Cálculo Numérico II

Fernando Quirós Gracián

Capítulo 1

Métodos de un paso

1.1 Introducción

El objetivo de esta primera parte del curso es encontrar soluciones aproximadas para el problema de valor inicial

$$\begin{cases} y'(x) = f(x, y(x)), \\ y(a) = \eta, \end{cases} \tag{1.1}$$

donde $f : D \subset \mathbb{R} \times \mathbb{R}^d \mapsto \mathbb{R}^d$ es una función continua y $(a, \eta) \in D$, siendo D un dominio. Estamos tratando con sistemas, y por tanto y, f y η tienen d componentes cada uno,

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

Este problema tiene una solución única bajo condiciones bastante generales, que supondremos a partir de ahora.

Teorema 1.1 (Euler, 1768). *Sea $f : D \subset \mathbb{R} \times \mathbb{R}^d \mapsto \mathbb{R}^d$ una función continua y $(a, \eta) \in D$. Supongamos que f satisface la condición de Lipschitz*

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

para algún $L \geq 0$. *Entonces hay un intervalo* $I = [a - \alpha, a + \alpha]$ *tal que existe una única solución de* (1.1) *en* I.

Aunque el problema (1.1) tenga solución, en la mayor parte de los casos no es posible encontrar una forma cerrada para la misma por métodos analíticos. Habrá que contentarse con obtener una aproximación numérica.

Buscaremos una solución de (1.1) en el intervalo $[a, b]$, donde a y b son finitos, con f satisfaciendo las hipótesis del teorema. Cuando haga falta, o convenga para simplificar razonamientos, haremos hipótesis adicionales de regularidad sobre f.

Todos los métodos numéricos para resolver (1.1) que veremos involucran la idea de *discretización*; esto es, el intervalo continuo $[a, b]$ se reemplaza por el conjunto discreto de puntos $\{x_n\}$ definido por

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

El parámetro h es la *longitud del paso*. De momento lo consideraremos constante, aunque conviene mencionar que gran parte de la potencia de los algoritmos modernos proviene de su capacidad de cambiar h automáticamente a medida que se realizan los cálculos.

Denotamos por y_n a la aproximación de la solución $y(x_n)$ de (1.1) en x_n,

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

Nuestro objetivo es encontrar una forma de producir una sucesión de valores $\{y_n\}$ que aproxime a la solución de (1.1) en el conjunto discreto de puntos $\{x_n\}$; tal sucesión constituye una *solución numérica* del problema (1.1). Los valores y_n se pueden interpolar para obtener una aproximación de $y(x)$ en puntos x que no pertenezcan al conjunto x_n.

Un *método numérico* será un procedimiento para producir la sucesión de valores aproximados $\{y_n\}$.

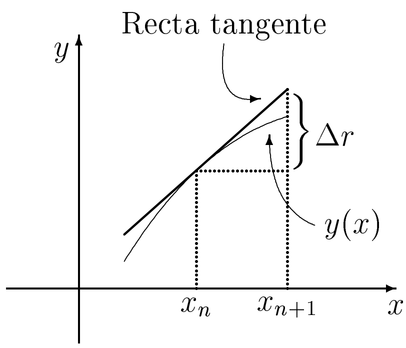
1.2 Método de Euler

Nuestro primer método es el *método de Euler*,

$$y_{n+1} = y_n + hf(x_n, y_n). \tag{1.2}$$

Vamos a llegar a él desde cuatro puntos de vista diferentes.

- *Punto de vista geométrico* ($d = 1$). Utilizamos la recta tangente en x_n como aproximación de la curva solución en $x_n \leq x \leq x_{n+1}$.



  Entonces, llamando Δr al incremento de la recta tangente en $[x_n, x_{n+1}]$, se tiene

$$\frac{\Delta r}{h} = y'(x_n) = f(x_n, y(x_n)),$$

$$y(x_{n+1}) - y(x_n) \approx \Delta r = hy'(x_n),$$

$$y(x_{n+1}) \approx y(x_n) + hf(x_n, y(x_n)).$$

  El método de Euler se obtiene imponiendo que la solución numérica satisfaga esta relación de forma exacta.

- *Serie de Taylor.* Desarrollamos $y(x_{n+1})$ alrededor de x_n y usamos la ecuación diferencial para obtener

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

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

Si despreciamos el término de error obtenemos el método de Euler. El *residuo* o cantidad despreciada,

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

recibe el nombre de *error de truncación*, o *error de discretización* del método. Es lo que le falta a la solución teórica para satisfacer la ecuación del método. También se puede ver como el error que se comete al dar un paso con el método, suponiendo que el valor en x_n era exacto.

- *Diferenciación numérica.* De la definición de derivada,

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

  de donde, despejando, de nuevo obtenemos el método de Euler.

- *Cuadratura numérica.* Integramos $y'(x) = f(x, y(x))$ en $[x_n, x_{n+1}]$, y obtenemos

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

  Para calcular la integral utilizamos la fórmula de cuadratura numérica

  $$\int_c^{c+h} g \approx hg(c),$$

  llamado regla rectangular izquierda, y llegamos al método de Euler.

Todos estos puntos de vista se utilizan para obtener otros métodos más precisos.

Para generar una solución numérica de (1.1) usando (1.2) necesitamos un *valor de arranque* y_0. Uno desearía tomar $y_0 = y(x_0)$, pero en general esto no es posible, debido a la precisión finita del ordenador. Así pues, habrá que contentarse con tomar una aproximación $y_0 \approx y(x_0)$.

¿Da el método de Euler una buena aproximación a la verdadera solución del problema (1.1)? Una medida posible de la bondad de la aproximación es el mayor error cometido

$$\max_{1\leq n\leq N} \|y(x_n) - y_n\|.$$

Definición 1.2. *Decimos que un método numérico es* convergente *si para todo problema* (1.1) *con solución suficientemente regular se tiene que*

$$\lim_{h\to 0^+} \max_{1\leq n\leq N} \|y(x_n) - y_n\| = 0 \qquad \textit{si} \lim_{h\to 0^+} \|y(x_0) - y_0\| = 0.$$

Estamos pidiendo que el valor de arranque y_0 aproxime bien al dato inicial $y(x_0)$; si esto no es así no hay por qué esperar que la solución numérica se parezca a la teórica.

Observación. El producto Nh permanece constante, $Nh = b - a$.

Se tiene el siguiente resultado de convergencia.

Teorema 1.3. *El método de Euler es convergente. Es más, se tiene la siguiente cota para el error*

$$\|y(x_n) - y_n\| \leq e^{(b-a)L}\|y(x_0) - y_0\| + \frac{C}{2L}\left(e^{(b-a)L} - 1\right) h,$$

donde $C = \max_{x\in[a,b]} \|y''(x)\|$ *y* $L > 0$ *es una constante de Lipschitz para* $f(x,y)$ *con respecto a* y.

Prueba. Para simplificar la prueba supondremos que $f \in C^1$; derivando la ecuación se ve inmediatamente que esto implica que $y \in C^2$.

Tenemos que estimar $e_n = y(x_n) - y_n$. Utilizando la fórmula de Taylor y la ecuación diferencial tenemos que

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

con $\overline{\xi}_n \in [x_n, x_{n+1}]$. Restándole a esta ecuación la ecuación (1.2) queda

$$e_{n+1} = e_n + h(f(x_n, y(x_n)) - f(x_n, y_n)) + \frac{h^2}{2} y''(\overline{\xi}_n).$$

Usando la desigualdad triangular y la condición de Lipschitz para f obtenemos

$$\begin{aligned} \|e_{n+1}\| &\leq \|e_n\| + h\|f(x_n, y(x_n)) - f(x_n, y_n)\| + Ch^2/2 \\ &\leq \|e_n\| + hL\|e_n\| + Ch^2/2, \end{aligned}$$

donde $C = \max_{x \in [a,b]} \|y''(x)\|$; es decir,

$$\|e_{n+1}\| \leq (1 + hL)\|e_n\| + Ch^2/2.$$

De aquí se tiene que

$$\begin{aligned} \|e_1\| &\leq (1 + hL)\|e_0\| + Ch^2/2, \\ \|e_2\| &\leq (1 + hL)\|e_1\| + Ch^2/2 \\ &\leq (1 + hL)^2\|e_0\| + \big((1 + hL) + 1\big)Ch^2/2, \\ &\vdots \\ \|e_n\| &\leq (1 + hL)\|e_{n-1}\| + Ch^2/2 \\ &\leq (1 + hL)^n\|e_0\| + \big((1 + hL)^{n-1} + \cdots + (1 + hL) + 1\big)Ch^2/2. \end{aligned}$$

Usando la fórmula de la suma de los n primeros términos de una progresión geométrica tenemos que

$$\|e_n\| \leq (1 + hL)^n\|e_0\| + \frac{Ch}{2L}\left((1 + hL)^n - 1\right).$$

Por otra parte, como hL es positivo, $1 + hL < e^{hL}$, y así

$$\|e_n\| \leq e^{nhL}\|e_0\| + \frac{Ch}{2L}\left(e^{nhL} - 1\right) = e^{(b-a)L}\|e_0\| + \frac{Ch}{2L}\left(e^{(b-a)L} - 1\right)$$

si $0 \leq n \leq N$. $\square$

Ejemplo. Aplicamos el método de Euler al problema

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

para calcular de forma aproximada $y(b) = \int_a^b g$. Tomando $y_0 = 0$ tenemos que

$$\begin{aligned} y(b) \approx y_N &= y_{N-1} + h\,g(x_{N-1}) = y_{N-2} + h\,g(x_{N-2}) + h\,g(x_{N-1}) \\ &= \cdots = \sum_{j=0}^{N-1} h\,g(x_j). \end{aligned}$$

Así pues,

$$|y(x_N) - y_N| = \left| \int_a^b g - \sum_{j=0}^{N-1} h\,g(x_j) \right|,$$

y la convergencia del método de Euler nos garantiza la convergencia de la suma de Riemann

$$\sum_{j=0}^{N-1} h\,g(x_j)$$

a la integral $\int_a^b g$.

Ejemplo. Consideramos el problema lineal

$$y'(x) = \lambda y(x), \qquad y(0) = \eta.$$

La solución exacta es $y(x) = \eta e^{\lambda x}$. Aplicando el método de Euler con $y_0 = \eta$ obtenemos

$$y_{n+1} = y_n + h\lambda y_n = (1 + h\lambda) y_n,$$

e iterando se llega a

$$y_n = (1 + h\lambda)^n \eta.$$

Así pues,

$$\lim_{h \to 0^+} y_N = \lim_{N \to \infty} \eta \left(1 + \frac{b\lambda}{N}\right)^N = \eta e^{\lambda b},$$

lo que prueba la convergencia del método de Euler directamente. Sin embargo esta prueba no da ninguna estimación del error.

La cota proporcionada por el Teorema 1.3 en general sobreestima el error en muchos ordenes de magnitud, y no se debe por tanto utilizar en la práctica. Sin embargo da una información importante, y es que el error es una $O(h)$, siempre y cuando $\|y(x_0) - y_0\| = O(h)$. Eso motiva la siguiente definición.

Definición 1.4. *Decimos que un método numérico es* convergente de orden p *si para todo problema* (1.1) *con solución suficientemente regular se tiene que*

$$\max_{1\leq n\leq N}\|y(x_n)-y_n\|=O(h^p) \quad \textit{si } \|y(x_0)-y_0\|=O(h^p) \quad \textit{cuando } h\to 0^+.$$

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

Ejemplo. Consideramos el problema

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

cuya solución exacta es $y(x)=x$. Aplicando Euler con valor de arranque $y_0=0$ tenemos $y_{n+1}=y_n+h$, y por tanto $y_n=nh=x_n$. Euler nos da por tanto la solución para este problema sin ningún error.

Si consideramos el problema

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

la solución exacta es $y(x)=x^2/2$. Al aplicar Euler con valor de arranque $y_0=0$ tenemos que

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

Es decir,

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

y por tanto

$$\max_{1\leq n\leq N}\|y(x_n)-y_n\|=\max_{1\leq n\leq N}\|x_nh/2\|=bh/2,$$

lo que demuestra que el método no es de orden 2.

Para este problema la función f satisface una condición de Lipschitz en y para cualquier $L \geq 0$; por tanto el Teorema 1.3 nos proporciona una estimación del error

$$\max_{1 \leq n \leq N} \|y(x_n) - y_n\| \leq (e^{bL} - 1)h/2L,$$

para cada $L > 0$. Pasando al límite $L \to 0^+$ llegamos a

$$\max_{1 \leq n \leq N} \|y(x_n) - y_n\| \leq bh/2,$$

y la cota es en este caso óptima.

1.3 Un método implícito: la regla del trapecio

El método de Euler aproxima la derivada en $[x_n, x_{n+1}]$ por una constante, concretamente por su valor en x_n. ¿Por qué privilegiar al punto x_n? ¿No será mejor tomar, por ejemplo, como aproximación constante de la derivada el promedio de sus valores en los extremos del intervalo? En ese caso

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

Esta aproximación conduce a la llamada *regla del trapecio*,

$$y_{n+1} = y_n + \frac{h}{2}\big(f(x_n, y_n) + f(x_{n+1}, y_{n+1})\big). \tag{1.3}$$

Veamos que converge con un orden mejor que el del método de Euler.

Teorema 1.5. *La regla del trapecio es convergente de orden 2.*

Prueba. Como en el método de Euler, definimos el error de truncación R_n como lo que le falta a la verdadera solución para satisfacer la ecuación del método, es decir

$$y(x_{n+1}) = y(x_n) + \frac{h}{2}\big(f(x_n, y(x_n)) + f(x_{n+1}, y(x_{n+1}))\big) + R_n. \tag{1.4}$$

Restando la ecuación del método (1.3) de la ecuación (1.4) obtenemos que el error $e_n = y(x_n) - y_n$ satisface la siguiente recurrencia

$$\begin{aligned} e_{n+1} = e_n + \tfrac{h}{2}\{&[f(x_n, y(x_n)) - f(x_n, y_n)] \\ &+[f(x_{n+1}, y(x_{n+1})) - f(x_{n+1}, y_{n+1})]\} + R_n. \end{aligned}$$

Usando la desigualdad triangular y la condición de Lipschitz para f obtenemos que

$$\|e_{n+1}\| \leq \|e_n\| + \frac{h}{2}(L\|e_n\| + L\|e_{n+1}\|) + \|R_n\|.$$

Suponiendo que $hL/2 < 1$ llegamos a

$$\|e_{n+1}\| \leq \left(\frac{1+\frac{hL}{2}}{1-\frac{hL}{2}}\right)\|e_n\| + \frac{\|R_n\|}{1-\frac{hL}{2}}.$$

Para proseguir necesitamos una estimación de $\|R_n\|$ que no dependa de n. Desarrollando por Taylor alrededor de $x = x_n$ y usando la ecuación diferencial tenemos que

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

Así pues,

$$\|e_{n+1}\| \leq \left(\frac{1+\frac{hL}{2}}{1-\frac{hL}{2}}\right)\|e_n\| + \frac{Ch^3}{1-\frac{hL}{2}} \qquad \text{con } C = \frac{5}{12}\max_{x\in[a,b]}\|y'''(x)\|.$$

Iterando, y usando la fórmula de la suma de los n primeros términos de una progresión geométrica tenemos que

$$\|e_n\| \leq \left(\frac{1+\frac{hL}{2}}{1-\frac{hL}{2}}\right)^n\|e_0\| + \frac{Ch^2}{L}\left(\left(\frac{1+\frac{hL}{2}}{1-\frac{hL}{2}}\right)^n - 1\right).$$

Usando ahora que

$$\frac{1+\frac{hL}{2}}{1-\frac{hL}{2}} = 1 + \frac{hL}{1-\frac{hL}{2}} < e^{hL/(1-\frac{hL}{2})},$$

concluimos que

$$\|e_n\| \leq e^{(b-a)L/(1-\frac{hL}{2})}\|e_0\| + \frac{Ch^2}{L}\left(e^{(b-a)L/(1-\frac{hL}{2})} - 1\right) \qquad \text{si } 0 \leq n \leq N,$$

de donde se deduce inmediatamente la convergencia de orden 2. □

La regla del trapecio tiene un orden de convergencia mayor que el método de Euler. Eso no significa que sea un método mejor, pues hay otra diferencia importante entre ambos métodos. El método de Euler es *explícito*; el valor y_{n+1} viene dado explícitamente en términos del valor anterior y_n y se puede calcular fácilmente mediante una evaluación de f y unas pocas operaciones aritméticas. Por el contrario, la regla del trapecio es un método *implícito*; para calcular y_{n+1} hay que resolver un sistema de ecuaciones no lineales, lo que en general es computacionalmente costoso.

En resumen, el orden no lo es todo. Aunque en un método de orden alto en principio hay que dar menos pasos, éstos pueden ser muy costosos; puede ser preferible un método de orden más bajo, en el que, a pesar de dar más pasos, cada uno de ellos lleve menos trabajo computacional.

Observación. A pesar de todo, los métodos implícitos son útiles a veces, por ejemplo para tratar problemas *stiff* (ver el capítulo 2).

1.4 Métodos de Taylor

El método de Euler es tan sólo de primer orden. Eso supone un inconveniente. Si queremos un error muy pequeño nos veremos forzados a tomar h muy pequeño, lo que supondrá tener que dar muchos pasos, con un coste computacional que puede ser inaceptable. La regla del trapecio es de segundo orden, pero a cambio es implícita. ¿Podemos construir métodos "mejores", de mayor orden que el de Euler, y que a la vez sean explícitos?

La idea para ello será tomar más términos en el desarrollo de Taylor, lo que da lugar a los llamados *métodos de Taylor*. Como ejemplo construimos el método de Taylor de orden 2, en el caso $d = 1$.

El desarrollo de Taylor produce

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

El término de error, de orden h^3, se desprecia. Para calcular $y'(x_n)$ e $y''(x_n)$ usamos la ecuación diferencial,

$$\begin{aligned}
&y'(x) = f(x, y(x)),\\
&y''(x) = f_x(x, y(x)) + f_y(x, y(x))y'(x) = f_x(x, y(x)) + f_y(x, y(x))f(x, y(x)).
\end{aligned}$$

Obtenemos el método

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

En este caso el error de truncación $R_n = \frac{h^3}{3!}y'''(\xi_n)$ satisface que

$$\|R_n\| \leq Ch^3,$$

y esperamos que el método sea de orden 2 (lo veremos más adelante).

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

Comparemos con el método de Euler. Los métodos de Taylor tienen mayor orden, luego habrá que dar menos pasos. A cambio, dar un paso es computacionalmente más caro. No sólo hay que evaluar f, sino también sus derivadas, lo que puede ser muy costoso, sobre todo si éstas son difíciles. Así, si no queremos mucha precisión y la función es complicada, Euler es preferible. Si, por el contrario, queremos bastante precisión y la función y sus derivadas no son complicadas, es preferible usar un método de Taylor de orden más alto. En resumen: no hay un método que sea “el mejor”. Cuál sea el más conveniente depende de la ecuación diferencial y de la precisión que queramos conseguir.

1.5 Métodos de un paso

El método de Euler, los métodos de Taylor y muchos otros, entre ellos los métodos de Runge-Kutta, que veremos más adelante, se pueden escribir como

$$y_{n+1} = y_n + h\phi(x_n, y_n, y_{n+1}; h), \tag{1.5}$$

donde ϕ es una función que depende de f, y que se denomina *función de incremento*. Los métodos de esta forma reciben el nombre de *métodos de un paso*, pues para calcular el paso y_{n+1}, sólo necesitamos conocer un paso, el anterior, y_n.

Si $\phi(x_n, y_n, y_{n+1}; h)$ no depende de y_{n+1} el método es *explícito*; en caso contrario será *implícito* y en cada paso tendremos que resolver un sistema no lineal para obtener y_{n+1}.

Ejemplos.

- El método de Euler es un método de un paso explícito con función de incremento $\phi(x_n, y_n, y_{n+1}; h) = f(x_n, y_n)$.
- La regla del trapecio es un método de un paso implícito con función de incremento $\phi(x_n, y_n, y_{n+1}; h) = \frac{1}{2}(f(x_n, y_n) + f(x_n + h, y_{n+1}))$.
- El método de Taylor de orden 2 es un método de un paso explícito con función de incremento (para $d = 1$) dada por $\phi(x_n, y_n, y_{n+1}; h) = f(x_n, y_n) + \frac{h}{2}(f_x(x_n, y_n) + f_y(x_n, y_n)f(x_n, y_n))$.

Queremos un teorema de convergencia para métodos de un paso. Para ello tenemos que estimar $e_n = y(x_n) - y_n$. Con el fin de imitar las pruebas de convergencia del método de Euler y de la regla del trapecio, definimos el error de truncación como lo que le falta a la verdadera solución, $y(x)$, para satisfacer la ecuación del método.

Definición 1.6. *El* error de truncación*, R_n, del método* (1.5) *se define por*

$$y(x_{n+1}) = y(x_n) + h\phi(x_n, y(x_n), y(x_{n+1}); h) + R_n. \tag{1.6}$$

Por conveniencia separamos un factor h y escribimos $R_n = h\tau_n$. Restando a la ecuación (1.6) la ecuación del método queda

$$e_{n+1} = e_n + h(\phi(x_n, y(x_n), y(x_{n+1}); h) - \phi(x_n, y_n, y_{n+1}; h)) + h\tau_n. \qquad (1.7)$$

Usando la desigualdad triangular obtenemos

$$\|e_{n+1}\| \le \|e_n\| + h\|\phi(x_n, y(x_n), y(x_{n+1}); h) - \phi(x_n, y_n, y_{n+1}; h)\| + h\|\tau_n\|.$$

Para poder seguir pedimos que ϕ satisfaga la condición de Lipschitz

$$\|\phi(x, y, z; h) - \phi(x, \hat{y}, \hat{z}; h)\| \le L(\|y - \hat{y}\| + \|z - \hat{z}\|). \qquad (1.8)$$

Nótese que la constante L puede depender de f. Tenemos ahora que

$$\|e_{n+1}\| \le \|e_n\| + hL(\|e_n\| + \|e_{n+1}\|) + h\|\tau_n\|.$$

Sea $\tau = \max_{0\le n\le N} \|\tau_n\|$. Entonces, si $hL < 1$,

$$\|e_{n+1}\| \le \left(\frac{1 + hL}{1 - hL}\right) \|e_n\| + \frac{h\tau}{1 - hL}.$$

A partir de esta desigualdad se tiene, repitiendo lo que hicimos para la regla del trapecio el siguiente resultado.

Teorema 1.7. *Si ϕ satisface la condición de Lipschitz* (1.8)*, entonces, para $hL < 1$ se tiene que*

$$\|y(x_n) - y_n\| \le e^{2(b-a)L/(1-hL)}\|y(x_0) - y_0\| + \frac{\tau}{2L}\left(e^{2(b-a)L/(1-hL)} - 1\right),$$

donde $\tau = \max_{0\le n\le N} \|\tau_n\|$. Por tanto, si $\lim_{h\to 0^+} \tau = 0$, el método converge. Si además $\tau = O(h^p)$, entonces el método converge con orden p.

Este teorema motiva la siguiente definición.

Definición 1.8. *Un método se dice* consistente *(respectivamente* consistente de orden *p) si para todo problema* (1.1) *con solución suficientemente regular*

$$\lim_{h\to 0^+} \tau = 0 \qquad (\textit{respectivamente } \tau = O(h^p)).$$

Hemos probado que un método de un paso con ϕ satisfaciendo la condición de Lipschitz (1.8) y consistente (consistente de orden p) es convergente (convergente de orden p).

¿Qué quiere decir la consistencia? Tenemos que

$$\frac{y(x_n+h)-y(x_n)}{h}=\phi(x_n,y(x_n),y(x_n+h);h)+\tau_n.$$

Tomamos $x_n=a+nh=\bar{x}$ fijo, y pasamos al límite $h\to 0^+$. La consistencia resulta ser equivalente a que $y'(\bar{x})=\phi(\bar{x},y(\bar{x}),y(\bar{x});0)$, es decir, a

$$\phi(x,y,y;0)=f(x,y).$$

Esta condición garantiza que estamos resolviendo la ecuación diferencial correcta, que estamos siendo *consistentes* con ella.

Ejemplos.

- Para el método de Euler $\phi(x,y,z;h)=f(x,y)$; usando la condición de Lipschitz en y para f se tiene que

  $$\|\phi(x,y,z;h)-\phi(x,\hat{y},\hat{z};h)\|=\|f(x,y)-f(x,\hat{y})\|\le L\|y-\hat{y}\|,$$

  de donde se sigue inmediatamente la condición (1.8). Por otra parte $\tau_n=\frac{h}{2}y''(\overline{\xi}_n)$; por tanto $\tau=O(h)$ si $y\in C^2([a,b])$. Combinando ambos hechos tenemos que el método de Euler es convergente de orden 1.
- Para la regla del trapecio $\phi(x,y,z;h)=\frac{1}{2}(f(x,y)+f(x+h,z))$; usando la condición de Lipschitz en y para f se tiene que

  $$\begin{aligned}\|\phi(x,y,z;h)-\phi(x,\hat{y},\hat{z};h)\|&\le\tfrac{1}{2}\|f(x,y)-f(x,\hat{y})\|\\&+\tfrac{1}{2}\|f(x+h,z)-f(x+h,\hat{z})\|\le\tfrac{L}{2}(\|y-\hat{y}\|+\|z-\hat{z}\|);\end{aligned}$$

  es decir, se cumple (1.8). Por otra parte

  $$\tau_n=\left(\frac{y'''(\overline{\xi}_n)}{3!}-\frac{y'''(\overline{\eta}_n)}{4}\right)h^2;$$

  así $\tau=O(h^2)$ si $y\in C^3([a,b])$, y por tanto la regla del trapecio es convergente de orden 2.

- Para el método de Taylor de orden k, $\tau_n = \frac{h^k}{(k+1)!} y^{(k+1)}(\overline{\xi}_n)$, y si la solución es suficientemente regular se tiene que $\tau = O(h^k)$.

1.6 Métodos de Runge-Kutta

Los métodos de Taylor son conceptualmente sencillos. Sin embargo:

- Exigen calcular derivadas de orden superior, lo que puede ser engorroso.
- Exigen evaluar dichas derivadas, lo que puede ser computacionalmente costoso.

Para evitar estas dificultades, los métodos de Runge-Kutta evalúan $f(x, y)$ en más puntos, intentando igualar la precisión de los métodos de Taylor. Concretamente toman

$$\phi(x_n, y_n, y_{n+1}; h) = \sum_{i=1}^{s} A_i f(\theta_i, \gamma_i)$$

de forma que $\phi(x_n, y_n, y_{n+1}; h)$ sea una media ponderada de las pendientes cerca de (x_n, y_n). La idea es elegir los parámetros A_i, θ_i, γ_i de manera que la función de incremento coincida lo más posible con la función de incremento de un método de Taylor.

Veamos por ejemplo cómo hacer esto con dos evaluaciones de la función y $d = 1$. Parece razonable tomar $(\theta_1, \gamma_1) = (x_n, y_n)$. Para θ_2 tomamos $\theta_2 = x_n + \alpha h$. Una elección sensata para γ_2 consiste en estimar el valor $y(\theta_2)$ avanzando con el método de Euler, es decir

$$\gamma_2 = y_n + \alpha h f(x_n, y_n).$$

Queda por tanto

$$y_{n+1} = y_n + h\big(A_1 f(x_n, y_n) + A_2 f(x_n + \alpha h, y_n + \alpha h f(x_n, y_n))\big).$$

¿Cómo elegir A_1, A_2 y α? Desarrollamos por Taylor en potencias de h y obtenemos

$$\begin{aligned} y_{n+1} = \quad & y_n + (A_1 + A_2)hf(x_n, y_n) \\ & + \alpha A_2 h^2 (f_x(x_n, y_n) + f_y(x_n, y_n)f(x_n, y_n)) + O(h^3). \end{aligned}$$

Si hacemos

$$A_1 + A_2 = 1, \qquad \alpha A_2 = 1/2,$$

entonces la función de incremento ϕ coincide hasta orden uno con la función de incremento ϕ^{T2} del método de Taylor de orden 2,

$$\phi(x_n, y_n, y_{n+1}; h) = \phi^{T2}(x_n, y_n, y_{n+1}; h) + O(h^2).$$

Veamos cuál es el orden de consistencia del método:

$$\begin{aligned} h\tau_n &= y(x_{n+1}) - y(x_n) - h\phi(x_n, y(x_n), y(x_{n+1}); h) \\ &= y(x_{n+1}) - y(x_n) - h\phi^{T2}(x_n, y(x_n), y(x_{n+1}); h) + O(h^3) \\ &= O(h^3). \end{aligned}$$

Concluimos que $\tau = O(h^2)$, es decir, el método es consistente de orden 2.

Por tanto, para ver que el método es convergente de orden 2 basta con comprobar que la función de incremento

$$\phi(x, y, z; h) = A_1 f(x, y) + A_2 f(x + \alpha h, y + \alpha h f(x, y)).$$

satisface la condición de Lipschitz (1.8). Esto se prueba fácilmente usando que f es Lipschitz en y.

Hemos obtenido $A_2 = \frac{1}{2\alpha}$, $A_1 = 1 - \frac{1}{2\alpha}$ y $\alpha \neq 0$ arbitrario. No tenemos un único método, sino toda una familia.

Ejemplos.

- Tomando $\alpha = 1/2$ obtenemos el llamado método de *Euler modificado*,

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

- Tomando $\alpha = 1$ se tiene el método de *Euler mejorado*, también conocido como método de *Heun de orden 2*,
$$y_{n+1} = y_n + \frac{h}{2}(f(x_n, y_n) + f(x_n + h, y_n + hf(x_n, y_n)).$$

En ambos casos se realizan dos evaluaciones de función, frente a las tres necesarias en el método de Taylor de orden 2 (a saber, $f(x_n, y_n)$, $f_x(x_n, y_n)$ y $f_y(x_n, y_n)$) cuando $d = 1$, siendo la cosa aún peor para problemas de dimensión mayor.

El método de Euler modificado se puede escribir alternativamente como

$$\begin{aligned} k_1 &= f(x_n, y_n), \\ k_2 &= f(x_n + \tfrac{h}{2}, y_n + \tfrac{h}{2}k_1), \\ y_{n+1} &= y_n + hk_2. \end{aligned}$$

En cuanto a Euler mejorado, se puede escribir como

$$\begin{aligned} k_1 &= f(x_n, y_n), \\ k_2 &= f(x_n + h, y_n + hk_1), \\ y_{n+1} &= y_n + \tfrac{h}{2}(k_1 + k_2). \end{aligned}$$

Siguiendo estos ejemplos, definimos un *método de Runge-Kutta de s etapas* para el problema (1.1) por

$$\begin{cases} k_i = f(x_n + c_i h, y_n + h \sum_{j=1}^{s} a_{ij} k_j), \qquad i = 1, 2, \ldots, s \\ y_{n+1} = y_n + h \sum_{j=1}^{s} b_j k_j. \end{cases} \tag{1.9}$$

Cada una de las evaluaciones de función k_i es una *etapa*.

El método (1.9) se representa por medio de su *tablero de Butcher*:

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

Ejemplos.

- El tablero del método de Euler modificado es

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

- El tablero del método de Euler mejorado es

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

Si escribimos

$$c = (c_1, c_2, \dots, c_s)^T, \qquad b = (b_1, b_2, \dots, b_s)^T, \qquad A = (a_{ij}),$$

el tablero se puede resumir como

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

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

Si el método no es explícito, es *implícito*. En general es necesario entonces resolver en cada paso un sistema no lineal para calcular los k_i. Este sistema tiene dimensión ds. Hay una situación intermedia; si $a_{ij} = 0$ para $j > i$, $i = 1, 2, \dots, s$, entonces cada k_i está definido individualmente por

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

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

1.7 Condiciones de orden

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

$$\phi(x_n, y_n, y_{n+1}; h) = \sum_{j=1}^{s} b_j k_j.$$

Si hacemos $h = 0$, entonces $k_i = f(x_n, y_n)$ para todo $i = 1, 2, \ldots, s$. Así,

$$\phi(x, y, y; 0) = \left(\sum_{j=1}^{s} b_j\right) f(x, y).$$

Por tanto, un método de Runge-Kutta es consistente si y sólo si

$$\sum_{j=1}^{s} b_j = 1. \tag{1.10}$$

Por otra parte, puesto que las etapas k_i son evaluaciones de la función f, no es difícil convencerse de que ϕ satisface la condición de Lipschitz (1.8) si f satisface una condición de Lipschitz en y. Así pues, por el Teorema 1.7, la condición de consistencia (1.10) es suficiente para garantizar la convergencia. Veamos que también es necesaria.

Teorema 1.9. *Un método de Runge-Kutta es convergente si y sólo si es consistente.*

Prueba. Si es convergente, en particular lo es para el problema

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

cuya solución es $y(x) = x$. El método, aplicado a este problema, se reduce a

$$y_{n+1} = y_n + h\sum_{j=1}^{s} b_j.$$

Tomando como valor de arranque $y_0 = 0$ tenemos que la solución numérica viene dada por

$$y_n = nh\sum_{j=1}^{s} b_j = x_n\sum_{j=1}^{s} b_j.$$

Por tanto

$$y(x_n) - y_n = x_n\left(1 - \sum_{j=1}^{s} b_j\right),$$

y la solución numérica converge a la teórica si y sólo si se cumple la condición de consistencia (1.10). $\square$

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

Supongamos que $y(x_n) = y_n$, la llamada *hipótesis de localización*. Usando (1.7) y el Teorema del Valor Medio tenemos que

$$R_n = \left(I - h\frac{\partial\phi}{\partial y}(x_n, y_n, \overline{\xi}; h)\right)(y(x_{n+1}) - y_{n+1}). \tag{1.11}$$

Esta igualdad nos dice que el término principal del desarrollo de $y(x_{n+1}) - y_{n+1}$ en potencias de h coincide con el de R_n. Por tanto $y(x_{n+1}) - y_{n+1} = O(h^{p+1})$ si y sólo si $R_n = O(h^{p+1})$.

Así pues, lo que hay que hacer es obtener desarrollos en potencias de h para $y(x_{n+1})$, la solución teórica, e y_{n+1}, la solución numérica. Si estos desarrollos coinciden hasta orden p, entonces $R_n = O(h^{p+1})$ y tendremos convergencia de orden p.

Para que los desarrollos de la solución numérica y la teórica sean iguales hasta orden p, es necesario que sus derivadas hasta orden p en $h = 0$ coincidan. Eso dará lugar a unas ciertas condiciones sobre los coeficientes del método, las llamadas *condiciones de orden*. Obtenerlas será nuestro siguiente objetivo.

Notación. Los superíndices en mayúsculas denotarán componentes, y variarán por tanto entre 1 y d. Los subíndices en minúsculas denotan etapas; varían entre 1 y s. Finalmente, los subíndices en mayúsculas significan derivación parcial, y también varían entre 1 y d. Ejemplo:

$$f^J_{KL} = \frac{\partial f^J}{\partial y^K \partial y^L}.$$

Si el sistema es autónomo, $y' = f(y)$, podemos obtener más simetría en las fórmulas sustituyendo k_i por g_i tal que $k_i = f(g_i)$. De esta forma el método (1.9) se transforma (para sistemas autónomos) en

$$\begin{cases} g_i^J = y_n^J + h \sum\limits_{j=1}^{s} a_{ij} f^J(g_j), & i = 1, \ldots, s, \\ y_{n+1}^J = y_n^J + h \sum\limits_{j=1}^{s} b_j f^J(g_j), & J = 1, \ldots, d. \end{cases} \tag{1.12}$$

Así, la única diferencia entre las definiciones de g_i^J e y_{n+1}^J es que las constantes a_{ij} en la primera fórmula pasan a ser las constantes b_j en la segunda. Esta semejanza nos permitirá simplificar notablemente algunos cálculos.

Si el sistema no es autónomo, pero el método de Runge-Kutta es *consistente* y satisface la *condición de suma por filas*

$$c_i = \sum_{j=1}^{s} a_{ij}, \qquad i = 1, 2 \ldots, s, \tag{1.13}$$

el método también se puede escribir en la forma (1.12). Para ello basta con definir

$$\begin{array}{lll} \hat{y}_n^1 = x_n, & \hat{f}^1 = 1, & \hat{g}_i^1 = x_n + c_i h, \\ \hat{y}_n^{J+1} = y_n^J, & \hat{f}^{J+1} = f^J, & \hat{g}_i^{J+1} = y_n^J + h \sum\limits_{j=1}^{s} a_{ij} k_j^J, \quad J = 1, \ldots, d. \end{array}$$

En lo que sigue escribiremos siempre los métodos en la forma (1.12). Por ello, para no dejar fuera los sistemas no autónomos, de aquí en adelante nos restringiremos a métodos que cumplan la condición de suma por filas.

Obtengamos las condiciones de orden 3. Tenemos que calcular las derivadas con respecto a h en $h = 0$ hasta orden 3 de la solución numérica, y_{n+1}^J, y compararlas con las de la teórica, $y^J(x_{n+1}) = y^J(x_n + h)$, que también tenemos que calcular.

Empezamos calculando las derivadas de la solución numérica (siempre bajo la hipótesis de localización). Aplicando la regla de Leibniz para la derivación de un producto se tiene que

$$(y_{n+1}^J)^{(q)}\big|_{h=0} = \sum_{j=1}^{s} b_j q \, (f^J(g_j))^{(q-1)}\big|_{h=0} . \tag{1.14}$$

Así, la derivada primera de la solución numérica es

$$\left(y_{n+1}^J\right)^{(1)}\Big|_{h=0} = \sum_{j=1}^{s} b_j \, (f^J(g_j))\big|_{h=0} = \left(\sum_{j=1}^{s} b_j\right) f^J(y_n). \tag{1.15}$$

La derivada segunda viene dada por

$$(y_{n+1}^J)^{(2)}\big|_{h=0} = 2\sum_{j=1}^{s} b_j \, (f^J(g_j))^{(1)}\big|_{h=0} .$$

Pero

$$\left(f^J(g_j)\right)^{(1)} = \sum_{K=1}^{d} f_K^J(g_j) \left(g_j^K\right)^{(1)} .$$

Usando la semejanza de las definiciones de g_i^J e y_{n+1}^J junto con (1.15) se tiene que

$$(g_i^J)^{(1)}\big|_{h=0} = \left(\sum_{j=1}^{s} a_{ij}\right) f^J(y_n), \tag{1.16}$$

por lo que concluimos que

$$\left(y_{n+1}^J\right)^{(2)}\Big|_{h=0} = 2\sum_{j,k=1}^{s} b_j a_{jk} \sum_{K=1}^{d} f_K^J(y_n) f^K(y_n). \tag{1.17}$$

Para calcular la derivada tercera necesitamos la derivada segunda de $f^J(g_j)$, que viene dada por

$$\left(f^J(g_j)\right)^{(2)} = \sum_{K,L=1}^{d} f_{KL}^J(g_j) \left(g_j^K\right)^{(1)} \left(g_j^L\right)^{(1)} + \sum_{K=1}^{d} f_K^J(g_j) \left(g_j^K\right)^{(2)} .$$

La derivada segunda de g_j^K se obtiene utilizando de nuevo la semejanza entre g_i^J e y_{n+1}^J. Finalmente llegamos a

$$\begin{aligned}\left(y_{n+1}^J\right)^{(3)}\Big|_{h=0} &= 3\sum_{j,k,l=1}^{s} b_j a_{jk} a_{jl} \sum_{K,L=1}^{d} f_{KL}^J(y_n) f^K(y_n) f^L(y_n) \\ &+ 6\sum_{j,k,l=1}^{s} b_j a_{jk} a_{kl} \sum_{K,L=1}^{d} f_K^J(y_n) f_L^K(y_n) f^L(y_n).\end{aligned} \tag{1.18}$$

En cuanto a la solución teórica, se tiene que

$$\left(y^J\right)^{(1)}(x_{n+1})\Big|_{h=0} = f^J(y(x_n+h))\big|_{h=0} = f^J(y_n),$$

$$\left(y^J\right)^{(2)}(x_{n+1})\Big|_{h=0} = \sum_{K=1}^{d} f_K^J(y_n)(y^K)^{(1)}(x_n) = \sum_{K=1}^{d} f_K^J(y_n) f^K(y_n),$$

$$\begin{aligned}\left(y^J\right)^{(3)}(x_{n+1})\Big|_{h=0} &= \sum_{K,L=1}^{d} f_{KL}^J(y_n) f^K(y_n)(y^L)^{(1)}(x_n) \\ &+ \sum_{K,L=1}^{d} f_K^J(y_n) f_L^K(y_n)(y^L)^{(1)}(y_n). \\ = \sum_{K,L=1}^{d} f_{KL}^J(y_n) f^K(y_n) f^L(y_n) &+ \sum_{K,L=1}^{d} f_K^J(y_n) f_L^K(y_n) f^L(y_n).\end{aligned}$$

Comparando las derivadas de la solución numérica y de la teórica obtenemos las *condiciones de orden* 3:

- Orden 1: $\sum_{j=1}^{s} b_j = 1,$
- Orden 2: $2\sum_{j,k=1}^{s} b_j a_{jk} = 1,$
- Orden 3: $3\sum_{j,k,l=1}^{s} b_j a_{jk} a_{jl} = 1, \qquad 6\sum_{j,k,l=1}^{s} b_j a_{jk} a_{kl} = 1.$

Un método de Runge-Kutta tendrá orden de consistencia 3 si y sólo si cumple estas cuatro condiciones.

Observación. La condición de consistencia coincide con la condición de consistencia de orden 1.

Observación. Algunas de estas condiciones se pueden simplificar usando la condición de suma por filas, quedando

$$
\begin{aligned}
&\bullet \text{ Orden 1:} \quad \sum_{j=1}^{s} b_j = 1, \\
&\bullet \text{ Orden 2:} \quad 2\sum_{j=1}^{s} b_j c_j = 1, \\
&\bullet \text{ Orden 3:} \quad 3\sum_{j=1}^{s} b_j c_j^2 = 1, \qquad 6\sum_{j,k=1}^{s} b_j a_{jk} c_k = 1.
\end{aligned}
\tag{1.19}
$$

Observación. Las sumas que aparecen en las condiciones de orden se pueden interpretar como productos de matrices, lo que puede ser cómodo a la hora de hacer cálculos. Por ejemplo, la condición $6 \sum_{j,k=1}^{s} b_j a_{jk} c_k = 1$ se puede reescribir como $6\, b^T Ac = 1$.

Ejemplo. El *método de Heun de orden 3*, de tablero

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

cumple la condición de suma por filas y las cuatro condiciones (1.19). Es por tanto de orden 3.

1.8 Árboles de Butcher

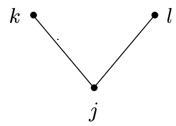
La continuación del proceso anterior para obtener condiciones de orden para ordenes más altos está clara desde el punto de vista teórico. Pero conduce rápidamente a fórmulas muy complicadas, cuya obtención es larga y engorrosa. Conviene utilizar una representación gráfica que simplifique las cuentas: los *árboles de Butcher* (1964).

Empezamos observando que en (1.18) los índices j, k, l aparecen emparejados en a_{jk}, a_{jl}, ... exactamente igual que los subíndices y superíndices J, K, L

en las expresiones f^J_{KL}, f^J_K, $\ldots$ Concretamente, en el primer sumando,

$$3\sum_{j,k,l=1}^{s} b_j a_{jk} a_{jl} \sum_{K,L=1}^{d} f^J_{KL}(y_n) f^K(y_n) f^L(y_n),$$

encontramos la asociación $k \mapsto j$, $l \mapsto j$, a través de a_{jk}, a_{jl} y de f^J_{KL}, yendo de derecha a izquierda en los primeros y de abajo hacia arriba en el último caso. Esta asociación se puede representar gráficamente como

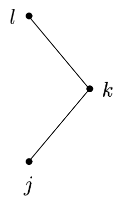
$t_{31} =$ 

(1.20)

En el segundo sumando

$$6\sum_{j,k,l=1}^{s} b_j a_{jk} a_{kl} \sum_{K,L=1}^{d} f^J_K(y_n) f^K_L(y_n) f^L(y_n)$$

encontramos la asociación $k \mapsto j$, $l \mapsto k$, que se puede representar como

$t_{32} =$ 

(1.21)

Observación. En la representación gráfica usamos siempre letras minúsculas. No obstante, si la asociación que estamos representando es la dada por los subíndices y superíndices en las expresiones f^J_{KL}, $\ldots$, habrá que interpretar los índices del gráfico como mayúsculas.

Los grafos t_{31} y t_{32} son ejemplos de *árboles etiquetados con raíz.*

Definición 1.10. *Sea A un conjunto ordenado de índices, $A = \{j_1 < j_2 < \ldots\}$. Denotamos por A_q al subconjunto de los q primeros índices, $A_q = \{j_1 < j_2 < \cdots < j_q\}$. Un* árbol etiquetado con raíz de orden q, *$q \geq 1$, es una aplicación*

$$t : A_q \setminus \{j_1\} \mapsto A_q \qquad \textit{tal que } t(j_i) < j_i, \qquad i = 2, \ldots, q.$$

Notación. Casi siempre, salvo que q sea un número indeterminado o grande, usaremos el conjunto ordenado de índices $A = \{j < k < l < m < \dots\}$.

La aplicación t se representa gráficamente de la siguiente manera: cada índice $j_i \in A_q$ se representa por un vértice, y la correspondencia $t(j_k) = j_i$ mediante una arista que une los vértices j_i y j_k, situándose el vértice j_i por debajo del j_k.

Ejemplos. El árbol de orden 3, $t(k) = j$, $t(l) = j$ es el árbol t_{31} representado en (1.20), y el árbol de orden 3, $t(l) = k$, $t(k) = j$ es el árbol t_{32} representado en (1.21).

El conjunto de todos los árboles etiquetados con raíz de orden q se denota por LT_q (por *labelled trees*). Decimos que j_i es el *hijo* de $t(j_i)$ y que $t(j_i)$ es el *padre* de j_i. Con este lenguaje la condición $t(j_i) < j_i$ dice que el padre está antes que el hijo en el árbol genealógico. El vértice j_1, antecesor de toda la dinastía, recibe el nombre de *raíz* de t. El orden q de un árbol etiquetado con raíz es igual al número de sus vértices.

Definición 1.11. *Se llama* diferencial elemental *asociado al árbol etiquetado* $t \in LT_q$ *a la expresión*

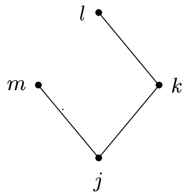
$$F^{J_1}(t)(y) = \sum_{J_2,\dots,J_q=1}^{d} \prod_{i=1}^{q} f^{J_i}_{t^{-1}(J_i)}(y).$$

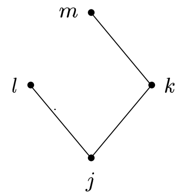
Es decir, el diferencial elemental de un árbol se obtiene sumando sobre todos los índices del árbol menos el de la raíz, siendo cada sumando el producto de q factores f, donde los superíndices recorren todos los vértices de t y los subíndices son los hijos correspondientes.

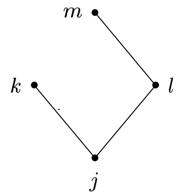
Ejemplos.

- $F^J(t_{31})(y) = \sum\limits_{K,L=1}^{d} f^J_{KL}(y) f^K(y) f^L(y)$;
- $F^J(t_{32})(y) = \sum\limits_{K,L=1}^{d} f^J_K(y) f^K_L(y) f^L(y)$.

Observación. Los tres árboles etiquetados







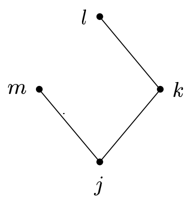
tienen el mismo aspecto topológico. Es más, sus diferenciales elementales

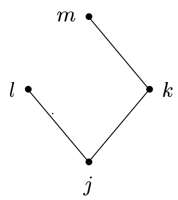
$$\sum_{K,L,M=1}^{d} f^J_{KM} f^K_L f^M f^L, \qquad \sum_{K,L,M=1}^{d} f^J_{KL} f^K_L f^L f^M, \qquad \sum_{K,L,M=1}^{d} f^J_{LK} f^L_M f^K f^M,$$

son todos el mismo (basta con cambiar los nombres de los índices de sumación). Esto motiva la siguiente definición.

Definición 1.12. *Dos árboles* $t, t' \in T_q$ *se dicen* equivalentes *si existe una permutación* $\sigma : A_q \mapsto A_q$ *tal que* $\sigma(j_1) = j_1$ *y* $t \circ \sigma = \sigma \circ t'$ *en* $A_q \setminus \{j_1\}$.

Ejemplo. Los árboles

$t =$ 

$t' =$ 

son equivalentes. La permutación que permite pasar de un árbol a otro es

$$\sigma = \begin{pmatrix} j & k & l & m \\ j & k & m & l \end{pmatrix}.$$

Obsérvese que la imagen por σ de una etiqueta de un nodo en el árbol t es la etiqueta que tiene ese mismo nodo en el árbol t'; con la permutación σ lo único que estamos haciendo es cambiar los nombres de los índices.

Definición 1.13. *Cada clase de equivalencia de árboles etiquetados con raíz de orden q constituye un* árbol con raíz de orden q.

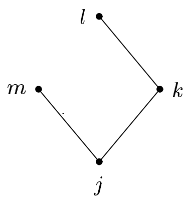
El conjunto de todos los árboles de orden q se denota por T_q. El *orden* de un árbol se define como el orden de un representante. Dado $t \in T_q$, denotamos por $\alpha(t)$ al número de elementos en la clase de equivalencia de t.

Definición 1.14. *Sea $t \in LT_q$, ($A_q = \{j_1 < \cdots < j_q\}$). Definimos $\phi_{j_1}(t)$ por*

$$\phi_{j_1}(t) = \sum_{j_1,\ldots,j_q=1}^{s} \prod_{i=2}^{q} a_{t(j_i)j_i}.$$

Es decir, $\phi_{j_1}(t)$ es la suma a todos los índices del árbol menos el de la raíz de los productos de las $a_{\partial(\text{aristas})}$, donde $\partial($  $) = ij$.

Ejemplo.

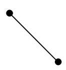
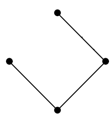
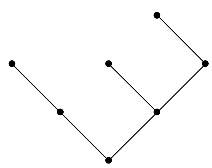
$t =$ 

$\phi_j(t) = \sum_{k,l,m=1}^{s} a_{jk}a_{kl}a_{jm}.$

Observación. Si t y t' son equivalentes entonces $\phi_j(t) = \phi_j(t')$. Nótese que única diferencia son los nombres de los índices de sumación, que son mudos.

Definición 1.15. *Denotamos por $t = [t_1, \ldots, t_m]$ al árbol que produce los árboles $t_1, \ldots, t_m$ cuando se le corta la raíz. Los árboles $t_1, \ldots, t_m$ son las* ramas *del árbol t.*

Ejemplo.

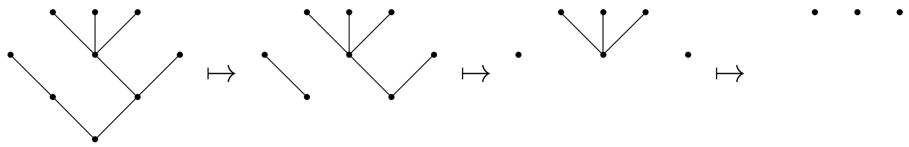
$t_1 =$  $t_2 =$  $t = [t_1, t_2]$ 

Definición 1.16. *Dado $t \in LT_q$ se define*

$$\begin{aligned} \gamma(t) &= 1 && \textit{si } q = 1; \\ \gamma(t) &= q\prod_{i=1}^{r} \gamma(t_i), && \textit{si } t = [t_1, \ldots, t_r],\ q > 1. \end{aligned}$$

Es decir, $\gamma(t)$ es el producto de $\rho(t)$ y de los órdenes de los árboles que van apareciendo al ir quitando las raíces.

Ejemplo.



$\gamma = 9\cdot$ $\quad 2\cdot 6\cdot$ $\quad 1\cdot 4\cdot 1\cdot$ $\quad 1\cdot 1\cdot 1$

Observación. Si t y t' son equivalentes entonces $\gamma(t) = \gamma(t')$.

Las derivadas de las soluciones teórica y numérica se pueden escribir de forma compacta usando el lenguaje de árboles, como se hace en los siguientes teoremas, debidos a Butcher. Posponemos las pruebas, que se pueden omitir en una primera lectura, hasta la sección siguiente.

Teorema 1.17. *La solución teórica satisface*

$$(y^J)^{(q)} = \sum_{t\in LT_q} F^J(t)(y).$$

Teorema 1.18. *La solución numérica satisface*

$$(y^J_{n+1})^{(q)}\big|_{h=0} = \sum_{t\in LT_q} \gamma(t) \sum_{j=1}^{s} b_j \phi_j(t) F^J(t)(y_n). \tag{1.22}$$

Combinando los teoremas 1.18 y 1.17 obtenemos las condiciones de orden.

Teorema 1.19. *Un método de Runge-Kutta es de orden p si y sólo si*

$$\gamma(t) \sum_{j=1}^{s} b_j \phi_j(t) = 1 \tag{1.23}$$

para todos los árboles $t \in T_q$ con $q \le p$.

Prueba. Comparando las derivadas de la solución numérica y de la teórica, vemos que si se cumple (1.23) para cada árbol $t \in LT_q$, el método es de orden p. Ahora bien, puesto que $\gamma(t)$ y $\phi_j(t)$ coinciden para todos los árboles de la misma clase, basta con comprobar las condiciones para los árboles sin etiquetar $t \in T_q$.

Sólo queda probar la implicación recíproca. Para ello basta con comprobar que los diferenciales elementales de los diferentes árboles son independientes. En efecto, dado $t \in T_q$, el sistema

$$y'_{j_i} = \prod_{j_r \in t^{-1}(j_i)} y_{j_r}, \qquad j_i = 1, \ldots, q,$$

verifica que $F^1(t)(y_n) = 1$ y $F^1(u)(y_n) = 0$ para todos los demás árboles $u \in T_q$. $\square$

Ejemplo. Consideramos el método de Runge-Kutta explícito de cuatro etapas de tablero

$$\begin{array}{c|cccc}
0 & & & & \\
1/2 & 1/2 & & & \\
-1 & 1/2 & -3/2 & & \\
1 & 0 & 4/3 & -1/3 & \\
\hline
 & 1/6 & 2/3 & 0 & 1/6.
\end{array} \tag{1.24}$$

Nótese que el tablero cumple la condición de suma por filas.

Hay un único árbol de orden 1,

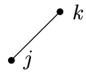
$$\tau = \quad \bullet\, j$$

que da lugar a la condición de orden 1

$$\sum_{j=1}^{s} b_j = 1,$$

que evidentemente se cumple para nuestro tablero.

De orden 2 tenemos el árbol

$t_{21} =$ 

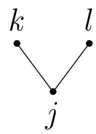
que produce la condición de orden $2 \sum_{j,k=1}^{s} b_j a_{jk} = 1$. Usando la condición de suma por filas la condición se transforma en

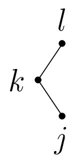
$$2 \sum_{j=1}^{s} b_j c_j = 1.$$

Nuestro tablero la satisface.

Obsérvación. El efecto de la condición de suma por filas es que desaparece la suma en el índice del nodo terminal (sin hijos) k, $\sum_{k=1}^{s} a_{jk}$, y se sustituye por la componente de c correspondiente al padre de k, en este caso c_j.

Hay dos árboles de orden 3,

$t_{31} =$ 

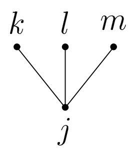
$t_{32} =$ 

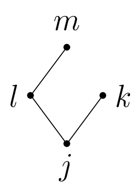
que dan lugar (usando la condición de suma por filas) a las condiciones de orden

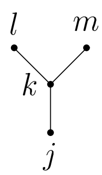
$$3 \sum_{j=1}^{s} b_j c_j^2 = 1, \qquad 6 \sum_{j,k=1}^{s} b_j a_{jk} c_k = 1.$$

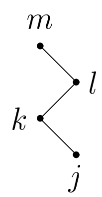
Nuestro método cumple ambas.

Hay cuatro árboles de orden 4,

$t_{41} =$ 

$t_{42} =$ 

$t_{43} =$ 

$t_{44} =$ 

Las condiciones correspondientes son

$$4\sum_{j=1}^{s} b_j c_j^3 = 1, \qquad 8\sum_{j,l=1}^{s} b_j c_j a_{jl} c_l = 1,$$
$$12\sum_{j,k=1}^{s} b_j a_{jk} c_k^2 = 1, \qquad 24\sum_{j,k,l=1}^{s} b_j a_{jk} a_{kl} c_l = 1.$$

Nuestro tablero cumple las condiciones correspondientes a t_{41} y t_{42}, pero no las otras dos, pues

$$\sum_{j,l=1}^{s} b_j c_j a_{jl} c_l = 1/6, \qquad \sum_{j,k=1}^{s} b_j a_{jk} c_k^2 = 0.$$

El método es por tanto de orden 3, pero no de orden 4.

1.9 Demostración de los teoremas de Butcher

En esta sección probamos los teoremas 1.17 y 1.18, que proporcionan respectivamente las derivadas de las soluciones teórica y numérica.

Prueba del teorema 1.17. La prueba es por inducción en q.

Empezamos viendo que el resultado es cierto para $q = 1$. En efecto, solo hay un árbol de orden $q = 1$, que se suele denotar por τ,

$$\tau = \quad \bullet\, j_1$$

para el que se tiene $F^{J_1}(\tau) = f^{J_1}$.

Supongamos ahora que el resultado es cierto para q. Entonces

$$(y^{J_1})^{(q)} = \sum_{t\in LT_q} \sum_{J_2,\dots,J_q=1}^{d} \prod_{i=1}^{q} f^{J_i}_{t^{-1}(J_i)}.$$

Por la regla de la cadena, $\left(f^{J_r}_{t^{-1}(J_r)}\right)' = \sum_{K=1}^{d} f^{J_r}_{t^{-1}(J_r)K}(y^K)^{(1)}$, así que derivando

$(y^{J_1})^{(q)}$ obtenemos

$$\begin{aligned}(y^{J_1})^{(q+1)} &= \sum_{t\in LT_q}\ \sum_{J_2,\dots,J_q=1}^{d}\ \sum_{r=1}^{q}\ \prod_{i=1,i\neq r}^{q} f^{J_i}_{t^{-1}(J_i)} \sum_{K=1}^{d} f^{J_r}_{t^{-1}(J_r)K}(y^K)^{(1)}\\ &= \sum_{t\in LT_q}\ \sum_{r=1}^{q}\ \sum_{J_2,\dots,J_{q+1}=1}^{d}\ \prod_{i=1,i\neq r}^{q} f^{J_i}_{t^{-1}(J_i)} f^{J_r}_{t^{-1}(J_r)J_{q+1}} f^{J_{q+1}}.\end{aligned}$$

La clave para proseguir es observar que todo árbol de orden $q+1$ se obtiene añadiendo en algún lugar una nueva arista a un árbol de orden q. Esto permite establecer la aplicación biyectiva

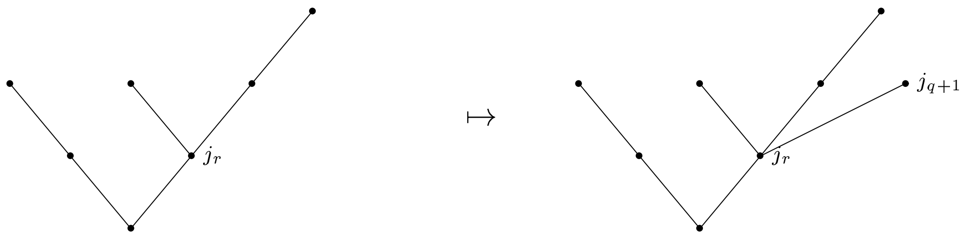
$$\begin{array}{ccc} LT_q\times A_q & \mapsto & LT_{q+1}\\ (t,j_r) & \mapsto & t', \end{array}$$

donde

$$t'|_{A_q}=t, \qquad t'(j_{q+1})=j_r.$$

Es decir, el árbol t' de orden $q+1$ se obtiene a partir del árbol t de orden q añadiendo una arista en el nodo j_r.

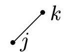
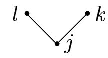
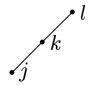
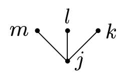
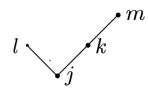
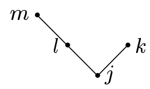
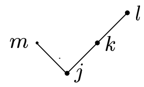
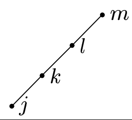
Veamos un ejemplo gráficamente:



Con ayuda de esta aplicación podemos escribir

$$\begin{aligned}(y^{J_1})^{(q+1)} &= \sum_{t'\in LT_{q+1}}\ \sum_{J_2,\dots,J_{q+1}=1}^{d}\ \prod_{i=1,i\neq r}^{q} f^{J_i}_{(t')^{-1}(J_i)} f^{J_r}_{(t')^{-1}(J_r)} f^{J_{q+1}}_{(t')^{-1}(J_{q+1})}\\ &= \sum_{t'\in LT_{q+1}}\ \sum_{J_2,\dots,J_{q+1}=1}^{d}\ \prod_{i=1}^{q+1} f^{J_i}_{(t')^{-1}(J_i)} = \sum_{t'\in LT_{q+1}} F^{J_1}(t')(y). \qquad \square\end{aligned}$$

Nuestro siguiente objetivo es calcular las derivadas de la solución numérica. Usaremos la fórmula (1.14), por lo que necesitamos calcular $(f^J(g_j))^{(q-1)}$, que es lo que hacemos a continuación.

Derivadas	Gráficamente
$(f^J(g_j))^{(0)} = f^J$	
$(f^J(g_j))^{(1)} = \sum_{K=1}^{d} f^J_K(g_j)(g_j^K)^{(1)}$	
$(f^J(g_j))^{(2)} = \sum_{K,L=1}^{d} f^J_{KL}(g_j)(g_j^K)^{(1)}(g_j^L)^{(1)}$	
$+ \sum_{K=1}^{d} f^J_K(g_j)(g_j^K)^{(2)}$	
$(f^J(g_j))^{(3)} = \sum_{K,L,M=1}^{d} f^J_{KLM}(g_j)(g_j^K)^{(1)}(g_j^L)^{(1)}(g_j^M)^{(1)}$	
$+ \sum_{K,L=1}^{d} f^J_{KL}(g_j)(g_j^K)^{(2)}(g_j^L)^{(1)}$	
$+ \sum_{K,L=1}^{d} f^J_{KL}(g_j)(g_j^K)^{(1)}(g_j^L)^{(2)}$	
$+ \sum_{K,M=1}^{d} f^J_{KM}(g_j)(g_j^K)^{(2)}(g_j^M)^{(1)}$	
$+ \sum_{K=1}^{d} f^J_K(g_j)(g_j^K)^{(3)}$	

Ahora debería estar claro el proceso. Cada vez que diferenciamos tenemos que:

(i) derivar el primer factor $f^J_{K\ldots}$; es decir, añadir una nueva rama a la raíz;

(ii) aumentar el número de derivadas de cada una de las g en una unidad; es decir, alargar la correspondiente rama.

Todos los árboles que se obtienen de esta forma son aquellos árboles "especiales" que no tienen ramificaciones salvo en la raíz; y se obtienen todos.

Definición 1.20. *Un árbol* $t \in LT_q$ *se dice* especial *si para* $i = 2, \ldots, q$, *se tiene que* $\#t^{-1}(j_i) \in \{0, 1\}$.

Al conjunto de árboles especiales se lo denota LS_q.

Tenemos el siguiente resultado, cuya prueba es inmediata empleando el lenguaje de árboles.

Lema 1.21. *(Fórmula de Faà di Bruno) Para* $q \geq 1$ *se tiene*

$$(f^J(g_j))^{(q-1)} = \sum_{\substack{u \in LS_q, \\ u = [u_1, \ldots, u_r]}} \sum_{K_1, \ldots, K_r}^{d} f^J_{K_1, \ldots, K_r}(g_j)(g_j^{K_1})^{(\rho(u_1))} \cdots (g_j^{K_r})^{(\rho(u_r))},$$

donde $r = r(u)$ *es el número de ramas que dejan la raíz y* $\rho(u_1), \ldots, \rho(u_r)$ *es el número de nodos en cada una de esas ramas.*

Observación. Se tiene que $q = 1 + \rho(u_1) + \cdots + \rho(u_r)$.

Prueba del teorema 1.18. Debido a la semejanza de las definiciones de g_i^J e y_{n+1}^J, probar (1.22) es equivalente a demostrar que

$$(g_i^J)^{(q)}\big|_{h=0} = \sum_{t \in LT_q} \gamma(t) \sum_{j=1}^{s} a_{ij} \phi_j(t) F^J(t)(y_n). \tag{1.25}$$

Probaremos esta última fórmula por inducción en q.

Veamos en primer lugar que el resultado es cierto para $q = 1$. Para el único árbol τ de orden 1 tenemos que $\phi_j(\tau) = 1$ y $F^J(\tau) = f^J$. Así que usando (1.16) tenemos que

$$\begin{aligned}
\sum_{t \in LT_1} \gamma(t) \sum_{j=1}^{s} a_{ij} \phi_j(t) F^J(t)(y_n) &= \sum_{j=1}^{s} a_{ij} \phi_j(\tau) F^J(\tau)(y_n) \\
&= \sum_{j=1}^{s} a_{ij} f^J(y_n) = (g_i^J)^{(1)}\big|_{h=0},
\end{aligned}$$

y por tanto el resultado es cierto para $q = 1$.

Supongamos ahora el resultado cierto para todo número de derivadas menor o igual que $q-1$, y veamos que es cierto para q derivadas.

Usando la regla de Leibniz y la fórmula de Faà di Bruno tenemos que

$$\begin{aligned}
(g_i^J)^{(q)}\big|_{h=0} &= q\sum_{j=1}^{s} a_{ij}\ (f^J(g_j))^{(q-1)}\big|_{y=y_n} \\
&= q\sum_{j=1}^{s} \quad a_{ij} \sum_{\substack{u\,\in\, LS_q,\\ u\,=\,[u_1,\ldots,u_r]}} \ \sum_{K_1,\ldots,K_r}^{d} f^J_{K_1,\ldots,K_r}(g_j)(g_j^{K_1})^{(\rho(u_1))}\cdots(g_j^{K_r})^{(\rho(u_r))}
\end{aligned}$$

Dado que $\rho(u_1),\ldots,\rho(u_r)<q$, podemos aplicar la hipótesis de inducción para obtener

$$\begin{aligned}
(g_i^J)^{(q)}\big|_{h=0} = &\sum_{\substack{u\,\in\, LS_q,\\ u\,=\,[u_1,\ldots,u_r]}} \ \sum_{t_1\in LT_{\rho(u_1)}}\cdots\sum_{t_r\in LT_{\rho(u_r)}} q\gamma(t_1)\cdots\gamma(t_r)\times \\
&\times\sum_j a_{ij}\sum_{k_1} a_{jk_1}\phi_{k_1}(t_1)\cdots\sum_{k_r} a_{jk_r}\phi_{k_r}(t_r)\sum_{K_1,\ldots,K_r}^{d} f^J_{K_1,\ldots,K_r}F^{K_1}(t_1)\cdots F^{K_r}(t_r).
\end{aligned}$$

Sea $u=[u_1,\ldots,u_r]\in LS_q$, y $t_l\in LT_{\rho(u_l)}$, para $l=1,\ldots,r$. Definimos una biyección

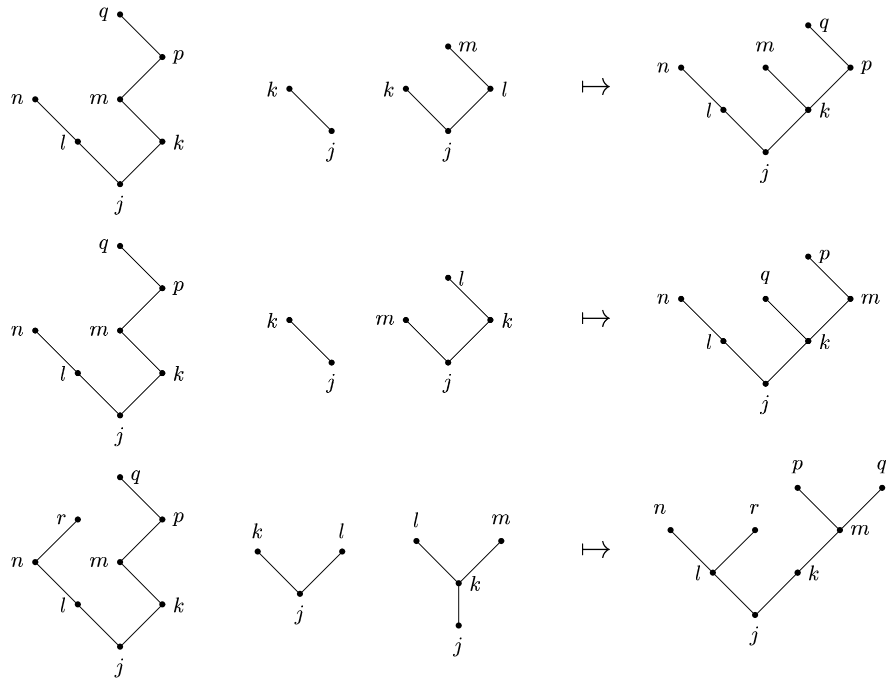
$$(u,t_1,\ldots,t_r)\mapsto t\in LT_q$$

de forma que

$$\begin{aligned}
&\gamma(t)=q\gamma(t_1)\cdots\gamma(t_r),\\
&\phi_j(t)=\sum_j a_{ij}\sum_{k_1} a_{jk_1}\phi_{k_1}(t_1)\cdots\sum_{k_r} a_{jk_r}\phi_{k_r}(t_r),\\
&F^J(t)=\sum_{K_1,\ldots,K_r}^{d} f^J_{K_1,\ldots,K_r}F^{K_1}(t_1)\cdots F^{K_r}(t_r).
\end{aligned}$$

El árbol t se obtiene sustituyendo las ramas de u por los árboles $t_1,\ldots,t_r$ y tomando las etiquetas correspondientes de forma natural, es decir, en el mismo orden.

Ejemplos.



Usando esta biyección llegamos finalmente a (1.25). □

1.10 Métodos explícitos. Orden obtenible

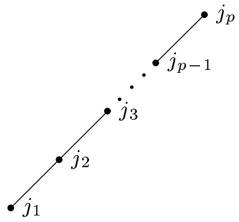
Hemos obtenido las condiciones para que un método de Runge-Kutta tenga un orden dado. Ahora dirigimos nuestra atención al problema de encontrar soluciones a dichas condiciones de orden. Nos preguntamos qué orden se puede lograr con un método explícito de s etapas.

Nuestro primer resultado (debido a Butcher) da una cota superior para el orden que se puede conseguir con un método explícito de s etapas. La prueba es un ejemplo de la potencia del enfoque de árboles de Butcher.

Teorema 1.22. *Un método de Runge-Kutta explícito de s etapas no puede tener orden mayor que s.*

Prueba. Supongamos que el método de s etapas tiene orden p. Consideramos el árbol de orden p

$t =$



Evidentemente $\gamma(t) = p\,!$ y por otra parte,

$$\sum_{j_1=1}^{s} b_{j_1}\phi_{j_1}(t) = \sum_{j_1,\ldots,j_p=1}^{s} b_{j_1} a_{j_1 j_2} a_{j_2 j_3} \cdots a_{j_{p-1} j_p}.$$

Como el método es explícito, $a_{ij} = 0$ para $j \geq i$. Así, $\sum\limits_{j_1=1}^{s} b_{j_1}\phi_{j_1}(t) = 0$ a menos que exista una sucesión $j_1, j_2, \ldots, j_p$ de enteros $1, 2, \ldots, s$ tal que

$$j_1 > j_2 > j_3 > \cdots > j_p.$$

Dicha sucesión sólo puede existir si $p \leq s$. En caso contrario

$$\gamma(t) \sum_{j_1=1}^{s} b_{j_1}\phi_{j_1}(t) = 0 \neq 1$$

y no se cumple la condición de orden correspondiente al árbol t. $\square$

¿Se puede conseguir orden $p = s$? Para $s = 1, 2, 3, 4$ se puede ver que sí, y es relativamente fácil encontrar familias de soluciones a las correspondientes condiciones de orden.

Ejemplo. Sea $s = 1$. La condición de orden 1 es en este caso $b_1 = 1$. El único método explícito de una etapa y orden 1 es por tanto el método de Euler.

Ejemplo. Sea $s = 2$. Las condiciones de orden 1 y 2 son en este caso $b_1 + b_2 = 1$ y $b_1 c_1 + b_2 c_2 = 1$ respectivamente (estamos suponiendo que se cumple la condición de suma por filas). Si el método es explícito entonces $c_1 = 0$, y la condición de orden 2 queda que $b_2 c_2 = 1/2$. Hay pues una familia uniparamétrica de métodos explícitos de 2 etapas y orden 2,

$$c_2 = \lambda, \qquad b_1 = 1 - \frac{1}{2\lambda}, \qquad b_2 = \frac{1}{2\lambda}, \qquad \lambda \in \mathbb{R},\ \lambda \neq 0.$$

¿Qué pasa con $s = 5$? Para que un método sea de orden 5 se tienen que satisfacer 17 condiciones de orden. Si $s = 5$ tenemos sólo 15 parámetros libres. Sin embargo, el sistema es no lineal, así que podría haber solución. Pero no es así.

Teorema 1.23. *Para $p \geq 5$ no existe ningún método de Runge-Kutta explícito de orden p con $s = p$ etapas.*

La prueba se puede encontrar en [HNW].

Este es un primer ejemplo de *barrera de Butcher*. Hay más barreras. Por ejemplo, se tiene el siguiente teorema.

Teorema 1.24. *(i) Para $p \geq 7$ no existe ningún método de Runge-Kutta explícito de orden p con $s = p + 1$ etapas.*

(ii) Para $p \geq 8$ no hay ningún método de Runge-Kutta explícito con $s = p + 2$ etapas.

En la siguiente tabla recogemos el estado del arte:

Orden	1	2	3	4	5	6	7	8	9	10
s mínimo	1	2	3	4	6	7	9	11	$12 \leq s \leq 17$	$13 \leq s \leq 17$

1.11 Estimaciones de error

Supongamos que $f \in C^p$. Por un lado tenemos por (1.9) que

$$R_n = y(x_n + h) - y(x_n) - h \sum_{i=1}^{s} b_i k_i. \tag{1.26}$$

Por otra parte, usando la fórmula de Taylor se tiene que

$$\begin{aligned} &y(x_n + h) = y(x_n) + y'(x_n)h + y''(x_n)\tfrac{h^2}{2} + \cdots + y^{(p+1)}(x_n + \overline{\theta}h)\tfrac{h^{p+1}}{(p+1)!}, \\ &k_i(h) = k_i(0) + k_i'(0)h + \cdots + k_i^{(p)}(\overline{\theta}_i h)\tfrac{h^p}{p!}. \end{aligned}$$

Si el método (1.12) satisface las condiciones de orden hasta orden p, entonces los términos hasta orden p en (1.26) se anulan. Se tiene entonces que

$$\|R_n\| \le h^{p+1}\left(\frac{1}{(p+1)!}\max_{t\in[0,1]}\|y^{(p+1)}(x_n+th)\| + \frac{1}{p!}\sum_{i=1}^{s}|b_i|\max_{t\in[0,1]}\|k_i^{(p)}(th)\|\right).$$

Incluso suponiendo conocidas las derivadas que aparecen en la fórmula anterior, esta cota es con frecuencia mucho mayor que el tamaño que realmente tiene el error de truncación, y no es por tanto adecuada para estimarlo. Es mucho más realista considerar el primer término distinto de cero del desarrollo de Taylor del error. El trabajo realizado hasta ahora nos da dicho término.

Teorema 1.25. *Si un método de Runge-Kutta satisface las condiciones de orden p, y $f \in C^{p+1}$, entonces el error de truncación satisface*

$$R_n^J = \frac{h^{p+1}}{(p+1)!}\sum_{t\in T_{p+1}}\alpha(t)e(t)F^J(t)(y_n) + O(h^{p+2}),$$

donde

$$e(t) = 1 - \gamma(t)\sum_{j=1}^{s} b_j\phi_j(t).$$

Las expresiones $e(t)$ reciben el nombre de *coeficientes de error.*

Ejemplo. Calculemos el término principal del error al aplicar el método de tablero (1.24) al problema lineal

$$y' = \Lambda y, \qquad y(0) = \eta, \tag{1.27}$$

donde Λ es una matriz $d \times d$ con elementos λ^{IJ} constantes. La verdadera solución es $y(x) = e^{x\Lambda}\eta$.

Se tiene que

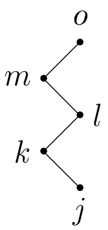
$$e(t_{41}) = 0, \quad e(t_{42}) = -1/3, \quad e(t_{43}) = 1, \quad e(t_{44}) = 0,$$

de manera que los únicos diferenciales elementales que hay que calcular son los correspondientes a los árboles t_{42} y t_{43},

$$F^J(t_{42}) = \sum_{K,L,M=1}^{d} f^J_{KM} f^K_L f^L f^M, \qquad F^J(t_{43}) = \sum_{K,L,M=1}^{d} f^J_K f^K_{LM} f^L f^M.$$

En nuestro caso f es lineal; cualquier derivada segunda suya se anula, y por tanto $F^J(t_{42}) = 0$, $F^J(t_{43}) = 0$. Concluimos que $R^J_n = O(h^5)$, y el método es al menos de orden 5 para problemas del tipo (1.27).

Veamos cuál es el término de orden 5. Todos los diferenciales que tengan derivadas de orden superior a uno se van a anular, por ser f lineal. Habrá derivadas de orden superior cuando haya alguna bifurcación en el árbol. Así pues, el único árbol de los nueve de orden 5 cuyo diferencial elemental no se anula es

$t_{59} =$ 

Teniendo en cuenta que $f^J_K = \lambda^{JK}$, obtenemos que

$$F^J(t_{59}) = \sum_{K,L,M,O=1}^{d} f^J_K f^K_L f^L_M f^M_O f^O = (\Lambda^5 y)^J.$$

Por otra parte

$$e(t_{59}) = 1 - \gamma(t_{59}) \sum_{j=1}^{s} b_j \phi_j(t_{59}).$$

El orden del árbol lineal t_{59} es mayor que el número de etapas. Así pues (ver la prueba del Teorema 1.22), $\sum_{j=1}^{s} b_j \phi_j(t_{59}) = 0$, y por tanto $e(t_{59}) = 1$. Como $\alpha(t_{59}) = 1$, concluimos finalmente que

$$R^J_n = \frac{h^5}{5!}(\Lambda^5 y_n)^J + O(h^{p+2}).$$

La estimación del error simplificada,

$$R_n^J \approx \frac{h^{p+1}}{(p+1)!} \sum_{t \in T_{p+1}} \alpha(t) e(t) F^J(t)(y_n), \tag{1.28}$$

que se contenta con el término principal del error, es también de poco interés práctico, pues requiere calcular varias derivadas parciales de orden superior. Y precisamente la principal ventaja de los métodos de Runge-Kutta (comparados con métodos de Taylor), es que no exigen calcular derivadas.

Sin embargo, necesitamos una estimación práctica para el error. Por un lado para asegurarnos de que los tamaños de paso h_i son suficientemente pequeños como para tener la precisión deseada en los resultados, y por otra parte para asegurarnos de que los tamaños de paso son suficientemente grandes como para evitar trabajo computacional innecesario.

1.12 Extrapolación de Richardson

La idea es utilizar de forma cuidadosa el comportamiento del error como función de h. Esta idea, debida a Richardson, se utiliza en otros contextos, por ejemplo para estimar el error en fórmulas de cuadratura numérica. Expliquémosla.

Sea $a(h)$ una aproximación a b cuando h tiende a 0. Queremos estimar el error $b - a(h)$ sin usar b, pues no lo conocemos. Supongamos que sabemos que b coincide con $a(h)$ hasta orden p,

$$b - a(h) = Ch^{p+1} + O(h^{p+2}).$$

Si aplicamos esta fórmula con paso $2h$, se tiene que

$$b - a(2h) = 2^p Ch^p + O(h^{p+1}).$$

de donde, eliminando Ch^{p+1} se tiene que

$$b = \frac{2^{p+1}a(h) - a(2h)}{2^{p+1} - 1} + O(h^{p+2}),$$

y de ahí la estimación de error, conocida como *estimación de Richardson*,

$$b - a(h) = \frac{a(h) - a(2h)}{2^{p+1} - 1} + O(h^{p+2}).$$

Observación. La cantidad

$$\frac{2^{p+1}a(h) - a(2h)}{2^{p+1} - 1},$$

es una aproximación a b de orden $p+1$.

¿Cómo aplicar la idea de Richardson para estimar el error de truncación? Supongamos dado un "valor inicial" $(x_n, y(x_n))$ y un tamaño de paso h. Si el método es de orden p, entonces

$$y(x_{n+1}) - y_{n+1} = \psi(y(x_n))h^{p+1} + O(h^{p+2}), \tag{1.29}$$

donde $\psi(y(x_n))h^{p+1}$ es el término principal del error de truncación,

$$\psi(y)h^{p+1} = \frac{h^{p+1}}{(p+1)!} \sum_{t \in T_{p+1}} \alpha(t)e(t)F^J(t)(y).$$

Damos otro paso de tamaño h. Si en un paso el error es aproximadamente $\psi(y(x_n))h^{p+1}$, en dos será aproximadamente $2\psi(y(x_n))h^{p+1}$. En efecto, probaremos más adelante que la aproximación numérica y_{n+2} verifica que

$$y(x_{n+2}) - y_{n+2} = 2\psi(y(x_n))h^{p+1} + O(h^{p+2}). \tag{1.30}$$

Por otra parte, si empezando en $(x_n, y(x_n))$ damos un paso de tamaño $2h$, la aproximación numérica resultante w satisface que

$$y(x_{n+2}) - w = \psi(y(x_n))2^{p+1}h^{p+1} + O(h^{p+2}).$$

Restando esta igualdad de la anterior y despejando se tiene que

$$2\psi(y(x_n))h^{p+1} = \frac{y_{n+2} - w}{2^p - 1} + O(h^{p+2}).$$

En resumen: hemos probado que el error de y_{n+2} satisface

$$y(x_{n+2}) - y_{n+2} = \frac{y_{n+2} - w}{2^p - 1} + O(h^{p+2}), \tag{1.31}$$

lo que da una forma simple de estimar el error. Por otra parte, podemos obtener un valor mejorado, $\bar{y}_{n+2}$, para la aproximación de $y(x_{n+2})$, mediante

$$\bar{y}_{n+2} = y_{n+2} + \frac{y_{n+2} - w}{2^p - 1}$$

aproximación que resulta ser de orden $p+1$.

Sólo queda probar (1.30). Para ello descomponemos $y(x_{n+2}) - y_{n+2}$ como suma de dos términos,

$$y(x_{n+2}) - y_{n+2} = (y(x_{n+2}) - \hat{y}(x_{n+2})) + (\hat{y}(x_{n+2}) - y_{n+2}), \tag{1.32}$$

donde $\hat{y}$ es la solución de

$$\begin{cases} \hat{y}' = f(x, \hat{y}), \\ \hat{y}(x_{n+1}) = y_{n+1}. \end{cases}$$

El primer sumando de (1.32) es la diferencia de dos soluciones exactas de la misma ecuación con distintos valores iniciales en $x = x_{n+1}$. Tenemos que

$$\begin{array}{l} y(x) = y(x_{n+1}) + \int_{x_{n+1}}^{x} f(s, y(s))\, ds, \\ \hat{y}(x) = y_{n+1} + \int_{x_{n+1}}^{x} f(s, \hat{y}(s))\, ds. \end{array}$$

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

$$\begin{array}{rcl} \|y(x) - \hat{y}(x)\| & \leq & \|y(x_{n+1}) - y_{n+1}\| + \int_{x_{n+1}}^{x} \|f(s, y(s)) - f(s, \hat{y}(s))\|\, ds \\ & \leq & \|y(x_{n+1}) - y_{n+1}\| + L \int_{x_{n+1}}^{x} \|y(s) - \hat{y}(s)\|\, ds. \end{array} \tag{1.33}$$

Para $\epsilon > 0$ arbitrario definimos

$$g_\epsilon(x) = \epsilon + \|y(x_{n+1}) - y_{n+1}\| + L \int_{x_{n+1}}^{x} \|y(s) - \hat{y}(s)\|\, ds.$$

Así pues

$$g_\epsilon'(x) = L\|y(x) - \hat{y}(x)\| \leq L g_\epsilon(x).$$

Dado que $g_\epsilon(x) > 0$ para todo $x \geq x_{n+1}$, tenemos la desigualdad $g'_\epsilon/g_\epsilon \leq L$, que integrada en (x_{n+1}, x) produce $g_\epsilon(x) \leq g_\epsilon(x_{n+1})e^{L(x-x_{n+1})}$. Usando una vez más que $\|y(x) - \hat{y}(x)\| \leq g_\epsilon(x)$, y pasando al límite $\epsilon \to 0^+$ llegamos a

$$\|y(x) - \hat{y}(x)\| \leq \|y(x_{n+1}) - y_{n+1}\|e^{L(x-x_{n+1})}. \tag{1.34}$$

Observación. Esta desigualdad prueba que el problema de valor inicial (1.1) tiene solución única, y que ésta dependen de manera continua del dato inicial.

Observación. El paso de la estimación (1.33) a la estimación puntual (1.34) es lo que se conoce en la literatura como Lema de Gronwall, [QV].

Procediendo como antes, pero pasando $y(x_{n+1}) - y_{n+1}$ al primer miembro y sustituyendo $x = x_{n+2}$,

$$\begin{aligned}
&\|y(x_{n+2}) - \hat{y}(x_{n+2}) - (y(x_{n+1}) - y_{n+1})\| \leq L \int_{x_{n+1}}^{x_{n+2}} \|y(s) - \hat{y}(s)\|\, ds \\
&\quad \leq L\|y(x_{n+1}) - y_{n+1}\| \int_{x_{n+1}}^{x_{n+2}} e^{L(s-x_{n+1})}\, ds = \|y(x_{n+1}) - y_{n+1}\|(e^{Lh} - 1) \\
&\quad = O(h^{p+2}).
\end{aligned}$$

Así, usando (1.29) tenemos que

$$y(x_{n+2}) - \hat{y}(x_{n+2}) = \psi(y(x_n))h^{p+1} + O(h^{p+2}). \tag{1.35}$$

Para estimar el segundo sumando observamos que

$$\hat{y}(x_{n+2}) - y_{n+2} = \psi(\hat{y}(x_{n+1}))h^{p+1} + O(h^{p+2}).$$

Dado que $\hat{y}(x_{n+1}) = y_{n+1} = y(x_n) + O(h)$, se tiene que

$$\hat{y}(x_{n+2}) - y_{n+2} = \psi(y(x_n))h^{p+1} + O(h^{p+2}). \tag{1.36}$$

Sumando (1.35) y (1.36) se obtiene finalmente (1.30).

1.13 Cambio de paso

Queremos escribir un código que ajuste automáticamente la longitud del paso de forma que el error local se ajuste a una tolerancia prescrita.

Una vez elegida una longitud de paso inicial h, el programa calcula dos pasos de tamaño h y un paso de tamaño $2h$. Se estima entonces el error en norma infinito usando (1.31) en la forma

$$\texttt{ERROR} = \frac{1}{2^p - 1} \max_{i=1,\dots,m} |y^i_{n+2} - w^i|.$$

También son de uso frecuente otras normas, como por ejemplo la euclídea.

Una vez estimado el error se compara con la tolerancia prescrita, TOL. Entonces, si $\texttt{ERROR} \leq \texttt{TOL}$, se aceptan los dos pasos calculados y la solución se avanza con y_{n+2}, o bien con $\bar{y}_{n+2}$ (en este caso se dice que se está haciendo extrapolación local). Se intenta ahora un nuevo paso con una longitud de paso h'. Dicha longitud se intenta escoger de forma óptima, es decir, la mayor que produzca un ERROR menor que TOL.

Queremos que

$$2\psi(y(x_{n+2}))(h')^{p+1} = \texttt{TOL}.$$

Pero

$$2\psi(y(x_{n+2}))(h')^{p+1} \approx 2\psi(y(x_n))h^{p+1} \left(\frac{h'}{h}\right)^{p+1} \approx \texttt{ERROR} \left(\frac{h'}{h}\right)^{p+1}.$$

Así,

$$h' = h \left(\frac{\texttt{TOL}}{\texttt{ERROR}}\right)^{\frac{1}{p+1}} \tag{1.37}$$

es la elección "óptima".

Para tener un buen código hay que poner un poco más de cuidado. Se suele multiplicar el segundo miembro de (1.37) por un factor de seguridad FAC, normalmente $\texttt{FAC} = 0.9$, de forma que el ERROR sea aceptable la siguiente vez con probabilidad alta. Tampoco se suelen permitir incrementos de paso muy grandes, con el fin de hacer el método más seguro. Así, por ejemplo podemos poner

$$h' = h \min\left(\mathtt{FACMAX}, \mathtt{FAC}\left(\frac{\mathtt{TOL}}{\mathtt{ERROR}}\right)^{\frac{1}{p+1}}\right).$$

`FACMAX` se suele tomar entre 1.5 y 5.

Si $\mathtt{ERROR} > \mathtt{TOL}$, los dos pasos se rechazan y se vuelven a repetir los cálculos con el nuevo tamaño de paso h'.

1.14 Pares encajados

La estimación de Richardson funciona bien, pero es computacionalmente cara; si un método de Runge-Kutta explícito tiene s etapas, se necesitan otras $s-1$ evaluaciones de función (la primera etapa en el cálculo de w ya se conocía) para estimar el error.

Se utiliza una idea alternativa, debida a Merson, que pasamos a explicar.

Sea un Runge-Kutta que cumple las condiciones de orden p

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

Si tuviéramos otro método de Runge-Kutta con las mismas etapas,

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

pero de distinto orden, $\hat{p} > p$, podemos calcular

$$\begin{aligned} y_{n+1} &= y_n + h \textstyle\sum_{i=1}^{s} b_i k_i, \\ \hat{y}_{n+1} &= y_n + h \textstyle\sum_{i=1}^{s} \hat{b}_i k_i. \end{aligned}$$

No hay que hacer nuevas evaluaciones de función, pues las etapas k_i coinciden.

Como siempre, supondremos que se cumple la hipótesis de localización, $y(x_n) = y_n$. Sabemos que los errores locales satisfacen

$$\begin{aligned} R_n &= y(x_{n+1}) - y_{n+1} + O(h^{p+2}), \\ \hat{R}_n &= y(x_{n+1}) - \hat{y}_{n+1} + O(h^{\hat{p}+2}). \end{aligned}$$

Eliminando $y(x_{n+1})$ y usando que $\hat{R}_n = O(h^{\hat{p}+1})$ y que $p+2 \leq \hat{p}+1$, se tiene que

$$R_n = \hat{y}_{n+1} - y_{n+1} + O(h^{p+2}).$$

Esto justifica la estimación

$$R_n \approx \hat{y}_{n+1} - y_{n+1} = h \sum_{i=1}^{s} (\hat{b}_i - b_i) k_i. \tag{1.38}$$

Así, podemos tomar

$$\mathtt{ERROR} = \max_{i=1,\ldots,m} |\hat{y}_{n+1}^i - y_{n+1}^i|,$$

y se sigue como en el caso de la extrapolación de Richardson.

Veamos dos ejemplos debidos a Fehlberg. El primero es el RKF2(3), de tablero

$$\begin{array}{c|ccc} 0 & 0 & & \\ 1 & 1 & & \\ \frac{1}{2} & \frac{1}{4} & \frac{1}{4} & \\ \hline y_{n+1} & \frac{1}{2} & \frac{1}{2} & 0 \\ \hline \hat{y}_{n+1} & \frac{1}{6} & \frac{1}{6} & \frac{4}{6}. \end{array}$$

El método de orden 2 que se usa para avanzar es en realidad de dos etapas, mientras que el de orden 3 que se usa para estimar es de tres etapas. El segundo par encajado que queremos presentar es el RKF2(3)B,

$$
\begin{array}{c|cccc}
0 & 0 & & & \\
\frac{1}{4} & \frac{1}{4} & & & \\
\frac{27}{40} & -\frac{189}{800} & \frac{729}{800} & & \\
1 & \frac{214}{891} & \frac{1}{33} & \frac{650}{891} & \\
\hline
y_{n+1} & \frac{214}{891} & \frac{1}{33} & \frac{650}{891} & 0 \\
\hline
\hat{y}_{n+1} & \frac{533}{2106} & 0 & \frac{800}{1053} & -\frac{1}{78}.
\end{array}
$$

El método de avance es de 3 etapas y orden 2, y el de estimación es de 4 etapas y orden 3.

Notación. "Nombre $p(q)$" quiere decir que el orden de y_{n+1} (el método que se utiliza para avanzar) es p, y el orden de $\hat{y}_{n+1}$ (el método que se usa para estimar) es q.

Comparando los métodos RKF2(3) y RKF2(3)B podríamos pensar que el primero es mucho mejor, por emplear una etapa menos para conseguir el mismo orden. Sin embargo, el método RKF2(3)B tiene la propiedad FSAL (First Same As Last),

$$a_{si} = b_i \qquad i = 1, \ldots, s.$$

Por tanto, recordando que es un método explícito,

$$y_{n+1} = y_n + h \sum_{i=1}^{s-1} b_i k_i = y_n + h \sum_{i=1}^{s-1} a_{si} k_i,$$

y se tiene que

$$k_s = f(t + h, y_n + h \sum_{i=1}^{s-1} a_{si} k_i)$$

coincide con la primera etapa del paso siguiente,

$$k_1 = f(t + h, y_{n+1}).$$

Es decir, la primera etapa de cada paso coincide con la última del paso anterior. Por tanto la primera etapa sólo hay que calcularla en el primer paso, y desde el punto de vista computacional es como si el método tuviera una etapa menos.

Aunque la estimación (1.38) da el error de truncación para el método del par encajado de orden más bajo, si h es suficientemente pequeño el error para el método de orden alto será aún menor; así que, ¿por qué no avanzar con el método de orden más alto? Esto es lo que hacen los métodos de *Dormand y Prince* DOPRI5(4) y DOPRI8(7), que son algunos de los mejores métodos existentes hasta el momento.

Habrá que tomar una precaución. Queremos que el error de cometido al avanzar, dado aproximadamente por (1.28), sea lo menor posible. Por tanto, si se avanza con el método de orden alto, en el diseño del algoritmo conviene minimizar los coeficientes de error de dicho método, en lugar de los del método de orden bajo, que al fin y al cabo sólo se usa para controlar el paso.

1.15 Diagramas de eficiencia

¿Cómo podemos comparar la eficiencia de distintos métodos? Una manera muy conveniente es elaborar un *diagrama de eficiencia*. En estos diagramas se representa, en escalas logarítmicas, la precisión conseguida, medida, por ejemplo, con la norma infinito del error en el punto final del intervalo de integración, frente al coste computacional, medido en evaluaciones de función, o en tiempo de CPU. En el caso de métodos de paso fijo podemos obtener diferentes puntos del diagrama cambiando el paso h. En el caso de métodos de paso variable lo que haremos para obtener distintos puntos es cambiar la tolerancia. Veamos algunos ejemplos.

En la Figura 1.1 representamos `error`≡norma infinito del error en $b = 2$ frente a `fc`≡número de evaluaciones de función, en escalas logarítmicas para el problema lineal escalar

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

Cuanta mayor precisión, más abajo estaremos en el diagrama, y cuanto

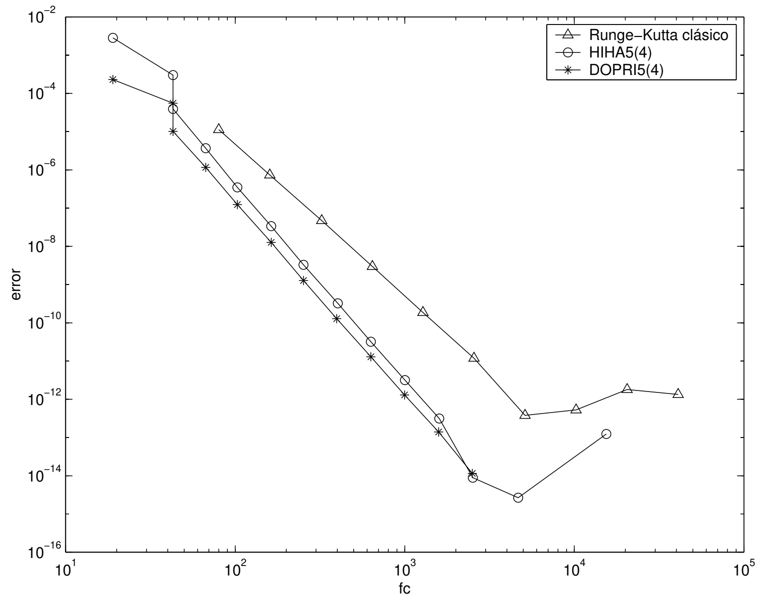


Figura 1.1: Diagrama de eficiencia para un problema lineal.

más trabajo realicemos, más a la derecha.

El diagrama muestra que, salvo para precisiones muy exigentes, próximas a 10^{-14} (donde el error de redondeo empieza a desempeñar un papel), a mayor precisión mayor trabajo. También queda claro que para este problema el método más eficiente es DOPRI5(4), si bien HIHA5(4) de Higham y Hall no le va a la zaga. El método de Runge-Kutta clásico es obviamente peor, siendo la diferencia más grande cuanto mayor sea la precisión requerida.

Los puntos parecen disponerse (salvo en casos extremos) a lo largo de rectas. Veamos una explicación cuando el método es de paso fijo. En este caso

$$\texttt{error} = \|y(x_N) - y_N\| \approx Ch^p = \frac{C(b-a)^p}{N^p}.$$

Usando que $\texttt{fc}= Ns$ (o bien $\texttt{fc}= N(s-1)$ si se tiene FSAL), obtenemos que

$$\texttt{error} = c\,\texttt{fc}^{-p},$$

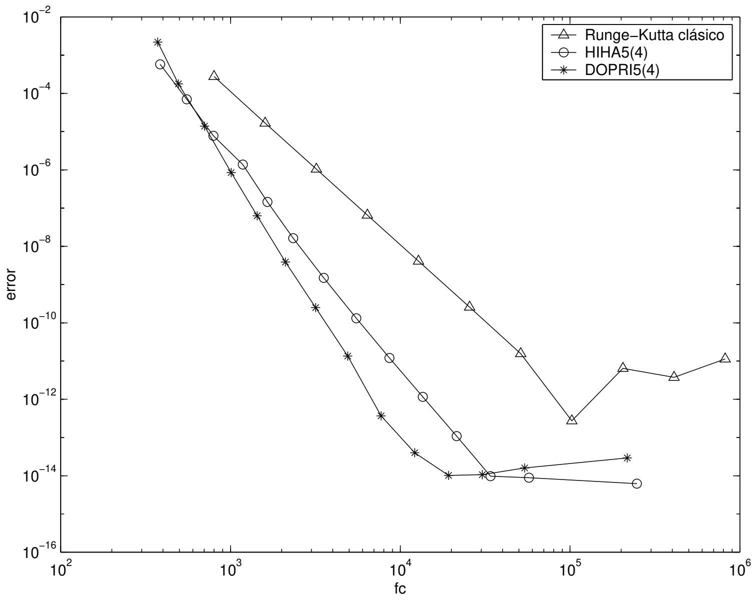


Figura 1.2: Diagrama de eficiencia para el Brusselator.

y tomando logaritmos

$$\log_{10}(\texttt{error}) = -p\log_{10}(\texttt{fc}) + \bar{c}.$$

La pendiente de la recta es $-p$. Esto produce una forma de obtener empíricamente el orden de convergencia, incluso para métodos de paso variable. Por ejemplo, en la gráfica vemos que el método de Runge-Kutta clásico tiene orden 4, mientras que DOPRI5(4) y HIHA5(4) tienen orden 5.

En la Figura 1.2 representamos `error` frente a `fc`, en escalas logarítmicas para el problema conocido como Brusselator, que modela ciertas reacciones químicas,

$$\begin{aligned} y_1' &= 1 + y_1^2 y_2 - 4y_1, \qquad && y_1(0) = 1.5,\\ y_2' &= 3y_1 - y_1^2 y_2, && y_2(0) = 3, \end{aligned}$$

con intervalo de integración $0 \leq x \leq 20$.

Nótese que el orden de convergencia empírico es 4 para Runge-Kutta, 6 para

HIHA5(4) y 7 para DOPRI5(4). Para todos los métodos, dada una precisión, necesitamos más trabajo para resolver el Brusselator que el problema lineal.

Para métodos de paso fijo hay otra forma de obtener empíricamente el orden de convergencia. Para ello integramos la ecuación en $[a,b]$ con paso h y $h/2$. Tenemos

$$\begin{aligned}\texttt{error}(h) &= \|y(x_N) - y_{h,N}\| \approx Ch^p,\\ \texttt{error}(h/2) &= \|y(x_N) - y_{\frac{h}{2},N}\| \approx Ch^p/2^p.\end{aligned}$$

Por tanto

$$\texttt{error}(h/2) \approx \frac{\texttt{error}(h)}{2^p},$$

y concluimos que

$$p \approx \frac{\log(\texttt{error}(h)) - \log(\texttt{error}(h/2))}{\log 2}.$$

De hecho el límite del segundo miembro cuando $h \to 0^+$ es p.

1.16 Comentarios y bibliografía

Se puede encontrar una prueba del Teorema de Euler en [HNW]. El primer capítulo de dicho libro presenta la teoría clásica de ecuaciones diferenciales mostrando su evolución histórica. Contiene además mucha información sobre el tema que no siempre se puede encontrar en otros libros. Por todo ello su lectura resulta muy recomendable. Otras posibles referencias para la teoría clásica son los libros que el alumno habrá manejado en la asignatura de Cálculo IV, como [BD] o [Si].

El método de Euler, predecesor de todos los demás métodos numéricos para ecuaciones diferenciales ordinarias, aparece en la mayoría de los libros de texto de introducción al cálculo numérico, por ejemplo en [A1], [A2], [JR] o [BF]. La regla del trapecio y los métodos de Taylor también se pueden encontrar en cualquiera de esos textos.

El libro [JR] contiene una prueba de convergencia de métodos de un paso parecida a la que damos nosotros, si bien restringida a métodos explícitos. Contiene también una introducción asequible a los métodos de Runge-Kutta.

Los métodos de Runge-Kutta aparecen por primera vez en un artículo de Runge en 1895 [R]. Posteriormente, en 1900, Heun [Heu] introdujo nuevos métodos. Fue Kutta en 1901 [Ku] quien formuló el esquema general de los que hoy se conocen como métodos de Runge-Kutta. Para saber más sobre la historia del análisis numérico se puede consultar [Go].

Las primeras ideas de la teoría de Butcher aparecen en un artículo de Merson en 1957; su desarrollo completo es debido a John Butcher, en una larga serie de artículos, el primero de los cuales es de 1963. En [Bu] se puede encontrar una exposición completa de la teoría, con el encanto de que está escrita por su creador. El libro contiene además una extensísima bibliografía sobre métodos numéricos para ecuaciones diferenciales ordinarias. En [L] se presenta una versión simplificada (sin pruebas) de la teoría. En [HNW] se expone el tema de forma razonablemente clara y completa.

Hay una manera alternativa de obtener las condiciones de orden debida a P. Albrecht. Usa algunas de las ideas que se emplean al estudiar métodos lineales multipaso. Se puede encontrar una descripción breve en [L].

La mejor referencia para cuestiones prácticas (estimación del error, selección de la longitud del paso, ...) es sin duda [HNW].

Existen métodos de Runge-Kutta específicos para ciertas clases de problemas, con la característica de que las soluciones numéricas retienen algunas propiedades especialmente importantes de las soluciones del problema original. Como ejemplos tenemos los métodos simplécticos para problemas hamiltonianos [SC], que preservan el hamiltoniano y los volúmenes, o los métodos que mantienen el carácter disipativo de un problema [DV].

Capítulo 2

Problemas stiff

2.1 ¿Qué es un problema stiff?

Intentemos resolver el problema lineal

$$y' = Ay, \quad y(0) = y_0, \qquad \text{con} \quad A = \begin{pmatrix} -100 & 1 \\ 0 & -\frac{1}{10} \end{pmatrix} \tag{2.1}$$

por el método de Euler. Se tiene que

$$y_{n+1} = (1 + hA)y_n,$$

y por inducción que

$$y_n = (I + hA)^n y_0, \qquad n = 0, 1, 2, \ldots.$$

Pasando a la base canónica de Jordan

$$A = VDV^{-1}, \quad \text{donde} \quad V = \begin{pmatrix} 1 & 1 \\ 0 & \frac{999}{10} \end{pmatrix} \quad \text{y} \quad D = \begin{pmatrix} -100 & 0 \\ 0 & -\frac{1}{10} \end{pmatrix},$$

de donde se obtiene que la solución exacta del problema viene dada por

$$y(x) = e^{xA} y_0 = V e^{xD} V^{-1} y_0, \quad x \geq 0, \quad \text{donde} \quad e^{xD} = \begin{pmatrix} e^{-100x} & 0 \\ 0 & e^{-\frac{x}{10}} \end{pmatrix}.$$

En otras palabras, existen dos vectores v_1 y v_2, dependientes de y_0, pero no de x, tales que

$$y(x) = e^{-100x} v_1 + e^{-x/10} v_2, \qquad x \geq 0.$$

La función e^{-100x} decae extraordinariamente más aprisa que $e^{-x/10}$. Por tanto, incluso para $x > 0$ pequeño, la contribución de v_1 es despreciable a todos los efectos, y $y(x) \approx e^{-x/10} v_2$. ¿Qué podemos decir de la solución numérica $\{y_n\}$ obtenida con el método de Euler? Tenemos que

$$y_n = V(I + hA)^n V^{-1} y_0, \qquad n = 0, 1, \dots,$$

y, dado que

$$(1 + hD)^n = \begin{pmatrix} (1 - 100h)^n & 0 \\ 0 & (1 - \frac{h}{10})^n \end{pmatrix},$$

se sigue que

$$y_n = (1 - 100h)^n v_1 + \left(1 - \frac{h}{10}\right)^n v_2, \qquad n = 0, 1, \dots.$$

Si $h > 1/50$, entonces $|1 - 100h| > 1$, y para n grande los iterados de Euler crecen en magnitud geométricamente, en contraste con el comportamiento asintótico de la verdadera solución.

Supongamos que escogemos una condición inicial idéntica a un autovector correspondiente al autovalor $-1/10$, por ejemplo

$$y_0 = \begin{pmatrix} 1 \\ \frac{999}{10} \end{pmatrix}. \tag{2.2}$$

En aritmética exacta $v_1 = 0$, $v_2 = y_0$ e $y_n = (1 - \frac{h}{10})^n y_0$. La solución numérica converge al vector cero cuando $n \to \infty$ para valores razonables de h, concretamente para $h < 20$. Así que podríamos esperar que todo fuera bien con

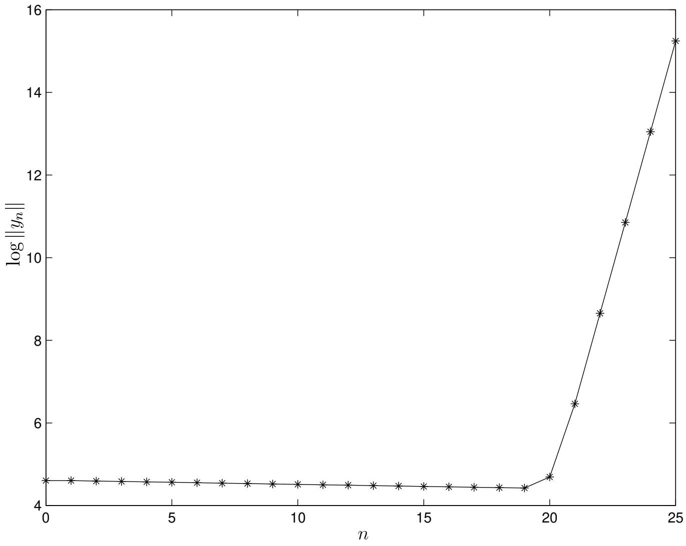


Figura 2.1: Logaritmo de la norma euclídea $\|y_n\|$ de los pasos de Euler, cuando se aplica el método al problema (2.1) con $h = 1/10$ y condición inicial igual al autovector "estable".

el método de Euler. No es así. Los ordenadores reales producen errores de redondeo y antes o después producirán una contribución distinta de cero a un autovector correspondiente al autovalor -100. En cuanto esto ocurre la componente "inestable" crece geométricamente como $(1 - 100h)^n$, y rápidamente, a menos que $h < 1/50$, eclipsa a la verdadera solución.

En la figura 2.1 se representa $\log\|y_n\|$ frente a n, para $n = 0, \ldots, 25$ con la condición inicial (2.2) y paso $h = 1/10$. El cálculo se ha realizado con Matlab. La norma de los diecinueve primeros paso decrece con la velocidad correcta, dictada por $(1 - \frac{h}{10})^n = (99/100)^n$. Sin embargo, a partir de ese momento las cosas se estropean, y después de un par de pasos la norma crece geométricamente, como $|1 - 100h|^n = 9^n$. Se puede comprobar en la figura que, en efecto, la pendiente de la curva es inicialmente $\log(99/100) \approx -0.0101$, pero pasa a ser $\log 9 \approx 2.1972$ cuando se llega al régimen inestable.

La anterior elección del dato inicial no es artificial. Si nos enfrentamos a un problema como (2.1) con una condición inicial arbitraria tendremos que emplear un tamaño de paso pequeño en el *régimen transitorio* inicial en el que la contribución del autovector "inestable" todavía es significativa. Una vez pasado este corto periodo, la solución viene esencialmente descrita por el autovector "estable", y uno se ve tentado a incrementar h. Sin embargo hay que resistirse a esta tentación pues, como hemos visto antes, aunque no lo parezca el autovector "inestable" siempre deja sentir su presencia.

Es importante comprender que este comportamiento no tiene nada que ver con el error local del método numérico; hay que disminuir el paso, no por consideraciones de precisión, sino por la inestabilidad.

No todos los métodos numérico muestran tanta sensibilidad a la estabilidad. Por ejemplo, si resolvemos (2.1) con la *regla del trapecio*

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

obtenemos

$$y_{n+1} = \left(I - \frac{h}{2}A\right)^{-1} \left(I + \frac{h}{2}A\right) y_0.$$

Dado que $(I - \frac{h}{2}A)^{-1}$ y $(I + \frac{h}{2}A)$ conmutan, el orden de multiplicación no importa, y se tiene que

$$y_n = \left(\left(I - \frac{h}{2}A\right)^{-1} \left(I + \frac{h}{2}A\right)\right)^n y_0, \qquad n = 0, 1 \ldots.$$

Usando como antes la forma canónica de Jordan de A y factorizando deducimos que

$$y_n = \left(\frac{1 - 50h}{1 + 50h}\right)^n v_1 + \left(\frac{1 - \frac{h}{20}}{1 + \frac{h}{20}}\right)^n v_2, \qquad n = 0, 1, \ldots.$$

Dado que

$$\left|\frac{1 - 50h}{1 + 50h}\right|, \ \left|\frac{1 - \frac{h}{20}}{1 + \frac{h}{20}}\right| < 1$$

para todo $h > 0$, deducimos que $\lim_{n\to\infty} y_n = 0$. Esto recupera el comportamiento asintótico correcto para la ecuación (2.1) independientemente del tamaño de h.

En otras palabras, la regla del trapecio no requiere ninguna restricción en el paso para evitar la inestabilidad. Esto no quiere decir, por supuesto, que cualquier h valga. Es necesario escoger $h > 0$ suficientemente pequeño como para que el error local sea razonablemente pequeño y estemos aproximando adecuadamente la verdadera solución. Sin embargo, no hay necesidad de disminuir h hasta un tamaño minúsculo con el fin de evitar el crecimiento incontrolado de componentes espurias de la solución.

La ecuación (2.1) es un ejemplo de problema *stiff* (rígido, en inglés). En la literatura se pueden encontrar varias definiciones de este concepto, pero quizás sea más interesante adoptar una definición operativa, si bien un poco vaga.

"Definición" 2.1. *Decimos que el sistema*

$$y' = f(x, y), \qquad x \geq x_0, \qquad y(x_0) = y_0,$$

es stiff *si su resolución numérica por algunos métodos exige una disminución significativa del paso de integración para evitar inestabilidades.*

Es evidente que ésta no es una definición matemática correcta; pero tampoco estamos intentando probar teoremas del tipo *si un sistema es stiff entonces. . . .* Lo importante de este concepto es que nos debe ayudar a elegir y diseñar métodos numéricos.

En nuestro ejemplo hemos visto uno de los mecanismos (de hecho, uno de los más frecuentes) que hacen que un problema sea *stiff*, el que haya modos con escalas y tiempos de vida muy diferentes en la solución. Es frecuente designar al cociente de los módulos de los autovalores de mayor y menor módulo de un sistema lineal (en el caso de un sistema general, los autovalores de la matriz jacobiana) como *coeficiente de rigidez* del sistema. Si dicho coeficiente es grande, el sistema suele ser *stiff*. No obstante, haremos notar que este análisis lineal puede fallar al aplicarlo a un sistema no lineal.

Una gran parte de los sistemas de ecuaciones diferenciales ordinarias que aparecen en problemas reales son *stiff*. Siempre que las ecuaciones modelen varios procesos con velocidades de evolución muy distintas, el problema será probablemente *stiff*. Esto sucede por ejemplo en problemas de cinética química, combustión, predicción meteorológica, biología y electrónica. Otra fuente muy importante de problemas *stiff* es el propio análisis numérico. Por ejemplo, las ecuaciones en derivadas parciales parabólicas se aproximan con frecuencia con sistemas de ecuaciones diferenciales ordinarias que resultan ser *stiff*.

2.2 Dominio de estabilidad lineal y A-estabilidad

Supongamos que aplicamos un método numérico con paso constante $h > 0$ a la ecuación lineal escalar

$$y' = \lambda y, \qquad x \geq 0, \qquad y(0) = 1, \tag{2.3}$$

donde $\lambda \in \mathbb{C}$. La solución exacta es $y(x) = e^{\lambda x}$, y por tanto $\lim_{x\to\infty} y(x) = 0$ si y sólo si $\operatorname{Re}\lambda < 0$.

Definición 2.2. *El* dominio de estabilidad lineal *(también llamado región de estabilidad absoluta) de un método numérico es el conjunto de todos los números $z = h\lambda$ con $h > 0$, $\lambda \in \mathbb{C}$, tales que la solución numérica de* (2.3) *verifica que* $\lim\limits_{n\to\infty} y_n = 0$.

Es decir, el dominio de estabilidad lineal es el conjunto de los $h\lambda$ para los que se recupera el comportamiento asintótico de la solución exacta cuando esta es estable. Lo denotaremos por $\mathcal{D}$.

Ejemplo. Al aplicar el método de Euler a (2.3) obtenemos la sucesión

$$y_n = (1 + h\lambda)^n, \qquad n = 0, 1, \ldots,$$

que es geométrica. Así pues, $\lim_{n\to\infty} y_n = 0$ si y sólo si $|1+h\lambda| < 1$. Concluimos que $\mathcal{D} = \{z \in \mathbb{C} : \ |1 + z| < 1\}$.

Ejemplo. Para la regla del trapecio tenemos que

$$y_n = \left(\frac{1 + \frac{1}{2}h\lambda}{1 - \frac{1}{2}h\lambda} \right)^n, \qquad n = 0, 1, \ldots,$$

que es una progresión geométrica. El dominio de estabilidad lineal es obviamente

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

Es fácil verificar entonces que $\mathcal{D} = \{z \in \mathbb{C} : \operatorname{Re} z < 0\}$.

¿Por qué estamos súbitamente interesados en una humilde ecuación escalar?

Supongamos que queremos resolver

$$y' = Ay, \qquad y(0) = y_0,$$

con A una matriz $d \times d$ arbitraria por Euler. La solución numérica es $y_n = (1 + hA)^n y_0$. Si existe una base de autovectores de A, entonces $A = VDV^{-1}$ con $D = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_d)$, y se tiene que $y_n = V(I + hD)^n V^{-1} y_0$. Por tanto existen vectores $v_1, v_2, \ldots, v_d \in \mathbb{C}^d$ tales que

$$y_n = \sum_{k=1}^{d} (1 + h\lambda_k)^n v_k, \qquad n = 0, 1, \ldots.$$

Supongamos que la solución exacta es asintóticamente estable. Esto sólo ocurre si $\operatorname{Re} \lambda_k < 0$ para todo $k = 1, 2, \ldots, d$. Para que la solución numérica de Euler imite este comportamiento se necesita que el paso $h > 0$ sea tal que

$$|1 + h\lambda_k| < 1, \qquad k = 1, \ldots, d.$$

Es decir, todos los productos $h\lambda_1, \ldots, h\lambda_d$ deben estar en el dominio de estabilidad lineal del método de Euler. Esto significa en la práctica que el tamaño del paso está determinado por aquel autovalor que sea peor desde el punto de vista de la estabilidad.

La restricción de que A tenga una base de autovectores se puede relajar, ya que la forma canónica de Jordan es suficiente.

El análisis se puede extender también fácilmente a sistemas homogéneos, $y' = Ay + a$.

La región de estabilidad lineal $\mathcal{D}$ es también importante en sistemas no lineales. Dado un sistema no lineal

$$y' = f(x, y), \qquad x \geq x_0, \qquad y(x_0) = y_0,$$

se suele pedir que en el paso n-ésimo

$$h\lambda_{n,1}, h\lambda_{n,2}, \ldots, h\lambda_{n,m} \in \mathcal{D},$$

donde $\lambda_{n,1}, \ldots, \lambda_{n,m}$ son los autovalores de la matriz jacobiana

$$J_n \equiv \frac{\partial f(x_n, y_n)}{\partial y}.$$

Esto se basa en la suposición de que localmente el comportamiento del sistema está bien modelado por la ecuación variacional

$$y' = y_n + J_n(y - y_n).$$

Hay que advertir que esto dista mucho de ser exacto.

Cuanto mayor sea la región de $\mathbb{C}^- := \{z \in \mathbb{C} : \operatorname{Re} z < 0\}$ cubierta por $\mathcal{D}$, menos restricciones para h debidas a posibles inestabilidades (recuérdense los ejemplos de Euler y la regla del trapecio). Eso motiva la siguiente definición.

Definición 2.3. *Un método es A-estable si*

$$\mathbb{C}^- \subseteq \mathcal{D}.$$

Es decir, si un método es A-estable podemos elegir la longitud de paso h (al menos para sistemas lineales) sólo por motivos de precisión, sin restricciones debidas a la inestabilidad. Estos métodos serán buenos para problemas *stiff*.

2.3 Estabilidad de métodos de Runge-Kutta

Queremos aplicar el método de Runge-Kutta (1.9) al problema escalar

$$y' = \lambda y, \qquad x \geq 0, \qquad y(0) = 1. \tag{2.4}$$

Para ello conviene reescribir (1.9) en la forma alternativa

$$\begin{cases} Y_i = y_n + h \sum\limits_{j=1}^{s} a_{ij} f(x_n + c_j h, Y_j), \\ y_{n+1} = y_n + h \sum\limits_{i=1}^{s} b_i f(x_n + c_i h, Y_i). \end{cases}$$

Obtenemos que

$$\begin{cases} Y_i = y_n + \lambda h \sum\limits_{j=1}^{s} a_{ij} Y_j, \\ y_{n+1} = y_n + \lambda h \sum\limits_{i=1}^{s} b_i Y_i. \end{cases} \tag{2.5}$$

Sean $Y, e \in \mathbb{R}^s$ dados por

$$Y = (Y_1, \ldots, Y_s)^T, \qquad e = (1, 1, \ldots, 1)^T.$$

Entonces podemos escribir (2.5) en la forma

$$\begin{cases} Y = e\, y_n + \lambda h A Y, \\ y_{n+1} = y_n + \lambda h b^T Y. \end{cases}$$

La primera de estas ecuaciones da

$$Y = (I - \lambda h A)^{-1} e y_n,$$

que introducido en la segunda produce

$$y_{n+1} = (1 + \lambda h b^T (I - \lambda h A)^{-1} e) y_n, \qquad n = 0, 1, \ldots,$$

es decir,

$$y_{n+1} = R(\lambda h) y_n,$$

donde

$$R(z) = 1 + zb^T(I - zA)^{-1}e, \qquad z \in \mathbb{C}. \tag{2.6}$$

La función $R(z)$ recibe el nombre de *función de estabilidad del método*, o también *función de amplificación*.

Obsérvese que $y_n = (R(\lambda h))^n$, así que de la definición de $\mathcal{D}$ se sigue inmediatamente que

$$\mathcal{D} = \{z \in \mathbb{C} : \ |R(z)| < 1\}.$$

Por tanto un método será A-estable si y sólo si $|R(z)| < 1$ en $\mathbb{C}^-$.

Notación. $\mathbb{P}_\alpha$ es el conjunto de todos los polinomios de grado menor o igual que α. $\mathbb{P}_{\alpha/\beta}$ es el conjunto de todas las funciones racionales $\hat{p}/\hat{q}$, con $\hat{p} \in \mathbb{P}_\alpha$ y $\hat{q} \in \mathbb{P}_\beta$.

Un sencillo análisis permite probar el siguiente lema.

Lema 2.4. *Para todo método de Runge-Kutta existe $R \in \mathbb{P}_{s/s}$ tal que la solución por dicho método del problema* (2.4) *viene dada por*

$$y_n = (R(\lambda h))^n.$$

Es más, si el método de Runge-Kutta es explícito, entonces $R \in \mathbb{P}_s$.

Prueba. Tenemos que probar que la función R dada por (2.6) pertenece a $\mathbb{P}_{s/s}$. Usamos que

$$(I - zA)^{-1} = \frac{(\mathrm{adj}\,(I - zA))^T}{\det(I - zA)}.$$

Cada componente de $I - zA$ es lineal en z. Por tanto, como cada elemento de $\mathrm{adj}\,(I - zA)$ es el determinante de una matriz $(s-1) \times (s-1)$, está en $\mathbb{P}_{s-1}$. Así pues,

$$b^T(\mathrm{adj}\,(I - zA))^T e \in \mathbb{P}_{s-1}.$$

Como $\det(I - zA) \in \mathbb{P}_s$, concluimos que $R \in \mathbb{P}_{s/s}$.

Si el método es explícito, entonces A es estrictamente triangular inferior y $I - zA$ es una matriz triangular inferior con unos a lo largo de la diagonal. Por tanto $\det(I - zA) = 1$, y R es un polinomio. $\square$

Corolario 2.5. *Ningún método de Runge-Kutta explícito puede ser A-estable.*

Prueba. La función de estabilidad de un método de Runge-Kutta explícito es un polinomio. Además $R(0) = 1$. Por otra parte, ningún polinomio, salvo la función constante $R(z) = c \in (-1, 1)$ está estrictamente acotado por uno en $\mathbb{C}^-$. $\square$

En el caso de métodos explícitos podemos decir aún más sobre la función de amplificación.

Teorema 2.6. *Si un método de Runge-Kutta explícito tiene orden de consistencia p, entonces*

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

Prueba. La solución exacta del (2.4) es $y(x) = e^{\lambda x}$. Por ser el método consistente de orden p tenemos que

$$y(h) - y_1 = O(h^{p+1}),$$

así que

$$y_1 = e^z + O(h^{p+1}) = 1 + z + \frac{z^2}{2!} + \cdots + \frac{z^p}{p!} + O(z^{p+1}),$$

donde $z = \lambda h$. Usando que $y_1 = R(z)y_0$ concluimos inmediatamente el resultado. $\square$

Como consecuencia tenemos que todos los métodos de Runge-Kutta explícitos con $p = s$ tienen como función de estabilidad

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

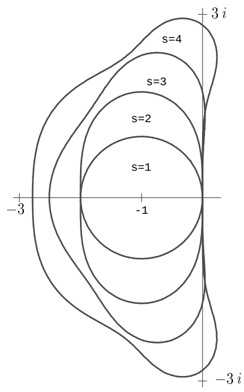


Figura 2.2: Dominio de estabilidad lineal de los métodos de Runge-Kutta explícitos de orden $p = s$.

En la figura 2.2 mostramos los correspondientes dominios de estabilidad lineal.

¿Qué podemos decir de los métodos implícitos? Veamos con un ejemplo que pueden ser A-estables.

Ejemplo. Consideramos el método

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

Su función de amplificación es

$$R(z) = \frac{1 + \frac{1}{3}z}{1 - \frac{2}{3}z + \frac{1}{6}z^2}.$$

Veamos si es A-estable. Representamos $z \in \mathbb{C}^-$ en coordenadas polares, $z = \rho e^{i\theta}$, con $\rho > 0$, $|\theta - \pi| < \pi/2$, y nos preguntamos si $|R(\rho e^{i\theta})| < 1$. Esto es equivalente a

$$\left|1 + \frac{1}{3}\rho e^{i\theta}\right|^2 < \left|1 - \frac{2}{3}\rho e^{i\theta} + \frac{1}{6}\rho^2 e^{2i\theta}\right|^2,$$

y por tanto a

$$1+\frac{2}{3}\rho\cos\theta+\frac{1}{9}\rho^2<1-\frac{4}{3}\rho\cos\theta+\rho^2\left(\frac{1}{3}\cos 2\theta+\frac{4}{9}\right)-\frac{2}{9}\rho^3\cos\theta+\frac{1}{36}\rho^4.$$

Reordenando términos, la condición $\rho e^{i\theta}\in\mathcal{D}$ pasa a ser

$$2\rho\left(1+\frac{1}{9}\rho^2\right)\cos\theta<\frac{1}{3}\rho^2\left(1+\cos 2\theta\right)+\frac{1}{36}\rho^4=\frac{2}{3}\rho^2\cos^2\theta+\frac{1}{36}\rho^4,$$

que se satisface para todos los $z\in\mathbb{C}^-$, puesto que $\cos\theta<0$ para todo z de este tipo. El método es por tanto A-estable.

Esto explica por qué son interesantes los métodos implícitos. Son más caros, pero, además de permitir romper las barreras de Butcher, algunos de ellos tienen regiones de estabilidad grandes, y pueden ser convenientes a la hora de tratar problemas *stiff*.

Sin embargo, no todos los métodos implícitos son A-estables.

Ejemplo. Consideramos el método

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

Su función de amplificación es

$$R(z)=\frac{1+\frac{2}{3}z+\frac{1}{6}z^2}{1-\frac{1}{3}z}.$$

Así pues, $\lim\limits_{|z|\to\infty}|R(z)|=\infty$, por lo que el método no es A-estable.

En realidad no hace falta comprobar todos los $z\in\mathbb{C}^-$ para ver si una función racional dada tiene su origen en un método A-estable (una $R(z)$ tal se dice A-aceptable). Ése es el contenido del siguiente lema, cuya prueba es una aplicación del principio del máximo.

Lema 2.7. *Sea $R(z)$ una función racional no constante. Entonces $|R(z)| < 1$ para todo $z \in \mathbb{C}^-$ si y sólo si todos los polos de R tienen parte real positiva y $|R(it)| \leq 1$ para todo $t \in \mathbb{R}$.*

Prueba. Si $|R(z)| < 1$ para todo $z \in \mathbb{C}^-$, por continuidad $|R(z)| \leq 1$ para todo $z \in \overline{\mathbb{C}^-}$. En particular R no puede tener polos en el semiplano cerrado izquierdo, y $|R(it)| \leq 1$ para todo $t \in \mathbb{R}$.

Para probar el recíproco observamos que si los polos están a la derecha del eje imaginario, entonces la función racional R es analítica en el cerrado $\overline{\mathbb{C}^-}$. Por tanto, al no ser R constante, $|R|$ no puede tener un máximo en el interior, y alcanza su máximo en la frontera. Parte de la frontera está en el infinito. Sin embargo, por ser R una función racional existe el límite (finito o infinito) $\lim_{|z|\to\infty} |R(z)|$, y es igual a $\lim_{t\to\infty} |R(it)|$. Por tanto, si $|R(it)| \leq 1$, $t \in \mathbb{R}$, entonces $|R(z)| < 1$ para todo $z \in \mathbb{C}^-$. $\square$

Ejemplo. Consideramos el método de Gauss-Legendre de dos etapas (orden 4),

$$\begin{array}{c|cc}
\frac{1}{2} - \frac{\sqrt{3}}{6} & \frac{1}{4} & \frac{1}{4} - \frac{\sqrt{3}}{6} \\
\frac{1}{2} + \frac{\sqrt{3}}{6} & \frac{1}{4} + \frac{\sqrt{3}}{6} & \frac{1}{4} \\
\hline
 & \frac{1}{2} & \frac{1}{2}.
\end{array}$$

Un sencillo cálculo muestra que

$$R(z) = \frac{1 + \frac{1}{2}z + \frac{1}{12}z^2}{1 - \frac{1}{2}z + \frac{1}{12}z^2}.$$

Los polos de esta función, $3 \pm i\sqrt{3}$, están en el semiplano abierto derecho, y $|R(it)| = 1$, $t \in \mathbb{R}$, por lo que el método es A-estable.

2.4 Comentarios y bibliografía

El capítulo dedicado en [L] a los problemas *stiff* y a la estabilidad lineal nos parece excelente. Más simple, pero aún más claro, nos parece [I]. La referencia más completa sobre este tema es [HW].

Los métodos de Runge-Kutta implícitos, de los que no hemos dicho nada, pueden ser interesantes a la hora de tratar problemas *stiff*. Se puede encontrar una breve introducción en las secciones 3.3 y 3.4 de [I], donde además se explica su conexión con los llamados métodos de colocación. Si se quiere saber más, habrá que consultar [HNW] y [HW].

Si se usan métodos implícitos (de Runge-Kutta o lineales multipaso), será necesario resolver sistemas no lineales de ecuaciones algebraicas. Se puede encontrar un tratamiento exhaustivo de los métodos con que se resuelven estos sistemas en el clásico [OR], así como en [F] y el más moderno [K].

La teoría de estabilidad no lineal es relativamente reciente. Se puede encontrar una introducción en [L] y un estudio más riguroso y completo en [DV].

Capítulo 3

Diferencias finitas

3.1 La fórmula de cinco puntos

En este capítulo y el siguiente mostraremos algunos métodos para resolver numéricamente ecuaciones en derivadas parciales. Consideraremos tan sólo problemas elípticos de segundo orden. No obstante, algunas de las ideas se pueden utilizar provechosamente para resolver otros problemas.

Nuestra primera familia de métodos usa un enfoque que ya hemos empleado con éxito para resolver ecuaciones diferenciales ordinarias: sustituir las derivadas por cocientes incrementales. Esto dará lugar a los *métodos de diferencias finitas.*

Sea $\Omega \subset \mathbb{R}^2$ un abierto acotado del plano, tal que todo punto de la frontera sea accesible desde el exterior por una curva continua simple. Consideramos la ecuación de Poisson con condiciones de frontera de tipo Dirichlet

$$-\Delta u = f \qquad \text{en } \Omega, \qquad u = g \qquad \text{en } \partial\Omega. \tag{3.1}$$

Supondremos que g es continua en $\overline{\Omega}$ y que existen números $C > 0$ y $0 < \alpha < 1$

tales que $|f(x) - f(y)| \leq C|x - y|^\alpha$ para todo $x, y \in \overline{\Omega}$ (diremos que f es α-Hölder continua en $\overline{\Omega}$, y escribiremos que $f \in C^{0,\alpha}(\overline{\Omega})$). Se tiene el siguiente resultado de existencia y unicidad.

Teorema 3.1. *Si $f \in C^{0,\alpha}(\overline{\Omega})$ y $g \in C(\overline{\Omega})$, el problema* (3.1) *tiene una única solución $u \in C^2(\Omega) \bigcap C(\overline{\Omega})$.*

Una herramienta fundamental en el análisis de esta ecuación es el principio del máximo.

Teorema 3.2 (Principio del máximo). *Sea $u \in C^2(\Omega) \bigcap C(\overline{\Omega})$.*

(i) Si $-\Delta u \leq 0$ para todo $(x, y) \in \Omega$, entonces $\max\limits_{\overline{\Omega}} u \leq \max\limits_{\partial\Omega} u$.

(ii) Si $-\Delta u \geq 0$ para todo $(x, y) \in \Omega$, entonces $\min\limits_{\overline{\Omega}} u \geq \min\limits_{\partial\Omega} u$.

Supongamos que el dominio sobre el que se formula la ecuación (3.1) es el rectángulo

$$\Omega = [a_1, a_2] \times [b_1, b_2].$$

Se puede ajustar una red uniforme formada por los puntos de coordenadas

$$(x_i, y_j) = (a_1 + ih, b_1 + jk)$$

para $0 \leq i \leq N$ y $0 \leq j \leq M$, donde los tamaños de paso h y k están determinados por

$$h = \frac{a_2 - a_1}{N}, \qquad k = \frac{b_2 - b_1}{M}.$$

Sobre los nodos de la red definimos una función U cuyo valor $U(x_i, y_j) = U_{ij}$ es una aproximación de la solución u en (x_i, y_j),

$$U_{ij} \approx u(x_i, y_j).$$

Nuestro objetivo es encontrar una forma de generar las aproximaciones U_{ij}. Para ello utilizaremos fórmulas de *diferencias finitas* que aproximen las derivadas a partir de valores de la función.

Nota. El término *diferencias finitas* es clásico, y se remonta a los tiempos en que, siguiendo a Leibniz (1646–1716), se concebía la diferencial df de una función f como diferencia "infinitamente pequeña" $f(x+dx)-f(x)$ entre los valores de f en dos puntos "infinitamente" próximos x y $x+dx$. Entonces se contraponía a la diferencial infinitesimal la diferencia finita $f(x+h)-f(x)$.

Sea $f \in C^4([x-h, x+h])$. Queremos obtener una expresión aproximada para la derivada segunda combinando valores de f en distintos puntos. Desarrollando por Taylor en torno a x se tiene que

$$f(x \pm h) = f(x) \pm hf'(x) + \frac{h^2}{2}f''(x) \pm \frac{h^3}{3!}f'''(x) + \frac{h^4}{4!}f^{(4)}(\xi_\pm),$$

donde ξ_+ y ξ_- son puntos intermedios. De esta forma

$$f''(x) = \frac{f(x+h) - 2f(x) + f(x-h)}{h^2} - \frac{h^2}{4!}(f^{(4)}(\xi_+) + f^{(4)}(\xi_-)).$$

Esto permite aproximar la derivada segunda por la *fórmula de diferencias centrales*

$$f''(x) \approx \frac{f(x+h) - 2f(x) + f(x-h)}{h^2}.$$

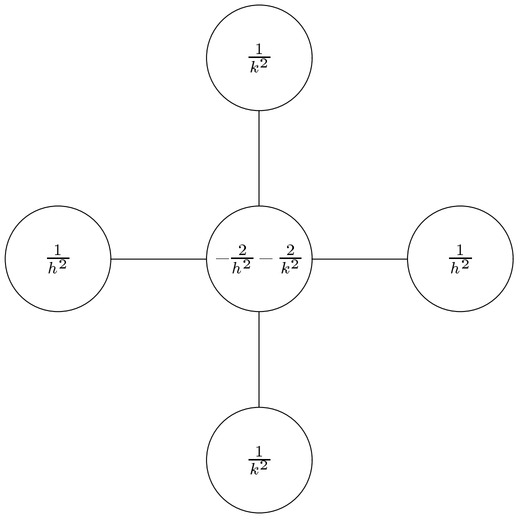
Si $u \in C^4(\Omega)$, podemos aplicar la fórmula de diferencias centrales a $u(x,y)$ para aproximar $u_{xx}(x,y)$ y $u_{yy}(x,y)$, obteniendo

$$\Delta u(x,y) = \Delta_{hk}u(x,y) + O(k^2 + h^2),$$

donde Δ_{hk} es el *operador Laplaciano discreto*,

$$\begin{aligned}\Delta_{hk}u(x,y) = \; & \frac{u(x+h,y) - 2u(x,y) + u(x-h,y)}{h^2} \\ & + \frac{u(x,y+k) - 2u(x,y) + u(x,y-k)}{k^2},\end{aligned}$$

también conocido como *fórmula de cinco puntos*. Se representa gráficamente mediante la *molécula computacional*



Si u es la solución de (3.1), en cada uno de los nodos interiores tenemos

$$f(x_i, y_j) \approx -\Delta_{hk} u(x_i, y_j) \approx \Delta_{hk} U_{ij}.$$

Definiendo $f_{ij} = f(x_i, y_j)$ y $g_{ij} = g(x_i, y_j)$, lo que haremos será pedir que U satisfaga el problema de Dirichlet discreto para la ecuación de Poisson

$$\begin{cases} -\Delta_{hk} U_{ij} = f_{ij} & \text{en los nodos interiores,} \\ U_{ij} = g_{ij} & \text{en los nodos de la frontera.} \end{cases} \tag{3.2}$$

Por tanto, para un nodo interior (x_i, y_j), $0 < i < N$, $0 < j < M$, tenemos

$$\frac{-U_{i+1,j} + 2U_{ij} - U_{i-1,j}}{h^2} + \frac{-U_{i,j+1} + 2U_{i,j} - U_{i,j-1}}{k^2} = f_{ij},$$

relación que no sólo involucra a U_{ij}, sino también a los valores en los nodos adyacentes, a la izquierda y a la derecha, arriba y abajo. Así, para calcular la solución aproximada es preciso resolver un sistema lineal de $(N-1) \times (M-1)$ ecuaciones cuyas incógnitas son todos los elementos de la matriz U_{ij}, excepto los que corresponden a los nodos de frontera, que ya se conocen.

El Laplaciano discreto reproduce algunas de las propiedades más importantes del operador Laplaciano, por ejemplo el principio del máximo discreto.

Teorema 3.3 (Principio del máximo discreto).

(i) Si $-\Delta_{hk}U_{ij} \leq 0$ *en todos los nodos interiores,* $\max\limits_{i,j} U_{ij} = \max\limits_{(x_i,y_j)\in\partial\Omega} U_{ij}$.

(ii) Si $-\Delta_{hk}U_{ij} \geq 0$ *en todos los nodos interiores,* $\min\limits_{i,j} U_{ij} = \min\limits_{(x_i,y_j)\in\partial\Omega} U_{ij}$.

Prueba. (i) Si $-\Delta_{hk}U_{ij} \leq 0$, entonces

$$\left(\frac{2}{h^2} + \frac{2}{k^2}\right) U_{ij} \leq \frac{1}{h^2}U_{i+1,j} + \frac{1}{h^2}U_{i-1,j} + \frac{1}{k^2}U_{i,j+1} + \frac{1}{k^2}U_{i,j-1}.$$

Se puede escribir esta desigualdad como

$$U_{ij} \leq \alpha_1 U_{i+1,j} + \alpha_2 U_{i-1,j} + \alpha_3 U_{i,j+1} + \alpha_4 U_{i,j-1},$$

donde $0 < \alpha_i < 1$ para $i = 1, \ldots, 4$ y $\sum_{i=1}^{4} \alpha_i = 1$. Es decir, el valor de U en un nodo interior es menor que una media aritmética ponderada de los valores en los cuatro nodos adyacentes. Por consiguiente el valor de U en el nodo interior no puede ser mayor que en los adyacentes y por tanto no hay máximos interiores.

(ii) Observamos que $-\Delta_{hk}(-U_{ij}) \leq 0$, y aplicamos el apartado (i) a $-U$. □

- Para calcular las aproximaciones U_{ij} tenemos que resolver el problema (3.2). ¿Tiene dicho problema solución única?

La respuesta es afirmativa: si hubiera dos soluciones, al aplicar el Teorema 3.3 a la diferencia de ambas se concluye que ésta es trivial.

- ¿Se parece la solución numérica a la teórica cuando $h, k \to 0$?

Siguiendo lo que hacíamos para ecuaciones diferenciales ordinarias, diremos que el método *converge* si

$$\lim_{h,k\to 0^+} \max_{i,j} |u(x_i, y_j) - U_{ij}| = 0.$$

Sea $R_{hk}u(x, y)$ el *residuo* o *error de truncación*, es decir, lo que le falta a la solución teórica para satisfacer la ecuación del método,

$$R_{hk}u(x, y) = -\Delta_{hk}u(x, y) - f(x, y).$$

El error $E = u - U$ es solución de

$$\begin{cases} -\Delta_{hk} E_{ij} = R_{hk} u(x_i, y_j) & \text{en los nodos interiores,} \\ E_{ij} = 0 & \text{en los nodos de frontera,} \end{cases}$$

que es el problema (3.2) con $f_{ij} = R_{hk} u(x_i, y_j)$ y $g_{ij} = 0$, un problema con dato Dirichlet homogéneo.

Lema 3.4. *La solución U_{ij} del problema (3.2) con $g_{ij} = 0$ satisface*

$$|U_{ij}| \leq \frac{1}{16} |f|_\infty (a^2 + b^2), \qquad a = a_2 - a_1, \ b = b_2 - b_1, \tag{3.3}$$

donde $|f|_\infty = \max\limits_{i,j} |f_{ij}|$.

Prueba. Sean (c_1, c_2) las coordenadas del centro del rectángulo $[a_1, b_1] \times [a_2, b_2]$. Consideramos la función auxiliar

$$w(x, y) = (x - c_1)^2 + (y - c_2)^2,$$

y denotamos $w_{ij} = w(x_i, y_j)$. La función w cumple que

$$0 \leq w_{ij} \leq \frac{1}{4}(a^2 + b^2).$$

Por construcción, el Laplaciano discreto y el Laplaciano coinciden en la clase de los polinomios de grado menor o igual que 2. Por consiguiente,

$$\Delta_{hk} w_{ij} = \Delta w(x_i, y_j) = 4,$$

y por tanto

$$\begin{aligned} -\Delta_{hk}(U_{ij} + \tfrac{1}{4}|f|_\infty w_{ij}) &= f_{ij} - |f|_\infty \leq 0, \\ -\Delta_{hk}(U_{ij} - \tfrac{1}{4}|f|_\infty w_{ij}) &= f_{ij} + |f|_\infty \geq 0, \end{aligned}$$

Usando los principios del máximo y del mínimo discreto se obtiene

$$\begin{aligned} U_{ij} + \tfrac{1}{4}|f|_\infty &\leq \max_{(x_i,y_j)\in\partial\Omega} \left\{ U_{ij} + \tfrac{1}{4}|f|_\infty w_{ij} \right\} = \tfrac{1}{4}|f|_\infty \max_{(x_i,y_j)\in\partial\Omega} w_{ij}, \\ U_{ij} - \tfrac{1}{4}|f|_\infty &\geq \max_{(x_i,y_j)\in\partial\Omega} \left\{ U_{ij} - \tfrac{1}{4}|f|_\infty w_{ij} \right\} = -\tfrac{1}{4}|f|_\infty \min_{(x_i,y_j)\in\partial\Omega} w_{ij}. \end{aligned}$$

De estas relaciones y de la acotación de w (nótese que $|x_i - c_1| \le a/2$ y $|y_i - c_2| \le b/2$) se deduce que

$$U_{ij} \le \tfrac{1}{4}|f|_\infty \max_{(x_i,y_j)\in\partial\Omega} w_{ij} \le \tfrac{1}{16}|f|_\infty(a^2+b^2),$$
$$U_{ij} \ge -\tfrac{1}{4}|f|_\infty \min_{(x_i,y_j)\in\partial\Omega} w_{ij} \ge -\tfrac{1}{16}|f|_\infty(a^2+b^2),$$

de donde se concluye (3.3). □

Aplicando el lema al problema satisfecho por el error, obtenemos que

$$|E| \le \frac{a^2+b^2}{16}|R_{hk}u|_\infty.$$

Ahora bien,

$$|R_{hk}u(x,y)| = |-\Delta_{hk}u(x,y) + \Delta u(x,y)| = O(h^2+k^2).$$

Así pues,

$$|R_{hk}u|_\infty = O(h^2+k^2),$$

y concluimos que

$$|E| = O(h^2+k^2).$$

Hemos probado de esta forma el siguiente resultado de convergencia.

Teorema 3.5. *La solución del problema de Dirichlet discreto para la ecuación de Poisson* (3.2) *converge a la solución del correspondiente problema continuo* (3.1).

- ¿Hay formas eficientes de resolver el sistema a que da lugar el método numérico?

Lo primero que hay que hacer es numerar adecuadamente los valores U_{ij} como incógnitas del sistema. Es habitual utilizar el siguiente orden lexicográfico,

$$(i_1,j_1) < (i_2,j_2) \Leftrightarrow \begin{cases} j_1 < j_2, \\ \quad \text{ó} \\ j_1 = j_2 \text{ e } i_1 < i_2. \end{cases}$$

La dimensión del sistema es $(N-1)(M-1) \times (N-1)(M-1)$, y es en general alta. No obstante, es fácil ver que el sistema se puede expresar en forma tridiagonal por bloques. Esto se puede aprovechar por un lado para reducir el almacenamiento de datos de la matriz de coeficientes, y por otro para simplificar el cálculo de la descomposición LU de la descomposición gaussiana (ver [M], capítulo 7). Sin embargo, para (3.2) resulta más sencillo resolver el sistema lineal mediante técnicas iterativas.

El método de diferencias finitas que acabamos de explicar se puede aplicar a otras ecuaciones, no necesariamente elípticas. Así mismo, se pueden usar distintas aproximaciones de diferencias finitas para las derivadas que aparezcan en la ecuación. De esta forma se obtienen, para un mismo problema, métodos diversos.

¿Qué sucede si el dominio no es rectangular? La dificultad se presenta al intentar incorporar las condiciones de contorno al sistema lineal. Esto se puede conseguir haciendo modificaciones del esquema, (ver [M]); sin embargo, los esquemas resultantes ya no son tan fáciles de programar, ni son tan eficientes. Es preferible utilizar una idea completamente diferente: los elementos finitos.

3.2 Comentarios y bibliografía

El lector que quiera saber más sobre la teoría analítica de la ecuación de Poisson puede consultar [P], o el libro clásico [GT]. Otra buena referencia es [H], que tiene la ventaja de combinar aspectos analíticos y numéricos.

En cuanto al método de diferencias finitas, aunque sólo lo hemos aplicado a la ecuación de Poisson, es útil para tratar otros muchos problemas, por ejemplo problemas de evolución (incluyendo problemas de primer orden). Se pueden encontrar introducciones sencillas en [I] y [M]. El segundo de estos libros tiene otras dos ventajas para el alumno: está en castellano y contiene bastantes problemas resueltos. Como referencias más completas y avanzadas tenemos los libros ya clásicos de [RM] y [MG], y los un poco más modernos [St] y [T].

El estudio detallado de los diferentes métodos de resolución de los sistemas lineales dispersos a que da lugar el método de diferencias finitas queda fuera del alcance del curso. En la sección 4.9 damos abundantes referencias sobre este tema.

Capítulo 4

Elementos finitos

4.1 Introducción

Presentaremos el método de los elementos finitos estudiando un problema concreto,

$$\begin{cases} -\Delta u + u = f & \text{en } \Omega \subset \mathbb{R}^d, \\ u = 0 & \text{en } \partial\Omega. \end{cases}$$

Para simplificar todavía más, nos centraremos inicialmente en dimensión $d = 1$. Tenemos así el problema de contorno

$$-u'' + u = f \quad \text{si } a < x < b, \qquad u(a) = u(b) = 0. \tag{4.1}$$

Nota. El problema con condiciones no homogéneas $u(a) = \alpha$, $u(b) = \beta$ se reduce a éste mediante el cambio de incógnita $\tilde{u} = u - u_0$, siendo u_0 cualquier función regular tal que $u_0(a) = \alpha$, $u_0(b) = \beta$ (tomar, por ejemplo, u_0 afín). La función $\tilde{u}$ es solución del problema homogéneo, eso sí, con una nueva f.

Definición 4.1. *Sea $f \in C([a, b])$. Una solución clásica del problema* (4.1) *es una función $u \in C^2((a, b)) \cap C([a, b])$ que verifica* (4.1) *para todo $x \in [a, b]$.*

Gracias a los resultados conocidos para ecuaciones diferenciales ordinarias tenemos el siguiente teorema de existencia y unicidad.

Teorema 4.2. *Dado $f \in C([a,b])$, existe una única solución clásica de* (4.1).

¿Cómo calcular la solución numéricamente? En este caso, como la geometría del dominio es sencilla, podemos utilizar un método de diferencias finitas. Por ejemplo, si consideramos el conjunto de puntos

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

y aproximamos la derivada segunda usando la fórmula de diferencias centradas, los valores u_n de las aproximaciones de la solución en los nodos x_n se obtienen resolviendo el sistema

$$-\frac{u_{n+1} - 2u_n + u_{n-1}}{h^2} + u_n = f_n \quad \text{si } 0 < n < N, \qquad u_0 = u_N = 0.$$

Los valores u_n se pueden interpolar para obtener una aproximación de $u(x)$ en puntos x que no pertenezcan al conjunto $\{x_n\}$. Si, por ejemplo, usamos interpolación lineal, entonces obtenemos como aproximación

$$u_h(x) = \sum_{n=1}^{N-1} u_n \varphi_n(x),$$

con

$$\varphi_n(x) = \begin{cases} \dfrac{x - x_{n-1}}{x_n - x_{n-1}}, & x \in [x_{n-1}, x_n], \\[2ex] \dfrac{x_{n+1} - x}{x_{n+1} - x_n}, & x \in [x_n, x_{n+1}], \\[2ex] 0, & \text{en el resto.} \end{cases} \tag{4.2}$$

Veamos lo que hemos conseguido desde una perspectiva distinta. Hemos obtenido una aproximación u_h a la verdadera solución dentro del espacio de dimensión finita

$$V_h = \{v \in C([a,b]) : v(a) = v(b) = 0,\ v\big|_{[x_{n-1},x_n]} \in \mathbb{P}_1([x_{n-1}, x_n]),\ 1 \le n \le N\}.$$

Éste será precisamente el punto de partida para el método de elementos finitos:

• Buscar una aproximación u_h a la solución dentro de un espacio de dimensión finita.

¿Cómo hacerlo? Elegimos M funciones linealmente independientes que satisfagan las condiciones de frontera, $\varphi_l(a) = \varphi_l(b) = 0$, $l = 1, \ldots, M$ y consideramos el espacio vectorial de dimensión finita generado por ellas,

$$V_h := \mathcal{L}(\{\varphi_1, \ldots, \varphi_M\}).$$

Buscamos una aproximación u_h a u dentro de V_h,

$$u_h(x) = \sum_{l=1}^{M} \gamma_l \varphi_l(x), \qquad a \leq x \leq b, \tag{4.3}$$

donde $\gamma_1, \ldots, \gamma_M$ son constantes reales. Nótese que, dado que todos los elementos de V_h satisfacen las condiciones de frontera, u_h también lo hace.

Elegiremos los números $\gamma_1, \ldots, \gamma_M$ de manera que $u_h \in V_h$ aproxime a u "lo mejor posible". ¿Qué queremos decir con esto?

Para cualquier elección de $\gamma_1, \ldots, \gamma_M$ consideramos el *defecto* o *residuo*

$$d_h(x) := -u_h''(x) + u_h(x) - f(x), \qquad a < x < b.$$

Nota. Para que el defecto esté bien definido necesitamos que $u_h \in C^2((a,b))$. Por tanto pedimos en principio que $\varphi_l \in C^2((a,b)) \cap C([a,b])$, aunque más adelante rebajaremos esta hipótesis.

Si u_h fuera la solución de (4.1), el defecto sería idénticamente cero. Por tanto, cuanto más próximo esté d_m a la función cero, mejor esperamos que sea nuestro candidato a solución.

Dado un producto escalar en $C([a,b])$ (espacio que incluye al de las soluciones clásicas) se puede escribir $C([a,b]) = V_h \oplus V_h^{\perp}$, lo que se traduce en la descomposición $u = u_h + u_h^\perp$, y la fórmula para el defecto es simplemente $d_h = (u_h^\perp)'' - u_h^\perp$. Como esperamos que V_h sea tan grande que contenga todas las funciones de interés, es natural imponer, para minimizar errores que la proyección de d_h sobre V_h sea nula, esto es, $d_h \in V_h^\perp$, o lo que es lo mismo

$$\langle d_h, \varphi_k \rangle = 0, \qquad k = 1, \dots, M. \tag{4.4}$$

Estas condiciones se llaman *ecuaciones de Galerkin.* También podemos reinterpretar lo anterior diciendo que consideramos que $V_h^\perp$ es el espacio de funciones despreciables, y lo que imponemos es que el "operador de error" $u_h^\perp \mapsto d_h = (u_h^\perp)'' - u_h^\perp$ aplique $V_h^\perp$ en $V_h^\perp$ para que una aproximación buena no dé lugar a un defecto grande.

Nuestro segundo principio es por tanto:

- Elegir la aproximación de forma que el defecto sea ortogonal al espacio V_h.

En nuestro caso podemos tomar como producto escalar en $C([a,b])$ el producto escalar de $L^2((a,b))$,

$$\langle v, w \rangle = \int_a^b vw.$$

Por otra parte, usando la representación (4.3) tenemos que

$$d_h = -\sum_{l=1}^M \gamma_l \varphi_l'' + \sum_{l=1}^M \gamma_l \varphi_l - f,$$

que sustituido en (4.4) produce

$$\sum_{l=1}^M \gamma_l \int_a^b (-\varphi_l'' \varphi_k + \varphi_l \varphi_k) = \int_a^b f \varphi_k, \tag{4.5}$$

para $k = 1, \dots, M$. Es decir, tenemos el sistema lineal

$$\sum_{l=1}^M a_{kl} \gamma_l = \int_a^b f \varphi_k, \qquad k = 1, \dots, M,$$

donde

$$a_{kl} := \int_a^b (-\varphi_l''\varphi_k + \varphi_l\varphi_k), \qquad k, l = 1, \ldots, M.$$

La resolución del sistema nos produce los γ_l y por tanto la aproximación numérica u_h. Éste es el llamado *método de Galerkin.*

Para la puesta en práctica del método hay que evaluar M^2 integrales, bien explícitamente, o bien por medio de fórmulas de cuadratura numérica, así como resolver un sistema lineal de dimensión M. Si M es un poco grande la tarea computacional, así como las necesidades de almacenamiento en memoria, pueden ser excesivas. Ahí es donde entra el principio que distingue al método de elementos finitos de otros métodos de Galerkin:

• Elegir las funciones φ_k de manera que $\operatorname{sop}\varphi_k \cap \operatorname{sop}\varphi_l = \emptyset$ para la mayor parte de las elecciones de $k, l = 1, \ldots, M$.

De esta forma, habremos reducido ampliamente tanto el número de integrales a calcular, como el número de elementos no nulos de la matriz del sistema. Ésta será *dispersa*, con el consiguiente ahorro de memoria y de número de operaciones para resolver el sistema.

Una elección excelente en este sentido sería tomar

$$V_h = \mathcal{L}(\{\varphi_1, \ldots, \varphi_{N-1}\}),$$

con φ_n como en (4.2). De esta forma, $\operatorname{sop}\varphi_k \cap \operatorname{sop}\varphi_l = \emptyset$ si $|k - l| > 1$. Pero hay un problema: estas funciones básicas φ_n no pertenecen a $C^2((a, b))$, y de hecho, ni siquiera a $C^1((a, b))$. Aun así, desearíamos poder trabajar con este espacio, pues la matriz de coeficientes sería tridiagonal.

¿Qué hacer? Podemos integrar por partes las ecuaciones de Galerkin (4.5), usando que las funciones φ_l se anulan en los extremos del intervalo, para obtener

$$\sum_{l=1}^{M} \gamma_l \int_a^b (\varphi_l'\varphi_k' + \varphi_l\varphi_k) = \int_a^b f\varphi_k, \tag{4.6}$$

para $k = 1, \ldots, M$. Es decir, tenemos el sistema lineal

$$\sum_{l=1}^{M} \hat{a}_{kl} \gamma_l = \int_a^b f \varphi_k, \qquad k = 1, \ldots, M,$$

donde

$$\hat{a}_{kl} := \int_a^b (\varphi_l' \varphi_k' + \varphi_l \varphi_k), \qquad k, l = 1, \ldots, M.$$

¿Qué hemos ganado? La integración por partes elimina una derivada, así que ya no hay que pedir que las funciones básicas tengan dos derivadas. De hecho, incluso pedir $\varphi_l \in C^1((a,b))$ es excesivo. Los coeficientes $\hat{a}_{kl}$ tienen sentido para funciones φ_l mucho más generales, por ejemplo para funciones derivables a trozos como las definidas en (4.2).

La integración por partes reduce los requisitos de regularidad sobre las φ_l. Cuanto menores sean dichos requisitos, mayor será la clase de funciones base que podemos considerar. Esto facilitará nuestra tarea de construir funciones base que se anulen en casi todo el intervalo. Tenemos así un nuevo principio:

- Integrar por partes para reducir lo más posible las exigencias de regularidad sobre el espacio V_h.

4.2 Un poco de análisis funcional. Espacios de Sobolev

Llegados a este punto, nos gustaría saber cuál es la regularidad mínima necesaria para que las ecuaciones (4.6), o equivalentemente los coeficientes $\hat{a}_{kl}$, tengan sentido. Esto nos lleva a estudiar los llamados *espacios de Sobolev*.

Sea $I = (a, b)$ un intervalo, no necesariamente acotado, y sea $p \in \mathbb{R}$ con $1 \leq p \leq \infty$. Denotamos por $C_c^1(I)$ al conjunto de las funciones derivables en I con soporte compacto en I.

Definición 4.3. *El espacio de Sobolev $W^{1,p}(I)$ se define por*

$$W^{1,p}(I) = \{u \in L^p(I) : \exists g \in L^p(I) \text{ tal que } \int_a^b u\varphi' = -\int_a^b g\varphi \quad \forall \varphi \in C_c^1(I)\}.$$

Notación. Se suele escribir $H^1(I) = W^{1,2}(I)$.

La función g desempeña el papel de u' cuando se integra por partes. Así, para cada $u \in W^{1,p}(I)$ se escribe $g = u'$, y se dice que g es la *derivada débil* de u. De esta forma, $W^{1,p}(I)$ es el espacio de las funciones de $L^p(I)$ cuya derivada (débil) está en $L^p(I)$.

Observación. Si $u \in C^1(I) \cap L^p(I)$ y $u' \in L^p(I)$ (aquí u' es la derivada usual de u), entonces $u \in W^{1,p}(I)$. Además la derivada usual de u coincide con la derivada de u en el sentido de $W^{1,p}$. En particular, si I está acotado, entonces $C^1(\overline{I}) \subset W^{1,p}(I)$ para todo $1 \leq p \leq \infty$.

Ejemplo. Consideramos la función $u(x) = |x|$ para $x \in (-1, 1)$. Dada $\varphi \in C_c^1(I)$ se tiene que

$$\int_I u\varphi' = -\int_{-1}^0 x\varphi' + \int_0^1 x\varphi' = \int_{-1}^0 \varphi + \int_0^1 \varphi,$$

donde la última igualdad proviene de integrar por partes teniendo en cuenta que $\varphi(\pm 1) = 0$. Por tanto,

$$\int_I u\varphi' = -\int_I g\varphi,$$

con

$$g(x) = \begin{cases} -1 & \text{si } -1 < x < 0, \\ 1 & \text{si } 0 < x < 1. \end{cases}$$

Evidentemente $g \in L^p(I)$ para todo $1 \leq p \leq \infty$. Así $u \in W^{1,p}(I)$ para todo $1 \leq p \leq \infty$, y la derivada débil de u es $u' = g$. Sin embargo, la función $u(x) = |x|$ no es derivable en $x = 0$ en el sentido usual, y por tanto $u \notin C^1(I)$.

De la misma forma se prueba que si el intervalo I está acotado, entonces toda función continua en $\overline{I}$ y derivable con continuidad a trozos en $\overline{I}$ pertenece a $W^{1,p}(I)$ para todo $1 \leq p \leq \infty$.

Notaciones. El espacio $W^{1,p}$ está dotado de la norma

$$\|u\|_{W^{1,p}} = (\|u\|_{L^p}^p + \|u'\|_{L^p}^p)^{1/p}.$$

En H^1 esta norma corresponde al producto escalar

$$\langle u, v\rangle_{H^1} = \langle u, v\rangle_{L^2} + \langle u', v'\rangle_{L^2}.$$

Teorema 4.4. *El espacio $W^{1,p}$ es un espacio de Banach para $1 \leq p \leq \infty$.*

Prueba. Sea $\{u_n\}$ una sucesión de Cauchy en $W^{1,p}$; entonces $\{u_n\}$ y $\{u'_n\}$ son sucesiones de Cauchy en L^p. Por la completitud $u_n \to u$ en L^p y $u'_n \to g$ en L^p. Se tiene

$$\int_a^b u_n\varphi' = -\int_a^b u'_n\varphi \qquad \forall\varphi \in C_c^1(I),$$

y en el límite

$$\int_a^b u\varphi' = -\int_a^b g\varphi \qquad \forall\varphi \in C_c^1(I),$$

Por tanto $u \in W^{1,p}$, $u' = g$ y $\|u_n - u\|_{W^{1,p}} \to 0$. □

Consecuentemente, $H^1(I)$ es un espacio de Hilbert. Dicho espacio parece ser el adecuado para plantear las ecuaciones (4.6), salvo porque sus elementos no satisfacen en general las condiciones de frontera homogéneas. Las funciones de $W^{1,p}$ están definidas en casi todo punto, y la frontera ∂I tiene medida cero, así que en principio no tiene sentido pedir que las funciones de la base φ_l, y por tanto la solución aproximada u_h, valgan cero en ∂I. Nos tendremos que contentar con que esa propiedad se verifique en un sentido débil. Para ello introducimos un nuevo espacio.

Definición 4.5. *Dado $1 \leq p < \infty$, se designa por $W_0^{1,p}(I)$ al cierre de $C_c^1(I)$ en $W^{1,p}(I)$. Se nota $H_0^1(I) = W_0^{1,2}(I)$.*

El espacio $W_0^{1,p}$ está dotado de la norma inducida por $W^{1,p}$ y con ésta norma es un espacio de Banach; el espacio H_0^1 está dotado del producto escalar inducido por H^1, y con este producto escalar es un espacio de Hilbert.

Pedir que una función u esté en $W_0^{1,p}(I)$ es una forma débil de pedir que $u = 0$ en ∂I, en el sentido de que podemos aproximar u tanto como queramos (en la norma de $W^{1,p}$) por funciones de $C_c^1(I)$ que tienen esa propiedad.

En realidad no es completamente satisfactorio que las condiciones de contorno se satisfagan tan sólo en un sentido débil. Afortunadamente, en dimensión $d = 1$ las funciones de $W^{1,p}(I)$ son mejores de lo que cabría esperar, lo que permitirá dar sentido al valor que toman en un punto.

Teorema 4.6. *Sea $u \in W^{1,p}(I)$; existe una función $\overline{u} \in C(\overline{I})$ tal que $u = \overline{u}$ c.t.p. en I; es decir, u tiene un representante continuo.*

Prueba. Basta tomar $\overline{u}(x) - \overline{u}(y) = \int_y^x u'(t)\, dt$, $\forall x, y \in I$. $\square$

En general se sustituirá sistemáticamente u por su representante continuo; y con el fin de no recargar la notación se designará también por u al representante continuo. Podemos por tanto hablar del valor de una función $u \in W^{1,p}(I)$ en un punto de $\overline{I}$. Nos estaremos refiriendo al valor de su representante continuo en dicho punto. En particular, las funciones de $W_0^{1,p}(I)$ se pueden caracterizar de la forma siguiente.

Teorema 4.7. *Sea $u \in W^{1,p}(I)$; entonces $u \in W_0^{1,p}$ si y solamente si $u = 0$ sobre ∂I.*

Así que ya tenemos el espacio que estábamos buscando: pediremos que $\varphi_l \in H_0^1$.

Para terminar la sección probamos, usando la caracterización que acabamos de dar, que en $W_0^{1,p}(I)$ la cantidad $\|u'\|_{L^p}$ es una norma equivalente a la norma de $W^{1,p}$.

Teorema 4.8 (Desigualdad de Poincaré). *Sea I acotado. Existe una constante C, dependiente de $|I|$, tal que*

$$\|u\|_{W^{1,p}} \leq C\|u'\|_{L^p} \qquad \forall u \in W_0^{1,p}.$$

Prueba. Para $u \in W_0^{1,p}(I)$ se tiene

$$|u(x)| = |u(x) - u(a)| = \left| \int_a^x u' \right| \leq \|u'\|_{L^1},$$

de donde se deduce el resultado gracias a la desigualdad de Hölder. $\square$

4.3 Soluciones débiles. Formulación de Galerkin

Conviene que revisemos lo que hemos hecho. Para tener más flexibilidad en la elección del espacio de dimensión finita V_h, hemos rebajado las hipótesis de regularidad sobre las funciones φ_l, pidiendo tan sólo que $\varphi_l \in H_0^1$. Con esto las ecuaciones de Galerkin (4.6) tienen perfecto sentido y permiten calcular la aproximación $u_h \in V_h$. Estas ecuaciones caracterizan a u_h como la única función en V_h (subespacio de dimensión finita de H_0^1) tal que

$$\int_a^b u_h'\varphi' + \int_a^b u_h\varphi = \int_a^b f\varphi \qquad \forall \varphi \in V_h. \tag{4.7}$$

Obsérvese que $u_h \in H_0^1(I)$, pero que en general $u_h \notin C^2(I)$. Es decir, la aproximación tiene menos regularidad que la verdadera solución. Esto, que en principio no es demasiado importante, produce sin embargo un punto débil en nuestros argumentos. Para llegar a las ecuaciones (4.6) hemos partido de las ecuaciones (4.4), que involucran al defecto d_h; y el defecto no está bien definido para funciones que sólo están en H_0^1, pero no en C^2. Esta situación parece inevitable, ¿acaso es posible escribir un problema en derivadas segundas sin usar derivadas segundas? Sorprendentemente la respuesta es afirmativa en muchos casos importantes, y conduce al concepto de *solución débil*.

Sea u una solución clásica de (4.1). Si multiplicamos la ecuación por una función $\varphi \in C_c^1(I)$ e integramos por partes se tiene que

$$\int_a^b (u'\varphi' + u\varphi) = \int_a^b f\varphi. \tag{4.8}$$

De hecho, por ser $C_c^1(I)$ denso en $H_0^1(I)$, la igualdad es cierta para cada $\varphi \in H_0^1(I)$. En efecto, dada $\varphi \in H_0^1(I)$, existe una sucesión $\{\varphi_n\}$ de funciones de $C_c^1(I)$ tal que $\|\varphi_n - \varphi\|_{H^1(I)} \to 0$ cuando $n \to \infty$. Para cada una de las funciones φ_n se tiene que

$$\int_a^b (u'\varphi_n' + u\varphi_n) = \int_a^b f\varphi_n. \tag{4.9}$$

Ahora bien, usando la desigualdad de Cauchy-Schwarz tenemos que

$$\left|\int_a^b (u'\varphi_n' + u\varphi_n) - \int_a^b (u'\varphi' + u\varphi)\right| = |\langle u, \varphi_n - \varphi\rangle_{H^1}| \leq \|u\|_{H_1}\|\varphi_n - \varphi\|_{H^1},$$
$$\left|\int_a^b f\varphi_n - \int_a^b f\varphi\right| = |\langle f, \varphi_n - \varphi\rangle_{L^2}| \leq \|f\|_{L^2}\|\varphi_n - \varphi\|_{L^2}.$$

Por tanto, pasando al límite en (4.9) obtenemos que (4.8) es cierta para cualquier $\varphi \in H_0^1(I)$.

En realidad para que (4.8) tenga sentido basta con que $f \in L^2(I)$ y $u \in H_0^1(I)$, lo que motiva la siguiente definición.

Definición 4.9. *Sea $f \in L^2(I)$. Una solución débil de* (4.1) *es una función $u \in H_0^1(I)$ que verifica* (4.8) *para toda función $\varphi \in H_0^1(I)$.*

Con esta definición las condiciones de contorno están implícitas en el espacio funcional.

Observación. Sea $f \in C([a,b])$. Hemos visto que la solución clásica es solución débil. Recíprocamente, si una solución débil es $C^2(I) \cap C(\overline{I})$, entonces es solución clásica. Para verlo, integramos por partes la expresión (4.8) y obtenemos que

$$\int_a^b (-u'' + u - f)\varphi = 0 \qquad \forall \varphi \in C_c^1(I).$$

Supongamos que $g = -u'' + u - f$ es distinto de cero para algún punto $x_0 \in I$. Para concretar supongamos que $g(x_0) > 0$. Por continuidad, existe $\varepsilon > 0$ tal que $g(x) > 0$ en $(x_0 - \varepsilon, x_0 + \varepsilon)$. Tomando $\varphi \in C_c^1(I)$ tal que $\varphi \geq 0$, $\varphi(x_0) > 0$, $\operatorname{sop}\varphi \subset (x_0 - \varepsilon, x_0 + \varepsilon)$, tenemos que

$$\int_a^b (-u'' + u - f)\varphi > 0,$$

una contradicción. Concluimos que

$$-u'' + u = f \qquad \text{si } a < x < b.$$

Las condiciones de contorno se verifican, por ser u una función de $H_0^1(I)$.

Sea V_h un subespacio de dimensión finita de $H_0^1(I)$. Si pedimos que se satisfaga la condición de solución débil (4.8), pero restringiéndonos a funciones u_h y φ del subespacio V_h (en lugar de considerar todo el espacio H_0^1), obtenemos (4.7), que es la llamada *forma de Galerkin del método de elementos finitos.* Nótese que para llegar a (4.7), que es equivalente a (4.6), no hemos usado para nada el defecto d_h, ni nada que nos exija una mayor regularidad para u_h.

Se podría objetar que de esta forma lo que hemos obtenido es una aproximación a una solución débil del problema. Sabemos que la solución clásica es también solución débil; pero, ¿no habrá más soluciones débiles? La respuesta viene dada por el siguiente teorema.

Teorema 4.10. *El problema* (4.1) *tiene una única solución débil.*

Prueba. Sean $u, v \in H_0^1$ dos soluciones débiles. De (4.8) se tiene que

$$\int_a^b (u-v)'\varphi' + \int_a^b (u-v)\varphi = 0 \qquad \forall \varphi \in H_0^1.$$

Tomando $\varphi = u - v$ se tiene que

$$\int_a^b ((u-v)')^2 + \int_a^b (u-v)^2 = 0,$$

y por tanto $u - v = 0$ c.t.p., es decir, $u = v$. $\square$

El método de elementos finitos da una aproximación de la solución débil incluso cuando ésta no es clásica. Pero, ¿sucede esto alguna vez? La respuesta es afirmativa. Un ejemplo sencillo es tomar el problema (4.1) con un dato $f \in L^2(I)$, pero $f \notin C([a,b])$. El problema tiene una solución débil (ver el teorema 4.12 más adelante), pero no es, evidentemente, clásica. Veamos que regularidad tiene una solución débil que corresponda a un dato con estas características. Sea u solución débil, entonces $u \in H_0^1$ y además

$$\int_a^b u'\varphi' = \int_a^b (f-u)\varphi \qquad \forall \varphi \in C_c^1.$$

Dado que $f - u \in L^2$, concluimos que $u' \in H^1$. La función u' está por tanto en el espacio $H^2(I)$, que se define por recurrencia como

$$H^2(I) = \{u \in H^1(I) : u' \in H^1(I)\}.$$

A la derivada débil de u', $(u')'$ se la llama derivada débil segunda de u; se denota por u''. En nuestro caso se tiene que $u'' = u - f$.

Resumiendo, si $f \in L^2$ entonces $u \in H^2$. ¿Qué pasa si $f \in C([a, b])$? Tenemos entonces que $u'' = u - f \in C([a, b])$, y por tanto u es solución clásica.

4.4 Minimización de funcionales. Formulación de Ritz

Un paradigma que se remonta a Euler es que el mundo físico se explica imponiendo que ciertas cantidades ("acciones") son mínimas; por ello no es extraño que podamos representar algunos problemas de ecuaciones ordinarias y parciales (que tratan de representar fenómenos físicos) a otros de minimizaciòn.

Sea $f \in L^2(I)$. Definimos un *funcional de energía* $\mathcal{J} : H_0^1 \mapsto \mathbb{R}$ mediante

$$\mathcal{J}(v) = \frac{1}{2}\int_a^b ((v')^2 + v^2) - \int_a^b fv.$$

Teorema 4.11. *(i) Si $u \in H_0^1$ es tal que $\mathcal{J}(u) \leq \mathcal{J}(v)$ para todo $v \in H_0^1(I)$, entonces u es solución débil de (4.1).*

(ii) Recíprocamente, si $u \in H_0^1(I)$ es solución débil de (4.1), entonces $\mathcal{J}(u) \leq \mathcal{J}(v)$ para todo $v \in H_0^1(I)$.

Prueba. (i) Dada $\varphi \in H_0^1$, definimos $g(\epsilon) = \mathcal{J}(u + \epsilon\varphi)$, $\epsilon \in \mathbb{R}$. Se tiene que $g(0) \leq g(\epsilon)$ para todo $\epsilon \in \mathbb{R}$. Si la derivada $g'(0)$ existe, se tendrá que $g'(0) = 0$. Ahora bien,

$$g(\epsilon) = \mathcal{J}(u) + \epsilon\int_a^b (u'\varphi' + u\varphi - f\varphi) + \frac{\epsilon^2}{2}\int_a^b ((\varphi')^2 + \varphi^2), \tag{4.10}$$

y por tanto

$$g'(0) = \int_a^b (u'\varphi' + u\varphi - f\varphi).$$

La condición $g'(0) = 0$ para todo $\varphi \in H_0^1$, es precisamente la condición de solución débil de (4.1).

(ii) Si $u \in H_0^1(I)$ es solución débil, entonces (4.10) se convierte en

$$\mathcal{J}(u + \epsilon\varphi) = \mathcal{J}(u) + \frac{\epsilon^2}{2}\int_a^b ((\varphi')^2 + \varphi^2) \geq \mathcal{J}(u) \qquad \forall \epsilon \in \mathbb{R},\ \varphi \in H_0^1.$$

Para cualquier $v \in H_0^1(I)$ tomamos $\varphi = v - u$ y $\epsilon = 1$ y obtenemos que $\mathcal{J}(v) \geq \mathcal{J}(u)$. $\square$

El área de la matemáticas que estudia problemas de minimización es el *cálculo de variaciones*. Hemos visto que minimizar el funcional es equivalente a encontrar una solución débil de una ecuación. Dicha ecuación se conoce como *ecuación de Euler-Lagrange* del problema de minimización.

A partir del teorema observamos que una forma alternativa de encontrar una solución aproximada al problema (4.1) es minimizar $\mathcal{J}(v)$ en un subespacio $V_h \subset H_0^1$ de dimensión finita; es decir, buscamos $u_h \in V_h$ tal que

$$\mathcal{J}(u_h) \leq \mathcal{J}(v_h) \qquad \forall v_h \in V_h.$$

Sea $\{\varphi_1, \ldots, \varphi_m\}$ una base de V_h. Cualquier $v_h \in V_h$ se puede escribir como

$$v_h = \sum_{l=1}^m \gamma_l \varphi_l,$$

y el valor del funcional será

$$\mathcal{J}(v_h) = \frac{1}{2}\sum_{k,l=1}^m \gamma_l\gamma_k \int_a^b (\varphi_l'\varphi_k' + \varphi_l\varphi_k) - \sum_{k=1}^m \gamma_k \int_a^b f\varphi_k.$$

Minimizar el funcional en V_h es por tanto equivalente a minimizar la forma cuadrática

$$Q(\gamma_1, \ldots, \gamma_m) = \sum_{k,l=1}^m \gamma_k \hat{a}_{kl}\gamma_l - \sum_{k=1}^m \gamma_k \int_a^b f\varphi_k, \qquad \hat{a}_{kl} = \int_a^b (\varphi_l'\varphi_k' + \varphi_l\varphi_k)$$

La condición $\nabla Q(\gamma_1, \ldots, \gamma_m) = 0$ es equivalente al sistema lineal (4.6). Obtenemos las ecuaciones de Galerkin partiendo de un punto de vista distinto (aunque equivalente). Esta forma de llegar al sistema (4.6) se conoce como *formulación de Ritz* del método de elementos finitos.

4.5 Formulación abstracta

Nos gustaría generalizar lo que hemos hecho para el problema (4.1) a otros problemas. No queremos repetir los mismos argumentos cada vez, así que introduciremos una formulación abstracta que nos permita tratar de forma unificada una amplia clase de problemas.

Empezamos por escribir la formulación débil del problema (4.1) de una manera susceptible de generalización. Tenemos un espacio de Hilbert $V = H_0^1$, una forma bilineal sobre $V \times V$ dada por

$$a(u, v) = \int_a^b (u'v' + uv),$$

y una forma lineal sobre V dada por

$$L(v) = \int_a^b fv.$$

Lo que hemos hecho ha sido encontrar una función $u \in V$ tal que

$$a(u, v) = L(v), \qquad \forall v \in V;$$

alternativamente, hemos encontrado $u \in V$ tal que $\mathcal{J}(u) = \min_{v \in V} \mathcal{J}(v)$, siendo

$$\mathcal{J}(v) = \frac{1}{2} a(v, v) - L(v).$$

¿Podemos hacer esto para otras formas a y L, definidas quizá sobre otro espacio de Hilbert distinto? La respuesta es afirmativa; pero habrá que pedir que las formas a y L tengan las propiedades que nos han permitido probar los teoremas en el caso del problema (4.1). Eso nos lleva a introducir algunos conceptos nuevos.

Sea V un espacio de Hilbert con producto escalar $\langle\cdot,\cdot\rangle_V$ y norma correspondiente $\|\cdot\|_V$. Sean $a(\cdot,\cdot)$ una forma bilineal sobre $V\times V$ y L una forma lineal sobre V tales que:

- $a(\cdot,\cdot)$ es *simétrica*;
- $a(\cdot,\cdot)$ es *continua*, es decir, existe una constante $\mu>0$ tal que

$$|a(u,v)|\leq \mu\|u\|_V\|v\|_V \qquad \forall u,v\in V;$$

- $a(\cdot,\cdot)$ es *coerciva*, es decir, existe una constante $\alpha>0$ tal que

$$a(u,u)\geq \alpha\|u\|_V^2 \qquad \forall v\in V;$$

- L es *continua*, esto es, existe una constante $\Lambda>0$ tal que

$$|L(v)|\leq \Lambda\|v\|_V \qquad \forall v\in V.$$

Consideramos los siguientes problemas abstractos:

Problema de minimización (M): encontrar $u\in V$ tal que

$$\mathcal{J}(u)=\min_{v\in V}\mathcal{J}(v),$$

donde

$$\mathcal{J}(v)=\frac{1}{2}a(v,v)-L(v).$$

Problema en forma débil (D): encontrar $u\in V$ tal que

$$a(u,v)=L(v) \qquad \forall v\in V. \tag{4.11}$$

Teorema 4.12. *Los problemas* (M) *y* (D) *son equivalentes; esto es,* $u \in V$ *es solución de* (M) *si y sólo si es solución de* (D)*. Estos problemas tienen una única solución, y se satisface la estimación de estabilidad*

$$\|u\|_V \leq \frac{\Lambda}{\alpha}. \tag{4.12}$$

Prueba. La existencia de solución de (D) se obtiene usando el teorema de Lax-Milgram, que es una variante del teorema de representación de Riesz (para los detalles, ver por ejemplo [B]).

La prueba de la equivalencia de (M) y (D) es análoga a la que dimos en el Teorema 4.11 para el problema (4.1). Probemos, por ejemplo, que si $u \in V$ satisface (M) entonces satisface (D). Sean $v \in V$ y $\epsilon \in \mathbb{R}$ arbitrarios. Entonces $u + \epsilon v \in V$, y por tanto

$$\mathcal{J}(u) \leq \mathcal{J}(u + \epsilon v) \qquad \forall \epsilon \in \mathbb{R}.$$

Definiendo $g(\epsilon) = \mathcal{J}(u + \epsilon v)$ vemos que g tiene un mínimo en $\epsilon = 0$. Ahora bien,

$$\begin{aligned} g(\epsilon) &= \tfrac{1}{2}a(u + \epsilon v, u + \epsilon v) - L(u + \epsilon v) \\ &= \tfrac{1}{2}a(u, u) + \tfrac{\epsilon}{2}a(u, v) + \tfrac{\epsilon}{2}a(v, u) + \tfrac{\epsilon^2}{2}a(v, v) - L(u) - \epsilon L(v) \\ &= \tfrac{1}{2}a(u, u) - L(u) + \epsilon a(u, v) - \epsilon L(v) + \tfrac{\epsilon^2}{2}a(v, v), \end{aligned} \tag{4.13}$$

donde hemos usado la simetría de $a(\cdot, \cdot)$. Se sigue que

$$0 = g'(0) = a(u, v) - L(v),$$

es decir, u es solución de (D).

Para probar la implicación contraria observamos que si $u \in V$ es solución de (D), entonces (4.13) se convierte en

$$\mathcal{J}(u + \epsilon v) = \mathcal{J}(u) + \frac{\epsilon^2}{2}a(v, v) \geq \mathcal{J}(u) \qquad \forall \epsilon \in \mathbb{R}, v \in V.$$

Dado $\varphi \in V$ arbitrario, tomamos $\varphi = v - u$ y $\epsilon = 1$, y tenemos que $\mathcal{J}(\varphi) \geq \mathcal{J}(u)$, es decir, u es solución de (M).

Para probar la estimación de estabilidad tomamos $v = u$ en (4.11) y usamos la coercividad de a y la continuidad de L para obtener que

$$\alpha \|u\|_V^2 \leq a(u, u) = L(u) \leq \Lambda \|u\|_V.$$

Dividiendo por $\|u\|_V \neq 0$ obtenemos (4.12).

Veamos la unicidad. Sean $u_1, u_2 \in V$ tales que

$$a(u_i, v) = L(v) \qquad \forall v \in V, \ i = 1, 2.$$

Restando tenemos que $u_1 - u_2 \in V$ satisface

$$a(u_1 - u_2, v) = 0 \qquad \forall v \in V.$$

Tomando $v = u_1 - u_2$ y usando la coercividad de a tenemos finalmente que

$$\alpha \|u_1 - u_2\|_V \leq a(u_1 - u_2, u_1 - u_2) = 0,$$

es decir, $u_1 = u_2$. $\square$

Observación. Se puede probar la existencia de una única solución $u \in V$ del problema (D) sin la condición de simetría sobre a. Sin embargo, en este caso no habrá un problema de minimización asociado.

Observación. Para que el problema de minimización (M) esté bien planteado no es necesario que V sea un espacio de Hilbert; basta con que sea un subconjunto convexo cerrado de un espacio de Hilbert. El problema (M) tendrá entonces solución única, aunque no será equivalente al problema (D) (ver el ejemplo de condiciones de Dirichlet no homogéneas más adelante).

Ejemplo. (Condiciones de Neumann). Consideramos el problema

$$-u'' + u = f \quad \text{si } a < x < b, \qquad u'(a) = \alpha, \qquad u'(b) = \beta.$$

Una solución clásica es una función $u \in C^2((a, b)) \cap C^1([a, b])$ que verifica la ecuación y las condiciones de contorno de forma puntual. Nótese que para dar sentido a las condiciones de contorno hemos tenido que pedir que la función sea derivable hasta la frontera.

Para escribir el problema en forma débil, multiplicamos la ecuación por $v \in H^1$ e integramos por partes. Imponiendo las condiciones de frontera, se tiene que

$$\int_a^b (u'v' + uv) = \int_a^b fv - \alpha v(a) + \beta v(b) \qquad \forall v \in H^1. \tag{4.14}$$

Una solución débil del problema de Neumann será una función $u \in H^1$ que verifique (4.14).

Observación. En realidad, para ser rigurosos en el razonamiento anterior tendríamos que haber multiplicado la ecuación por una función $v \in C_c^1(\mathbb{R})$ y después haber utilizado un argumento de densidad para probar el resultado para $v \in H^1(I)$ general. El resultado de densidad que se necesita, cuya prueba se puede encontrar por ejemplo en [B], es el siguiente.

Lema 4.13. *Sea $u \in W^{1,p}(I)$ con $1 \leq p < \infty$. Entonces existe una sucesión $\{u_n\}$ en $C_c^\infty(\mathbb{R})$ tal que $u_{n|I} \to u$ en $W^{1,p}(I)$.*

En posteriores ejemplos omitiremos este tipo de argumentos de densidad.

Queremos formular el problema en forma abstracta. Para ello tomamos $V = H^1$, $a(u,v) = \int_a^b (u'v' + uv)$ y $L(v) = \int_a^b fv - \alpha v(a) + \beta v(b)$. Se puede probar que $a(\cdot,\cdot)$ y $L(\cdot)$ están en las hipótesis del Teorema 4.12; por tanto existe una única solución débil del problema y además coincide con la función $u \in H^1$ que minimiza el funcional

$$\mathcal{J}(v) = \frac{1}{2}\int_a^b ((v')^2 + v^2) - \int_a^b fv + \alpha v(a) + \beta v(b).$$

Obsérvese que no se exige explícitamente que la solución satisfaga las condiciones de frontera. Las condiciones aparecen implícitamente en la forma lineal L, y a través de ella en el funcional de energía. En el caso de datos Dirichlet homogéneos, por el contrario, se imponen explícitamente los datos de contorno a la solución, exigiendo que ésta pertenezca al espacio funcional H_0^1. Las condiciones de frontera que no necesitan ser impuestas explícitamente se llaman *condiciones de frontera naturales*; las que sí lo necesitan se llaman *condiciones de frontera esenciales.*

Para finalizar el ejemplo, probaremos que una solución débil que pertenezca a $C^2(I) \cap C^1(\overline{I})$ es solución clásica. Nótese que la función está por hipótesis en el espacio correcto. Empezamos probando que se satisface la ecuación para todo $x \in I$. Para ello, consideramos (4.14) para el caso particular de $v \in C_c^1(I)$. Integrando por partes se tiene que

$$\int_a^b (-u'' + u - f)v = 0.$$

Argumentando como lo hicimos en el caso de datos de Dirichlet homogéneos se obtiene que $-u'' + u = f$ para todo $x \in I$.

Una vez probado esto volvemos a (4.14), y tomamos ahora una función $v \in C^1(\overline{I})$ tal que $v(a) = 0$, $v(b) = 1$. Integrando por partes se tiene que

$$\int_a^b (-u'' + u - f)v = \beta - v(b).$$

Dado que ya hemos probado que se verifica la ecuación, se tiene que $\beta - v(b) = 0$, es decir, la condición de contorno en $x = b$. Para probar la condición de contorno en $x = a$ hay que tomar $v \in C^1(\overline{I})$ tal que $v(a) = 1$, $v(b) = 0$.

Ejemplo. (Condiciones de contorno mixtas). Consideramos el problema

$$-u'' + u = f \quad \text{si } a < x < b, \qquad u(a) = 0, \qquad u'(b) = \beta. \tag{4.15}$$

La condición Dirichlet es esencial, y tiene que aparecer en el espacio funcional. Así que para obtener la formulación débil multiplicamos la ecuación por $v \in H := \{v \in H^1 : \ v(a) = 0\}$. Tras integrar por partes, imponiendo las condiciones de frontera, llegamos a

$$\int_a^b (u'v' + uv) = \int_a^b fv + \beta v(b) \qquad \forall v \in H. \tag{4.16}$$

Una solución débil del problema (4.15) será una función $u \in H$ que verifique (4.16).

Para escribir el problema en forma abstracta tomamos $V = H$, $a(u,v) = \int_a^b (u'v' + uv)$, $L(v) = \int_a^b fv + \beta v(b)$. La condición natural $u'(b) = \beta$ aparece en el funcional, y no en el espacio H.

Ejemplo. (Condición de Robin). Consideramos el problema

$$-u'' + u = f \quad \text{si } a < x < b, \qquad u'(a) - ku(a) = 0, \qquad u(b) = 0, \tag{4.17}$$

donde $k \geq 0$ es un dato del problema. Necesitamos un espacio funcional que recoja la condición de frontera esencial; $H := \{v \in H^1 : \ v(b) = 0\}$ cumple ese requisito. Una solución débil de (4.17) será una función $u \in H$ que verifique

$$\int_a^b (u'v' + uv) + ku(a)v(a) = \int_a^b fv \qquad \forall v \in H.$$

La condición de frontera de tipo Robin es una condición natural, pues aparece en el funcional; lo hace a través de la forma bilineal $a(u,v) = \int_a^b (u'v' + uv) + ku(a)v(a)$, siendo $L(v) = \int_a^b fv$. Nótese la diferencia con el problema de Neumann, donde la condición de frontera aparecía en la forma lineal L.

Hasta aquí hemos tratado problemas en una dimensión. Para tratar problemas en más dimensiones tenemos que introducir los espacios de Sobolev correspondientes.

Definición 4.14. *Sea $\Omega \subset \mathbb{R}^d$ abierto. El espacio de Sobolev $H^1(\Omega)$ se define como*

$$\begin{aligned} H^1(\Omega) = \{u \in L^2(\Omega) : \exists g_1, \dots, g_d \in L^2(\Omega) \\ \textit{tales que } \textstyle\int_\Omega u \frac{\partial \varphi}{\partial x_i} = -\int_\Omega g_i \varphi \ \forall \varphi \in C_c^1(\Omega) \ \forall i = 1, \dots, d\}. \end{aligned}$$

El espacio $H^1(\Omega)$ está dotado del producto escalar

$$\langle u, v \rangle_{H^1} = \langle u, v \rangle_{L^2} + \sum_{i=1}^d \left\langle \frac{\partial u}{\partial x_i}, \frac{\partial v}{\partial x_i} \right\rangle_{L^2},$$

y de la norma inducida por él

$$\|u\|_{H^1} = \left(\|u\|_{L^2}^2 + \sum_{i=1}^{d} \left\| \frac{\partial u}{\partial x_i} \right\|_{L^2}^2 \right)^{1/2}.$$

Con esta norma es un espacio de Hilbert.

También necesitamos un espacio que recoja las condiciones Dirichlet (que son esenciales).

Definición 4.15. *Sea $\Omega \subset \mathbb{R}^d$ abierto. El espacio $H_0^1(\Omega)$ es la clausura de $C_c^1(\Omega)$ en $H^1(\Omega)$.*

El espacio H_0^1 es un espacio de Hilbert con la norma que hereda de H^1.

Para $d > 1$ no es cierto que las funciones de H^1 tengan un representante continuo. ¿En qué sentido verifican entonces que $u = 0$ para $x \in \partial\Omega$? Esto lleva a la *teoría de trazas*. Éste es un asunto bastante delicado, y no lo trataremos aquí. Digamos simplemente que existe un operador lineal y continuo

$$tr : H^1(\Omega) \mapsto L^2(\partial\Omega)$$

tal que $tr(u)$ coincide con $u_{|\partial\Omega}$ cuando u es una función continua. De esta forma el espacio de Sobolev $H_0^1(\Omega)$ verifica que

$$H_0^1(\Omega) = \{u \in H^1(\Omega) : tr(u) = 0 \text{ c.t.p. en } \partial\Omega\}.$$

La continuidad del operador garantiza que

$$\|tr(u)\|_{L^2(\partial\Omega)} \leq C\|u\|_{H^1(\Omega)}.$$

Por fortuna estas complicaciones técnicas no nos afectan en la práctica, y fundamentalmente sirven para asegurar que existe un espacio funcional en el que construir la teoría.

Veamos un par de ejemplos con $d > 1$. Para pasar a la formulación débil será fundamental la fórmula de Green (desempeñará el papel de la integración por partes)

$$\int_\Omega v\Delta u = -\int_\Omega \nabla v \cdot \nabla u + \int_{\partial\Omega} v\nabla u \cdot n,$$

donde $n = n(x)$ es el vector normal exterior al dominio en el punto $x \in \partial\Omega$. Esta fórmula se obtiene a partir de la igualdad

$$\operatorname{div}(v\nabla u) = \nabla v \cdot \nabla u + v\Delta u$$

integrando en Ω y aplicando el teorema de la divergencia.

Ejemplo. (Problema de Dirichlet homogéneo d-dimensional). Consideramos el problema

$$\begin{cases} -\Delta u + u = f & \text{en } \Omega \subset \mathbb{R}^d, \\ u = 0 & \text{en } \partial\Omega. \end{cases}$$

Multiplicamos por $v \in H_0^1$, aplicamos la fórmula de Green y las condiciones de contorno y obtenemos

$$\int_\Omega \nabla u \cdot \nabla v + \int_\Omega uv = \int_\Omega fv \qquad \forall v \in H_0^1(\Omega),$$

que es la forma débil del problema. Para aplicar el Teorema 4.12 tomamos $V = H_0^1(\Omega)$, $a(u,v) = \int_\Omega \nabla u \cdot \nabla v$ y $L(v) = \int_\Omega fv$. Veamos que se cumplen las condiciones del teorema. En efecto, $a(\cdot,\cdot)$ es obviamente bilineal y simétrica. Por otra parte,

$$a(u,u) = \|u\|_{H_0^1(\Omega)},$$

luego $a(\cdot,\cdot)$ es coerciva. La continuidad es consecuencia de la desigualdad de Cauchy-Schwarz, pues

$$a(u,v) = \langle u, v\rangle_{H_0^1(\Omega)}.$$

La linealidad de L es también trivial. La continuidad es consecuencia de la desigualdad de Cauchy-Schwarz en $L^2(\Omega)$,

$$|L(v)| = \left|\int_\Omega fv\right| \le \|f\|_{L^2(\Omega)}\|v\|_{L^2(\Omega)} \le \|f\|_{L^2(\Omega)}\|v\|_{H_0^1(\Omega)} \qquad \forall v \in H_0^1(\Omega).$$

Ejemplo. (Problema de Neumann d-dimensional). Consideramos el problema

$$\begin{cases} -\Delta u + u = f & \text{en } \Omega \subset \mathbb{R}^d, \\ \frac{\partial u}{\partial n} = g & \text{en } \partial\Omega. \end{cases}$$

Multiplicamos por $v \in H^1$ (recordemos que la condición de Neumann es natural), aplicamos la fórmula de Green y las condiciones de contorno y obtenemos

$$\int_\Omega \nabla u \cdot \nabla v + \int_\Omega uv = \int_\Omega f + \int_{\partial\Omega} gv \qquad \forall v \in H^1(\Omega),$$

que es la forma débil del problema. Para aplicar el Teorema 4.12 tomamos $V = H^1(\Omega)$, $a(u,v) = \int_\Omega \nabla u \cdot \nabla v$ y $L(v) = \int_\Omega fv + \int_{\partial\Omega} gv$. La forma bilineal $a(\cdot,\cdot)$ cumple las condiciones del teorema (copiar del ejemplo anterior). Para ver la continuidad de L observamos que

$$|L(v)| \le \left| \int_\Omega fv \right| + \left| \int_{\partial\Omega} gv \right|.$$

Para probar la continuidad del segundo sumando aplicamos la desigualdad de Cauchy-Schwarz en $L^2(\partial\Omega)$ y la continuidad del operador traza,

$$\left| \int_{\partial\Omega} gv \right| \le \|g\|_{L^2(\partial\Omega)} \|v\|_{L^2(\partial\Omega)} \le C\|g\|_{L^2(\partial\Omega)} \|v\|_{H^1(\Omega)} \qquad \forall v \in H^1(\Omega).$$

Ejemplo. (Problema de Dirichlet no homogéneo d-dimensional). Consideramos el problema

$$\begin{cases} -\Delta u + u = f & \text{en } \Omega \subset \mathbb{R}^d, \\ u = g & \text{en } \partial\Omega. \end{cases} \tag{4.18}$$

La condición de Dirichlet debería aparecer en el espacio funcional. Así pues, nos gustaría tomar

$$V = \{v \in H^1(\Omega) : \, v = g \text{ en } \partial\Omega\}.$$

El conjunto V no es un espacio vectorial, pero es un subconjunto convexo cerrado del espacio de Hilbert $H^1(\Omega)$, lo que será suficiente para nuestros propósitos. Sin embargo, si multiplicamos la ecuación por funciones test $v \in V$ y aplicamos la fórmula de Green, aparecerá un término de la forma $\int_{\partial\Omega} g \frac{\partial u}{\partial n}$, que no se puede incorporar ni a la forma bilineal ni a la lineal.

Con el fin de que no aparezca el término de borde molesto, utilizamos como tests funciones $v \in H_0^1(\Omega)$. Obtenemos que

$$\int_\Omega (\nabla u \cdot \nabla v + uv) = \int_\Omega fv \qquad \forall v \in H_0^1(\Omega). \tag{4.19}$$

Esto sugiere la siguiente definición: una solución débil del problema (4.18) es una función $u \in V$ que verifica (4.19). Tomando $a(u,v) = \int_\Omega \nabla u \cdot \nabla v$ y $L(v) = \int_\Omega fv$ el problema débil se puede escribir en forma abstracta: encontrar una función $u \in V$ que verifica que

$$a(u,v) = L(v) \qquad \forall v \in H_0^1(\Omega).$$

¿Cómo garantizar la existencia y unicidad de solución débil? En $H^1(\Omega)$ la forma bilineal $a(\cdot,\cdot)$ es continua y coerciva y la forma lineal $L(\cdot)$ es continua. Sin embargo, no podemos aplicar el teorema 4.12, pues, por una parte la solución u y los tests v no están en el mismo espacio, y por otra el conjunto V no es un espacio de Hilbert. Así que adoptamos una estrategia distinta, e intentamos ver la solución débil como solución de un problema de minimización. Para ello consideramos el funcional

$$\mathcal{J}(v) = \frac{1}{2} a(v,v) - L(v),$$

que está bien definido para funciones $v \in V$. Es fácil ver (repitiendo los argumentos de la prueba del teorema 4.12), que una función $u \in V$ es solución débil de (4.18) si y sólo si

$$\mathcal{J}(u) = \min_{v \in V} \mathcal{J}(v).$$

Así que el problema débil es equivalente al problema de encontrar un minimizante para $\mathcal{J}$ en V, que es un subconjunto convexo cerrado del espacio de Hilbert $H^1(\Omega)$. Por tanto el problema de minimización tiene una única solución. Concluimos que hay una única solución débil de (4.18).

4.6 Discretización. Convergencia

Sea ahora V_h un subespacio de V de dimensión finita M. Podemos formular los siguientes problemas discretos, análogos a los problemas (M) y (D):

- Problema (M_h): encontrar $u_h \in V_h$ tal que

$$\mathcal{J}(u_h) \leq \mathcal{J}(v) \qquad \forall v \in V_h.$$

- Problema (D_h): encontrar $u_h \in V_h$ tal que

$$a(u_h, v) = L(v) \qquad \forall v \in V_h. \tag{4.20}$$

Sea $\{\varphi_1, \dots, \varphi_M\}$ una base de V_h, de forma que cualquier $u \in V_h$ se puede representar en la forma

$$u = \sum_{j=1}^{M} \gamma_j \varphi_j, \qquad \gamma_j \in \mathbb{R}.$$

Usando la bilinealidad y la simetría de a, así como la linealidad de L, vemos que ambos problemas son equivalentes a resolver el sistema lineal

$$\sum_{i=1}^{M} a(\varphi_i, \varphi_j)\gamma_j = L(\varphi_i), \qquad i = 1, \dots, M.$$

Definiendo $\gamma = (\gamma_i) \in \mathbb{R}^M$, $b = (b_i) \in \mathbb{R}^M$ con $b_i = L(\varphi_i)$, $A = (a_{ij})$ una matriz $M \times M$ con elementos $a_{ij} = a(\varphi_i, \varphi_j)$, podemos escribir el sistema en forma matricial como

$$A\gamma = b.$$

La matriz A recibe el nombre de *matriz de rigidez* y b es el *vector de carga*. Estos nombres provienen de las aplicaciones más antiguas del método de elementos finitos, que correspondían a cálculos de estructuras mecánicas.

Lema 4.16. *La matriz de rigidez A es simétrica y definida positiva.*

Prueba. La simetría se hereda de la simetría de la forma bilineal a. Que A sea definida positiva es consecuencia de la coercividad. $\square$

Podemos probar el siguiente resultado básico.

Teorema 4.17. *Existe una única solución de los problemas equivalentes* (M_h) *y* (D_h)*. Además se tiene la siguiente estimación de estabilidad*

$$\|u_h\|_V \le \frac{\Lambda}{\alpha}.$$

Prueba. A es definida positiva, y por tanto no singular, lo que prueba la existencia y unicidad. Para probar la estimación de estabilidad se toma $v = u_h$ en (4.20); la coercividad de a y la continuidad de L producen

$$\alpha\|u_h\|_V^2 \le a(u_h, u_h) = L(u_h) \le \Lambda\|u_h\|_V.$$

Dividiendo por $\|u_h\|_V \neq 0$ se obtiene el resultado. Si $\|u_h\|_V = 0$ se cumple trivialmente. $\square$

A continuación obtenemos una estimación para el error.

Teorema 4.18 (Lema de Céa). *Sea* $u \in V$ *solución de* (D) *y* $u_h \in V_h$ *solución de* (D_h) *con* $V_h \subset V$*. Entonces*

$$\|u - u_h\|_V \le \frac{\mu}{\alpha}\|u - w\|_V \qquad \forall w \in V_h.$$

Prueba. Dado que $V_h \subset V$, de (4.11) tenemos en particular que

$$a(u, v) = L(v) \qquad \forall v \in V_h;$$

restando de aquí (4.20) llegamos a

$$a(u - u_h, v) = 0 \qquad \forall v \in V_h. \tag{4.21}$$

Para $w \in V_h$ arbitrario definimos $v = u_h - w$. Entonces $v \in V_h$. Usando la coercividad y la continuidad de a y la ecuación (4.21) tenemos

$$\begin{aligned}\alpha\|u - u_h\|_V^2 &\le a(u - u_h, u - u_h) = a(u - u_h, u - u_h) + a(u - u_h, v)\\ &= a(u - u_h, u - u_h + v) = a(u - u_h, u - w) \le \mu\|u - u_h\|_V\|u - w\|_V.\end{aligned}$$

Dividiendo por $\|u-u_h\|_V \neq 0$ se obtiene el resultado. $\square$

Podemos definir una nueva norma $\|\cdot\|_a$ en V, llamada *norma de energía*, mediante

$$\|v\|_a^2 = a(v,v), \qquad v \in V.$$

La coercividad y la continuidad de a nos dicen que esta norma es equivalente a la norma $\|\cdot\|_V$. La norma $\|\cdot\|_a$ proviene del producto escalar

$$\langle v, w\rangle_a = a(v,w).$$

La ecuación (4.21) se puede escribir

$$\langle u-u_h, v\rangle_a = 0 \qquad \forall v \in V_h, \tag{4.22}$$

es decir, la solución de elementos finitos u_h es la proyección de la solución exacta u sobre V_h con respecto al producto escalar $\langle\cdot,\cdot\rangle_a$.

Por otra parte, la propiedad (4.22) nos dice que $\langle u-u_h, u_h-v\rangle_a = 0$ para todo $v \in V_h$. Por tanto, el teorema de Pitágoras nos dice que

$$\|u-u_h\|_a^2 + \|u_h-v\|_a^2 = \|u-v\|_a^2,$$

de donde

$$\|u-u_h\|_a \leq \|u-v\|_a \qquad \forall v \in V_h;$$

es decir, u_h es el elemento de V_h más próximo a u en la norma de energía.

El lema de Céa nos da una condición suficiente para la convergencia del método de elementos finitos.

Corolario 4.19. *Sea $\{V_h\}_{h>0}$ una familia de subespacios de V de dimensión finita tales que para todo $u \in V$ se tiene que*

$$\lim_{h\to 0} \inf_{v\in V_h} \|u-u_h\|_V = 0; \tag{4.23}$$

entonces el método de elementos finitos converge, es decir,

$$\lim_{h\to 0} \|u-u_h\|_V = 0.$$

Observación. Para que el método converja basta con que la condición (4.23) se cumpla para la verdadera solución.

Nótese que hemos reducido la cuestión de la convergencia del método de elementos finitos a una cuestión de teoría de aproximación: ¿Cómo de lejos de V_h están las funciones de V?

4.7 Espacios de elementos finitos

Queremos elegir una familia de subespacios $\{V_h\}_{h>0}$ de forma que se verifique (4.23) al menos cuando u es la verdadera solución.

Una elección común es tomar espacios de funciones polinómicas a trozos sobre subdivisiones o "triangulaciones" $T_h = \{K\}$ de un dominio acotado $\Omega \subset \mathbb{R}^d$, $d = 1, 2, 3$, en *elementos* K. Para $d = 1$ los elementos K serán intervalos, para $d = 2$, triángulos o cuadriláteros y para $d = 3$ tetraedros.

Si estamos estudiando un problema de contorno de segundo orden, V_h tendrá que estar contenido en $H^1(\Omega)$. Dado que el espacio V_h está formado por funciones polinómicas a trozos, se tiene que

$$V_h \subset H^1(\Omega) \Leftrightarrow V_h \subset C(\overline{\Omega}).$$

Para definir un espacio V_h del tipo mencionado tendremos que especificar:

- la triangulación $T_h = \{K\}$ del dominio Ω;
- la naturaleza de las funciones v de V_h sobre cada elemento K (lineal, cuadrática, cúbica, etc);
- una base de V_h.

4.7.1 Caso de una variable espacial

Dada una partición

$$a = x_0 < x_1 < \cdots < x_N = b,$$

del intervalo $[0, L]$ denotaremos

$$h_j = x_j - x_{j-1}, \qquad j = 1, \ldots, N, \qquad h = \max_{1 \leq j \leq N} h_j.$$

Nuestros elementos K serán en este caso los intervalos $[x_{i-1}, x_i]$.

Para $r = 1, 2, \ldots$, consideramos los espacios $S_{h,r}$ de funciones continuas en $[a, b]$ que restringidas a cada intervalo $[x_{j-1}, x_j]$ son polinomios de grado menor o igual que r, esto es

$$S_{h,r} = \{v \in C([a, b]) : \; v_{|_{[x_{j-1}, x_j]}} \in \mathbb{P}_r([x_{j-1}, x_j]), \; j = 1, \ldots, N\}.$$

La dimensión de $S_{h,r}$ es fácil de calcular. Una función $\phi \in S_{h,r}$ depende en principio de $(r + 1)N$ parámetros, al haber N intervalos y, en cada uno de ellos ϕ es un polinomio de grado r, que depende de $r + 1$ parámetros. Para mantener la continuidad en cada nodo interno x_j de la partición hay que imponer la condición $\phi(x_j+) = \phi(x_j-)$. Los nodos internos son los $N-1$ nodos $x_1, \ldots, x_{N-1}$. La dimensión de $S_{h,r}$ es por tanto $N(r+1) - (N-1) = rN + 1$.

Todavía tenemos que indicar qué parámetros vamos a utilizar para describir las funciones de V_h. Eso es equivalente a escoger una base en V_h. Veamos cómo hacerlo en el caso de elementos lineales ($r = 1$). En este caso consideramos los nodos $x_0, \ldots, x_N$ y consideramos las llamadas *funciones base nodales*:

$$\varphi_n(x) = \begin{cases} \dfrac{x - x_{n-1}}{x_n - x_{n-1}}, & x \in [x_{n-1}, x_n], \\[2ex] \dfrac{x_{n+1} - x}{x_{n+1} - x_n}, & x \in [x_n, x_{n+1}], \qquad n = 1, \ldots, N-1, \\[2ex] 0, & \text{en el resto.} \end{cases}$$

$$\varphi_0(x) = \begin{cases} \dfrac{x_1 - x}{x_1 - x_0}, & x \in [x_0, x_1], \\ 0, & \text{en el resto.} \end{cases}$$

$$\varphi_N(x) = \begin{cases} \dfrac{x - x_{N-1}}{x_N - x_{N-1}}, & x \in [x_{N-1}, x_N], \\ 0, & \text{en el resto.} \end{cases}$$

Obviamente cualquier función $v_h \in S_{h,1}$ se puede escribir como

$$v_h = \sum_{n=0}^{N} \gamma_n \phi_n.$$

Dado que $\phi_i(x_j) = \delta_{ij}$ tenemos que

$$\gamma_n = v_h(x_n).$$

Así pues, los $N+1$ parámetros mediante los cuales se describe la función v_h son sus valores en los $N+1$ nodos; de ahí el nombre de base nodal.

En el caso en que $r > 1$ el número de grados de libertad es mayor que el de nodos. Lo que se hace es añadir nodos intermedios (denominados nodos fraccionarios), entre los de la partición, dados por

$$x_{i+j/r} = x_i + \frac{j}{r} h_i, \qquad i = 0, \ldots, N-1, \qquad j = 1, \ldots, r-1.$$

De esta forma se tienen tantos nodos como grados de libertad $(rN+1)$, y se pueden elegir como parámetros para describir cualquier función de $S_{h,r}$ sus valores en los nodos. Esto definirá una base, la base nodal, caracterizada porque las funciones de la base toman el valor 1 en el nodo (fraccionario o entero) al que están asociadas y el valor 0 en el resto de los nodos.

Estudiemos la convergencia del método. Nos limitaremos al caso $r = 1$; es decir, $V_h = S_{h,1}$.

Observamos que $\inf\limits_{v \in V_h} \|u - v\|_V \leq \|u - u_I\|_V$, donde $u_I \in V_h$ es el interpolante lineal a trozos de u en los nodos $x_0, \ldots, x_N$ de la partición. Se tiene el siguiente resultado de aproximación.

Lema 4.20. *Si $u \in C^2([a,b])$, entonces*

$$\max_{[x_{j-1}, x_j]} |u - u_I| \leq \frac{1}{8} h_j^2 \max_{[x_{j-1}, x_j]} |u''|, \qquad \max_{[x_{j-1}, x_j]} |u' - u_I'| \leq h_j \max_{[x_{j-1}, x_j]} |u''|,$$

y por tanto

$$\|u - u_I\|_{H^1(I)} \le Ch \max_{[a,b]} |u''|.$$

Prueba. Consideramos la diferencia $\Delta(x) = u(x) - u_I(x)$ sobre el intervalo $[x_{j-1}, x_j]$. Dado que Δ se anula en ambos extremos del intervalo, debe existir al menos un punto z en el que $\Delta'(z) = 0$. Entonces para todo x,

$$\Delta'(x) = \int_z^x \Delta''(y)\, dy.$$

Ahora bien, como u_I es lineal, $\Delta'' = u''$, y se tiene que

$$|\Delta'(x)| = \left| \int_z^x u''(y)\, dy \right| \le h_j \max_{[x_{j-1},x_j]} |u''|, \qquad x \in [x_{j-1}, x_j],$$

que es la segunda de las estimaciones.

El máximo de $|\Delta(x)|$ en $[x_{j-1}, x_j]$, ocurrirá en un punto en el que la derivada se anula, $\Delta'(z) = 0$. Miramos en qué mitad del intervalo está z; supongamos, por ejemplo, que está más próximo al extremo derecho, de manera que $x_j - z \le h_j/2$. Desarrollando en serie de Taylor alrededor de z,

$$\Delta(x_j) = \Delta(z) + (x_j - z)\Delta'(z) + \frac{1}{2}(x_j - z)^2 \Delta''(w),$$

donde $z < w < x_j$. Usando que Δ se anula en el extremo x_j y que $\Delta'' = u''$,

$$|\Delta(z)| = \left| \frac{1}{2}(x_j - z)^2 \Delta''(w) \right| \le \frac{1}{8} h_j^2 \max_{[x_{j-1},x_j]} |u''|.$$

Combinando las dos primeras estimaciones se obtiene inmediatamente la estimación del error de interpolación en norma H^1

$$\begin{aligned}
\|u - u_I\|_{H^1(I)} &= \left(\sum_{j=1}^{N} \int_{x_{j-1}}^{x_j} [(u - u_I)^2 + (u' - u_I')^2] \right)^{1/2} \\
&\le \left[\left(\tfrac{h^4(b-a)}{64} + h^2(b-a) \right) (\max_{[a,b]} |u''|)^2 \right]^{1/2} \le Ch \max_{[a,b]} |u''|. \quad \square
\end{aligned}$$

El lema garantiza la convergencia de la solución de elementos finitos a la verdadera solución u si ésta es tal que $u \in C^2([a,b])$.

Ejemplo. Sea u la solución (clásica) del problema (4.1) con dato $f \in C([a,b])$. Por ser $u'' = u - f$, y ser $u \in C([a,b])$ tenemos a posteriori que $u'' \in C([a,b])$. Por tanto tenemos convergencia de la solución de elementos finitos a la solución clásica.

El resultado del lema 4.20 no es en realidad el más adecuado para nuestros fines. El factor h es correcto, y refleja la tasa observada de decrecimiento del error cuando se refina la partición. La imperfección está en el otro factor, $\max |u''|$. Las soluciones débiles u de problemas elípticos de segundo orden no verifican en general que $u \in C^2([a,b])$, sino que $u \in H^2(I)$. La estimación idónea para el error de interpolación se da a continuación.

Lema 4.21. *Si $u \in H^2(I)$, entonces*

$$\|u - u_I\|_{L^2(I)} \leq Ch^2 \|u''\|_{L^2(I)}, \qquad \|u' - u_I'\|_{L^2(I)} \leq Ch \|u''\|_{L^2(I)},$$

y por tanto

$$\|u - u_I\|_{H^1(I)} \leq Ch \|u''\|_{L^2(I)}.$$

Se puede encontrar una prueba, basada en desarrollos en serie de Fourier, en [SF].

Combinando este resultado con el lema de Céa tenemos el siguiente resultado de convergencia.

Corolario 4.22. *Sea $I = (a,b)$, $u \in V \subset H^1(I)$ solución de (D) y $u_h \in V_h$ solución de (D_h) con $V_h \subset V$. Si $u \in H^2(I)$, entonces*

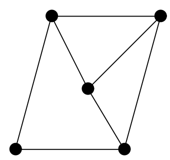
$$\|u - u_h\|_{H^1(I)} \leq Ch \|u''\|_{L^2(I)}.$$

4.7.2 Caso de dos variables espaciales

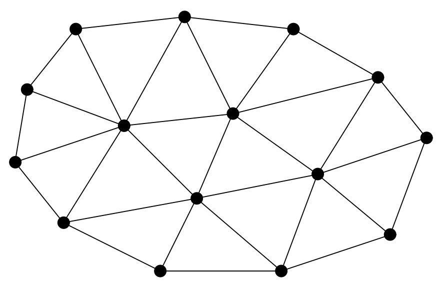
Para simplificar supondremos que $\partial\Omega$ es poligonal, en cuyo caso diremos que Ω es un dominio poligonal (si el dominio Ω no es poligonal, podemos aproximarlo por un dominio que sí lo sea). Hacemos una triangulación del dominio Ω, subdividiéndolo en un conjunto $T_h = \{K_j\}_{j=1}^{n_t}$ de triángulos K_j que satisfacen

$$\overline{\Omega} = \bigcup_{k=1}^{n_t} K_j.$$

siendo la intersección $K_j \cap K_l$, $j \neq l$, o bien vacía, o bien un lado común, o bien un vértice común. Así pues, la única restricción es que un vértice de un triángulo no puede estar en mitad de una arista de otro; en otras palabras, una configuración como



(donde la posición de los vértices se enfatiza con un '$\bullet$') no está permitida. A continuación se muestra una triangulación que se ajusta a las reglas.



Introducimos el *parámetro de malla*

$$h = \max_{K_j \in T_h} \operatorname{diam}(K_j),$$

donde $\operatorname{diam}(K_j)$ = diámetro de K_j = longitud del mayor lado de K_j.

Para $r = 1, 2, \ldots$, consideramos los espacios $S_{h,r}$ de funciones continuas en $\overline{\Omega}$ que restringidas a cada triángulo K_j son polinomios de grado menor o igual que r, esto es

$$S_{h,r} = \{v \in C(\overline{\Omega}) :\ v_{|_{K_j}} \in \mathbb{P}_r(K_j),\ j = 1, \ldots, n_t\}.$$

Como en el caso unidimensional, queremos caracterizar las funciones de $S_{h,r}$ por sus valores en un conjunto de nodos

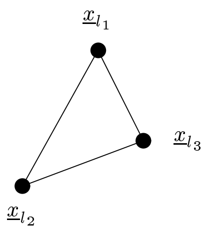
$$\underline{x}_1, \ldots, \underline{x}_M.$$

Veamos como hacerlo en el caso de funciones lineales ($r = 1$). En este caso los nodos son los vértices de la triangulación.

Empezamos por probar que las funciones de $S_{h,1}$, y en particular las funciones base, quedan completamente determinadas por sus valores nodales. Sobre el triángulo K_l las funciones $v \in S_{h,1}$ son de la forma

$$v_{|_{K_l}}(x, y) = a_l + b_l x + c_l y.$$

Supongamos que el triángulo K_l en cuestión es



de nodos $\underline{x}_{l_1} = (x_{l_1}, y_{l_1})$, $\underline{x}_{l_2} = (x_{l_2}, y_{l_2})$ y $\underline{x}_{l_3} = (x_{l_3}, y_{l_3})$. La función v toma valores v_{l_1}, v_{l_2}, v_{l_3} en los nodos $\underline{x}_{l_1}$, $\underline{x}_{l_2}$, $\underline{x}_{l_3}$, lo que da lugar al sistema lineal

$$a_l + b_l x_{l_i} + c_l y_{l_i} = v_{l_i}, \qquad i = 1, 2, 3.$$

Como los vértices no son colineales, el sistema es no singular y se puede resolver fácilmente, produciendo a_l, b_l y c_l.

Para que la función v resultante este bien definida necesitamos que las restricciones de v a triángulos K_i, K_j con una arista común coincidan sobre la arista. Dichas restricciones son funciones lineales en un segmento, y están unívocamente determinadas por los valores en los extremos del segmento. Como los extremos del segmento son nodos, los valores de ambas restricciones coinciden en ellos (son valores nodales, que están prescritos); se concluye que las dos restricciones coinciden a lo largo de la arista.

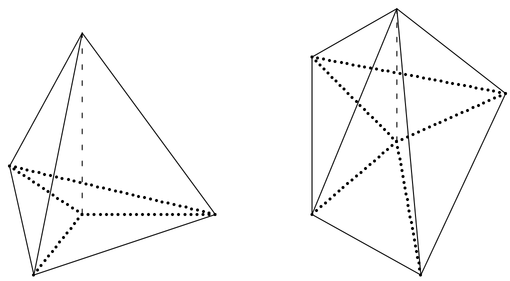
La base conveniente para describir las funciones de $S_{h,1}$ por sus valores nodales es la base nodal; las funciones de dicha base, $\varphi_1, \ldots, \varphi_M \in S_{h,1}$, se definen por la propiedad

$$\varphi_i(\underline{x}_j) = \delta_{ij}.$$

De esta forma cualquier $v \in S_{h,l}$ se expresa como

$$v = \sum_{j=1}^{M} v(\underline{x}_j)\varphi_j.$$

El soporte de la función base asociada a un nodo es la unión de los triángulos de los que dicho nodo es vértice. En efecto, para el resto de los triángulos toma el valor cero en los nodos, y es por tanto nula. A continuación mostramos tres ejemplos de funciones nodales, dibujadas sobre su soporte.



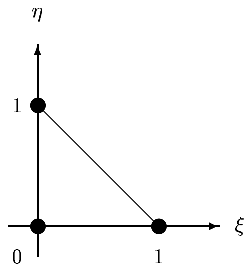
Por motivos evidentes, las funciones base reciben también el nombre de funciones pirámide. El número de elementos en cada soporte puede cambiar de vértice a vértice, y la descripción de las funciones base nodales, aunque sencilla en teoría, acaba convirtiéndose en algo engorroso e inconveniente desde el punto de vista práctico.

Sin embargo, restringidas a cada triángulo las funciones base tienen una expresión sencilla. Como veremos más adelante, esto es todo lo que se necesita en la práctica.

Las únicas funciones de la base nodal que no son idénticamente nulas sobre un triángulo dado son las correspondientes a los nodos del triángulo. Estas funciones base se expresan en términos de las funciones base para un triángulo muy especial, el llamado *elemento de referencia* K_0. Éste es el triángulo de vértices

$$\underline{\xi}_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \qquad \underline{\xi}_2 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \qquad \underline{\xi}_3 = \begin{bmatrix} 0 \\ 0 \end{bmatrix},$$

que representamos en la figura en el plano ξ, η.



Para cada $i = 1, 2, 3$ denotamos por N_i a la función base del triángulo de referencia asociada al nodo $\underline{\xi}_i$, y N a la aplicación

$$N(\xi, \eta) = \begin{bmatrix} N_1(\xi, \eta) \\ N_2(\xi, \eta) \\ N_3(\xi, \eta) \end{bmatrix}.$$

Las funciones N_i son muy fáciles de calcular. Se tiene que

$$N(\xi, \eta) = \begin{bmatrix} 1 - \xi - \eta \\ \xi \\ \eta \end{bmatrix}.$$

El triángulo K, de vértices

$$\underline{x}_1 = \begin{bmatrix} x_1 \\ y_1 \end{bmatrix}, \qquad \underline{x}_2 = \begin{bmatrix} x_2 \\ y_2 \end{bmatrix}, \qquad \underline{x}_3 = \begin{bmatrix} x_3 \\ y_3 \end{bmatrix},$$

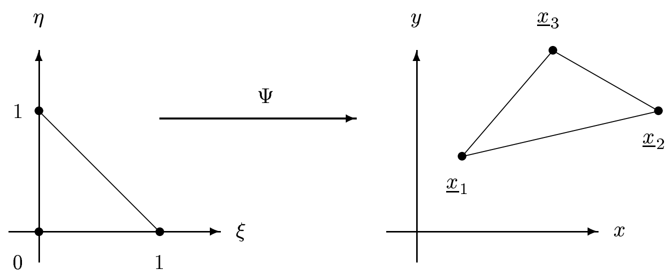
se obtiene a partir del triángulo de referencia K_0 mediante la transformación afín Ψ dada por

$$\underline{\xi} \mapsto \underline{x} = \Psi(\underline{\xi}) = \underline{x}_1 + [\underline{x}_2 - \underline{x}_1, \underline{x}_3 - \underline{x}_1]\underline{\xi},$$

donde, obviamente, $[\underline{x}_2 - \underline{x}_1, \underline{x}_3 - \underline{x}_1]$ es la matriz cuyas columnas son los vectores $\underline{x}_2 - \underline{x}_1$ y $\underline{x}_3 - \underline{x}_1$, aplicación que expresada en función de sus componentes se escribe como

$$\underline{\xi} = \begin{bmatrix} \xi \\ \eta \end{bmatrix} \mapsto \underline{x} = \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} x_1 \\ y_1 \end{bmatrix} + \begin{bmatrix} x_2 - x_1 & x_3 - x_1 \\ y_2 - y_1 & y_3 - y_1 \end{bmatrix} \begin{bmatrix} \xi \\ \eta \end{bmatrix}.$$

Gráficamente,



La aplicación

$$\varphi(x,y) = \begin{bmatrix} \varphi_1(x,y) \\ \varphi_2(x,y) \\ \varphi_3(x,y) \end{bmatrix}$$

de las funciones base correspondientes a nodos del triángulo K se expresa entonces como

$$\varphi = N \circ \Psi^{-1},$$

cuando nos restringimos a dicho triángulo.

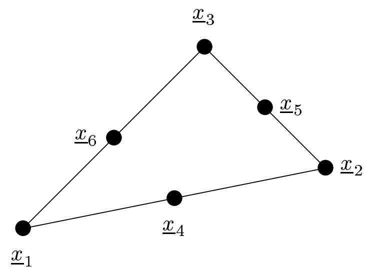
Recordando que $N_i(\underline{\xi}_j) = \delta_{ij}$, se tiene que Ψ se puede expresar también en términos de las funciones N_i, mediante

$$\Psi(\underline{\xi}) = \underline{x}_1 N_1(\underline{\xi}) + \underline{x}_2 N_2(\underline{\xi}) + \underline{x}_3 N_3(\underline{\xi}).$$

Sea ahora $r = 2$; las funciones $v \in S_{h,2}$ son cuadráticas sobre cada triángulo K,

$$v_{|T}(x,y) = a + bx + cy + dx^2 + exy + fy^2.$$

Hay por tanto seis parámetros; para que una función quede completamente determinada sobre el triángulo habrá que especificar su valor en seis puntos. Además de los vértices habrá que considerar otros tres nodos. Se suelen tomar los puntos medios:



La numeración estándar de los nodos es la de la figura: primero los tres vértices en sentido positivo, y luego, en el mismo sentido, los tres puntos medios, empezando por el del primer lado (el de los vértices primero y segundo).

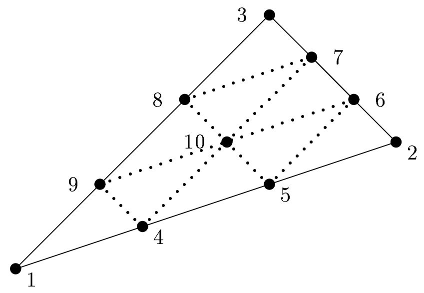
En este caso habrá seis funciones base nodales en el elemento de referencia, N_i, $i = 1, \ldots, 6$. La aplicación

$$\varphi(x, y) = \begin{bmatrix} \varphi_1(x, y) \\ \vdots \\ \varphi_6(x, y) \end{bmatrix}$$

de las funciones base correspondientes a nodos del triángulo K se expresa también en este caso como $\varphi = N \circ \Psi^{-1}$, cuando nos restringimos a dicho triángulo, donde

$$N(\xi, \eta) = \begin{bmatrix} N_1(\xi, \eta) \\ \vdots \\ N_6(\xi, \eta) \end{bmatrix}.$$

Al ir tomando r más grande hará falta tomar cada vez más nodos. Por ejemplo, si $r = 3$ hay diez parámetros en cada triángulo, y habrá que especificar los valores de la función en diez nodos; se toman los de la figura:



La numeración estándar es la de la figura. Obsérvese que cada arista queda dividida en tres partes iguales, y que el baricentro del triángulo también es un nodo.

Para finalizar esta sección estudiamos la convergencia en el caso $r = 1$. Para $K \in T_h$ definimos

$$\begin{aligned} h_K &= \text{diámetro de } K, \\ \rho_K &= \text{diámetro de la circunferencia inscrita en } K, \\ h &= \max_{K \in T_h} h_K. \end{aligned}$$

En realidad no vamos a usar una triangulación, sino una familia de triangulaciones $\{T_h\}$, caracterizadas por el parámetro h. Supondremos que existe una constante positiva β independiente de la triangulación $T_h \in \{T_h\}$, esto es, independiente de h, tal que

$$\frac{\rho_K}{h_K} \geq \beta \qquad \forall K \in T_h.$$

Esta condición significa que no se permite que los ángulos de los triángulos se hagan arbitrariamente pequeños; la constante β es una medida del ángulo más pequeño en cualquier $K \in T_h$ para cualquier $T_h \in \{T_h\}$.

Sean $\underline{x}_i$, $i = 1, \ldots, M$, los nodos de T_h. Dada una función $u \in C(\overline{\Omega})$, definimos el interpolante lineal a trozos u_I como la única función de $S_{h,1}$ tal que

$$u(\underline{x}_i) = u_I(\underline{x}_i) \qquad \forall i = 1, \ldots, M.$$

Sea

$$H^2(\Omega) = \{u \in H^1(\Omega) : \frac{\partial u}{\partial x_i} \in H^1(\Omega),\ i = 1, \ldots, d\}.$$

Se tiene la siguiente estimación para el error de interpolación.

Lema 4.23. *Si $u \in H^2(\Omega)$, entonces*

$$\|u - u_I\|_{L^2(\Omega)} \le Ch^2 \left(\sum_{i,j=1}^{d} \int_\Omega \left| \frac{\partial^2 u}{\partial x_i \partial x_j} \right|^2 \right)^{1/2},$$

$$\|u' - u_I'\|_{L^2(\Omega)} \le \frac{Ch}{\beta} \left(\sum_{i,j=1}^{d} \int_\Omega \left| \frac{\partial^2 u}{\partial x_i \partial x_j} \right|^2 \right)^{1/2},$$

y por tanto

$$\|u - u_I\|_{H^1(\Omega)} \le h \left(C_1 h^2 + \frac{C_2}{\beta^2} \right)^{1/2} \left(\sum_{i,j=1}^{d} \int_\Omega \left| \frac{\partial^2 u}{\partial x_i \partial x_j} \right|^2 \right)^{1/2}.$$

Se puede encontrar una prueba de este lema (con hipótesis adicionales sobre u) en [J].

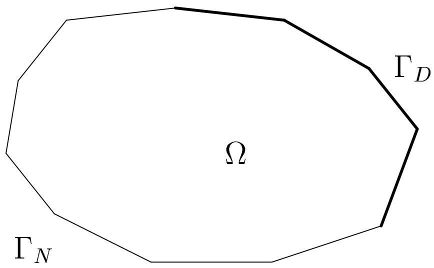
Observación. Los experimentos numéricos muestran que la estimación da la dependencia de β correcta. Esto tiene consecuencias prácticas importantes. Cuanto menor sea β, peor será la aproximación. Habrá que intentar, por tanto, que la triangulación no contenga triángulos muy "delgados".

Gracias al lema se tiene convergencia de las soluciones del método de elementos finitos a la verdadera solución, siempre y cuando ésta tenga suficiente regularidad; concretamente se necesita que $u \in H^2(\Omega)$. Esta regularidad se tiene para muchos problemas elípticos si el dominio Ω es convexo.

4.8 Cálculos con elementos finitos. Programación

En esta sección concretaremos como se usan todos los conceptos anteriores a la hora de hacer cálculos con el ordenador. Nos limitaremos al caso de dos dimensiones y funciones lineales sobre triángulos. Ilustraremos las distintas ideas que vayamos introduciendo por medio de un problema modelo. Entre otras cosas indicaremos instrucciones (o conjuntos de instrucciones) de MatLab que realicen las diversas tareas que vayamos explicando. No intentaremos hacer el mejor programa posible, sino uno que permita ver claramente cuáles son las ideas del método de elementos finitos. Concluiremos la sección explicando sucintamente qué hay que hacer para tratar otros problemas.

Sea Ω el dominio de la figura,



La frontera de Ω está descompuesta en dos partes disjuntas, Γ_D y Γ_N. La primera de ellas aparece marcada en la figura por medio de un trazo más grueso.

Nuestro problema modelo es el siguiente:

Dados $f \in L^2(\Omega)$ y $g \in L^2(\Gamma_N)$, encontrar una aproximación numérica de la solución débil del problema

$$\begin{cases} -\Delta u + \alpha u = f & \text{en } \Omega, \\ u = 0, & \text{en } \Gamma_D, \\ \frac{\partial u}{\partial n} = g & \text{en } \Gamma_N, \end{cases} \tag{4.24}$$

donde $\alpha \geq 0$ es un parámetro real. Si $\alpha = 0$ recuperamos la ecuación de Poisson, que es el caso que trataremos en los experimentos numéricos.

Según hemos visto, una solución débil de (4.24) es una función u de

$$V = \{v \in H^1(\Omega) : \ v = 0 \text{ c.t.p. en } \Gamma_D\}$$

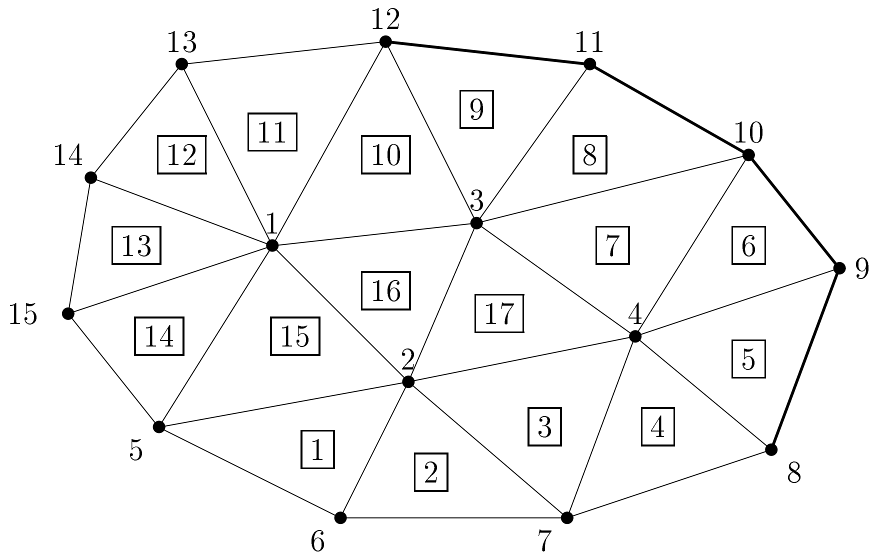
tal que

$$a(u, v) = L(v) \qquad \forall v \in V,$$

donde

$$a(u,v) = \int_\Omega (\nabla u \cdot \nabla v + \alpha u v), \qquad L(v) = \int_\Omega f v + \int_{\Gamma_N} g v.$$

Para resolver el problema de forma aproximada lo primero que tenemos que hacer es triangular el dominio. Veamos una posible triangulación.



En esta figura hemos numerado los nodos y los triángulos; los índices de los triángulos están recuadrados.

Hay un total de $M = 15$ nodos,

$$\underline{x}_1, \dots, \underline{x}_M,$$

y $n_t = 17$ triángulos,

$$K_1, \cdots, K_{n_t}.$$

La base nodal consta así de M funciones, φ_i, $i = 1, \dots, M$, que verifican que

$$\varphi_i(\underline{x}_j) = \delta_{ij}.$$

La aproximación de elementos finitos es una función $u_h \in V_h \subset V$,

$$V_h = \{v \in C(\overline{\Omega}) : v(\underline{x}) = 0 \text{ si } \underline{x} \in \Gamma_D,\, v_{|_{K_j}} \in \mathbb{P}_1(K_j),\ j = 1, \dots, n_t\},$$

tal que

$$a(u_h, v) = L(v) \qquad \forall v \in V_h.$$

Cualquier $v \in V_h$ se puede representar en la forma

$$v = \sum_{j=1}^{M} v(\underline{x}_j)\varphi_j.$$

Ahora bien, si $v \in V_h$, entonces $v(\underline{x}_j) = 0$ para $\underline{x}_j \in \Gamma_D$, y por tanto

$$v = \sum_{j \notin I_D} v(\underline{x}_j)\varphi_j, \qquad I_D = \{j : \underline{x}_j \in \Gamma_D\}.$$

Así que podemos tomar como base de V_h las funciones φ_i, $i = 1, \ldots, M$ con $i \notin I_D$.

Denotando $u_j = u_h(\underline{x}_j)$ tenemos que la solución de elementos finitos viene dada por la solución del sistema lineal

$$\sum_{j \notin I_D} a(\varphi_i, \varphi_j) u_j = L(\varphi_i), \qquad i \notin I_D. \tag{4.25}$$

Las incógnitas son los valores nodales u_j con $j \notin I_D$.

Deberíamos por tanto construir la matriz de rigidez de elementos

$$a(\varphi_i, \varphi_j), \qquad i, j \notin I_D.$$

Sin embargo, como veremos más adelante, la matriz que se puede construir de forma sistemática y sencilla es la matriz "completa",

$$A = (a_{ij}), \qquad i, j = 1, \ldots, M, \qquad a_{ij} = a(\varphi_i, \varphi_j).$$

Así pues, lo razonable será construir la matriz completa A y extraer de ella la matriz de coeficientes del sistema (4.25) eliminando las filas y columnas de A correspondientes a índices de I_D. Ahora bien, para el ordenador esta extracción es extraordinariamente costosa, pues obliga a trasvasar datos de un sector de memoria a otro. Por ello, lo que se hace en la práctica es completar las ecuaciones (4.25) añadiendo las ecuaciones

$$u_j = 0, \qquad j \in I_D. \tag{4.26}$$

Denotemos por D_0 a la matriz diagonal $M \times M$ cuyas componentes diagonales d_{ii} son 1 si $i \notin I_D$ y 0 si $i \in I_D$. Entonces para un vector $v \in \mathbb{R}^M$, el producto $D_0 v$ es simplemente hacer 0 las componentes de v correspondientes a índices de I_D.

Sea $\underline{u}_h = (u_1, \ldots, u_M)^T$. El sistema de ecuaciones (4.25)–(4.26) se escribe como

$$A_0 \underline{u}_h = D_0 b, \tag{4.27}$$

donde $A_0 = D_0 A D_0 + (I - D_0)$, e I es la matriz identidad de $\mathbb{R}^M$, esto es, A_0 es la matriz que se obtiene al hacer cero las filas y las columnas de A correspondientes a índices de I_D y a continuación dando valor 1 a los elementos de la diagonal principal que correspondan a columnas cuyos índices estén en I_D.

Teniendo en cuenta todo esto, descompondremos la programación en las siguientes tareas:

- Suministrar los datos al programa, en particular la triangulación.
- Calcular la matriz de rigidez.
- Calcular el vector de carga.
- Construir el sistema reducido (4.27) a partir del sistema completo, para tomar en cuenta las condiciones de Dirichlet.
- Resolver el sistema resultante.

4.8.1 Datos de la triangulación

¿Cómo se almacena la triangulación en el ordenador? La idea es muy simple. Se utilizan dos matrices. La primera de ellas, llamémosla `T`, es una matriz de elementos *enteros* que tiene una columna por cada triángulo, y en la columna j-ésima tres componentes que guardan los índices de los vértices del triángulo j-ésimo.

En la triangulación de nuestro ejemplo vemos que los índices de los vértices del primer triángulo son 2, 5 y 6, luego las tres componentes de la primera columna de `T` serán

$$\begin{bmatrix} 2 \\ 5 \\ 6 \end{bmatrix}.$$

Para el segundo triángulo los índices de sus vértices son 2, 6 y 7, luego las tres componentes de las dos primeras columnas de `T` serán

$$\begin{bmatrix} 2 & 2 \\ 5 & 6 \\ 6 & 7 \end{bmatrix}.$$

Siguiendo con este proceso llegamos a que la matriz `T` de la triangulación de nuestro ejemplo es

$$\begin{bmatrix} 2 & 2 & 2 & 4 & 4 & 4 & 4 & 3 & 3 & 3 & 1 & 1 & 1 & 1 & 1 & 1 & 3 \\ 5 & 6 & 7 & 7 & 8 & 9 & 10 & 10 & 11 & 12 & 12 & 13 & 14 & 15 & 5 & 2 & 2 \\ 6 & 7 & 4 & 8 & 9 & 10 & 3 & 11 & 12 & 1 & 13 & 14 & 15 & 5 & 2 & 3 & 4 \end{bmatrix}.$$

Hay varios aspectos que señalar. En cada triángulo hemos colocado los índices ordenados en sentido positivo (antihorario); esta ordenación no tiene por qué coincidir con el orden creciente de los nodos. Por ejemplo, la tercera columna de `T` es

$$\begin{bmatrix} 2 \\ 7 \\ 4 \end{bmatrix},$$

en vez de $[2, 4, 7]^T$. Obsérvese también que para los triángulos que tienen un lado en la frontera, éste es siempre el que corresponde al lado cuyos extremos son el segundo y el tercer vértice. Por otra parte, no hay triángulos con dos lados de frontera; así el primer vértice de cada triángulo siempre es interior a Ω.

No todas estas convenciones son habituales. Sin embargo, si se mantienen se pueden utilizar provechosamente para simplificar la programación.

La segunda matriz necesaria para describir la triangulación es una matriz que denotamos por `Z`, que tiene una columna por cada nodo, y que en la columna j-ésima guarda las coordenadas x, y del nodo j-ésimo. En el ejemplo de la figura la matriz `Z` sería

$$\mathtt{Z} = \begin{bmatrix} x_1 & x_2 & \cdots & x_{15} \\ y_1 & y_2 & \cdots & y_{15} \end{bmatrix}.$$

Con estas dos matrices queda perfectamente descrita la triangulación, al menos desde el punto de vista de los cálculos.

También tenemos que almacenar de alguna forma los datos de frontera, en nuestro caso los lados en los que se tienen datos de tipo Dirichlet (los lados que pertenecen a Γ_D) y los lados en los que se tienen datos de tipo Neumann (los que pertenecen a Γ_N). Esta información se guarda en dos vectores, que llamaremos `ID` e `IN`.

El vector `ID` (resp. `IN`) contiene los índices de los triángulos que tienen un lado en Γ_D (resp. Γ_N).

En nuestro ejemplo las matrices serían

$$\mathtt{ID} = \begin{bmatrix} 5 & 6 & 8 & 9 \end{bmatrix}, \qquad \mathtt{IN} = \begin{bmatrix} 1 & 2 & 4 & 11 & 12 & 13 & 14 \end{bmatrix}.$$

Conociendo el triángulo al que pertenece una arista, se conocen sus nodos de frontera gracias al convenio que hemos seguido: si el triángulo es el `j`-ésimo, los nodos de frontera serán `T(j,2)` y `T(j,3)`.

Una vez construidas las matrices se le pasan al programa como argumentos.

MatLab. La primera línea de nuestro programa de MatLab podría ser

```
function u=mef(T,Z,ID,IN,alpha,fsm,dn)
```

Los argumentos `alpha`, `fsm` y `dn` se utilizan para pasar al programa el valor del parámetro α y los nombres de la función f del segundo miembro de la ecuación y de la función g del dato de Neumann.

El programa, que se almacenará en un fichero con extensión `.m`, devolverá como resultado la matriz columna `u`, que contendrá los valores nodales de la solución de elementos finitos.

4.8.2 Construcción de la matriz de rigidez

Veamos ahora cómo construir la matriz de rigidez A, de elementos

$$a_{ij} = a(\varphi_i, \varphi_j), \qquad i = 1, \dots, M.$$

Como vimos, las funciones de la base nodal, φ_i, admiten una expresión sencilla cuando nos restringimos a cada triángulo. Esto nos lleva a nuestra primera idea básica.

• Calcular la matriz de rigidez como suma de contribuciones de cada triángulo.

Para ello expresamos la forma bilineal a como suma de formas bilineales restringidas a los triángulos:

$$a = \sum_{j=1}^{n_t} a_{K_j},$$

donde a_{K_j} es la forma bilineal a restringida al triángulo K_j. En nuestro ejemplo,

$$a(u,v) = \int_\Omega (\nabla u \cdot \nabla v + \alpha uv) = \sum_{j=1}^{n_t} \int_{K_j} (\nabla u \cdot \nabla v + \alpha uv),$$

de manera que

$$a_{K_j}(u,v) = \int_{K_j} (\nabla u \cdot \nabla v + \alpha uv),$$

En términos de la matriz de rigidez la descomposición se traduce en

$$A = \sum_{j=1}^{n_t} A_{K_j}$$

donde las matrices A_{K_j} son las matrices de las formas bilineales a_{K_j} en la base $\varphi_1, \ldots, \varphi_M$, y son por tanto matrices $M \times M$. Calcularemos las matrices A_{K_j} para luego *ensamblarlas* para formar la matriz A.

Sobre cada triángulo K_j (de nodos $\underline{x}_{j_1}$, $\underline{x}_{j_2}$, $\underline{x}_{j_3}$) sólo hay 3 funciones base no nulas, las correspondientes a los nodos del triángulo (φ_{j_1}, φ_{j_2}, φ_{j_3}). Por tanto, la matriz A_{K_j} tiene a lo sumo 3^2 elementos no nulos, que son los que se encuentran en la intersección de las filas y columnas cuyos índices corresponden a los nodos del triángulo.

Los 3^2 elementos de la matriz A_{K_j} que pueden ser no nulos conforman una matriz E_j dada por

$$E_j = \begin{bmatrix} a_{K_j}(\varphi_{j_1}, \varphi_{j_1}) & a_{K_j}(\varphi_{j_1}, \varphi_{j_2}) & a_{K_j}(\varphi_{j_1}, \varphi_{j_3}) \\ a_{K_j}(\varphi_{j_2}, \varphi_{j_1}) & a_{K_j}(\varphi_{j_2}, \varphi_{j_2}) & a_{K_j}(\varphi_{j_2}, \varphi_{j_3}) \\ a_{K_j}(\varphi_{j_3}, \varphi_{j_1}) & a_{K_j}(\varphi_{j_3}, \varphi_{j_2}) & a_{K_j}(\varphi_{j_3}, \varphi_{j_3}) \end{bmatrix}.$$

Obsérvese que E_j es la matriz de la forma bilineal a_{K_j} restringida al espacio generado por las funciones base que son no nulas en el triángulo K_j, y expresada en esa base.

Las matriz E_j (de dimensión 3×3) es la *matriz elemental* del triángulo (elemento) K_j; la correspondiente matriz A_{K_j} (de dimensión $M \times M$) es la *contribución de la matriz elemental* E_j a la matriz A.

En nuestro ejemplo los únicos elementos que pueden ser no nulos de la matriz A_{K_1} son los que ocupan la posición (i, j) para $i, j = 2, 5, 6$, y la matriz E_1 correspondiente viene dada por

$$E_1 = \begin{bmatrix} a_{K_1}(\varphi_2, \varphi_2) & a_{K_1}(\varphi_2, \varphi_5) & a_{K_1}(\varphi_2, \varphi_6) \\ a_{K_1}(\varphi_5, \varphi_2) & a_{K_1}(\varphi_5, \varphi_5) & a_{K_1}(\varphi_5, \varphi_6) \\ a_{K_1}(\varphi_6, \varphi_2) & a_{K_1}(\varphi_6, \varphi_5) & a_{K_1}(\varphi_6, \varphi_6) \end{bmatrix}.$$

Observación. Los elementos de la matriz E_j no tienen por qué coincidir con los correspondientes elementos de la matriz A; si un lado $[\underline{x}_i, \underline{x}_j]$ corresponde a dos triángulos, el elemento de A correspondiente, $a(\varphi_i, \varphi_j)$, recibirá contribuciones de las matrices E_j de ambos. Así, en nuestro ejemplo,

$$a_{25} = a(\varphi_2, \varphi_5) = \sum_{j=1}^{N} a_{K_j}(\varphi_2, \varphi_5) = a_{K_1}(\varphi_2, \varphi_5) + a_{K_{15}}(\varphi_2, \varphi_5).$$

Análogamente, si un nodo $\underline{x}_i$ pertenece a varios triángulos, el elemento correspondiente $a(\varphi_i, \varphi_i)$ tiene contribuciones de todos ellos.

Observación. La matriz E_j no se calcula necesariamente con la misma ordenación con que se suma a A. En nuestro ejemplo, en el caso del tercer triángulo los programas de elementos finitos calcularían la matriz

$$\begin{bmatrix} a_{K_3}(\varphi_2, \varphi_2) & a_{K_1}(\varphi_2, \varphi_7) & a_{K_1}(\varphi_2, \varphi_4) \\ a_{K_3}(\varphi_7, \varphi_2) & a_{K_1}(\varphi_7, \varphi_7) & a_{K_1}(\varphi_7, \varphi_4) \\ a_{K_3}(\varphi_4, \varphi_2) & a_{K_1}(\varphi_4, \varphi_7) & a_{K_1}(\varphi_4, \varphi_4) \end{bmatrix},$$

pero, a la hora de sumar estos elementos a los de A se reordenan para obtener la matriz E_3. Afortunadamente, usando MatLab no nos tendremos que preocupar por estos detalles.

Veamos cómo se hace el cálculo de E_j en nuestro ejemplo. Para no recargar la notación, abandonamos la j. Los elementos de la matriz elemental E correspondiente al triángulo K son

$$a_K(\varphi_i, \varphi_k) = \int_K (\varphi_{i,x}\varphi_{k,x} + \varphi_{i,y}\varphi_{k,y} + \alpha\varphi_i\varphi_k), \qquad i, k = 1, 2, 3.$$

En principio podríamos calcular cada uno de estos nueve elementos por separado. Sin embargo:

- El cálculo de la matriz E no se efectúa componente a componente, sino que se calcula toda la matriz a la vez.

Con el fin de expresar E en una forma adecuada para este propósito definimos

$$\varphi = \begin{bmatrix} \varphi_1 \\ \varphi_2 \\ \varphi_3 \end{bmatrix},$$

donde φ_i, $i = 1, 2, 3$, es la función de la base nodal asociada al vértice $\underline{x}_i$, $i = 1, 2, 3$, del triángulo K. Esto nos permite escribir

$$E = \int_K (\varphi_x \varphi_x^T + \varphi_y \varphi_y^T + \alpha \varphi \varphi^T).$$

Como hemos visto, las funciones base restringidas al triángulo K se pueden expresar de forma muy sencilla en términos de las funciones de base en el elemento de referencia, mediante

$$\varphi = N \circ \Psi^{-1}.$$

Esto motiva la siguiente idea.

• Referir las integrales, mediante un cambio de variables, al elemento de referencia.

Aplicando la regla de la cadena

$$\varphi_x = N_\xi \xi_x + N_\eta \eta_x, \qquad \varphi_y = N_\xi \xi_y + N_\eta \eta_y,$$

por lo que los productos $\varphi_x \varphi_x^T$, $\varphi_y \varphi_y^T$ se expresan como

$$\begin{aligned} \varphi_x \varphi_x^T &= \xi_x^2 N_\xi N_\xi^T + \xi_x \eta_x (N_\xi N_\eta^T + N_\eta N_\xi^T) + \eta_x^2 N_\eta N_\eta^T, \\ \varphi_y \varphi_y^T &= \xi_y^2 N_\xi N_\xi^T + \xi_y \eta_y (N_\xi N_\eta^T + N_\eta N_\xi^T) + \eta_y^2 N_\eta N_\eta^T, \end{aligned}$$

que sumados dan

$$\varphi_x \varphi_x^T + \varphi_y \varphi_y^T = (\xi_x^2 + \xi_y^2) N_\xi N_\xi^T + (\xi_x \eta_x + \xi_y \eta_y)(N_\xi N_\eta^T + N_\eta N_\xi^T) + (\eta_x^2 + \eta_y^2) N_\eta N_\eta^T.$$

Para calcular $\xi_x, \xi_y, \eta_x, \eta_y$ observamos que son precisamente las componentes de $D\Psi^{-1}$. Recordando que

$$\Psi(\underline{\xi}) = \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} x_1 \\ y_1 \end{bmatrix} + \begin{bmatrix} x_2 - x_1 & x_3 - x_1 \\ y_2 - y_1 & y_3 - y_1 \end{bmatrix} \begin{bmatrix} \xi \\ \eta \end{bmatrix},$$

tenemos que

$$D\Psi = \begin{bmatrix} x_\xi & x_\eta \\ y_\xi & y_\eta \end{bmatrix} = \begin{bmatrix} x_2 - x_1 & x_3 - x_1 \\ y_2 - y_1 & y_3 - y_1 \end{bmatrix},$$

y de aquí que

$$D\Psi^{-1} = \begin{bmatrix} \xi_x & \xi_y \\ \eta_x & \eta_y \end{bmatrix} = \frac{1}{J} \begin{bmatrix} y_3 - y_1 & -(x_3 - x_1) \\ -(y_2 - y_1) & x_2 - x_1 \end{bmatrix},$$

donde J es el jacobiano de la aplicación Ψ,

$$J = (x_2 - x_1)(y_3 - y_1) - (x_3 - x_1)(y_2 - y_1).$$

Vemos que tanto J como ξ_x, ξ_y, η_x y η_y son constantes.

Definiendo

$$\begin{aligned} S_1 &= \textstyle\int_{K_0} N_\xi N_\xi^T, & S_2 &= \textstyle\int_{K_0} (N_\xi N_\eta^T + N_\eta N_\xi^T), \\ S_3 &= \textstyle\int_{K_0} N_\eta N_\eta^T, & S_4 &= \textstyle\int_{K_0} N N^T, \end{aligned}$$

tenemos que

$$E = aS_1 + bS_2 + cS_3 + \alpha d S_4,$$

donde

$$\begin{aligned} a &= |J|(\xi_x^2 + \xi_y^2) = \frac{(y_3 - y_1)^2 + (x_3 - x_1)^2}{|J|}, \\ b &= |J|(\xi_x \eta_x + \xi_y \eta_y) = -\frac{(y_3 - y_1)(y_2 - y_1) + (x_3 - x_1)(x_2 - x_1)}{|J|}, \\ c &= |J|(\eta_x^2 + \eta_y^2) = \frac{(y_2 - y_1)^2 + (x_2 - x_1)^2}{|J|}, \\ d &= |J|. \end{aligned}$$

Observación. Si se han numerado los nodos con la orientación positiva (antihoraria), la aplicación Ψ mantiene la orientación, y por tanto $J > 0$.

Es decir, para cada triángulo K, la matriz elemental de $a_K(u, v)$ es combinación lineal de las mismas cuatro matrices S_i, $i = 1, \ldots, 4$, con coeficientes a, b, c, d que dependen exclusivamente de las coordenadas x, y de los vértices del triángulo K. Por tanto, para cada triángulo en una triangulación, sólo hay que calcular los coeficientes a, b, c y d, y con ellos combinar las matrices S_i. Estas matrices reciben el nombre de *matrices básicas.*

Las matrices básicas son

$$S_1 = \tfrac{1}{2}\begin{bmatrix} 1 & -1 & 0 \\ -1 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \qquad S_2 = \tfrac{1}{2}\begin{bmatrix} 2 & -1 & -1 \\ -1 & 0 & 1 \\ -1 & 1 & 0 \end{bmatrix},$$

$$S_3 = \tfrac{1}{2}\begin{bmatrix} 1 & 0 & -1 \\ 0 & 0 & 0 \\ -1 & 0 & 1 \end{bmatrix}, \qquad S_4 = \tfrac{1}{24}\begin{bmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{bmatrix}.$$

Observación. Si en el funcional aparecen términos cruzados, $u_x u_y$, es necesario disponer de las matrices S_{12} y S_{12}^T, donde

$$S_{12} = \int_{K_0} N_\xi N_\eta^T = \frac{1}{2}\begin{bmatrix} 1 & 0 & -1 \\ -1 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}.$$

MatLab. Al principio del programa, después del encabezamiento, habrá que introducir las matrices básicas. Por ejemplo, para S_1 escribimos

```
S1=[1,-1,0;-1,1,0;0,0,0]/2;
```

De la misma forma se introducen las demás matrices, que denotaremos dentro del programa por `S2`, `S3` y `S4`.

También habrá que incluir un bucle que ensamble la matriz de rigidez. La secuencia de instrucciones puede ser parecida a ésta:

```
% reservamos espacio para la matriz de rigidez
M=size(Z,2); % número de nodos
A=sparse(M,M); % creo la matriz de rigidez (dispersa)
```

```
% ensamblado de la matriz de rigidez
for j=1:size(T,2); % recorremos los triángulos
   I=T(1:3,j); % obtenemos los vértices del triángulo
   x=Z(1,I); % coord. x de los vértices
   y=Z(2,I); % coord. y de los vértices
   % calculamos el jacobiano
   J= (x(2)-x(1))*(y(3)-y(1))-(x(3)-x(1))*(y(2)-y(1));
   % calculamos los coeficientes a, b, c, d.  Para b utilizo
   % el nombre bb para no confundirlo con el vector de carga
   a=((y(3)-y(1))*(y(3)-y(1))+(x(3)-x(1))*(x(3)-x(1)))/J;
   bb=-((y(3)-y(1))*(y(2)-y(1))+(x(3)-x(1))*(x(2)-x(1)))/J;
   c=((y(2)-y(1))*(y(2)-y(1))+(x(2)-x(1))*(x(2)-x(1)))/J;
   d=J;
   Se=a*S1+bb*S2+c*S3+alpha*d*S4; % matriz elemental
   % se suma la matriz elemental a las componentes
   % de A que correspondan
   A(I,I)=A(I,I)+Se;
end
```

4.8.3 Construcción del vector de carga

A continuación veremos como construir el vector de carga

$$b = \begin{bmatrix} L(\varphi_1) \\ \vdots \\ L(\varphi_M) \end{bmatrix}.$$

En nuestro problema modelo $L(v) = \int_\Omega fv + \int_{\Gamma_N} gv$, por lo que definiendo

$$\Phi = \begin{bmatrix} \varphi_1 \\ \vdots \\ \varphi_M \end{bmatrix},$$

se tiene que

$$b = \underline{f}_h + \underline{g}_h, \qquad \text{con} \quad \underline{f}_h = \int_\Omega f\Phi, \qquad \underline{g}_h = \int_{\Gamma_N} g\Phi.$$

Calcularemos por separado la contribución al vector de carga de cada uno de los sumandos.

Contribución de f. Las ideas son similares a las que se aplicaron al construir la matriz de rigidez. El cálculo se hace como suma de contribuciones de cada triángulo:

$$\underline{f}_h = \sum_{j=1}^{N} \int_{K_j} f\Phi.$$

Cada uno de los vectores

$$\int_{K_j} f\Phi$$

tiene todas sus componentes no nulas, salvo quizás las 3 componentes que corresponden a las 3 funciones base no nulas sobre el elemento K_j. De esta forma, el procedimiento para calcular $\underline{f}_h$ consiste en para cada triángulo K_j, cuyos nodos tienen índices j_1, j_2, j_3, calcular el vector de 3 componentes

$$f_{K_j} = \int_{K_j} f \begin{bmatrix} \varphi_{j_1} \\ \varphi_{j_2} \\ \varphi_{j_3} \end{bmatrix},$$

y sumar estos valores a las componentes j_1, j_2, j_3 de $\underline{f}_h$ (en la práctica se suman a las componentes de b).

Sea pues el triángulo K de vértices $\underline{x}_1, \underline{x}_2, \underline{x}_3$. Queremos expresar la integral sobre K como una integral sobre el triángulo de referencia K_0. Recordemos que $\varphi = N \circ \Psi^{-1}$, y que por tanto

$$\varphi \circ \Psi = N,$$

por lo que

$$f_K = \int_K f\varphi = |J| \int_{K_0} f(\Psi(\underline{\xi}))N(\underline{\xi})\, d\xi d\eta,$$

donde J es el determinante jacobiano de la transformación Ψ, que es positivo bajo nuestra hipótesis acerca de la orientación de los nodos.

Normalmente estas integrales no pueden expresarse en términos de funciones elementales, por lo que es necesario recurrir a fórmulas de cuadratura numérica. En general usamos una fórmula de s nodos

$$\underline{c}_1, \ldots, \underline{c}_s,$$

y sus correspondientes pesos, que agrupamos en el vector

$$w = [w_1, \ldots, w_s]^T.$$

Con esta fórmula,

$$\int_{K_0} fg \approx f(\underline{c}_j) w_j g(\underline{c}_j) = [f(\underline{c}_1), \ldots, f(\underline{c}_s)] \begin{bmatrix} w_1 & & \\ & \ddots & \\ & & w_s \end{bmatrix} \begin{bmatrix} g(\underline{c}_1) \\ \vdots \\ g(\underline{c}_s) \end{bmatrix}.$$

Denotando entonces por $\underline{N}$ la matriz de dimensión $3 \times s$ cuyas columnas son $N(\underline{c}_j)$, esto es,

$$\underline{N} = [N(\underline{c}_1), \ldots, N(\underline{c}_s)] = \begin{bmatrix} N_1(\underline{c}_1) & \ldots & N_1(\underline{c}_s) \\ N_2(\underline{c}_1) & \ldots & N_2(\underline{c}_s) \\ N_3(\underline{c}_1) & \ldots & N_3(\underline{c}_s) \end{bmatrix},$$

y por f_q al vector cuyas componentes son los valores de f en los transformados $\Psi(\underline{c}_j)$ de los nodos de cuadratura, esto es

$$f_q = \begin{bmatrix} f(\Psi(\underline{c}_1)) \\ \vdots \\ f(\Psi(\underline{c}_s)) \end{bmatrix}$$

podemos pues aproximar

$$f_K = J \int_{K_0} N(\underline{\xi}) f(\Psi(\underline{\xi}))\, d\xi d\eta \approx J \underline{N}\, \mathrm{diag}(w) f_q, \tag{4.28}$$

donde diag(w) es la matriz diagonal cuyos elementos diagonales son las componentes del vector w.

Los transformados de los nodos de cuadratura se calculan usando que

$$\Psi(\underline{\xi}) = \underline{x}_1 N_1(\underline{\xi}) + \underline{x}_2 N_2(\underline{\xi}) + \underline{x}_3 N_3(\underline{\xi}),$$

de donde

$$\Psi(\underline{c}_i) = [\underline{x}_1 \ \underline{x}_2 \ \underline{x}_3] \begin{bmatrix} N_1(\underline{c}_i) \\ N_2(\underline{c}_i) \\ N_3(\underline{c}_i) \end{bmatrix},$$

es decir,

$$[\Psi(\underline{c}_1) \ \ldots \ \Psi(\underline{c}_s)] = [\underline{x}_1 \ \underline{x}_2 \ \underline{x}_3]\underline{N}.$$

MatLab. Tenemos un procedimiento sistemático:

1. Se calcula la matriz $\underline{N}$.
2. Para cada triángulo K_j, j=1,...,N,
   (a) se transforman los nodos de cuadratura;
   (b) se evalúa f en los transformados de los nodos de cuadratura para formar el vector f_q;
   (c) se aproxima f_{K_j} mediante (4.28)

Usaremos la fórmula de cuadratura gaussiana definida por los siguientes coeficientes:

i	ξ_i	η_i	w_i
1	$1/3$	$1/3$	0.1125
2	$(6+\sqrt{15})/21$	$(6+\sqrt{15})/2$	$(155+\sqrt{15})/2400$
3	$(9-2\sqrt{15})/21$	$(6+\sqrt{15})/2$	$(155+\sqrt{15})/2400$
4	$(6+\sqrt{15})/21$	$(9-2\sqrt{15})/21$	$(155+\sqrt{15})/2400$
5	$(6-\sqrt{15})/21$	$(6-\sqrt{15})/21$	$(155-\sqrt{15})/2400$
6	$(9+2\sqrt{15})/21$	$(6-\sqrt{15})/21$	$(155-\sqrt{15})/2400$
7	$(6-\sqrt{15})/21$	$(9+2\sqrt{15})/21$	$(155-\sqrt{15})/2400$

Esta fórmula, que se usa con frecuencia, es exacta en el triángulo de referencia K_0 para polinomios de grado menor o igual que 5.

En el programa principal tendremos que incluir un conjunto de instrucciones parecido al siguiente:

```
M=size(Z,2); % número de nodos
b=zeros(M,1); % reservo sitio para el vector de carga
% pesos de la cuadratura
w=[0.1125,...
(155+sqrt(15))/2400,(155+sqrt(15))/2400,...
(155+sqrt(15))/2400,(155-sqrt(15))/2400,...
(155-sqrt(15))/2400,(155-sqrt(15))/2400];
% matriz Nc:  valores de N en los nodos de la cuadratura
Nc=[1/3,-(96+23*sqrt(15))/42,...
-(102+17*sqrt(15))/42,(6+sqrt(15))/21,...
(9+2*sqrt(15))/21,(6-sqrt(15))/21,(6-sqrt(15))/21;
1/3,(6+sqrt(15))/21,(9-2*sqrt(15))/21,(6+sqrt(15))/21,...
(6-sqrt(15))/21,(9+2*sqrt(15))/21,(6-sqrt(15))/21;
1/3,(6+sqrt(15))/2,(6+sqrt(15))/2,(9-2*sqrt(15))/21,...
(6-sqrt(15))/21,(6-sqrt(15))/21,(9+2*sqrt(15))/21];
% ensamblado del vector de carga
for j=1:size(T,2); % recorremos los triángulos
   I=T(1:3,j); % obtenemos los nodos del triángulo
   x=Z(1,I); % coord. x de los vértices
   y=Z(2,I); % coord. y de los vértices
   % calculamos el jacobiano
   J= (x(2)-x(1))*(y(3)-y(1))-(x(3)-x(1))*(y(2)-y(1));
   % imágenes por Psi de los nodos de cuadratura
   Psicx=x*Nc; % coord. x de las imágenes
   Psicy=y*Nc; % coord. y de las imágenes
```

```
    % fq:  valor de f (denotada por fsm dentro del programa)
    % en las transformadas de los nodos.  f se debe definir
    % externamente y debe devolver un vector columna
    fq=feval(fsm,Psicx,Psicy);
    % sumamos la contribución del triángulo a las
    % componentes de b que correspondan
    b(I)=b(I)+J*Nc*diag(w)*fq;
end
```

Como se ha indicado, la función f del lado derecho de la ecuación se debe definir externamente. Esto se hará por medio de un fichero `.m` independiente. Si, por ejemplo, el lado derecho es, $f(x,y) = xy$ entonces el fichero podría ser

```
function func=fsm(x,y); % los argumentos son matrices
func=(x.*y)'; % la multiplicación es componente a componente
% la salida debe ser un vector columna
```

Al llamar al programa principal habrá que introducir como penúltimo argumento el nombre del fichero, que no tiene por qué ser `fsm`, entre comillas.

Contribución de los datos Neumann. A continuación estudiamos la contribución al vector de carga del término procedente de los datos de Neumann, $\underline{g}_h$. Escribiendo Γ_N como unión de segmentos, $\Gamma_N = l_1 \cup l_2 \cdots \cup l_{n_s}$, podemos expresar $\underline{g}_h$ como suma de contribuciones de los distintos segmentos,

$$\underline{g}_h = \int_{\Gamma_N} g\Phi = \sum_{j=1}^{n_s} \int_{l_j} g\Phi.$$

Cada uno de los vectores

$$\int_{l_j} g\Phi$$

tiene todas sus componentes nulas excepto quizá aquellas que correspondan a los nodos $\underline{x}_{j_1}$, $\underline{x}_{j_2}$ que están sobre el segmento l_j. Así pues, para construir el vector $\underline{g}_h$ calculamos el vector de 2 componentes

$$g_{l_j} = \int_{l_j} g \begin{bmatrix} \varphi_{j_1} \\ \varphi_{j_2} \end{bmatrix}$$

para cada segmento l_j, y sumamos los valores obtenidos a las componentes j_1, j_2 de $\underline{g}_h$ (en la práctica se suman a las componentes j_1, j_2 de b).

Abandonamos la j para no recargar la notación. Sea pues un segmento l de Γ_N, de extremos $\underline{x}_1$, $\underline{x}_2$, con funciones de base asociadas φ_1, φ_2. Definimos

$$\varphi = \begin{bmatrix} \varphi_1 \\ \varphi_2 \end{bmatrix}.$$

Parametrizando l mediante la aplicación $\gamma : [0,1] \mapsto \mathbb{R}^2$,

$$\gamma(\xi) = \underline{x}_1 + \xi(\underline{x}_2 - \underline{x}_1),$$

se tiene que

$$g_l = \int_l g\varphi \, ds = \|\underline{x}_2 - \underline{x}_1\| \int_0^1 g(\gamma(\xi))L(\xi)\, d\xi, \tag{4.29}$$

donde $L = [L_1, L_2]^T$ se define como $L = \varphi \circ \gamma$. Nótese que L_1 y L_2 son funciones lineales de ξ, y que $L_1(0) = L_2(1) = 1$, $L_1(1) = L_2(0) = 0$. Por tanto,

$$L_1(\xi) = 1 - \xi, \qquad L_2(\xi) = \xi.$$

Normalmente las integrales (4.29) no pueden expresarse en términos de funciones elementales, por lo que de nuevo es necesario recurrir a fórmulas de cuadratura. Sean $c_1, \ldots, c_s$, los nodos de una de dichas fórmulas de cuadratura en $[0,1]$, y $w_1, \ldots, w_s$ los correspondientes pesos, que agrupamos en el vector

$$w = [w_1, \ldots, w_s]^T.$$

Denotando por L la matriz cuyas columnas son los valores de L en cada uno de los nodos de cuadratura c_j, esto es,

$$\underline{L} = [L(c_1), \cdots, L(c_s)],$$

y por g_q el vector

$$g_q = \begin{bmatrix} g(\gamma(c_1)) \\ \vdots \\ g(\gamma(c_s)) \end{bmatrix},$$

tenemos entonces que el vector g_l se aproxima como

$$g_l = \|\underline{x}_2 - \underline{x}_1\| \int_0^1 g(\gamma(\xi))L(\xi)\, d\xi \approx \|\underline{x}_2 - \underline{x}_1\| \underline{L}\, \mathrm{diag}(w) g_q. \tag{4.30}$$

Para calcular las transformadas $\gamma(c_i)$ de los nodos de cuadratura observamos que $\gamma(\xi) = \underline{x}_1 L_1(\xi) + \underline{x}_2 L_2(\xi)$, de donde

$$[\gamma(c_1)\ \ldots\ \gamma(c_s)] = [\underline{x}_1\ \underline{x}_2]\underline{L}.$$

Una situación bastante frecuente es aquélla en que g es constante. En ese caso

$$g_l = \int_l g\varphi\, ds = g\|\underline{x}_2 - \underline{x}_1\| \int_0^1 L\, d\xi = g\|\underline{x}_2 - \underline{x}_1\| \begin{bmatrix} \frac{1}{2} \\ \frac{1}{2} \end{bmatrix}. \tag{4.31}$$

MatLab. Teniendo la matriz `IN` (con los lados correspondientes a datos Neumann) y la matriz `Z` (con las coordenadas de cada nodo), el procedimiento sistemático para sumar el vector $\underline{g}_h$ al vector b es como sigue:

1. Se calcula la matriz $\underline{L}$.
2. Para cada segmento l_j, $j = 1, \ldots, n_s$,
   (a) se transforman los nodos de cuadratura c_j;
   (b) se evalúa g en los transformados para formar el vector g_q;
   (c) se aproxima g_{l_j} mediante (4.30) (o (4.31) si g es constante).

Como regla de cuadratura usamos la de Simpson, con nodos y pesos

c_j	0	1/2	1
w_j	1/6	4/6	1/6

Las instrucciones de MatLab podrían ser:

```
wborde=[1/6 4/6 1/6]; % pesos de la cuadratura
% les llamamos wborde para no confundirlos con los pesos w
% utilizados para calcular la contribución de f.
Lc=[1,1/2,0;0,1/2,1]; % valores de L en nodos de cuadratura
% ensamblado del vector de carga
for j=1:length(IN); % recorremos los lados
   I=T(2:3,IN(j)); % obtenemos los nodos del lado
   x=Z(1,I); % coord. x de los vértices
   y=Z(2,I); % coord. y de los vértices
   % imágenes por gamma de los nodos de cuadratura
   gammacx=x*Lc; % coordenada x de la transformada
   gammacy=y*Lc; % coordenada y de la transformada
   % gq:  valor de g (denotada por dn dentro del programa)
   % en las transformadas de los nodos.  g se debe definir
   % externamente, y debe devolver un vector columna
   gq=feval(dn,gammacx,gammacy);
   % longitud del intervalo
   lg= sqrt((x(2)-x(1))*(x(2)-x(1))+(y(2)-y(1))*(y(2)-y(1)));
   % sumamos la contribución del segmento
   % a las componentes de b que correspondan
   b(I)=b(I)+lg*Lc*diag(wborde)*gq;
end
```

La función g también deberá ser definida externamente por medio de un fichero `.m` independiente. Al utilizar el programa principal habrá que pasarle el nombre de dicho fichero (entre comillas) como último argumento.

4.8.4 Condiciones Dirichlet y resolución del sistema lineal

Una vez construidos la matriz de rigidez A y el vector de carga b, tenemos que formar a partir de ellos el sistema (4.27), que tiene en cuenta las condiciones de Dirichlet homogéneas. Para ello:

- Se hacen cero las filas y columnas de A, y las filas de b correspondientes a índices de `ID`.
- Se coloca en los elementos diagonales de A correspondientes a índices de `ID` el valor 1.

MatLab.

```
% vector con los índices de nodos con datos Dirichlet
I=reshape(T(2:3,ID),2*length(ID),1)
% hacemos ceros las filas y columnas de A de índices de ID
A(I,:)=zeros(size(A(I,:))); % las filas
A(:,I)=zeros(size(A(:,I))); % las columnas
% creamos un vector que tenga unos en índices de ID, y ceros
% en las demás posiciones.
v=zeros(M,1);
v(I)=1;
% unos en la diagonal en las posiciones de los índices de ID
A=A+spdiags(v,0,M,M); % matriz diagonal dispersa
% ceros en las componentes de b de índices de ID
b(I)=0;
```

La resolución del sistema lineal

$$Au = b$$

es en general una tarea compleja, difícil de programar, pues hay que aprovechar el hecho de que la matriz A es dispersa, lo que obliga a usar algoritmos especiales para ese tipo de matrices. Sin embargo, nosotros no nos tendremos que preocupar por eso, pues el MatLab realizará la tarea por nosotros.

MatLab. El sistema se resuelve por medio de la instrucción

```
u=A\b
```

En [I] se describen de forma sencilla algunos métodos para resolver sistemas lineales dispersos.

4.8.5 Resultados numéricos

Uniendo adecuadamente los sectores de código que hemos ido suministrando, se construye sin mucha dificultad un programa de MatLab que permite resolver nuestro problema modelo (y otros muchos). Veamos algunos resultados numéricos obtenidos con dicho programa.

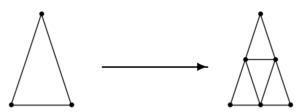
Empezamos por asegurarnos de que el programa funciona correctamente. Para ello resolvemos un problema cuya solución conocemos. Concretamente consideramos la función $u(x,y) = x(x^2 - x - y^2 + y)$, que es la solución de

$$\begin{cases} -\Delta u + \alpha u = f, & (x,y) \in \Omega, \\ u(x,y) = 0, & (x,y) \in \Gamma_D, \\ \frac{\partial u}{\partial n} = g, & (x,y) \in \Gamma_N, \end{cases}$$

con $\Omega = (0,1) \times (0,1)$, $\Gamma_D = \{(x,y) \in \mathbb{R}^2 : x = 0, 0 \leq y \leq 1\}$, $\Gamma_N = \partial\Omega \setminus \Gamma_D$, $\alpha = 1$, $f = 2 - 4x$ y

$$g(x,y) = \begin{cases} y^2 - y & \text{si } x = 0,\ 0 < y < 1, \\ -x & \text{si } y = 0 \text{ o } y = 1,\ 0 < x < 1. \end{cases}$$

En la figura 4.1 mostramos las soluciones numéricas correspondientes a dos triangulaciones. La segunda triangulación es un *refinamiento* de la primera, obtenida descomponiendo cada triángulo en cuatro, mediante la unión de los puntos medios de cada lado:



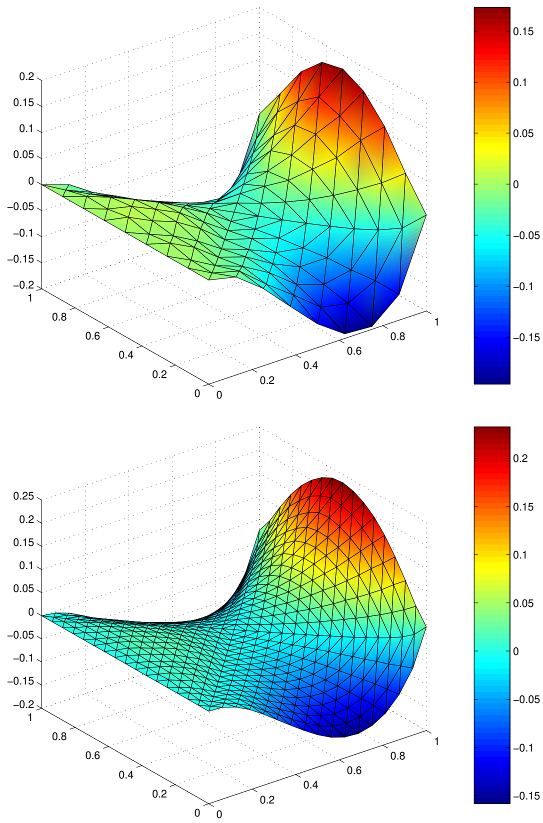


Figura 4.1: Ejemplo en un cuadrado. Soluciones numéricas.

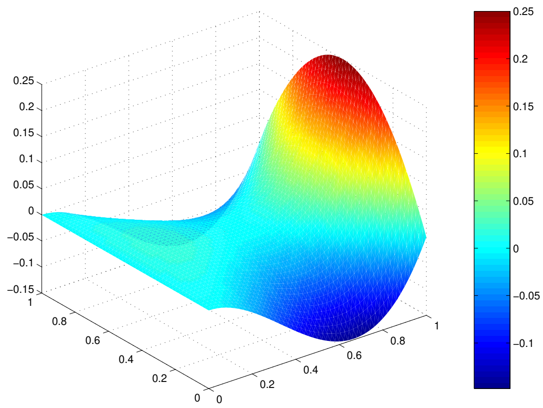


Figura 4.2: Ejemplo en un cuadrado. Solución teórica.

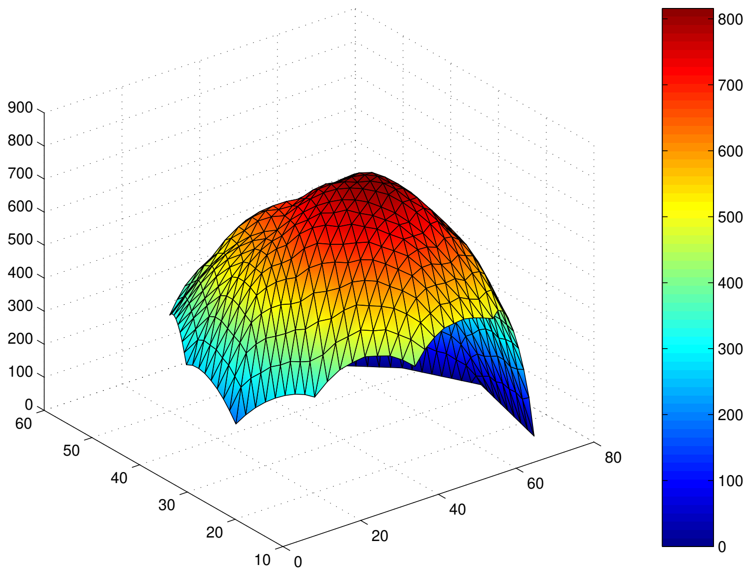


Figura 4.3: Problema modelo. Superficie solución.

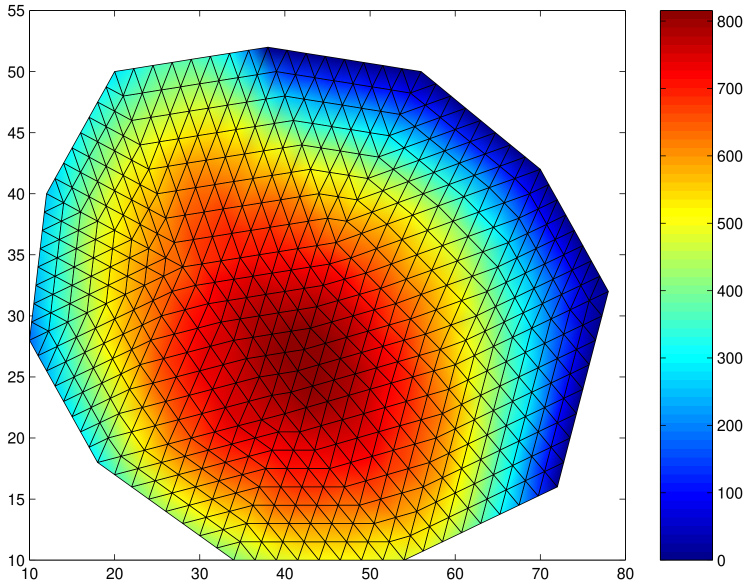


Figura 4.4: Problema modelo. Representación plana.

En la figura 4.2 se muestra la verdadera solución. Comparando las dos figuras observamos que la solución numérica se parece, a medida que vamos refinando la triangulación, cada vez más a la teórica, y que el parecido entre ambas es muy alto.

A continuación resolvemos nuestro problema modelo con $\alpha = 0$, $f = 4$ y $g = -0.6\sqrt{|(100-x)(100-y)|}$. En este caso no conocemos la solución explícitamente, y tenemos que recurrir al cálculo numérico si queremos tener una idea cuantitativa de cómo es. En la figura 4.3 se representa la superficie solución. Este problema se puede ver como un modelo para una distribución estacionaria de temperaturas en un medio en el que hay una fuente de calor (representada por f), un flujo de calor en la frontera (representado por g), y una zona de la frontera, Γ_D, que se mantiene a cero grados. Los problemas de este tipo admiten una representación plana, mediante un código de colores, muy conveniente. Damos un ejemplo en la figura 4.4.

4.8.6 Condiciones Dirichlet no homogéneas

Supongamos ahora que en lugar de condiciones Dirichlet homogéneas sobre la parte Γ_D de la frontera tenemos condiciones Dirichlet más generales,

$$u(\underline{x}) = \rho(\underline{x}), \qquad \underline{x} \in \Gamma_D.$$

Lo que se hace es miminimizar el funcional en el convexo cerrado

$$V_h = \{v \in C(\overline{\Omega}) : v_{|_{K_j}} \in \mathbb{P}_1(K_j),\, v(\underline{x}_j) = \rho(\underline{x}_j),\, j \in I_D\}.$$

Denotando $u_j = u_h(\underline{x}_j)$ tenemos que la solución de elementos finitos viene dada por la solución del sistema lineal

$$\sum_{j=1}^{M} a(\varphi_i, \varphi_j) u_j = L(\varphi_i), \qquad i \notin I_D.$$

Los valores nodales u_j con $j \in I_D$ no son incógnitas, así que los pasamos al segundo miembro, quedando que

$$\sum_{j \notin I_D} a_{ij} u_j = b_i - \sum_{j \in I_D} a_{ij} u_j, \qquad i \notin I_D.$$

Así pues, hay dos diferencias con el caso homogéneo: (i) hay que modificar el vector de carga, y (ii) los valores en los nodos $\underline{x}_j$, $j \in I_D$ no son cero, sino que vienen dados por $u_j = \rho(\underline{x}_j)$. Por lo demás el procedimiento a seguir es como en el caso homogéneo.

MatLab. Una vez construidos la matriz de rigidez completa A, de elementos $a_{ij} = a(\varphi_i, \varphi_j)$, $i, j = 1, \ldots, M$ y el vector de carga b, de elementos $b_i = L(\varphi_i)$, $i = 1, \ldots, M$, habrá que hacer lo siguiente:

- Se construye un vector u (no hay inconveniente en usar el mismo nombre que para el vector solución) cuyas componentes son $u_j = 0$ si $j \notin$`ID`, $u_j = \rho(\underline{x}_j)$ si $j \in$`ID`.
- Se resta a b el vector Au.
- Se hacen cero las filas y columnas de A correspondientes a índices de `ID`.
- Se coloca en los elementos diagonales de A correspondientes a índices de `ID` el valor 1.
- Se colocan los valores de u de índices de `ID` en las posiciones correspondientes del vector b.

El conjunto de instrucciones de MatLab podría ser:

```
% vector con los índices de nodos con datos Dirichlet
I=reshape(T(2:3,ID),2*length(ID),1)
% construimos el vector u de datos Dirichlet
u=zeros(M,1);
u(I)=rho(Z(1,I),Z(2,I)); % rho se define externamente
% se modifica el vector de carga
b=b-A*u;
```

```
% hacemos ceros las filas y columnas de A de índices de ID
A(I,:)=zeros(size(A(I,:))); % las filas
A(:,I)=zeros(size(A(:,I))); % las columnas
% creamos un vector que tenga unos en índices de ID, y ceros
% en las demás posiciones.
v=zeros(M,1);
v(I)=1;
% unos en la diagonal en las posiciones de los índices de ID
A=A+spdiags(v,0,M,M); % matriz diagonal dispersa
% ceros en las componentes de b de índices de ID
b(I)=u(I);
```

El sistema se resuelve como en el caso homogéneo.

4.8.7 Datos de Robin

Supongamos que estamos estudiando un problema con condiciones de Robin en una parte Γ_R de la frontera, por ejemplo,

$$\begin{cases} -\Delta u + u = f & \text{en } \Omega \subset \mathbb{R}^2, \\ \frac{\partial u}{\partial n} + u = g & \text{en } \Gamma_R, \\ u = 0 & \text{en } \partial\Omega \setminus \Gamma_R. \end{cases}$$

El espacio de Hilbert adecuado para tratar este problema es

$$V = \{v \in H^1(\Omega) : \, v = 0 \text{ c.t.p. en } \partial\Omega \setminus \Gamma_R\}.$$

Multiplicando la ecuación por una función de V e integrando por partes se tiene, tras imponer las condiciones de contorno, que

$$\int_\Omega (\nabla u \cdot \nabla v + uv) + \int_{\Gamma_R} uv = \int_\Omega fv + \int_{\Gamma_R} gv.$$

En la forma bilineal aparece una integral sobre Γ_R ,

$$r(u,v) = \int_{\Gamma_R} uv,$$

que contribuirá a la matriz de rigidez con una matriz R, de dimensión $M \times M$ con elementos

$$r_{ij} = \int_{\Gamma_R} \varphi_i \varphi_j.$$

Escribiendo Γ_R como unión de segmentos, $\Gamma_R = l_1 \cup l_2 \cup \cdots \cup l_{n_r}$, tenemos que

$$r_{ij} = \sum_{k=1}^{n_r} b_{l_k}(\varphi_i, \varphi_j), \qquad \text{donde} \;\; b_{l_k}(u,v) = \int_{l_k} uv.$$

Para cada segmento l_k todos los valores $b_{l_k}(\varphi_i, \varphi_j)$ son cero excepto quizás aquellos que correspondan a las funciones base φ_{k_1}, φ_{k_2} asociadas a los nodos $\underline{x}_{k_1}$, $\underline{x}_{k_2}$ del segmento l_k. Así, para calcular la contribución del segmento l_k a la matriz R sólo es necesario calcular la matriz

$$R_k = \int_{l_k} \begin{bmatrix} \varphi_{k_1} \\ \varphi_{k_2} \end{bmatrix} [\varphi_{k_1} \;\; \varphi_{k_2}] \, ds,$$

y añadir sus valores a la submatriz de R que se obtiene al intersecar las filas y columnas de índices k_1, k_2. Parametrizando l mediante $\gamma : [0,1] \mapsto \mathbb{R}^2$, $\gamma(\xi) = \underline{x}_1 + \xi(\underline{x}_2 - \underline{x}_1)$, la matriz R_k se expresa como

$$R_k = \|\underline{x}_2 - \underline{x}_1\| \int_0^1 LL^T \, d\xi,$$

donde $L = \varphi \circ \gamma$. Por tanto, el procedimiento sistemático para sumar R a los otros términos de la matriz de rigidez, y obtener así A, consiste en calcular primero

$$\mathtt{G} = \int_0^1 LL^T \, d\xi = \frac{1}{6} \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix},$$

y después, para cada lado l_k, multiplicar esta matriz por la longitud del lado, y sumar sus elementos a las correspondientes componentes de A.

MatLab. Habrá que crear una matriz `IR` con los índices de los triángulos tienen lados en Γ_R. Las instrucciones para recoger las contribuciones de los segmentos con condiciones de Robin podrían ser las siguientes:

```
G=[2 1
1 2]/6;
for j=1:length(IR); % recorremos los triángulos
   I=T(2:3,IR(j)); % obtenemos los vértices del triángulo
   x=Z(1,I); % coord. x de los vértices
   x=Z(2,I); % coord. y de los vértices
   % calculamos la longitud del lado
   lg= sqrt((x(2)-x(1))*(x(2)-x(1))+(y(2)-y(1))*(y(2)-y(1)));
   Se=lg*G; % matriz elemental
   % se suma la matriz elemental a las componentes
   % de A que correspondan
   A(I,I)=A(I,I)+Se;
end
```

4.8.8 Funciones polinómicas a trozos de orden más alto

Hemos descrito cómo se programan los elementos finitos en el caso en que se utilicen elementos triangulares y funciones continuas, lineales sobre cada triángulo. En el caso de funciones continuas, cuadráticas o cúbicas sobre cada triángulo, las cosas se hacen esencialmente de la misma forma, aunque los cálculos son bastante más complejos.

La contribución E del triángulo K a la matriz de rigidez se calcula como combinación de matrices básicas,

$$E = aS_1 + bS_2 + cS_3 + \alpha dS_4,$$

igual que se hacia en el caso de funciones lineales, tomando los coeficientes a, b, c, d los mismos valores que en aquel caso. Las matrices básicas también se definen igual,

$$\begin{array}{ll} S_1 = \int_{K_0} N_\xi N_\xi^T, & S_2 = \int_{K_0} (N_\xi N_\eta^T + N_\eta N_\xi^T) \\ S_3 = \int_{K_0} N_\eta N_\eta^T, & S_4 = \int_{K_0} N N^T. \end{array}$$

La diferencia estriba en que cambia el número de nodos en cada triángulo y en consecuencia el número de componentes de N, que pasa a ser 6 en el caso de funciones cuadráticas y 10 en el caso de funciones cúbicas. De esta forma las matrices elementales son matrices 6×6 en el caso de funciones cuadráticas y 10×10 en el caso de funciones cúbicas. De la misma forma, la matriz $\underline{N} = [N(\underline{c}_1), \ldots, N(\underline{c}_s)]$ que aparece al construir el vector de carga tendrá más filas, no habiendo por lo demás cambios especialmente significativos.

Las adaptaciones necesarias para tratar las condiciones de frontera son también más o menos obvias.

4.9 Comentarios y bibliografía

El análisis funcional que utilizamos en el capítulo (espacios de Sobolev y teoría variacional de ecuaciones elípticas) está muy bien explicado en [B].

Probablemente el texto de introducción a los elementos finitos más fácil de leer y con una crítica excelente sea [J]. Vino a sustituir al clásico "Strang-Fix" [SF], quizás el primer libro que explicaba los elementos finitos "para matemáticos", y que sigue resultando magnífico como libro de consulta .

Otro dos libros muy accesibles son [BCO] y [CO1], que son los dos primeros volúmenes de una "enciclopedia" de seis pequeños tomos dedicada a los elementos finitos. Los restantes volúmenes, [CO2], [CO3], [OC1], [OC2], más especializados, también son recomendables. El último volumen [CO3] trata sobre elementos finitos aplicados a la mecánica de fluidos. Este tema se trata de una forma sencilla en [Pi].

Algo más antiguo que [J], aunque con aplicaciones más cuidadas y más completo es [MW]. El libro de Ciarlet [C] trata en profundidad el análisis de la convergencia del método de elementos finitos. El libro de Zienkiewicz [Z], además de ser un clásico, está muy orientado a las aplicaciones, y es extraordinariamente completo en la discusión de todo tipo de elementos para el cálculo de estructuras. Tiene además el atractivo del punto de vista originario de los ingenieros. Hay una versión abreviada y simplificada, [ZM]. Otros dos textos con buena crítica son [AB] y [BS]. Para captar la amplitud del tema se recomienda [QV].

Sobre la programación del método hay unas notas excelentes (no publicadas) del profesor Bosco García Archilla (U. Sevilla). El libro de Schwartz [Sc], que es muy completo en este aspecto, es indispensable si se quiere de verdad aprender a programar el método de elementos finitos.

Por falta de tiempo y por ser un tema excesivamente técnico, no se trata en el curso la resolución de los sistemas lineales dispersos que aparecen al aplicar el método de elementos finitos. Se puede encontrar una introducción clara y breve, tanto a los métodos directos como a los iterativos para la resolución de este tipo de sistemas en [I]. Como referencias más avanzadas tenemos [DER] y [GL] para métodos directos, y [Ax], [HY], [N] o el clásico [V] para métodos iterativos. Dentro de los métodos iterativos se pueden utilizar técnicas multimalla. De nuevo [I] proporciona una breve introducción. Hay varios buenos libros introductorios sobre la materia, por ejemplo [BR], [H2] o [W].

Como referencia general para temas de álgebra lineal numérica recomendamos el excelente libro [GV], y a un nivel más básico el estupendo [TB].

Capítulo 5

Métodos lineales multipaso

5.1 Introducción

Según hemos visto en el capítulo anterior, el método de Euler se puede obtener partiendo de la aproximación de la derivada $y'(x)$ por el cociente incremental

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

Esta aproximación es de orden 1. Una idea para obtener un método mejor es usar una aproximación de la derivada más precisa, por ejemplo, la fórmula de diferencias centradas

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

que es una aproximación de orden 2. Aplicándola en $x = x_{n+1}$ se tiene que

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

lo que da lugar a la iteración

$$y_{n+2} = y_n + 2hf(x_{n+1}, y_{n+1}), \tag{5.1}$$

que se conoce como método *leap-frog.* Como vemos, para calcular y_{n+2} se utilizan *dos* valores anteriores, y_n e y_{n+1}, y no uno sólo, como en los métodos del capítulo anterior. Diremos que es un *método de dos pasos.*

Este método se puede obtener también a partir de una fórmula de cuadratura. Para ello observamos que

$$y(x_{n+2}) - y(x_n) = \int_{x_n}^{x_{n+2}} y'(x)\, dx, \tag{5.2}$$

y usamos la regla del punto medio para aproximar la integral, obteniendo que

$$y(x_{n+2}) - y(x_n) \approx 2hy'(x_{n+1}),$$

lo que da lugar al método (5.1), que también se conoce como *regla del punto medio.*

Si aproximamos la integral (5.2) usando la regla de Simpson obtenemos que

$$y(x_{n+2}) - y(x_n) \approx \frac{h}{3}(y'(x_n) + 4y'(x_{n+1}) + y'(x_{n+2})).$$

Esta aproximación conduce al método numérico,

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

conocido como *regla de Simpson*, que también es un método de dos pasos.

Tanto la regla del punto medio como la regla de Simpson son *métodos lineales de dos pasos.* En general, un *método lineal de k-pasos* obtiene las aproximaciones y_n a la solución del problema de valor inicial (1.1) como solución de la recurrencia

$$\sum_{j=0}^{k} \alpha_j y_{n+j} = h \sum_{j=0}^{k} \beta_j f(x_{n+j}, y_{n+j}), \qquad n = 0, \ldots, N-k. \tag{5.3}$$

Las constantes α_j y β_j son los *coeficientes del método* y satisfacen

$$\alpha_k = 1, \qquad |\alpha_0| + |\beta_0| \neq 0.$$

El método se dice *lineal* porque y_{n+k} es una combinación lineal de y_{n+j} y $f(x_{n+j}, y_{n+j})$; y se dice de k *pasos* porque se necesitan k valores anteriores $y_n, \ldots, y_{n+k-1}$ para calcular y_{n+k}.

La condición $\alpha_k = 1$ elimina la arbitrariedad que surge del hecho de que podemos multiplicar ambos lados de (5.3) sin cambiar el método. La condición $|\alpha_0| + |\beta_0| \neq 0$ elimina la posibilidad de que α_0 y β_0 sean ambas cero, excluyendo así métodos como por ejemplo

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

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

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

Notación. Para abreviar se suele escribir $f_n = f(x_n, y_n)$.

Una forma alternativa de especificar un método lineal multipaso es dando su *primer* y *segundo polinomio característicos*, ρ y σ, que se definen mediante

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

donde $\zeta \in \mathbb{C}$ es una variable muda.

Si $\beta_k = 0$, el método se dice *explícito.* En este caso se puede despejar explícitamente cada y_{n+k} en términos de los k valores anteriores $y_n, \ldots, y_{n+k-1}$. Si se han almacenado los valores $f_n, f_{n+1}, \ldots, f_{n+k-2}$, entonces una única evaluación de función, f_{n+k-1}, producirá y_{n+k}.

Si $\beta_k \neq 0$, el método es *implícito.* Para obtener y_{n+k} en cada paso tendremos que resolver la ecuación no lineal

$$y_{n+k} = h\beta_k f(x_{n+k}, y_{n+k}) + g, \tag{5.4}$$

donde g es una función conocida de los valores ya calculados,

$$g = \sum_{j=0}^{k-1} (h\beta_j f(x_{n+j}, y_{n+j}) - \alpha_j y_{n+j}).$$

Si el lado derecho de la igualdad (5.4) tiene una constante de Lipschitz M con respecto a y_{n+k} tal que $M < 1$, entonces este sistema de ecuaciones tiene una única solución y_{n+k}, que se puede aproximar tanto como se desee por medio de la iteración

$$y_{n+k}^{[\nu+1]} = h\beta_k f(x_{n+k}, y_{n+k}^{[\nu]}) + g, \qquad y_{n+k}^{[0]} \text{ arbitrario.} \tag{5.5}$$

Por el Teorema de la Aplicación Contractiva se cumple que

$$\lim_{\nu\to\infty} y_{n+k}^{[\nu]} = y_{n+k}.$$

Nótese que siempre que f sea Lipschitz podemos forzar que se cumpla la condición $M < 1$ tomando un valor de h suficientemente pequeño. Si éste fuera inaceptable se debería optar por otro método iterativo, como el de Newton.

Los métodos lineales multipaso presentan una dificultad computacional. Para aplicarlos es necesario disponer de los k *valores de arranque* $y_0, \dots, y_{k-1}$, y sin embargo el problema de valor inicial (1.1) sólo proporciona y_0. Si $k \geq 2$ habrá que obtener los $k-1$ valores de arranque restantes por algún otro procedimiento, como puede ser por desarrollo de Taylor, utilizando un método de Runge-Kutta, etc. Estos valores de arranque no coinciden por tanto en general con los valores de la solución teórica; tendrán un cierto error, y no hay por qué esperar que un método converja si no son suficientemente precisos.

Como ahora hay k valores de arranque, tenemos que modificar la definición de convergencia que se introdujo en el capítulo anterior.

Definición 5.1. *Un método lineal de k pasos se dice* convergente *(respectivamente* convergente de orden p*)* si para todo problema (1.1) con solución suficientemente regular se tiene que

$$\lim_{h\to 0^+} \max_{k\leq n\leq N} \|y(x_n) - y_n\| = 0, \qquad \text{si } \lim_{h\to 0^+} \max_{0\leq n\leq k-1} \|y(x_n) - y_n\| = 0,$$

(respectivamente

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

cuando $h \to 0^+$).

Observación. La condición $\lim\limits_{h\to 0^+} \max\limits_{0\le n\le k-1} \|y(x_n) - y_n\| = 0$ sobre los valores de arranque es equivalente a pedir que $\lim\limits_{h\to 0^+} y_n = y(x_0)$ para $n = 0, \dots, k-1$.

¿Qué condiciones deben satisfacer los coeficientes de un método lineal multipaso (5.3) para que éste converja? Si converge, ¿con qué orden lo hace? Intentaremos responder a estas preguntas en las siguientes secciones.

5.2 Consistencia

Queremos proceder como lo hacíamos en el caso de los métodos de un paso. La situación ideal sería que la solución y del problema de valor inicial (1.1) satisficiese las ecuaciones del método (5.3), pero en general esto no será así, y habrá un residuo o *error de truncación*

$$\begin{aligned} R_n &= \sum_{j=0}^{k} (\alpha_j y(x_{n+j}) - h\beta_j f(x_{n+j}, y(x_{n+j})) \\ &= \sum_{j=0}^{k} (\alpha_j y(x_{n+j}) - h\beta_j y'(x_{n+j})), \qquad n = 0, \dots, N-k, \end{aligned}$$

que, como en el caso de métodos de un paso, es lo que le falta a la solución teórica para satisfacer la ecuación del método.

Definición 5.2. *Un método lineal multipaso se dice consistente de orden p si para todo problema de valor inicial* (1.1) *con solución y suficientemente regular se tiene que*

$$\max_{0\le n\le N-k} \|R_n\| = O(h^{p+1})$$

cuando $h \to 0^+$.

Determinar las condiciones de orden es mucho más sencillo para un método lineal multipaso que para un método de Runge-Kutta. En efecto, si tomamos desarrollos de Taylor para $y(x_{n+j}) = y(x_n + jh)$ e $y'(x_{n+j}) = y'(x_n + jh)$ alrededor del punto $x = x_n$,

$$y(x_{n+j}) = y(x_n) + y'(x_n)jh + \cdots + \frac{y^{(p+1)}(x_n)(jh)^{p+1}}{(p+1)!} + O(h^{p+2}),$$
$$y'(x_{n+j}) = y'(x_n) + y''(x_n)jh + \cdots + \frac{y^{(p+1)}(x_n)(jh)^{p}}{p!} + O(h^{p+1}),$$

obtenemos que

$$R_n = C_0 y(x_n) + C_1 y'(x_n)h + \cdots + C_{p+1}y^{(p+1)}(x_n)h^{p+1} + O(h^{p+2}),$$

donde

$$\begin{aligned} C_0 &= \sum_{j=0}^{k} \alpha_j, \\ C_1 &= \sum_{j=0}^{k} \alpha_j j - \sum_{j=0}^{k} \beta_j, \\ &\vdots \\ C_q &= \frac{1}{q!}\left(\sum_{j=0}^{k} \alpha_j j^q - q \sum_{j=0}^{k} \beta_j j^{q-1} \right), \qquad q > 1. \end{aligned}$$

Así pues, un método lineal multipaso será consistente de orden p si y sólo si $C_j = 0$, $j = 0, \dots, p$. El valor C_{p+1} recibe el nombre de *constante de error.*

Ejemplo. El método

$$y_{n+2} + y_{n+1} - 2y_n = \frac{h}{4}(f_{n+2} + 8f_{n+1} + 3f_n)$$

es un método lineal de dos pasos con

$$\alpha_0 = -2, \quad \alpha_1 = 1, \quad \alpha_2 = 1, \qquad \beta_0 = 3/4, \quad \beta_1 = 8/4, \quad \beta_2 = 1/4.$$

Un sencillo cálculo muestra que $C_0 = C_1 = C_2 = C_3 = 0$ y que $C_4 = 1/4!$. Así pues, el método es de orden 3 y la constante de error es $C_4 = 1/4!$.

No es necesario efectuar el desarrollo de Taylor en $x = x_n$. Se puede (y se debe) efectuar en otro punto si hay posibilidad de aprovechar simetrías.

Ejemplo. Consideramos la regla de Simpson,

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

Si desarrollamos alrededor del punto medio $x = x_{n+1}$ queda un desarrollo en potencias impares de h,

$$\begin{aligned} R_n &= y(x_{n+2}) - y(x_n) - \tfrac{h}{3}(y'(x_n) + 4y'(x_{n+1}) + y'(x_{n+2})) = \\ &= (2y'(x_{n+1})h + \tfrac{1}{3}y'''(x_{n+1})h^3 + \tfrac{2}{5!}y^{(v)}(x_{n+1})h^5 + \cdots) \\ &\quad -\tfrac{h}{3}(6y'(x_{n+1}) + y'''(x_{n+1})h^2 + \tfrac{2}{4!}y^{(v)}(x_{n+1})h^4 + \cdots). \end{aligned}$$

Concluimos que $R_n = O(h^5)$, y el método es por tanto consistente de orden 4.

Ejemplo. Consideramos la regla del trapecio

$$y_{n+1} - y_n = \frac{h}{2}(f_{n+1} + f_n).$$

Si desarrollamos alrededor del punto medio $x = x_{n+\frac{1}{2}} := x_n + \frac{h}{2}$, de nuevo queda un desarrollo en potencias impares de h,

$$\begin{aligned} R_n &= y(x_{n+1}) - y(x_n) - \tfrac{h}{2}(y'(x_n) + y'(x_{n+1})) = \\ &= (y'(x_{n+\frac{1}{2}})h + \tfrac{1}{24}y'''(x_{n+\frac{1}{2}})h^3 + \cdots) \\ &\quad -\tfrac{h}{2}(2y'(x_{n+\frac{1}{2}}) + \tfrac{1}{4}y'''(x_{n+\frac{1}{2}})h^2 + \cdots). \end{aligned}$$

Concluimos que $R_n = O(h^3)$, y el método es por tanto consistente de orden 2.

Lo que tienen en común los dos ejemplos anteriores es que ambos métodos son simétricos, en el sentido de que

$$\alpha_j = -\alpha_{k-j}, \qquad \beta_j = \beta_{k-j}.$$

Es fácil comprobar, desarrollando alrededor del punto medio, que los métodos simétricos tienen orden par. Así pues, si un método es simétrico sólo será necesario considerar las condiciones de orden impar, pues las de orden par se cumplen necesariamente.

El siguiente teorema, que caracteriza el orden en términos de los polinomios característicos puede resultar útil.

Teorema 5.3. *El método lineal multipaso* (5.3) *es consistente de orden* $p \geq 1$ *si y sólo si existe una constante* $c \neq 0$ *tal que cuando* $\zeta \to 1$

$$\rho(\zeta) - \sigma(\zeta) \log \zeta = c(\zeta - 1)^{p+1} + O(|\zeta - 1|^{p+2}).$$

Prueba. Sea $\zeta = e^w$; por tanto $\zeta \to 1$ corresponde a $w \to 0$. Desarrollando en serie de Taylor tenemos que

$$\begin{aligned}
\rho(e^w) - w\sigma(e^w) &= \sum_{j=0}^{k} \alpha_j e^{jw} - w \sum_{j=0}^{k} \beta_j e^{jw} \\
&= \sum_{j=0}^{k} \alpha_j \left(\sum_{q=0}^{\infty} \frac{j^q w^q}{q!} \right) - w \sum_{j=0}^{k} \beta_j \left(\sum_{q=0}^{\infty} \frac{j^q w^q}{q!} \right) \\
&= \sum_{q=0}^{\infty} \frac{1}{q!} \left(\sum_{j=0}^{k} \alpha_j j^q \right) w^q - \sum_{q=1}^{\infty} \frac{1}{(q-1)!} \left(\sum_{j=0}^{k} \beta_j j^{q-1} \right) w^q.
\end{aligned}$$

Por tanto,

$$\rho(e^w) - w\sigma(e^w) = cw^{p+1} + O(w^{p+2})$$

para algún $c \neq 0$ si y sólo si el método es de orden p. Volviendo a la variable ζ y usando que

$$\log \zeta = (\zeta - 1) + O((\zeta - 1)^2) \qquad \text{para } \zeta \to 1,$$

se concluye el resultado. $\square$

Ejemplo. Consideramos el llamado método de Adams-Bashforth de dos pasos,

$$y_{n+2} - y_{n+1} = \frac{h}{2}(3f_{n+1} - f_n).$$

Al aplicar el teorema convendrá expresar las cosas en términos de la variable $\mu = \zeta - 1$. Tenemos que

$$\begin{aligned}
\rho(\zeta) - \sigma(\zeta) \log \zeta &= (\zeta^2 - \zeta) - \left(\tfrac{3}{2}\zeta - \tfrac{1}{2}\right) \log \zeta \\
&= (\mu + \mu^2) - \left(1 + \tfrac{3}{2}\mu\right) \left(\mu - \tfrac{1}{2}\mu^2 + \tfrac{1}{3}\mu^3 + \cdots\right) \\
&= \tfrac{5}{12}\mu^3 + O(\mu^4);
\end{aligned}$$

el método es consistente de orden 2.

Como en el caso de los métodos de Runge-Kutta, la consistencia es necesaria para la convergencia.

Teorema 5.4. *Si un método es convergente, entonces es consistente de orden al menos 1.*

Prueba. Si es convergente, en particular lo es para el problema

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

cuya solución es $y(x) = 1$. El método lineal multipaso (5.3) aplicado a este problema produce la ecuación en diferencias

$$\sum_{j=0}^{k} \alpha_j y_{n+j} = 0, \qquad n = 0, \dots, N-k. \tag{5.6}$$

Tomamos valores de arranque $y_n = 1$, $n = 0, \dots, k-1$. Por ser el método convergente tenemos que $\lim\limits_{h \to 0^+} y_n = 1$, $j = k, \dots, N-k$. Así pues, pasando al límite $h \to 0^+$ en (5.6) obtenemos que

$$\sum_{j=0}^{k} \alpha_j = 0,$$

o lo que es lo mismo, $C_0 = 0$.

Para ver que C_1 también es cero consideramos el problema

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

cuya solución es $y(x) = x$. El método, aplicado a este problema, se reduce a

$$\sum_{j=0}^{k} \alpha_j y_{n+j} = h \sum_{j=0}^{k} \beta_j,$$

que es una ecuación en diferencias no homogénea. Es fácil ver que una solución de esta ecuación es la sucesión $y_n = nhM$, $n = 0, 1, \dots$, donde

$$M = \frac{\sum\limits_{j=0}^{k} \beta_j}{\sum\limits_{j=0}^{k} j\alpha_j}.$$

(Veremos en breve que para un método convergente $\sum_{j=0}^{k} j\alpha_j \neq 0$, de manera que M está bien definido.) Si tomamos como valores de arranque $y_i(h) = ihM$, $i = 0, \ldots, k-1$, entonces el método producirá la sucesión $y_n = nhM = x_n M$. Por tanto,

$$y(x_n) - y_n = x_n - x_n M = x_n(1 - M),$$

y para que haya convergencia se tiene que tener que $M = 1$, es decir

$$\sum_{j=0}^{k} \beta_j = \sum_{j=0}^{k} j\alpha_j,$$

o lo que es lo mismo, $C_1 = 0$. □

Observación. En términos del primer y segundo polinomio característico la condición de consistencia de orden al menos 1 se escribe

$$\rho(1) = 0, \qquad \rho'(1) = \sigma(1).$$

5.3 0-estabilidad

La consistencia de orden 1 es necesaria para la convergencia. ¿Es también suficiente? Veamos que no con un ejemplo.

Consideramos el método

$$y_{n+2} + y_{n+1} - 2y_n = h(5f_{n+1} - 2f_n),$$

que es consistente de orden 1. Se lo aplicamos al problema más simple que uno pueda imaginar,

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

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

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

La solución general de esta ecuación en diferencias se puede hallar fácilmente. Se buscan soluciones de la forma $y_n = \lambda^n$, y se tiene que λ debe verificar que $\lambda^2 + \lambda - 2 = 0$. Así λ debe ser 1 o -2, lo que da lugar a las soluciones $y_n = 1$, $y_n = (-2)^n$. La solución general del problema será una combinación lineal de estas dos soluciones:

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

Consideramos la solución

$$y_n = h(-2)^n.$$

Obsérvese que

$$y_0 = h \to 0 = y(x_0), \quad y_1 = -2h \to 0 = y(x_0) \qquad \text{cuando } h \to 0^+.$$

Sin embargo,

$$\|y(x_N) - y_N\| = h2^N = \frac{b-a}{N}2^N \to \infty.$$

Así pues, al contrario de lo que sucede para métodos de un paso, en el caso de métodos multipaso la consistencia de un método no garantiza su convergencia.

La divergencia se debe a la inestabilidad del método numérico, es decir, de la ecuación en diferencias. La solución numérica con datos de arranque $y_0 = y_1 = 0$ coincide con la teórica; sin embargo, una pequeña perturbación en los datos hace que la solución numérica explote.

Para que un método numérico sea de alguna utilidad debería ser estable, imitando la estabilidad del problema (1.1) al cual intenta aproximar (recordemos que, bajo las condiciones del teorema de existencia y unicidad, las soluciones de (1.1) dependen de forma continua de los datos).

Definición 5.5. *Un método lineal multipaso se dice* 0-estable *si para cada problema* (1.1) *con f Lipschitz, y para cada dos sucesiones* $\{u_n\}_{n=0}^N$ *y* $\{v_n\}_{n=0}^N$

satisfaciendo para $0 \leq n \leq N-k$

$$\begin{aligned}\sum_{j=0}^{k} \alpha_j u_{n+j} - h\sum_{j=0}^{k} \beta_j f(x_{n+j}, u_{n+j}) = h\delta_n,\\ \sum_{j=0}^{k} \alpha_j v_{n+j} - h\sum_{j=0}^{k} \beta_j f(x_{n+j}, v_{n+j}) = h\gamma_n,\end{aligned}$$

se tiene que

$$\max_{k\leq n\leq N} \|u_n - v_n\| = O\left(\max_{0\leq j\leq k-1} \|u_j - v_j\| + \max_{0\leq j\leq N-k} \|\delta_j - \gamma_j\|\right).$$

La 0-estabilidad es la versión discreta de la dependencia continua de los datos.

Teorema 5.6. *Si un método lineal multipaso es consistente de orden p y 0-estable, entonces es convergente de orden p.*

Prueba. Por un lado tenemos que

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

y por otro que

$$\sum_{j=0}^{k} \alpha_j y(x_{n+j}) = h\sum_{j=0}^{k} \beta_j f(x_{n+j}, y(x_{n+j})) + h\tau_n,$$

donde $\tau_n = R_n/h$. Pero el método es 0-estable, así que

$$\max_{k\leq n\leq N} \|y(x_n) - y_n\| = O\left(\max_{0\leq j\leq k-1} \|y(x_j) - y_j\| + \max_{0\leq n\leq N-k} \|\tau_n\|\right)$$

cuando $h \to 0^+$. Como el método es consistente de orden p, se tiene que $\|\tau_n\| = O(h^p)$. Concluimos que

$$\max_{k\leq n\leq N} \|y(x_n) - y_n\| = O(h^p) \qquad \text{si} \max_{0\leq j\leq k-1} \|y(x_j) - y_j\| = O(h^p);$$

es decir, el método es convergente de orden p. $\square$

Nos preguntamos ahora que condiciones debe satisfacer un método lineal multipaso para ser 0-estable.

Definición 5.7. *Se dice que un método lineal multipaso satisface la* condición de la raíz *si todas las raíces del primer polinomio característico tienen módulo menor o igual que 1, y aquellas que tienen módulo 1 son simples.*

Observación. Si un método es consistente, por el Teorema 5.3 una de las raíces de $\rho(\zeta)$ debe ser 1. Ésta es la *raíz principal*, a la que se etiqueta como ζ_1. Las restantes raíces, ζ_i, $i = 2, \dots, k$, son las llamadas *raíces espurias*; surgen porque decidimos representar un sistema diferencial de primer orden por un sistema en diferencias de orden k. Obviamente, los métodos lineales multipaso con $k = 1$ no tienen raíces espurias, y satisfacen por tanto la condición de la raíz.

Teorema 5.8. *Un método lineal multipaso es* 0*-estable si y sólo si satisface la condición de la raíz.*

Prueba. Consideramos el problema (1.1) con $f = 0$ y tomamos dos soluciones u_n y v_n de la ecuación en diferencias sin perturbar

$$\sum_{j=0}^{k} \alpha_j u_{n+j} = 0. \tag{5.7}$$

Si el método es 0-estable, existe una constante C tal que

$$\max_{k \le j \le N} \|u_n - v_n\| \le C \max_{0 \le j \le k-1} \|u_j - v_j\|.$$

Veamos que esto es imposible si no se satisface la condición de la raíz.

(i) Supongamos que existe ξ tal que $|\xi| > 1$ y $\rho(\xi) = 0$. Las sucesiones $u_n = 0$ y $v_n = h\xi^n$ son solución de la ecuación en diferencias (5.7). Tenemos que

$$\|u_N - v_N\| = h|\xi|^N = \frac{b-a}{N}|\xi|^N \to \infty,$$

mientras que,

$$\max_{0 \le j \le k-1} \|u_j - v_j\| = \max_{0 \le j \le k-1} h|\xi|^j \le \frac{b-a}{N}|\xi|^{k-1}.$$

Así pues, el método no es 0-estable.

(ii) Supongamos que existe ξ tal que $|\xi| = 1$, $\rho(\xi) = 0 = \rho'(\xi)$. Las sucesiones $u_n = 0$ y $v_n = \sqrt{h}n\xi^n$ son solución de la ecuación en diferencias (5.7) (nótese que ξ es raíz doble). Entonces

$$\|u_N - v_N\| = \sqrt{h}N|\xi|^N = \sqrt{b-a}\sqrt{N} \to \infty,$$

mientras que

$$\max_{0\le j\le k-1} \|u_j - v_j\| = \max_{0\le j\le k-1} \sqrt{h}j|\xi|^j \le \frac{\sqrt{b-a}}{\sqrt{N}}(k-1),$$

y el método no es 0-estable.

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

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

$$\begin{aligned} &Y_n = (y_{n+k-1}, \dots, y_n)^T, && n = 0, \dots, N-k+1, \\ &F(Y_n) = (\textstyle\sum_{j=0}^k \beta_j f_{n+j}, 0, \dots, 0)^T, && n = 0, \dots, N-k+1. \end{aligned}$$

Nótese que estamos diciendo que

$$\eta = \sum_{j=0}^{k} \beta_j f(x_{n+j}, y_{n+j})$$

es función sólo de $y_n, \dots, y_{n+k-1}$, y no de y_{n+k} (los valores x_n se consideran fijos). Para verlo observamos que

$$\eta = \sum_{j=0}^{k-1} \beta_j f(x_{n+j}, y_{n+j}) + \beta_k f\left(x_{n+k}, h\eta - \sum_{j=0}^{k-1} \alpha_j y_{n+j}\right),$$

ecuación implícita que define a η como función de $y_n, \dots, y_{n+k-1}$.

Con esta notación podemos escribir el método como

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

donde A es la *matriz compañera* del primer polinomio característico $\rho(\zeta) = \alpha_0 + \alpha_1\zeta + \cdots + \alpha_{k-1}\zeta^{k-1} + \zeta^k$,

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

Iterando el proceso tenemos que

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

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

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

y restando tendremos que

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

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

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

tomando normas y utilizando que f es Lipschitz con constante L tenemos que

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

siendo B es una cota de los módulos de los β_j, desigualdad que se puede escribir como

$$\phi_n \leq A + hLMB\sum_{l=0}^{n-1} \phi_l,$$

definiendo ϕ_n y A de la manera obvia.

Sea $\psi_n = A + hLMB \sum_{l=0}^{n-1} \phi_l$. Se tiene que

$$\psi_{n+1} - \psi_n = hLMB\phi_n \leq hLMB\psi_n,$$

de donde concluimos que

$$\phi_n \leq \psi_n \leq (1 + hLMB)^{n-1}\psi_1 \leq e^{(n-1)hLMB}(A + \|U_0 - V_0\|),$$

y de aquí la 0-estabilidad.

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

Finalmente nos preguntamos si es cierto el recíproco del Teorema 5.6.

Teorema 5.9 (Teorema de equivalencia de Dahlquist). *Un método lineal multipaso es convergente si y sólo si es consistente de orden al menos 1 y* 0-*estable.*

Prueba. Por los teoremas 5.4, 5.6 y 5.8, basta con probar que

$$\text{convergencia} \Rightarrow \text{ criterio de la raíz.}$$

Supongamos que no se satisface el criterio de la raíz. Consideramos el problema

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

cuya solución es $y(x) = 0$.

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

$$\|y(x_N) - y_N\| = h|\xi|^N = \frac{b-a}{N}|\xi|^N \to \infty \quad \text{cuando } h \to 0^+,$$

y el método no es convergente.

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

$$\|y(x_N) - y_N\| = \sqrt{h}N|\xi|^N = \sqrt{b-a}\sqrt{N} \to \infty \quad \text{cuando } h \to 0^+,$$

y el método no es convergente. □

¿Cuál es el mejor orden que podemos conseguir para un método, para un número de pasos k dado? Puesto que un método lineal de k pasos viene determinado por $2k+1$ coeficientes, y las condiciones de orden de consistencia son relaciones lineales de estos coeficientes, existen métodos lineales de k pasos con orden de consistencia $p = 2k$; reciben el nombre de *maximales*. Sin embargo no sirven para nada, puesto que no son convergentes.

Teorema 5.10 (Primera barrera de Dahlquist (1956)). *El orden de convergencia p de un método lineal de k pasos 0-estable satisface:*

- $p \le k+2$ *si* k *es par;*
- $p \le k+1$ *si* k *es impar;*
- $p \le k$ *si* $\beta_k/\alpha_k \le 0$ *(en particular si es explícito).*

La prueba se puede encontrar en [HNW].

Veremos en el próximo capítulo que los métodos con $p = k+2$, llamados *optimales*, tampoco funcionan bien en la práctica. Así pues, el mejor orden que se puede alcanzar con un método de k pasos que sirva para algo es $p = k+1$.

5.4 Comentarios y bibliografía

La prueba del Teorema de Equivalencia y el establecimiento de la primera barrera por Germund Dahlquist, en 1956 y 1959 respectivamente, fueron hitos importantes en la historia del análisis numérico. No sólo son resultados de gran importancia intrínseca, sino que sirvieron para dar al análisis numérico el rango de disciplina matemática *de verdad*, e impulsaron el rigor matemático en el pensamiento numérico. La monografía clásica de Henrici [He], dedicada a explicar los resultados de Dahlquist, sigue siendo un modelo de exposición clara y bonita.

Los libros [I] y [JR] tratan el tema de forma mucho más sencilla (también menos completa). Un poco más avanzado resulta [L], donde se estudia el Teorema de equivalencia (sin dar todas las pruebas) en una forma más general, aplicable a métodos multipaso no necesariamente lineales. Ese mismo punto de vista se adopta en el también clásico [IK]. De nuevo es en [HNW] donde podemos encontrar más información. Además de pruebas completas (aunque a veces diferentes) de todo el material que presentamos, contiene información adicional sobre, por ejemplo, los métodos de Nordsieck (una clase de métodos lineales multipaso especialmente adecuada para cambiar la longitud del paso) o aspectos prácticos de programación.

En [L] se puede encontrar todo un capítulo dedicado a los pares predictor-corrector, de los que aquí apenas hemos dicho lo que son.

Los métodos lineales multipaso de hoy en día no sólo tienen paso variable, sino también *orden variable*. Son los llamados algoritmos VSVO (Variable Step Variable Order). En [L] se explica la estructura de este tipo de algoritmos. El lector interesado en conocer detalles sobre algoritmos concretos podrá encontrarlos por ejemplo en [G] y [SG].

Bibliografía

[A1] Atkinson, K.E. (1989), *An Introduction to Numerical Analysis* (2nd ed.), John Wiley & Sons, New York.

[A2] Atkinson, K.E. (1993), *Elementary Numerical Analysis* (2nd ed.), John Wiley & Sons, New York.

[ABD] Aubanell, A.; Bensey, A.; Deshalms, A. (1993) *Útiles básicos del cálculo numérico*, Labor, Barcelona.

[Ax] Axelsson, O. (1994), *Iterative Solution Methods*, Cambridge University Press, Cambridge.

[AB] Axelsson, O.; Barker, V.A. (1984), *Finite Element Solution of Boundary Value Problems: Theory and Computation*, Academic Press, Orlando, FL.

[BCO] Becker, E.B.; Carey, G.F.; Oden, J.T. (1981), *Finite elements. An Introduction. Volume I*, Prentice Hall, New Jersey.

[B] Brézis, H. (1984), *Análisis Funcional*, Alianza Editorial, Madrid.

[BR] Briggs, W.L. (1987), *A Multigrid Tutorial*, SIAM, Philadelphia.

[BD] Boyce, W.E.; DiPrima, R.C. (1986), *Elementary Differential Equations and Boundary Value Problems* (4th ed.), Wiley, New York.

[BS] Brenner, S.C.; Scott, L.R. (1994), The Mathematical Theory of Finite Element Methods, Springer-Verlag, New-York.

[BF] Burden, R.L.; Faires, J.D. (1998), *Análisis Numérico* (6a ed.), International Thompson Editores, Madrid.

[Bu] Butcher, J.C. (1987), *The Numerical Analysis of Ordinary Differential Equations*, John Wiley, New York.

[CO1] Carey, G.F.; Oden, J.T. (1983), *Finite elements: A Second Course. Volume II*, Prentice Hall, New Jersey.

[CO2] Carey, G.F.; Oden, J.T. (1984), *Finite elements: Computational Aspects. Volume III*, Prentice Hall, New Jersey.

[CO3] Carey, G.F.; Oden, J.T. (1986), *Finite elements: Fluid Mechanics. Volume VI*, Prentice Hall, New Jersey.

[C] Ciarlet, P.G. (1978), *The Finite Element Method for Elliptic Problems*, North-Holland, Amsterdam.

[CB] Conte, S.D.; de Boor, C. (1981), *Elementary Numerical Analysis. An algorithmic approach* (3rd ed.), McGraw-Hill, Singapore.

[CM] Crouzeix, M.; Mignot, A.L. (1986), *Exercises d'analyse numérique des équations différentielles*, Masson, Paris.

[DB] Dahlquist, G.; Björk, Å. (1974), *Numerical Methods*, Prentice Hall, Englewood Cliffs, NJ.

[Da] Datta, B.N. (1995), *Numerical Linear Algebra and Applications*, Brooks/Cole, Pacific Grove CA, USA.

[DV] Dekker, K.; Verwer, J.G. (1984), *Stability of Runge-Kutta Methods for Stiff Nonlinear Differential Equations*, North-Holland, Amsterdam.

[DER] Duff, I.S.; Erisman, A.M.; Reid, J.K. (1986), *Direct Methods for Sparse Matrices*, Oxford University Press, Oxford.

[F] Fletcher, R. (1987), *Practical Methods of Optimization* (2nd ed.), Wiley, London.

[G] Gear, C.W. (1971), *Numerical Initial Value Problems in Ordinary Differential Equations*, Prentice-Hall, Englewood Cliffs, NJ.

[GL] George, A.; Liu, J.W.-H. (1981), *Computer Solution of Large Sparse Positive Definite Systems*, Prentice-Hall, Englewood Cliffs, NJ.

[GT] Gilbarg, D.; Trudinger, N.S. (1983), *Elliptic Partial Differential Equations of the Second Order* (2a ed.), Springer-Verlag.

[Go] Goldstine, H.H. (1977), *A History of Numerical Analysis*, Springer-Verlag, Berlin.

[GV] Golub, G.H.; Van Loan, C.F. (1996) *Matrix Computations* (3rd ed.), John Hopkins U. Press, Baltimore.

[H] Hackbusch, W. (1992), *Elliptic Differential Equations: Theory and Numerical Treatment*, Springer-Verlag, Berlin.

[H2] Hackbusch, W. (1985), *Multigrid Methods and Applications*, Springer-Verlag, Berlin.

[HY] Hageman, L.A.; Young, D.M. (1981), *Applied Iterative Methods*, Academic Press, New York.

[HNW] Hairer, E.; Nørsett, S.P.; Wanner, G. (1991), *Solving Ordinary Differential Equations I: Nonstiff Problems* (2nd ed.), Springer-Verlag, Berlin.

[HW] Hairer, E.; Wanner, G. (1991), *Solving Ordinary Differential Equations II: Stiff Problems and Differential-algebraic Equations*, Springer-Verlag, Berlin.

[HH] Hämmerlin, G.; Hoffman, K.-H. (1991), *Numerical Mathematics*, Springer-Verlag, Berlin.

[He] Henrici, P. (1962), *Discrete Variable Methods in Ordinary Differential Equations*, Wiley, New York.

[Heu] Heun, K. (1900), *Neue Methode zur approximativen Integration der Differentialgleichungen einer unabhängigen Veränderlichen*, Z. Math. Phys. **45**, 23–38.

[I] Iserles, A. (1996), *A First Course in the Numerical Analysis of Differential Equations*, Cambridge University Press, Cambridge.

[IK] Isaacson, E.; Keller, H.B. (1966), *Analysis of Numerical Methods*, John Wiley & Sons, New York.

[J] Johnson, C. (1987), *Numerical Solution of Partial Differential Equations by the Finite Element Method*, Cambridge University Press, Cambridge.

[JR] Johnson, L.W.; Riess, R.D. (1982), *Numerical Analysis*, Addison Wesley, London.

[K] Kelley, C.T. (1995), *Iterative Methods for Linear an Nonlinear Equations*, SIAM, Philadelphia.

[KC] Kincaid, D.; Cheney, W. (1996), *Numerical Analysis*, Brooks/Cole, Pacific Grove CA, USA.

[Ku] Kutta, W. (1901), *Beitrag zur näherungsweisen Integration totaler Differentialgleichungen*, Z. Math. Phys. **46**, 435–453.

[L] Lambert, J.D. (1991), *Numerical Methods for Ordinary Differential Systems*, Wiley, London.

[MF] Mathews, J.H.; Fink, K.D. (2000), *Métodos Numéricos con Matlab*, Prentice Hall, Madrid.

[MG] Mitchell, A.R.; Griffiths, D.F. (1980), *The Finite Difference Method in Partial Differential Equations*, Wiley, London.

[MW] Mitchell, A.R.; Wait, R. (1977), *The Finite Element Method in Partial Differential Equations*, Wiley, London.

[M] Moreno González, C. (1999), *Cálculo Numérico II*, Universidad Nacional de Educación a Distancia, Madrid.

[N] Nevanlinna, O. (1993), *Convergence of Iterations for Linear Equations*, Birkhäuser, Basel.

[OC1] Oden, J.T.; Carey, G.F. (1983), *Finite elements: Mathematical Aspects. Volume IV*, Prentice Hall, New Jersey.

[OC2] Oden, J.T.; Carey, G.F. (1984), *Finite elements: Special Problems in Solid Mechanics. Volume V*, Prentice Hall, New Jersey.

[OR] Ortega, J.M.; Rheinholdt, W.C. (1970), *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, New York.

[P] Peral, I. (1995), *Primer Curso de Ecuaciones en Derivadas Parciales*, Addison-Wesley/U. Autónoma de Madrid, Madrid.

[Pi] Pironneau, O. (1989), *Finite Element Methods for Fluids*, John Wiley & Sons, Chichester.

[Po] Pozrikidis, C. (1988), *Numerical Computation in Science and Engineering*, Oxford University Press, Oxford.

[QV] Quarteroni, A.; Valli, A. (1997), *Numerical Approximation of Partial Differential Equations* (2nd ed.), Springer-Verlag, Berlin.

[RT] Rabier, P.; J.-M. Thomas (1993), *Exercises d'analyse numérique des équations aux dérivées partielles*, Masson, Paris.

[RM] Richtmyer, R.D.; Morton, K.W. (1967), *Difference Methods for Initial-Value Problems*, Interscience, New York.

[R] Runge, C. (1895), *Ueber die numerische Auflösung von Differentialgleichungen*, Math. Ann. **46**, 167–178.

[S] Sanz-Serna, J.M. (1998), *Diez lecciones de cálculo numérico*, Secretariado de Publicaciones e Intercambio Científico, Universidad de Valladolid.

[SC] Sanz-Serna, J.M.; Calvo, M.P. (1994), *Numerical Hamiltonian Problems*, Chapman-Hall, London.

[SG] Shampine, L.F.; Gordon, M.K. (1975), *Computer Solution of Ordinary Differential Equations*, W.H. Freeman, San Francisco.

[Sc] Schwartz, H.R. (1998), *Finite Element Methods*, Academic Press, London.

[Sc2] Schwartz, H.R. (1989), *Numerical Analysis. A Comprehensive Introduction*, John Wiley & Sons, Chichester.

[Si] Simmons, G.F. (1993), *Ecuaciones Diferenciales con Aplicaciones y Notas Históricas* (2a ed.), McGraw-Hill, Madrid.

[Ste] Stewart, G. (1996) *Afternotes on Numerical Analysis*, SIAM, Philadelphia.

[SB] Stoer, J.; Bulirsch, R. (1993), *Introduction to Numerical Analysis* (2nd ed.), Springer-Verlag, Berlin.

[Str] Strang, G. (1982), *Algebra lineal y sus aplicaciones*, Fondo Educativo Interamericano, México.

[SF] Strang, G., Fix, G.J. (1973), *An Analysis of the Finite Element Method*, Prentice-Hall, Englewood Cliffs, NJ.

[St] Strikwerda, J.C. (1989), *Finite Difference Schemes and Partial Differential Equations*, Brooks/Cole Publishing, Pacific Grove, CA.

[T] Thomas, J.W. (1995), *Numerical Partial Differential Equations*, Springer-Verlag, Berlin.

[TB] Trefethen, L.N., Bau, D. III (1997), *Numerical Linear Algebra*, SIAM, Philadelphia.

[V] Varga, R.S. (1962), *Matrix Iterative Analysis*, Prentice-Hall, Englewood Cliffs, NJ.

[W] Wesseling, P. (1992), *An Introduction to Multigrid Methods*, Wiley, Chichester.

[Z] Zienkiewicz, O.C. (1977), *The finite Element Method* (3rd ed.), McGraw-Hill, London.

[ZM] Zienkiewicz, O.C.; Morgan, K. (1983), *Finite Elements and Approximation*, John Wiley & Sons, New York.