Cálculo Numérico II

Problemas resueltos

(y resúmenes de teoría)

Fernando Chamizo Lorente

Estas notas constan de tres partes: la primera es una colección de problemas, la segunda contiene sus soluciones, y la tercera, que podría entenderse como un apéndice, es un resumen de los temas de teoría. Todas ellas están asociadas al curso de cálculo numérico impartido en 4º de Matemáticas de la UAM el primer semestre del año académico 2001/2002. En dicho curso se siguieron las notas de F. Quirós tituladas "Cálculo Numérico II", de las cuales se toman, salvo pequeños cambios, la división en secciones y capítulos.

Los enunciados de los problemas son en su mayoría originales, con ciertas excepciones. Varios han sido tomados de hojas de problemas del curso anterior; algunos del libro de E. Hairer, S.P. Nørsett y G. Wanner "Solving ordinary differential equations 1. Nonstiff problems", casi siempre con modificaciones; y uno del texto de C. Moreno González "Cálculo Numérico II". Cuando sea posible, se citará la fuente en la solución. De todas formas la originalidad de los problemas debe entenderse en sentido laxo, ya que muchos están basados en resultados bien conocidos. Los ejercicios están señalados con tres números separados por puntos. El primero y el segundo indican el capítulo y la sección a la que pertenecen, y el tercero su orden dentro de la lista. El símbolo Ⓛ precediendo algunos ejercicios indica que son de naturaleza práctica y la mayor parte de las veces requieren emplear un ordenador. La dificultad de algunos problemas o apartados está indicada con cierto número y tamaño de asteriscos. Un problema de dos asteriscos puede ser dificilísimo incluso para los mejores alumnos de la clase.

La segunda parte incluye las soluciones completas de todos los problemas excepto de los prácticos. No se han revisado de forma tan exhaustiva como los enunciados, por lo cual no es de extrañar que contengan cierto número de erratas. Incluso podría haberse deslizado en algún caso un error en el planteamiento. Cualquier corrección será bienvenida e incorporada en las posibles versiones futuras.

Los resúmenes de teoría, como se puede leer en sus primeras líneas, fueron concebidos como ayuda al estudio para el curso durante el que se elaboraron. Dependen fuertemente de la versión de las notas de F. Quirós que se distribuyó entre los alumnos. Obviamente las consideraciones acerca de los requisitos mínimos o de las fórmulas que conviene memorizar, no pueden ser sacadas del contexto del curso particular al que se refieren. Teniendo en cuen-

ta que el temario de esta asignatura sufre continuos cambios, aunque parece acercarse por fin a ciertas líneas básicas, la ayuda al estudio que puedan representar estos resúmenes en el futuro, es muy relativa.

Para terminar, hay algunas cuestiones de tipo organizativo que podrían resultar de interés para los que impartan más adelante esta asignatura. Cada semana hubo cuatro horas de teoría y cuatro de prácticas, las últimas divididas en dos turnos, de modo que cada alumno generalmente sólo asistía a dos de ellas. La parte teórica se evaluó con un examen final, mientras que para la parte prácticas se especificaron dos trabajos a lo largo del curso. La calificación final se compuso en un 60% de la parte teórica y en un 40% de la práctica. Además se exigió que fuera aprobada cada parte por separado.

El curso Cálculo Numérico II del año 2001/2002 contuvo los capítulos y secciones que se indican a continuación. Las señaladas con ⊘ no están cubiertas por los problemas. La primera porque es de carácter práctico, y las otras porque el último capítulo fue minimizado y se consideraron suficientes los ejemplos de clase.

1. Métodos de un paso

1. Introducción
2. Método de Euler
3. Un método implícito: la regla del trapecio
4. Métodos de Taylor
5. Métodos de un paso
6. Métodos de Runge-Kutta
7. Condiciones de orden
8. Árboles de Butcher
9. Demostración de los teoremas de Butcher
10. Métodos explícitos. Orden obtenible
11. Estimaciones de error
12. Extrapolación de Richardson
13. Cambio de paso
14. Pares encajados
⊘15. Diagramas de eficiencia

2. Problemas stiff

1. ¿Qué es un problema stiff?
2. Dominio de estabilidad lineal y A-estabilidad
3. Estabilidad de métodos de Runge-Kutta

3. Diferencias finitas

1. La fórmula de cinco puntos

4. Elementos finitos

1. Introducción al método de elementos finitos
2. Formulación abstracta débil y variacional
3. Programación y espacios de elementos finitos en dos variables
4. Breve introducción a la convergencia

5. Métodos lineales multipaso

⊘1. Introducción
⊘2. Consistencia
⊘3. 0-estabilidad

Problemas

1. Métodos de un paso

1.1.1. Comprobar que $y' + y^2 = 0$, $y(0) = 2$ no admite una solución continua en $(-1, 1)$, a pesar de que $f(t) = -t^2$ es Lipschitz en cada intervalo acotado. Explicar por qué esto no contradice el teorema de existencia.

1.1.2. a) Comprobar que $f(x) = \sqrt{|x|}$ no es Lipschitz en ningún intervalo que contenga al cero.

b) Hallar tres soluciones de $y' = |y|^{1/2}$, $y(0) = 0$, en $I = (-1, 1)$.

*c) Probar que el problema anterior admite una única solución C^2.

1.1.3. 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 $Y = (y, y', \ldots, y^{n-1)})$ y A una matriz $n \times n$.

1.1.4. Sea $y' = Ay$ con $y \in \mathbb{R}^n$ y $A \in \mathcal{M}_{n\times n}(\mathbb{R})$. Probar que la constante de Lipschitz correspondiente es $L = \max \sqrt{\lambda_j}$ donde los λ_j son los autovalores de $A^t \cdot A$. (Recuérdese que toda matriz simétrica diagonaliza en una base ortonormal).

1.2.1. Aplicar el método de Euler con $h = 1/4$, $y_0 = 1$, para aproximar $y(1)$ donde y es la solución de $y' = -y^2$, $y(0) = 1$. Comparar el error real con el previsto en el teorema correspondiente.

1.2.2. En la sección anterior vimos que $y' = |y|^{1/2}$, $y(0) = 0$ tiene muchas soluciones (de hecho infinitas), ¿a cuál de ellas converge el método de Euler? Suponiendo que a causa de ciertos errores de redondeo en realidad se aplica el método de Euler con $y_0 = 10^{-10}$, ¿a qué solución converge?

1.2.3. a) Hallar una fórmula para la y_n cuando se aplica el método de Euler a $y' = \alpha y + \beta$, $y(0) = 0$, con $y_0 = 0$.

b) Calcular la verdadera solución y comprobar que $|y(b) - y_N| \to 0$ cuando $h \to 0$ con $x_N = b > 0$.

*c) En las condiciones del apartado anterior demostrar que si $\alpha \neq 0$ se cumple $|y(b) - y_N| = O(h)$ pero $|y(b) - y_N| = O(h^p)$ no se cumple para ningún $p > 1$.

1.2.4. Repetir el problema anterior para el sistema

$$\begin{aligned}(y^1)' &= 2y^1 + 2y^2\\ (y^2)' &= \qquad 2y^2\end{aligned}$$

con $y(0) = (1,1)^T$.

1.2.5. Probar que el método de Euler con $y_0 = 0$ aplicado a $y' = 1 + x^2$, $y(0) = 0$ no es de orden 2. *Indicación*: $1^2 + 2^2 + 3^2 + \cdots + k^2 = (2k+1)(1 + 2 + 3 + \ldots k)/3$.

1.2.6. Probar que el método de Euler con $y_0 = 0$ aplicado en $[0,1]$ a $y' = x(1-x)$, $y(0) = 0$ cumple $|y(1) - y_N| \leq h^2$ y sin embargo no es de orden dos.

1.2.7. Un conductor por una carretera comarcal ve una señal que indica "Reduzca a 50", un rato después otra que indica "Reduzca a 40", más adelante se encuentra, ya a paso de tortuga, con otras análogas de 30, 20 y 10. Finalmente, una última señal indica "Bienvenidos a Reduzca".

a) Explicar qué relación guarda el chiste (¿?) anterior con el método de Euler aplicado a $(50 - x)y' = 1$ (interpretar x como la distancia e y como el tiempo) y cuánto tardará el conductor en llegar a Reduzca si hace caso a todas las señales.

b) Si hubiera infinitas señales, espaciadas infinitesimalmente, ¿llegaría alguna vez a Reduzca?

1.2.8. Dado el problema $y' = y$, $y(0) = 1$, sea y_n el resultado de la n-ésima iteración al aplicar el método de Euler con paso h (e $y_0 = 1$) y sea u_n lo mismo para paso $h/2$. Demostrar que $y_n < u_{2n} < y(x_n)$.

*__1.2.9__ Probar que el método de Euler aplicado a $y'\sqrt{1-x} = 1$, $y(0) = 0$ tiene orden estrictamente $1/2$ cuando se aplica en $[0,1]$, es decir, que $|y(1) - y_N| = O(h^{1/2})$ pero $|y(1) - y_N| \neq O(h^p)$ para $p > 1/2$. Explicar cómo es esto posible si habíamos probado que el método de Euler es de orden uno. *Indicación*: Utilizar el teorema del valor medio para probar $\sqrt{n+1} - \sqrt{n} < 1/\sqrt{4n} < \sqrt{n} - \sqrt{n-1}$.

Ⓛ **1.2.10.** Considérese el problema con $d = 2$

$$y' = \begin{pmatrix} -68 & -4 \\ 67 & 3 \end{pmatrix} y, \qquad y(0) = \begin{pmatrix} 0 \\ -63 \end{pmatrix}.$$

Aproximar $y(1)$ mediante el método de Euler con $h = 1/30$, $y_0 = y(0)$ y comparar el resultado con la solución real $y(1) = (1'4715\ldots, -24'6479\ldots)^T$. Tratar de explicar la gran disparidad en el resultado. (Más adelante diremos que estos sistema en los que el método de Euler nos puede dar sorpresas son *stiff*).

Ⓛ **1.2.11.** Aproximar $y(2)$ donde y es la solución de $y' = |y|^{1/2}$, $y(0) = 0$, utilizando el método de Euler con $y_0 = 0$ e $y_0 = 10^{-3}$ y h pequeño en cada caso. Explicar la diferencia en los resultados.

1.3.1. Hallar una fórmula para y_n cuando se aplica la regla del trapecio a ecuaciones de la forma $y' = ay + b$ con $y(0) = 0$.

1.3.2. Aplicar la regla del trapecio en $[0, 1]$ a $y' = 2y$, $y(0) = 1$ probando que $\sqrt[N]{y_N} = e^{2/N} + O(N^{-3})$. Deducir de ahí que $|y(1) - y_N| = O(h^2)$. *Indicación*: La desigualdad $1 + t \leq e^t$ implica $(1 + O(N^{-3}))^N = 1 + O(N^{-2})$.

1.3.3. Recuérdese que la regla del trapecio en Cálculo Numérico I era una formula de cuadratura numérica que aproximaba $\int_a^b f(t)\ dt$ (digamos con $f > 0$ para fijar ideas) por el área de la colección de trapecios de vértices $(x_i, 0)$, $(x_i, f(x_i))$, $(x_{i+1}, 0)$, $(x_{i+1}, f(x_i))$. Explicar por qué recibe este nombre también el método implícito de esta sección.

1.3.4. Siguiendo la idea del ejercicio anterior, explicar por qué la regla de Simpson no da lugar a un método unipaso, es decir, en el que y_{n+1} sólo dependa de y_n, de x_n y de h.

*__1.3.5__ Demostrar que cuando la regla del trapecio se aplica a problemas escalares del tipo $y' = f(x)$ en $[0, 1]$ se cumple que $|y(x_n) - y_n| < 0'25\, h^2 \sup |f''|$.

Ⓛ **1.3.6.** Aplicar la regla del trapecio con $h = 0'2$ para aproximar $y(1)$ donde y está determinada por $y' = x + y^2 x$, $y(0) = 1$. Comparar el resultado con la solución real.

Ⓛ **1.3.7.** Combinar el método de Newton para resolver ecuaciones (esto es, $x_{n+1} = x_n - f(x_n)/f'(x_n)$) con la regla del trapecio en un programa FORTRAN que dado un $y_0 \in (-1, 1)$ aproxime, de esta manera, la solución de $y' = \cos y$, $y(0) = y_0$.

1.4.1. Comprobar que el método de Taylor de orden 3 aplicado al problema $y' = 2xy$, $y(0) = 1$, lleva a la iteración

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

1.4.2. Construir el método de Taylor general de orden 3 para $d = 1$.

1.4.3. Construir el método de Taylor general de orden 2 para $d = 2$.

1.4.4. a) Escribir la fórmula iterativa correspondiente al método de Taylor de orden k para la ecuación (vectorial) $y' = Ay$ con A una matriz constante $n \times n$.

*b) Comprobar que formalmente cuando $k = \infty$ basta tomar $N = 1$ (una iteración) para obtener la solución exacta.

1.4.5. Dada la función $p : \mathbb{R} \longrightarrow (-1, 1)$

$$p(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2}\, dt,$$

se define $y = y(x)$ como la función inversa de p (tal función es muy útil en Estadística).

a) Demostrar que tal función inversa existe y satisface la ecuación diferencial

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

b) Escribir la iteración correspondiente al método de Taylor de orden tres.

Ⓛ **1.4.6.** Escribir un programa FORTRAN que dado un $\alpha \in (-\pi/2, \pi/2)$ aproxime la solución de la ecuación del péndulo

$$y'' + \operatorname{sen} y = 0, \qquad y'(0) = 0, \qquad y(0) = \alpha,$$

empleando el método de Euler y el de Taylor de orden dos (para ello hay que comenzar escribiendo la ecuación como una de primer orden). En Mecánica es conocido que el primer valor positivo de x con $y'(x) = 0$ (el *semiperiodo*) viene dado con gran aproximación por la fórmula

$$\pi(1 + \frac{1}{4}\operatorname{sen}^2 \frac{\alpha}{2} + \frac{9}{64}\operatorname{sen}^4 \frac{\alpha}{2}).$$

Utilizar este resultado para comprobar la precisión de ambos métodos según varía la longitud del paso.

1.5.1. Hallar las constantes α y β de forma que

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

sea consistente de orden uno o consistente de orden dos.

1.5.2. a) Hallar el orden de consistencia del método

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

b) Comprobar (suponiendo, como simpre, que f es Lipschitz en la segunda variable) que el método satisface las condiciones del teorema de convergencia.

1.5.3. El método de Euler modificado es un método de un paso con

$$\phi(x, y, z; h) = f(x + \frac{h}{2}, y + \frac{h}{2}f(x, y)).$$

a) Demostrar que tiene orden de convergencia dos.

b) Demostrar, buscando algún contraejemplo, que el orden no es mayor que dos.

1.5.4. El propósito de este problema es mostrar que los métodos de un paso no son los únicos posibles. Históricamente uno de los primeros fue el *método de Picard*, en el que no se discretiza sino que se considera una sucesión de funciones, $\{y_n(x)\}_{n=1}^{\infty}$, que tratan de aproximar a la verdadera solución y están definidas por

$$y_{n+1}(x) = y(a) + \int_a^x f(t, y_n(t))\, dt$$

a) Aplicar el método de Picard a $y' = y$, $y(0) = 1$ con y_1 la función que vale constantemente uno.

*b) Aplicar el teorema de la aplicación contractiva (o del punto fijo) al método de Picard para probar un teorema de existencia y unicidad para el problema $y' = f(x, y)$, $y(a) = \eta$.

Ⓛ **1.5.5.** Elaborar un programa FORTRAN que aplique el método de Euler modificado a la ecuación $y' = -x^2 y$, $y(0) = 1$, y que muestre el tamaño real del error al aproximar $y(2)$ en relación con la cota para el error obtenida con el teorema correspondiente.

Ⓛ **1.5.6.** Elegir un método de orden dos y escribir un programa FORTRAN que lo aplique para estimar $y(3)$ en cada uno de los problemas

$$y' = \begin{cases} y^2 & \text{si } y \le 2 \\ 1 & \text{si } y > 2 \end{cases} \qquad y' = \begin{cases} y^2 & \text{si } y \le 2 \\ 4 & \text{si } y > 2 \end{cases} \qquad y' = \begin{cases} y^2 & \text{si } y \le 2 \\ -4 + 4y & \text{si } y > 2 \end{cases}$$

con $y_0 = y(0) = 1$ en todos los casos. Comparar los resulados obtenidos con las soluciones exactas $4'5$, 12 y $e^{10} + 1$, según se modifica el paso, tratando de deducir empíricamente en cada caso cuál parece ser el orden de convergencia del algoritmo.

1.6.1. Explicar por qué el método de Euler modificado es en general mejor que la regla del trapecio. ¿Es la regla del trapecio un método de Runge-Kutta?

1.6.2. Escribir el tablero que corresponde al siguiente método creado por Heun

$$y_{n+1} = y_n + \frac{h}{4}\left[f(x_n, y_n) + 3f\left(x_n + \frac{2h}{3}, y_n + \frac{2h}{3} f(x_n + \frac{h}{3}, y_n + \frac{h}{3} f(x_n, y_n))\right)\right]$$

1.6.3. Aplicar el método del problema anterior a $y' = y$, $y(0) = y_0 = 1$ hallando y_n y comprobar directamente que para este problema es de orden 3. *Indicación*: Emplear que $1 + h + h^2/2 + h^3/6 = e^h + O(h^4)$ y que $(1 + x)^n = 1 + O(nx)$ cuando $n|x| \le 1$.

1.6.4. En los métodos de Runge-Kutta implícitos se debe resolver el sistema, en general no lineal,

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

con $f : \mathbb{R} \times \mathbb{R}^d \longrightarrow \mathbb{R}^d$. En este problema probaremos que dados x_n, y_n, el sistema tiene solución única cuando h es suficientemente pequeño.

a) Sea $F : \mathbb{R}^{sd} \longrightarrow \mathbb{R}^{sd}$ dados por $F = (F^1, \ldots, F^s)$ con

$$F^i(k_1, \ldots, k_s) = f(x_n + c_i h, y_n + h\sum_{j=1}^{s} a_{ij}k_j).$$

Probar que F satisface la condición de Lipschitz con constante $s^2 hL \max |a_{ij}|$ (donde L es la constante Lipschitz de f).

b) Deducir del apartado anterior y del teorema de la aplicación contractiva que el sistema del enunciado tiene solución única cuando h es pequeño.

1.6.5. A los métodos implícitos con $a_{ij} = 0$ si $j > i$ se les llama *semiexplícitos*. Explicar qué ventaja tiene con respecto a los implícitos generales.

Ⓛ **1.6.6.** La solución de

$$\begin{pmatrix} y^1 \\ y^2 \end{pmatrix}' = \begin{pmatrix} 10 & 13 \\ -8 & -10 \end{pmatrix} \begin{pmatrix} y^1 \\ y^2 \end{pmatrix}, \qquad y(0) = \begin{pmatrix} 3 \\ 7 \end{pmatrix}$$

es periódica de periodo π, en particular $\|y(\pi) - y(0)\| = 0$, $\|y(4\pi) - y(0)\| = 0$. Comprobar en cuánto difieren estas ecuaciones de ser ciertas cuando se emplea el método de Euler y cuando se emplea el método de Euler mejorado, en ambos casos con $h = \pi/100$.

1.7.1. Escribir las condiciones de orden 1, 2 y 3 en términos de $\vec{b}$, $\vec{\mathbf{1}} = (1, 1, \ldots, 1)^T$ y las matrices A y $B = \mathrm{diag}(\vec{b})$ (la matriz diagonal con $b_{ii} = b_i$), sin hacer referencia a sus componentes.

1.7.2. a) Escribir el tablero del método

$$k_1 = f(x_n + \frac{h}{2}, y_n), \qquad k_2 = f(x_n + \frac{h}{2}, y_n + hk_1), \qquad y_{n+1} = y_n + \frac{h}{2}(k_1 + k_2)$$

y comprobar que no satisface la condición de suma por filas y es de orden 2.

b) Hallar todos los métodos de tablero

c_1	0	0
c_2	$(2\beta)^{-1}$	0
	b_1	b_2

con $\beta \neq 0$, que no satisfagan la condición de suma por filas y tengan orden dos.

c) Explicar por qué ninguno de los métodos del apartado anterior tiene ninguna ventaja respecto al correspondiente con $c_1 = 0$ y $c_2 = (2\beta)^{-1}$ que sí satisface la condición de suma por filas.

1.7.3. a) Comprobar que la condición de suma por filas equivale a decir que el polinomio de Taylor de orden uno de k_i (como función de h y bajo la hipótesis de localización) es $y'(x_n) + c_i y''(x_n)h$.

b) Explicar la siguiente frase tomada de [HNW] p. 132: "[La condición de suma por filas] fue supuesta por Kutta sin ningún comentario y expresa que todos los puntos donde f se evalúa son aproximaciones de primer orden a la solución".

Nota: En los ejercicios posteriores supondremos siempre que se verifica la condición de suma por filas, sin advertirlo cada vez.

1.7.4. Construir todos los métodos de orden 2 de la forma

0			
c_2	c_2		
c_3	0	c_3	
	0	0	1

Estudiar si alguno tiene orden 3.

1.7.5. Encontrar todos los tableros posibles de los métodos de Runge-Kutta explícitos de dos etapas con orden 1 y con orden 2. Hallar también todos los implícitos de una etapa con orden 2.

1.7.6. Determinar el orden del método,

$1/4$	$1/8$	$1/8$
$3/4$	$3/8$	$3/8$
	$1/2$	$1/2$

1.7.7. Comprobar que el siguiente método implícito es de orden (al menos) 3

$$\begin{array}{c|cc} (3+\sqrt{3})/6 & (3+\sqrt{3})/6 & 0 \\ (3-\sqrt{3})/6 & -\sqrt{3}/3 & (3+\sqrt{3})/6 \\ \hline & 1/2 & 1/2 \end{array}$$

1.7.8. Deducir las condiciones de orden tres sin mirar a la teoría para problemas escalares de la forma $y' = f(y)$.

1.7.9. Hallar todos los métodos de orden 2 con dos etapas y con $c_2 = c_1$, $b_1 = b_2$. ¿Tiene alguno orden 3?

1.7.10. Encontrar una solución para las condiciones de orden 3 de manera que dé lugar a un método explícito de 3 etapas (y orden 3) con $c_2 = c_3$ y $b_2 = b_3$. El método resultante se conoce como método de Nyström.

1.7.11. Probar que la *regla implícita del punto medio*

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

es un método de Runge-Kutta de una etapa. Escribir su tablero y calcular su orden.

Ⓛ **1.7.12.** Elaborar un programa FORTRAN que aproxime $y(1)$ empleando la regla implícita del punto medio al problema $y' = -40y$, $y(0) = 1$ con $h = 0'1$. Tratar de explicar por qué el resultado no es bueno.

1.8.1. Probar 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 pero no 4.

1.8.2. Consideramos el método de Runge-Kutta explícito de tablero

0				
1/4	1/4			
1/2	0	1/2		
1	1	−2	2	
	1/6	0	4/6	1/6

Obtener el orden del método.

1.8.3. Escoger tres árboles de orden 5 y comprobar si se cumplen las condiciones de orden correspondientes para el método dado por el tablero

0					
1/3	1/3				
1/3	1/6	1/6			
1/2	1/8	0	3/8		
1	1/2	0	−3/2	2	
	1/10	0	3/10	2/5	1/5

1.8.4. Hallar el único método de orden 4 con tablero

0				
c_2	c_2			
c_2	0	c_2		
c_4	0	0	c_4	
	b_1	b_2	b_2	b_1

comprobando que realmente tiene el orden asegurado. *Nota*: El método obtenido se llama método de Runge-Kutta por antonomasia.

1.8.5. Probar que la llamada *regla* 3/8, cuyo tablero es

0				
1/3	1/3			
2/3	−1/3	1		
1	1	−1	1	
	1/8	3/8	3/8	1/8

es de orden 4.

1.8.6. Probar que el método de Hammer-Hollingsworth

$(3-\sqrt{3})/6$	$1/4$	$(3-2\sqrt{3})/12$
$(3+\sqrt{3})/6$	$(3+2\sqrt{3})/12$	$1/4$
	$1/2$	$1/2$

satisface las condiciones de orden 4. Se puede probar que en cuanto a orden es el mejor método de Runge-Kutta con dos etapas, pero ¿qué desventaja tiene? *Indicación*: Los cálculos son largos pero se abrevian un poco escribiendo el tablero en términos de $\alpha = (3-\sqrt{3})/6$ y usando cuando sea necesario que $\alpha(1-\alpha) = 1/6$.

1.8.7. Verificar que el método

0						
1/3	1/3					
2/5	4/25	6/25				
1	1/4	-3	15/4			
2/3	6/81	90/81	-50/81	8/81		
4/5	7/30	18/30	-5/30	4/30	0	
	48/192	0	125/192	0	-81/192	100/192

no es de orden 5.

1.8.8. Sabiendo que el siguiente método es de orden 4 calcular α y β

0				
α	α			
$\beta/4$	0	$\beta/4$		
$3-\beta$	$3-\beta$	$-(2\alpha)^{-1}$	$(2\alpha)^{-1}$	
	1/6	0	4/6	1/6

1.8.9. Comprobar las condiciones de orden 4 para el *método semiexplícito de Butcher*

0	0	0	0
1/2	1/4	1/4	0
1	0	1	0
	1/6	4/6	1/6

Ⓛ **1.8.10.** Escribir un programa FORTRAN que aplique el método semiexplícito de Butcher a ecuaciones escalares lineales, es decir, de la forma $y' = g_1(x)y + g_2(x)$ con $g_1, g_2 : \mathbb{R} \longrightarrow \mathbb{R}$.

Ⓛ **1.8.11.** Escribir un programa FORTRAN que dado el tablero de un método decida si es de orden 5.

1.9.1. Considérense los árboles de LT_q etiquetados con los enteros de 1 a q con el orden habitual. Estudiando los valores posibles que pueden tomar $t(2)$, $t(3)$, $\ldots$, $t(q)$, probar que Card $LT_q = (q-1)!$

1.9.2. Según el problema anterior hay seis árboles en LT_4. Representarlos gráficamente y decir cuáles de ellos corresponden a un mismo árbol (sin etiquetas) de T_4.

1.9.3. Utilizando que Card $LT_q = (q-1)!$ dar una cota superior para el número de condiciones de orden que debe satisfacer un método para ser de orden 8. *Nota*: El número exacto se calculará en un problema posterior.

1.9.4. Hallar el único árbol de T_q con $\gamma(t) = q$ y escribir el diferencial elemental correspondiente y la condición de orden.

1.9.5. ¿Es inyectiva para todo q la función $\gamma : T_q \longrightarrow \mathbb{N}$? *Indicación*: Para $q \leq 4$ no hay contraejemplo.

1.9.6. Consideremos el problema bidimensional $(y^1, y^2)' = (1, g)$ con $g = g(y^1, y^2)$.

a) Hallar un árbol de orden 5 con $\gamma(t) = 30$ y otro con $\gamma(t) = 40$ y escribir las diferenciales elementales correspondientes al problema del enunciado.

b) Deducir del apartado anterior que para este tipo de problemas las condiciones de orden 5 son suficientes pero no son necesarias.

c) ¿Por qué el apartado anterior no contradice el teorema correspondiente (de Butcher) afirmando que un método es de orden p si y sólo si se cumplen las condiciones de orden?

***1.9.7** Hallar una fórmula para la derivada $n-$ésima de la composición de dos funciones $f, g : \mathbb{R} \longrightarrow \mathbb{R}$ y comprobarla para los casos $n = 1, 2, 3$. *Indicación*: Simplemente recuérdese la fórmula de Faà di Bruno.

1.9.8. Sea $A_q =$ Card T_q.

a) Suponiendo conocido que (para $x \to 0$)

$$A_1 + A_2 x + \cdots + A_{k+1} x^k = (1-x)^{-A_1}(1-x^2)^{-A_2} \cdots (1-x^k)^{-A_k} + O(x^{k+1}),$$

deducir sin necesidad de dibujar ningún árbol

$$A_1 = 1, \quad A_2 = 1, \quad A_3 = 2, \quad A_4 = 4, \quad A_5 = 9, \quad A_6 = 20.$$

Indicación: Emplear para los cálculos $(1-r)^{-1} = 1 + r + r^2 + r^3 + \ldots$ y $(1-r)^{-\alpha} = 1 + \alpha r + O(r^2)$.

b) La lista anterior continúa con $A_7 = 48$ y $A_8 = 115$. ¿Cuántas condiciones de orden debe satisfacer un método de Runge-Kutta para ser de orden 8?

****1.9.9** Demostrar la fórmula del apartado a) del ejercicio anterior.

Ⓛ **1.9.10.** Escribir un programa FORTRAN que al introducir un árbol de orden 6 y un tablero, decida si se cumple la correspondiente condición de orden.

1.10.1. Probar que si existiera un método de orden 10 explícito de 10 etapas con $b_{10} = 1$ entonces se debería cumplir $\max(|a_{i+1\,i}|) > 0'18$.

1.10.2. Usando los valores mencionados en un problema anterior para Card T_q con $1 \leq q \leq 8$, hallar el número de ecuaciones y de incógnitas que tienen las condiciones de orden correspondientes a un método explícito genérico de s etapas. Deducir del resultado cuáles son los valores naturales del mínimo de s para cada $1 \leq q \leq 8$. ¿Por qué esto no prueba rigurosamente las barreras de Butcher?

1.10.3. Supongamos un método de Runge-Kutta explícito de orden 5 con seis etapas y tal que $a_{65} = 0$. Demostrar que $b_6 a_{64} + b_5 a_{54} \neq 0$. *Indicación*: Hallar la condición de orden correspondiente al árbol lineal.

1.10.4. Tratar de explicar por qué en los tableros que sólo incluyen números racionales, los denominadores o al menos sus factores suelen aparecer repetidos. Concretamente, ¿por qué en los b_j no aparece nunca un denominador no trivial que sea coprimo con los demás?

1.10.5. Demostrar que si en un método de Runge-Kutta explícito con todos los c_i distintos el orden coincide con el número de etapas s, entonces los c_i determinan los b_i. En particular, conocida A sólo hay un posible $\vec{b}$. *Indicación*: Considérense las condiciones de orden $q = 1, 2, \ldots, s$ correspondientes al árbol que cumple $\gamma(t) = q$.

1.11.1. 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.

1.11.2. Probar que si en el problema anterior todos los elementos de M son nulos excepto m_{12}, m_{23} y m_{34}, entonces el orden del método es 4. ¿No contradice esto el teorema que afirmaba que un método explícito de s etapas tiene a lo más orden s?

****1.11.3** Hallar un método que sea de orden 4 para ecuaciones escalares de la forma $y' = f(y)$ y no lo sea en general para sistemas.

Ⓛ **1.11.4.** Comprobar numéricamente que el método de orden 2 antes citado converge más rápido para sistemas lineales con coeficientes constantes que para otros problemas.

1.12.1. Supongamos que en un problema aproximamos $y(b)$ usando sucesivamente pasos h, $2h$ y $4h$. Sean r_h, r_{2h}, r_{4h} los resultados obtenidos. Probar que para h suficientemente pequeño el orden p del método verifica

$$p \approx \frac{\log \|r_{4h} - r_{2h}\| - \log \|r_{2h} - r_h\|}{\log 2}$$

1.12.2. Al aplicar cierto método con diferentes pasos se ha obtenido la tabla

h	y_N
$0'125$	$3'30678344$
$0'0625$	$3'31658745$
$0'03125$	$3'31920958$
$0'015625$	$3'31988621$

a) Estimar el orden usando la fórmula del problema anterior.

b) Utilizar la extrapolación de Richardson para estimar el error cometido en los tres últimos valores.

1.12.3. En el problema anterior la solución real era $3'320116923\ldots$. Estimar de nuevo el orden usando ahora este dato.

1.12.4. a) Escribir la fórmula correspondiente a aplicar el método de Euler dos veces con paso $h/2$.

b) Usar la idea de Richardson para combinar a) con el método de Euler de paso h, obteniendo un método de orden 2. ¿Qué método es?

1.12.5. Explicar cómo utilizar el procedimiento del problema anterior para demostrar que hay métodos de Runge-Kutta explícitos de orden arbitrariamente grande. ¿Por qué esta construcción no es demasiado útil en la práctica?

1.12.6. Para el problema $y'(x) = y$, con $y(0) = 1$, el método de Runge-Kutta explícito de cuatro etapas de tablero

$$\begin{array}{c|cccc} 0 & & & & \\ 1/2 & 1/2 & & & \\ -1 & 1/2 & -3/2 & & \\ 1 & 0 & 4/3 & -1/3 & \\ \hline & 1/6 & 2/3 & 0 & 1/6 \end{array}$$

ha producido la siguiente tabla de errores al aproximar la solución en cierto punto

h	error
$0'3$	$1'28856 \cdot 10^{-4}$
$0'15$	$9'24133 \cdot 10^{-6}$
$0'075$	$5'77639 \cdot 10^{-7}$
$0'0375$	$3'74758 \cdot 10^{-8}$

a) Determinar el orden teórico del método usando árboles de Butcher y hallar el orden empírico empleando la tabla anterior.

b) Explicar la discrepancia entre ambos resultados.

1.12.7. Para un método de Runge-Kutta se obtiene la siguiente tabla de errores

h	error
1/11	0′02353
1/12	0′01703
1/13	0′01262
1/14	0′00955

h	error
1/15	0′00736
1/16	0′00577
1/17	0′00458
1/18	0′00368

Deducir el orden del método.

1.12.8. Supongamos que hemos programado un método de Runge-Kutta que aproxima $y(b)$ para algún problema escalar y muestra en pantalla el error cometido sin valor absoluto. Si al reducir el paso el error se va reduciendo y en algún momento cambia de signo, explicar por qué probablemente estemos viendo los errores de redondeo.

Ⓛ **1.12.9.** Escribir un programa FORTRAN que aplique un método de orden 3 con precisión simple a un problema del que se conozca la solución, y muestre una tabla con el error en función de h. Tratar de deducir a partir de qué h influyen los errores de redondeo.

Ⓛ **1.12.10.** Elaborar un programa FORTRAN que a partir de una tabla de pasos y errores estime el orden usando regresión lineal.

1.13.1. Al tomar dos pasos con $h = 0'1$ y uno con $\widetilde{h} = 0'2$ en un método de orden 3, obtenemos $y_2 = 2'1745$, $\widetilde{y}_2 = 2'1508$. Si admitimos un error máximo de $2 \cdot 10^{-3}$, ¿qué paso debiéramos tomar?

1.13.2. Considérese el problema $y' = y + \operatorname{sen}(xy)$, $y(0) = 1$. Con ayuda de una calculadora aplicar el método de Euler con $h = 0'1$ cambiando el paso por medio de extrapolación de Richardson con un error local tolerado de $0'01$ y un factor de seguridad $0'9$. Continuar los cálculos hasta que $x_n > 0'34$.

Ⓛ **1.13.3.** Escribir un programa FORTRAN que realice los cálculos del problema anterior.

1.14.1. Consideramos el par encajado de tres etapas cuyos únicos elementos no nulos son $c_2 = a_{21} = 1/2$, $c_3 = a_{33} = b_2 = 1$, $\widehat{b}_1 = 1/6$, $\widehat{b}_2 = 2/3$, $\widehat{b}_3 = 1/6$. Hallar el orden de ambos métodos y señalar si son explícitos o implícitos. *Nota*: En el resto de los problemas supondremos que todos los pares encajados son explícitos.

1.14.2. Construir todos los pares encajados 1(2) con dos etapas.

1.14.3. Explicar por qué si en un par encajado de s etapas $a_{si} = b_i$ y $b_s = 0$ entonces se reduce el número de operaciones. Encontrar los métodos del problema anterior que satisfacen este requerimiento.

1.14.4. Hallar algún par encajado de tipo 1(2) con tres etapas satisfaciendo $a_{3i} = b_i$ y $b_3 = 0$.

1.14.5. Repetir el problema enunciado anteriormente en el que se empleaba extrapolación de Richardson aplicada a $y' = y + \text{sen}(xy)$ para cambiar el paso, pero usando ahora en su lugar un par encajado 1(2). Comparar el trabajo computacional en ambos casos.

1.14.6. Probar que

0				
1/4	1/4			
1/2	0	1/2		
1	1	−2	2	
	1	−2	2	0
	1/6	0	4/6	1/6

es un par encajado 2(4).

1.14.7. Hallar todos los pares encajados 2(3) de la forma

$$\begin{array}{c|ccc} 0 & & & \\ 1 & 1 & & \\ 1/2 & a_{31} & a_{32} & \\ \hline & b_1 & b_2 & 0 \\ \hline & \widehat{b}_1 & \widehat{b}_2 & \widehat{b}_3 \end{array}$$

Ⓛ **1.14.8.** Usando algún par encajado, elaborar un programa FORTRAN que aplique el correspondiente método de paso variable a $y' = \cos(x/(1+y^2))$, $y(0) = 0$ con una tolerancia de $0'0001$ en el error local.

2. Problemas stiff

2.1.1. Caracterizar el valor de la condición inicial $y(0)$ en el problema

$$y' = \begin{pmatrix} -5 & -4 \\ 2 & 1 \end{pmatrix} y$$

para que al aplicar el método de Euler con $h = 1$ se cumpla $\lim y_n = 0$.

2.1.2. Probar el teorema de los círculos de Gershgorin que afirma que los autovalores de una matriz $A = (a_{ij})$ pertenecen a la unión de los círculos en el plano complejo dados por $D_i = \{z \ : \ |z - a_{ii}| \leq \sum_{j \neq i} |a_{ij}|\}$.

2.1.3. Considérese el problema $y' = Ay$, $y(0) = y_0$ con $A \in M_{2\times 2}(\mathbb{R})$ diagonalizable en $\mathbb{R}$. Demostrar que si los elementos de A son menores que uno en valor absoluto y $\lim_{x\to+\infty} y(x) = 0$, entonces al aplicar el método de Euler con cualquier $h < 1$ se cumple $\lim y_n = 0$. *Indicación*: Utilizar el teorema de los círculos de Gershgorin.

2.1.4. Considérese el problema

$$y' = \begin{pmatrix} 242 & 324 \\ -183 & -245 \end{pmatrix} y.$$

Probar que cualquiera que sea la condición inicial y el paso $h < 1$, al aplicar el método de Euler se tiene $\lim y_n = \lim_{x\to+\infty} y(x) = 0$. ¿Cómo es posible si los coeficientes son muy grandes?

Ⓛ **2.1.5.** Escribir un programa FORTRAN que calcule la solución numérica de

$$y' = \begin{pmatrix} -68 & -4 \\ 67 & 3 \end{pmatrix} y.$$

en el intervalo $[0, 1]$ por medio del método de Euler con $h = 0'04$, primero con el valor inicial $y(0) = (0, -63)^T$ y después con $y(0) = (-4, 67)^T$. Explicar los resultados.

2.2.1. Hallar el dominio de estabilidad lineal del método

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

2.2.2. 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 180^o alrededor de dicho punto), con respecto a la recta $\mathrm{Im}\, z = 0$, y con respecto a la recta $\mathrm{Re}\, z = -1$.

2.2.3. Se dice que un método es $L-$estable si al aplicarlo a $y' = \lambda y$ se cumple $y_{n+1} = R(\lambda h) y_n$ con $R(z) \to 0$ cuando $\mathrm{Re}\, z \to -\infty$.

a) Probar 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 $-\mathrm{Re}\, \lambda$ es grande en comparación con h^{-1}.

2.2.4. 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 $\lim y_n = 0$, donde y_n es la solución numérica obtenida al aplicar el método de Runge-Kutta clásico de cuatro etapas y orden cuatro.

Ⓛ **2.2.5.** Escribir un programa FORTRAN que compruebe el resultado del problema anterior.

2.3.1. Hallar la función de amplificación del método de Euler mejorado.

2.3.2. Demostrar que para un método de Runge-Kutta consistente el dominio de estabilidad lineal es no vacío y que el origen siempre pertenece a su cierre o adherencia. De hecho pertenece a la frontera del cierre. *Indicación*: $R(z) = 1 + z + O(z^2)$.

2.3.3. Usando la función de amplificación, tratar de dar una prueba del teorema que afirmaba que no hay métodos explícitos de orden p y menos de p etapas

*__2.3.4__ Probar que fijada una matriz cuadrada A se cumple para $z \to 0$

$$(I - zA)^{-1} = I + zA + z^2A^2 + \cdots + z^kA^k + O(z^{k+1}),$$

donde aquí la notación $O(z^{k+1})$ indica una matriz tal que todos sus elementos son $O(z^{k+1})$.

2.3.5. Demostrar que un método de Runge-Kutta es de orden 2 si y sólo si $R(z) = 1 + z + z^2/2 + O(z^3)$. *Indicación*: Se puede utilizar el problema anterior con $k = 1$.

2.3.6. 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 $R(z) = 1 + z + z^2/2 + z^3/6$.

*__2.3.7__ Demostrar que para cualquier $p > 2$ es posible encontrar métodos que no son de orden p y sin embargo satisfacen $R(z) = 1 + z + z^2/2! + \cdots + z^p/p! + O(z^{p+1})$.

2.3.8. Probar que si en un método de Runge-Kutta de dos etapas y orden dos se cumple $a_{11} = 1$, $a_{12} = 0$, entonces $\lim_{z\to\infty} R(z) \neq 0$.

2.3.9. Comprobar que la región de estabilidad del método

$$\begin{array}{c|ccc} 0 & & & \\ 1/2 & 1/2 & & \\ 3/4 & 0 & 3/4 & \\ \hline & 2/9 & 3/9 & 4/9 \end{array}$$

tiene intersección no vacía con el eje imaginario Re $z = 0$.

*__2.3.10__ Sea un método de Runge-Kutta consistente y explícito salvo porque $a_{11} \neq 0$. Demostrar que si su dominio de estabilidad lineal es $\mathcal{D} = \{\text{Re } z < 0\}$ entonces su orden es exactamente 2.

2.3.11. 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.

2.3.12. Estudiar si es $A-$estable el método de tablero

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

2.3.13. Comprobar si es $A-$estable el método de tablero

$$\begin{array}{c|cc} (3+\sqrt{3})/6 & (3+\sqrt{3})/6 & 0 \\ (3-\sqrt{3})/6 & -\sqrt{3}/3 & (3+\sqrt{3})/6 \\ \hline & 1/2 & 1/2 \end{array}$$

Indicación: Para simplificar un poco las operaciones conviene escribir $\alpha = (3+\sqrt{3})/6$ y emplear en los cálculos finales que $\alpha^2 = \alpha - 1/6$.

2.3.14. Se dice que una función racional $f(x) = P(x)/Q(x)$ es la *aproximante de Padé* (q,q) *de la exponencial* si $\text{gr}(P),\ \text{gr}(Q) \leq q$ y $f(x) = e^x + O(x^{2q+1})$. Se puede probar que para cada $q \in \mathbb{N}$, tal aproximante existe y es única.

a) Calcular la aproximante de Padé $(2,2)$ de la exponencial.

b) ¿Cuál es la función de amplificación para un método como el de Gauss-Legendre de dos etapas y orden 4?

*c) Demostrar que si un método tiene s etapas y orden $2s$, los coeficientes de su función de amplificación son números racionales.

Ⓛ **2.3.15.** Cuando una calculadora o un ordenador aproximan internamente e^x basta con que lo hagan para $x \in [0, \log 2]$ ya que $e^{m \log 2} = 2^m$ es un número exacto en base 2. Escribir un programa que dibuje en dicho intervalo las funciones $|f_1(x) - e^x|$ y $|f_2(x) - e^x|$, estimando el máximo de cada una de ellas, donde f_1 es la aproximación de Taylor de orden 4 y f_2 es la aproximante de Padé $(1 + x/2 + x^2/12)/(1 - x/2 + x^2/12)$. Decidir cuál de estas aproximaciones es mejor y cuántas cifras significativas de la exponencial puede dar.

3. Diferencias finitas

3.1.1. Sea el problema en una dimensión

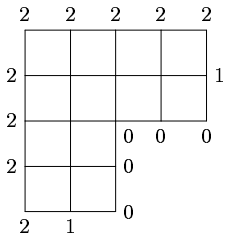
$$u'' = f \qquad u(a) = u(b) = 0.$$

Aplicar el método de diferencias finitas para reducirlo a un sistema lineal $A\vec{U} = \vec{b}$. Comprobar que $\det(A) \neq 0$ calculando el determinante explícitamente.

3.1.2. Resolver el problema

$$\begin{cases} \Delta u &= 0 & \text{en } \Omega \\ u &= g & \text{en } \partial\Omega \end{cases}$$

usando la fórmula de los cinco puntos con $h = k = 1$ para el dominio Ω de la figura y los valores de g indicados en ella.



Si se incrementase el valor de g en cada uno de los puntos del contorno en una milésima, determinar cómo se modificaría el valor de la solución.

3.1.3. Considérese la ecuación de Poisson $-\Delta u = f$ en $Q = [0,1] \times [0,1]$ con $u|_{\partial Q} = 0$. Demostrar que si discretizamos con $N = M$ (número de divisiones en X= número de divisiones en Y) y numeramos los nodos de izquierda a derecha y de arriba a abajo (como al escribir en los renglones de un cuaderno), entonces la matriz correspondiente al método de diferencias finitas con la fórmula de los cinco puntos es

$$-N^2 \begin{pmatrix} T & I & & \\ I & T & I & \\ & I & T & \ddots \\ & & \ddots & T \end{pmatrix}$$

donde I es la matriz identidad $(N-1) \times (N-1)$ y T es la matriz tridiagonal de las mismas dimensiones con $t_{ii} = -4$ y $t_{ij} = 1$ para $|i-j| = 1$.

3.1.4. Toda función u suficientemente regular que se anule en la frontera del cuadrado $Q = [0,3] \times [0,3]$ se puede escribir de forma única como una serie funcional del tipo

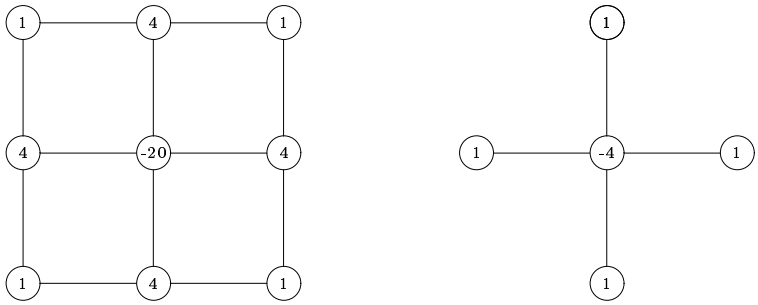
$$u(x,y) = \sum_{n,m=1}^{\infty} a_{n\,m} \phi_{nm}(x,y) \qquad \text{con } \phi_{nm}(x,y) = \operatorname{sen}(\frac{\pi}{3} nx) \operatorname{sen}(\frac{\pi}{3} my).$$

a) Dando por supuesto este resultado y desconsiderando problemas de convergencia, probar que la ecuación $\Delta u + \lambda u = 0$, $u|_{\partial Q} = 0$, tiene solución única para $\lambda < 2\pi^2/9$ y tiene infinitas soluciones para $\lambda = 2\pi^2/9$.

b) Hallar el menor λ para el que el esquema de diferencias finitas $\Delta_{hk} U_{ij} + \lambda U_{ij} = 0$ con $h = k = 1$ tenga infinitas soluciones y comparar el resultado con el del apartado anterior.

3.1.5. Considérese la ecuación de Laplace, $\Delta u = 0$, en el cuadrado $Q = [0,1] \times [0,1]$ con las condiciones de contorno $\frac{\partial u}{\partial \mathbf{n}} = -1$ si $x = 1$, $0 < y < 1$, $u = 0$ en el resto de la frontera de Q. Aproximar la solución utilizando la fórmula de cinco puntos con $h = k = 1/3$. *Nota*: Recuérdese que la derivada normal $\frac{\partial u}{\partial \mathbf{n}}$ se define como $\nabla u \cdot \mathbf{n}$ donde $\mathbf{n}$ es la normal exterior unitaria.

3.1.6. La llamada *fórmula de los nueve puntos* y la fórmula de los cinco puntos responden a las moléculas computacionales



Sean $6\Delta_9$ y Δ_5 los operadores discretos correspondientes con $h = k$.

a) Probar sin mirar a la teoría que $\Delta_5 u - \Delta u = O(h^2)$ para $u \in C^4$.

*b) Probar que $\Delta_9 u - \Delta u = \frac{h^2}{12}\Delta(\Delta u) + O(h^4)$ para $u \in C^6$.

c) Usando el apartado anterior, demostrar que $\Delta_9 u + (\lambda - \lambda^2 h^2/12)u = 0$ es una aproximación de orden 4 de la ecuación $\Delta u + \lambda u = 0$. Es decir, que si u verifica $\Delta u + \lambda u = 0$, entonces $\Delta_9 u + (\lambda - \lambda^2 h^2/12)u = O(h^4)$.

3.1.7. La ecuación $-\Delta u = f$ en $Q = [0,1] \times [0,1]$, $u|_{\partial Q} = 0$, tiene como solución $u(x,y) = (x - x^2)(y - y^2)$ cuando $f(x,y) = 2(x + y - x^2 - y^2)$ y $u(x,y) = \pi^{-2}\operatorname{sen}(\pi x)\operatorname{sen}(\pi y)$ cuando $f(x,y) = 2\operatorname{sen}(\pi x)\operatorname{sen}(\pi y)$. Explicar por qué en el primer caso, al aplicar la fórmula de cinco puntos con $h = k = 1/N$ el error cometido (el máximo de $|u(x_i, y_j) - U_{ij}|$) es cero. En el segundo caso se tiene la siguiente tabla de errores:

N	5	6	7	8	9
Error	$3'076 \cdot 10^{-3}$	$2'347 \cdot 10^{-3}$	$1'633 \cdot 10^{-3}$	$1'312 \cdot 10^{-3}$	$1'004 \cdot 10^{-3}$

Dando por supuesto que Error$\approx Ch^p$, calcular el orden empírico.

***3.1.8** Sea $L[u] = \partial^{2n}u/\partial x^{2n} + \partial^{2n}u/\partial y^{2n}$. Probar que para $u \in C^{2n+2}$

$$L[u](x_i, y_j) = \frac{1}{h^{2n}} \sum_{m=0}^{2n} (-1)^m \binom{2n}{m} \big(u(x_{i+n-m}, y_j) + u(x_i, y_{j+n-m})\big) + O(h^2).$$

Indicación: Puede ser útil considerar la fórmula elemental $(2 \text{ senh } t)^{2n} = \sum_{m=0}^{2n} (-1)^m \binom{2n}{m} e^{2(n-m)t}$ y sus derivadas en $t = 0$.

*__3.1.9__ Supongamos que se aplica el método de diferencias finitas al problema $-\Delta u = f$ en $R = [a,b] \times [c,d]$ con $h = k = 1/N$, de manera que sólo se

dan condiciones en las fronteras superior e inferior: $u(x,c) = g_1(x)$, $u(x,d) = g_2(x)$ mientras que en los segmentos verticales se aplica la fórmula de los cinco puntos definiendo $U_{N+1\,j} = U_{1\,N-j}$, $U_{-1\,j} = U_{N\,N-j}$. Demostrar que el sistema lineal que se obtiene es siempre compatible determinado. Tratar de encontrar una configuración geométrica que represente estas identificaciones de los nodos.

3.1.10. Explicar cómo se puede aplicar el método de diferencias finitas para aproximar la solución de la ecuación de ondas

$$\begin{aligned} &\frac{\partial^2 u}{\partial t^2} - \frac{\partial^2 u}{\partial x^2} = 0 \qquad t > 0,\ 0 < x < 1 \\ &u(x,0) = f(x), \quad \frac{\partial u}{\partial t}(x,0) = g(x) \\ &u(0,t) = u(1,t) = 0 \end{aligned}$$

3.1.11. Mediante la regla de la cadena (empleando fórmulas del tipo $\frac{\partial u}{\partial r} = \frac{\partial u}{\partial x}\frac{\partial x}{\partial r} + \frac{\partial u}{\partial y}\frac{\partial y}{\partial r}$), demostrar que en coordenadas polares el laplaciano se escribe como

$$\Delta u = \frac{\partial^2 u}{\partial r^2} + \frac{1}{r}\frac{\partial u}{\partial r} + \frac{1}{r^2}\frac{\partial^2 u}{\partial \theta^2}.$$

Utilizar esta expresión y una discretización en la que los incrementos de r y θ sean 1 y $\pi/6$, para escribir un esquema de diferencias finitas que aproxime la solución de la ecuación de Laplace $\Delta u = 0$, $u|_{\partial\Omega} = \sqrt{x^2+y^2}$ donde $\Omega = \{(x,y) \,:\, 1 \le x^2+y^2 \le 16,\ x \ge 0,\ y \ge 0\}$. *Nota*: Se obtiene un sistema que no se pide resolver. Aun así el ejercicio es largo.

3.1.12. Cuando se emplea el método de diferencias finitas para resolver la ecuación de Poisson en un dominio curvo Ω difícil de transformar en un cuadrado con un cambio sencillo, se consideran los *nodos de frontera* obtenidos por la intersección de las rectas horizontales y verticales de la malla con $\partial\Omega$. Cada nodo interior estará rodeado por otros cuatro nodos, ya sean interiores o de frontera. Procediendo como en la fórmula de cinco puntos, se puede encontrar una combinación lineal de los valores de la función en ellos que aproxime el laplaciano. Aplicar esta idea al problema $\Delta u = 0$, $u|_{\partial\Omega} = 3x - x^2 - y^2$ donde $\Omega = \{(x,y) \,:\, x^2+y^2 \le 3x,\ y \ge 0\}$, tomando $h = k = 1$.

Ⓛ **3.1.13.** Escribir un programa MATLAB o FORTRAN (lo primero es más sencillo) que aplique el método de diferencias finitas al problema de contorno (unidimensional)

$$-u'' + u = f, \qquad u(a) = c, \quad u(b) = d.$$

Tratar de estimar empíricamente cuál es el orden de la aproximación.

Ⓛ **3.1.14.** Escribir un programa MATLAB que aproxime la solución de la ecuación de Poisson $-\Delta u = f$ en $Q = [0,1] \times [0,1]$ con $u|_{\partial Q} = 0$ para cualquier $h = k = 1/N$. *Indicación*: En uno de los problemas anteriores se describe explícitamente la matriz del sistema resultante.

Ⓛ **3.1.15.** Rehacer el programa del ejercicio anterior para que permita condiciones de frontera generales del tipo $u|_{\partial Q} = g$.

4. Elementos finitos

4.1.1. Dado el problema $-u'' + u = 1$, $u(0) = u(1) = 0$; hallar las aproximaciones de la solución que se obtienen al imponer la condición de Galerkin con cada uno de los siguientes conjuntos de funciones base

a) $\mathcal{B}_1 = \{\phi(x - 1/3), \phi(x - 2/3)\}$ con $\phi(x) = \max(0, 1/3 - |x|)$.

b) $\mathcal{B}_2 = \{\text{sen}(\pi x), \text{sen}(2\pi x), \text{sen}(3\pi x)\}$.

c) $\mathcal{B}_3 = \{x(1-x), x^2(1-x)\}$.

4.1.2. Aplicar el método de elementos finitos a $-u'' + u = x$, $u(0) = u'(1) = 0$, considerando los nodos $x_j = j/3$, $j = 0, 1, 2, 3$ (tómese en x_1 y x_2 la función tejado habitual y en x_3 sólo la mitad izquierda de ella).

4.1.3. Repetir el problema anterior pero ahora con las condiciones $u(0) = 0$, $u'(1) = 1$. Comprobar en ambos problemas cuán cerca está la solución aproximada de cumplir la condición sobre $u'(1)$.

4.1.4. Hallar la matriz de rigidez y el vector de carga cuando se aplica el método de elementos finitos a $-u'' + (x+2)u = 1$, $u(-1) = u(1) = 0$ considerando los nodos $x_j = j/2$ con $j = -2, -1, 0, 1, 2$. Para simplificar los cálculos empléese que $\int x\, \phi_1^2 = -\int x\, \phi_3^2 = 8 \int x\, \phi_1\phi_2 = -8 \int x\, \phi_2\phi_3 = -\frac{1}{6}$.

4.1.5. Hallar la forma general de la matriz de rigidez correspondiente al problema $-u''+u = f$, $u(0) = u'(1) = 0$ para N nodos igualmente espaciados.

4.1.6. Sea el problema $-(xu')' = f$, $u(1) = u(2) = 0$. Supongamos que escogiendo las funciones base $\{\phi(x), \phi(x-1/4), \phi(x-1/2)\}$ con cierta ϕ tal que sop $\phi = [1, 3/2]$, la matriz de rigidez es

$$\begin{pmatrix} 239/48 & -17/12 & 0 \\ -17/12 & 97/16 & -5/3 \\ 0 & -5/3 & 343/48 \end{pmatrix}$$

Calcular la matriz de rigidez cuando se consideran quince funciones base, $\tilde{\phi}_j$, obtenidas a partir de las anteriores de la forma natural con homotecias y traslaciones. *Indicación*: Probar primero que $\tilde{\phi}_j(x) = \phi(4x - \frac{11+j}{4})$.

4.1.7. Considérense los nodos $0 = x_0 < x_1 < \cdots < x_N = 1$ con $x_{i+1} - x_i = 1/N$. Probar que la matriz correspondiente al esquema de diferencias finitas para el problema $-u'' = f$, $u(0) = u(1) = 0$; coincide con la matriz de rigidez al aplicar el método de elementos finitos (salvo multiplicación por una constante).

4.1.8. Probar que si f es constante en el problema anterior entonces el método de diferencias finitas y el de elementos finitos dan el valor exacto de la solución en los nodos. *Nota*: Aunque $\sum \gamma_l \phi_l$ y la solución u coincidan en los nodos porque $\gamma_l = u(x_l)$, es obvio que no lo hacen en todos los puntos porque la primera función es lineal a trozos y la segunda no lo es (para $f \not\equiv 0$). Este fenómeno está relacionado con la llamada *superconvergencia*.

4.1.9. Sean las funciones cuadráticas a trozos

$$\phi_1 = \begin{cases} 4x(1-x) & \text{si } 0 \le x \le 1 \\ 0 & \text{en otro caso} \end{cases} \qquad \phi_2 = \begin{cases} x(2x-1) & \text{si } 0 \le x \le 1 \\ (2x-3)(x-2) & \text{si } 1 \le x \le 2 \\ 0 & \text{en otro caso} \end{cases}$$

Aplicar el método de elementos finitos con las funciones base $\phi_1(x)$, $\phi_2(x)$ y $\phi_1(x-1)$, al problema $-u'' + u = x^2 - 2x - 2$, $u(0) = u(2) = 0$ y comprobar que la solución obtenida coincide con la exacta. *Indicación*: Para simplificar los cálculos es conveniente notar que ϕ_2 es simétrica por $x = 1$. Utilícese también que $\int \phi_1^2 = 2\int \phi_2^2 = 8\int \phi_1\phi_2 = \frac{8}{15}$.

Ⓛ **4.1.10.** Representar gráficamente las funciones obtenidas en cada uno de los apartados del primer problema y comparar el resultado con la solución exacta $u = 1 - \cosh x + \alpha \operatorname{senh} x$ con $\alpha = (\cosh 1 - 1)/\operatorname{senh} 1 = 0'4621171\ldots$

Ⓛ **4.1.11.** Elaborar un programa MATLAB que aplique el método de elementos finitos a los problemas de contorno de la forma $-u'' + \alpha u = f$, $u(a) = u(b) = 0$, y dibuje la gráfica correspondiente.

4.2.1. Hallar una función $u \in C^2([0, 1/2])$ con $u(0) = u(1/2) = 0$ tal que para toda v con estas mismas características se cumpla

$$\int_0^{1/2} (u'v' + (4u - 1)v) = 0.$$

4.2.2. Escribir el problema $-xu'' - u' + 2u = 1$, $u(1) = u'(3) = 0$ en su formulación débil y aplicarle el método de elementos finitos empleando sólo dos funciones base (de tipo "tejado", la segunda dividida por la mitad).

4.2.3. Se considera el problema en *forma divergencia* en $\mathbb{R}^3$

$$-\operatorname{div}(A(x)\nabla u) = f \qquad \text{en} \quad \Omega \subset \mathbb{R}^3$$

donde A es una matriz definida positiva 3×3. Escribir su formulación débil bajo las condiciones de Dirichlet y de Neumann. *Indicación*: Calcular primero $\operatorname{div}(vA(x)\nabla u)$.

4.2.4. Dado el problema de contorno $u'' = 2$, $u(1) = u(-1) = 0$, cuya solución es $u(x) = x^2 - 1$; escribir su formulación variacional y concluir que para toda $u \in C^2$ con $u(1) = u(-1) = 0$ se cumple

$$\frac{8}{3} + \int_{-1}^{1} \left((u')^2 + 4u\right) \geq 0.$$

*__4.2.5__ Tratar de demostrar la desigualdad anterior directamente, sin emplear la formulación variacional, sólo con argumentos de cálculo elemental.

4.2.6. Sea el problema $-a(x)u'' + b(x)u' + c(x)u = f$ donde a y c son funciones estrictamente positivas. Demostrar que multiplicando por una función adecuada se puede escribir en la forma $-(\tilde{a}(x)u')' + \tilde{c}(x)u = \tilde{f}$ y hallar su formulación débil bajo las condiciones $u(a) = u(b) = 0$.

4.2.7. Considérese la ecuación biarmónica en $\Omega \subset \mathbb{R}^2$

$$\frac{\partial^4 u}{\partial x^4} + 2\frac{\partial^4 u}{\partial x^2 \partial y^2} + \frac{\partial^4 u}{\partial y^4} = f$$

bajo las condiciones $u = \frac{\partial u}{\partial \mathbf{n}} = 0$ en $\partial\Omega$. Demostrar que su formulación débil es

$$\int_\Omega \Delta u \Delta v = \int_\Omega f v$$

para v en cierto espacio de funciones (en realidad en $H_0^2(\Omega)$) satisfaciendo las mismas condiciones de frontera. *Indicación*: Calcular primero $\Delta(\Delta u)$.

4.2.8. Para $n = 1, 2, 3, \ldots$ sean las funciones

$$f_n(x) = \begin{cases} nx & \text{si } x \in [-\frac{1}{2n}, \frac{1}{2n}] \\ \frac{1}{2}\,\mathrm{sgn}(x) & \text{si } x \in [-1,1] - [-\frac{1}{2n}, \frac{1}{2n}] \end{cases}$$

Demostrar que f_n tiene derivada débil que está uniformemente en L^1, y que para cualquier función $g \in C([-1,1])$ se cumple

$$\lim_{n\to\infty} \int_{-1}^{1} f_n'\, g = g(0).$$

Deducir que no existe ninguna función de L^1 que sea el límite (en L^1) de f_n'. *Nota*: No obstante, la teoría de distribuciones da un sentido a este límite y se dice que f_n' converge a la distribución delta de Dirac.

Ⓛ **4.2.9.** Según un problema anterior, las funciones con $u(-1) = u(1) = 0$ deben satisfacer $\frac{8}{3} + \int_{-1}^{1} \left((u')^2 + 4u\right) \geq 0$. Escribir un program FORTRAN que dada una función $w \in C^2$ compruebe aproximadamente esta desigualdad para $u = (x^2 - 1)w(x)$. Utilícese la regla de Simpson para aproximar la integral y el cociente incremental para aproximar la derivada.

Ⓛ **4.2.10.** Si se toma $u = \sum_{n=1}^{N} a_n \cos(\frac{\pi}{2}(2n+1)x)$ en el problema anterior, entonces la desigualdad integral se transforma en una desigualdad que involucra los a_n. Escribir un programa FORTRAN que busque para cada N los a_n (digamos con $a_n \in [-3/2, 3/2]$) para los que la expresión obtenida es mínima. Dibujar la gráfica de la función u resultante y comprobar que se parece a la de $x^2 - 1$.

4.3.1. Sean $\phi_1,\ \phi_2,\ \phi_3$ las funciones base (pirámide) asociadas a los nodos de cierto triángulo K demostrar que $\phi_1 + \phi_2 + \phi_3 \equiv 1$ en K.

4.3.2. Aplicar el método de elementos finitos a $-\Delta u + u = 1$ en el cuadrado $|x| + |y| \leq 1$ con condiciones de Dirichlet homogéneas en el lado superior derecho y condiciones de Neumann homogéneas en los otros. Utilícense sólo dos triángulos: el correspondiente a $x \geq 0$ y el correspondiente a $x \leq 0$.

4.3.3. Utilizar el método de elementos finitos en $Q = [0,2] \times [0,2]$ con los cuatro triángulos determinados por las aristas de Q y el punto $(1,1)$, para aproximar la solución del problema

$$-\Delta u = 0, \qquad u(0,y) = u(x,0) = 0, \qquad \frac{\partial u}{\partial x}(2,y) = 0,\ \frac{\partial u}{\partial y}(x,2) = 1.$$

Indicación: Aprovechar el hecho de que sólo dos de las funciones base participan realmente en el problema.

4.3.4. Repetir el problema anterior ahora bajo las condiciones $u(0,y) = y$, $\frac{\partial u}{\partial y}(x,2) = 0$, $u(x,0) = 0$, $u(2,y) = 0$.

4.3.5. Sea la función definida por $\phi(x,y) = (1 - x\ \mathrm{sgn}(x))(1 - y\ \mathrm{sgn}(y))$ en $Q = [-1,1] \times [-1,1]$ y por cero en otro caso. Con ella se pueden considerar elementos finitos cuadrangulares en lugar de triangulares. Aproximar la solución de $-\Delta u + u = xy$, $\frac{\partial u}{\partial n}\big|_{\partial\Omega} = 0$, con $\Omega = [0,1] \times [0,1]$, utilizando como funciones base las restricciones a Ω de $\phi(x,y)$, $\phi(x-1,y)$, $\phi(x-1,y-1)$ y $\phi(x,y-1)$.

4.3.6. Considérese la *ecuación del calor*

$$\frac{\partial u}{\partial t} - \frac{\partial^2 u}{\partial x^2} = f, \qquad u(0,t) = u(1,t) = 0, \qquad u(x,0) = g(x).$$

Para aplicar el método de elementos finitos a este tipo de problemas, se impone la condición de Galerkin con $u(x,t) = \sum \gamma_l(t)\phi_l(x)$ y ϕ_l las funciones tejado habituales. Es decir, se supone que los coeficientes γ_l dependen de t. Demostrar que esto lleva a una ecuación diferencial de la forma

$$A_1\vec{\gamma}'(t) + A_2\vec{\gamma}(t) = \vec{b} \qquad \text{con} \quad \gamma_j(0) = g(x_j)$$

donde A_1 y A_2 son matrices. Calcularlas cuando se toman las funciones base asociadas a la discretización $x_j = j/3$ con $j = 0, 1, 2, 3$.

Ⓛ **4.3.7.** Elaborar un programa MATLAB que permita aplicar el método de elementos finitos a $-\Delta u + u = f$, $\frac{\partial u}{\partial n}\big|_{\partial Q} = 0$, en $Q = [0, N] \times [0, N]$ utilizando como elementos la partición en cuadrados unidad y como funciones base los trasladados de la función $\phi(x, y) = (1 - x\ \mathrm{sgn}(x))(1 - y\ \mathrm{sgn}(y))$ restringida a $Q = [-1, 1] \times [-1, 1]$, que aparecía en un problema anterior.

Ⓛ **4.3.8.** Modificar el programa de elementos finitos para que admita condiciones de Dirichlet no homogéneas y comprobar numéricamente, sustituyendo diferentes funciones en la frontera, el principio del máximo para la ecuación $-\Delta u = 0$. *Nota*: En este problema y en el siguiente se supone que se tiene el programa de elementos finitos que se pedía en la práctica.

Ⓛ **4.3.9.** Modificar el programa para que admita condiciones de Robin de la forma $u + \alpha\frac{\partial u}{\partial n} = 0$. Comprobar numéricamente que cuando α crece, la solución se parece a la obtenida bajo las condiciones de Neumann homogéneas.

4.4.1. Considérese el espacio $W = \{f \in C([0, \pi]) \ : \ f(0) = f(\pi) = 0\}$ con el producto escalar usual (de L^2), y sea V el subespacio generado por $\{\mathrm{sen}\, x, \mathrm{sen}(2x), \mathrm{sen}(3x)\}$. Hallar la distancia de $g(x) = x(\pi - x)$ a V.

4.4.2. Sea $Q_{[0,N]} = \{f \in C([0, N]) \ : \ f(0) = f(N) = 0,\ f\big|_{[j,j+1]} =$ función cuadrática, para todo $0 \le j < N\}$. Demostrar que $Q_{[0,N]}$ está generado por traslaciones enteras de las funciones

$$\phi_1 = \begin{cases} 4x(1-x) & \text{si } 0 \le x \le 1 \\ 0 & \text{en otro caso} \end{cases} \qquad \phi_2 = \begin{cases} x(2x-1) & \text{si } 0 \le x \le 1 \\ (2x-3)(x-2) & \text{si } 1 \le x \le 2 \\ 0 & \text{en otro caso} \end{cases}$$

4.4.3. Utilizando el resultado anterior demostrar que la solución de $-u''+(x+1)u = x^3+x^2+x-2$, $u'(0)=0$, $u'(10)=20$ coincide exactamente con la obtenida por el método de elementos finitos (con las funciones base indicadas).

Soluciones

1. Métodos de un paso

1.1.1. Para resolver la ecuación, basta separar las variables

$$y' + y^2 = 0 \;\Rightarrow\; -\frac{y'}{y^2} = 1 \;\Rightarrow\; \frac{1}{y} = x + \frac{1}{2} \;\Rightarrow\; y = \frac{1}{x + 1/2}.$$

La solución "explota" en $x = -1/2$. No se contradice el teorema de existencia porque éste sólo aseguraba la existencia de solución en un entorno (de tamaño indeterminado) de la condición inicial.

1.1.2. a) Si se cumpliera $|\sqrt{|x|} - 0| \leq K|x - 0|$, dividiendo entre $|x|$ y tomando límites cuando $x \to 0$, se tendría una contradicción.

b) La más obvia es $y_1 \equiv 0$. Resolviendo la ecuación de la forma habitual (suponiendo por ejemplo $y > 0$), se tiene que $y(x) = (x - K)^2/4$ es solución de $y' = |y|^{1/2}$ en $x > K$; lo que permite fabricar nuevas soluciones pegando un trozo de y_1. Por ejemplo

$$y_2(x) = \begin{cases} 0 & \text{si } x \leq 0 \\ x^2/4 & \text{si } x \geq 0 \end{cases} \qquad y_3(x) = \begin{cases} 0 & \text{si } x \leq 1/2 \\ (x - 1/2)^2/4 & \text{si } x \geq 1/2 \end{cases}$$

*c) Obviamente lo que hay que probar es que la única solución C^2 es la idénticamente nula.

Supóngase que existe una solución C^2 de $y' = |y|^{1/2}$, $y(0) = 0$ con $y(a) \neq 0$ para algún $a \in (-1, 1)$ con $y(a) > 0$ (el otro caso es análogo y se menciona más adelante). En un entorno de A, y no se anula y se tiene

$$y' = |y|^{1/2} \;\Rightarrow\; y' y^{1/2} = 1 \;\Rightarrow\; y(x) = (x - K)^2/4, \quad x > K.$$

Por construcción, la solución no nula C^2 buscada debe coincidir con $(x - K)^2/4$ para $x > K$ (nótese que mientras $y > 0$ se tiene unicidad). Si $K < 0$ se llega a una contradicción con $y(0) = 0$. Si $K > 0$ necesariamente y es nula en el intervalo $[0, K]$ por ser creciente, $y' \geq 0$ y cumplir $y(0) = y(K) = 0$, pero entonces $y'' = 0$ en el interior de dicho intervalo mientras que $y'' = 1/2$ si $x > K$, es decir, $y \notin C^2$.

En definitva, el único caso posible es $K = 0$. Si se cumpliera $y(x) = 0$ para $x < 0$, por se creciente se llegaría de nuevo a que $\lim_{x \to 0^-} y'' = 0 \neq \lim_{x \to 0^+} y'' = 1/2$. Con un argumento simétrico al del párrafo anterior, se tiene que si $y(\widetilde{a}) < 0$ para algún $\widetilde{a} < 0$, entonces y coincide con $-(x - \widetilde{K})^2/4$

para $x < \widetilde{K}$ y además $\widetilde{K} = 0$. Esto lleva la la contradicción final $\lim_{x\to 0^-} y'' = -1/2 \neq \lim_{x\to 0^+} y'' = 1/2$.

1.1.3. Basta elegir la matriz

$$A = \begin{pmatrix} 0 & 1 & 0 & 0 & \dots & 0 \\ 0 & 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \dots & 1 \\ -a_0 & -a_1 & -a_2 & -a_3 & \dots & -a_{n-1} \end{pmatrix}$$

1.1.4. Hay que probar

$$\sup_{y\neq\widetilde{y}} \frac{\|Ay - A\widetilde{y}\|}{\|y - \widetilde{y}\|} = \max \sqrt{\lambda_j}.$$

Claramente el primer miembro se puede escribir como

$$\sup_{\vec{x}\neq\vec{0}} \frac{\|A\vec{x}\|}{\|\vec{x}\|} = \sup_{\|\vec{u}\|=1} \|A\vec{u}\| = \sup_{\|\vec{u}\|=1} \sqrt{\vec{u}^t A^t A \vec{u}}.$$

Diagonalizando $A^t A$ en una base ortonormal, se tiene que existe una matriz ortogonal U y una matriz diagonal (no negativa por serlo $A^t A$) tales que

$$\sup_{\vec{x}\neq\vec{0}} \frac{\|A\vec{x}\|}{\|\vec{x}\|} = \sup_{\|\vec{u}\|=1} \sqrt{\vec{u}^t U^t D U \vec{u}} = \sup_{\|U^{-1}\vec{u}\|=1} \sqrt{\vec{v}^t D \vec{v}}.$$

Empleando que $\|U^{-1}\vec{v}\| = \|\vec{v}\|$, por ser U ortogonal, se obtiene fácilmente el resultado deseado.

1.2.1. Al aplicar el método de Euler hay que iterar

$$y_{n+1} = y_n - h y_n^2 \qquad \text{con } y_0 = 1.$$

Lo que lleva a

$$y_1 = \frac{3}{4}, \qquad y_2 = \frac{34}{64}, \qquad y_3 = \frac{8463}{16384}, \qquad y_4 = 0'449836\dots$$

La solución exacta es $y(x) = 1/(1+x)$, con lo cual el error es en valor absoluto $0'050163\ldots$ Con el teorema correspondiente se obtiene sin embargo la cota (se ha tomado $L = 2 = \sup|\partial f/\partial y|$ en $x \in [0,1]$)

$$\frac{C}{2L}(e^{L(b-a)} - 1)h = \frac{2}{2\cdot 2}(e^{2(1-0)} - 1)\cdot\frac{1}{4} = 0'79863\ldots$$

Lo cual es unas 16 veces mayor que el error real.

1.2.2. Obviamente converge a cero al elegir $y_0 = 0$. Si se elige $y_0 = 10^{-10}$, por el teorema de convergencia, el método de Euler debe converger en un entorno de la condición inicial, a la única solución correspondiente de

$$\begin{cases} y' = |y|^{1/2} \\ y(0) = 10^{-10} \end{cases}$$

Es decir, a $y(x) = (x + 2\cdot 10^{-5})^2/4$.

1.2.3. a) Directamente, se obtiene por el método de Euler

$$y_{n+1} = (1+\alpha h)y_n + h\beta \qquad \text{con} \quad y_0 = 0.$$

Iterando se llega a

$$y_n = \big((1+\alpha hh)^{n-1} + (1+\alpha hh)^{n-2} + \cdots + 1\big)h\beta,$$

que, sumando la progresión geométrica es, para $\alpha \neq 0$,

$$y_n = \frac{\beta}{\alpha}\big((1+\alpha h)^n - 1\big).$$

Si $\alpha = 0$ se obtiene fácilmente $y_n = nh\beta$.

b) Dividiendo ambos miembros entre $y + \beta/\alpha$ e integrando, se llega a que la verdadera solución es

$$y(x) = \frac{\beta}{\alpha}(e^{\alpha x} - 1).$$

Del apartado anterior

$$y_N = \frac{\beta}{\alpha}\big((1+\alpha b/N)^N - 1\big), \qquad h = (b-0)/N,$$

que tomando límites $N \to \infty$ coincide con $y(b)$.

*c) Se demuestra que

$$L = \lim_{N\to\infty} |y(b) - y_N| N$$

es una constante no nula, se deducirá lo primero (recuérdese, $h = b/N$), y también lo segundo, porque en ese caso

$$\lim_{h\to 0} \frac{|y(b) - y_N|}{h^p} = \lim_{N\to\infty} \frac{|y(b) - y_N| N^p}{b^p} = b^{-p} L \lim_{N\to\infty} N^{p-1} = \infty.$$

Para calcular rápidamente L lo mejor es usar que $(1+\alpha b/N)^N = e^{N\log(1+\alpha b/N)}$ junto con las aproximaciones de Taylor $\log(1+x) = x - x^2/2 + O(x^3)$, $e^x = 1 + x + O(x^2)$, cuando $x \to 0$. Con esto

$$\begin{aligned} L &= \lim |e^{\alpha b} - (1+\alpha b/N)^N| N = \lim |e^{\alpha b} - e^{\alpha b} \cdot e^{-\alpha^2 b^2/(2N)} \cdot e^{O(N^{-2})}| N \\ &= \lim \big| e^{\alpha b} - e^{\alpha b}\big(1 - \frac{\alpha^2 b^2}{2N} + O(N^{-2})\big)\big| N = \frac{\alpha^2 b^2}{2} e^{\alpha b} \neq 0. \end{aligned}$$

1.2.4. La iteración es ahora

$$y_{n+1} = y_n + hAy_n = (I + hA)y_n \qquad \text{con } y_0 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

Esto conduce a

$$y_n = \begin{pmatrix} 1+2h & 2h \\ 0 & 1+2h \end{pmatrix}^n \begin{pmatrix} 1 \\ 1 \end{pmatrix} = (1+2h)^n \begin{pmatrix} 1 & 2h/(1+2h) \\ 0 & 1 \end{pmatrix}^n \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

Fácilmente de aquí se deduce

$$y_n^1 = (1+2h)^n (1 + \frac{2hn}{1+2h}), \qquad y_n^2 = (1+2h)^n.$$

La segunda ecuación del enunciado implica $y^2 = e^{2x}$, mientras que sustituyendo en la primera, mediante las técnicas habituales se llega a $y^1 = e^{2x}(1+2x)$.

Si $h = b/N$, se sigue que $y_N^1 \to y^1(b)$, $y_N^2 \to y^2(b)$ cuando $n \to \infty$.

El mismo argumento del problema anterior prueba que $|y^2(b) - y_N^2| \neq O(h^p)$ para $p > 1$, en particular $\|y(b) - y_N\| \neq O(h^p)$. Por otra parte, $\|y(b) - y_N\| = O(h)$ es consecuencia del teorema de convergencia.

1.2.5. A partir de $y_{n+1} = y_n + h(1 + x_n^2)$ se obtiene al iterar con $y_0 = 0$,

$$\begin{aligned} y_N =& h(1 + 0^2) + h(1 + h^2) + \cdots + h(1 + (N-1)^2) \\ =& Nh + h^3 \frac{(2N-1)}{3} \frac{(N-1)N}{2} \end{aligned}$$

La solución exacta en b es $y(b) = b + b^3/3$, sin embargo la solución aproximada con $h = b/N$ es

$$y_N = b + b^3\big(\frac{1}{3} - \frac{1}{2N} + \frac{1}{6N^2}\big).$$

Con lo cual $y(b) - y_N \neq O(N^{-2})$.

1.2.6. Procediendo como en el problema anterior, con $b = 1$,

$$\begin{aligned} y_N =& h^2(0 + 1 + \cdots + N - 1) - h^3(1 + 0^2 + 1^2 + \cdots + (N-1)^2) \\ =& h^2 \frac{N(N-1)}{2} - h^3 \frac{(2N-1)}{3} \frac{(N-1)N}{2} \\ =& \frac{1}{2} - \frac{1}{2N} - \big(\frac{1}{3} - \frac{1}{2N} - \frac{1}{6N^2}\big) = \frac{1}{6} - \frac{1}{6N^2} \end{aligned}$$

Como $y(1) = 1/6$, se deduce $|y(1) - y_N| = h^2/6$. Si el cálculo anterior se hace con $b = 2$, tomando $h = 2/N$, se llega de la misma forma a $y_N = y(1) + 2/N + O(N^{-2})$, por consiguiente no se cumple $|y(1) - y_N| = O(h^2)$.

1.2.7. a) Sea $x_i = 10i$, $0 \leq i \leq 5$, con lo cual las indicaciones estarán en $s_i = 50 - x_i$ espaciadas una distancia $h = 10$. Al ir de s_0 a s_1, la velocidad es $s_0 = 50$, y el tiempo que se tarda es

$$y_1 = y_0 + \frac{h}{50 - x_0} \qquad \text{con } y_0 = 0.$$

De la misma forma, al ir de s_1 a s_2, la velocidad es $s_1 = 50 - x_1$ y el tiempo transcurrido es

$$y_2 = y_1 + \frac{h}{50 - x_1}.$$

Y así sucesivamente. Por tanto el tiempo total, y_5, es el resultado de aplicar el método de Euler a $y' = 1/(50 - x)$, $y(0) = 0$, con $h = 10$. Al hacer las cuentas se obtiene

$$y_5 = 0 + \frac{10}{50} + \frac{10}{40} + \frac{10}{30} + \frac{10}{20} + \frac{10}{10} = 2\text{h } 17\text{min}.$$

b) Esto sería como aplicar el método de Euler cuando $h \to 0$, con lo cual el tiempo en función del espacio recorrido $y = y(x)$ vendría dado por la solución exacta de la ecuación diferencial anterior. Ésta es $y(x) = -\log(1 - x/50)$ y se comprueba que $\lim_{x\to 50^-} y(x) = +\infty$. Es decir, no llega nunca.

1.2.8. Es muy sencillo obtener $y_n = (1+h)^n$, $u_n = (1+h/2)^{2n}$. Por tanto todo lo que hay que probar es

$$(1+h)^n < (1+h/2)^{2n} < e^{nh}.$$

Extrayendo raíces $n-$ésimas

$$1 + h < 1 + h/2 < e^h.$$

La primera desigualdad es obvia y la siguiente es consecuencia del desarrollo $e^h = 1 + h + h^2/2 +$ (términos positivos).

*__1.2.9__ Iterando la fórmula del método de Euler se obtiene

$$\begin{aligned} y_N =& h\Big(\frac{1}{\sqrt{1}} + \frac{1}{\sqrt{1-1/N}} + \frac{1}{\sqrt{1-2/N}} + \cdots + \frac{1}{\sqrt{1-(N-1)/N}}\Big) \\ =& h\sqrt{N} + h\sqrt{N}\Big(\frac{1}{\sqrt{2}} + \frac{1}{\sqrt{3}} + \cdots + \frac{1}{\sqrt{N}}\Big) \end{aligned}$$

La indicación se sigue directamente del teorema del valor medio aplicado a $f(x) = 1/\sqrt{4x}$ en $[n, n+1]$ y $[n-1, n]$. Al utilizarla en el término entre paréntesis se llega a

$$(\sqrt{3} - \sqrt{2}) + (\sqrt{4} - \sqrt{3}) + \cdots + (\sqrt{N+1} - \sqrt{N}) < \frac{y_N - h\sqrt{N}}{2h\sqrt{N}} < (\sqrt{2} - \sqrt{1}) + (\sqrt{3} - \sqrt{2}) + \cdots + (\sqrt{N} - \sqrt{N-1}).$$

Simplificando las sumas telescópicas y operando

$$2h\sqrt{N^2+N} - 2h\sqrt{2N} < y_N - h\sqrt{N} < 2hN - 2h\sqrt{N}.$$

Sustituyendo $N = 1/h$ y sumando $h^{1/2} - 2$ a todos los términos

$$2\sqrt{1+h} - 2 - (2\sqrt{2} - 1)H^{1/2} < y_N - 2 < -h^{1/2}.$$

Como $2\sqrt{1+h}-2=O(h)$, se tiene que $|y_N-2|/h^p\to\infty$ cuando $p>1/2$ (nótese que $y(1)=\int_0^1 dx/\sqrt{1-x}=2$). Esto no contradice que el método de Euler sea de orden 1, ya que la función $f(x,y)=1/\sqrt{1-x}$ no está ni siquiera bien definida en un entorno de $x=1$, con lo cual no se tiene suficiente regularidad para que se pueda aplicar el teorema de convergencia.

Ⓛ **1.2.10.** (Omitido).

Ⓛ **1.2.11.** (Omitido).

1.3.1. Una aplicación directa de la definición lleva a

$$y_{n+1}=y_n+h\frac{(ay_n+b)+(ay_{n+1}+b)}{2}.$$

Al despejar se obtiene la fórmula de recurrencia

$$y_{n+1}=\frac{1+ah/2}{1-ah/2}y_n+\frac{bh}{1-ah/2},\qquad \text{con } y_0=0.$$

Iterando se sigue

$$y_n=\frac{bh}{1-ah/2}\left(1+(\frac{1+ah/2}{1-ah/2})^1+\cdots++(\frac{1+ah/2}{1-ah/2})^{n-1}\right).$$

Sumando la progresión geométrica

$$y_n=\frac{b}{a}\left((\frac{1+ah/2}{1-ah/2})^n-1\right).$$

1.3.2. Al aplicar la fórmula

$$y_{n+1}=y_n+h(y_{n+1}+y_n)\ \Rightarrow\ y_{n+1}=\frac{1+h}{1-h}y_n.$$

Cuando $y_0=0$ y $h=1/N$, se obtiene

$$y_N=(\frac{1+1/N}{1-1/N})^N,$$

y de la conocida fórmula $1/(1-x)=1+x+x^2+O(x^3)$ se deduce $\sqrt[N]{y_N}=1+2/N+2/N^2+O(N^{-3})$ simplemente por Taylor. Así pues

$$\sqrt[N]{y_N}=e^{2/N}(1+O(N^{-3}))$$

y la fórmula de la indicación implica $y_N = e^2(1+O(N^{-2}))$, o equivalentemente $|y(1) - y_N| = O(h^2)$, ya que $y(1) = e^2$ y $h = 1/N$.

La desigualdad de la indicación es consecuencia de que $e^t - t - 1$ tenga un mínimo en $t = 0$. De aquí $(1+t)^N - 1 \leq e^{Nt} - 1$ y al sustituir t por $O(N^{-3})$ se obtiene $(1 + O(N^{-3}))^N - 1 = O(N^{-2})$.

1.3.3. El problema $y' = f(x, y)$ se puede escribir en el intervalo $[x_n, x_{n+1}]$, integrando, como

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

Al aplicar la regla de cuadratura del trapecio se obtendría

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

y lo único que resta es escribir $y_n \approx y(x_n)$, $y_{n+1} \approx y(x_{n+1})$.

1.3.4. La regla de Simpson involucraría y_n e y_{n+1} para hallar y_{n+2}, ya que se basa en la fórmula de cuadratura

$$\int_{x_n}^{x_{n+2}} g \approx \frac{x_{n+2} - x_n}{6}\big(g(x_n) + 4g(x_{n+1}) + g(x_{n+2})\big).$$

No sería un método de un paso sino de dos.

*__1.3.5__ En cada paso

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

mientras que la solución real satisface

$$y(x_{n+1}) = y(x_n) + \int_{x_n}^{x_{n+1}} f(t)dt.$$

Restando ambas ecuaciones

$$y(x_{n+1}) - y_{n+1} = y(x_n) - y_n + R_n$$

con

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

Sumando todas estas igualdades para $n = 0, 1, \ldots, N-1$, se tiene que el error está acotado por $h \sup |R_n|$, siempre que $y_0 = y(x_0)$. Así que basta acotar R_n. Para ello considérese para cada n la función

$$g(t) = \int_{x_n}^{x_n+t} f - \frac{t}{2}\big(f(x_n) + f(x_n+t)\big).$$

Obviamente $R_n = g(h)$. Además es fácil comprobar que $g(0) = g'(0) = 0$ y $|g''(t)| \leq 0'5|t| \sup |f''|$. Por la fórmula de Taylor con error se tiene

$$|g(t)| \leq 0'25|t|^3 \sup |f''|,$$

de donde se deduce la estimación para R_n.

Ⓛ **1.3.6.** (Omitido).

Ⓛ **1.3.7.** (Omitido).

1.4.1. Derivando la ecuación

$$y'' = 2y + 2xy' = 2y + 4x^2 y \;\Rightarrow\; \frac{y''}{2!} = (1 + 2x^2)y.$$

De la misma forma

$$y''' = 2y' + 8xy + 4x^2 y' = 4xy + 8xy + 8x^3 y \;\Rightarrow\; \frac{y'''}{3!} = 2xy(3 + 2x^2).$$

Lo que da lugar a la iteración del enunciado.

1.4.2. Derivando dos veces la ecuación $y' = f(x, y)$, se obtiene

$$\begin{aligned} y'' =& f_x + f_y y' = f_x + f_y f \\ y''' =& f_{xx} + f_{xy} y' + (f_{xy} + f_{yy} y') f + f_y (f_x + f_y y') \\ =& f_{xx} + 2 f_{xy} f + f_{yy} f^2 + f_y f_x + f_y^2 f \end{aligned}$$

Llamando $d_{n,2}$ y $d_{n,3}$ a estas expresiones sustituyendo x e y por x_n e y_n, el método de Taylor es de la forma

$$y_{n+1} = y_n + h f(x_n, y_n) + d_{n,2} \frac{h^2}{2!} + d_{n,3} \frac{h^3}{3!}.$$

1.4.3. Sea el problema $y' = f(x,y)$ con $y = (y^1, y^2)^t$, $f = (f^1, f^2)^t$. Aplicando la regla de la cadena

$$\begin{pmatrix} y^1 \\ y^2 \end{pmatrix}'' = \begin{pmatrix} f^1_x & f^1_{y^1} & f^1_{y^2} \\ f^2_x & f^2_{y^1} & f^2_{y^2} \end{pmatrix} \begin{pmatrix} 1 \\ (y^1)' \\ (y^2)' \end{pmatrix} = \begin{pmatrix} f^1_x + f^1_{y^1} f^1 + f^1_{y^2} f^2 \\ f^2_x + f^2_{y^1} f^1 + f^2_{y^2} f^2 \end{pmatrix}$$

Sean $d^1_{n,2}$ y $d^2_{n,2}$ las coordenadas de este vector al sustituir x, y^1 e y^2 por x_n, y^1_n e y^2_n; entonces el método de Taylor es

$$\begin{pmatrix} y^1_{n+1} \\ y^2_{n+1} \end{pmatrix} = \begin{pmatrix} y^1_n \\ y^2_n \end{pmatrix} + \begin{pmatrix} f^1(x_n, y^1_n, y^2_n) \\ f^2(x_n, y^1_n, y^2_n) \end{pmatrix} h + \begin{pmatrix} d^1_{n,2} \\ d^2_{n,2} \end{pmatrix} \frac{h^2}{2}.$$

1.4.4. a) Derivando k veces, $y^{k+1)} = Ay^{k)}$ de donde inductivamente $y^{k)} = A^k y$. El método de Taylor será

$$y_{n+1} = y_n + hAy_n + \frac{h^2}{2!}A^2 y_n + \frac{h^3}{3!}A^3 y_n + \cdots + \frac{h^k}{k!}A^k y_n.$$

*b) Si $k = \infty$, se tiene

$$y_{n+1} = e^{hA} y_n = e^{A(x_1 - x_0)} y_n.$$

Pero la solución exacta de $y' = Ay$ es $y(x) = e^{A(x-x_0)} y(x_0)$, por tanto se cumple $y_1 = y(x_1)$.

1.4.5. (Hairer p. 50)

a) Como $p' > 0$, p es una función estrictamente creciente, por otro lado, $\int_0^{\pm\infty} e^{-t^2} dt = \pm\sqrt{\pi}/2$ implica que es sobreyectiva. Así pues tiene una función inversa $y = y(x)$. Por la regla de derivación de funciones inversas $y'(x) = 1/p(y(x))$, lo que da lugar a la ecuación diferencial del enunciado. Es evidente que $y(0) = 0$.

b) Derivando la ecuación

$$\begin{aligned} y'' &= \sqrt{\pi} e^{y^2} y' y = \frac{\pi}{2} y e^{2y^2} \\ y''' &= \frac{\pi}{2} y' e^{2y^2} + \frac{\pi}{2} y e^{2y^2} 4yy' = \frac{\pi^{3/2}}{4} e^{3y^2}(1 + 4y^2). \end{aligned}$$

Por consiguiente el método de Taylor buscado es

$$y_{n+1} = y_n + \frac{\sqrt{\pi}}{2} e^{y_n^2} h + \frac{\pi}{4} y_n e^{2y_n^2} h^2 + \frac{\pi^{3/2}}{24} e^{3y_n^2} (1 + 4y_n^2) h^3.$$

Ⓛ **1.4.6.** (Omitido).

1.5.1. Considérese, para simplificar, sólo el caso escalar. Nótese que

$$y' = f, \qquad y'' = f_x + f_y y' = f_x + f_y f.$$

Por tanto el polinomio de Taylor de grado 2 de $y(x+h) - y(x)$ en $h = 0$ es

$$y(x) + fh + (f_x + f_y f)\frac{h^2}{2}.$$

Mientras que el de grado 1 de $f(x+h, y(x+h))$ es

$$f + (f_x + f_y y')h = f + (f_x + f_y f)h.$$

De aquí, los coeficientes de Taylor de grados 1 y 2 en $h = 0$ del error de truncación son

$$c_1 = f - \alpha f - \beta f, \qquad c_2 = \frac{1}{2}(f_x + f_y f) - \alpha(f_x + f_y f).$$

Para que el orden de consistencia sea 1 basta con que $\alpha + \beta = 1$. Y para que sea 2 hay que imponer además $1/2 - \alpha = 0$, es decir, $\alpha = \beta = 1/2$.

1.5.2. a) Como en el problema anterior, nos limitaremos al caso escalar. El polinomio de Taylor de grado 1 de $f(x + \frac{h}{7}, y + \frac{h}{7} f(x,y))$ en $h = 0$ es

$$f + (\frac{1}{7} f_x + \frac{1}{7} f_y f)h.$$

Comparando con el desarrollo de Taylor de $y(x+h) - y(x)$ (véase el problema anterior), se tiene que los coeficientes de Taylor del error de truncación de grados 1 y 2 se anulan

$$c_1 = f + \frac{5}{2} f - \frac{7}{2} f = 0, \qquad c_2 = \frac{1}{2}(f_x + f_y f) - \frac{7}{2}(\frac{1}{7} f_x + \frac{1}{7} f_y f) = 0.$$

No es difícil ver, sin hacer los cálculos, que el de grado 3 no se anula, ya que en el coeficiente correspondiente de $y(x+h)-y(x)$ no aparecen potencias de 7 en el denominador y en el de $f(x+\frac{h}{7}, y+\frac{h}{7}f(x,y))$ sí aparecen. Por tanto el orden de consistencia es 2 y no mayor.

b) Como el método es obviamente consistente, basta notar que la composición de funciones Lipschitz es también Lipschitz.

1.5.3. a) Al desarrollar ϕ por Taylor en $h=0$ hasta grado 1 se tiene

$$f + (\frac{1}{2}f_x + \frac{1}{2}f_y f)h.$$

El polinomio de Taylor de grado 2 de $y(x+h)-y(x)$ es $fh+(f_x+f_y f)h^2/2$, con lo cual el orden de consistencia, y por tanto el de convergencia es 2.

b) Aplicándolo a $y'=x^2$, $y(0)=0$, se tiene

$$y_{n+1} = y_n + h(x_n + \frac{h}{2})^2, \qquad y_0 = 0.$$

Iterando con $x_n = nh$, $h = 1/N$,

$$\begin{aligned} y_N =& \big(\frac{1}{2}\big)^2 h^3 + \big(\frac{3}{2}\big)^2 h^3 + \cdots + \big(N-\frac{1}{2}\big)^2 h^3 \\ =& \frac{1}{4N^3}(1^2+3^2+\cdots+(2N-1)^2) \\ =& \frac{1}{4N^3}\cdot\frac{8N^3-2N}{6} = \frac{1}{3} - \frac{1}{12N^2} = y(1) - \frac{h^2}{12}. \end{aligned}$$

Con lo cual $\lim_{h\to 0} h^{-p}|y(1)-y_N| = \infty$ para $p>2$.

1.5.4. (Hairer p. 45)

a) La recurrencia es

$$y_{n+1}(x) = 1 + \int_0^x y_n(t)\, dt,$$

que a partir de $y_1 \equiv 1$ produce

$$y_2(x) = 1+x, \qquad y_3(x) = 1+x+\frac{x^2}{2}, \ldots \qquad y_n(x) = 1+x+\frac{x^2}{2}+\cdots+\frac{x^{n-1}}{(n-1)!},$$

es decir, los polinomios de Taylor de la exponencial en $x=0$.

*b) La fórmula del método de Picard implica

$$\begin{aligned}|y_{n+1}(x) - y_n(x)| \leq & \int_a^x |f(t, y_n(t)) - f(t, y_{n-1}(t))|\, dt \\ \leq & L \int_a^x |y_n(t) - y_{n-1}(t)|\, dt\end{aligned}$$

donde L es la constante Lipschitz en la segunda variable. Partiendo de $y_1 \equiv \eta$, se tiene $y_2(x) = \eta + (x-a)K$, donde K es una constante. La acotación anterior implica, iterando,

$$|y_{n+1}(x) - y_n(x)| \leq L^{n-1} K \frac{|x-a|^n}{n!}.$$

En particular la serie

$$y_1(x) + \sum_{n=1}^{\infty} (y_{n+1}(x) - y_n(x))$$

converge uniformemente (test M de Weierstrass) a una función $y(x)$. Las sumas parciales de esta serie son $y_k(x)$, con lo cual queda probado que $y_k(x) \to y(x)$ uniformemente. Esto permite tomar límites en la relación

$$y_{n+1}(x) = y(a) + \int_a^x f(t, y_n(t))\, dt,$$

siempre que f sea continua. De donde y es una solución de $y' = f(x, y)$ con $y(a) = \eta$ porque $y_k(a) = \eta$.

Es decir, se ha probado que si f es continua y Lipschitz en la segunda variable se tiene existencia de solución en el problema $y' = f(x, y)$, $y(a) = \eta$.

La unicidad se puede obtener a partir de una acotación como la inicial, porque si hubiera dos soluciones distintas: y, $\widetilde{y}$, se cumpliría

$$\sup_{t \in [a,x]} |y(t) - \widetilde{y}(t)| \leq L|x-a| \sup_{t \in [a,x]} |y(t) - \widetilde{y}(t)|,$$

lo cual es una contradicción para $L|x-a| < 1$, es decir, en un pequeño entorno de a. Con lo cual, localmente, hay solución única.

Ⓛ **1.5.5.** (Omitido).

Ⓛ **1.5.6.** (Omitido).

1.6.1. Con lo visto hasta ahora, el método de Euler modificado es "mejor" que la regla del trapecio porque es del mismo orden y es explícito (y por tanto computacionalmente más sencillo).

Escribiendo

$$k_1 = f(x_n, y_n), \qquad k_2 = f(x_n+h, y_n+\frac{h}{2}k_1+\frac{h}{2}k_2), \qquad y_{n+1} = y_n+\frac{h}{2}(k_1+k_2),$$

se tiene $k_2 = 2(y_{n+1} - y_n) - hf(x_n, y_n)$, que sustituyendo en la fórmula para k_2 implica

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

lo que después de operar da lugar a la regla del trapecio.

1.6.2. El método en términos de los k_i es

$$\begin{aligned} k_1 &= f(x_n, y_n) \\ k_2 &= f(x_n + \frac{h}{3}, y_n + \frac{h}{3}k_1) \\ k_3 &= f(x_n + \frac{2h}{3}, y_n + \frac{2h}{3}k_2) \\ y_{n+1} &= y_n + \frac{h}{4}(k_1 + 3k_3) \end{aligned}$$

Por tanto el tablero es

0			
1/3	1/3		
2/3	0	2/3	
	1/4	0	3/4

1.6.3. Al aplicarlo con $f(x, y) = y$ se obtiene

$$\begin{aligned} y_{n+1} &= y_n + \frac{h}{4}\big(y_n + 3\big(y_n + \frac{2h}{3}(y_n + \frac{h}{3}y_n)\big)\big) \\ &= \big(1 + h + \frac{h^2}{2} + \frac{h^3}{6}\big)y_n \end{aligned}$$

Por tanto

$$\begin{aligned} y_N =& \big(1 + h + \frac{h^2}{2} + \frac{h^3}{6}\big)^N = \big(e^h + O(h^4)\big)^N = e^{nh}\big(1 + O(h^4)\big)^N \\ =& e^{x_N}\big(1 + O(Nh^4)\big) = e^{x_N}\big(1 + O(h^3)\big). \end{aligned}$$

Así pues $|y_N - y(x_N)| = O(h^3)$.

1.6.4. a) Evidentemente

$$\|F(\vec{k}') - F(\vec{k})\| \leq s \max_i \|F^i(\vec{k}') - F^i(\vec{k})\|.$$

Por tanto existe cierto i tal que, por ser f Lipschitz,

$$\|F(\vec{k}') - F(\vec{k})\| \leq sL\|h\sum_{j=1}^{s} a_{ij}(k'_j - k_j)\|.$$

Nótese que $k'_j, k_j \in \mathbb{R}^d$, son vectores. Aplicando la desigualdad de Cauchy-Schwarz a cada una de las coordenadas de $\sum_{j=1}^{s} a_{ij}(k'_j - k_j)$, se deduce

$$\begin{aligned} \|\sum_{j=1}^{s} a_{ij}(k'_j - k_j)\| &\leq \big(\sum_{j=1}^{s} a_{ij}\big)\sum_j \|k'_j - k_j\|^2 \\ &\leq s \max |a_{ij}|^2 \|\vec{k}' - \vec{k}\|^2. \end{aligned}$$

Extrayendo raíces cuadradas y sustituyendo, se llega a mejorar la constante Lipschitz del enunciado en un factor $s^{1/2}$.

b) Según el primer apartado, para $h < s^{-2}L^{-1}/\max|a_{ij}|$ la función tiene constante de Lipschitz menor que uno, y por el teorema de la aplicación contractiva debe existir un único punto fijo.

1.6.5. Nótese que para hallar k_1 hay que resolver una ecuación del tipo $k_1 = F_1(k_1)$. Una vez resuelta, se puede hallar k_2 resolviendo otra ecuación del mismo tipo (porque k_1 es ya conocido), etc. Es decir, en cada caso la ecuación (en general no lineal) que hay que resolver tiene dimensión d igual al número de coordenadas de los k_i. Sin embargo en los métodos implícitos generales hay que resolver un sistema de dimensión sd, lo que computacionalmente es más costoso.

Ⓛ **1.6.6.** (Omitido).

1.7.1. Las condiciones de orden 1 y 2 se escriben respectivamente como

$$\vec{b}\cdot\vec{1}=1 \qquad \text{y} \qquad 2\vec{b}^t A\vec{1}=1$$

(nótese que $A\vec{1}$ es el vector cuya coordenada $j-$ésima es $\sum_k a_{jk}$). Por otro lado

$$3\sum_{j,k,l} b_j a_{jk} a_{jl}=1 \;\Leftrightarrow\; 3\sum_{j,k,l}\widetilde{a}_{ki} b_{ij} a_{jl}=1,$$

donde $\widetilde{a}_{ki}=a_{ki}$ son los elementos de A^t. entonces las condiciones de orden 3 son, con esta notación,

$$3\,\vec{1}^t A^t B A\vec{1}=1 \qquad \text{y} \qquad 6\,\vec{b}^t A^2\vec{1}=1.$$

1.7.2. a) El tablero es

$1/2$		
$1/2$	1	
	$1/2$	$1/2$

Y está a la vista que c_i no es la suma de la fila $i-$ésima, $i=1,2$. El orden se debe calcular con la definición porque se probaron las condiciones de orden cuando se satisfacía la condición de suma por filas. Desarrollando por Taylor en $h=0$

$$k_1^J=f^J+\frac{1}{2}f_x^J h+O(h^2), \qquad k_2^J=f^J+(\frac{1}{2}f_x^J+\sum_K f_K^J f^K)h+O(h^2)$$

donde f_K indica la derivada de f con respecto a la $K-$ésima componente de y. Por otra parte

$$y^J(x_n+h)-y^J(x_n)=f^J h+(f_x^J+\sum_K f_K^J f^K)\frac{h^2}{2}+O(h^3).$$

Lo cual coincide hasta orden 2 con $\frac{h}{2}(k_1^J+k_2^J)$. Así pues $R_n=O(h^3)$ y el orden es 2.

b) Un cálculo similar al anterior prueba en este caso

$$k_1^J=f^J+c_1 f_x^J h+O(h^2), \qquad k_2^J=f^J+(c_2 f_x^J+(2\beta)^{-1}\sum_K f_K^J f^K)h+O(h^2).$$

Para que $h(b_1 k_1^J + b_2 k_2^J)$ coincida hasta orden 2 con $y^J(x_n+h) - y^J(x_n)$ debe cumplirse, por tanto,

$$b_1 + b_2 = 1, \qquad b_1 c_1 + b_2 c_2 = \frac{1}{2}, \qquad (2\beta)^{-1} b_2 = \frac{1}{2}.$$

De donde se deduce

$$b_1 = 1 - \beta, \qquad b_2 = \beta, c_2 = (\frac{1}{2} - (1-\beta)c_1)/\beta$$

con c_1 arbitrario cumpliendo $c_1 \neq 0$ para que no se verifique la condición de suma por filas.

c) Porque tienen el mismo orden y son igual de simples computacionalmente.

1.7.3. a) De la definición $k_i = f(x_n + c_i h, y_n + h\sum_j a_{ij} k_j)$, se deduce, entendiendo k_i como función de h

$$k_i(0) = f, \qquad k_i'(0) = f_x c_i + \sum_K f_K \sum_j a_{ij} f^K.$$

Obviamente $k_i(0) = y'(x_n)$, y $\sum_j a_{ij} = c_i$ si y sólo si $k_i'(0) = y''(x_n)$.

b) La función se evalúa en puntos de la forma $(x_n + c_i h, y_n + h\sum_j a_{ij} k_j)$. La segunda coordenada se aproxima hasta orden uno (bajo la hipótesis de localización) por $y(x_n) + h\sum_j a_{ij} f$, mientras que la función evaluada en la primera coordenada, $y(x_n + c_i h)$, se aproxima por $y(x_n) + y'(x_n) c_i h = y(x_n) + h c_i f$.

1.7.4. (Hairer p. 142) La condición de orden 1 se cumple trivialmente, mientras que la de orden 2 implica $c_3 = 1/2$. Ninguno de estos métodos tiene orden 3 ya que $3\sum b_j c_j^2 \neq 1$.

1.7.5. En el primer caso, los tableros son de la forma

c_1		
c_2	α	
	b_1	b_2

La condición de orden 1 implica $b_1 + b_2 = 1$ y la de orden 2, $2\alpha b_2 = 1$. Con lo cual todos los métodos de este tipo de órdenes 1 y 2 responden a los tableros

$$\begin{array}{c|cc} 0 & & \\ \alpha & \alpha & \\ \hline & b_1 & 1-b_1 \end{array} \qquad \begin{array}{c|cc} 0 & & \\ \alpha & \alpha & \\ \hline & 1-\frac{1}{2\alpha} & \frac{1}{2\alpha} \end{array}$$

Para un método implícito de una etapa,

$$\begin{array}{c|c} c_1 & c_1 \\ \hline & b_1 \end{array}$$

la condición de orden 1 implica $b_1 = 1$ y la de orden 2 implica $c_1 = 1/2$.

1.7.6. El método es de orden al menos 2 porque

$$\sum b_j = \frac{1}{2} + \frac{1}{2} = 1 \qquad \text{y} \qquad 2\sum b_j c_j = 2\big(\frac{1}{2}\cdot\frac{1}{4} + \frac{1}{2}\cdot\frac{3}{4}\big) = 1.$$

Y no es de orden 3 porque

$$3\sum b_j c_j^2 = 3\big(\frac{1}{2}\cdot\frac{1}{4^2} + \frac{1}{2}\cdot\frac{3^2}{4^2}\big) \neq 1.$$

1.7.7. Las condiciones de orden 1 y 2 se reducen a un cálculo sencillo

$$\sum b_j = \frac{1}{2} + \frac{1}{2} = 1 \qquad \text{y} \qquad 2\sum b_j c_j = 2\big(\frac{3+\sqrt{3}}{12} + \frac{3+\sqrt{3}}{12}\big) = 1.$$

La condiciones de orden 3 llevan a cálculos un poco más largos. Se verificará la primera sin cambios, $3\sum b_j c_j^2 = 1$, y la segunda en forma matricial, $6\vec{b}^t A\vec{c} = 6\sum b_j a_{jk} c_k = 1$.

$$\begin{aligned} 3\sum b_j c_j^2 =& \frac{3}{2}\big(\big(\frac{3+\sqrt{3}}{6}\big)^2 + \big(\frac{3-\sqrt{3}}{6}\big)^2\big) = 1. \\ 6\,\vec{b}^t A\vec{c} =& 6\begin{pmatrix}\frac{1}{2} & \frac{1}{2}\end{pmatrix}\begin{pmatrix}(3+\sqrt{3})/6 & 0 \\ -\sqrt{3}/3 & (3+\sqrt{3})/6\end{pmatrix}\begin{pmatrix}(3+\sqrt{3})/6 \\ (3-\sqrt{3})/6\end{pmatrix} = 1. \end{aligned}$$

1.7.8. De $y' = f(y)$ se deduce $y'' = f'f$, $y''' = f''f^2 + (f')^2 f$, y por tanto

$$y(x_n + h) - y(x_n) = fh + f'f\frac{h^2}{2} + (f''f^2 + (f')^2 f)\frac{h^3}{6} + O(h^4).$$

Escribiendo $k_i = f(g_i)$ con $g_i = y(x_n) + h\sum_j a_{ij} f(g_j)$ se tiene que el error de truncación es

$$R_n = y(x_n + h) - y(x_n) - h\sum_j b_j f(g_j).$$

Así que basta con calcular el desarrollo de Taylor de orden 2 de $\sum b_j f(g_j)$, sustituir en esta fórmula y comparar los coeficientes de Taylor con los de $y(x_n + h) - y(x_n)$.

Por definición $g_i(0) = y(x_n)$ (se entiende g_i como función de h) y derivando

$$g_i' = \sum_j a_{ij} f(g_j) + h\sum_j a_{ij} f'(g_j) g_j'$$
$$g_i'' = 2\sum_j a_{ij} f'(g_j) g_j' + h\sum_j a_{ij} (f'(g_j) g_j')'$$

que implican en $h = 0$, $g_i'(0) = \sum_j a_{ij} f$, $g_i''(0) = 2\sum_{j,k} a_{ij} a_{jk} f' f$. Donde se supone que f y f' están sustituidas en $(x_n, y(x_n))$. Entonces los coeficientes de Taylor de $\sum b_j f(g_j)$ de órdenes 0, 1 y 2 son

$$\begin{aligned} c_0 &= \sum b_j f \\ c_1 &= (\sum b_j f'(g_j) g_j')(0) = \sum_{j,k} a_{jk} f' f \\ c_2 &= \frac{1}{2}\sum b_j (f''(g_j)(g_j')^2 + f'(g_j) g_j'')(0) \\ &= \frac{1}{2}\sum b_j \big((\sum_k a_{jk} f)^2 f'' + 2 f' \sum_{k,l} a_{jk} a_{kl} f' f\big) \end{aligned}$$

Sustituyendo en la fórmula para el error de truncación, los términos de orden 1, 2 y 3 se anularán si se cumplen las respectivas condiciones de orden.

1.7.9. El tablero será de la forma

c_1	α	$c_1 - \alpha$
c_1	β	$c_1 - \beta$
	b_1	b_1

Las condiciones de orden 1 y 2 implican $2b_1 = 1$, $4b_1 c_1 = 1$; de donde $b_1 = c_1 = 1/2$. Ninguno de estos métodos tiene orden 3 porque $3(b_1 c_1^2 + b_2 c_2^2) = 3/4 \neq 1$.

1.7.10. El tablero será del tipo

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

Las condiciones de orden 1, 2 y 3 dan lugar a las ecuaciones

$$b_1 + 2b_2 = 1, \qquad 4b_2c_2 = 1, \qquad 6b_2c_2^2 = 1, \qquad 6b_2a_{32}c_2 = 1.$$

Despejando b_2 en todas las ecuaciones e igualando, se llega a

$$c_2 = \frac{2}{3}, \qquad b_2 = \frac{3}{8}, \qquad b_1 = \frac{1}{4}, \qquad a_{32} = \frac{2}{3}.$$

Para que se cumpla la condición de suma por filas se debe tener además $a_{31} = 0$.

1.7.11. Definiendo $k_1 = f\big(x_n + \frac{h}{2}, \frac{y_n+y_{n+1}}{2}\big)$ se puede escribir $y_{n+1} - y_n = hk_1$ y

$$k_1 = f\big(x_n + \frac{h}{2}, y_n + \frac{1}{2}(y_{n+1} - y_n)\big) = f\big(x_n + \frac{h}{2}, y_n + \frac{h}{2}k_1\big).$$

Por consiguiente es un método de una etapa con $c_1 = 1/2$, $a_{11} = 1/2$ y $b_1 = 1$.

Ⓛ **1.7.12.** (Omitido).

1.8.1. Las condiciones de orden 1, 2 y 3 se verifican con sencillos cálculos

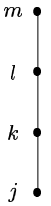
$$\begin{aligned} \text{orden } 1 \;&\longrightarrow\; \frac{2}{9} + \frac{1}{3} + \frac{4}{9} = 1 \\ \text{orden } 2 \;&\longrightarrow\; 2\big(\frac{1}{3}\cdot\frac{1}{2} + \frac{4}{9}\cdot\frac{3}{4}\big) = 1 \\ \text{orden } 3 \;&\longrightarrow\; 3\big(\frac{1}{3}\cdot\frac{1}{2^2} + \frac{4}{9}\cdot\frac{3^2}{4^2}\big) = 1 \\ &\qquad\quad 6\big(\frac{1}{3}\cdot\frac{1}{2}\cdot 0 + \frac{4}{9}\cdot\frac{3}{4}\cdot\frac{1}{2}\big) = 1 \end{aligned}$$

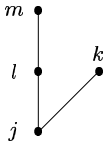
La condición de orden 4 asociada al árbol lineal es

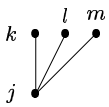
$$24 \sum_{j,k,l,m} b_j a_{jk} a_{kl} a_{lm} = 1.$$

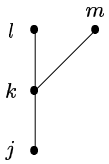
Como el método es explícito, necesariamente los sumandos no nulos corresponden a $j > k > l > m$. Pero es imposible elegir cuatro enteros entre 1 y $s = 3$, por lo que el primer miembro se anula.

1.8.2. Por el mismo argumento del problema anterior, el orden no puede ser 5. Por otra parte, es fácil comprobar que se verifican las condiciones de orden 1, 2 y 3. Hay que centrarse, por consiguiente, en comprobar las de orden 4. Habrá una condición asociada a cada uno de los árboles de orden 4.

$A =$ 

$B =$ 

$C =$ 

$D =$ 

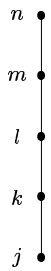
Utilizando la definición, se tiene $\gamma(A) = 24$, $\gamma(B) = 8$, $\gamma(C) = 4$, $\gamma(D) = 12$, y

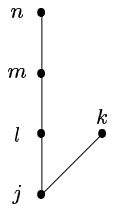
$$\phi_j(A) = \sum_{k,l,m} a_{jk} a_{kl} a_{lm} = \sum_{k,l} a_{jk} a_{kl} c_l \qquad \phi_j(B) = \sum_{k,l,m} a_{jk} a_{jl} a_{lm} = \sum_{l} c_j a_{jl} c_l$$

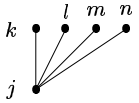
$$\phi_j(C) = \sum_{k,l,m} a_{jm} a_{jl} a_{jk} = c_j^3 \qquad \phi_j(D) = \sum_{k,l,m} a_{jk} a_{km} a_{kl} = \sum_{k} a_{jk} c_k^2$$

En cada caso, unos sencillos cálculos prueban que $\gamma(t) \sum b_j \phi_j(t) = 1$, y el método es de orden 4.

1.8.3. Eligiendo los árboles

$A =$ 

$B =$ 

$C =$ 

se tienen las correspondientes condiciones de orden

$$\begin{aligned} A &\longrightarrow 120\sum b_j a_{jk}a_{kl}a_{lm}a_{mn} = 1 \\ B &\longrightarrow 30\sum b_j a_{jk}a_{jl}a_{lm}a_{mn} = 1 \\ C &\longrightarrow 5\sum b_j a_{jk}a_{jl}a_{jm}a_{jn} = 1 \end{aligned}$$

Como el método es explícito, la única contribución no nula a la suma de la condición asociada a A corresponde a $j=5$, $k=4$, $l=3$, $m=2$, $n=1$. Sustituyendo, se cumple la igualdad. Utilizando la condición de suma por filas, comprobar la condición de orden para B se reduce a calcular

$$30\sum b_j c_j a_{jl}a_{lm}c_m = 30(b_4c_4a_{43}a_{32}c_2 + b_5c_5(a_{54}a_{43}c_3 + a_{53}a_{32}c_2)).$$

Como es diferente de 1, no se cumple. Tampoco se cumple para C porque

$$5\sum b_j c_j^4 = 5\big(\frac{3}{10}\cdot\frac{1}{3^4} + \frac{2}{5}\cdot\frac{1}{2^4} + \frac{1}{5}\big) \neq 1.$$

1.8.4. Las condiciones de orden 1, 2 y 3 llevan al sistema no lineal

$$b_1 + b_2 = \frac{1}{2}, \quad 2b_2c_2 + b_1c_4 = \frac{1}{2}, \quad 2b_2c_2^2 + b_1c_4^2 = \frac{1}{3}, \quad b_2c_2^2 + b_1c_4c_2 = \frac{1}{6}.$$

Multiplicando la última ecuación por 2 y restándole la anterior se llega a $b_1c_4(c_4-2c_2)=0$. Las soluciones $b_1=0$, $c_4=0$ son claramente incompatibles con la condición de orden 4 correspondiente al árbol lineal $24b_1c_4c_2^2=1$. Así pues la única posibilidad es $c_4=2c_2$ que sustituida en la segunda ecuación y dividiendo entre la primea, permite deducir $c_2=1/2$, y de aquí $c_4=1$. Utilizando de nuevo la primera y la tercera se obtiene finalmente $b_1=1/6$, $b_2=1/3$. Por la particular forma del método (con muchos ceros en su tablero) la verificación de que el orden es realmente 4 se reduce a los sencillos cálculos

$$24b_1c_4c_2^2 = 8(b_2c_2^3 + b_1c_4^2c_2) = 4(b_2c_2^3 + b_2c_2^3 + b_1c_4^3) = 12(b_2c_2^3 + b_1c_4c_2^3) = 1.$$

1.8.5. De nuevo sse supondrán comprobadas las condiciones hasta orden 3. Según se había visto en un problema anterior, las condiciones de orden 4 son

$$24\sum b_j a_{jk}a_{kl}c_l = 1,\ 8\sum b_j c_j a_{jl}c_l = 1,\ 4\sum b_j c_j^3 = 1,\ 12\sum b_j a_{jk}c_k^2 = 1.$$

Despreciando los términos nulos, para este método basta verificar

$$24b_4a_{43}a_{32}c_2 = 1, \qquad 8(b_3c_3a_{32}c_2 + b_4c_4(a_{43}c_3 + a_{42}c_2)) = 1,$$
$$4(b_2c_2^3 + b_3c_3^3 + b_4c_4^3) = 1, \qquad 12(b_3a_{32}c_2^2 + b_4(a_{43}c_3^2 + a_{42}c_2^2)) = 1.$$

Al sustituir los datos del tablero, todas estas igualdades se cumplen.

1.8.6. La desventaja obvia es que es implícito. El tablero con la notación de la indicación es

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

Las condiciones de orden 1 y 2 se verifican trivialmente. También se cumplen las de orden 3:

$$3\sum b_jc_j^2 = 3\cdot\frac{1}{2}(\alpha^2 + (1-\alpha)^2) = \frac{3}{2}(1 - 2\alpha(1-\alpha)) = 1$$
$$6\sum b_ja_{jk}c_k = 3\sum a_{jk}c_k = 3((1-\alpha)c_1 + \alpha c_2) = 6\alpha(1-\alpha) = 1.$$

Las de orden 4 llevan a cálculos más extensos. La correspondiente al árbol lineal, $24\sum b_ja_{jk}a_{kl}c_l = 1$, puede escribirse matricialmente como $24\vec{b}^tAA\vec{c} = 1$, dando lugar a

$$\begin{pmatrix} 12 & 12 \end{pmatrix}\begin{pmatrix} 1/4 & \alpha - 1/4 \\ 3/4-\alpha & 1/4 \end{pmatrix}\begin{pmatrix} 1/4 & \alpha - 1/4 \\ 3/4-\alpha & 1/4 \end{pmatrix}\begin{pmatrix} \alpha \\ 1-\alpha \end{pmatrix}$$
$$= \begin{pmatrix} 3-3\alpha & 3\alpha \end{pmatrix}\begin{pmatrix} 6\alpha - 4\alpha^2 - 1 \\ 2\alpha - 4\alpha^2 + 1 \end{pmatrix}.$$

Utilizando adecuadamente la relación $\alpha(1-\alpha) = 1/6$, el último vector columna es $\frac{1}{3}\begin{pmatrix} 6\alpha - 1 & 5 - 6\alpha \end{pmatrix}^t$, que al multiplicar y al emplear de nuevo la relación anterior, da el resultado esperado.

La segunda condición de orden, $8\sum b_jc_ja_{jl}c_l = 1$, en este caso se puede escribir como $4\vec{c}^{\,t}A\vec{c} = 1$ que se verifica con cálculos como los anteriores

$$\begin{pmatrix} \alpha & 1-\alpha \end{pmatrix}\begin{pmatrix} 1 & 4\alpha - 1 \\ 3-4\alpha & 1 \end{pmatrix}\begin{pmatrix} \alpha \\ 1-\alpha \end{pmatrix} = \begin{pmatrix} \alpha & 1-\alpha \end{pmatrix}\begin{pmatrix} 2\alpha - 1/3 \\ 5/3 - 2\alpha \end{pmatrix} = 1.$$

La tercera condición de orden, $4\sum b_jc_j^3 = 1$, equivale a $2(\alpha^3 + \overline{\alpha}^3) = 1$ donde $\overline{\alpha}$ es el conjugado real de α. Una manera de comprobar esta igualdad

sin hacer cálculos es pensar en el resultado de desarrollar los cubos con el binomio de Newton, deduciendo que $2(\alpha^3+\overline{\alpha}^3)$ es necesariamente un número entero, como $\overline{\alpha} < 0'8$ y $\alpha < 0'3$, este entero debe ser 1.

Finalmente, la última condición de orden 4 lleva a los cálculos

$$\begin{aligned} 12\sum b_j a_{jk} c_k^2 =& 12(b_1 a_{11} + b_2 a_{21})c_1^2 + 12(b_1 a_{12} + b_2 a_{22})c_2^2 \\ =& 6(1-\alpha)\alpha^2 + 6\alpha(1-\alpha)^2 = 6\alpha(1-\alpha) = 1 \end{aligned}$$

y por tanto también se satisface.

1.8.7. (Hairer p. 155) Aunque hay condiciones de orden inferior que no se satisfacen, es sencillo considerar la condición de orden 5 correspondiente al árbol con $\gamma = 5$.

$$5\sum b_j c_j^4 = 5\Big(\frac{125}{192}\cdot\frac{2^4}{5^4} - \frac{81}{192}\cdot\frac{2^4}{3^4}\frac{100}{192}\cdot\frac{4^4}{5^4}\Big) \neq 1.$$

1.8.8. La condición de orden 3, $3\sum b_j c_j^2 = 1$ da lugar a la ecuación $\beta^2/4 + (3-\beta)^2 = 2$ cuyas raíces son $\beta = 2$ y $\beta = 14/5$. Esta última raíz implicaría $c_3 = 7/10$, $c_4 = 1/5$, lo que contradiría la condición de orden 4, $4\sum b_j c_j^3 = 1$; por tanto $\beta = 2$, $c_3 = 1/2$, $c_4 = 1$. Utilizando otra de las condiciones de orden 4

$$8\sum b_j c_j a_{jl} c_l = 8\cdot\frac{4}{6}\cdot\frac{1}{2}\cdot\frac{1}{2}\cdot\alpha + 8\cdot\frac{1}{6}\cdot 1\cdot\big((2\alpha)^{-1}\cdot\frac{1}{2} - (2\alpha)^{-1}\cdot\alpha\big) = 1.$$

De donde $\alpha = 1$ o $\alpha = 1/4$. Al sustituir $\alpha = 1$ el tablero resultante no verifica la condición de orden 4, $12\sum b_j a_{jk} c_k^2 = 1$, por tanto $\alpha = 1/4$.

1.8.9. Todo lo que hay que hacer son las siguientes comprobaciones

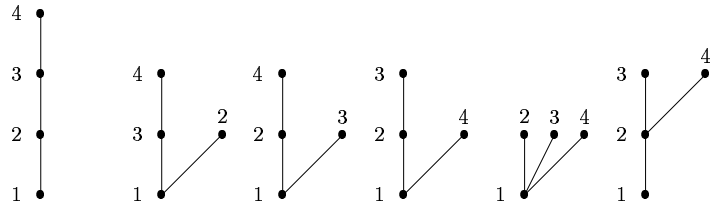
$$\begin{aligned} 24\sum b_j a_{jk} a_{kl} c_l &= 24(b_2 a_{22} a_{22} c_2 + b_3 a_{32} a_{22} c_2) = 1 \\ 8\sum b_j c_j a_{jl} c_l &= 8(b_2 c_2 a_{22} c_2 + b_3 c_3 a_{32} c_2) = 1 \\ 4\sum b_j c_j^3 &= 4(b_2 c_2^3 + b_3 c_3^3) = 1 \\ 12\sum b_j a_{jk} c_k^2 &= 12(b_2 a_{22} c_2^2 + b_3 a_{32} c_2^2) = 1. \end{aligned}$$

Ⓛ **1.8.10.** (Omitido).

Ⓛ **1.8.11.** (Omitido).

1.9.1. (Hairer p. 155) Necesariamente $t(2) = 1$, $t(3) \in \{1, 2\}$ y $t(r) \in \{1, 2, 3, \ldots, r-1\}$, en general. Si para cada $t(r)$ hay $r-1$ posibilidades, Card $LT_q = \prod_{r=2}^{q} = (q-1)!$, ya que cada árbol etiquetado está determinado por los valores de la función $t(r)$.

1.9.2. Hay que representar árboles etiquetados con 1,2, 3 y 4 de manera que el orden de las etiquetas respete las direcciones ascendente de las ramas. Las posibilidades son



Obviamente el segundo, tercero y cuarto representan un mismo árbol.

1.9.3. Como Card $LT_q = (q-1)!$, hay a lo más $(q-1)!$ condiciones de orden q. Para que un método sea de orden 8 debe satisfacerlas todas con $1 \leq q \leq 8$. Por tanto

$$\text{n}^{\text{o}} \text{ de cond.} \leq 0! + 1! + 2! + \cdots + 7! = 5913.$$

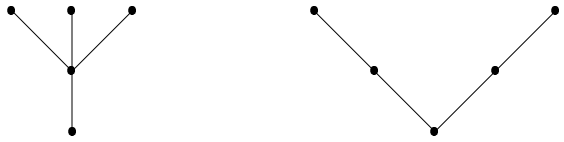
1.9.4. Evidentemente el único árbol de orden q con $\gamma(t) = q$ es el que tiene todos sus vértices unidos directamente con la raíz. La condición de orden será

$$q \sum b_{j1} a_{j_1 j_2} a_{j_1 j_3} \ldots a_{j_1 j_q} = q \sum b_j c_j^{q-1} = 1,$$

y el diferencial elemental tiene la J_1-ésima coordenada

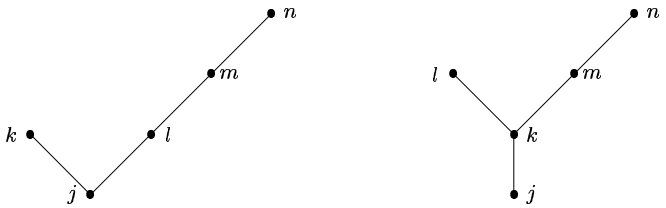
$$F^{J_1}(t) = \sum_{J_2, \ldots, J_q} f^{J_1}_{J_2 \ldots J_q} f^{J_2} f^{J_3} \ldots f^{J_q}.$$

1.9.5. No, un contraejemplo para $q = 5$ lo conforman los árboles:



ya que para ambos $\gamma(t) = 20$.

1.9.6. (Hairer p. 154)
a) Los siguientes árboles cumplen $\gamma(A) = 30$, $\gamma(B) = 40$.



Los diferenciales elementales son

$$F^J(A) = \sum_{K,L,M,N} f^J_{KL} f^L_M f^M_N f^K f^N = \sum_{K,N} f^J_{K2} f^2_2 f^2_N f^K f^N$$
$$F^J(A) = \sum_{K,L,M,N} f^J_K f^K_{LM} f^M_N f^L f^N = \sum_{L,N} f^J_2 f^2_{L2} f^2_N f^L f^N$$

Donde se ha usado para simplifcar que $f^1 = 1$ y por tanto sus derivadas se anulan. Por esta misma razón $F^1(A) = F^1(B) = 0$. Desarrollando las sumas y sustituyendo $f^1 = 1$, $f^2 = g$

$$F^2(A) = F^2(B) = g_1 g_{12}(g_1 + g_2 g) + g_2 g_{22} g(g_1 + g_2 g).$$

b) Supuestas las condiciones hasta orden 4, el método será de orden 5 si y sólo si

$$\sum_{t \in LT_5} \big(1 - \gamma(t) \sum_j b_j \phi_j(t)\big) F^J(t) = 0.$$

Si la condición de orden 5 se cumple para todos los árboles diferentes de A y B (y de sus equivalentes), como $F(A) = F(B)$, el método será de orden 5

siempre que

$$\big(1 - 30\sum_j b_j\phi_j(A)\big) + \big(1 - 40\sum_j b_j\phi_j(B)\big) = 0.$$

Y esta igualdad puede darse cuando $30\sum_j b_j\phi_j(A) \neq 1$ y $40\sum_j b_j\phi_j(B) \neq 1$. Es decir, las condiciones de orden son suficientes pero no son necesarias.

c) El teorema es para problemas generales (en la definición inicial de orden de consistencia se pedían ciertas propiedades para todo problema suficientemente regular); sin embargo pueden existir casos particulares con características especiales.

***1.9.7** La fórmula de Faà di Bruno venía exactamente de expresar la derivada de una composición con el lenguaje de los árboles, por tanto

$$(f \circ g)^{(n)} = \sum_{\substack{u \in LS_{n+1} \\ u=[u_1,u_2,\ldots,u_r]}} (f^{(r)} \circ g) g^{(s_1)} g^{(s_2)} \ldots g^{(s_r)}$$

donde s_i es el orden de u_i.

Para $n = 1$ la única posibilidad es que $r = 1$ y u_1 sea el árbol trivial (de un solo vértice), lo que corresponde a

$$(f \circ g)' = (f' \circ g) g'.$$

Para $n = 2$, si $r = 1$, u_1 es el árbol de una sola arista; y si $r = 2$, u_1 y u_2 son ambos el árbol trivial. La fórmula anterior implica entonces

$$(f \circ g)'' = (f' \circ g) g'' + (f'' \circ g) g' g'.$$

Para $n = 3$, si $r = 1$, u_1 es el árbol lineal de dos aristas; si $r = 2$, u_1 y u_2 deben ser uno el árbol trivial y otro el árbol de una arista (nótese que hay tres formas de etiquetar el árbol resultante); y si $r = 3$, u_1 u_2 y u_3 son árboles triviales. Esto corrobora la fórmula

$$(f \circ g)''' = (f' \circ g) g''' + 3(f'' \circ g) g' g'' + (f''' \circ g) g' g' g'.$$

1.9.8. (Hairer p. 154)

a) Al tomar límites $x \to 0$ se obtiene inmediatamente $A_1 = 1$. Con $k = 1$, $1 + A_2 x = (1-x)^{-1} + O(x^2)$ implica $A_2 = 1$. Para $k = 2$

$$1 + x + A_3 x^2 = (1-x)^{-1}(1-x^2)^{-1} = (1+x+x^2)(1+x^2) + O(x^3) \;\Rightarrow\; A_3 = 2.$$

Así se procede inductivamente en general, desarrollando $(1-x)^{-A_1}(1-x^2)^{-A_2}\cdots(1-x^k)^{-A_k}$ y buscando el coeficiente de grado k que será A_{k+1}. El cálculo más largo corresponde naturalmente a A_6 y está basado en el desarrollo

$$(1-x)^{-1}(1-x^2)^{-1}(1-x^3)^{-2}(1-x^4)^{-4}(1-x^5)^{-9} = (1+x+x^2+x^3+x^4 + x^5)\cdot(1+x+x^4)(1+2x^3)(1+4x^4)(1+9x^5) + O(x^6).$$

El coeficientes de x^5 en este producto se localiza fácilmente utilizando que las únicas formas de escribir 5 como suma de los exponentes que aparecen en estos paréntesis son : $0+0+0+0+5$, $1+0+0+4+0$, $2+0+3+0+0$, $0+2+3+0+0$, $1+4+0+0+0$, $3+2+0+0+0$ y $5+0+0+0+0$. Localizando los correspondientes productos de coeficientes se obtiene $A_5 = 9+4+2+2+1+1+1 = 20$.

b) Para que sea de orden 8 debe satisfacer las condiciones de orden q para $1 \le q \le 8$, y según el apartado anterior hay un total de $1+1+2+4+9+20+48+115 = 200$ condiciones de orden.

****1.9.9** Al quitar la raíz a un árbol de orden $k+1$ queda un "bosque" de árboles (que se suponen con raíz). Sea r_i, $1 \le i \le k$, el número de estos árboles que son de orden i. Debe cumplirse

$$1r_1 + 2r_2 + 3r_3 + \cdots + kr_k = \sum_{i=1}^{k} r_i = k.$$

El número de formas de elegir $r_1, r_2, \ldots, r_k$ que cumplan esta igualdad viene dado por el coeficiente de x^k en

$$(1 + x^{1\cdot 1} + x^{1\cdot 2} + x^{1\cdot 3}\ldots)(1 + x^{2\cdot 1} + x^{2\cdot 2} + x^{2\cdot 3}\ldots)\cdots(1 + x^{k\cdot 1} + x^{k\cdot 2}\ldots) = (1-x)^{-1}(1-x^2)^{-1}\cdots(1-x^k)^{-1}.$$

Sin embargo al elegir los r_i el árbol inicial no está totalmente determinado, ya que hay A_i posibles tipos distintos de árboles de orden i. Llamando s_{ij} al número de los r_i árboles de orden i que son del tipo j, $1 \le j \le A_i$; se tiene

que el número total de árboles de orden $k+1$ es el número de posibles s_{ij} tales que

$$\sum_{i=1}^{k} i(s_{i1} + s_{i2} + \cdots + s_{iA_i}) = k.$$

Procediendo como antes, se tiene que el número de posibles elecciones de los s_{ij} es el coeficiente de x^k en $((1-x)^{-1})^{A_1}((1-x^2)^{-1})^{A_2}\cdots((1-x^k)^{-1})^{A_k}$.

Ⓛ **1.9.10.** (Omitido).

1.10.1. Para el árbol lineal la condición de orden 10 correspondiente es

$$1 = 10!\, b_{10} \prod_{i=1}^{9} a_{i+1\,i} \leq 10!\, (\max |a_{i+1\,i}|)^9.$$

Despejando se obtiene el resultado buscado.

1.10.2. Hay Card T_i condiciones de orden i, y para que un método sea de orden q debe satisfacerlas todas con $1 \leq i \leq q$. Por tanto el número de ecuaciones, E_q, es

$$E_1 = 1,\ E_2 = 2,\ E_3 = 4,\ E_4 = 8,\ E_5 = 17,\ E_6 = 37,\ E_7 = 85,\ E_8 = 200.$$

Las incógnitas son todos los elementos a_{ij}, $i > j$, y los b_i del tablero. Esto es, hay

$$I_s = (1 + 2 + 3 + \cdots + (s-1)) + s = \frac{s(s+1)}{2}$$

incógnitas. Sería de esperar que existieran tableros que verificasen las ecuaciones si el número de incógnitas I_s fuera mayor o igual que el de ecuaciones E_q. Los valores mínimos de s para que se cumpla esta relación son

$$\begin{array}{ll} q = 1 \mapsto s = 1\ (E_1 = 1,\ I_1 = 1) & q = 5 \mapsto s = 6\ (E_5 = 17,\ I_6 = 21) \\ q = 2 \mapsto s = 2\ (E_2 = 1,\ I_2 = 3) & q = 6 \mapsto s = 9\ (E_6 = 37,\ I_9 = 45) \\ q = 3 \mapsto s = 3\ (E_3 = 1,\ I_3 = 6) & q = 7 \mapsto s = 13\ (E_7 = 85,\ I_{13} = 91) \\ q = 4 \mapsto s = 4\ (E_4 = 1,\ I_4 = 10) & q = 8 \mapsto s = 20\ (E_8 = 200,\ I_{20} = 210) \end{array}$$

Aunque esto es coherente con lo que asegura la barrera de Burcher en su forma más simple: para $s > 4$ no hay métodos de orden s y s etapas; no es una prueba ni siquiera en estos casos particulares; ya que en un sistema de

ecuaciones (no lineal en este caso), que el número de ecuaciones sea mayor o menor que el de incógnitas no asegura la existencia o inexistencia de solución.

1.10.3. La condición de orden correspondiente al árbol lineal de orden 5 es

$$120(b_5a_{54}a_{43}a_{32}a_{21} + b_6a_{64}a_{43}a_{32}a_{21}) = 1.$$

Sin más que sacar factor común $a_{43}a_{32}a_{21}$, se obtiene el resultado deseado.

1.10.4. Las condiciones de orden imponen que ciertas funciones polinómicas de coeficientes enteros evaluadas en los elementos del tablero sean 1. Esto implica que los denominadores siempre se deben simplificar, lo que sugiere que deben aparecer en varias fracciones. Más concretamente, la condición de orden 1 es $b_1 + b_2 + \ldots b_s = 1$ y si b_i fuera una fracción irreducible con denominador D coprimo con los de el resto de los b_j, entonces al reducir a común denominador se llegaría a una contradicción (el denominador sería múltiplo de D y el numerador no).

1.10.5. Las condiciones de orden correspondientes a los árboles con $\gamma(t) = q$ (todos los vértices unidos a la raíz) son

$$\begin{aligned} q = 1 &\longrightarrow \sum b_j = 1 \\ q = 2 &\longrightarrow \sum b_j c_j = 1/2 \\ \ldots &\quad \ldots\ldots \quad \ldots \\ q = s &\longrightarrow \sum b_j c_j^{s-1} = 1/s \end{aligned}$$

Dados los c_j distintos, el sistema es compatible determinado porque el determinante de la matriz de coeficientes es el de Vandermonde $\prod_{i<j}(c_j - c_i) \neq 0$.

1.11.1. a) Las condiciones de orden 1 y 2, $\sum b_j = 1$, $2\sum b_j c_j = 1$, se verifican, mientras que $3\sum b_j c_j^2 = 1$, que es de orden 3, no se cumple.

b) Para todo árbol t distinto del lineal, el diferencial elemental $F^J(t)$ se anula porque si hay alguna ramificación dará lugar a factores con derivadas segundas de f. En la fórmula para el desarrollo del error de truncación (con $p = 2$), el diferencial elemental del árbol lineal aparece multiplicado por $1 - 6\sum b_j a_{jk} c_k$ que para este método se anula. Así pues, $R_n = 0 \cdot h^{2+1} + O(h^4)$ y el método es de orden 3 para estos problemas lineales.

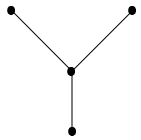
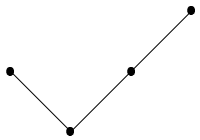
1.11.2. Como se había visto, todos los diferenciales elementales excepto el correspondiente al árbol lineal se anulan. En el caso de orden 4 este diferencial será

$$F^J = \sum_{K,L,M} f^J_K f^K_L f^L_M f^M = \sum_{k,l,r} m_{jk} m_{kl} m_{lr} \sum_s m_{rs} y^s.$$

Esto coincide con la $J = j$-ésima coordenada de $M \cdot M \cdot M \cdot My$. Por otra parte es fácil comprobar que todas las matrices del eneunciado cumplen que M^4 es la matriz nula. Así que se tiene como antes $R_n = 0 \cdot h^{3+1} + O(h^5)$ y el método es de orden 4 para este problema.

Esto no contradice el teorema mencionado porque el orden 4 es sólo en casos especiales. En general es sólo 2.

****1.11.3** Considérense dos árboles de orden 4 tales que para problemas escalares los diferenciales elementales no sean independientes. Por ejemplo

$A =$  $B =$ 

cumplen $F(A) = F(B) = f''f'f^2$. Para resolver el problema basta encontrar un método de orden 3 que satisfaga las condiciones de orden 4 correspondientes a todos los árboles distintos de A y B, que no las cumpla para A y B; y que sin embargo

$$\big(1 - \gamma(A) \sum_j b_j \phi_j(A)\big) + \big(1 - \gamma(B) \sum_j b_j \phi_j(B)\big) = 0.$$

Con ello se asegura que el desarrollo del error de truncación para problemas escalares sea $R_n = 0 \cdot h^{3+1} + O(h^5)$, mientras que como no todas las condiciones de orden 4 se cumplen, según el teorema de Butcher, sólo se tiene $R_n = O(h^{3+1})$ y no $R_n = O(h^{4+1})$.

Para conseguir un método que satisfaga muchas condiciones de orden hasta orden 4, es conveniente escoger los b_j y c_j de algún método de este orden (con ellos se tendrá automáticamente $l \sum b_j c_j^{l-1} = 1$). Por ejemplo

0					
1/3	1/3				
1/3	a_{31}	a_{32}			
1/2	a_{41}	a_{42}	a_{43}		
1	a_{51}	a_{52}	a_{53}	a_{54}	
	1/6	0	0	2/3	1/6

Las condiciones de orden 1 y 2 se cumplen automáticamente, así como las de orden 3 y 4 correspondientea a los árboles con $\gamma(t) = 3$ y $\gamma(t) = 4$, respectivamente. Faltan imponer las de orden 3 y 4 para los árboles lineales así como la relación antes mencionada entre las condiciones asociadas a A y B. Es decir,

$$6\sum b_j a_{jk} c_k = 1 \qquad 24\sum b_j a_{jk} a_{kl} c_l = 1$$
$$12\sum b_j a_{jk} c_k^2 + 8\sum b_j c_j a_{jk} c_k = 1$$

Además hay que imponer las condiciones de suma por fila

$$a_{31} + a_{32} = \frac{1}{3}, \qquad a_{41} + a_{42} + a_{43} = \frac{1}{2}, \qquad a_{51} + a_{52} + a_{53} + a_{54} = 1.$$

Esto hacen 6 ecuaciones con 9 incógnitas, así que es de esperar que podamos elegir tres de éstas como parámetros. Sean por ejemplo $a_{32} = a_{43} = a_{54} = 1$. Esto implica (de $a_{31} + a_{32} = 1/3$) que $a_{31} = -2/3$. Sustituyendo en las ecuaciones anteriores se obtiene un sistema lineal de 5 ecuaciones y 5 incógnitas

$$4(a_{42} + 1) + (a_{52} + a_{53}) = \frac{3}{2} \qquad a_{41} + a_{42} + 1 = \frac{1}{2}$$
$$32(a_{42} + 1) + 12(a_{52} + a_{53}) = 15 \qquad a_{51} + a_{52} + a_{53} + 1 = 1$$
$$4(a_{42} + 1) + 16 + 4a_{53} = 3$$

Al resolverlo se llega al tablero

0					
1/3	1/3				
1/3	−2/3	1			
1/2	−3/16	−5/16	1		
1	5/4	43/16	−63/16	1	
	1/6	0	0	2/3	1/6

Ⓛ **1.11.4.** (Omitido).

1.12.1. Considerando el desarrollo de Taylor de r_h en $h = 0$, cada coordenada del error $y(b) - r_h$ será de la forma $Ch^p + O(h^{p+1})$ (por definición de orden de convergencia) y lo mismo para $y(b) - r_{2h}$ cambiando h por $2h$. Restando se obtiene

$$\|r_{2h} - r_h\| = K(2^p - 1)h^p + O(h^{p+1}), \qquad K \neq 0.$$

Cambiando de nuevo h por $2h$ y dividiendo las igualdades correspondientes

$$\frac{\|r_{4h} - r_{2h}\|}{\|r_{2h} - r_h\|} = \frac{2^p + O(h)}{1 + O(h)} = 2^p + O(h).$$

Tomando logaritmos y límites cuando $h \to 0$ se tiene la fórmula deseada.

1.12.2. a) Tomando los valores de h menores, cabe esperar una mejor aproximación. Se estima entonces el error por

$$\frac{\log|y_N(0'0625) - y_N(0'03125)| - \log|y_N(0'03125) - y_N(0'015625)|}{\log 2} \approx 2.$$

b) Si el orden es 2, para h pequeño

$$y_{N,h} \approx y(x_N) + Ch^2, \qquad y_{N,2h} \approx y(x_N) + C(2h)^2.$$

De donde

$$|y(x_N) - y_{N,h}| \approx |C|h^2 \approx \frac{|y_{N,h} - y_{N,2h}|}{3}.$$

Al sustituir los tres últimos valores de h se obtienen respectivamente como estimaciones del error $3'27 \cdot 10^{-3}$, $1'31 \cdot 10^{-3}$ y $3'38 \cdot 10^{-4}$.

1.12.3. Si $|y(x_N) - y_{N,h}| \approx Kh^p$, $K \neq 0$; se tiene

$$p \approx \frac{\log|y(x_N) - y_{N,2h}| - \log|y(x_N) - y_{N,h}|}{\log 2}.$$

Con $h = 0'015625$ se consigue $p \approx 1'976\ldots$ que es coherente con la anterior estimación $p = 2$.

1.12.4. a) Denotando $y_{n+\frac{1}{2}}$ a la aproximación de Euler en $x_{n+\frac{1}{2}} = x_n + h/2$,

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

Sustituyendo,

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

b) Sea $y_{n+1,h}$ la fórmula para el método de Euler con paso h y sea $y_{n+1,h/2}$ la fórmula del apartado anterior. Según la extrapolación de Richardson, para $p = 1$ (el orden del método de Euler)

$$y_{n+1,h/2} + \frac{y_{n+1,h/2} - y_{n+1,h}}{2^p - 1}$$

es una aproximación de mayor orden. Por tanto el método

$$\begin{aligned} y_{n+1} =& 2y_{n+1,h/2} - y_{n+1,h} \\ =& y_n + hf(x_n + \frac{h}{2}, y_n + \frac{h}{2} f(x_n, y_n)) \end{aligned}$$

debe ser de orden (al menos) 2. De hecho es el método de Euler modificado.

1.12.5. Como en el ejercicio anterior, si $y_{n+1,h}$ corresponde a aplicar un método de orden p una vez con paso h mientras que $y_{n+1,h/2}$ corresponde a aplicarlo dos con paso $h/2$, entonces

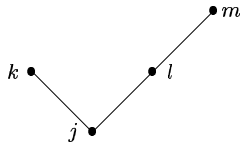
$$y_{n+1,h/2} + \frac{y_{n+1,h/2} - y_{n+1,h}}{2^p - 1}$$

da lugar a un método de orden (al menos) $p + 1$. Iterando el procedimiento se conseguiría uno de orden $p + 2$, $p + 3$, etc. En la práctica esto no produce método de orden alto que sean interesantes porque cabe esperar que en cada incremento de orden se duplique el número de evaluaciones de función, lo que da lugar a un crecimiento exponencial del número de etapas en comparación con el orden.

1.12.6. a) Las condiciones de orden , 2 y 3 se verifican con los cálculos

$$\frac{1}{6} + \frac{2}{3} + \frac{1}{6} = 1, \ 2(\frac{1}{3} + \frac{1}{6}) = 1, \ 3(\frac{2}{3} \cdot \frac{1}{4} + \frac{1}{6}) = 1, \ 6 \cdot \frac{1}{6}(\frac{4}{3} \cdot \frac{1}{2} + \frac{1}{3}) = 1.$$

Sin embargo, la de orden 4 correspondiente al árbol



no se verifica:

$$8\sum b_j c_j a_{jl} c_l = 8b_4(a_{42}c_2 + a_{43}c_3) \neq 1.$$

Con $h = 0'0375$, se tiene la aprosimación para el orden empírico

$$\frac{\log \texttt{error}(2h) - \log \texttt{error}(h)}{\log 2} = 3'946\ldots$$

Es decir, el orden teórico es 3 y el empírico 4.

b) No hay contradicción en ello porque el orden empírico se calculaba usando datos de un problema concreto en un punto dado, y sin embargo la definición de orden (teórico) requería estimaciones que se cumpliesen para todo problema suficientemente regular.

1.12.7. Si $\texttt{error}(h) \approx Ch^p$, $C \neq 0$, entonces el logaritmo del error es una función lineal del logaritmo de h con pendiente p. Se puede hacer regresión lineal con los logaritmos de los valores de la tabla o simplemente usar las dos correspondientes a los menores valores de h y la fórmula

$$\text{pendiente} = \frac{\Delta y}{\Delta x} = \frac{\log 0'00458 - \log 0'00368}{\log(1/17) - \log(1/18)} = 3'8\ldots$$

Y el orden empírico es 4.

1.12.8. Si h es suficientemente pequeño y los errores de redondeo no intervienen, el error debe ser como Ch^p, $C \neq 0$, y por tanto su signo debe coincidir siempre con el de C.

Ⓛ **1.12.9.** (Omitido).

Ⓛ **1.12.10.** (Omitido).

1.13.1. Según la extrapolación de Richardson, la estimación del error es

$$\frac{y_2 - \widetilde{y}_2}{2^3 - 1} \approx 3'4 \cdot 10^{-3}.$$

Y la fórmula del cambio de paso (sin factor de seguridad) sugiere tomar un paso de tamaño a lo más

$$h' = 0'1\Big(\frac{2 \cdot 10^{-3}}{3'4 \cdot 10^{-3}}\Big)^{1/(3+1)} = 0'087\ldots$$

1.13.2. La fórmula para el método de Euler con paso h, en este caso es

$$y_{n+1} = (1+h)y_n + h\,\mathrm{sen}(x_n y_n).$$

Y las fórmulas que aproximan el error (local) y cambian el paso son respectivamente

$$e \approx \frac{y_{n+2} - \widetilde{y}_{n+2}}{2^1 - 1}, \qquad h' = 0'9h\sqrt{\frac{0'01}{e}}$$

donde $\widetilde{y}_{n+2}$ es el resultado de aplicar el método de Euler una sola vez, a partir de (x_n, y_n), pero con paso $2h$. Es decir

$$\widetilde{y}_{n+2} = (1+2h)y_n + 2h\,\mathrm{sen}(x_n y_n).$$

Partiendo de x_0, $y_0 = 1$ con $h = 0'1$, se obtiene

$$\begin{array}{llll} x_1 = 0'1 & y_1 = 1'1 & & \\ x_2 = 0'2 & y_2 = 1'2209\ldots & \widetilde{y}_2 = 1'2 & \Rightarrow\ e \approx 0'0209 \end{array}$$

Como el error supera el tolerado, se deben descartar estas dos iteraciones y repetirlas con el nuevo paso $h' = 0'0621\ldots$ Con ello

$$\begin{array}{llll} x_1 = 0'0621\ldots & y_1 = 1'0621\ldots & & \\ x_2 = 0'1242\ldots & y_2 = 1'1322\ldots & \widetilde{y}_2 = 1'1242\ldots & \Rightarrow\ e \approx 0'0079 \end{array}$$

Este valor es admisible y se puede disminuir el paso hasta $h' = 0'0556\ldots$ Lo que da lugar a

$$\begin{array}{llll} x_3 = 0'1869\ldots & y_3 = 1'2112\ldots & & \\ x_4 = 0'2496\ldots & y_4 = 1'3020\ldots & \widetilde{y}_4 = 1'2917\ldots & \Rightarrow\ e \approx 0'00103 \end{array}$$

De nuevo hay descartar estos resultados y repetirlos con $h' = 0'0556\ldots$ Obteniéndose

$$x_3 = 0'1798\ldots \quad y_3 = 1'2030\ldots$$
$$x_4 = 0'2355\ldots \quad y_4 = 1'2818\ldots \quad \widetilde{y}_4 = 1'2737\ldots \;\Rightarrow\; e \approx 0'0080$$

Se puede aumentar ligeramente el paso hasta $h' = 0'0557\ldots$ Las últimas iteraciones son

$$x_5 = 0'2912\ldots \quad y_5 = 1'3698\ldots$$
$$x_6 = 0'3464\ldots \quad y_6 = 1'4677\ldots \quad \widetilde{y}_6 = 1'4577\ldots \;\Rightarrow\; e \approx 0'0099$$

Ⓛ **1.13.3.** (Omitido).

1.14.1. Según los datos del problema, el tablero es

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

El primer método es de orden 2 (las condiciones de orden se cumplen trivialmente) y explícito ya que podría escribirse su tablero como (la última etapa es superflua porque $b_3 = 0$)

0		
1/2	1/2	
	0	1

El segundo es implícito porque $a_{33} \neq 0$. De nuevo, las condiciones hasta orden 3 son muy sencillas de comprobar. El método no es de orden 4 porque

$$24 \sum b_j a_{jk} a_{kl} c_l = 24 \cdot \frac{1}{6} \cdot 1 \cdot 1 \neq 1.$$

1.14.2. Los métodos de orden 1 y de orden 2 responden a los tableros

0		
c_2	c_2	
	b_1	$1 - b_1$

0		
c_2	c_2	
	$1 - \frac{1}{2c_2}$	$\frac{1}{2c_2}$

con $2c_2(1-b_1)\neq 1$. Lo que da lugar a los pares encajados

0		
c_2	c_2	
	b_1	$1-b_1$
	$1-\frac{1}{2c_2}$	$\frac{1}{2c_2}$

1.14.3. Si $b_s=0$ simplemente no hace falta multiplicar por él. Esto no constituye por sí mismo una gran ventaja, pero si además $a_{si}=b_i$ entonces

$$k_s=f(x_n+h\sum b_i, y_n+h\sum b_ik_i)=f(x_{n+1},y_{n+1}),$$

de modo que k_s coincide con el k_1 de la siguiente iteración. A efectos del número de operaciones todo funciona como si se tuviera una etapa menos.

Si en el problema anterior se impone $b_1=c_2$, $1-b_1=0$, el tablero es

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

1.14.4. Partiendo del tablero

0			
c_2	c_2		
b_1+b_2	b_1	b_2	
	b_1	b_2	0
	$\widehat{b}_1$	$\widehat{b}_2$	$\widehat{b}_3$

las condiciones de orden 1 para el primer método y las de orden 1 y 2 para el segundo se traducen en las ecuaciones

$$b_1+b_2=1,\qquad \widehat{b}_1+\widehat{b}_2+\widehat{b}_3=1,\qquad 2(c_2\widehat{b}_2+(b_1+b_2)\widehat{b}_3)=1.$$

Una de cuyas soluciones es $c_2=1$, $b_1=b_2=1/2$, $\widehat{b}_1=\widehat{b}_3=1/2$, $\widehat{b}_2=0$. Lo que corresponde al tablero

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

1.14.5. Considérese como método de avance el de Euler y como método para controlar el error el de Euler modificado. El cambio de paso vendrá dado como antes por $h' = 0'9h\sqrt{0'01/e}$.

Partiendo de $x_0 = 0$, $y_0 = 1$, $\widetilde{y}_0 = 1$, se tiene

$$x_1 = 0'1, \quad y_1 = 1'1, \quad \widetilde{y}_1 = 1'1102\ldots \Rightarrow\ e \approx 0'0102,\ h' = 0'0889\ldots$$

Como el error excede $0'01$, hay que descartar estos cálculos y repetirlos con h'. Se sigue iterando (cambiando el paso cada vez) hasta que el error exceda el tolerado

$$\begin{aligned}
&x_1 = 0'0839\ldots\ y_1 = 1'0889\ldots\ \widetilde{y}_1 = 1'0969\cdots \Rightarrow e \approx 0'0080\ h' = 0'0889\ldots\\
&x_2 = 0'1778\ldots\ y_2 = 1'1944\ldots\ \widetilde{y}_1 = 1'2034\cdots \Rightarrow e \approx 0'0090\ h' = 0'0844\ldots\\
&x_3 = 0'2623\ldots\ y_3 = 1'3130\ldots\ \widetilde{y}_3 = 1'3232\cdots \Rightarrow e \approx 0'0102\ h' = 0'0751\ldots
\end{aligned}$$

Descartando la última iteración y repitiéndola con $h' = 0'0751\ldots$ se obtiene

$$\begin{aligned}
&x_3 = 0'2530\ldots\ y_3 = 1'3000\ldots\ \widetilde{y}_3 = 1'3081\cdots \Rightarrow e \approx 0'0080\ h' = 0'0752\ldots\\
&x_4 = 0'3283\ldots\ y_4 = 1'4222\ldots\ \widetilde{y}_4 = 1'4315\cdots \Rightarrow e \approx 0'0092\ h' = 0'0702\ldots\\
&x_5 = 0'3986\ldots\ y_5 = 1'5537\ldots\ \widetilde{y}_5 = 1'5629\cdots \Rightarrow e \approx 0'0091
\end{aligned}$$

La cantidad de operaciones es menor en este caso.

1.14.6. Por un problema anterior, el método "grande" (el de cuatro etapas) tiene orden 4. Basta por tanto ver que el método "pequeño" tiene orden 2, lo cual se reduce a las simples comprobaciones $\sum b_j = 2\sum b_j c_j = 1$. Además no tiene orden 3 porque $3\sum b_j c_j^2 \neq 1$.

1.14.7. Las condicones de orden 1 y 2 aplicadas a ambos métodos implican

$$b_1 + b_2 = 1, \quad 2b_2 = 1, \quad \widehat{b}_1 + \widehat{b}_2 + \widehat{b}_3 = 1, \quad 2\widehat{b}_2 + \widehat{b}_3 = 1.$$

Por tanto $b_1 = b_2 = 1/2$, $\widehat{b}_3 = 1 - 2\widehat{b}_2$, $\widehat{b}_1 = \widehat{b}_2$. Si a estas ecuaciones se les añade la condición de orden 3 dada por $3\sum \widehat{b}_j c_j^2 = 1$, se tiene $3\widehat{b}_2 + \frac{3}{4}\widehat{b}_3 = 1$ y de aquí $\widehat{b}_1 = \widehat{b}_2 = 1/6$, $\widehat{b}_3 = 4/6$. La otra condición de orden 3, $6\sum \widehat{b}_j a_{jk} c_k = 1$, impone $6\ \widehat{b}_3 a_{22} c_2 = 1$, lo que lleva al único tablero (usando la condición de suma por filas)

0			
1	1		
1/2	1/4	1/4	
	1/2	1/2	0
	1/6	1/6	4/6

Ⓛ **1.14.8.** (Omitido).

2. Problemas stiff

2.1.1. Los autovalores de la matriz son $\lambda_1 = -1$ y $\lambda_2 = -3$. Buscando soluciones no triviales de $(A - \lambda_i I)\vec{v}_i = \vec{0}$ se sigue que $\vec{v}_1 = (1, -1)^t$ y $\vec{v}_2 = (2, -1)^t$ forman una base de autovectores. Es decir

$$A = \begin{pmatrix} -5 & -4 \\ 2 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 2 \\ -1 & -1 \end{pmatrix} \begin{pmatrix} -1 & 0 \\ 0 & -3 \end{pmatrix} \begin{pmatrix} 1 & 2 \\ -1 & -1 \end{pmatrix}^{-1} = PDP^{-1}.$$

El método de Euler con $h = 1$ es

$$y_{n+1} = y_n + Ay_n = (I + hA)y_n = P(I + D)P^{-1}y_n,$$

que iterado implica

$$y_n = P(I + D)^n P^{-1} y(0) = P \begin{pmatrix} 0 & 0 \\ 0 & (-2)^n \end{pmatrix} P^{-1} y(0).$$

Si $P^{-1}y(0)$ es un vector de cordenadas λ y μ entonces

$$y_n = P \begin{pmatrix} 0 \\ (-2)^n \mu \end{pmatrix},$$

y la única forma de que se cumpla $\lim y_n = 0$ es que $\mu = 0$. Además

$$P^{-1}y(0) = \begin{pmatrix} \lambda \\ 0 \end{pmatrix} \Leftrightarrow y(0) = \lambda P \begin{pmatrix} 1 \\ 0 \end{pmatrix} \Leftrightarrow y(0) = \lambda \vec{v}_1.$$

En ese caso y_n es idénticamente nulo para $n > 1$.

2.1.2. Sea $\vec{v}$ el autovector de un autovalor λ y sea v_i su coordenada mayor en valor absoluto. La $i-$ésima coordenada de la ecuación $(A-\lambda I)\vec{v}=\vec{0}$ es

$$a_{i1}v_1+\cdots+a_{ii-1}v_{i-1}+(a_{ii}-\lambda)v_i+a_{ii+1}v_{i+1}+\cdots+a_{in}v_n=0,$$

que aplicando la desigualdad triangular lleva a

$$|\lambda-a_{ii}||v_i|\leq|a_{i1}v_1+\cdots+a_{ii-1}v_{i-1}+a_{ii+1}v_{i+1}+\cdots+a_{in}v_n|\leq|v_i|\sum_{j\neq i}|a_{ij}|.$$

Simplificando v_i, se tiene que $\lambda\in D_i$.

2.1.3. Según el teorema de los círculos de Gershgorin, los autovalores $\lambda_1,\ \lambda_2$ cumplen $|\lambda_i-r_i|\leq 1$ para ciertos $r_1,r_2\in(-1,1)$, por tanto $|\lambda_i|<2$. La solución exacta y la $n-$ésima iteración del método de Euler son, respectivamente,

$$y(x)=P\begin{pmatrix}e^{\lambda_1 x}&0\\0&e^{\lambda_2 x}\end{pmatrix}\vec{C},\qquad y_n=P\begin{pmatrix}(1+\lambda_1 h)^n&0\\0&(1+\lambda_2 h)^n\end{pmatrix}\vec{C}$$

con $\vec{C}=P^{-1}y(0)$. Si $\lim_{x\to+\infty}y(x)=0$ y la $i-$ésima coordenada $(i=1,2)$ es no nula, necesariamente $\lambda_i<0$, y como $|\lambda_i|<2$, $(1+\lambda_i h)^n\to 0$ para cualquier $h<1$. Obviamente, si la $i-$ésima coordenada es nula, $(1+\lambda_i h)^n$ aparece multiplicado por cero. En cualquier caso $\lim_{x\to+\infty}y(x)=0$ implica $\lim y_n=0$.

2.1.4. Los autovalores son -1 y -2 con lo cual al aplicar el método de Euler

$$y_{n+1}=(I+hA)y_n=P\left(I+h\begin{pmatrix}-2&0\\0&-1\end{pmatrix}\right)P^{-1}y_n.$$

Y como antes,

$$y_n=P\begin{pmatrix}(1-2h)^n&0\\0&(1-h)^n\end{pmatrix}P^{-1}y_0$$

que tiende a cero para todo $0<h<1$. La razón es que los autovalores de una matriz con elementos grandes pueden ser pequeños.

Ⓛ **2.1.5.** (Omitido).

2.2.1. Al aplicar el método a $y' = \lambda y$,

$$y_{n+1} = y_n + \lambda h \frac{y_n + y_{n+1}}{2} \;\Rightarrow\; y_{n+1} = \frac{2 + h\lambda}{2 - h\lambda} y_n.$$

Por consiguiente el dominio de estabilidad lineal es

$$\mathcal{D} = \left\{ z \in \mathbb{C} \;:\; \left| \frac{2+z}{2-z} \right| < 1 \right\} = \{ \mathrm{Re}\, z < 0 \}.$$

La última igualdad se puede deducir algebraicamente escribiendo $z = x + iy$ o de manera geométrica representando los números complejos como vectores: z está en el semiplano izquierdo si y sólo si z y 2 forman un ángulo obtuso, lo que equivale (recuérdese la regla del paralelogramo) a que $2 + z$ mida menos que $2 - z$.

2.2.2. El dominio de estabilidad lineal del método de Euler mejorado es

$$\mathcal{D} = \{ z \in \mathbb{C} \;:\; |R(z)| < 1 \} \qquad \text{con } R(z) = 1 + z + \frac{z^2}{2}.$$

Un giro de 180º alrededor de $z = -1$ viene dado por $z \mapsto -2 - z$. Como $R(-2 - z) = R(z)$ (lo cual se reduce a un simple cálculo) el dominio de estabilidad lineal es invariante por este tipo de giros.

De la misma forma, $|R(z)| = |R(\overline{z})|$ así que también es simétrico por $\mathrm{Im}\, z = 0$. Finalmente, tras unos cálculos

$$R(-1 + r + iy) = \frac{r^2 + 1 - y^2}{2} + iry.$$

Cuyo módulo es invariante al cambiar r por $-r$. De modo que el dominio de estabilidad lineal también es invariante por la transformación $-1 + r + iy \mapsto -1 - r + iy$ que corresponde a una simetría por la recta $\mathrm{Re}\, z = -1$.

2.2.3. a) La regla del trapecio y el método implícito del enunciado implican, al ser aplicados a $y' = \lambda y$, las fórmulas de recurrencia

$$y_{n+1} = \frac{2 + h\lambda}{2 - h\lambda} y_n, \qquad y_{n+1} = \frac{1}{2 - h\lambda} y_n.$$

Evidentemente $(2 + z)/(2 - z) \to -1$ y $1/(1 - z) \to 0$.

b) La verdadera solución de $y' = \lambda y$ cumple

$$\frac{y(x_{n+1})}{y(x_n)} = \frac{e^{\lambda(x_n+h)}y(0)}{e^{\lambda x_n}y(0)} = e^{\lambda h}.$$

Si $-\text{Re }\lambda$ es grande en comparación con h^{-1}, $e^{\lambda h}$ es muy pequeño, con un decaimiento rápido según Re $\lambda h \to -\infty$. Los métodos $L-$estables imitan este decaimiento ya que

$$\frac{y_{n+1}}{y_n} = R(\lambda h) \to 0 \qquad \text{cuando Re } \lambda h \to -\infty.$$

2.2.4. Los autovalores de la matriz del enunciado son -20 y -1. En la base de los autovectores este problema vectorial se "desacopla" en los dos problemas escalares $y' = -20y$, $y' = -y$. Por tanto el supremo pedido es

$$H = \sup\{h \ : \ |R(-20h)| < 1, \ |R(-h)| < 1\}.$$

La función de amplificación viene dada por $R(x) = 1+x+x^2/2!+x^3/3!+x^4/4!$ que es un polinomio que tiende a $+\infty$ cuando $x \to -\infty$ y tiende a 1^- cuando $x \to 0^-$; así que existirá un valor mínimo $x_0 < 0$ con $R(x_0) = 1$, tal que $R(x) < 1$ en $(-\infty, x_0)$ y $0 < R(x) < 1$ en $(x_0, x_0 + \epsilon]$. Si además $|R(x_0/20)| < 1$, se tiene necesariamente $H = -x_0/20$. En definitiva, hay que buscar la menor raíz negativa, x_0, de la ecuación

$$1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} = 1.$$

Lo que equivale, eliminando la raíz trivial $x = 0$, a hallar la menor raíz de

$$x^3 + 4x^2 + 12x + 24 = 0.$$

El primer miembro es una función creciente f, así que $f(x) = 0$ tiene una sola raíz. Con bisecciones sucesivas

$$f(-3) < 0 < f(-2), \quad f(-2'8) < 0 < f(-2'5), \quad f(-2'79) < 0 < f(-2'78).$$

Por tanto $x_0 = -2'79 + \delta$ con $0 < \delta < 0'01$ (un cálculo comprueba que $|R(x_0/20)| < 1$). De modo que

$$H = \frac{2'79}{20} + \widetilde{\delta} = 0'1395 + \widetilde{\delta} \qquad \text{con } |\widetilde{\delta}| < 0'0005.$$

Esto aproxima H con un error relativo menor que el 1%.

Ⓛ **2.2.5.** (Omitido).

2.3.1. Basta emplear la fórmula que define el método de Euler mejorado o utilizar que es un método explícito de dos etapas y orden 2 para concluir que

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

2.3.2. Si un método es consistente $\vec{b}^t\vec{1} = 1$, y por tanto

$$R(z) = 1 + z\vec{b}^t(I - Az)^{-1}\vec{1} = 1 + z\vec{b}^t\vec{1} + O(z^2) = 1 + z + O(z^2),$$

como afirma la indiciación.

Según esta fórmula, para $\epsilon > 0$ suficientemente pequeño, se tiene

$$R(-\epsilon) = 1 - \epsilon + O(\epsilon^2) < 1 < 1 + \epsilon + O(\epsilon^2) = R(\epsilon).$$

Por tanto $-\epsilon \in \mathcal{D}$ y $\epsilon \notin \mathcal{D}$ donde $\mathcal{D}$ es el dominio de estabilidad lineal. De manera que $0 \in \overline{\mathcal{D}}$ y $0 \in \mathrm{Fr}(\mathcal{D})$. Además considerando un pequeño entorno abierto de ϵ disjunto con $\mathcal{D}$, se tiene $0 \in \overline{\mathcal{D}} \cap \overline{(\mathbb{C} - \overline{\mathcal{D}})} = \mathrm{Fr}(\overline{\mathcal{D}})$.

2.3.3. Si un método es de orden p entonces

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

Por otra parte, la función de amplificación de un método explícito de s etapas es un polinomio de grado menor o igual que s. Estas dos propiedades son compatibles sólo si $p \leq s$.

⋆**2.3.4** Se tiene la identidad

$$(I - zA)(I + zA + z^2A^2 + \cdots + z^kA^k) = I - z^{k+1}A^{k+1}.$$

Invirtiendo y despejando $(I - zA)^{-1}$

$$\begin{aligned}(I - zA)^{-1} =& (I + zA + z^2A^2 + \cdots + z^kA^k)(I - z^{k+1}A^{k+1})^{-1} \\ =& (I + zA + z^2A^2 + \cdots + z^kA^k)(I + O(z^{k+1})) \\ =& I + zA + z^2A^2 + \cdots + z^kA^k + O(z^{k+1}).\end{aligned}$$

Donde la igualdad $(I - z^{k+1}A^{k+1})^{-1} = I + O(z^{k+1})$ se puede deducir, por ejemplo, de la fórmula habitual para el cálculo de la inversa.

2.3.5. Según el problema anterior con $k = 1$,

$$1 + z\vec{b}^t(I - zA)^{-1}\vec{1} = 1 + z + \frac{z^2}{2} + O(z^3)$$

equivale a

$$+z\vec{b}^t(I + zA)\vec{1} = 1 + z + \frac{z^2}{2}.$$

Igualando coeficientes, esto es lo mismo que las condiciones de orden 1 y 2: $\vec{b}^t\vec{1} = 1$, $2\vec{b}^tA\vec{1} = 1$.

2.3.6. El método no es de orden 3 porque no se cumple $3\sum b_jc_j^2 = 1$.
Para calcular la función de amplificación, se considera

$$(I - zA)^{-1}\vec{1} = \begin{pmatrix} 1 & 0 & 0 \\ z/2 & 1 & 0 \\ z^2/6 & z/3 & 1 \end{pmatrix}\vec{1} = \begin{pmatrix} 1 \\ 1 + z/2 \\ 1 + z/3 + z^2/6 \end{pmatrix}.$$

Por tanto

$$R(z) = 1 + z\Big(-\frac{1}{3} + \frac{1}{3}(1 + \frac{z}{2}) + (1 + \frac{z}{3} + \frac{z^2}{6})\Big) = 1 + z + \frac{z^2}{2} + \frac{z^3}{6}.$$

*__2.3.7__ Siempre que un método sea de orden p para el problema escalar $y' = \lambda y$, se tendrá $r(z) = 1 + z + \cdots + z^p/p! + O(z^{p+1})$. Por ello basta demostrar que hay métodos que tienen orden p para dicho problema pero no en general.

Sea el método explícito con p etapas que tiene $a_{ij} = 1$ si $i < j$. Los diferenciales elementales para el problema $y' = \lambda y$ son todos nulos excepto los que corresponden a los árboles lineales. Si t es el árbol lineal de orden r

$$\phi_j(t) = \sum_{j_2 > \cdots > j_r} a_{jj_2}a_{j_2j_3}\ldots a_{j_{r-1}j_r} = \binom{j-1}{r-1},$$

entendiendo este número combinatorio como cero si $r > j$. Por tanto eligiendo los b_j de manera que

$$r!\sum_{j=1}^{p} b_j\binom{j-1}{r-1} = 1, \qquad r = 1, 2, \ldots, p;$$

el método tendrá orden p para $y' = \lambda y$. Este sistema tiene solución (su matriz es triangular no singular). Además el método obtenido no tiene orden p en general porque fallan las condiciones de orden. Por ejemplo

$$3\sum_{j=1}^{p} b_j c_j^2 = 3\sum_{j=1}^{p} b_j (j-1)^2 = 3\sum_{j=1}^{p} b_j \left(2\binom{j-1}{2} + \binom{j-1}{1}\right),$$

que según la elección de los b_j es $6 \cdot 1/3! + 3 \cdot 1/2! \neq 1$.

2.3.8. El tablero del método será de la forma

$$\begin{array}{c|cc} 1 & 1 & 0 \\ \alpha+\beta & \alpha & \beta \\ \hline & 1-b & b \end{array}$$

Se puede suponer $b \neq 0$ y $\alpha + \beta \neq 1$, ya que en otro caso no podría verificar la condición de orden 2.

La función de amplificación $R(z)$ es

$$\begin{aligned} 1 + z\vec{b}^t \begin{pmatrix} 1-z & 0 \\ -\alpha z & 1-\beta z \end{pmatrix}^{-1} \vec{1} &= 1 + \frac{z\vec{b}^t}{(1-z)(1-\beta z)} \begin{pmatrix} 1-\beta z & 0 \\ \alpha z & 1-z \end{pmatrix} \vec{1} \\ &= 1 + z\frac{1 + (b(\alpha-1) - \beta(1-b))z}{(1-z)(1-\beta z)}. \end{aligned}$$

Si $\beta = 0$, por las reducciones iniciales $b \neq 0$ y $\alpha \neq 1$, de manera que $\lim R(z) = \infty$. Si $\beta \neq 0$

$$\lim R(z) = 1 + \frac{b(\alpha-1) - \beta(1-b)}{\beta} = \frac{1 - b + b(\alpha+\beta) - 1}{\beta} = \frac{-1}{2\beta} \neq 0,$$

donde se ha usado la condición de orden 2, $1 - b + b(\alpha + \beta) = 1/2$.

2.3.9. Tras algunos cálculos se comprueba que el método es de orden 3. Como es explícito, la función de amplificación es

$$R(z) = 1 + z + \frac{z^2}{2} + \frac{z^3}{6}.$$

De modo que

$$R(it) = 1 - \frac{t^2}{2} + it\big(1 - \frac{t^2}{6}\big).$$

Así pues, por ejemplo $|R(\pm i\sqrt{2})| = 2\sqrt{2}/3 < 1$, y $\pm i\sqrt{2} \in \mathcal{D}$.

*__2.3.10__ La función de amplificación será de la forma

$$R(z) = \frac{P(z)}{1 - a_{11}z}$$

donde P es un polinomio real.

Si gr $P > 1$ entonces $\lim_{z\to\infty} R(z) = \infty$ y el dominio de estabilidad lineal $\mathcal{D}$ es acotado. Si gr $P = 0$ entonces $\mathcal{D}$ es el exterior de una circunferencia $|1 - za_{11}| = \text{cte}$. Por tanto la única posibilidad es gr $P = 1$. Además como $R(0) = 1$, debe ser

$$R(z) = \frac{1 + \beta z}{1 - a_{11}z}.$$

Si $\mathcal{D} = \{\text{Re } z < 0\}$, necesariamente $|R(it)| = 1$ para $t \in \mathbb{R}$, con lo cual $\beta = \pm a_{11}$ (para ello basta tomar $t \to \infty$). El caso $\beta = -a_{11}$ lleva a la contradicción $\mathcal{D} = \emptyset$. Por tanto la única posibilidad es $\beta = a_{11}$ y

$$R(z) = \frac{1 + a_{11}z}{1 - a_{11}z} = -1 + \frac{2}{1 - a_{11}z} = 1 + 2a_{11}z + 2a_{11}^2z^2 + 2a_{11}^3z^3 + O(z^4).$$

Si el método es consistente, $R(z) = 1 + z + O(z^2)$ (porque $\vec{b}^t\vec{1} = 1$), de modo que $a_{11} = 1/2$ y

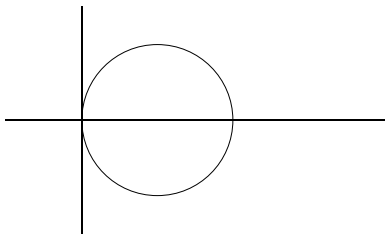
$$R(z) = \frac{2 + z}{2 - z} = 1 + z + \frac{z^2}{2} + \frac{z^3}{4} + O(z^4).$$

Por un problema anterior, $R(z) = 1 + z + z^2/2 + O(z^3)$ implica que el método es de orden 2. No es de orden 3 porque $R(z) \neq 1 + z + z^2/2 + z^3/6 + O(z^4)$.

__2.3.11.__ La función de amplificación es

$$R(z) = 1 + z\begin{pmatrix}0 & 1\end{pmatrix}\begin{pmatrix}1 - z/2 & -z/2 \\ z/2 & 1 - 3z/2\end{pmatrix}^{-1}\begin{pmatrix}1 \\ 1\end{pmatrix}$$

que después de efectuar los cálculos lleva a $R(z) = 1/(1 - z)$. El dominio de estabilidad lineal es la región exterior a la circunferencia $|z - 1| = 1$



Por tanto $\mathcal{D} \supset \{\mathrm{Re}\, z < 0\}$ y el método es $A-$estable.

2.3.12. Tras algunos cálculos se prueba que la función de amplificación es

$$R(z) = 1 + z \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -z/2 & 1 - z/2 \end{pmatrix}^{-1} \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \frac{2+z}{2-z}.$$

Según se vio en un problema anterior la región $|(2+z)/(2-z)| < 1$ coincide exactamente con el semiplano izquierdo, y en consecuencia el método es $A-$estable.

2.3.13. Siguiendo la indicación, el tablero se puede escribir como

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

donde $\overline{\alpha}$ es el conjugado real de α.

La función de amplificación es

$$R(z) = 1 + z \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \end{pmatrix} \frac{1}{(1-\alpha z)^2} \begin{pmatrix} 1 - \alpha z & 0 \\ (\overline{\alpha} - \alpha) z & 1 - \alpha z \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix},$$

que después de operar da lugar a

$$R(z) = \frac{1 + (1-2\alpha)z + (\alpha^2 - 2\alpha + 1/2)z^2}{(1-\alpha z)^2} = \frac{1 + (1-2\alpha)z + (1/3 - \alpha)z^2}{(1-\alpha z)^2}.$$

Esta función meromorfa tiene su único polo en el semiplano derecho (y está acotada en el izquierdo). Por el principio del máximo será $A-$estable si y sólo si para todo $t \in \mathbb{R}$

$$|1 + (1-2\alpha)it + (\alpha - 1/3)t^2| \leq |1 - \alpha i t|^2.$$

Elevando al cuadrado en ambos miembros

$$1+2\big(\alpha-\frac{1}{3}\big)t^2+\big(\alpha-\frac{1}{3}\big)^2t^4+(1-2\alpha)^2t^2\le 1+2\alpha^2t^2+\alpha^4t^4.$$

Nótese que $2(\alpha-1/3)+(1-2\alpha)^2=2\alpha^2+(2\alpha^2-2\alpha+1/3)=2\alpha^2$, con lo cual los términos en t^2 se simplifican y la desigualdad se reduce a $(\alpha-1/3)^2t^4\le\alpha^4t^4$ que es trivialmente cierta porque $(\alpha-1/3)^2<0'3<\alpha^4$.

2.3.14. a) Se puede suponer (dividiendo por una constante) que $P(0)=Q(0)=1$, de modo que todo lo que hay que hacer es hallar a, b, c y d tales que

$$\frac{1+ax+bx^2}{1+cx+dx^2}=1+x+\frac{x^2}{2}+\frac{x^3}{6}+\frac{x^4}{24}+O(x^5).$$

Al multiplicar el segundo miembro por $1+cx+dx^2$, se tiene, salvo términos $O(x^5)$,

$$1+(c+1)x+(d+c+\frac{1}{2})x^2+(d+\frac{1}{2}c+\frac{1}{6})x^3+(\frac{1}{2}d+\frac{1}{6}c+\frac{1}{24})x^4.$$

Así pues hay que resolver el sistema lineal

$$\begin{aligned} a&=c+1 & b &= d+c+\frac{1}{2}\\ 0&=d+\frac{1}{2}c+\frac{1}{6} & 0 &= \frac{1}{2}d+\frac{1}{6}c+\frac{1}{24}\end{aligned}$$

La solución única es $a=-c=1/2$, $b=d=1/12$, y la aproximante de Padé correspondiente es

$$f(x)=\frac{1+x/2+x^2/12}{1-x/2+x^2/12}.$$

b) Por ser de dos etapas, $R(z)=P(z)/Q(z)$ con P y Q polinomios de grado menor o igual que 2. Y por ser de orden 4

$$R(z)=1+z+\frac{z^2}{2}+\frac{z^3}{6}+\frac{z^4}{24}+O(z^5).$$

Por la unicidad de las aproximantes de Padé, $R(z)$ es la función del apartado anterior.

*c) Según el razonamiento anterior, en estas condiciones, $R(z)$ debe ser la aproximante de Padé (s,s) de la exponencial. Procediendo como en el

primer apartado, los coeficientes de ésta se pueden calcular como solución de un sistema lineal compatible determinado (por la existencia y unicidad) con coeficientes racionales. Por la regla de Cramer, la solución también es racional.

Ⓛ **2.3.15.** (Omitido).

3. Diferencias finitas

3.1.1. Sea U_i la aproximación de $u(x_i)$, con la notación habitual $a = x_0$, $b = X_N$, $h = (b-a)/N$. Del "cociente incremental" para la derivada segunda, se tiene

$$\frac{U_{i+1} + U_{i-1} - 2U_i}{h^2} = f(x_i), \qquad i = 1, 2, \ldots N-1;$$

donde se define, como es natural, $U_0 = U_N = 0$. Esto se puede escribir como el sistema con matriz tridiagonal

$$h^{-2} \begin{pmatrix} -2 & 1 & & & \\ 1 & -2 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & -2 & 1 \\ & & & 1 & -2 \end{pmatrix} \begin{pmatrix} U_1 \\ U_2 \\ \vdots \\ U_{N-2} \\ U_{N-1} \end{pmatrix} = \begin{pmatrix} f(x_1) \\ f(x_2) \\ \vdots \\ f(x_{N-2}) \\ f(x_{N-1}) \end{pmatrix}$$

Sea M_{N-1} la matriz de coeficientes de este sistema sin el factor h^{-2}. Desarrollando por la primera columna se tiene

$$\det M_{N-1} = -2 \det M_{N-2} - \det M_{N-3}.$$

Partiendo de los valores iniciales $\det M_1 = -2$, $\det M_2 = 3$, no es difícil probar por inducción $\det M_{N-1} = (-1)^{N-1} N$. Por tanto $\det A = (-h^{-2})^{N-1} N$.

3.1.2. (C. Moreno p. 227). Numerando los cinco nodos interiores de izquierda a derecha y de arriba a abajo; y aplicando en cada uno de ellos la fórmula de los cinco puntos, se obtiene el sistema

$$U_2 + U_4 + 4 - 4U_1 = 0, \qquad U_1 + U_3 + 2 - 4U_2 = 0, \qquad U_2 + 3 - 4U_3 = 0,$$
$$U_1 + U_5 + 2 - 4U_4 = 0, \qquad U_4 + 3 - 4U_5 = 0.$$

Resolviéndolo, se llega a la solución

$$(U_1, U_2, U_3, U_4) = \frac{1}{26}(41, 30, 27, 30, 27).$$

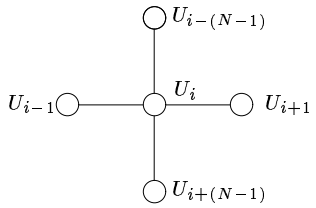
Por la simetría del problema se puede anticipar $U_2 = U_4$ y $U_3 = U_5$, lo que simplifica la resolución del sistema.

Sea $\widetilde{U}_i$ la solución cuando se incrementan los valores de contorno en una milésima, entonces $\Delta_{hk}(\widetilde{U} - U) = 0$, con $(\widetilde{U} - U)\big|_{\partial\Omega} = 0'001$ en todos los nodos. La solución de este sistema es obviamente $\widetilde{U} - U = 0'001$, es decir, $\widetilde{U} = U + 0'001$.

3.1.3. Con la numeración indicada, el $i-$ésimo nodo no está al lado de la frontera si $i \neq (N-1)n$ (frontera derecha), $i \neq 1 + (N-1)n$ (frontera izquierda), $i \notin [1, N-1]$ (frontera superior) e $i \notin ((N-1)^2-(N-1), (N-1)^2]$ (frontera inferior). En ese caso la ecuación $i-$ésima resultante al aplicar la fórmula de los cinco puntos es

$$-N^2(U_{i-1} + U_{i+1} + U_{i-(N-1)} + U_{i+(N-1)} - 4U_i) = f_i,$$

ya que proviene del esquema



Esto quiere decir que para los i considerados, si $A = (a_{rs})$ es la matriz buscada,

$$a_{ii} = 4N^2, \quad a_{ii-1} = a_{ii+1} = a_{ii-(N-1)} = a_{ii+(N-1)} = -N^2.$$

Lo que se ajusta a la forma indicada en el enunciado.

Al considerar el resto de los valores de i y aplicar la fórmula de los cinco puntos, por la condición $u\big|_{\partial\Omega} = 0$, la única diferencia es que los nodos que caen en la frontera no aparecen. No es difícil convencerse de que éstos corresponden a los ceros entre los bloques que componen la matriz o a los límites de ésta. Veamos dos ejemplos (fronteras superior y derecha), los otros dos son completamente análogos.

Los i con $1 < i < N-1$ están en la frontera superior y la ecuación asociada a ellos es

$$-N^2(U_{i-1} + U_{i+1} + U_{i+(N-1)} - 4U_i) = f_i.$$

El término ausente con respecto al caso anterior, $U_{i-(N-1)}$, simplemente indica que la matriz comienza en la columna 1, mientras que $i-(N-1) < 0$. Por otro lado, los i con $i = (N-1)n$, $1 < n < N-1$, están en la frontera derecha y la ecuación correspondiente es

$$-N^2(U_{i-1} + U_{i-(N-1)} + U_{i+(N-1)} - 4U_i) = f_i.$$

El término ausuente esta vez, U_{i+1}, es responsable de la anulación de los elementos a_{ii+1} cuando $i = (N-1)n$, $1 < n < N-1$. Estos ceros están situados en la parte inferior derecha de cada bloque T.

3.1.4. a) Un sencillo cálculo prueba

$$\Delta\phi_{nm} = -\frac{\pi^2}{9}(n^2+m^2)\phi_{nm}.$$

De forma que, suponiendo la convergencia, $\Delta u + \lambda u = 0$ implica

$$\sum_{m,n=1}^{\infty} a_{nm}\big(\lambda - \frac{\pi^2}{9}(n^2+m^2)\big)\phi_{nm} = 0.$$

Por la unicidad de la representación, $a_{nm}(\lambda - \pi^2(n^2+m^2)/9) = 0$. Si $\lambda < 2\pi^2/9$ entonces $\lambda - \pi^2(n^2+m^2)/9 < 0$ y necesariamente $a_{nm} = 0$, por lo que $u = 0$. Por otro lado, para cualquier $C \in \mathbb{R}$, $C\phi_{11}$ es solución cuando $\lambda = 2\pi^2/9$.

b) Con la notación del problema anterior, con $N = 3$, todo lo que hay que hacer es hallar el menor λ para el que el siguiente sistema tenga infinitas soluciones (el factor extra $1/3^2$ se debe a que Q es $[0,3]^2$ y no $[0,1]^2$).

$$\frac{1}{3^2}9\begin{pmatrix} T & I \\ I & T \end{pmatrix}\vec{U} + \lambda\vec{U} = \vec{0}.$$

El λ buscado será, por tanto, el mayor autovalor de

$$-\begin{pmatrix} T & I \\ I & T \end{pmatrix} = \begin{pmatrix} 4 & -1 & -1 & 0 \\ -1 & 4 & 0 & -1 \\ -1 & 0 & 4 & -1 \\ 0 & -1 & -1 & 4 \end{pmatrix}$$

que después de hacer los cálculos es $\lambda = 2$. El resultado no está demasiado lejos de $2\pi^2/9 = 2'193\ldots$ que se obtenía en el problema inicial sin aproximar por diferencias finitas.

3.1.5. Como la normal en la frontera derecha apunta en la dirección de x, se cumple $\frac{\partial u}{\partial \mathbf{n}} = \frac{\partial u}{\partial x}$.

Sea U_{ij} la aproximación por diferencias finitas en el nodo $(x_i, y_j) = (i/3, j/3)$, $i = 1, 2, 3$; $j = 1, 2$.

En los nodos interiores (x_1, y_1), (x_1, y_2), (x_2, y_1), (x_2, y_2), la fórmula de los cinco puntos conduce a las ecuaciones

$$\begin{aligned} &U_{12} + U_{21} - 4U_{11}, && U_{11} + U_{31} + U_{22} - 4U_{21} = 0, \\ &U_{11} + U_{22} - 4U_{12}, && U_{12} + U_{32} + U_{21} - 4U_{22} = 0. \end{aligned}$$

Faltan dos ecuaciones para poder tener un sistema determinado. Éstas se pueden obtener aproximando por el cociente incremental la condición de frontera $\frac{\partial u}{\partial x} = -1$ en (x_3, y_1) y (x_3, y_2). Lo que da lugar a

$$\frac{U_{31} - U_{21}}{1/3} = -1, \qquad \frac{U_{32} - U_{22}}{1/3} = -1.$$

Resolviendo el sistema obtenido (lo cual puede simplificarse drásticamente usando la simetría del problema concluyendo $U_{i1} = U_{i2}$), se llega a

$$(U_{11}, U_{12}, U_{21}, U_{22}, U_{31}, U_{32}) = -\frac{1}{15}(1, 1, 3, 3, 8, 8).$$

3.1.6. a) Se tiene los desarrollos de Taylor

$$\begin{aligned} f(x+h) &= f(x) + f'(x)h + \frac{f''(x)}{2}h^2 + \frac{f'''(x)}{6}h^3 + O(h^4) \\ f(x-h) &= f(x) - f'(x)h + \frac{f''(x)}{2}h^2 - \frac{f'''(x)}{6}h^3 + O(h^4) \end{aligned}$$

Por tanto

$$\frac{f(x+h) + f(x-h) - 2f(x)}{h^2} = f''(x) + O(h^2).$$

Y de aquí es fácil obtener $\Delta_5 u - \Delta u = O(h^2)$ aproximando con esta fórmula cada una de las parciales segundas.

*b) Procediendo como en el apartado anterior pero ahora con $f(x) = u(x, y_0)$, se tiene

$$f(x_0 + h) + f(x_0 - h) = 2u + \frac{\partial^2 u}{\partial x^2} h^2 + \frac{1}{12} \frac{\partial^4 u}{\partial x^4} h^4 + O(h^6)$$

donde u y sus derivadas están evaluadas en (x_0, y_0). Procediendo de la misma forma con $f(y) = u(x_0, y)$, se llega a que la contribución de los puntos medios de los lados en el esquema cuadrado de la figura es

$$\begin{aligned} 4\big(u(x_0 + h, y_0) + u(x_0 - h, y_0) + u(x_0, y_0 + h) + u(x_0, y_0 + h)\big) \\ = 16u + 4\Delta u h^2 + \frac{1}{3}\big(\frac{\partial^4 u}{\partial x^4} + \frac{\partial^4 u}{\partial y^4}\big) h^4 + O(h^6). \end{aligned}$$

Al considerar $f(h) = u(x_0 + h, y_0 + h)$ se tiene,

$$f(h) + f(-h) = 2u + \sum \text{parc. segundas } h^2 + \frac{1}{12} \sum \text{parc. cuartas } h^4 + O(h^6).$$

Y al considerar $f(h) = u(x_0 + h, y_0 - h)$, se obtiene el mismo resultado salvo que las parciales segundas cruzadas tendrán un signo menos, y las derivadas cuartas también tendrán este signo en el caso en que se derive un número impar de veces con respecto a cada variable (es decir, una vez con respecto a una y tres veces con respecto a la otra). Teniendo en cuenta todos estos signos y sumando el resultado a la ecuación anterior se llega a

$$\begin{aligned} u(x_0 + h, y_0 + h) + u(x_0 - h, y_0 - h) + u(x_0 + h, y_0 - h) + u(x_0 - h, y_0 + h) \\ = 4u + 2\Delta u h^2 + \frac{1}{6}\big(\frac{\partial^4 u}{\partial x^4} + \frac{\partial^4 u}{\partial y^4} + 6\frac{\partial^4 u}{\partial x^2 \partial y^2}\big) h^4 + O(h^6). \end{aligned}$$

Y esto es la contribución de los vértices del esquema de la figura.

Sumando todas las contribuciones se tiene que

$$\begin{aligned} 6h^2 \Delta_9 u = & -20u + 16u + 4\Delta u h^2 + \frac{1}{3}\big(\frac{\partial^4 u}{\partial x^4} + \frac{\partial^4 u}{\partial y^4}\big) h^4 \\ & + 4u + 2\Delta u h^2 + \frac{1}{6}\big(\frac{\partial^4 u}{\partial x^4} + \frac{\partial^4 u}{\partial y^4} + 6\frac{\partial^4 u}{\partial x^2 \partial y^2}\big) h^4 + O(h^6) \\ = & 6\Delta u h^2 + \frac{1}{2}\big(\frac{\partial^4 u}{\partial x^4} + \frac{\partial^4 u}{\partial y^4} + 2\frac{\partial^4 u}{\partial x^2 \partial y^2}\big) h^4 + O(h^6). \end{aligned}$$

Éste es el resultado buscado ya que $\Delta(\Delta u)$ es la combinación de derivadas cuartas que aparece en el paréntesis.

c) Si u es solución de $\Delta u + \lambda u = 0$ entonces $\Delta(\Delta u) = -\lambda \Delta u$, y por el apartado anterior

$$\Delta_9 u - \Delta u = \frac{h^2}{12}(-\lambda)\Delta u + O(h^4).$$

Sustituyendo también $\Delta u = -\lambda u$, se tiene la igualdad esperada.

3.1.7. El desarrollo de Taylor hasta orden 2 de $f(x+h)$ y $f(x-h)$ con término de error, implica

$$f(x+h) + f(x-h) - 2f(x) = f''(x)h^2 + O(h^3 \sup |f'''|),$$

de modo que para polinomios de segundo grado, cualquiera que sea h

$$\frac{f(x+h) + f(x-h) - 2f(x)}{h^2} = f''(x).$$

Consecuentemente para $u(x,y) = (x - x^2)(y - y^2)$, que es un polinomio de segundo grado en cada variable, la fórmula de los cinco puntos es exacta. Al verificarse $\Delta_{hk} u = \Delta u$ en los nodos, la solución real coincide en ellos con la obtenida al aplicar el esquema de diferencias finitas.

Si Error$\approx C/N^p$ entonces $\log \text{Error} \approx \log C - p \log N$, de manera que los logaritmos de los valores de la tabla se deben ajustar a una recta de pendiente $-p$. Uno puede conseguir mejores resultados usando regresión lineal, pero una aproximación es

$$-p = \frac{\Delta y}{\Delta x} = \frac{\log(1'004 \cdot 10^{-3}) - \log(1'312 \cdot 10^{-3})}{\log 9 - \log 8} = -2'27\ldots$$

que sugiere que el orden es 2.

*__3.1.8__ Desarrollando por Taylor en (x_i, y_j)

$$u(x_{i+n-m}, y_j) = u(x_i + (n-m)h, y_j) = \sum_{r=0}^{2n+1} \frac{1}{r!} \frac{\partial^r u}{\partial x^r} (n-m)^r h^r + O(h^{2n+2}).$$

Procediendo de la misma manera con $u(x_i, y_{j+n-m})$ y sustituyendo en la fórmula del enunciado, basta probar que para toda función $f \in C^{2n+2}$ se

cumple

$$f^{(2n)}(t) = \frac{1}{h^{2n}} \sum_{m=0}^{2n} (-1)^m \binom{2n}{m} \sum_{r=0}^{2n+1} \frac{(n-m)^r h^r}{r!} f^{(r)}(t).$$

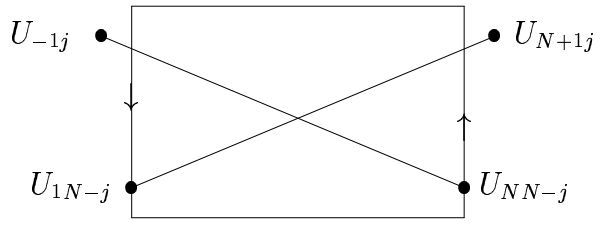
Nótese que $(-1)^m \binom{2n}{m} (n-m)^{2n+1}$ toma valores opuestos cuando $m = k$ y $m = 2n-k$, por tanto la contribución de los términos con $r = 2n+1$ es nula, y la identidad anterior equivale a

$$\sum_{m=0}^{2n} (-1)^m \binom{2n}{m} \frac{(n-m)^r}{r!} = \begin{cases} 0 & \text{si } 0 \le r < 2n \\ 1 & \text{si } r = 2n \end{cases}$$

Esto se sigue de la fórmula de la indicación derivando r veces, ya que $\operatorname{senh} t = t + O(t^3)$ implica $\operatorname{senh}^{2n} t = t^{2n} + O(t^{2n+2})$ de modo que

$$\left.\frac{d^r}{dt^r}\right|_{t=0} (\operatorname{senh}^{2n} t) = \begin{cases} 0 & \text{si } 0 \le r < 2n \\ (2n)! & \text{si } r = 2n \end{cases}$$

*__3.1.9__ Con las definiciones indicadas es como si la simétrica de la frontera izquierda estuviera a continuación de la derecha o viceversa



Geométricamente esta situación se puede representar entendiendo que en vez de resolver el problema en un rectángulo, se esta resolviendo en una banda de Möbius.

Con esta representación en mente, la prueba de la existencia y unicidad seguirá las mismas líneas que en el caso de un rectángulo, gracias a que se cumple de manera análoga el principio del máximo (la fórmula de los cinco puntos para $\Delta u = 0$ implica que el valor en cada nodo es promedio de los de los alrededores con la topología de la banda de Möbius). Las únicas fronteras que persisten son las de la parte de arriba y abajo, ya que las frontera izquierda y derecha han desaparecido al identificar.

3.1.10. Sea U_{ij}, como es habitual, la aproximación buscada de $u(x_i, t_j)$ con $x_i = ih$, $h = 1/N$, $t_j = jk$. Las condiciones de contorno se pueden aproximar como

$$U_{i0} = f(x_i), \qquad \frac{U_{i1} - U_{i0}}{k} = g(x_i), \qquad U_{0j} = U_{Nj} = 0.$$

Por otra parte, la ecuación se aproxima por

$$\frac{U_{ij+1} + U_{ij-1} - 2U_{ij}}{k^2} = \frac{U_{i+1j} + U_{i-1j} - 2U_{ij}}{h^2}.$$

Nótese que las condiciones de contorno dan inmediatamente los valores en las dos primeras "filas", U_{i0} y U_{i1}; mientras que la ecuación permite calcular los valores en la fila $i+1-$ésima, U_{i+1j}, en función de los valores en las dos filas anteriores. De esta forma, el método de diferencias finitas conduce a un sencillo método iterativo.

3.1.11. Por la regla de la cadena, con $x = r\cos\theta$, $y = r\,\mathrm{sen}\,\theta$, se obtiene

$$\frac{\partial u}{\partial r} = \cos\theta\,\frac{\partial u}{\partial x} + \mathrm{sen}\,\theta\,\frac{\partial u}{\partial y}, \qquad \frac{\partial u}{\partial \theta} = -r\,\mathrm{sen}\,\theta\,\frac{\partial u}{\partial x} + r\cos\theta\,\frac{\partial u}{\partial y}.$$

Despejando en estas fórmulas, para cualquier función v

$$\frac{\partial v}{\partial x} = \cos\theta\,\frac{\partial v}{\partial r} - \frac{\mathrm{sen}\,\theta}{r}\,\frac{\partial v}{\partial \theta}, \qquad \frac{\partial v}{\partial y} = \mathrm{sen}\,\theta\,\frac{\partial v}{\partial r} + \frac{\cos\theta}{r}\,\frac{\partial u}{\partial \theta}.$$

De modo que

$$\frac{\partial^2 u}{\partial x^2} = \frac{\partial}{\partial x}\left(\cos\theta\,\frac{\partial u}{\partial r} - \frac{\mathrm{sen}\,\theta}{r}\,\frac{\partial u}{\partial \theta}\right).$$

Utilizando de nuevo la fórmula para $\partial v/\partial x$ pero ahora sustituyendo v por la función entre paréntesis, se obtiene después de operar

$$\frac{\partial^2 u}{\partial x^2} = \mathrm{sen}^2\,\theta\,\frac{\partial^2 u}{\partial r^2} + \frac{\cos^2\theta}{r}\,\frac{\partial u}{\partial r} + \frac{\mathrm{sen}\,\theta\,\cos\theta}{r}\,\frac{\partial^2 u}{\partial r\partial\theta} + \frac{\cos^2\theta}{r^2}\,\frac{\partial^2 u}{\partial\theta^2} - \frac{\mathrm{sen}(2\theta)}{r^2}\,\frac{\partial u}{\partial\theta}.$$

El eje Y se obtiene a partir del eje X mediante un giro de ángulo $\pi/2$, por lo que la fórmula para $\partial^2 u/\partial y^2$ será igual salvo el cambio $\theta \mapsto \theta + \pi/2$. Sumando ambas expresiones se llega al resultado deseado.

Al discretizar de la manera indicada, los nodos serán en coordenadas polares (r, θ)

$$(r_1, \theta_1) = (3, \frac{\pi}{3}), \quad (r_2, \theta_2) = (3, \frac{\pi}{6}), \quad (r_3, \theta_3) = (2, \frac{\pi}{3}), \quad (r_4, \theta_4) = (2, \frac{\pi}{6}).$$

Sean U_1, U_2, U_3, U_4 las aproximaciones de los valores de u en estos nodos. Las condiciones de frontera implican $u(4, \theta) = 4$, $u(1, \theta) = 1$, $u(r, 0) = u(r, \pi/2) = r$. De manera que

$$\frac{\partial^2 u}{\partial r^2}(r_1, \theta_1) \approx \frac{u(4, \theta_1) + u(r_3, \theta_3) - 2u(r_1, \theta_1)}{1^2} \approx 4 + U_3 - 2U_1.$$

Y de la misma forma,

$$\frac{1}{r_1^2}\frac{\partial^2 u}{\partial \theta^2}(r_1, \theta_1) \approx \frac{U_2 + 3 - 2U_1}{3^2(\pi/3)^2}, \quad \frac{1}{r_1}\frac{\partial u}{\partial r}(r_1, \theta_1) \approx \frac{4 - U_1}{3 \cdot 1}.$$

La ecuación resultante es

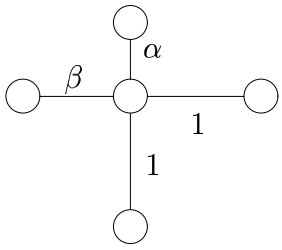
$$4 + U_3 - 2U_1 + \frac{4 - U_1}{3} + \frac{U_2 + 3 - 2U_1}{\pi^2/3} = 0.$$

Análogamente, en (r_3, θ_3) se obtiene

$$1 + U_1 - 2U_3 + \frac{U_1 - U_3}{2} + \frac{2 + U_4 - 2U_3}{2^2\pi^2/3^2} = 0.$$

Por la simetría del problema a través de la bisectriz del primer cuadrante ($\theta \mapsto \pi/2 - \theta$) se tiene $U_1 = U_2$ y $U_3 = U_4$; con estas igualdades no es necesario escribir las ecuaciones para los otros nodos (que se obtendría empleándolas en las ecuaciones ya obtenidas).

3.1.12. Los únicos nodos interiores son $(x_1, y_1) = (1, 1)$ y $(x_2, y_2) = (2, 1)$. al considerar la intersección del dominio con $x = 1$ e $y = 1$ se concluye que el nodo 1 está rodeado por el nodo 2 y los nodos de frontera $(1, \sqrt{2})$, $((3 - \sqrt{5})/2, 1)$ y $(1, 0)$. De forma que hay que crear una fórmula de cinco puntos en la que dos de los "brazos" midan 1 y los otros $\alpha = \sqrt{2} - 1$ y $\beta = (\sqrt{5} - 1)/2$.



Por Taylor

$$f(y+\alpha) \approx f(y) + f'(y)\alpha + f''(y)\alpha^2/2, \quad f(y-1) \approx f(y) - f'(y) + f''(y)/2.$$

Y por tanto

$$f''(y) \approx \frac{f(y+\alpha) + \alpha f(y-1) - (1+\alpha)f(y)}{(\alpha^2+\alpha)/2}.$$

De la misma forma

$$g''(x) \approx \frac{\beta g(x+1) + g(x-\beta) - (1+\beta)g(x)}{(\beta^2+\beta)/2}.$$

Por consiguiente, se obtiene la ecuación asociada al nodo 1

$$\frac{\beta U_2 + 0 - (1+\beta)U_1}{(\beta^2+\beta)/2} + \frac{0 + \alpha\cdot 2 - (1+\alpha)U_1}{(\alpha^2+\alpha)/2} = 0.$$

La ecuación en el segundo nodo se construiría de forma similar, pero la simetría del problema por la recta $x = 3/2$, asegura que $U_1 = U_2$ lo que lleva a $U_1 = U_2 = \sqrt{2} - 1$.

Ⓛ **3.1.13.** (Omitido).

Ⓛ **3.1.14.** (Omitido).

Ⓛ **3.1.15.** (Omitido).

4. Elementos finitos

4.1.1. Al integrar por partes la condición de Galerkin

$$\langle -u'' + u - 1, \phi_k \rangle = 0,$$

y susitutir $u = \sum \gamma_l \phi_l$, se llega en todos los casos a (nótese que $\phi_k(0) = \phi_k(1) = 0$)

$$\sum_l a_{kl}\gamma_l = b_k \quad \text{con} \quad a_{kl} = \int (\phi_k'\phi_l' + \phi_k\phi_l), \quad b_k = \int \phi_k$$

En cada apartado basta usar esta relación con las funciones que se indican.

a) La invariancia de la ecuación por traslaciones implica

$$a_{11} = a_{22} = \int ((\phi_1')^2 + (\phi_1)^2) = \int_{-1/3}^{1/3} 1\,dx + 2\int_0^{1/3} x^2 dx = \frac{56}{81}.$$

De la misma forma

$$a_{12} = a_{21} = -\int_{-1/3}^{1/3} 1\,dx + \int \phi\phi_1 = -\frac{2}{3} + \int_0^{1/3} (\frac{1}{3} - x)x\ dx = -\frac{107}{162}.$$

Calculando el área del triángulo limitado por ϕ, $b_1 = b_2 = 1/9$. De manera que hay que resolver el sistema

$$\begin{pmatrix} 56/81 & -53/162 \\ -53/162 & 56/81 \end{pmatrix} \begin{pmatrix} \gamma_1 \\ \gamma_2 \end{pmatrix} = \begin{pmatrix} 1/9 \\ 1/9 \end{pmatrix},$$

obteniéndose $\gamma_1 = \gamma_2 = 18/59$.

b) Es bien conocido del análisis de Fourier que en $[0,1]$ los senos y cosenos (por separado) con frecuencias dadas por múltiplos enteros de π son ortogonales, de modo que $a_{ll} = 0$ si $k \neq l$. Por otra parte

$$a_{ii} = \int_0^1 (i^2\pi^2 \cos^2(\pi i x) + \operatorname{sen}^2(\pi i x)) = \frac{i^2\pi^2 + 1}{2}, \qquad i = 1, 2, 3.$$

Además $b_1 = 2/\pi$, $b_2 = 0$, $b_3 = 2/(3\pi)$. Por tanto

$$\gamma_1 = \frac{4}{\pi(\pi^2+1)}, \qquad \gamma_2 = 0, \qquad \gamma_3 = \frac{4}{3\pi(9\pi^2+1)}.$$

c) En este caso no es posible aprovechar ninguna simetría y los cálculos son más extensos, obteniéndose

$$a_{11} = \int_0^1 \left((1-2x)^2 + x^2(1-x)^2\right) dx = \frac{11}{30}$$

$$a_{12} = \int_0^1 \left((1-2x)(2x-3x^2) + x^3(1-x)^2\right) dx = \frac{11}{60}$$

$$a_{22} = \int_0^1 \left((2x-3x^2)^2 + x^4(1-x)^2\right) dx = \frac{1}{7}$$

Por otra parte $b_1 = 1/6$, $b_2 = 1/12$. De modo que se llega al sistema

$$\begin{pmatrix} 11/30 & 11/60 \\ 11/60 & 1/7 \end{pmatrix} \begin{pmatrix} \gamma_1 \\ \gamma_2 \end{pmatrix} = \begin{pmatrix} 1/6 \\ 1/12 \end{pmatrix}$$

cuya solución es $\gamma_1 = 5/11$, $\gamma_2 = 0$.

4.1.2. Procediendo como antes, hay que calcular

$$a_{kl} = \int (\phi_k' \phi_l' + \phi_k \phi_l), \qquad b_k = \int x\phi_k$$

y después resolver $\sum a_{kl}\gamma_l = b_k$.

Está claro que $a_{11} = a_{22} = \frac{1}{2}a_{33}$ y que $a_{12} = a_{23} = a_{21} = a_{32}$, siendo el resto de los a_{kl} nulos. Del apartado a) del problema anterior se sigue $a_{11} = 56/9$ y $a_{12} = -53/18$. Con la función ϕ allí introducida

$$b_1 = \int_0^{2/3} x\phi_1 dx = \int_{-1/3}^{1/3} 3(u + \frac{1}{3})\phi \, dx = \int_{-1/3}^{1/3} \phi = \frac{1}{9}.$$

De la misma forma, $b_2 = 2/9$. Para b_3 se puede usar que $\phi_3(x) = 3x - 2$ en su soporte y se deduce $b_3 = 4/27$. En definitiva, el sistema resulante es

$$\begin{pmatrix} 56/9 & -53/18 & 0 \\ -53/18 & 56/9 & -53/18 \\ 0 & -53/18 & 28/9 \end{pmatrix} \begin{pmatrix} \gamma_1 \\ \gamma_2 \\ \gamma_3 \end{pmatrix} = \begin{pmatrix} 1/9 \\ 2/9 \\ 4/27 \end{pmatrix}$$

cuya solución aproximada es

$$(\gamma_1, \gamma_2, \gamma_3) = (0'11402\ldots, 0'20322\ldots, 0'23995\ldots).$$

4.1.3. Al integrar por partes ahora el término de frontera $u_k'(1)\phi_k(1)$ no se anula para $k = 3$, de modo que

$$b_3 = \phi_3(1) + \int_0^1 x\phi_3 = 1 + \frac{4}{27} = \frac{31}{27}.$$

Al resolver el sistema anterior con este cambio, la solución pasa a ser

$$(\gamma_1, \gamma_2, \gamma_3) = (1/3, 2/3, 1).$$

La solución aproximada $\widetilde{u} = \frac{1}{3}\phi_1 + \frac{2}{3}\phi_2 + \phi_3$ cumple $\widetilde{u}'(1^-) = \frac{2}{3}(-3) + 1 \cdot 3 = 1$. Es decir, satisface en este caso exactamente la condición de frontera. En el problema anterior, al hacer la misma comprobación se tiene

$$\widetilde{u}'(1^-) = (-3) \cdot 0'20322 \cdots + 3 \cdot 0'23995 \cdots = 0'1101 \ldots$$

4.1.4. La condición de Galerkin lleva a $A\vec{\gamma} = \vec{b}$ con

$$a_{kl} = \int_{-1}^{1} (\phi_k' \phi_l' + (x+2)\phi_k \phi_l) \quad \text{y} \quad b_1 = b_2 = b_3 = \frac{1}{2}.$$

Con cálculos sencillos se puede evaluar la parte de a_{kl}, obteniéndose

$$\int_{-1}^{1} (\phi_k' \phi_l' + 2\phi_k \phi_l) = \begin{cases} 14/3 & \text{si } k = l \\ -11/6 & \text{si } |k-l| = 1 \end{cases}$$

Basta añadir la parte correspondiente a $\int x\phi_k\phi_l$. Estos valores se dan en la indicación, excepto $\int x\phi_2^2$ que es trivialmente nulo. De esta forma

$$A = \begin{pmatrix} 14/3 & -11/6 & 0 \\ -11/6 & 14/3 & -11/6 \\ 0 & -11/6 & 14/3 \end{pmatrix} + \begin{pmatrix} -1/6 & -1/48 & 0 \\ -1/48 & 0 & 1/48 \\ 0 & 1/48 & 1/6 \end{pmatrix}.$$

Es decir

$$A = \begin{pmatrix} 9/2 & -89/48 & 0 \\ -89/48 & 14/3 & -29/16 \\ 0 & -29/16 & 29/6 \end{pmatrix}$$

Al resolver el sistema se obtiene

$$(\gamma_1, \gamma_2, \gamma_3) = (0'22534 \ldots, 0'27723 \ldots, 0'20741 \ldots).$$

4.1.5. Los elementos de la matriz de rigidez vienen dados por

$$a_{kl} = \int (\phi_k' \phi_l' + \phi_k \phi_l).$$

Como la última función base está cortada por la mitad,

$$a_{11} = a_{22} = \cdots = \frac{1}{2} a_{NN}.$$

Además

$$a_{12} = a_{21} = a_{23} = a_{32} = \cdots = a_{N-1N} = a_{NN-1}.$$

El resto de los elementos son nulos. Estas cantidades son sencillas de calcular directamente o con un cambio de variable. El resultado es

$$a_{11} = \int_0^{2/N} N^2 dx + 2\int_0^{1/N} N^2 x^2 dx = 2N + \frac{2}{3N}$$
$$a_{12} = \int_{1/N}^{2/N} -N \cdot N \; dx + 2\int_0^{1/N} Nx(1-Nx)\; dx = -N + \frac{1}{6N}.$$

4.1.6. Se puede asociar a cada función base el nodo que está en el punto medio de su soporte: $x_j = 1+j/4$; $j = 1, 2, 3$. Para que haya quince funciones base se deberían considerar los nodos $x_j = 1+j/16$; $j = 1, 2, \ldots, 15$. Es decir, la función base $\widetilde{\phi}_j$ debe ser ϕ ajustada al intervalo $[x_{j-1}, x_{j+1}]$. La homotecia combinada con una traslación que pasa $[x_{j-1}, x_{j+1}] = [(j+15)/16, (j+17)/16]$ a $[1, 3/2]$ es $x \mapsto 4x - (11+j)/4$, lo cual prueba $\widetilde{\phi}_j(x) = \phi(4x - \frac{11+j}{4})$.

La condición de Galerkin en este caso implica

$$\widetilde{a}_{kl} = \int x\widetilde{\phi}'_k\widetilde{\phi}'_l.$$

Sustituyendo y cambiando la variable

$$\widetilde{a}_{kl} = \int x\phi'\big(4x - \frac{11+k}{4}\big)\; 4\phi'\big(4x - \frac{11+l}{4}\big)\; dx$$
$$= \int \big(u + \frac{11+k}{4}\big)\phi'(u)\phi'\big(u - \frac{k-l}{4}\big)\; du.$$

Los únicos elementos no nulos cuando $l \geq k$ son

$$\widetilde{a}_{kk} = \int u(\phi'_1)^2\; du + \frac{11+k}{4}\int (\phi'_1)^2\; du$$
$$\widetilde{a}_{kk+1} = \int u\phi'_1\phi'_2\; du + \frac{11+k}{4}\int \phi'_1\phi'_2\; du$$

donde $\phi_1(x) = \phi(x)$ y $\phi_2(x) = \phi(x - 1/4)$.

La matriz de rigidez del enunciado implica

$$\int u(\phi'_1)^2\; du = \frac{239}{48} \qquad \text{y} \qquad \int u\phi'_1\phi'_2\; du = -\frac{17}{12}.$$

Para hallar las integrales restantes se puede emplear un cambio de variable

$$\frac{97}{16} = \int u(\phi_1')^2\, du = \int (x+\frac{1}{4})(\phi_1')^2\, dx = \frac{239}{48} + \frac{1}{4}\int (\phi_1')^2$$
$$-\frac{5}{3} = \int u\phi_2'\phi_3'\, du = \int (x+\frac{1}{4})\phi_1'\phi_2'\, dx = -\frac{17}{12} + \frac{1}{4}\int \phi_1'\phi_2'.$$

Despejando, $\int(\phi_1')^2 = 13/12$ y $\int \phi_1'\phi_2' = -1$. Y sustituyendo, se obtiene finalmente

$$\widetilde{a}_{kk} = \frac{239}{48} + \frac{11+k}{4}\cdot\frac{13}{12} = \frac{382+13k}{48}$$
$$\widetilde{a}_{kk+1} = \widetilde{a}_{k+1k} = -\frac{17}{12} + \frac{11+k}{4}\cdot(-1) = -\frac{50+3k}{12}.$$

El resto de los elemntos de la matriz de rigidez son nulos.

4.1.7. Al aplicar el método de diferencias finitas el sistema que hay que resolver es

$$-\frac{U_{n+1}+U_{n-1}-2U_n}{1/N^2} = f_n$$

con $n = 1, 2, \ldots, N-1$ y definiendo $U_0 = U_{N-1} = 0$. La matriz correspondiente es tridiagonal simétrica, con $a_{ii} = 2N^2$ y $a_{ii+1} = -N^2$.

Por otra parte al aplicar el método de elementos finitos, en este caso $a_{kl} = \int \phi_k'\phi_l'$ de modo que la matriz vuelve a ser tridiagonal simétrica con

$$a_{ii} = \int_0^{2/N} N^2\, dx = 2N, \qquad a_{ii+1} = \int_{1/N}^{2/N} -N^2\, dx = -N,$$

que coincide con lo anterior salvo un factor N.

4.1.8. Con la notación del problema anterior $f_n = C$ y el vector de carga tiene de coordenadas $b_i = \int C\phi_i = C/N$. Por tanto los sistemas obtenidos al aplicar los métodos de diferencias finitas y de elementos finitos son iguales (salvo simplificar un factor N) y se tiene $\gamma_n = U_n$. Por otra parte, la solución exacta es $u(x) = -Cx(1-x)/2$, un polinomio de segundo grado, de modo que la fórmula siguiente es exacta

$$u''(x_n) = \frac{u(x_{n+1}) + u(x_{n-1}) - 2u(x_n)}{1/N^2},$$

y, en consecuencia, al aplicar el método de diferencias finitas $U_n = u(x_n)$. Evidentemente,

$$\sum \gamma_l \phi_l(x_n) = \sum U_l \phi_l(x_n) = U_n \cdot 1 = u(x_n).$$

4.1.9. Escribiendo $\phi_3(x) = \phi_1(x-1)$, los elementos de la matriz de rigidez y el vector de carga vienen dados por

$$a_{kl} = \int (\phi_k' \phi_l' + \phi_k \phi_l) \qquad \text{y} \qquad b_k = \int (x^2 - 2x - 2)\phi_k.$$

Efectuando los cálculos

$$\begin{aligned} a_{11} &= \int_0^1 ((4-8x)^2 + \phi_1^2)\, dx = \frac{16}{3} + \frac{8}{15} = \frac{88}{15} \\ a_{12} &= \int_0^1 ((4-8x)(4x-1) + \phi_1\phi_2)\, dx = -\frac{8}{3} + \frac{1}{15} = -\frac{13}{5} \\ a_{22} &= 2\int_0^1 ((4x-1)^2 + \phi_2^2)\, dx = \frac{14}{3} + \frac{4}{15} = \frac{74}{15}. \end{aligned}$$

Además $a_{23} = a_{12} = a_{21} = a_{32}$ y $a_{13} = a_{31} = 0$. Las componentes del vector de carga son

$$\begin{aligned} b_1 &= \int_0^1 4x(1-x)(x^2-2x-2)\, dx = -\frac{9}{5} \\ b_2 &= 2\int_0^1 x(2x-1)(x^2-2x-2)\, dx = -\frac{31}{30}, \qquad b_3 = b_1. \end{aligned}$$

Donde se ha usado la simetría por $x = 1$ de $x^2 - 2x - 2$.

Al resolver el sistema $A\vec{\gamma} = \vec{b}$ se obtiene $\gamma_1 = \gamma_3 = -3/4$, $\gamma_2 = -1$. Para $x \in [0,1]$

$$\sum \gamma_l \phi_l(x) = -\frac{3}{4} 4x(1-x) - x(2x-1) = x^2 - 2x.$$

Por la simetría, esta fórmula tiene validez en todo $[0,2]$. La función $u(x) = x^2 - 2x$ es evidentemente la solución exacta.

Ⓛ **4.1.10.** (Omitido).

Ⓛ **4.1.11.** (Omitido).

4.2.1. Esta ecuación no es más que la formulación débil del problema de contorno

$$-u'' + 4u - 1 = 0, \qquad u(0) = u(1/2) = 0.$$

Para resolverlo, nótese que $-u'' + 4u = 0 \Rightarrow u = Ae^{2x} + Be^{-2x}$, y que una solución particular de la no homogénea es $u = 1/4$. De modo que basta ajustar A y B en

$$u(x) = \frac{1}{4} + Ae^{2x} + Be^{-2x}$$

para que se cumplan las condiciones de contorno. El resultado es $1/A = -4(1+e)$, $1/B = -4(1+e^{-1})$.

4.2.2. Integrando por partes, la formulación débil es

$$a(u,v) = L(v) \qquad \text{con} \qquad a(u,v) = \int (xu'v' + 2uv), \qquad L(v) = v$$

donde $u, v \in H^1$ se "anulan" en 1. Para efectuar la integración por partes es cómodo notar que $-xu'' - u' = -(xu')'$. Para aplicar el método de elementos finitos hay que calcular

$$\begin{aligned}
a(\phi_1,\phi_1) &= \int_1^3 x \cdot 1^2 \, dx + 4\int_1^2 (x-1)^2 \, dx = \frac{16}{3} \\
a(\phi_1,\phi_2) &= \int_2^3 x \cdot (-1) \cdot 1 \, dx + 2\int_2^3 (3-x)(x-2) \, dx = -\frac{13}{6} \\
a(\phi_2,\phi_2) &= \int_2^3 (x \cdot 1 \cdot 1 + 2(2-x)^2) \, dx = \frac{19}{6}.
\end{aligned}$$

Por otra parte $L(\phi_1) = 1$, $L(\phi_2) = 1/2$. Así pues hay que resolver

$$\begin{pmatrix} 16/3 & -13/6 \\ -13/6 & 19/6 \end{pmatrix} \begin{pmatrix} \gamma_1 \\ \gamma_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 1/2 \end{pmatrix}$$

cuya solución es $\gamma_1 = 153/439$, $\gamma_2 = 174/439$.

4.2.3. Por la regla del producto

$$\operatorname{div}(vA(x)\nabla u) = \sum_{i=1}^{3} \left(\frac{\partial v}{\partial x_i} \, (A(x)\nabla u)_i + v \frac{\partial}{\partial x_i} (A(x)\nabla u)_i \right).$$

Donde el subíndide i indica la coordenada $i-$ésima.

De forma que

$$\operatorname{div}(vA(x)\nabla u) = (\nabla v)^t A(x)\nabla u + v \operatorname{div}(A(x)\nabla u).$$

Por esta igualdad y el teorema de la divergencia

$$-\int_\Omega v \operatorname{div}(A(x)\nabla u) = \int_\Omega (\nabla v)^t A(x)\nabla u + \int_{\partial\Omega} vA(x)\nabla u.$$

Para imponer las condiciones de contorno se procede igual que con el laplaciano. Por ejemplo, bajo condiciones homogéneas de Dirichlet o de Neumann, esto es, $u\big|_{\partial\Omega} = 0$ o $\frac{\partial u}{\partial \mathbf{n}}\big|_{\partial\Omega} = 0$; (trabajando en $H_0^1(\Omega)$ o $H^1(\Omega)$) se tiene

$$a(u,v) = \int_\Omega (\nabla v)^t A(x)\nabla u \qquad \text{y} \qquad L(v) = \int_\Omega fv.$$

Las condiciones de Neumann no homogéneas añadirían un término de frontera a $L(v)$, como en el caso del laplaciano.

4.2.4. Cambiando de signo e integrando por partes, se tiene $a(u,v) = \int u'v'$, $L(v) = -2\int v$, y la formulación variacional afirma que la expresión

$$\frac{1}{2}a(u,u) - L(u) = \frac{1}{2}\int (u')^2 + 2\int u$$

alcanza un mínimo para $u = x^2 + 1$. Es decir,

$$\frac{1}{2}\int_{-1}^{1} \left((u')^2 + 4u\right) \geq \frac{1}{2}\int_{-1}^{1} \left((2x)^2 + 4(x^2-1)\right) dx.$$

Al operar se obtiene la desigualdad deseada.

*__4.2.5__ Si u es una función como en el enunciado, sea $w = u - x^2 + 1$, entonces

$$\begin{aligned}\int_{-1}^{1} \left((u')^2 + 4u\right) &= \int_{-1}^{1} \left((w')^2 + 4xw' + 4w + 5x^2 - 4\right)\, dx \\ &= \frac{8}{3} + \int_{-1}^{1} \left((w')^2 + 4(xw)'\right) \\ &\geq \frac{8}{3} + \int_{-1}^{1} (xw)' = \frac{8}{3} + 4w(1) + 4w(-1) = \frac{8}{3}.\end{aligned}$$

4.2.6. Derivando

$$-(\tilde{a}(x)u')' = -\tilde{a}(x)u'' - \tilde{a}'(x)u',$$

de forma que basta probar que existe $\tilde{a}$ tal que $(-\tilde{a}, -\tilde{a}')$ y $(-a, b)$ son proporcionales, esto es,

$$0 = \begin{vmatrix} -\tilde{a} & -a \\ -\tilde{a}' & b \end{vmatrix} = \tilde{a}'a - \tilde{a}b.$$

Lo cual lleva a escoger $\tilde{a} = e^{-\int b/a}$, y la función por la que hay que multiplicar es $\tilde{a}/a$.

Integrando por partes se llega a la formulación débil $a(u, v) = L(v)$ con

$$a(u, v) = \int (\tilde{a}u'v' + \tilde{c}uv) \qquad \text{y} \qquad L(v) = \int \tilde{f}v.$$

4.2.7. Siguiendo la indicación

$$\Delta(\Delta u) = \Delta\left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\right) = \frac{\partial^4 u}{\partial x^4} + 2\frac{\partial^4 u}{\partial x^2 \partial y^2} + \frac{\partial^4 u}{\partial y^4}.$$

De manera que la ecuación del problema es $\Delta(\Delta u) = f$. Multiplicando por v y aplicando la identidad de Green se deduce

$$\int_\Omega v\ \Delta(\Delta u) = -\int_\Omega \nabla u \cdot \nabla(\Delta u) + \int_{\partial\Omega} v\nabla(\Delta u) = -\int_\Omega \nabla(\Delta u) \cdot \nabla v.$$

Aplicando de nuevo la identidad de Green a la última expresión, pero ahora para despejar $\int \nabla \cdot \nabla$, se tiene

$$\int_\Omega v\ \Delta(\Delta u) = \int_\Omega \Delta u \Delta v - \int_{\partial\Omega} \Delta u \nabla v = \int_\Omega \Delta u \Delta v.$$

El término de frontera se anula porque $\nabla v \cdot \mathbf{n} = \partial v/\partial \mathbf{n} = 0$.

4.2.8. La derivada débil de f_n es

$$f_n'(x) = \begin{cases} n & \text{si } x \in [-\frac{1}{2n}, \frac{1}{2n}] \\ 0 & \text{si } x \in [-1,1] - [-\frac{1}{2n}, \frac{1}{2n}] \end{cases}$$

Para comprobarlo rigurosamente con la definición, basta notar que para cualquier función $\phi \in C_0^1$, se tiene

$$\int f_n \phi' = \int_{-1}^{-1/2n} \frac{1}{2}\phi' + \int_{1/2n}^{1} \frac{1}{2}\phi' + n\int_{-1/2n}^{1/2n} x\phi' = -n\int_{-1/2n}^{1/2n} \phi,$$

donde se ha integrado por partes el último sumando. Con ello $\int f_n\phi' = -\int f_n'\phi$ y f_n' es realmente la derivada débil.

Es evidente que $\|f_n'\|_1 = 1$, así que f_n' está uniformemente en L^1.

Por el teorema del valor medio para integrales

$$\int_{-1/2n}^{1/2n} g = \frac{1}{n} g(\xi_n) \qquad \text{con} \qquad \xi_n \in [-\frac{1}{2n}, \frac{1}{2n}].$$

Por tanto

$$\lim_{n\to\infty} \int_{-1}^{1} f_n' g = \lim_{n\to\infty} n \int_{-1/2n}^{1/2n} g = \lim_{n\to\infty} n \cdot \frac{1}{n}\, g(\xi_n) = g(0).$$

Si existiera $F \in L^1$ que fuera límite de f_n', se tendría $\int Fg = g(0)$ para cualquier $g \in C([-1,1])$. En particular $\int F = 1$ y por ejemplo

$$\int F(x)\,(1 - e^{-nx^2})\,dx = 0.$$

Pero tomando límites, esto contradice el teorema de la convergencia dominada porque $1 - e^{-nx^2} \to 1$ en casi todo punto.

Ⓛ **4.2.9.** (Omitido).

Ⓛ **4.2.10.** (Omitido).

4.3.1. La función $\phi_1 + \phi_2 + \phi_3 : \mathbb{R}^2 \longrightarrow \mathbb{R}$ es lineal en K y en cada uno de sus tres vértices vale 1, por tanto debe ser constantemente 1.

4.3.2. Sea la numeración de los nodos como sigue:

$$1 \to (-1,0), \qquad 2 \to (0,-1), \qquad 3 \to (1,0), \qquad 4 \to (0,1).$$

La formulación débil lleva a

$$a(u,v) = \int_\Omega (\nabla u \cdot \nabla v + uv) \qquad \text{y} \qquad L(v) = \int_\Omega v.$$

Como u se anula en los nodos 3 y 4 la solución por el método de elementos finitos será de la forma $u = \gamma_1\phi_1 + \gamma_2\phi_2$ y todo lo que hay que hacer es resolver el sistema

$$\begin{pmatrix} a(\phi_1,\phi_1) & a(\phi_1,\phi_2) \\ a(\phi_2,\phi_1) & a(\phi_2,\phi_2) \end{pmatrix} \begin{pmatrix} \gamma_1 \\ \gamma_2 \end{pmatrix} = \begin{pmatrix} L(\phi_1) \\ L(\phi_2) \end{pmatrix}.$$

Para hacer los cálculos, sea K el triángulo 1-2-4. Como $\phi_1 : K \longrightarrow \mathbb{R}$ es una función lineal que vale 1 en el nodo 1 y 0 en los otros dos, se tiene $\phi_1(x,y) = -x$ en K. De la misma forma $\phi_4(x,y) = (1+x-y)/2$ en K. Con ello,

$$\begin{aligned} a(\phi_1,\phi_1) &= \int_K (1+x^2)\, dx\, dy = 1 + \int_{-1}^{0}\int_{-x-1}^{x+1} x^2\, dx\, dy = \frac{7}{6} \\ a(\phi_1,\phi_2) &= -\frac{1}{2} - \int_{-1}^{0}\int_{-x-1}^{x+1} \frac{x}{2}(1+x-y)\, dx\, dy = -\frac{5}{12} \\ a(\phi_2,\phi_2) &= 1 + \int_{-1}^{0}\int_{-x-1}^{x+1} \frac{1}{2}(1+x-y)^2\, dx\, dy = \frac{4}{3}. \end{aligned}$$

Por otra parte, con la fórmula para el volumen del tetraedro $L(\phi_1) = 1/3$, $L(\phi_2) = 2/3$. Al sustituir en el sistema y resolverlo, se tiene $\gamma_1 = 104/199$, $\gamma_2 = 132/199$.

Otra forma de hacer el problema es utilizando las matrices básicas. Tras unos cálculos tediosos, se puede obtener que las matrices que corresponden al triángulo de la derecha y de la izquierda, numerados como 2-3-4 y 1-2-4, son respectivamente

$$\frac{1}{12}\begin{pmatrix} 8 & -5 & 1 \\ -5 & 14 & -5 \\ 1 & -5 & 8 \end{pmatrix} \qquad \text{y} \qquad \frac{1}{12}\begin{pmatrix} 14 & -5 & -5 \\ -5 & 8 & 1 \\ -5 & 1 & 8 \end{pmatrix}.$$

Al ensamblar la matriz de rigidez, las dos últimas filas y columnas se pueden despreciar por la condición de Dirichlet, y las otras dos dan lugar a la matriz 2×2 del sistema anteriormente resuelto.

4.3.3. Considérese la numeración de los nodos

$$1 \to (1,1), \qquad 2 \to (2,2), \qquad 3 \to (0,2), \qquad 4 \to (0,0), \qquad 5 \to (2,0).$$

La condición $u(0,y) = u(x,0) = 0$ implica que sólo las funciones asociadas a los dos primeros nodos participan en la solución. Es decir, $u = \gamma_1\phi_1 + \gamma_2\phi_2$.

La formulación débil tiene como forma bilineal y lineal a

$$a(u,v) = \int_\Omega \nabla u \cdot \nabla v \qquad \text{y} \qquad L(v) = \int_0^2 v(x,2)\, dx.$$

Sea K el triángulo 1-5-2. En él $\phi_1(x,y) = 2 - x$ y $\phi_2(x,y) = (x+y-2)/2$. Con ello

$$\begin{aligned} a(\phi_1,\phi_1) =& 4 \int_K \|\nabla\phi_1\|^2 = 4 \\ a(\phi_1,\phi_2) =& 2 \int_K (-1,0)\cdot(\frac{1}{2},\frac{1}{2}) = -1 \\ a(\phi_2,\phi_2) =& 2 \int_K \|\nabla\phi_2\|^2 = 1 \end{aligned}$$

Obviamente $L(\phi_1) = 0$ y $L(\phi_2) = \int_0^2 x/2\, dx = 1$. Por tanto hay que resolver el sistema

$$\begin{pmatrix} 4 & -1 \\ -1 & 1 \end{pmatrix} \begin{pmatrix} \gamma_1 \\ \gamma_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}.$$

La solución es $\gamma_1 = 1/3$, $\gamma_2 = 4/3$.

4.3.4. Con la numeración de los nodos del problema anterior, se tiene que las condiciones de Dirichlet implican $\gamma_2 = \gamma_4 = 0$ y $\gamma_3 = 2$ (nótese que $u(0,2) = 2$). De modo que sólo γ_1 es una incógnita. La ecuación asociada al primer nodo es

$$a(\phi_1,\phi_1)\gamma_1 = L(\phi_1) - a(\phi_1,\phi_3)\gamma_3.$$

Según los cálculos del problema anterior $a(\phi_1,\phi_1) = 4$, y por la simetría del dibujo $a(\phi_1,\phi_3) = a(\phi_1,\phi_2) = -1$. Además $L \equiv 0$, con lo cual $\gamma_1 = 1/2$.

4.3.5. Las funciones base en Ω responden a las fórmulas

$$\phi_1 = (1-x)(1-y), \qquad \phi_2 = x(1-y), \qquad \phi_3 = xy, \qquad \phi_4 = (1-x)y.$$

Con la numeración de los vértices de Ω

$$1 \to (0,0) \qquad 2 \to (1,0) \qquad 3 \to (1,1), \qquad 4 \to (0,1);$$

se tiene que ϕ_i vale 1 en el vértice $i-$ésimo y cero en el resto.

La matriz de rigidez y el vector de carga vendrán dados por

$$a_{kl} = \int_\Omega (\nabla\phi_k \cdot \nabla\phi_l + \phi_k\phi_l), \qquad b_k = \int_\Omega xy\phi_k.$$

Las funciones base se obtienen unas de otras girando el cuadrado Ω, lo que establece las simetrías

$$a_{11} = a_{22} = a_{33} = a_{44}, \qquad a_{12} = a_{23} = a_{34} = a_{41}, \qquad a_{13} = a_{24}.$$

El cálculo de estas cantidades es sencillo:

$$a_{11} = \int_\Omega \left((1-y)^2 + (1-x)^2\right)\, dx\, dy + \int_\Omega (1-x)^2(1-y)^2\, dx\, dy = \frac{7}{9}$$

$$a_{12} = \int_\Omega \left(-(y-1)^2 + x(1-x)\right)\, dx\, dy + \int_\Omega x(1-x)(1-y)^2\, dx\, dy = -\frac{1}{9}$$

$$a_{13} = \int_\Omega \left((y-1)y + (x-1)x\right)\, dx\, dy + \int_\Omega xy(1-x)(1-y)\, dx\, dy = -\frac{11}{36}$$

Por otra parte

$$b_1 = \int_\Omega xy\phi_1 = \int_0^1\int_0^1 xy(1-x)(1-y)\, dx\, dy = \frac{1}{36}$$

$$b_2 = \int_\Omega xy\phi_2 = \int_0^1\int_0^1 x^2y(1-y)\, dx\, dy = \frac{1}{18}$$

$$b_3 = \int_\Omega xy\phi_3 = \int_0^1\int_0^1 x^2y^2\, dx\, dy = \frac{1}{9}, \qquad b_4 = b_2.$$

Y la solución de

$$\begin{pmatrix} 7/9 & -1/9 & -11/36 & -1/9 \\ -1/9 & 7/9 & -1/9 & -11/36 \\ -11/36 & -1/9 & 7/9 & -1/9 \\ -1/9 & -11/36 & -1/9 & 7/9 \end{pmatrix} \begin{pmatrix} \gamma_1 \\ \gamma_2 \\ \gamma_3 \\ \gamma_4 \end{pmatrix} = \begin{pmatrix} 1/36 \\ 1/18 \\ 1/9 \\ 1/18 \end{pmatrix}$$

es $\gamma_1 = 0'22153\ldots$, $\gamma_3 = 0'29846\ldots$, $\gamma_2 = \gamma_4 = 0'24$.

4.3.6. La condición de Galerkin es

$$\int_0^1 \Big(-\frac{\partial^2 u}{\partial x^2} - f + \frac{\partial u}{\partial t}\Big)\phi_k(x)\, dx = 0.$$

E integrando por partes el término que contiene la parcial segunda

$$\int_0^1 \left(\frac{\partial u}{\partial x}\phi_k' + \frac{\partial u}{\partial t}\phi_k\right) dx = \int f\phi_k\, dx.$$

Al sustituir $u(x,t) = \sum \gamma_l(t)\phi_l(x)$ se llega a

$$\sum_l \gamma_l \int_0^1 \phi_l'\phi_k' + \sum_l \gamma_l' \int_0^1 \phi_l\phi_k = \int f\phi_k.$$

De forma que las matrices A_1 y A_2 tienen como elementos

$$a_{kl}^1 = \int_0^1 \phi_l\phi_k \qquad \text{y} \qquad a_{kl}^2 = \int_0^1 \phi_l'\phi_k'$$

respectivamente.

Unos cálculos muy sencillos permiten deducir que en el ejemplo que se indica

$$A_1 = \begin{pmatrix} 2/9 & 1/18 \\ 1/18 & 2/9 \end{pmatrix} \qquad \text{y} \qquad A_2 = \begin{pmatrix} 6 & -3 \\ -3 & 6 \end{pmatrix}.$$

Ⓛ **4.3.7.** (Omitido).

Ⓛ **4.3.8.** (Omitido).

Ⓛ **4.3.9.** (Omitido).

4.4.1. La distancia de g a V será $\|g-u\|$ donde u es la proyección ortogonal de g en V. Para calcularla hay que que imponer

$$\langle g-u, \operatorname{sen} x\rangle = \langle g-u, \operatorname{sen}(2x)\rangle = \langle g-u, \operatorname{sen}(3x)\rangle = 0$$

con $u = \lambda \operatorname{sen} x + \mu \operatorname{sen}(2x) + \nu \operatorname{sen}(3x)$, porque $u \in V$. Empleando la ortogonalidad de $\operatorname{sen}(jx)$ en $[0, \pi]$, esto implica

$$\lambda = \frac{2}{\pi} \int_0^\pi g \operatorname{sen} x \, dx = \frac{8}{\pi}, \qquad \mu = \frac{2}{\pi} \int_0^\pi g \operatorname{sen}(2x) \, dx = 0,$$
$$\nu = \frac{2}{\pi} \int_0^\pi g \operatorname{sen}(3x) \, dx = \frac{8}{27\pi}.$$

Y la distancia deseada es

$$d = \left(\int_0^1 (g - \lambda \operatorname{sen} x - \nu \operatorname{sen}(3x))^2 \, dx \right)^{1/2}$$
$$= \left(\int_0^1 g^2 - \lambda^2 \int_0^1 \operatorname{sen}^2 x \, dx - \nu^2 \int_0^1 \operatorname{sen}^2(3x) \, dx \right)^{1/2},$$

donde se ha usado de nuevo la ortogonalidad. Sustituyendo,

$$d = \left(\frac{\pi^5}{30} - \frac{32}{\pi} - \frac{32}{729\pi} \right)^{1/2} = 0'027701\ldots$$

4.4.2. Sean ϕ_3, ϕ_5, ϕ_7,... los trasladados en una, dos, tres,... unidades de ϕ_1; y sean ϕ_2, ϕ_4, ϕ_6,... lo mismo para ϕ_2. Dada $f \in Q_{[0,N]}$ sea

$$F(x) = \sum_{l=1}^{2N-1} f(l/2) \phi_l(x).$$

Obviamente $F \in Q_{[0,N]}$. Además es fácil comprobar que $F(n/2) = f(n/2)$ para $n/2 \in [0, N]$, $n \in \mathbb{Z}$ (porque $\phi_1(1/2) = \phi_2(1) = 1$, $\phi_1(0) = \phi_1(1) = \phi_2(1/2) = 0$). Así que en cada intervalo $[j, j+1]$, F y f son funciones cuadráticas que toman el mismo valor en los puntos j, $(j+1)/2$ y $j+1$. Por tanto deben coincidir.

4.4.3. Una vez que el problema anterior asegura que las funciones base generan todas las funciones cuadráticas a trozos de $Q_{[0,N]}$, basta comprobar que la solución $u \in Q_{[0,10]}$, ya que por el lema de Céa la aproximación será óptima.

Por razones de regularidad, cabe esperar una solución que no sólo sea cuadrática a trozos, sino que una fórmula del tipo $u = Ax^2 + Bx + C$ sea válida en todo su dominio de definición. Sustituyendo se tiene

$$-u'' + (x+1)u = -2A + Ax^3 + Bx^2 + Cx + Ax^2 + Bx + C.$$

Identificando coeficientes se obtiene $u = x^2 + 1$ que claramente satisface las condiciones de contorno.

Resúmenes de Teoría

Estos resúmenes no sustituyen a los apuntes del curso pero sí conforman una referencia completa de todos los temas tratados. Su propósito es que sirvan de ayuda al estudio en una asignatura tan extensa como ésta. Para ello en cada sección se indican escuetamente las ideas principales (las palabras clave están señaladas con letra inclinada). A continuación se añaden algunos comentarios acerca de la importancia relativa de los resultados y sus relaciones, así como indicaciones acerca de las fórmulas que es conveniente memorizar. Al final de cada capítulo se incluye una lista con los números de los problemas más representativos de las colecciones distribuidas a lo largo del curso. En el último capítulo se ha omitido por ser suficiente con los ejemplos vistos en clase.

1 Métodos de un paso

1.1 Introducción

Ideas y resultados básicos:

- Los problemas $y' = f(x, y)$ con f bajo una débil condición de regularidad tienen solución única en un pequeño entorno una vez especificado el valor inicial $y(a)$.
- Para resolver numéricamente estos problemas se considera habitualmente una *discretización* $a = x_0 < x_1 < x_2 < \cdots < x_N = b$ de los valores de x y se aplican métodos que aproximan $y(x_n)$ por cierto valor y_n.

Comentarios:

Esta sección tiene un propósito auxiliar y preliminar. Lo más concreto que se podría precisar es que la "débil condición de regularidad" en la idea anterior es Lipschitz en la segunda variable y continua en ambas.

1.2 Método de Euler

Ideas y resultados básicos:

- El *método de Euler* es el más sencillo de los de un paso. Se reduce simplemente a aproximar la derivada por el cociente incremental. Es decir, se rige por la fórmula $y_{n+1} = y_n + hf(x_n, y_n)$.
- Se dice que un método es *convergente* si a base de tomar la discretización suficientemente fina ($h = x_{i+1} - x_i \to 0$) y el valor inicial muy

cercano al del problema teórico ($y_0 \to y(x_0)$), se obtiene que los y_n están tan cerca como se quiera de $y(x_n)$. Además se dice que es *convergente de orden* p si para problemas suficientemente regulares esta cercanía se puede acotar superiormente en un intervalo por algo del orden de la potencia p-ésima de h, es decir $\|y(x_n) - y_n\| = O(h^p)$.

- El método de Euler es convergente de orden 1 (y no de orden mayor).

Comentarios:

Hay que conocer de memoria la fórmula del método de Euler pero no es necesario saber la cota teórica para $\|y(x_n) - y_n\|$ que asegura su convergencia (y que es bastante pobre en la práctica). No hay que confundirse debido a la ambigüedad con que se usa habitualmente el término *orden de convergencia*. Según la definición, la cota $O(h^p)$ debe cumplirse cuando f es suficientemente regular, lo que no excluye que en algún caso particular se obtenga una cota mejor. Con el abuso de notación obvio se dice que el orden de convergencia es mayor para ese problema. En el sentido contrario, se pueden encontrar ejemplos con poca regularidad (por ejemplo con f no diferenciable) de manera que el método de Euler no tenga siquiera orden uno.

1.3 Un método implícito: la regla del trapecio

Ideas y resultados básicos:

- La regla del trapecio es un método de orden de convergencia 2 en el que y_{n+1} no aparece explícitamente despejado.

Comentarios:

El propósito de esta sección es simplemente mostrar que hay métodos mejores en cuanto a orden pero peores en cuanto al coste computacional (más adelante se verá que la estabilidad entra también en este balance). No hay nada especial que memorizar pero no cuesta ningún esfuerzo recordar la fórmula de la regla del trapecio.

1.4 Métodos de Taylor

Ideas y resultados básicos:

- Los *métodos de Taylor* consisten en tomar como y_{n+1} el polinomio de Taylor en h de $y(x_n + h)$ hasta cierto orden escribiendo $y^{(k)}$ en términos de las derivadas de f y sustituyendo $y(x_n)$ por y_n.

- Se pueden conseguir métodos de Taylor de orden arbitrariamente alto.

Comentarios:
Hay que saber construir métodos de Taylor de orden bajo. En cuanto se examinen algunos ejemplos se verá que son en general métodos muy costosos computacionalmente cuando el problema es mínimamente complejo.

1.5 Métodos de un paso

Ideas y resultados básicos:

- Se llaman *métodos de un paso* a aquellos en los que y_{n+1} se expresa en términos de h y de los datos del paso anterior, x_n e y_n. Si y_{n+1} está explícitamente despejado se dice que son *explícitos*, y en caso contrario, *implícitos*.
- Se llama *error de truncación*, R_n, a la cantidad que habría que sumarle a la ecuación que define un método de un paso para que la verdadera solución la satisficiera sustituyendo y_n por $y(x_n)$ e y_{n+1} por $y(x_{n+1})$. En un método explícito, el error de truncación no es más que el "error local" $y(x_{n+1}) - y_{n+1}$ cuando se supone que en el paso anterior no ha habido error, es decir, la llamada *hipótesis de localización* $y_n = y(x_n)$.
- Se dice que un método es *consistente de orden* p o que tiene *orden de consistencia* p si $R_n = O(h^{p+1})$.
- Si un método es consistente de orden p entonces es convergente de orden p (bajo cierta débil condición de regularidad que se da por supuesta).

Comentarios:
De nuevo no es necesario aprenderse el teorema que acota efectivamente $\|y(x_n) - y_n\|$ en términos del error de truncación, h y algunas constantes. Es importante entender la definición de orden de consistencia porque desde el punto de vista teórico adquirirá todo el protagonismo frente al de convergencia.

1.6 Métodos de Runge-Kutta

Ideas y resultados básicos:

- Los *métodos de Runge-Kutta* son métodos de un paso $y_{n+1} = y_n + h\phi(x_n, y_n, y_{n+1}; h)$ donde la función ϕ es una media ponderada de la función

f del problema evaluada en puntos que a su vez dependen de x_n, y_n y de medias ponderadas de f. Más concretamente, son métodos de la forma

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

$$y_{n+1} = y_n + h \sum_{j=1}^{s} b_j k_j.$$

- Un método de Runge-Kutta se representa por su *tablero de Butcher*

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

y se dice que s es el *número de etapas*.

- Un método de Runge-Kutta es explícito cuando $a_{ij} = 0$ para todo $j \geq i$ y es implícito en caso contrario.

Comentarios:

Es necesario (y muy sencillo) saber pasar de un método a su tablero y viceversa. Los tableros o fórmulas correspondientes a los métodos clásicos como el de Euler mejorado y Euler modificado no es necesario memorizarlos.

1.7 Condiciones de orden

Ideas y resultados básicos:

- Una hipótesis natural considerada por Kutta es la *condición de suma por filas* $c_i = \sum_{j=1}^{s} a_{ij}$ que se dará por supuesto en lo sucesivo. También se supondrá para simplificar las demostraciones, que se trata con sistemas autónomos ($y' = f(y)$), pero esto no constituye una restricción esencial.

- Al desarrollar por Taylor en h la fórmula para el error de truncación se obtiene que la definición de orden (de consistencia) para $p = 1$, 2 y 3 se traduce en sendas ecuaciones que involucran los elementos del tablero de Butcher.

Comentarios:

Es muy conveniente entender bien el argumento que lleva a las condiciones de orden al menos en los casos 1 y 2. Aunque no es totalmente imprescindible aprenderse las ecuaciones que definen estas condiciones de orden porque son consecuencia de la teoría general de los árboles de Butcher, es poco creíble que después de pasar por el curso uno sea capaz de olvidar las fórmulas $\sum b_j = 1$ y $2\sum b_j c_j = 1$.

1.8 Árboles de Butcher

Ideas y resultados básicos:

• Un *árbol con raíz* es un grafo que podemos entender que representa un árbol genealógico donde la raíz corresponde al antecesor común de toda la estirpe. Si se distingue a los *vértices* (individuos) con nombres distintos se dice que el árbol está etiquetado. El orden de un árbol es su número de vértices (de individuos de la estirpe).

• La función *padre* que asigna a cada nombre, excepto al de la raíz, el nombre de su padre está definida unívocamente (cada individuo tiene un solo padre). Sin embargo al asignar a cada nombre el de su hijo no se obtiene una función bien definida porque a veces está multivaluada (cada individuo puede tener varios hijos) y a veces no definida (un individuo puede no tener ningún hijo).

• Hay dos funciones muy importantes asociadas a un árbol en el ámbito de los métodos de Runge-Kutta:

-La función γ asigna a cada árbol el producto de los órdenes de los árboles que van quedando al eliminar la raíz y repetir el proceso con lo árboles resultantes.

- Sea A un matriz $s \times s$. Nombrando los vértices del árbol distintos de la raíz con las etiquetas $k,\ l,\ m,\ n \ldots$ y la raíz como j, se considera el producto $a_{t(k)k} a_{t(l)l} a_{t(m)m} \ldots$ donde $t(i)$ indica el padre de i. La función ϕ_j asigna al árbol la suma de estos productos cuando $k,\ l,\ m,\ n \ldots$ varían entre 1 y s.

• Se cumple el *teorema de Butcher*: Un método de Runge-Kutta es de orden p si y sólo si se cumple

$$\gamma \sum_{j=1}^{s} b_j \phi_j = 1$$

para todos los árboles de orden menor o igual que p, donde los b_j y los a_{rs}

usados para comprobar la igualdad anterior son los dados por el tablero del método.

• Dado un problema $y' = f(y)$ y un árbol cuya raíz está etiquetada con J y el resto de los vértices con K, L, M, $N\ldots$ se considera el producto $f^J_{t^{-1}(J)} f^K_{t^{-1}(K)} f^L_{t^{-1}(L)} \ldots$ donde $t^{-1}(I)$ indica los hijos de I (pueden ser el conjunto vacío). Se llama *diferencial elemental* a la suma de los productos correspondientes cuando K, L, M, $N\ldots$ varía entre 1 y d (el número de variables de f) y se conviene que los superíndices indican el número de coordenada de f y los subíndices las variables con respecto a las que se derivan.

Comentarios:

El teorema de Butcher es un resultado central del curso que hay que saber aplicar y memorizar inexcusablemente. Esto incluye las definiciones y cálculos de ϕ_j y de γ. Incluso si uno no quiere memorizar la definición del diferencial elemental, que es importante desde el punto de vista teórico, al menos debe ser capaz de interpretarla correctamente al verla escrita, lo que implica conocer la notación empleada.

1.9 Demostración de los teoremas de Butcher

Ideas y resultados básicos:

• Para probar $R_n = O(h^{p+1})$ se desarrolla R_n por Taylor (en h), lo que en los métodos de Runge-Kutta conlleva aplicaciones sucesivas de la regla de la cadena que se pueden traducir en el lenguaje de los árboles con fórmulas como la de Faà di Bruno

Comentarios:

Fuera de la idea general mencionada, esta sección es opcional, y sólo se recomienda como una segunda lectura para los alumnos interesados. Nótese, no obstante, que los problemas propuestos en esta sección son una prolongación de los la anterior y, por tanto, todavía aconsejables.

1.10 Métodos explícitos. Orden obtenible

Ideas y resultados básicos:

• En un método de Runge-Kutta explícito el orden no puede superar el número de etapas.

• Si un método de Runge-Kutta explícito tiene orden $p \geq 7$ entonces tiene más de $p+1$ etapas, y si $p \geq 8$ entonces tiene más de $p+2$ etapas.

Comentarios:

De los dos resultados anteriores antes mencionados, basta familiarizarse con el primero y ser capaz de deducirlo como una simple consecuencia de la teoría de los árboles de Butcher.

1.11 Estimaciones de error

Ideas y resultados básicos:

• El término principal del error de truncación de un método de orden p es h^{p+1} multiplicado por un coeficiente que se expresa como una combinación lineal de los productos de $1-\gamma\sum b_j\phi_j$ por el diferencial elemental para todos los árboles de orden $p+1$.

• Si este coeficiente se anula para un problema particular, el término principal deja de serlo y el orden es mayor de lo esperado.

Comentarios:

Dada la fórmula para el término principal del error de truncación (que no es necesario memorizar) hay que saber aplicarla a un problema concreto sencillo. Esta sección muestra claramente la ambigüedad ya señalada al emplear el término *orden*: el teorema de Butcher plantea condiciones necesarias y suficientes para que un método sea de orden p con la definición original (para todo problema suficientemente regular), pero si no se cumplen, el orden puede seguir siendo p si se restringe la definición a un problema particular.

1.12 Extrapolación de Richardson

Ideas y resultados básicos:

• Si $A(h)$ es una aproximación de orden α de B, en el sentido de que $B = A(h)+Ch^\alpha+O(h^{\alpha+1})$, entonces el "error" $B-A(h)$ es $\big(A(h)-A(2h)\big)/(2^\alpha-1)+O(h^{\alpha+1})$. Esta idea se puede aplicar para estimar el error de truncación (que es "local") y el error global.

Comentarios:

La idea que conduce a la fórmula que sustenta la extrapolación de Richardson es suficientemente sencilla para que uno sea capaz de improvisarla. En

los apuntes se hace un análisis cuidadoso, que en este curso se puede considerar opcional, de cómo se aplica la extrapolación de Richardson al estudio del error de truncación. Por otra parte, la estimación del orden empírico para métodos de paso fijo, que en clase se vio dentro de esta sección, en los apuntes se menciona brevemente en la sección 1.15.

1.13 Cambio de paso

Ideas y resultados básicos:

• Es mucho más conveniente no mantener fijo el tamaño del paso, h, haciéndolo variar en función del error que se estima que se está cometiendo.

• En un método de orden p si en cierto x_n se estima que el error "local" cometido es $\mathtt{e}_{\mathtt{est}}$ y supera al error tolerado $\mathtt{e}_{\mathtt{tol}}$, entonces la longitud del paso se debe reducir en un factor de al menos $(\mathtt{e}_{\mathtt{tol}}/\mathtt{e}_{\mathtt{est}})^{1/(p+1)}$, repitiéndose los cálculos de la última iteración. Este mismo factor sirve para aumentar la longitud del paso si el error estimado fuera demasiado pequeño en comparación con el tolerado.

Comentarios:

La fórmula para la reducción del paso es una simple consecuencia de la aproximación Ch^{p+1} para el error de truncación que puede interpretarse como asociado al error cometido en cada paso. Así $Ch^{p+1} \approx \mathtt{e}_{\mathtt{est}}$, $C(h')^{p+1} \approx \mathtt{e}_{\mathtt{tol}}$ $\Rightarrow h' = h(\mathtt{e}_{\mathtt{tol}}/\mathtt{e}_{\mathtt{est}})^{1/(p+1)}$. La introducción de un factor de seguridad es muy natural para no estar en el límite de las aproximaciones.

1.14 Pares encajados

Ideas y resultados básicos:

• Un *par encajado* son dos métodos de Runge-Kutta que comparten la matriz A (y por tanto $\vec{c}$) en su tablero, de manera que uno tenga orden mayor que el otro.

• La diferencia entre los valores obtenidos con ambos métodos se toma como estimación del error y se cambia el paso consecuentemente.

Comentarios:

Los pares encajados no son muy profundos desde el punto de vista teórico pero constituyen los mejores métodos en muchas situaciones para resolver

numéricamente ecuaciones diferenciales ordinarias, por ello es imprescindible conocer qué son y cómo se emplean.

1.15 Diagramas de eficiencia

Ideas y resultados básicos:

• En un método de paso fijo el orden, p, se puede estimar empíricamente utilizando que el logaritmo del error es aproximadamente una función lineal del logaritmo de h con pendiente p. También se puede retomar la idea bajo la extrapolación de Richardson obteniendo $p \approx \big(\log(e(h)) - \log(e(h/2))\big)/\log 2$.

• En los métodos de paso variable (pares encajados) la relación lineal aproximada con pendiente p es la del logaritmo del error en función del número de pasos.

Comentarios:

De nuevo, como en la sección 1.12, es muy fácil improvisar aproximaciones empíricas del orden en el caso de paso fijo a partir de la hipótesis de que el error se comporta aproximadamente como Ch^p.

◁ ◇ ▷

Problemas recomendados del capítulo 1:

1.1, 2.3, 2.5, 2.6, 3.1, 4.1, 4.3, 5.1, 5.2, 5.3, 6.2, 6.3, 7.4, 7.5, 7.8, 8.1, 8.2, 8.4, 8.8, 9.3, 9.4, 9.6, 10.2, 10.3, 11.1, 12.1, 12.4, 12.6, 13.1, 14.2, 14.4, 14.7.

2 Problemas stiff

2.1 ¿Qué es un problema stiff?

Ideas y resultados básicos:

• Existen problemas que son estables desde el punto de vista teórico (pequeñas variaciones en las condiciones iniciales no modifican el comportamiento asintótico de la solución) pero que se comportan de forma muy inestable cuando se les aplica ciertos métodos numéricos. Son los problemas *stiff*.

Comentarios:

No hay definiciones rigurosas ni resultados concretos generales en esta sección, pero es importante entender perfectamente el ejemplo que aparece en los apuntes, así como los que se vieron en clase.

2.2 Dominio de estabilidad lineal y A-estabilidad

Ideas y resultados básicos:

- El *dominio de estabilidad lineal* de un método es el conjunto de los $z = h\lambda$, $\lambda \in \mathbb{C}$, tales que la solución numérica de $y' = \lambda y$, $y(0) = 1$, cumple $\lim_{n\to\infty} y_n = 0$ cuando el método se aplica con tamaño del paso h.
- Se dice que un método es *A-estable* si su dominio estabilidad lineal incluye el semiplano izquierdo de $\mathbb{C}$. De modo que si un método es A-estable entonces la solución numérica de $y' = \lambda y$ con $\mathrm{Re}\,\lambda < 0$ aproxima a la solución teórica en "el infinito". Es decir, $\lim_{x\to+\infty} y(x) = \lim y_n = 0$.

Comentarios:

Aunque en la siguiente sección se verá un método general para decidir la A-estabilidad de los métodos de Runge-Kutta y hallar el dominio de estabilidad lineal, hay que saber las definiciones originales de estos conceptos y aplicarlas en casos concretos.

2.3 Estabilidad de métodos de Runge-Kutta

Ideas y resultados básicos:

- En los métodos de Runge-Kutta el dominio de estabilidad lineal se escribe como $\{z \in \mathbb{C} \ : \ |R(z)| < 1\}$ donde R es cierta función racional llamada *función de amplificación*.
- La función de amplificación responde a la fórmula $R(z) = 1 + z\vec{b}^T(I - Az)^{-1}\vec{1}$ donde $\vec{1}$ es el vector cuyas coordenadas son todas 1. De aquí se puede deducir que el grado del numerador y del denominador de R está acotado por el número de etapas.
- Los métodos de Runge-Kutta explícitos no son A-estables.
- La función de amplificación de un método de Runge-Kutta de orden (de consistencia) p verifica $R(z) = 1 + z/1! + z^2/2! + \cdots + z^p/p! + O(z^{p+1})$. Si el método es explícito de p etapas la igualdad se cumple sin el término $O(z^{p+1})$.

• Para decidir la A-estabilidad de un método de Runge-Kutta no hace falta conocer el aspecto del dominio de estabilidad lineal, basta comprobar que la función de amplificación tiene todos sus polos, si los hubiera, en el semiplano derecho, y tiene módulo acotado por 1 en el eje imaginario.

Comentarios:

No es necesario memorizar la fórmula de la función de amplificación pero sí saber utilizarla para decidir la A-estabilidad. También hay que entender con claridad por qué el dominio de estabilidad lineal se escribe en términos de ella y sus propiedades asintóticas.

◁ ◇ ▷

Problemas recomendados del capítulo 2:

1.1, 1.4, 2.1, 2.2, 2.4, 3.1, 3.5, 3.6, 3.9, 3.11, 3.12.

3 Diferencias finitas

3.1 La fórmula de cinco puntos

Ideas y resultados básicos:

• El método de diferencias finitas consiste esencialmente en sustituir derivadas parciales por cocientes incrementales evaluados en ciertos nodos que constituyen una discretización del dominio considerado.

• Si se discretiza la x con paso tamaño h y la y con paso de tamaño k, considerándose la red uniforme $(x_i, y_j) = (a + ih, b + jk)$, entonces el laplaciano $\Delta u(x_i, y_j)$ se puede aproximar por el laplaciano discreto dado por la fórmula de cinco puntos

$$\begin{aligned}\Delta_{hk} u(x_i, y_j) =& \frac{u(x_{i+1}, y_j) - 2u(x_i, y_j) + u(x_{i-1}, y_j)}{h^2} \\ &+ \frac{u(x_i, y_{j+1}) - 2u(x_i, y_j) + u(x_i, y_{j-1})}{k^2}.\end{aligned}$$

De hecho se tiene $\Delta u(x_i, y_j) = \Delta_{hk} u(x_i, y_j) + O(h^2 + k^2)$.

• El laplaciano discreto, aplicado a una función discreta definida en los nodos, U_{ij}, comparte con el laplaciano continuo los principios del máximo

para funciones subarmónicas ($\Delta u \geq 0$) y del mínimo para funciones superarmónicas ($\Delta u \leq 0$).

• La ecuación de Poisson $-\Delta u = f$, $u\big|_{\partial R} = g$, en un rectángulo $R = [a,c] \times [b,d]$ da lugar, aplicando el método de diferencias finitas, al problema $-\Delta_{hk} U_{ij} = f_{ij}$, $U_{ij}\big|_{\partial R} = g_{ij}$. Por el principio del máximo (y del mínimo), el sistema lineal resultante es compatible determinado: tiene solución única.

• Utilizando que el laplaciano discreto aproxima al laplaciano continuo hasta orden 2, se puede probar que la aproximación por diferencias finitas de la solución de la ecuación de Poisson en un rectángulo converge a la verdadera solución cuando la discretización se va haciendo más fina. Concretamente se cumple $\max |u(x_i, y_j) - U_{ij}| = O(h^2 + k^2)$.

Comentarios:

Es fundamental entender la sencilla idea en la que se basa el método de diferencias finitas y ser capaz de aplicarla incluso a ecuaciones diferentes de la de Poisson. La fórmula de cinco puntos es un simple caso particular que todos los alumnos deberían ser capaces de deducir (con término de error). El principio del máximo (y del mínimo) discreto así como la unicidad de la solución del esquema de diferencias finitas para la ecuación de Poisson son suficientemente asequibles para ser capaz de reproducirlos sin ayuda adicional. La prueba de la convergencia de la solución numérica a la teórica es un poco más compleja y no se requiere más que conocer la existencia de ese resultado.

◁ ◇ ▷

Problemas recomendados del capítulo 3:

1.1, 1.2, 1.4, 1.5, 1.7, 1.10, 1.12.

4 Elementos finitos

Este capítulo es el único en el cual lo visto en las clases de teoría no se ajusta fielmente a lo que aparece en las notas impresas del curso, habiendo algunos cambios de orden y reducciones al mínimo de algunas secciones. Esencialmente en las clase de teoría se trataron cuatro temas:

4.1- Introducción al método de elementos finitos.

4.2- Formulación abstracta débil y variacional.

4.3- Programación y espacios de elementos finitos en dos variables.

4.4- Breve introducción a la convergencia.
En seguida se pasará a resumirlos y quedarán claros los contenidos de cada uno de ellos. No obstante, para evitar confusiones, se indica aquí la correspondencia de esta nueva división en secciones con respecto a la de los apuntes:

La nueva sección 4.1 corresponde esencialmente a la antigua pero extendida con ejemplos resueltos detalladamente en dimensión uno (que enlazan con la parte introductoria de 4.7.1). La sección 4.2 abarca las secciones 4.2, 4.3, 4.4 y 4.5 de los apuntes pero teniendo en cuenta que la parte teórica referente a las derivadas débiles y los espacios de Sobolev queda minimizada. Nótese que en los apuntes la sección 4.5 contiene a las secciones 4.3 y 4.4 que no son más que ejemplos particulares. La sección 4.3 corresponde a la 4.8 en el caso bidimensional y a una buena parte de la 4.7.2. Finalmente la sección 4.4 es una breve mención de los resultado teóricos de 4.6 y de la parte final de 4.7.1 y 4.7.2.

4.1 Introducción al método de elementos finitos

Ideas y resultados básicos:

• El método de elementos finitos aplicado a un problema de contorno trata de aproximar la solución dentro de un espacio (de funciones) de dimensión finita, esencialmente a través del esquema siguiente:

- Se escoge el espacio $V = \mathcal{L}(\phi_1, \phi_2, \ldots, \phi_M)$ donde ϕ_i son funciones (*funciones base*) sencillas (típicamente lineales a trozos) asociadas a una partición del dominio.

- Se supone momentáneamente que $u \in V$, escribiendo $u = \sum \gamma_l \phi_l$, y si $L[u] = 0$ es el problema que se quiere resolver, se impone la *condición de Galerkin*

$$\langle L[u], \phi_k \rangle = 0 \qquad k = 1, 2, \ldots, M$$

integrando por partes, evitando así problemas de regularidad.

- Se escriben las condiciones resultantes como un sistema lineal $A\vec{\gamma} = \vec{b}$ (A = matriz de rigidez, $\vec{b}$ = vector de carga). Una vez resuelto, la solución aproximada será $\sum \gamma_l \phi_l$.

• En el caso de problemas de contorno en un intervalo, la eleccción natural es subdividir el intervalo en partes iguales y considerar V generado por *funciones tejado* que valen 1 en cierto nodo y 0 en los demás, con el producto escalar "usual" $\langle f, g \rangle = \int fg$.

• Un cambio en el número de partes, y por tanto en las funciones consideradas, se traduce en problemas sencillos (lineales) en un cambio de variable en las integrales que definen los elementos de la matriz de rigidez.

Comentarios:

Se habrán aprendido los contenidos de esta sección si se es capaz de aplicar el método de elementos finitos a problemas de contorno lineales sencillos (digamos del tipo $-u'' + \alpha u = f$, $u(a) = u(b) = 0$); sabiendo incrementar el número de funciones base sin necesidad de repetir los cálculos para hallar los elementos de la matriz de rigidez.

4.2 Formulación abstracta débil y variacional

Ideas y resultados básicos:

• Los problemas a los que se aplica el método de elementos finitos se pueden escribir en la *formulación débil*

$$a(u, v) = L(v) \qquad \forall v \in V$$

donde a es una forma bilineal (y simétrica, continua y coerciva) y L es una forma lineal (y continua); ambas definidas en el espacio de Hilbert V que contiene a la solución u.

• Una vez que se expresa un problema en su formulación débil, el método de elementos finitos se reduce a resolver $A\vec{\gamma} = \vec{b}$ donde $a_{kl} = a(\phi_k, \phi_l)$ y $b_k = L(\phi_k)$. La solución aproximada es $\sum \gamma_l \phi_l$.

• La formulación débil equivale a la *formulación variacional* y expresa la solución u como el mínimo de un funcional, concretamente de $\frac{1}{2}a(u, u) - L(u)$.

• Los espacios de Hilbert con los que se trabaja en las formulaciones débil y variacional son espacios de Sobolev.

• Los problemas del tipo $-\Delta u + \alpha u = f$ en una o varias dimensiones admiten formulaciones débiles cuando se imponen condiciones de Dirichlet homogéneas ($u|_{\partial\Omega} = 0$) o de Neumann ($\frac{\partial u}{\partial n}\big|_{\partial\Omega} = 0$).

Comentarios:

Dado un problema suficientemente sencillo (especialmente $-\Delta u + \alpha u = f$) hay que saber expresarlo en su formulación débil. La formulación variacional en este curso queda relegada a un segundo plano, pero no cuesta apenas esfuerzo adicional el aprenderla. Como ya se ha mencionado, se minimiza,

prácticamente se elimina, la parte que se refiere a la teoría de espacios de Sobolev (que no pertenece a esta asignatura). Esencialmente lo único que hay que tener en mente es que el espacio de Hilbert natural en el que se trabaja es H^1 que tiene por norma $\|u\|_{H^1}^2 = \int u^2 + (u')^2$ donde u' es la *derivada débil*. Esta derivada extiende a la habitual permitiendo derivar funciones lineales a trozos, como el valor absoluto.

4.3 Programación y espacios de elementos finitos en dos variables

Ideas y resultados básicos:

- Para aplicar el método de elementos finitos (lineales) en un dominio (poligonal) de $\mathbb{R}^2$ se triangula y se consideran como funciones base las *funciones pirámide* correspondientes.
- Para cada triángulo hay a lo más tres funciones base cuyos soportes lo incluyen. Estas funciones se pueden expresar, con un cambio de variables adecuado, en términos de ciertas funciones pirámide canónicas sencillas N_1, N_2, N_3.
- Tras el susodicho cambio de variable la contribución de la matriz de rigidez de cada rectángulo se expresa como una combinación lineal de ciertas *matrices básicas* cuyos elementos son todos los productos escalares de N_i por N_j o de sus derivadas parciales.
- El cálculo del vector de carga lleva a problemas de cuadratura en una y dos dimensiones.

Comentarios:

Obviamente la programación tiene más que ver con las clases prácticas que con la de teoría, sin embargo es necesario saber las ideas que fundamentan lo que se programa. En este caso, simplemente hay que entender bien el significado de los puntos antes mencionados.

4.4 Breve introducción a la convergencia

Ideas y resultados básicos:

- En el método de elementos finitos se intenta aproximar una función u que pertenece a un espacio de dimensión infinita (en el que la formulación débil tiene sentido) por otra que pertenece a un espacio de dimensión finita

(el generado por las funciones base). El lema de Céa asegura que el error en esta aproximación está acotado, salvo constantes, por la distancia de u al espacio de dimensión finita.

• Para intervalos en $\mathbb{R}$ y dominios poligonales convexos en $\mathbb{R}^2$, toda función con regularidad suficiente se aproxima bien por los espacios de funciones lineales "a trozos" generados por las funciones tejado o pirámide. En particular el método de elementos finitos convergerá cuando se garantice la condición de regularidad requerida de la solución.

Comentarios:

El resultado realmente importante es el sencillo lema de Céa que hay que conocer al menos en la forma indicada en la idea anterior. Respecto a los resultados de aproximación por funciones lineales (4.20, 4.21, 4.22 y 4.23 en las notas) no es necesario conocer ninguno.

◁ ◇ ▷

Problemas recomendados del capítulo 4:
1.1, 1.2, 1.3, 1.6, 1.7, 2.1, 2.2, 2.3, 2.4, 3.1, 3.2, 3.5, 3.6, 4.1.

5 Métodos lineales multipaso

5.1 Introducción

Ideas y resultados básicos:

• Los *métodos lineales multipaso* (*de k-pasos*) responden a la fórmula

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

Es decir, para hallar y_m hay que conocer los k valores anteriores: y_{m-1}, y_{m-2},$\ldots$,y_{m-k}.

• A un método como el anterior se le asocian los *polinomios característicos*

$$\rho(z) = \sum_{j=0}^{k} \alpha_j z^j \qquad \text{y} \qquad \sigma(z) = \sum_{j=0}^{k} \beta_j z^j.$$

• La definición de *convergencia* y *orden de convergencia* es similar a la de los métodos de un paso salvo que ahora no sólo hay que suponer que y_0 aproxima al valor inicial $y(x_0)$, sino que y_j aproxima a $y(x_j)$ para los k primeros valores.

Comentarios:

Hay que saber la forma general de los métodos lineales de k-pasos y la sencilla definición de los polinomios característicos. No hay que hacer ningún esfuerzo adicional para aprender lo relativo a la convergencia una vez que se conoce este concepto para los métodos de un paso.

5.2 Consistencia

Ideas y resultados básicos:

• El *error de truncación* y el *orden de consistencia* se definen como en el caso de los métodos de un paso.

• El orden de consistencia se puede hallar con la definición o calculando un p que verifique $\rho(1+z)-\sigma(1+z)\log(1+z) = Cz^{p+1}+O(z^{p+2})$ con $C \neq 0$ (cuando $z \to 0$).

Comentarios:

De nuevo las definiciones correspondientes no requieren un esfuerzo adicional porque ya son conocidas del primer capítulo. El criterio citado para calcular el orden de consistencia no es necesario memorizarlo siempre que uno sepa aplicar la definición, lo cual es aquí mucho más fácil que para los métodos de un paso. El teorema 5.4 que en las notas aparece en esta sección se suprime en favor del teorema principal de la sección siguiente.

5.3 0-estabilidad

Ideas y resultados básicos:

• La consistencia de orden p no implica la convergencia.

• Se dice que un método es 0-*estable*, esencialmente, si al partir de datos inciales parecidos se obtienen resultados parecidos. Esto es equivalente a que las raíces de ρ sean de módulo menor que uno o simples de módulo uno.

• Se cumple el *teorema de equivalencia de Dalquist*: Para los métodos lineales multipaso la convergencia equivale a la consistencia de orden $p \geq 1$ junto con la 0-estabilidad.

Comentarios:

Basta con conocer el teorema de Dalquist y saber aplicarlo para decidir la convergencia de un método. La prueba simplemente se esbozará en clase. No es necesario conocer los resultados parciales, teoremas 5.4, 5.6 y 5.8, que se emplean en los apuntes para demostrarlo.

◁ ◇ ▷