simpson de dos pasos para mallas uniformes,

$$y_{n+2} - y_n = \frac{h}{3}\left(f(x_{n+2}, y_{n+2}) + 4f(x_{n+1}, y_{n+1}) + f(x_n, y_n)\right),$$

usando el método de Newton a la hora de resolver ecuaciones no lineales. Hay que programarlo de manera que funcione para sistemas y de forma que se haga el menor número posible de evaluaciones de función en cada paso. El segundo valor de arranque, y_1, se calculará mediante el método de Runge-Kutta explícito de tablero

$$\begin{array}{c|ccc} 0 & & & \\ 1/3 & 1/3 & & \\ 2/3 & 0 & 2/3 & \\ \hline & 1/4 & 0 & 3/4, \end{array}$$

un método de orden 3 que se conoce como método de Heun.

Capítulo 6

Selección automática del paso

Hasta ahora nos hemos limitado a considerar mallas uniformes. Esto presenta un inconveniente importante: si la solución varía mucho en alguna zona del intervalo de integración, nos vemos obligados a dar pasos pequeños en *todo* el intervalo, con el consiguiente coste computacional. Sería preferible adaptar la malla a la solución, haciéndola más fina en aquellas regiones donde varíe rápidamente y más gruesa donde cambie poco. La dificultad estriba en que no conocemos la solución (¡estamos intentando calcularla!), y por tanto no sabemos a priori dónde necesitamos una malla más fina. ¿Qué hacer? Una posibilidad es que sea el propio método quien elija la malla, a medida que calcula la solución, de manera que el error cometido esté siempre bajo control.

6.1. Control del error global a través del error local

Supongamos que ya hemos calculado $y_0, \ldots, y_n$. Queremos avanzar de x_n a x_{n+1} con una longitud de paso $h_n = x_{n+1} - x_n$ tal que el *error global*,

$$y(x_{n+1}) - y_{n+1},$$

tenga tamaño menor que una tolerancia dada. La dificultad estriba en que no es posible controlar el error global directamente por medio de h_n. Lo que sí se puede controlar en muchos casos es el error cometido al dar el paso suponiendo que el valor y_n era exacto. Este error, conocido como *error local*, viene dado por

$$u_n(x_{n+1}) - y_{n+1},$$

donde u_n, la *solución local*, es la solución del problema de valor inicial

$$u_n'(x) = f(x, u_n(x)), \qquad u_n(x_n) = y_n.$$

¿Podemos controlar el error global a través del error local? El siguiente resultado muestra que sí.

Teorema 6.1. *Sea* $\mu = \max_{0\le n\le N-1}(\|u_n(x_{n+1}) - y_{n+1}\|/h_n)$ *y* L *una constante de Lipschitz para* f *con respecto a su segunda variable. Entonces*

$$\|y(x_n) - y_n\| \le e^{L(b-a)}\|y(x_0) - y_0\| + e^{L(b-a)}(b-a)\mu, \qquad n = 1, \dots, N.$$

Demostración. Se basa en la identidad

$$y(x_{n+1}) - y_{n+1} = (y(x_{n+1}) - u_n(x_{n+1})) + (u_n(x_{n+1}) - y_{n+1}).$$

Por la dependencia continua, lema 1.2,

$$\|y(x_{n+1}) - u_n(x_{n+1})\| \le e^{Lh_n}\|y(x_n) - y_n\|,$$

y por tanto

$$\|y(x_{n+1}) - y_{n+1}\| \le e^{Lh_n}\|y(x_n) - y_n\| + \mu h_n.$$

A partir de aquí se prueba fácilmente, por inducción, que

$$\|y(x_n) - y_n\| \le e^{L(x_n - x_0)}\|y(x_0) - y_0\| + e^{L(x_n - x_0)}(x_n - x_0)\mu,$$

de donde se deduce inmediatamente el resultado. □

Observación. La prueba tiene dos ingredientes: el control del error local y un control del crecimiento de la diferencia entre dos soluciones de la EDO. Esto muestra una limitación fundamental de cualquier método numérico: no importa cuán preciso sea al aproximar la solución en un paso dado, si el PVI es inestable, esto es, si la diferencia entre dos soluciones crece rápidamente con x, acabaremos calculando soluciones numéricas $\{y_n\}$ que estarán lejos de los verdaderos valores $y(x_n)$. ♠

Problemas

1. Queremos resolver el problema de valor inicial (PVI) en el intervalo $[0, 1]$ con función del lado derecho

$$f(x, y) = (x + \operatorname{sen} y^2, \frac{x^2}{2} + \cos y^1)^T,$$

y dato inicial $\eta = (0, 0)^T$. Suponiendo que el ordenador tiene precisión infinita, ¿qué tolerancia debemos tomar para el error local por unidad de paso para garantizar que el máximo error cometido sea a lo sumo 10^{-5}?

2. Consideramos la ecuación diferencial

$$y'(x) = -3(y(x) - \operatorname{sen} x) + \cos x, \qquad x \in [0, 2\pi].$$

   (a) Determinar las soluciones correspondientes a datos iniciales $y(0) = 0$ e $y(0) = \varepsilon$.

   (b) Supongamos que calculamos numéricamente la solución correspondiente al dato inicial $y(0) = 0$ con un método que selecciona automáticamente el paso, imponiendo una tolerancia para el error local por unidad de paso de 10^{-10}. Debido a la precisión finita del ordenador se comete un error en el dato inicial de 10^{-12}. ¿De cuántas cifras decimales de la solución nos podemos fiar?

3. Consideramos un problema de valor inicial en el que la función f del lado derecho satisface la condición de Lipschitz unilateral

$$< f(x, y) - f(x, \hat{y}), y - \hat{y} > \leq l\|y - \hat{y}\|_2^2$$

con constante $l < 0$. Demostrar que

$$\|y(x_n) - y_n\|_2 \leq e^{l(b-a)}\|y(x_0) - y_0\|_2 + (b-a)\mu, \qquad n = 1, \ldots, N,$$

siendo μ una cota para la norma euclídea del error local por unidad de paso.

6.2. Estimación del error local (métodos de un paso)

Motivados por este teorema, queremos escribir un código que elija automáticamente la longitud del paso de forma que la norma del error local por unidad de paso,

$$\|u_n(x_{n+1}) - y_{n+1}\|/h_n,$$

se ajuste a una tolerancia prescrita. Será por tanto necesario tener una forma de estimar esta cantidad.

Si un método de un paso es consistente de orden p y $f \in C^{p+1}$, se puede probar fácilmente que el error local es de orden h_n^{p+1}; es más, existe una función ψ_f tal que

$$u_n(x_{n+1}) - y_{n+1} = \psi_f(x_n, y_n)h_n^{p+1} + O(h_n^{p+2}).$$

Por consiguiente, podemos estimar el error local por medio del término principal de este desarrollo,

$$\texttt{errorl} \approx \|\psi_f(x_n, y_n)\|h_n^{p+1}.$$

Ejemplo. Supongamos que aplicamos Euler a un problema escalar. El error local satisface

$$u_n(x_{n+1}) - y_{n+1} = \frac{u_n''(x_n)}{2}\, h_n^2 + \frac{u_n'''(\xi_n)}{3!}\, h_n^3,$$

y podemos estimarlo por el término principal,

$$\psi_f(x_n, y_n)h_n^2 = \big(f_x(x_n, y_n) + f_y(x_n, y_n)f(x_n, y_n)\big)h_n^2/2.$$

La estimación supone dos evaluaciones de función, $f_x(x_n, y_n)$ y $f_y(x_n, y_n)$, ($f(x_n, y_n)$ ya se había evaluado para calcular y_{n+1}), un coste excesivo, pues no es razonable que estimar el error, una tarea en cierto sentido auxiliar, cueste el doble que calcular la solución numérica. ♣

El término principal del error local, $\psi_f(x_n, y_n)h_n^{p+1}$, se puede calcular, pero resulta demasiado costoso (véase el ejemplo anterior). Tendremos que pensar en otra forma más barata de estimarlo.

Supongamos que calculamos una segunda aproximación $\hat{y}_{n+1} \approx y(x_{n+1})$ por medio de un método de un paso más preciso,

$$u_n(x_{n+1}) - \hat{y}_{n+1} = \hat{\psi}_f(x_n, y_n)h_n^{\hat{p}+1} + O(h_n^{\hat{p}+2}), \qquad \hat{p} \geq p+1.$$

Por consiguiente,

$$\begin{aligned} \hat{y}_{n+1} - y_{n+1} &= \hat{y}_{n+1} - u_n(x_{n+1}) + u_n(x_{n+1}) - y_{n+1} \\ &= \psi_f(x_n, y_n)h_n^{p+1} + O(h_n^{p+2}), \end{aligned}$$

así que podemos estimar el error local por medio de

$$\texttt{errorl} \approx \|\psi_f(x_n, y_n)\|h_n^{p+1} \approx \|\hat{y}_{n+1} - y_{n+1}\|. \tag{6.1}$$

Una vez estimado, el error se compara con $\texttt{tol}h_n$, donde `tol` es la tolerancia prescrita para el error local por unidad de paso. Si $\texttt{errorl} > \texttt{tol}h_n$, el paso se rechaza, y se vuelve a calcular con un nueva longitud h_n' tal que

$$\|\psi_f(x_n, y_n)\|(h_n')^{p+1} = \texttt{tol}h_n'.$$

Dado que

$$\texttt{tol} = \|\psi_f(x_n, y_n)\|(h_n')^p = \|\psi_f(x_n, y_n)\|h_n^p \left(\frac{h_n'}{h_n}\right)^p \approx \frac{\texttt{errorl}}{h_n}\left(\frac{h_n'}{h_n}\right)^p,$$

vemos que la elección "óptima" es

$$h_n' = h_n\left(\frac{\texttt{tol}h_n}{\texttt{errorl}}\right)^{\frac{1}{p}}. \tag{6.2}$$

El proceso se repite hasta que $\texttt{errorl} \leq \texttt{tol}h_n$. En ese momento el paso se acepta. Se intenta ahora un nuevo paso con la longitud óptima (6.2).

Para tener un buen código hay que poner un poco más de cuidado. Se suele multiplicar el segundo miembro de (6.2) por un factor de seguridad `FAC`, normalmente $\texttt{FAC} = 0{,}9$, de forma que el `errorl` sea aceptable la siguiente vez con probabilidad alta. Tampoco se suelen permitir incrementos de paso muy grandes, con el fin de hacer el método más seguro. También es costumbre fijar una longitud de paso máxima, `HMAX`. Así, por ejemplo podemos poner

$$h_n' = \text{mín}\left(\texttt{HMAX}, h_n\,\text{mín}\left(\texttt{FACMAX}, \texttt{FAC}\left(\frac{\texttt{tol}h_n}{\texttt{errorl}}\right)^{\frac{1}{p}}\right)\right).$$

`FACMAX` se suele tomar entre 1,5 y 5.

Problemas

1. Calcular el término principal del desarrollo del error local para el método de Euler modificado. ¿Cuántas evaluaciones de función extra tendremos que hacer para estimar el error local calculando directamente esta cantidad?
2. Repetir el problema anterior para la regla del trapecio.

6.3. Pares encajados

¿Cómo elegir el segundo método para que el coste adicional de estimar el error no sea muy grande? Utilizamos una idea debida a Merson.

Sea un método de Runge-Kutta de orden p de tablero

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

La solución numérica por este método es

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

Si tuviéramos otro método de Runge-Kutta con las mismas etapas,

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

pero de distinto orden, $\hat{p} > p$, podríamos calcular la solución

$$\hat{y}_{n+1} = y_n + h_n \sum_{i=1}^{s} \hat{b}_i k_i$$

sin coste adicional, pues las etapas k_i coinciden. Esta solución más precisa se puede utilizar para estimar el error cometido con el primer método,

$$\texttt{errorl} \approx h_n \| \sum_{i=1}^{k} (\hat{b}_i - b_i) k_i \|. \tag{6.3}$$

A un par de métodos con estas características se le conoce como *par encajado*.

Ejemplo. El par encajado

$$\begin{array}{c|cc} 0 & & \\ 1 & 1 & \\ \hline y_{n+1} & 1 & 0 \\ \hline \hat{y}_{n+1} & \frac{1}{2} & \frac{1}{2} \end{array}$$

avanza con el método de Euler y estima con el método de Euler mejorado. Hemos dicho que no hay coste adicional, pero esto es un poco falaz, porque el método de Euler no usa la segunda etapa, mientras que el método que utilizamos para estimar, Euler mejorado, sí lo hace. Ahora bien, el método de avance es un método explícito que, además de cumplir la condición de suma por filas, tiene también la propiedad FSAL (First Same As Last),

$$a_{si} = b_i \qquad i = 1, \ldots, s.$$

En los métodos con esta propiedad la última etapa del paso n viene dada por

$$k_s(x_n, y_n; h_n) = f(x_n + c_s h_n, y_n + h_n \sum_{i=1}^{s-1} a_{si} k_i) = f(x_n + h_n, y_n + h_n \sum_{i=1}^{s-1} b_i k_i),$$

donde hemos usado la propiedad FSAL y las condiciones de suma por filas y de consistencia para probar que $c_s = \sum_{j=1}^{s} a_{sj} = \sum_{j=1}^{s} b_j = 1$. Por otra parte, por ser el método explícito, la primera etapa del paso $n+1$ viene dada por

$$k_1(x_{n+1}, y_{n+1}; h_{n+1}) = f(x_{n+1}, y_{n+1}) = f(x_n + h_n, y_n + h_n \sum_{i=1}^{s-1} b_i k_i).$$

Es decir, la primera etapa de cada paso coincide con la última del paso anterior. Por tanto la primera etapa sólo hay que calcularla en el primer paso, y desde el punto de vista computacional es como si el método tuviera una etapa menos. En el ejemplo que nos ocupa tenemos en la práctica una evaluación de función por paso, ¡lo mismo que para el método de Euler! ♣

Aunque la estimación (6.3) da el error de truncación para el método del par encajado de orden más bajo, si h_n es suficientemente pequeño el error para el método de orden alto será aún menor; así que, ¿por qué no avanzar con el método de orden más alto? A esto se le llama hacer *extrapolación local*. La desventaja es que si el método de avance original tenía la propiedad FSAL, ésta se pierde, con el coste computacional extra que ello supone.

Ejemplo. El par encajado

0		
1	1	
y_{n+1}	$\frac{1}{2}$	$\frac{1}{2}$
$\hat{y}_{n+1}$	1	0

avanza con el método de Euler mejorado y estima con el método de Euler. En este caso el método de avance es de orden 2, pero a cambio hay que hacer dos evaluaciones de función por paso. ♣

Notación. Un par encajado es de tipo $p(q)$ si el orden de y_{n+1} (el método que se utiliza para avanzar) es p, y el orden de $\hat{y}_{n+1}$ (el método que se usa para estimar) es q.

Problemas

1. Construir todos los pares encajados explícitos 1(2) con dos etapas, que satisfagan la condición de suma por filas y tales que el método de avance tenga la propiedad FSAL.
2. Hallar todos los pares encajados 2(3) de la forma

$$\begin{array}{c|ccc} 0 & & & \\ 1 & 1 & & \\ 1/2 & a_{31} & a_{32} & \\ \hline & b_1 & b_2 & 0 \\ \hline & \widehat{b}_1 & \widehat{b}_2 & \widehat{b}_3. \end{array}$$

   ¿Hay alguno tal que el método de avance cumpla la propiedad FSAL? ¿Y alguno que cumpla la condición de suma por filas?
3. Consideramos el par encajado

$$\begin{array}{c|ccc} 0 & & & \\ 1 & 1/2 & 1/2 & \\ 1/2 & 3/8 & 1/8 & \\ \hline & 1/6 & 1/6 & 4/6 \\ \hline & 1/2 & 1/2 & 0. \end{array}$$

   (i) Demostrar que el método de avance (implícito de tres etapas) es de orden 3. *Indicación:* Recuérdese que la condición

$$p! \sum_{j_1,\ldots,j_p=1}^{s} b_{j_1} a_{j_1 j_2} a_{j_2 j_3} \ldots a_{j_{p-1} j_p} = 1$$

   es necesaria para que un método de RK tenga orden p.

   (ii) Demostrar que el método de estimación (de dos etapas) tiene orden de consistencia 2.

6.4. Experimentos numéricos

A continuación se incluye el listado de un programa que calcula soluciones numéricas de PVI mediante el par encajado Euler/Euler mejorado que se ha descrito en la sección anterior.

```
function [x,y,evf] = parencajado12(ld,intervalo,y0,tol)

 % ld: función del lado derecho.
 % intervalo: vector fila que contiene el intervalo de integración.
 % y0: vector columna que contiene el dato inicial.
 % tol: tolerancia por unidad de paso 'tol'.
 % x: puntos de la malla en los que se ha calculado la solución.
 % y: solución numérica.
 % evf: evaluaciones de función

 a=intervalo(1); b=intervalo(2); hmax = (b-a)/10; facmax = 5; fac = 0.9;
 x(1) = a; y(:,1) = y0; n=1; evalfun = 0;
 k1 = feval(ld,x(n),y(:,n)); % Solamente se calcula en la primera etapa (FSAL)
 evalfun = evalfun+1;
 h = hmax;
 while x(n)<b
     k2 = feval(ld, x(n)+h, y(:,n)+h*k1);
     evalfun = evalfun+1;
     errorp = norm((k2-k1)/2,inf);
     if errorp<=tol % condicion que controla la tolerancia del error
         x(n+1) = x(n)+h;
         y(:,n+1) = y(:,n)+h*k1;
         n = n+1;
         k1 = k2;
     end
     if errorp==0 % condicion que impide la division por cero
         h = min ([hmax, h*facmax]);
     else
         h = min ([hmax, h*min([facmax, fac*(tol/errorp)^(1/1)])]);
     end
     if x(n)+h > b % condicion que impide escapar de la malla
         h = b-x(n);
     end
 end
```

Programa 6.3: Par encajado Euler / Euler mejorado.

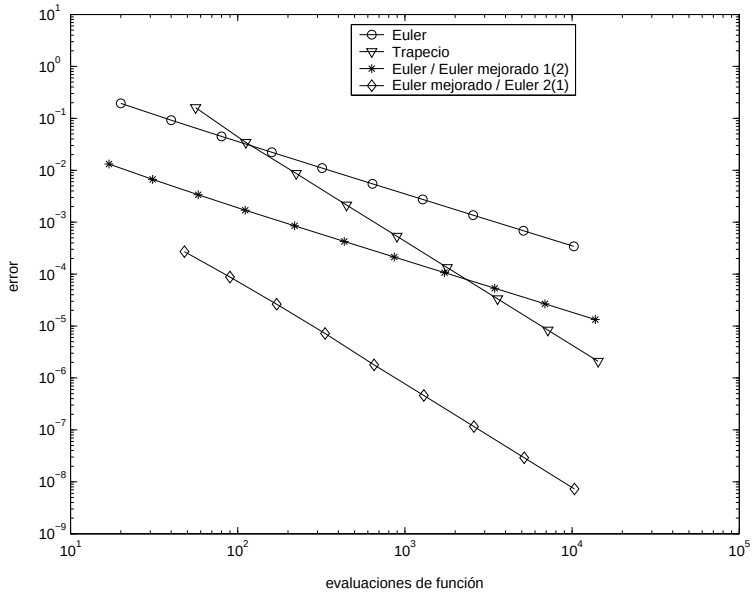

Figura 6.1: Diagramas de eficiencia para el Problema 1.

En la figura 6.1 representamos los diagramas de eficiencia para el problema lineal

$$y'(x) = Ay(x) + B(x) \quad \text{para } 0 \leq x \leq 10, \qquad y(0) = (2,3)^T,$$

$$A = \begin{pmatrix} -2 & 1 \\ 1 & -2 \end{pmatrix}, \qquad B(x) = \begin{pmatrix} 2\,\mathrm{sen}\, x \\ 2(\cos x - \mathrm{sen}\, x) \end{pmatrix},$$

cuya solución es

$$y = 2e^{-x} \begin{pmatrix} 1 \\ 1 \end{pmatrix} + \begin{pmatrix} \mathrm{sen}\, x \\ \cos x \end{pmatrix}.$$

correspondientes a los dos pares encajados que hemos descrito en la sección anterior, así como al método de Euler y la regla del trapecio, estos dos últimos con malla prefijada uniforme. Siendo ambos de orden 1, el par encajado Euler/Euler mejorado 1(2) trabaja diez veces menos que el método de Euler para conseguir un mismo error. Si comparamos la regla del trapecio con el par Euler mejorado/Euler 2(1), dos métodos de orden 2, se tiene un factor de

ahorro de coste similar. Por otra parte, los órdenes empíricos responden a lo esperado, es decir, orden 1 para el par Euler/Euler mejorado 1(2) y orden 2 cuando se hace extrapolación local. A pesar de ser tan sólo de orden 1, el par Euler/Euler mejorado 1(2) es capaz de competir con la regla del trapecio, a menos que queramos errores menores que 10^{-4}.

Problemas

1. Programar una función de Matlab que calcule soluciones numéricas para sistemas con el par encajado (debido a Fehlberg) rkf2(3), de tablero

$$\begin{array}{c|ccc} 0 & & & \\ 1 & 1 & & \\ \frac{1}{2} & \frac{1}{4} & \frac{1}{4} & \\ \hline y_{n+1} & \frac{1}{2} & \frac{1}{2} & 0 \\ \hline \hat{y}_{n+1} & \frac{1}{6} & \frac{1}{6} & \frac{4}{6}. \end{array}$$

   La estructura de las variables de entrada y salida será la misma que la del programa `parencajado12` que se os ha suministrado.

2. Repetir el ejercicio anterior para el par encajado rk2(3)b, también debido a Fehlberg, de tablero

$$\begin{array}{c|cccc} 0 & & & & \\ \frac{1}{4} & \frac{1}{4} & & & \\ \frac{27}{40} & -\frac{189}{800} & \frac{729}{800} & & \\ 1 & \frac{214}{891} & \frac{1}{33} & \frac{650}{891} & \\ \hline y_{n+1} & \frac{214}{891} & \frac{1}{33} & \frac{650}{891} & 0 \\ \hline \hat{y}_{n+1} & \frac{533}{2106} & 0 & \frac{800}{1053} & -\frac{1}{78}. \end{array}$$

   Al programar el método es importante aprovechar la propiedad FSAL.

3. Modificar el programa `parencajado12` que se os ha suministrado de manera que haga extrapolación local. Hacer un diagrama de eficiencia para el método de Euler con paso fijo, el par Euler/Euler mejorado (programa `parencajado12`) y el par Euler mejorado/Euler (extrapolación local) aplicados al problema del péndulo (usar `ode45` para calcular la solución "exacta").
4. Programar el par encajado, debido a Dormand y Prince, rkdp54, de tablero

$$
\begin{array}{c|ccccccc}
0 & & & & & & & \\
\frac{1}{5} & \frac{1}{5} & & & & & & \\
\frac{3}{10} & \frac{3}{40} & \frac{9}{40} & & & & & \\
\frac{4}{5} & \frac{44}{45} & -\frac{56}{15} & \frac{32}{9} & & & & \\
\frac{8}{9} & \frac{19372}{6561} & -\frac{25360}{2187} & \frac{64448}{6561} & -\frac{212}{729} & & & \\
1 & \frac{9017}{3168} & -\frac{355}{33} & \frac{46732}{5247} & \frac{49}{176} & -\frac{5103}{18656} & & \\
1 & \frac{35}{384} & 0 & \frac{500}{1113} & \frac{125}{192} & -\frac{2187}{6784} & \frac{11}{84} & \\
\hline
y_{n+1} & \frac{35}{384} & 0 & \frac{500}{1113} & \frac{125}{192} & -\frac{2187}{6784} & \frac{11}{84} & \\
\hline
\hat{y}_{n+1} & \frac{5179}{57600} & 0 & \frac{7571}{16695} & \frac{393}{640} & -\frac{92097}{339200} & \frac{187}{2100} & \frac{1}{40}.
\end{array}
$$

   Al programar el método es importante aprovechar la propiedad FSAL. Hacer un diagrama de eficiencia que compare el rendimiento de esta función y el del integrador de Matlab `ode45` al aplicarlos al problema del péndulo.
5. Programar el par encajado

$$
\begin{array}{c|ccc}
0 & & & \\
1 & 1/2 & 1/2 & \\
1/2 & 3/8 & 1/8 & \\
\hline
 & 1/6 & 1/6 & 4/6 \\
\hline
 & 1/2 & 1/2 & 0,
\end{array}
$$

resolviendo los sistemas no lineales resultantes, bien por iteración de punto fijo, bien por el método de Newton.

Capítulo 7

Problemas stiff

7.1. ¿Qué es un problema stiff?

Sea $p \in C^\infty([0,\infty))$ una función de variación lenta (para ser más precisos, pediremos que p'' tenga un tamaño moderado). Para todo $\lambda \in \mathbb{C}$ consideramos el problema lineal escalar[1]

$$\begin{cases} y'(x) = \lambda(y(x) - p(x)) + p'(x), & x \geq 0, \\ y(0) = \eta, \end{cases} \tag{7.1}$$

cuya solución es

$$y(x) = p(x) + (y(0) - p(0))e^{\lambda x}.$$

Al aplicarle el método de Euler obtenemos la recurrencia

$$y_{n+1} = (1 + h\lambda)y_n + h\big(-\lambda p(x_n) + p'(x_n)\big).$$

La solución teórica satisface la misma recurrencia, pero con una perturbación,

$$y(x_{n+1}) = (1 + h\lambda)y(x_n) + h\big(-\lambda p(x_n) + p'(x_n)\big) + R_n.$$

[1] Cualquier ecuación lineal, $y'(x) = \lambda y(x) + g(x)$, se puede escribir en esta forma, siendo $p = p(x)$ cualquier solución particular de la ecuación.

La perturbación (el residuo) satisface

$$R_n = \frac{h^2}{2} y''(\overline{\xi}_n).$$

Restando, vemos que el en el punto x_n, $e_n = y(x_n) - y_n$, resuelve la recurrencia[2]

$$e_{n+1} = (1 + h\lambda)e_n + R_n.$$

Así, al error cometido en el punto x_{n+1} contribuyen, por un lado el residuo, y por otro el error cometido en el punto anterior amplificado por un factor $1 + h\lambda$. Por tanto, para que el error global permanezca pequeño hacen falta dos cosas:

- Que los residuos R_n sean pequeños. Esto impone una restricción sobre h de la forma $h \leq h_p$.
- Que el módulo del factor de amplificación $1 + h\lambda$ sea menor que 1. Esto impone una restricción de la forma $h \leq h_e$.

Los subíndices p y e indican que las restricciones tienen que ver, respectivamente, con la precisión y la estabilidad.

No todos los métodos tienen restricciones por motivos de estabilidad. Consideremos por ejemplo la regla del trapecio. Al aplicársela a nuestro problema obtenemos

$$\left(1 - \frac{h\lambda}{2}\right) y_{n+1} = \left(1 + \frac{h\lambda}{2}\right) y_n + \frac{h}{2}\big(- \lambda p(x_n) + p'(x_n) - \lambda p(x_{n+1}) + p'(x_{n+1})\big),$$

mientras que la solución teórica satisface

$$\left(1 - \frac{h\lambda}{2}\right) y(x_{n+1}) = \left(1 + \frac{h\lambda}{2}\right) y(x_n) + \frac{h}{2}\big(- \lambda p(x_n) + p'(x_n) - \lambda p(x_{n+1}) + p'(x_{n+1})\big) + R_n,$$

[2]La solución de esta recurrencia es $e_n = (1 + h\lambda)^n e_0 + \sum_{m=1}^{n} (1 + h\lambda)^{n-m} R_{m-1}$.

con un residuo dado por

$$R_n = \left(\frac{y'''(\overline{\xi}_n)}{3!} - \frac{y'''(\overline{\eta}_n)}{4} \right) h^3.$$

Así pues, en este caso el error en el punto x_n es solución de la recurrencia

$$e_{n+1} = \left(\frac{1 + \frac{h\lambda}{2}}{1 - \frac{h\lambda}{2}} \right) e_n + \frac{R_n}{1 - \frac{h\lambda}{2}}.$$

Si Re $\lambda < 0$, entonces

$$\left| \frac{1 + \frac{h\lambda}{2}}{1 - \frac{h\lambda}{2}} \right| < 1,$$

y no hay ninguna restricción por motivos de estabilidad.

¿De qué depende el valor de h_p? De tres cosas:

- el método que se usa, que determina la forma del residuo;
- la suavidad de la solución, que determina el tamaño de la derivada que aparece en el residuo;
- la precisión requerida.

Por el contrario, el valor de h_e sólo depende del método y del valor de λ.

¿Qué restricción es más fuerte, la impuesta por la precisión o la impuesta por la estabilidad? Eso depende del problema (y del método). Veámoslo con nuestro ejemplo.

En función del tamaño de Re λ, podemos distinguir tres casos:

- Si Re $\lambda \gg 1$, dos soluciones cualesquiera se separan muy rápidamente con x: el problema es muy inestable. No se puede esperar que *ningún* método numérico aplicado a un problema como éste funcione bien, ya que los errores que se produzcan crecerán rápidamente con x. Observemos que la derivada segunda,

$$y''(x) = p''(x) + (\eta - p(0))\lambda^2 e^{\lambda x},$$

es muy grande incluso para x pequeño. Si se usa el método de Euler, haría falta tomar un h tan pequeño, tanto para hacer pequeño el residuo como para evitar inestabilidades, que el cálculo efectivo será impracticable.

- Si $|\operatorname{Re}\lambda|$ es pequeña, las curvas solución del problema son más o menos paralelas en intervalos de x moderados, y se dice que el problema tiene estabilidad neutra. La derivada segunda tiene un tamaño moderado (en intervalos de x moderados) y las restricciones impuestas por la precisión y la estabilidad serán ambas poco exigentes para cualquier método razonable.
- Si $\operatorname{Re}\lambda \ll -1$, todas las curvas solución se parecen al cabo de muy poco tiempo. La ecuación es muy estable. Pasada una corta etapa *transitoria*, la solución se parece mucho a $p(x)$. Para x pequeño (en la fase transitoria), y'' es grande, debido al término $\lambda^2 e^{\lambda x}$, lo que fuerza a h a ser muy pequeño si queremos mantener pequeño el residuo. Cuando x crece, $\lambda^2 e^{\lambda x} \to 0$ rápidamente, y se tiene que $y''(x) \approx p''(x)$, lo que permite que h sea grande si lo único que nos preocupa es hacer pequeño el residuo. Sin embargo, para que $|1 + h\lambda| < 1$ el paso ha de ser muy pequeño. En este caso es la estabilidad la que impone una mayor limitación sobre el tamaño del paso.

Como acabamos de ver, para algunos problemas puede suceder que la restricción para evitar inestabilidades con algunos métodos sea mucho más exigente que la restricción que impone la precisión. Estos son precisamente los problemas *stiff* (rígidos, en inglés).

"Definición" 7.1. *Decimos que el sistema*

$$y'(x) = f(x, y(x)), \qquad x \geq x_0, \qquad y(x_0) = y_0,$$

es stiff *si su resolución numérica por algunos métodos exige una disminución significativa del paso de integración para evitar inestabilidades.*

Es evidente que ésta no es una definición matemática correcta; pero tampoco estamos intentando demostrar teoremas del tipo *si un sistema es stiff*

entonces... Lo importante de este concepto es que nos debe ayudar a elegir y diseñar métodos numéricos. En particular, para obtener una solución numérica de un problema stiff parece razonable usar un método que, como la regla del trapecio, no imponga restricciones por razones de estabilidad.

Problemas

1. Aplicamos el método de Euler mejorado al problema (7.1). ¿Impone alguna limitación al paso h por motivos de estabilidad? En caso afirmativo, ¿cuál?
2. Repetir el problema anterior para el método de Euler implícito, $y_{n+1} = y_n + hf(x_{n+1}, y_{n+1})$.

7.2. Dominio de estabilidad lineal y A-estabilidad

En la sección anterior hemos aplicado el método de Euler y la Regla del Trapecio al problema (7.1) con paso $h > 0$. En ambos casos, el factor de amplificación del error coincide con el factor de amplificación (al pasar de x_n a x_{n+1}) de la solución numérica de la ecuación lineal escalar[3]

$$y' = \lambda y, \quad x \geq 0, \qquad y(0) = 1. \tag{7.2}$$

Si, como es de desear, el factor de amplificación tiene módulo menor que 1, entonces $\lim_{n\to\infty} y_n = 0$, y recíprocamente. Esto es cierto no sólo para los dos métodos mencionados, sino para otros muchos métodos (incluyendo los métodos de Runge-Kutta y los métodos lineales multipaso), lo que motiva la siguiente definición.

Definición 7.2. *El* dominio de estabilidad lineal *(también llamado región de estabilidad absoluta), $\mathcal{D}$, de un método numérico es el conjunto de todos los números $z = h\lambda$ con $h > 0$, $\lambda \in \mathbb{C}$, tales que la solución numérica de* (7.2) *verifica que* $\lim_{n\to\infty} y_n = 0$.

[3]La solución exacta es $y(x) = e^{\lambda x}$, y por tanto $\lim_{x\to\infty} y(x) = 0$ si y sólo si $\operatorname{Re}\lambda < 0$.

De acuerdo con esta definición, para que un método numérico aplicado al problema (7.1) no amplifique errores, tendremos que pedir que la longitud del paso sea tal que $h\lambda$ caiga en el dominio de estabilidad lineal del método.

Ejemplo. Al aplicar el método de Euler a (7.2) obtenemos la sucesión

$$y_n = (1 + h\lambda)^n, \qquad n = 0, 1, \ldots,$$

que es geométrica. Así pues, $\lim_{n\to\infty} y_n = 0$ si y sólo si $|1 + h\lambda| < 1$. Concluimos que $\mathcal{D} = \{z \in \mathbb{C} : |1 + z| < 1\}$. ♣

Ejemplo. Para la regla del trapecio tenemos que

$$y_n = \left(\frac{1 + \frac{1}{2}h\lambda}{1 - \frac{1}{2}h\lambda}\right)^n, \qquad n = 0, 1, \ldots,$$

que es una progresión geométrica. El dominio de estabilidad lineal es obviamente

$$\mathcal{D} = \left\{ z \in \mathbb{C} : \left|\frac{1 + \frac{1}{2}z}{1 - \frac{1}{2}z}\right| < 1 \right\}.$$

Es fácil verificar entonces que $\mathcal{D} = \{z \in \mathbb{C} : \operatorname{Re} z < 0\}$. ♣

El concepto de estabilidad lineal tiene, como veremos a continuación, un alcance mayor del que podría parecer en un principio.

Supongamos que queremos resolver un *sistema* lineal

$$\begin{cases} y'(x) = A(y(x) - p(x)) + p'(x), & x \geq 0, \\ y(0) = \eta, \end{cases}$$

donde A una matriz $d \times d$ arbitraria, por el método de Euler (para otros métodos numéricos se puede hacer un razonamiento parecido). El error global satisface la recurrencia

$$e_{n+1} = (I + hA)e_n + R_n.$$

Para que el método no amplifique errores queremos que $\|I + hA\| < 1$. La condición necesaria y suficiente para que esto sea cierto es que $|1 + h\lambda_i| < 1$ para

todos los autovalores λ_i, $i = 1, \ldots, d$ de la matriz A. Es decir, todos los productos $h\lambda_1, \ldots, h\lambda_d$ deben estar en el dominio de estabilidad lineal del método de Euler. Esto significa en la práctica que el tamaño del paso está limitado por aquel autovalor que sea peor desde el punto de vista de la estabilidad.

El dominio de estabilidad lineal $\mathcal{D}$ es también importante para sistemas no lineales

$$y' = f(x, y(x)), \qquad x \geq x_0, \qquad y(x_0) = \eta.$$

Desarrollando por Taylor, es fácil ver que las soluciones de la ecuación que están cerca de la solución $p(x)$ del problema (que son las que pueden intervenir al intentar resolver el problema numéricamente) verifican

$$\begin{aligned} y'(x) &\approx D_y f(x, p(x))(y(x) - p(x)) + f(x, p(x)) \\ &= D_y f(x, p(x))(y(x) - p(x)) + p'(x), \end{aligned}$$

donde la matriz jacobiana $D_y f$ tiene en su entrada (i, j) la derivada parcial de la i-ésima componente de f con respecto a la componente j-ésima de y. Si $D_y f(x, p(x))$ varía lentamente con x, podemos aproximarla (localmente en x) por una matriz constante, con lo que caemos en uno de los casos ya estudiados antes. Así, deberíamos pedir que en el paso n-ésimo

$$h\lambda_{n,1}, h\lambda_{n,2}, \ldots, h\lambda_{n,d} \in \mathcal{D},$$

donde $\lambda_{n,1}, \ldots, \lambda_{n,d}$ son los autovalores de la matriz jacobiana $D_y f$ evaluada en el punto (x_n, y_n).

Cuanto mayor sea la región de $\mathbb{C}^- := \{z \in \mathbb{C} : \operatorname{Re} z < 0\}$ cubierta por $\mathcal{D}$, menos restricciones para h debidas a posibles inestabilidades (recuérdense los ejemplos de Euler y la regla del trapecio). Eso motiva la siguiente definición.

Definición 7.3. *Un método es A-estable si*

$$\mathbb{C}^- \subseteq \mathcal{D}.$$

Es decir, si un método es A-estable podemos elegir la longitud de paso h (al menos para sistemas lineales) sólo por motivos de precisión, sin restricciones debidas a la inestabilidad. Estos métodos serán buenos para problemas *stiff*.

Problemas

1. Determinar el dominio de estabilidad lineal del método de colocación de parámetro $c_1 = 1/2$, (3.13). ¿Es A-estable?
2. Considérese el dominio de estabilidad lineal para el método de Euler mejorado. Estudiar si es simétrico con respecto al punto $z = -1$ (es decir, invariante por giros de 180º alrededor de dicho punto), con respecto a la recta Im $z = 0$, y con respecto a la recta Re $z = -1$.
3. Hallar el supremo de los $h > 0$ tales que al aplicar el método de Euler al problema
$$y' = \begin{pmatrix} -1 & 1 \\ -1 & -1 \end{pmatrix} y,$$
se cumple que $\lim_{n\to\infty} y_n = 0$ (para cada h fijado) cualquiera que sea la condición inicial $y(0) = y_0 \in \mathbb{R}^2$.
4. Al aplicar el método de Euler al PVI
$$y'(x) = -\lambda y(x), \quad x \in [0, 10], \qquad y(0) = 1, \tag{7.3}$$
donde λ es un número real positivo, obtenemos el diagrama que aparece en la Figura 7.1. ¿Cuál es el valor de λ?
5. Al aplicar el método de Euler al PVI
$$y'(x) = \begin{pmatrix} -\lambda & 1 \\ 0 & -1/10 \end{pmatrix} y(x), \quad x \in [0, 10], \qquad y(0) = (1, 999/10)^T, \tag{7.4}$$
donde λ es un número real positivo, obtenemos el diagrama que aparece en la Figura 7.2. ¿Cuál es el valor de λ?
6. Consideramos el θ-método
$$y_{n+1} = y_n + h(\theta f_n + (1 - \theta) f_{n+1}).$$

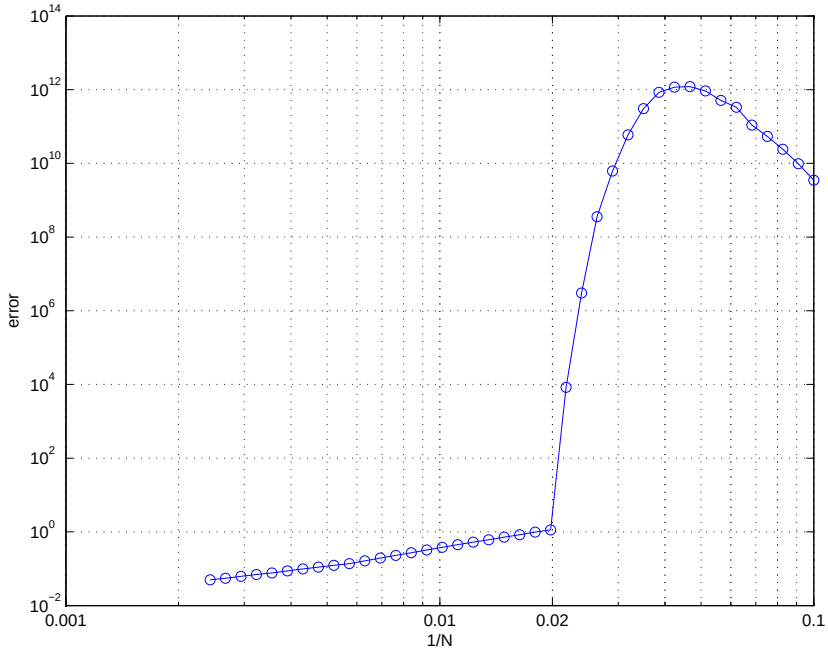

Figura 7.1: Diagrama de eficiencia para el método de Euler aplicado al problema (7.3).

(a) Determinar para qué valores de θ es A-estable.

(b) Fijado $h > 0$, ¿es posible elegir θ para que la solución numérica producida por este método para el PVI

$$y' = \lambda y \quad \text{en } [a, b], \qquad y(a) = \eta,$$

coincida con la solución exacta? Con el valor de θ así elegido, ¿se obtiene un método A-estable?

7. Se dice que un método es $L-$estable si es A-estable y, además, al aplicarlo a $y' = \lambda y$ se cumple $y_{n+1} = R(\lambda h)y_n$ con $R(z) \to 0$ cuando Re $z \to -\infty$.
a) Demostrar que la regla del trapecio no es $L-$estable mientras que el método implícito $y_{n+1} = y_n + hf(x_{n+1}, y_{n+1})$ sí lo es.
b) Explicar en qué sentido los métodos $L-$estables imitan mejor el decaimiento rápido de la solución de $y' = \lambda y$ cuando $-$Re λ es grande en comparación con h^{-1}.

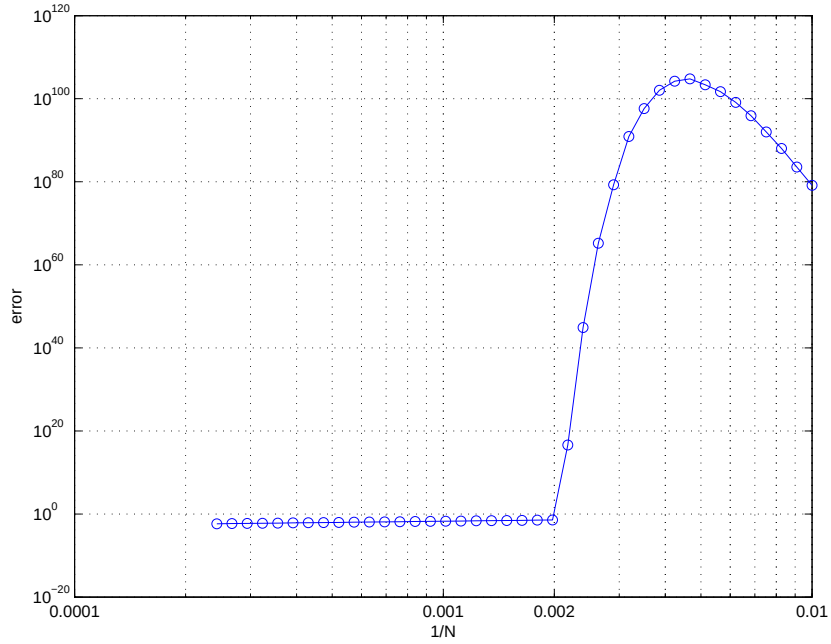

Figura 7.2: Diagrama de eficiencia para el método de Euler aplicado al problema (7.4).

7.3. A-estabilidad de métodos de Runge-Kutta

Para estudiar el dominio de estabilidad lineal $\mathcal{D}$ tenemos que aplicar el método de Runge-Kutta (RK) con paso fijo h al problema escalar

$$y' = \lambda y, \qquad x \geq 0, \qquad y(0) = 1. \tag{7.5}$$

Para ello conviene reescribir (RK) en la forma alternativa

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

Obtenemos que

$$\begin{cases} Y_i = y_n + \lambda h \sum_{j=1}^{s} a_{ij} Y_j, \\ y_{n+1} = y_n + \lambda h \sum_{i=1}^{s} b_i Y_i. \end{cases} \tag{7.6}$$

Sean $Y, e \in \mathbb{R}^s$ dados por

$$Y = (Y_1, \ldots, Y_s)^T, \qquad e = (1, 1, \ldots, 1)^T.$$

Entonces podemos escribir (7.6) en la forma

$$\begin{cases} Y = e\, y_n + \lambda h A Y, \\ y_{n+1} = y_n + \lambda h b^T Y. \end{cases}$$

La primera de estas ecuaciones da

$$Y = (I - \lambda h A)^{-1} e y_n,$$

que introducido en la segunda produce

$$y_{n+1} = (1 + \lambda h b^T (I - \lambda h A)^{-1} e) y_n, \qquad n = 0, 1, \ldots,$$

es decir,

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

donde

$$R(z) = 1 + z b^T (I - zA)^{-1} e, \qquad z \in \mathbb{C}. \tag{7.7}$$

La función $R(z)$ recibe el nombre de *función de estabilidad del método*, o también *función de amplificación.*

Obsérvese que $y_n = (R(\lambda h))^n y_0$, así que de la definición del dominio de estabilidad lineal $\mathcal{D}$ se sigue inmediatamente que

$$\mathcal{D} = \{z \in \mathbb{C} : |R(z)| < 1\}.$$

Por tanto, un método será A-estable si y sólo si $|R(z)| < 1$ en $\mathbb{C}^-$.

Notación. $\mathbb{P}_\alpha$ es el conjunto de todos los polinomios de grado menor o igual que α. $\mathbb{P}_{\alpha/\beta}$ es el conjunto de todas las funciones racionales p/q, con $p \in \mathbb{P}_\alpha$ y $q \in \mathbb{P}_\beta$.

Un sencillo análisis permite demostrar el siguiente lema.

Lema 7.4. *La función de amplificación R de un método de Runge-Kutta de s etapas satisface $R \in \mathbb{P}_{s/s}$. Si el método es explícito entonces $R \in \mathbb{P}_s$.*

Demostración. Tenemos que demostrar que la función R dada por (7.7) pertenece a $\mathbb{P}_{s/s}$. Usamos que

$$(I - zA)^{-1} = \frac{(\text{adj}\,(I - zA))^T}{\det(I - zA)}.$$

Cada elemento de $I - zA$ es lineal en z. Por tanto, como cada elemento de $\text{adj}\,(I - zA)$ es el determinante de una submatriz $(s-1) \times (s-1)$, está en $\mathbb{P}_{s-1}$. Así pues,

$$b^T(\text{adj}\,(I - zA))^T e \in \mathbb{P}_{s-1}.$$

Como $\det(I - zA) \in \mathbb{P}_s$, concluimos que $R \in \mathbb{P}_{s/s}$.

Si el método es explícito, entonces A es estrictamente triangular inferior y $I - zA$ es una matriz triangular inferior con unos a lo largo de la diagonal. Por tanto $\det(I - zA) = 1$, y R es un polinomio. □

Corolario 7.5. *Ningún método de Runge-Kutta explícito puede ser A-estable.*

Demostración. La función de estabilidad de un método de Runge-Kutta explícito es un polinomio. Además $R(0) = 1$. Por otra parte, ningún polinomio, salvo la función constante $R(z) = c$, $|c| < 1$, está estrictamente acotado por uno en $\mathbb{C}^-$. □

¿Qué podemos decir de los métodos implícitos? Veamos con un ejemplo que pueden ser A-estables.

Ejemplo. Sea el método de dos etapas y orden 3 dado por el tablero

$$\begin{array}{c|cc} 0 & \frac{1}{4} & -\frac{1}{4} \\ \frac{2}{3} & \frac{1}{4} & \frac{5}{12} \\ \hline & \frac{1}{4} & \frac{3}{4}. \end{array}$$

Su función de amplificación es

$$R(z) = \frac{1 + \frac{1}{3}z}{1 - \frac{2}{3}z + \frac{1}{6}z^2}.$$

Veamos si es A-estable. Representamos $z \in \mathbb{C}^-$ en coordenadas polares, $z = \rho e^{i\theta}$, con $\rho > 0$, $|\theta - \pi| < \pi/2$, y nos preguntamos si $|R(\rho e^{i\theta})| < 1$. Esto es equivalente a

$$\left|1 + \frac{1}{3}\rho e^{i\theta}\right|^2 < \left|1 - \frac{2}{3}\rho e^{i\theta} + \frac{1}{6}\rho^2 e^{2i\theta}\right|^2,$$

y por tanto a

$$1 + \frac{2}{3}\rho\cos\theta + \frac{1}{9}\rho^2 < 1 - \frac{4}{3}\rho\cos\theta + \rho^2\left(\frac{1}{3}\cos 2\theta + \frac{4}{9}\right) - \frac{2}{9}\rho^3\cos\theta + \frac{1}{36}\rho^4.$$

Reordenando términos, la condición $\rho e^{i\theta} \in \mathcal{D}$ pasa a ser

$$2\rho\left(1 + \frac{1}{9}\rho^2\right)\cos\theta < \frac{1}{3}\rho^2(1 + \cos 2\theta) + \frac{1}{36}\rho^4 = \frac{2}{3}\rho^2\cos^2\theta + \frac{1}{36}\rho^4,$$

que se satisface para todos los $z \in \mathbb{C}^-$, puesto que $\cos\theta < 0$ para todo z de este tipo. El método es por tanto A-estable. ♣

El ejemplo anterior explica por qué son interesantes los métodos implícitos. Son más caros, pero, además de permitir un orden mayor para un número de etapas dado, algunos de ellos tienen regiones de estabilidad grandes, y pueden ser convenientes a la hora de tratar problemas *stiff*.

Sin embargo, no todos los métodos implícitos son A-estables.

Ejemplo. Sea el método de dos etapas y orden 1 definido por el tablero

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

Su función de amplificación es

$$R(z) = \frac{1 + \frac{2}{3}z + \frac{1}{6}z^2}{1 - \frac{1}{3}z}.$$

Así pues, $\lim_{|z|\to\infty} |R(z)| = \infty$, por lo que el método no es A-estable. ♣

En realidad no hace falta comprobar todos los $z \in \mathbb{C}^-$ para ver si una función racional dada tiene su origen en un método A-estable (una $R(z)$ tal se dice A-aceptable). Ése es el contenido del siguiente lema, cuya demostración es una aplicación del principio del máximo.

Lema 7.6. *Sea $R(z)$ una función racional no constante. Entonces $|R(z)| < 1$ para todo $z \in \mathbb{C}^-$ si y sólo si todos los polos de R tienen parte real positiva y $|R(it)| \leq 1$ para todo $t \in \mathbb{R}$.*

Demostración. Si $|R(z)| < 1$ para todo $z \in \mathbb{C}^-$, por continuidad $|R(z)| \leq 1$ para todo $z \in \overline{\mathbb{C}^-}$. En particular R no puede tener polos en el semiplano cerrado izquierdo, y $|R(it)| \leq 1$ para todo $t \in \mathbb{R}$.

Para demostrar el recíproco observamos que si los polos están a la derecha del eje imaginario, entonces la función racional R es analítica en el cerrado $\overline{\mathbb{C}^-}$. Por tanto, al no ser R constante, $|R|$ no puede tener un máximo en el interior, y alcanza su máximo en la frontera. Parte de la frontera está en el infinito. Sin embargo, por ser R una función racional existe el límite (finito o infinito) $\lim_{|z|\to\infty} |R(z)|$, y es igual a $\lim_{t\to\infty} |R(it)|$. Por tanto, si $|R(it)| \leq 1$, $t \in \mathbb{R}$, entonces $|R(z)| < 1$ para todo $z \in \mathbb{C}^-$. □

Ejemplo. Consideramos el método de Gauss-Legendre de dos etapas (orden 4),

$$\begin{array}{c|cc} \frac{1}{2} - \frac{\sqrt{3}}{6} & \frac{1}{4} & \frac{1}{4} - \frac{\sqrt{3}}{6} \\ \frac{1}{2} + \frac{\sqrt{3}}{6} & \frac{1}{4} + \frac{\sqrt{3}}{6} & \frac{1}{4} \\ \hline & \frac{1}{2} & \frac{1}{2}. \end{array}$$

Un sencillo cálculo muestra que

$$R(z) = \frac{1 + \frac{1}{2}z + \frac{1}{12}z^2}{1 - \frac{1}{2}z + \frac{1}{12}z^2}.$$

Los polos de esta función, $3 \pm i\sqrt{3}$, están en el semiplano abierto derecho, y $|R(it)| = 1$, $t \in \mathbb{R}$, por lo que el método es A-estable. ♣

Problemas

1. Demostrar que si un método de Runge-Kutta tiene orden de consistencia p, entonces
$$R(z) = 1 + z + \frac{z^2}{2!} + \cdots + \frac{z^p}{p!} + O(z^{p+1}).$$
Como consecuencia tenemos que todos los métodos de Runge-Kutta explícitos con $p = s$ tienen como función de estabilidad
$$R(z) = 1 + z + \frac{z^2}{2!} + \cdots + \frac{z^s}{s!}.$$
2. Hallar la función de amplificación del método de Euler mejorado.
3. Determinar si es A-estable el método de Runge-Kutta cuya función de amplificación viene dada por:
$$R(z) = \frac{1 + \frac{2z}{3} + \frac{z^2}{6}}{1 - \frac{z}{3}}.$$
4. Considérese el problema
$$y' = \begin{pmatrix} -10 & 9 \\ 10 & -11 \end{pmatrix} y.$$
Calcular aproximadamente (con un error menor que el 1 %) el supremo de los h que pueden emplearse para que con cualquier condición inicial $\lim_{n\to\infty} y_n = 0$, donde y_n es la solución numérica obtenida al aplicar el método de Runge-Kutta clásico de cuatro etapas y orden cuatro.
5. Estudiar si es $A-$estable el método de tablero
$$\begin{array}{c|cc} 0 & 0 & 0 \\ 1 & 1/2 & 1/2 \\ \hline & 1/2 & 1/2 \end{array}$$
y hallar la función de incremento correspondiente, $\phi_f(x_n, y_n; h)$, comprobando que satisface una condición de Lipschitz con respecto a su segunda variable para h pequeño.

6. Consideramos el método de Runge-Kutta de tablero

$$\begin{array}{c|ccc} 0 & & & \\ 1/2 & 1/2 & & \\ 1/3 & 0 & 1/3 & \\ \hline & -1/3 & 1/3 & 1. \end{array}$$

   (i) Comprobar que no es de orden 3.

   (ii) Sabiendo que la función de amplificación es $R(z) = 1 + z + z^2/2 + z^3/6$, demostrar que cuando aplicamos el método al problema lineal escalar $y' = \lambda y$, $y(0) = 1$, se tiene convergencia de orden 3.

7. Supongamos que un método de Runge-Kutta tiene función de amplificación $R(z) = \frac{1+\frac{2z}{3}+z^2}{1-\frac{z}{3}}$. Demostrar que su orden de consistencia no puede ser superior a 1.

8. Comprobar que el método de tablero

$$\begin{array}{c|ccc} 0 & & & \\ 1/2 & 1/2 & & \\ 1/3 & 0 & 1/3 & \\ \hline & -1/3 & 1/3 & 1 \end{array}$$

   no es de orden 3 y sin embargo $R(z) = 1 + z + z^2/2 + z^3/6$.

9. Para cada $\beta \in \mathbb{R}$, consideramos el método Runge-Kutta semi-implícito

$$\begin{aligned} k_1 &= f\big(x_n + \beta\, h, y_n + h\,\beta\, k_1\big), \\ k_2 &= f\big(x_n + (1+\beta)\, h, y_n + h\, k_1 + h\,\beta\, k_2\big), \\ y_{n+1} &= y_n + h\left(\left(\frac{1}{2} + \beta\right) k_1 + \left(\frac{1}{2} - \beta\right) k_2\right). \end{aligned}$$

   (a) Determinar su orden de consistencia, comprobando que es independiente de β.

   (b) Determinar los valores de β para los cuales es A-estable.

10. Demostrar que el método de tablero

$$\begin{array}{c|ccc} 0 & & & \\ 1/2 & 1/2 & & \\ 3/4 & 0 & 3/4 & \\ \hline & 2/9 & 1/3 & 4/9 \end{array}$$

es de orden 3. Comprobar que la región de estabilidad del método tiene intersección no vacía con el eje imaginario $\mathrm{Re}\, z = 0$.

11. Dado el método de tablero

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

estudiar su $A-$estabilidad y tratar de dibujar el dominio correspondiente.

7.4. Estabilidad lineal de métodos lineales multipaso

Queremos aplicar el método lineal multipaso

$$\sum_{j=0}^{k} \alpha_j y_{n+j} = h \sum_{j=0}^{k} \beta_j f_{n+j} \tag{7.8}$$

al problema escalar

$$y' = \lambda y, \qquad \lambda \in \mathbb{C}, \qquad y(0) = 1. \tag{7.9}$$

Observación. Los valores de arranque $y_1, \ldots, y_{k-1}$ no vienen dados por la condición inicial. Al estudiar la estabilidad lineal pediremos que $\lim_{n\to\infty} y_n = 0$ para cualquier elección posible de los valores de arranque. ♠

El resultado de aplicar (7.8) a (7.9) es

$$\sum_{j=0}^{k} (\alpha_j - h\lambda\beta_j) y_{n+j} = 0. \tag{7.10}$$

Ésta es una ecuación en diferencias lineal, homogénea y con coeficientes constantes. Para resolverla consideramos el polinomio característico de la ecuación

en diferencias. Haciendo $z = h\lambda$, viene dado por

$$\Pi(r, z) = \sum_{j=0}^{k} (\alpha_j - z\beta_j) r^j, \qquad z \in \mathbb{C}.$$

En términos del primer y segundo polinomios característicos del método se escribe

$$\Pi(r, z) = \rho(r) - z\sigma(r),$$

y recibe el nombre de *polinomio de estabilidad* del método lineal multipaso.

Lema 7.7. *Supongamos que los ceros (como función de r) de $\Pi(r, z)$ son $r_1(z)$, $r_2(z)$, $\ldots, r_{q(z)}(z)$, con multiplicidades $m_1(z), m_2(z), \ldots, m_{q(z)}(z)$ respectivamente ($\sum_{i=1}^{q(z)} m_i(z) = k$). Entonces*

$$\mathcal{D} = \{z \in \mathbb{C} : \ |r_i(z)| < 1, \ i = 1, \ldots, q(z)\}.$$

Demostración. La solución general de (7.10) es

$$y_n = \sum_{i=1}^{q(z)} \left(\sum_{j=0}^{m_i(z)-1} c_{ij} n^j \right) (r_i(z))^n, \qquad n = 0, 1, \ldots.$$

Las k constantes c_{ij} quedan determinadas de manera única por los k valores de arranque $y_0, \ldots, y_{k-1}$, y no dependen de n. Así, el comportamiento de y_n está determinado por la magnitud de los números $r_i(z)$, $i = 1, \ldots, q(z)$. Si todos están en el interior del disco unidad del plano complejo, entonces sus potencias decaen más aprisa que cualquier polinomio en n, y $\lim_{n\to\infty} y_n = 0$.

Por otro lado, sea $|r_i(z)| \geq 1$; existen valores de arranque tales que $c_{i0} \neq 0$, y para dichos valores es imposible que $\lim_{n\to\infty} y_n = 0$. □

Ejemplo. El método de Euler implícito es un método lineal multipaso con $\rho(r) = r - 1$ y $\sigma(r) = r$. El polinomio de estabilidad lineal $\Pi(r, z) = r - 1 - zr$ tiene un único cero, $r = 1/(1 - z)$, que tiene módulo menor que 1 si y sólo si $|1 - z| > 1$. Así pues, $\mathcal{D} = \{z \in \mathbb{C} : \ |z - 1| > 1\}$, y el método es A-estable. ♣

La determinación del dominio de estabilidad lineal de un método lineal multipaso mediante la aplicación directa del lema 7.7 presenta dos dificultades: (i) hay que hallar las raíces del polinomio característico; (ii) hay que determinar los valores de z para los que dichas raíces tienen módulo menor que 1. Así que intentaremos aplicar otra técnica.

Las raíces de un polinomio son funciones continuas de sus coeficientes. Por consiguiente, $\mathrm{Fr}\,(\mathcal{D}) \subset \mathcal{F}$, donde

$$\mathcal{F} := \{z \in \mathbb{C} : \exists \text{ una raíz de } \Pi(r, z) \text{ de módulo } 1\}.$$

En general se tiene la inclusión y no la igualdad; puede haber valores $z \in \mathbb{C}$ tales que el polinomio $\Pi(r, z)$ tenga una raíz de módulo 1, alguna raíz de módulo menor que 1 y alguna raíz de módulo mayor que 1. Dicho z no pertenece a $\mathrm{Fr}\,(\mathcal{D})$. Nótese por otra parte que $\mathcal{D} \cap \mathcal{F} = \emptyset$.

Si $z \in \mathcal{F}$, existe $\theta \in [0, 2\pi]$ tal que

$$\Pi(e^{i\theta}, z) = \rho(e^{i\theta}) - z\sigma(e^{i\theta}) = 0;$$

es decir,

$$z = \frac{\rho(e^{i\theta})}{\sigma(e^{i\theta})}, \qquad 0 \le \theta \le 2\pi.$$

Tenemos por tanto que $\mathrm{Fr}\,(\mathcal{D})$ está contenida en un conjunto, $\mathcal{F}$, que es la imagen de una curva cerrada que divide al plano complejo en regiones. Cada una de estas regiones tiene que estar, bien contenida en $\mathcal{D}$, bien contenida en $\mathbb{C} \setminus \mathcal{D}$; para ver cuál de los dos casos se da, bastará por tanto con estudiar un único punto.

Ejemplo. La regla de Simpson es un método lineal de dos pasos con $\rho(r) = r^2 - 1$, $\sigma(r) = \frac{1}{3}(r^2 + 4r + 1)$. Por lo tanto,

$$\mathcal{F} = \{z \in \mathbb{C} : \, z = \frac{3\,\mathrm{sen}\,\theta}{\cos\theta + 2}\, i, \ \theta \in [0, 2\pi]\}.$$

Puesto que la función $3\,\mathrm{sen}\,\theta/(\cos\theta + 2)$ varía entre $-\sqrt{3}$ y $\sqrt{3}$, tenemos que $\mathcal{F}$ es el segmento $[-\sqrt{3}i, \sqrt{3}i]$. Hay, por tanto, dos posibilidades: $\mathcal{D} = \emptyset$ o

$\mathcal{D} = \mathbb{C} \setminus \mathcal{F}$. Las raíces de $\Pi(r, -3) = 2r^2 + 4r$ son $r_1(-3) = 0$ y $r_2(-3) = -2$. Por consiguiente $z = -3 \notin \mathcal{D}$, de donde concluimos que $\mathcal{D} = \emptyset$. ♣

Observación. La regla de Simpson tiene orden $p = k + 2$, el mayor obtenible para un método 0-estable (primera barrera de Dahlquist). Sin embargo no es útil, pues $\mathcal{D} = \emptyset$. Los métodos 0-estables con orden de convergencia $p = k + 2$, como la regla de Simpson, se llaman *optimales*, En general todos ellos tienen regiones de estabilidad que son o bien vacías, o bien esencialmente inútiles por no contener el eje real negativo en un entorno del origen. ♠

¿Hay métodos lineales multipaso A-estables? Explícitos, no, pues se tiene el siguiente resultado.

Teorema 7.8. *Todo método lineal multipaso explícito convergente tiene región de estabilidad absoluta acotada.*

Demostración. Como el método es explícito, $\beta_k = 0$, y se tiene que

$$\Pi(r, z) = r^k + \sum_{j=0}^{k-1} (\alpha_j - z\beta_j) r^j.$$

Por otra parte,

$$\Pi(r, z) = (r - r_1(z)) \dots (r - r_k(z)) = r^k + \sum_{j=0}^{k-1} s_j(z) r^j,$$

donde

$$s_j(z) = (-1)^{k-j} \sum_{i_1 < i_2 < \cdots < i_{k-j}} r_{i_1}(z) \cdots r_{i_{k-j}}(z).$$

Comparando ambas expresiones tenemos que $s_j(z) = \alpha_j - z\beta_j$. Como el método converge (con orden al menos 1), $0 \neq \rho'(1) = \sum_{j=0}^{k} j\alpha_j = \sum_{j=0}^{k} \beta_j$ y por tanto existe algún $j \in \{0, \dots, k-1\}$ tal que $\beta_j \neq 0$. Así, $s_j(z) \to \infty$ cuando $z \to \infty$, de modo que existe una raíz $r_i(z)$ tal que $\lim_{z \to \infty} r_i(z) = \infty$. □

¿Tenemos mejor suerte con los implícitos? Sí: hay ejemplos de métodos lineales multipaso implícitos A-estables de órdenes 1 (el método de Euler implícito) y 2 (la regla del trapecio). Sin embargo, no se puede conseguir nada mejor.

Teorema 7.9 (Segunda barrera de Dahlquist, 1957)**.** *El mayor orden de consistencia de un método lineal multipaso A-estable es $p = 2$.*

Los métodos de Runge-Kutta implícitos A-estables no tienen esta restricción sobre el orden. Recordemos por ejemplo que el método de Gauss-Legendre de dos etapas tiene orden de consistencia 4 y sin embargo es A-estable. Así que se podría pensar que los métodos lineales multipaso son inferiores a los de Runge-Kutta en lo concerniente a la estabilidad lineal. Esto no es completamente cierto. Hay métodos lineales multipaso que, a pesar de no ser A-estables, son $A(\alpha)$-*estables*: existe un $\alpha \in (0, \pi]$ tal que la cuña infinita

$$\Gamma_\alpha := \{\rho e^{i\theta} : \rho > 0, |\theta - \pi| < \alpha\}$$

está contenida en $\mathcal{D}$. En otras palabras, si todos los autovalores están en Γ_α, aunque estén muy lejos del origen el método no obliga a disminuir el paso por motivos de estabilidad.

Problemas

1. Consideramos el método lineal multipaso $y_{n+2} - y_n = 2hf_{n+1}$ (leap-frog). Demostrar que su dominio de estabilidad lineal es vacío.
2. El objetivo de este problema es demostrar que la región de estabilidad absoluta de un método lineal multipaso convergente no puede contener el eje real positivo en un entorno del origen.

   (i) Demostrar que hay una única raíz, $r_1(z)$, de $\Pi(r, z)$ con la propiedad de que $r_1(z) \to 1$ cuando $z \to 0$.

   (ii) Consideramos el problema (7.9) con $\lambda \in \mathbb{R}^+$. Demostrar que el residuo satisface $R_n = O(h^2)$ cuando $h \to 0^+$ y deducir de ello que $\Pi(\exp(z), z) = O(z^2)$ cuando $z \to 0$, $z \in \mathbb{R}^+$.

(iii) Usar que $\Pi(r,z) = (1-z\beta_k)(r-r_1)(r-r_2)\cdots(r-r_k)$ para concluir que $r_1(z) = e^z + O(z^2) = 1 + z + O(z^2)$ cuando $z \to 0$, $z \in \mathbb{R}^+$.

3. Determinar si es A-estable la fórmula BDF de dos pasos,

$$y_{n+2} - \frac{4}{3}y_{n+1} + \frac{1}{3}y_n = \frac{2}{3}hf_{n+2}.$$

Indicación. Demostrar que $\mathrm{Re}\left(\frac{\rho(e^{i\theta})}{\sigma(e^{i\theta})}\right) \geq 0$.

4. Decidir razonadamente si las siguientes afirmaciones son verdaderas o falsas.

(i) El método lineal multipaso cuyo primer polinomio característico es $\rho(\zeta) = (\zeta^2+1)(\zeta-1)$ y cuyo segundo polinomio característico es $\sigma(\zeta) = (5\zeta^3 + 7\zeta^2 + 7\zeta + 5)/12$ tiene orden de consistencia 5.

(ii) El método lineal multipaso $y_{n+3} - \frac{18}{11}y_{n+2} + \frac{9}{11}y_{n+1} - \frac{2}{11}y_n = \frac{6}{11}hf_{n+3}$ tiene orden de consistencia 3 y es A-estable.

(iii) El método lineal multipaso $y_{n+2} - y_n = \frac{h}{3}(f_n + 4f_{n+1} + f_{n+2})$ es convergente de orden 3, pero no de orden 4.

5. Demostrar que la región de estabilidad absoluta $\mathcal{D}$ del método

$$y_{n+2} - y_n = \frac{h}{2}(f_{n+1} + 3f_n)$$

es el interior del círculo de centro $(-2/3, 0)$ y de radio $2/3$.

6. Demostrar que si un MLM de polinomios característicos ρ y σ satisface

$$\mathrm{Re}\,\left(\rho(e^{i\theta})\sigma(e^{-i\theta})\right) = 0, \qquad \forall\theta \in [0, 2\pi], \tag{7.11}$$

entonces o bien $\mathcal{D}$ es vacía o bien $\mathcal{D} = \{z \in \mathbb{C} : \mathrm{Re}\,(z) < 0\}$. Demostrar que el método lineal de dos pasos 0-estable y de orden al menos 2 más general que satisface (7.11) es de la forma

$$y_{n+2} - y_n = h(\beta f_{n+2} + 2(1-\beta)f_{n+1} + \beta f_n),$$

y que es A-estable si y sólo si $\beta > 1/2$.

7. Sea el método

$$y_{n+2} - (1+\alpha)y_{n+1} + \alpha y_n = \frac{h}{2}(1-\alpha)(f_{n+1} + f_n), \qquad -1 < \alpha < 1.$$

(i) Demostrar que el orden es independiente de α.

(ii) Expresar la curva $z = \rho(e^{i\theta})/\sigma(e^{i\theta})$ de la forma $y^2 = F(x)$ si $z = x + iy$.

(iii) Determinar la región de estabilidad absoluta.

(iv) Deducir que el intervalo de estabilidad absoluta es independiente de α.

7.5. Experimentos numéricos

Consideramos los problemas lineales

$$y'(x) = Ay(x) + B(x) \qquad \text{para } 0 \le x \le 10, \qquad y(0) = (2,3)^T,$$

con

$$\text{(Problema 1)} \qquad A = \begin{pmatrix} -2 & 1 \\ 1 & -2 \end{pmatrix}, \qquad B(x) = \begin{pmatrix} 2\,\mathrm{sen}\,x \\ 2(\cos x - \mathrm{sen}\,x) \end{pmatrix},$$

$$\text{(Problema 2)} \qquad A = \begin{pmatrix} -2 & 1 \\ 998 & -999 \end{pmatrix}, \qquad B(x) = \begin{pmatrix} 2\,\mathrm{sen}\,x \\ 999(\cos x - \mathrm{sen}\,x) \end{pmatrix}.$$

En ambos casos la solución es

$$y = 2e^{-x}\begin{pmatrix} 1 \\ 1 \end{pmatrix} + \begin{pmatrix} \mathrm{sen}\,x \\ \cos x \end{pmatrix}.$$

En la figura 7.3 mostramos el diagrama de eficiencia correspondiente a aplicar al Problema 1 algunos de los métodos introducidos en los capítulos anteriores.

En el rango de costes considerado, el método de Euler es computacionalmente más caro que los demás métodos estudiados. Entre los métodos de orden 2, el más costoso, para un nivel de error dado, es la regla del trapecio.

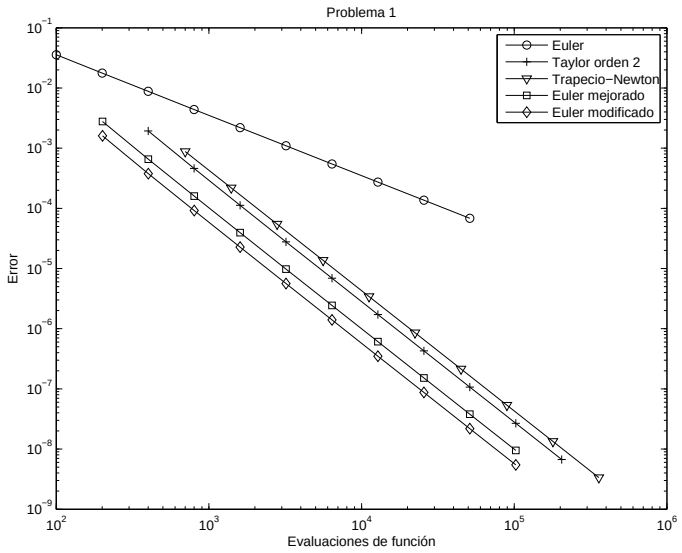

Figura 7.3: Diagrama de eficiencia para el problema 1.

Esto se debe a que en cada paso hay que aplicar el método de Newton, con el coste computacional que ello supone. Aunque en este problema concreto no es muy grande (es un problema lineal, y con una iteración Newton nos va a dar la solución exacta), es suficiente para que no pueda competir con los demás métodos. Entre los métodos de orden 2 considerados, el más caro es el de Taylor, por culpa de todas las derivadas que hay que evaluar. Euler mejorado y Euler modificado (con una ligera ventaja para este último) son los más eficientes.

Veamos ahora qué pasa con el Problema 2. Los resultados, en forma de diagramas de eficiencia, se muestran en la figura 7.4. Para todos los métodos considerados, salvo para la regla del trapecio-Newton, cuando N es "pequeño" los métodos funcionan muy mal, y no los representamos. ¿Cuál es la diferencia entre los problemas 1 y 2? Está claro que no es la regularidad de la solución, pues los dos problemas tienen la misma. La diferencia reside en que el Problema 2 es *stiff*. Para que los métodos funcionen bien tenemos que tomar un h que garantice que nos metemos en el dominio de estabilidad lineal del método. Salvo

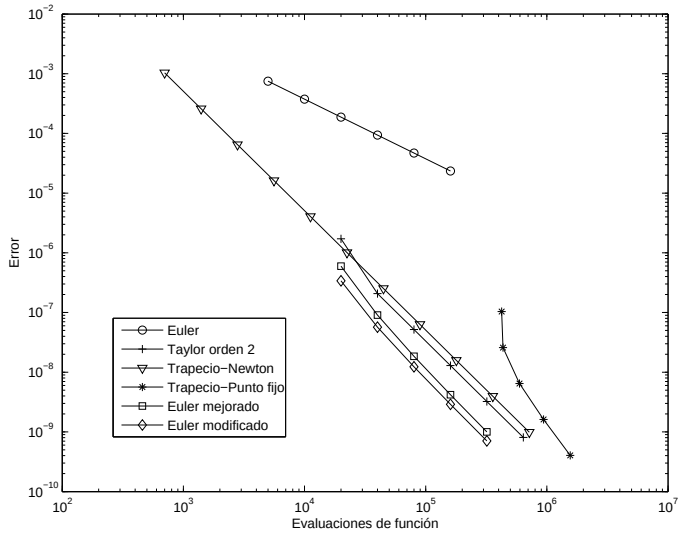

Figura 7.4: Diagrama de eficiencia para el Problema 2.

en el caso de la regla del trapecio, esto introducirá una restricción significativa. El problema con la regla del trapecio con iteración de punto fijo es de una naturaleza distinta: para N pequeño la iteración no converge. La razón es que el problema tiene una constante de Lipschitz relativamente grande, y la iteración de punto fijo no converge a menos que $Lh/2 < 1$.

Problemas

1. Se considera el problema escalar

$$y'(x) = -5xy^2(x) + \frac{5}{x} - \frac{1}{x^2}, \quad x \in [1, 25], \qquad y(1) = 1,$$

   cuya solución exacta es $y(x) = 1/x$. Resolverlo numéricamente con el método de Euler con paso fijo $h = 0{,}19$ y $h = 0{,}21$. Hacer un dibujo en el que aparezcan superpuestas ambas soluciones numéricas. Explicar lo observado.

2. Se considera el problema escalar

$$y'(x) = \lambda(y(x) - \operatorname{sen} x) + \cos x, \quad x \in [0, 1], \qquad y(0) = 1,$$

cuya solución exacta es $y(x) = e^{\lambda x} + \operatorname{sen} x$. Resolverlo numéricamente con el método de Adams-Basforth de dos pasos (el segundo valor de arranque se calculará con el método de Euler) y con la fórmula BDF de dos pasos (el segundo valor de arranque se calculará con el método de Euler implícito) con $h = 0,01$ para $\lambda = 10$, $\lambda = -10$ y $\lambda = -500$. Dibujar las soluciones obtenidas y explicar lo observado.