Análisis Numérico III

Apuntes
Curso Código 525442

Dr. Raimund Bürger
Profesor Titular
Departamento de Ingeniería Matemática
& Centro de Investigación en Ingeniería Matemática (CI^2MA)
Facultad de Ciencias Físicas y Matemáticas
Universidad de Concepción
Casilla 160-C
Concepción, Chile

15 de abril de 2020

Índice general

Literatura

1. H. Alder & E. Figueroa, *Introducción al Análisis Numérico*, Fac. de Cs. Físicas y Matemáticas, UdeC, 1995.
2. K. Burrage & W.H. Hundsdorfer, The order of algebraically stable Runge-Kutta methods, *BIT* **27** (1987) 62–71.
3. M.C. Bustos, M. Campos & G.N. Gatica, *Análisis Numérico II*, Dirección de Docencia, UdeC, 1995.
4. P. Deuflhard & F. Bornemann, *Scientific Computing with Ordinary Differential Equations*, Springer-Verlag, New York, 2002.
5. K. Graf Finck von Finckenstein. *Einführung in die Numerische Mathematik, Band 2*, Carl Hanser Verlag, München 1978.
6. W. Gautschi, *Numerical Analysis: An Introduction*, Birkhäuser, Boston, 1997.
7. R.D. Grigorieff, *Numerik gewöhnlicher Differentialgleichungen*, Vol. I, Teubner, Stuttgart, 1972.
8. M.H. Holmes, *Introduction to Numerical Methods in Differential Equations*, Springer-Verlag, New York, 2007.
9. W. Hundsdorfer & J.G. Verwer, *Numerical Solution of Time-Dependent Advection-Diffusion-Reaction Equations*, Springer-Verlag, Berlin, 2003.
10. P. Kaps & P. Rentrop, *Genralized Runge-Kutta methods of order 4 with stepsize control for stiff ODEs*, *Numer. Math.* **33** (1979) 55–68.
11. P. Knabner & L. Angermann, *Numerical Methods for Elliptic and Parabolic Differential Equations*, Springer-Verlag, New York, 2003.
12. S. Larsson & V. Thomée, *Partial Differential Equations with Numerical Methods*, Springer-Verlag, Berlin, 2003.
13. R.J. Le Veque, *Finite Difference Methods for Ordinary and Partial Differential Equations*, SIAM, Philadelphia, USA, 2007.
14. K.W. Morton & D.F. Mayers, *Numerical Solution of Partial Differential Equations*, Second Edition, Cambridge University Press, Cambridge, UK, 2005.
15. R.D. Richtmyer & K.W. Morton, *Difference Methods for Initial-Value Problems*, Wiley, New York, 1967.
16. J. Stoer & R. Bulirsch, *Numerische Mathematik 2*, Third Edition, Springer-Verlag, Berlin, 1990.
17. J.W. Thomas, *Numerical Partial Differential Equations. Finite Difference Methods*, Second Corrrected Printing, Springer-Verlag, Nueva York, 1998.
18. W. Törnig & P. Spellucci, *Numerische Mathematik für Ingenieure und Physiker. Band 2: Numerische Methoden der Analysis*, Second Ed., Springer-Verlag, Berlin, 1990.

Capítulo 1

Problemas de valores iniciales para ecuaciones diferenciales ordinarias (Parte I)

En este capítulo tratamos la solución numérica de sistemas de ecuaciones diferenciales ordinarias (EDOs) de primer orden. En general, un tal sistema está dado por

$$\begin{aligned} y_1' &= f_1(x, y_1, \ldots, y_n), \\ y_2' &= f_2(x, y_1, \ldots, y_n), \\ &\vdots \\ y_n' &= f_n(x, y_1, \ldots, y_n), \end{aligned} \tag{1.1}$$

Cada sistema de funciones

$$y_1 = y_1(x), \ldots, y_n = y_n(x), \quad y_i \in C^1(a, b), \quad i = 1, \ldots, n$$

que satisface (1.1) idénticamente se llama *solución* de (1.1). En general, se consideran problemas de valores iniciales:

$$y_i' = f_i(x, y_1, \ldots, y_n), \quad y_i(a) = y_a^i, \quad i = 1, \ldots, n. \tag{1.2}$$

No trataremos aquí problemas de existencia y unicidad de soluciones de (1.2), lo cual es tópico de cursos especializados de EDOs. Usaremos la siguiente notación compacta:

$$\boldsymbol{y} := \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad \boldsymbol{y}_a := \begin{pmatrix} y_a^1 \\ \vdots \\ y_a^n \end{pmatrix}, \quad \boldsymbol{f}(x, \boldsymbol{y}) = \begin{pmatrix} f_1(x, y_1, \ldots, y_n) \\ \vdots \\ f_n(x, y_1, \ldots, y_n) \end{pmatrix} = \begin{pmatrix} f_1(x, \boldsymbol{y}) \\ \vdots \\ f_n(x, \boldsymbol{y}) \end{pmatrix}.$$

Entonces (1.2) se escribe como

$$\boldsymbol{y}' = \boldsymbol{f}(x, \boldsymbol{y}), \quad \boldsymbol{y}(a) = \boldsymbol{y}_a. \tag{1.3}$$

Cada problema de valores iniciales de una ecuación diferencial ordinaria de orden n puede ser reducido a un problema del tipo (1.2) o (1.3). Para un problema de valores iniciales de una ecuación diferencial ordinaria dada por

$$y^{(n)} := \frac{\mathrm{d}^n y}{\mathrm{d}x^n} = f\big(x, y, y', \ldots, y^{(n-1)}\big), \quad y^{(j)}(a) = y_a^j, \quad j = 0, \ldots, n-1, \tag{1.4}$$

podemos identificar la j-ésima derivada $y^{(j)}$ de la función escalar $y = y(x)$ con la $(j+1)$-ésima componente y_{j+1} de un vector $\boldsymbol{y}(x) = (y_1(x), \ldots, y_n(x))^{\mathrm{T}}$ de n funciones escalares,

$$y^{(j)} = y_{j+1}, \quad j = 0, \ldots, n-1,$$

y así convertir (1.4) al siguiente sistema de n ecuaciones diferenciales ordinarias escalares de primer orden:

$$\begin{aligned} y_1' &= y_2, \\ y_2' &= y_3, \\ &\vdots \\ y_{n-1}' &= y_n, \\ y_n' &= f(x, y_1, \ldots, y_n), \end{aligned}$$

es decir, obtenemos un sistema de ecuaciones diferenciales ordinarias del tipo (1.3) con

$$\boldsymbol{y}(a) = \begin{pmatrix} y_a^0 \\ y_a^1 \\ \vdots \\ y_a^{n-1} \end{pmatrix}, \quad \boldsymbol{f}(\boldsymbol{y}) = \begin{pmatrix} y_2 \\ \vdots \\ y_n \\ f(x, y_1, \ldots, y_n) \end{pmatrix}.$$

La componente $y_1(x)$ de la solución del sistema corresponde a la solución $y(x)$ del problema de valores iniciales (1.4).

1.1. Métodos de paso simple explícitos

1.1.1. Método general. Para la aproximación numérica del problema (1.2) o (1.3), se subdivide el intervalo $[a, b]$ en N subintervalos del tamaño

$$h := \frac{b-a}{N} > 0,$$

el cual se llama *tamaño de paso*. Los puntos

$$x_i = a + ih, \quad i = 0, \ldots, N$$

se llaman *puntos de malla*; la totalidad de los puntos para un valor de $h > 0$ se llama *malla* G_h. Ahora queremos calcular aproximaciones

$$\boldsymbol{y}_i^h := \big(y_{1i}^h, \ldots, y_{ni}^h\big)^{\mathrm{T}}$$

de los valores exactos

$$\boldsymbol{y}(x_i) = \big(y_1(x_i), \ldots, y_n(x_i)\big)^{\mathrm{T}}$$

de la solución en los puntos de la malla. Para tal efecto, los métodos mas simples son *métodos de paso simple explícitos*. Estos métodos permiten la computación de $\boldsymbol{y}_{i+1}^h$ solamente de $\boldsymbol{y}_i^h$ (para un valor de h dado). Usando una función vectorial $\boldsymbol{\Phi} := (\Phi_1, \ldots, \Phi_n)^{\mathrm{T}}$, obtenemos el método general

$$\begin{aligned} \boldsymbol{y}_{i+1}^h &= \boldsymbol{y}_i^h + h\boldsymbol{\Phi}\big(x_i, \boldsymbol{y}_i^h; h\big), \quad i = 0, \ldots, N-1, \\ \boldsymbol{y}_0^h &= \boldsymbol{y}_a. \end{aligned} \tag{1.5}$$

En detalle,

$$y_{1,i+1}^h = y_{1i}^h + h\Phi_1\big(x_i, y_{1i}^h, \ldots, y_{ni}^h; h\big),$$

$$\begin{aligned} &\vdots \\ y^h_{n,i+1} &= y^h_{ni} + h\Phi_n\big(x_i, y^h_{1i}, \dots, y^h_{ni}; h\big), \\ y^h_{k0} &= y^k_a, \quad k = 1, \dots, n. \end{aligned}$$

El sistema (1.5) es un sistema de ecuaciones de diferencias. Obviamente, hay que elegir la función $\boldsymbol{\Phi}$ de tal forma que los vectores $\boldsymbol{y}^h_i$ efectivamente son aproximaciones de $\boldsymbol{y}(x_i)$, en un sentido que se especificará más abajo.

En el caso más simple, el sistema (1.5) resulta del sistema de ecuaciones diferenciales ordinarias a través de remplazar todas las derivadas por cuocientes de diferencias. El método (1.5) se llama *método de diferencias finitas de paso simple.*

1.1.2. Método de Euler. Consideremos el caso $n = 1$, o sea una ecuación diferencial ordinaria del tipo

$$y' = f(x, y). \tag{1.6}$$

Si $y = y(x)$ es una solución suficientemente suave de (1.6), tenemos la fórmula de Taylor

$$y(x+h) = \sum_{k=0}^{m} \frac{h^k}{k!} y^{(k)}(x) + \frac{h^{m+1}}{(m+1)!} y^{(m+1)}(x+\theta h), \quad \theta \in [0,1]. \tag{1.7}$$

Para $x = x_i$ y no considerando el término $\frac{h^{m+1}}{(m+1)!} y^{(m+1)}(x+\theta h)$ de (1.7), podemos calcular $y(x_i + h)$ de $y(x_i)$, dado que (omitiendo los argumentos (x) y (x, y) en el lado izquierdo y derecho, respectivamente)

$$y' = f, \tag{1.8}$$

$$y'' = f_x + f f_y, \tag{1.9}$$

$$y''' = f_{xx} + 2f_{xy} f + f_{yy} f^2 + f_x f_y + f (f_y)^2,$$

etc. Este método es poco útil, salvo en aquellos casos donde podemos desarrollar la función f en una series de potencias. Sin embargo, para $m = 1$ tenemos

$$y(x_i + h) = y(x_i) + h f\big(x_i, y(x_i)\big) + \frac{h^2}{2} y''(x_i + \theta h).$$

Despreciando el término $\frac{h^2}{2} y''(x_i + \theta h)$, llegamos al método

$$y^h_{i+1} = y^h_i + h f\big(x_i, y^h_i\big), \quad i = 0, 1, \dots, N-1. \tag{1.10}$$

Este método es un método de paso simple con

$$\Phi(x, y; h) = f(x, y),$$

o sea, la función Φ no depende de h.

Para el caso del sistema (1.3), un cálculo análogo entrega el método

$$\boldsymbol{y}^h_{i+1} = \boldsymbol{y}^h_i + h\boldsymbol{f}\big(x_i, \boldsymbol{y}^h_i\big), \quad i = 0, \dots, N-1, \tag{1.11}$$

donde

$$\boldsymbol{\Phi}(x, \boldsymbol{y}; h) = \boldsymbol{f}(x, \boldsymbol{y}).$$

En ambos casos, (1.10) y (1.11), podemos reescribir el método como

$$\frac{1}{h}\big(\boldsymbol{y}_{i+1}^h - \boldsymbol{y}_i^h\big) = \boldsymbol{f}\big(x_i, \boldsymbol{y}_i^h\big), \quad i = 0, \dots, N-1,$$

lo que ilustra la idea básica de la construcción. El método (1.10) o (1.11) se llama *método de Euler explícito* o *método de trazado poligonal*. Este método es el más simple posible. (Por supuesto, el término "método de Euler explícito" indica que existe también una versión implícita, la cual vamos a conocer más adelante.)

1.1.3. Métodos de Taylor. Los métodos del tipo Taylor son basados en la fórmula ya indicada, (1.7), donde las derivadas de y son remplazadas por evaluaciones de la función f y ciertos términos que involucran sus derivadas parciales. Por ejemplo, partiendo del desarrollo en series de Taylor truncado

$$y(x+h) = y(x) + hy'(x) + \frac{h^2}{2}y''(x) + \mathcal{O}(h^3)$$

y utilizando (1.8) y (1.9), obtenemos

$$y(x+h) = y(x) + hf\big(x, y(x)\big) + \frac{h^2}{2}\Big(f_x\big(x, y(x)\big) + f\big(x, y(x)\big) f_y\big(x, y(x)\big)\Big) + \mathcal{O}(h^3),$$

de lo cual sigue el método de Taylor de segundo orden

$$y_{i+1}^h = y_i^h + hf\big(x_i, y_i^h\big) + \frac{h^2}{2}\Big(f_x\big(x_i, y_i^h\big) + f\big(x_i, y_i^h\big) f_y\big(x_i, y_i^h\big)\Big).$$

Como mencionamos anteriormente, la gran desventaja de estos métodos es la necesidad de evaluar numéricamente (o por derivación automática o simbólica) las derivadas parciales de la función f. Salvo por casos excepcionales, no es aceptable el esfuerzo computacional requerido por tales métodos. Conoceremos ahora los métodos del tipo Runge-Kutta, que requieren solamente la evaluación de la función f misma.

1.1.4. Métodos de Runge-Kutta. Los métodos que vamos a introducir ahora no se distinguen en los casos de ecuaciones escalares y de sistemas de ecuaciones, por lo tanto los presentamos desde el principio para el caso de sistemas.

Supongamos que la solución $\boldsymbol{y} = \boldsymbol{y}(x)$ del problema de valores iniciales (1.3) es conocida en x, y queremos evaluarla en $x + h$, $h > 0$. Utilizando el Teorema Principal del Cálculo Diferencial e Integral, tenemos

$$\boldsymbol{y}(x+h) = \boldsymbol{y}(x) + h\int_0^1 \boldsymbol{y}'(x+\tau h)\,\mathrm{d}\tau. \tag{1.12}$$

La integral se aproxima por una fórmula de cuadratura, donde los nodos α_i y los pesos γ_i, $i = 1, \dots, m$, se refieren al intervalo de integración $[0, 1]$:

$$\int_0^1 \boldsymbol{g}(w)\,\mathrm{d}w \approx \sum_{i=1}^m \gamma_i \boldsymbol{g}(\alpha_i); \tag{1.13}$$

dado que la fórmula de cuadratura (1.13) debe ser exacta por lo menos para $\boldsymbol{g} \equiv$ const., obtenemos la restricción

$$\sum_{i=1}^{m} \gamma_i = 1. \tag{1.14}$$

Aproximando la integral en (1.12) por la fórmula de cuadratura (1.13), obtenemos

$$\boldsymbol{y}(x+h) \approx \boldsymbol{y}(x) + h \sum_{i=1}^{m} \gamma_i \boldsymbol{y}'(x_i + \alpha_i h). \tag{1.15}$$

Usando la ecuación diferencial $\boldsymbol{y}' = \boldsymbol{f}(x, \boldsymbol{y})$, (1.15) se convierte en

$$\boldsymbol{y}(x+h) \approx \boldsymbol{y}(x) + h \sum_{i=1}^{m} \gamma_i \boldsymbol{f}\big(x + \alpha_i h, \boldsymbol{y}(x + \alpha_i h)\big).$$

Obviamente, para $\alpha_i \neq 0$ los valores $\boldsymbol{y}(x + \alpha_i h)$ son desconocidos. Para su aproximación, usamos nuevamente el ya mencionado Teorema Principal, que esta vez entrega

$$\boldsymbol{y}(x + \alpha_i h) = \boldsymbol{y}(x) + h \int_0^{\alpha_i} \boldsymbol{y}'(x + \tau h)\, \mathrm{d}\tau, \quad i = 1, \ldots, m. \tag{1.16}$$

Nuevamente queremos aproximar las integrales en (1.16) por fórmulas de cuadratura. Para no generar una cascada infinita de nuevos valores de $\boldsymbol{y}$ desconocidos, exigimos que los nodos de las nuevas fórmulas de cuadratura sean los mismos valores $\alpha_1, \ldots, \alpha_m$ de (1.13), o sea para la aproximacoón de la integral en (1.16) sobre el intervalo $[0, \alpha_i] \subset [0, 1]$ usamos una fórmula del tipo

$$\int_0^{\alpha_i} \boldsymbol{g}(w)\, \mathrm{d}w \approx \sum_{j=1}^{m} \beta_{ij} \boldsymbol{g}(\alpha_j), \quad i = 1, \ldots, m,$$

donde los pesos β_{ij} aún están libres, y los nodos α_j dados no necesariamente deben pertencer al intervalo $[0, \alpha_i]$. Obviamente, se debe satisfacer la restricción

$$\alpha_i = \sum_{j=1}^{m} \beta_{ij}, \quad i = 1, \ldots, m. \tag{1.17}$$

Las condiciones (1.14) y (1.17) aseguran que la cuadratura es exacta para $\boldsymbol{g} \equiv$ const.

En virtud de lo anterior, un *método de Runge-Kutta de m pasos* está dado por las fórmulas

$$\boldsymbol{y}^h(x+h) = \boldsymbol{y}^h(x) + h \sum_{i=1}^{m} \gamma_i \boldsymbol{k}_i, \tag{1.18}$$

$$\boldsymbol{k}_i = \boldsymbol{f}\left(x + \alpha_i h, \boldsymbol{y}^h(x) + h \sum_{j=1}^{m} \beta_{ij} \boldsymbol{k}_j\right), \quad i = 1, \ldots, m. \tag{1.19}$$

En general, (1.19) representa un sistema de m ecuaciones no lineales para m vectores $\boldsymbol{k}_1, \ldots, \boldsymbol{k}_m$, cada uno con n componentes escalares. Incluso para una ecuación escalar, tenemos un sistema de m ecuaciones (algebraicas) de la dimensión m.

Para el caso $\alpha_1 = 0$ y $\beta_{ij} = 0$ para $j \geqslant i$, (1.19) no es intrinsicamente un sistema de ecuaciones no lineales, ya que en este caso podemos calcular $\boldsymbol{k}_2$ explícitamente de $\boldsymbol{k}_1 = \boldsymbol{f}(x, \boldsymbol{y}(x))$, $\boldsymbol{k}_3$ de $\boldsymbol{k}_1$ y $\boldsymbol{k}_2$, etc. En este caso, se habla de un *método de Runge-Kutta explícito.*

Dado que las fórmulas (1.18) y (1.19) son las mismas para todos los métodos de Runge-Kutta, pero hay una gran cantidad de coeficientes α_i, γ_i y β_{ij} propuestos en la litaratura, es muy común representar un método de Runge-Kutta a través del llamado *diagrama de Butcher:*

$$\begin{array}{c|cccc} \alpha_1 & \beta_{11} & \beta_{12} & \cdots & \beta_{1m} \\ \alpha_2 & \beta_{21} & \ddots & & \vdots \\ \vdots & \vdots & & \ddots & \vdots \\ \alpha_m & \beta_{m1} & \cdots & \cdots & \beta_{mm} \\ \hline & \gamma_1 & \gamma_2 & \cdots & \gamma_m \end{array} . \tag{1.20}$$

Ejemplo 1.1. *Para $m = 3$, los* métodos de Heun *están dados por*

$$\begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ 1/2 & 1/2 & 0 & 0 \\ 1 & -1 & 2 & 0 \\ \hline & 1/6 & 2/3 & 1/6 \end{array} \quad \text{y} \quad \begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ 1/3 & 1/3 & 0 & 0 \\ 2/3 & 0 & 2/3 & 0 \\ \hline & 1/4 & 0 & 3/4 \end{array} .$$

En detalle, combinando (1.18) *y* (1.19) *con el diagrama de Butcher, podemos, por ejemplo, escribir el segundo método como*

$$\boldsymbol{y}_{i+1}^h = \boldsymbol{y}_i^h + h\left(\frac{1}{4}\boldsymbol{k}_1^{(i)} + \frac{3}{4}\boldsymbol{k}_3^{(i)}\right),$$

donde las cantidades $\boldsymbol{k}_1^{(i)}$, $\boldsymbol{k}_2^{(i)}$ y $\boldsymbol{k}_3^{(i)}$, que corresponden a la i-ésima iteración, están dadas por

$$\begin{aligned} \boldsymbol{k}_1^{(i)} &= \boldsymbol{f}\big(x_i, \boldsymbol{y}_i^h\big), \\ \boldsymbol{k}_2^{(i)} &= \boldsymbol{f}\left(x_i + \frac{h}{3}, \boldsymbol{y}_i^h + \frac{h}{3}\boldsymbol{k}_1^{(i)}\right), \\ \boldsymbol{k}_3^{(i)} &= \boldsymbol{f}\left(x_i + \frac{2h}{3}, \boldsymbol{y}_i^h + \frac{2h}{3}\boldsymbol{k}_2^{(i)}\right). \end{aligned}$$

Ejemplo 1.2. *El* método clásico de Runge-Kutta de cuatro pasos *($m = 4$) corresponde al diagrama*

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

es decir, a la fórmula

$$\boldsymbol{y}_{i+1}^h = \boldsymbol{y}_i^h + \frac{h}{6}\big(\boldsymbol{k}_1^{(i)} + 2\boldsymbol{k}_2^{(i)} + 2\boldsymbol{k}_3^{(i)} + \boldsymbol{k}_4^{(i)}\big)$$

(o equivalentemente, al método (1.5) *con*

$$\boldsymbol{\Phi}\big(x_i, \boldsymbol{y}_i^h; h\big) = \frac{1}{6}\big(\boldsymbol{k}_1^{(i)} + 2\boldsymbol{k}_2^{(i)} + 2\boldsymbol{k}_3^{(i)} + \boldsymbol{k}_4^{(i)}\big),$$

donde

$$\begin{aligned}
\boldsymbol{k}_1^{(i)} &= \boldsymbol{f}\big(x_i, \boldsymbol{y}_i^h\big),\\
\boldsymbol{k}_2^{(i)} &= \boldsymbol{f}\left(x_i + \frac{h}{2}, \boldsymbol{y}_i^h + \frac{h}{2}\boldsymbol{k}_1^{(i)}\right),\\
\boldsymbol{k}_3^{(i)} &= \boldsymbol{f}\left(x_i + \frac{h}{2}, \boldsymbol{y}_i^h + \frac{h}{2}\boldsymbol{k}_2^{(i)}\right),\\
\boldsymbol{k}_4^{(i)} &= \boldsymbol{f}\big(x_{i+1}, \boldsymbol{y}_i^h + h\boldsymbol{k}_3^{(i)}\big).
\end{aligned}$$

1.2. Consistencia y convergencia

1.2.1. Métodos consistentes. Obviamente, queremos saber bajo qué condiciones los valores numéricos generados por un método de paso simple aproximan la solución exacta en los puntos de la malla, y queremos discutir la precisión de la aproximacón. Para tal efecto, estudiamos el problema

$$\boldsymbol{y}' = \boldsymbol{f}(x, \boldsymbol{y}), \quad \boldsymbol{y}(a) = \boldsymbol{y}_a, \tag{1.21}$$

y el método de paso simple

$$\boldsymbol{y}_{i+1}^h = \boldsymbol{y}_i^h + h\boldsymbol{\Phi}\big(x_i, \boldsymbol{y}_i^h; h\big), \quad i = 0, \ldots, N-1; \quad \boldsymbol{y}_0^h = \boldsymbol{y}_a. \tag{1.22}$$

Obviamente, los valores $\boldsymbol{y}_i^h$ son aproximaciones de los valores reales de la solución $\boldsymbol{y}(x_i)$, $i = 1, \ldots, N$, sólo si el sistema de ecuaciones de diferencias finitas (1.22), también es una aproximación del sistema de ecuaciones diferenciales (1.21). Para precisar este requerimiento, definimos lo siguiente.

Definición 1.1. *Se dice que la función vectorial* $\boldsymbol{\Phi} = \boldsymbol{\Phi}(x, \boldsymbol{y}; h)$ satisface la hipótesis de continuidad *si* $\boldsymbol{\Phi}$ *depende de forma continua de sus* $n+2$ *argumentos escalares en*

$$\mathcal{R}_{h_0} := \big\{(x, y_1, \ldots, y_n, h) \in \mathbb{R}^{n+2} \mid a \leqslant x \leqslant b,\ y_1, \ldots, y_n \in \mathbb{R},\ 0 \leqslant h \leqslant h_0\big\},$$

donde h_0 *es una constante suficientemente pequeña.*

Definición 1.2. *El esquema* (1.22) *se llama* consistente al sistema de ecuaciones diferenciales ordinarias (1.21) *si*

$$\boldsymbol{\Phi}(x, \boldsymbol{y}; 0) = \boldsymbol{f}(x, \boldsymbol{y}), \quad x \in [a, b], \quad \boldsymbol{y} \in \mathbb{R}^n.$$

En lo siguiente, se supone que el método definido a través de la función $\boldsymbol{\Phi}$ satisface la hipótesis de continuidad, y que es consistente. Si $\boldsymbol{y} = \boldsymbol{y}(x)$ es una solución exacta del sistema $\boldsymbol{y}' = \boldsymbol{f}(x, \boldsymbol{y})$, sabemos que

$$\begin{aligned}
&\lim_{h\to 0}\left\{\frac{1}{h}\big(\boldsymbol{y}(x+h) - \boldsymbol{y}(x)\big) - \boldsymbol{\Phi}\big(x, \boldsymbol{y}(x); h\big)\right\}\\
&= \lim_{h\to 0}\left\{\frac{1}{h}\big(\boldsymbol{y}(x+h) - \boldsymbol{y}(x)\big) - \boldsymbol{\Phi}\big(x, \boldsymbol{y}(x); h\big)\right\} - \big(\boldsymbol{y}'(x) - \boldsymbol{f}\big(x, \boldsymbol{y}(x)\big)\big)
\end{aligned}$$

$$= \lim_{h\to 0}\left\{\frac{1}{h}\big(\boldsymbol{y}(x+h)-\boldsymbol{y}(x)\big)-\boldsymbol{y}'(x)\right\}-\lim_{h\to 0}\Big\{\boldsymbol{\Phi}\big(x,\boldsymbol{y}(x);h\big)-\boldsymbol{f}\big(x,\boldsymbol{y}(x)\big)\Big\}=0.$$

Entonces, cada solución del sistema $\boldsymbol{y}'=\boldsymbol{f}(x,\boldsymbol{y})$ satisface asintoticamente para $h\to 0$ el sistema de ecuaciones de diferencias finitas. Debido a la continuidad de

$$\boldsymbol{\varrho}\big(x,\boldsymbol{y}(x);h\big):=\frac{1}{h}\big(\boldsymbol{y}(x+h)-\boldsymbol{y}(x)\big)-\boldsymbol{\Phi}\big(x,\boldsymbol{y}(x);h\big) \tag{1.23}$$

con respecto a h para $h\to 0$, podemos considerar cada solución del sistema de ecuaciones diferenciales ordinarias como una solución aproximada del sistema de ecuaciones de diferencias (1.22). La cantidad $\boldsymbol{\varrho}(x,\boldsymbol{y}(x);h)$ definida en (1.23) se llama *error de truncación* del método de paso simple.

Para obtener un método de gran precisión, se exige que para $h\to 0$, (1.23) desparace como h^p, con $p>0$ lo más grande posible.

Definición 1.3. *El método* (1.22) *se llama* consistente del orden p *si p es el mayor número positivo tal que para cada solución $\boldsymbol{y}(x)$ de $\boldsymbol{y}'=\boldsymbol{f}(x,\boldsymbol{y})$ suficientemente diferenciable, y para todo $x\in[a,b]$ tenemos*

$$\boldsymbol{\varrho}\big(x,\boldsymbol{y}(x);h\big)=\mathcal{O}(h^p)\quad\text{cuando } h\to 0. \tag{1.24}$$

En (1.24), $\mathcal{O}(h^p)$ denota un vector cuyos componentes son funciones de h, y cuya norma puede ser acotada por Mh^p, donde M es una constante que no depende de h, pero que puede depender de la norma considerada. Un método consistente del orden p también se llama *método del orden p.*

1.2.2. El orden de consistencia de algunos métodos de paso simple. Nos restringimos al caso $n=1$, es decir, al caso escalar, y suponemos que la función f es suficientemente diferenciable.

Ejemplo 1.3. *Analizando el método de Euler explícito, tenemos*

$$\begin{aligned}\varrho\big(x,y(x);h\big)&=\frac{1}{h}\big(y(x+h)-y(x)\big)-\Phi\big(x,y(x);h\big)\\&=\frac{1}{h}\big(y(x+h)-y(x)\big)-f\big(x,y(x)\big)=\mathcal{O}(h),\end{aligned}$$

dado que

$$\begin{aligned}y(x+h)&=y(x)+hy'(x)+\mathcal{O}(h^2)\\&=y(x)+hf\big(x,y(x)\big)+\mathcal{O}(h^2);\end{aligned}$$

por lo tanto, el método de Euler (es decir, el método de trazado poligonal) es del primer orden.

Ejemplo 1.4. *Consideremos la familia de métodos dada por*

$$\boldsymbol{\Phi}(x,\boldsymbol{y};h)=(1-c)\boldsymbol{f}(x,\boldsymbol{y})+c\boldsymbol{f}\left(x+\frac{h}{2c},\boldsymbol{y}+\frac{h}{2c}\boldsymbol{f}(x,\boldsymbol{y})\right), \tag{1.25}$$

donde $c \in \mathbb{R}$ *es un parámetro. Para* $c = 1/2$ *se recupera el* método de trazado poligonal mejorado *con la representación explícita*

$$\boldsymbol{y}_{i+1}^h = \boldsymbol{y}_i^h + \frac{h}{2}\Big[\boldsymbol{f}\big(x_i, \boldsymbol{y}_i^h\big) + \boldsymbol{f}\Big(x_{i+1}, \boldsymbol{y}_i^h + h\boldsymbol{f}\big(x, \boldsymbol{y}_i^h\big)\Big)\Big];$$

para $c = 1$ *obtenemos el* método del trazado poligonal modificado

$$\boldsymbol{y}_{i+1}^h = \boldsymbol{y}_i^h + h\boldsymbol{f}\left(x_i + \frac{h}{2}, \boldsymbol{y}_i^h + \frac{h}{2}\boldsymbol{f}\big(x_i, \boldsymbol{y}_i^h\big)\right).$$

Para el analisis del error de truncación de (1.25)*, volvemos al caso escalar y notamos que*

$$f\left(x + \frac{h}{2c}, y + \frac{h}{2c}f(x,y)\right) = f(x,y) + \frac{h}{2c}\big(f_x(x,y) + f_y(x,y)f(x,y)\big) + \mathcal{O}(h^2),$$

es decir, obtenemos el desarrollo del método en series de Taylor (con respecto a h*)*

$$\Phi\big(x, y(x); h\big) = f\big(x, y(x)\big) + \frac{h}{2}\Big(f_x\big(x, y(x)\big) + f_y\big(x, y(x)\big)f\big(x, y(x)\big)\Big) + \mathcal{O}(h^2).$$

Por otro lado, desarrollando la solución exacta *en series de Taylor obtenemos*

$$\frac{1}{h}\big(y(x+h) - y(x)\big) = f\big(x, y(x)\big) + \frac{h}{2}\Big(f_x\big(x, y(x)\big) + f_y\big(x, y(x)\big)f\big(x, y(x)\big)\Big) + \mathcal{O}(h^2),$$

es decir, usando la definición (1.23) *del error de truncación, obtenemos aquí*

$$\varrho\big(x, y(x); h\big) = \mathcal{O}(h^2) \quad \text{cuando } h \to 0.$$

Esto significa que para el caso $c \neq 0$ *todos los métodos son de segundo orden.*

Se puede demostrar analogamente ques los esquemas de Heun y el método clásico de Runge-Kutta (con $m = 4$) son de los ordenes 3 y 4, respectivamente (Tarea).

1.2.3. El orden de convergencia. A parte de la consistencia del método exigimos también que los valores $\boldsymbol{y}_i^h$ aproximan mejor los valores de la solución exacta $y(x_i)$ si se achica el tamaño de paso h. En consecuencia, queremos que los valores $\boldsymbol{y}_i^h$ converjan a los valores exactos $\boldsymbol{y}(x_i)$ cuando $h \to 0$. Un método con esta propiedad se llama *convergente*. Hay que tomar en cuenta que debido a

$$i = \frac{x_i - a}{h},$$

el número i crece cuando achicamos h.

Definición 1.4. *Un método de paso simple* (1.22) *se llama* convergente en $x \in [a, b]$ *si*

$$\lim_{\substack{h \to 0 \\ a + ih \to x \in [a,b]}} \big(\boldsymbol{y}_i^h - \boldsymbol{y}(x)\big) = 0.$$

Además, el método se llama convergente del orden p *si*

$$\boldsymbol{y}_i^h - \boldsymbol{y}(x) = \mathcal{O}(h^p), \quad h \to 0, \quad i = 1, \ldots, N,$$

donde $\mathcal{O}(h)$ *denota un vector cuyos componentes dependen de* h*.*

Resulta que para los métodos de paso simple, la consistencia junto con la propiedad adicional que se especifica en la definición abajo asegura la convergencia.

Definición 1.5. *Se dice que una función* $f = f(x, y)$ satisface una condición de Lipschitz *o* es Lipschitz continua *en un dominio* $[a, b] \ni x$ *con respecto a* $y \in I$ *si existe una constante* L *(la* constante de Lipschitz*) tal que para todo* $y^*, y^{**} \in I$ *y* $x \in [a, b]$,

$$\big|f(x, y^*) - f(x, y^{**})\big| \leqslant L|y^* - y^{**}|.$$

Más general, una función φ *satisface una condición de Lipschitz en un dominio* $\mathcal{G} \subset \mathbb{R}^n$ *con respecto a la variable* x_k *si*

$$\big|\varphi\big(x_1, \dots, x_{k-1}, x_k^{(1)}, x_{k+1}, \dots, x_n\big) - \varphi\big(x_1, \dots, x_{k-1}, x_k^{(2)}, x_{k+1}, \dots, x_n\big)\big| \leqslant L\big|x_k^{(1)} - x_k^{(2)}\big|$$

para todo $(x_1, \dots, x_{k-1}, x_k^{(1)}, x_{k+1}, \dots, x_n)^{\mathrm{T}}, (x_1, \dots, x_{k-1}, x_k^{(2)}, x_{k+1}, \dots, x_n)^{\mathrm{T}} \in \mathcal{G}$.

Ahora se supone que Φ como función de x, y y h es Lipschitz continua con respecto a y. Precisamente, sea

$$\begin{aligned} &\big|\Phi(x, y^*; h) - \Phi(x, y^{**}; h)\big| \leqslant L|y^* - y^{**}| \\ &\text{para todo } x \in [a, b],\ y^*, y^{**} \in \mathbb{R}, \text{ y } h \in [0, h_0]. \end{aligned} \tag{1.26}$$

Teorema 1.1. *Consideremos un método de paso simple dado para* $n = 1$ *mediante la función* $\Phi = \Phi(x, y; h)$. *Supongamos que*

(a) *la función* Φ *satisface la hipótesis de continuidad (Defínción 1.1),*
(b) *el método es consistente del orden* p, *y*
(c) *la función* Φ *satisface la condición de Lipschitz* (1.26).

En este caso, es método converge del orden p.

Demostración. El método es especificado por

$$y_{i+1}^h = y_i^h + h\Phi\big(x_i, y_i^h; h\big); \tag{1.27}$$

por otro lado, la suposición del orden de consistencia implica que la solución exacta $y = y(x)$ satisface

$$y(x_{i+1}) = y(x_i) + h\Phi\big(x_i, y(x_i); h\big) + \mathcal{O}(h^{p+1}). \tag{1.28}$$

Definiendo

$$\varepsilon_i := y_i^h - y(x_i),$$

restando (1.28) de (1.27) y tomando en cuenta que

$$\big|\mathcal{O}(h^{p+1})\big| \leqslant Mh^{p+1}$$

con una constante M que no depende de h, obtenemos la desigualdad

$$|\varepsilon_{i+1}| \leqslant |\varepsilon_i| + h\big|\Phi\big(x_i, y_i^h; h\big) - \Phi\big(x_i, y(x_i); h\big)\big| + Mh^{p+1}. \tag{1.29}$$

En virtud de (1.26) y

$$\varepsilon_0 = y_0^h - y(x_0) = 0,$$

obtenemos de (1.29) la desigualdad de diferencias

$$|\varepsilon_{i+1}| \leqslant (1 + hL)|\varepsilon_i| + Mh^{p+1}, \quad |\varepsilon_0| = 0. \tag{1.30}$$

Los valores ε_i satisfacen

$$|\varepsilon_i| \leqslant \eta_i, \quad i = 0, \ldots, N, \tag{1.31}$$

donde los valores η_i, $i = 0, \ldots, N$ son recursivamente definidos por la ecuación de diferencias

$$\eta_{i+1} = (1 + hL)\eta_i + Mh^{p+1}, \quad \eta_0 = 0. \tag{1.32}$$

(Para demostrar (1.31), podemos proceder por inducción, partiendo de $|\varepsilon_0| = \eta_0 = 0$; ahora si $|\varepsilon_i| \leqslant \eta_i$ para $i = 0, \ldots, k$ con $h > 0$, (1.30) y (1.32) implican

$$\begin{aligned} |\varepsilon_{k+1}| &\leqslant (1 + hL)|\varepsilon_k| + Mh^{p+1} \\ &\leqslant (1 + hL)\eta_k + Mh^{p+1} = \eta_{k+1}.) \end{aligned}$$

El problema (1.32) es un problema de valores iniciales de una ecuación de diferencias lineal de primer orden con coeficientes constantes. Su solución es análoga a la solución de una ecuación diferencial del mismo tipo, es decir es compuesta por la suma de la solución general de la ecuación homogénea más una solución particular de la ecuación no homogénea.

La solución de la ecuación homogénea

$$\bar{\eta}_{i+1} = (1 + hL)\bar{\eta}_i$$

está dada por

$$\bar{\eta}_i = C(1 + hL)^i = C \exp\big(i \ln(1 + hL)\big).$$

El planteo $\tilde{\eta}_i = \alpha = \text{const.}$ para la ecuación no homogénea nos lleva a

$$\alpha = (1 + hL)\alpha + Mh^{p+1} \Longleftrightarrow \alpha = -\frac{Mh^p}{L},$$

entonces la solución general de (1.32) es

$$\eta_i = \bar{\eta}_i + \tilde{\eta}_i = C(1 + hL)^i - \frac{Mh^p}{L};$$

usando

$$\eta_0 = 0 = C - \frac{Mh^p}{L}$$

obtenemos

$$\eta_i = \frac{Mh^p}{L}\big((1 + hL)^i - 1\big)$$

y finalmente

$$\begin{aligned} \big|y_i^h - y(x_i)\big| = |\varepsilon_i| \leqslant \eta_i &= \frac{Mh^p}{L}\big((1 + hL)^i - 1\big) \\ &\leqslant h^p \frac{M}{L}\big(\mathrm{e}^{iLh} - 1\big) \\ &\leqslant h^p \frac{M}{L}\big(\mathrm{e}^{L(b-a)} - 1\big) = \mathcal{O}(h^p). \end{aligned} \tag{1.33}$$

■

Lamentablemente es imposible acotar M, ni siquiera su orden de magnitud, en situaciones prácticas, asi que la desigualdad (1.33) es poco apta para estimar el error.

La condición de Lipschitz y el Teorema 1.1 pueden ser generalizados para sistemas de ecuaciones.

Definición 1.6. *La función* $\boldsymbol{\Phi} : \mathcal{R}_{h_0} \to \mathbb{R}^n$ satisface una condición de Lipschitz *en* $\mathcal{R}_{h_0}$ *con respecto a* $y_1, \ldots, y_n$ *si la siguiente desigualdad es válida en* $\mathcal{R}_{h_0}$ *con una constante* L*:*

$$\big\|\boldsymbol{\Phi}(x, \boldsymbol{y}; h) - \boldsymbol{\Phi}(x, \boldsymbol{y}^*; h)\big\|_2 \leqslant L\|\boldsymbol{y} - \boldsymbol{y}^*\|_2. \tag{1.34}$$

Teorema 1.2. *Consideremos un método de paso simple dado para* $n \geqslant 1$ *mediante la función vectorial* $\boldsymbol{\Phi} = \boldsymbol{\Phi}(x, \boldsymbol{y}; h)$*. Supongamos que*

(a) *la función* $\boldsymbol{\Phi}$ *satisface la hipótesis de continuidad (Defínción 1.1),*
(b) *el método es consistende del orden* p*, y*
(c) *la función* $\boldsymbol{\Phi}$ *satisface la condición de Lipschitz* (1.34)*.*

En este caso, es método converge del orden p*, o sea existe una constante* C *tal que*

$$\big\|\boldsymbol{y}_i^h - \boldsymbol{y}(x_i)\big\|_2 \leqslant Ch^p.$$

1.3. Métodos de pasos múltiples

En el caso de métodos de paso simple podemos aumentar el orden del método solamente si aumentamos el número de evaluaciones de f (o incluso de ciertas derivadas de f) en cada paso, lo que significa que el orden del método general puede ser mejorado solamente por un gran esfuerzo computacional. (El orden del método debe ser alto para permitir la solución de un problema de valores iniciales de una ecuación diferencial ordinaria con tamaños de paso largos).

Para mejorar esta situación, obervamos primero que si en los puntos $x_i, x_{i-1}, \ldots, x_{i-p}$ ya hemos generado aproximaciones $\boldsymbol{y}_i^h, \ldots, \boldsymbol{y}_{i-p}^h$ de $\boldsymbol{y}(x_i), \ldots, \boldsymbol{y}(x_{i-p})$, podemos aprovechar de toda esa información al calcular la aproximación asociada con x_{i+1}, es decir, el vector $\boldsymbol{y}_{i+1}^h$. Claramente debemos usar un valor de $p+1$, es decir, del número de evaluaciones requeridas para ejecutar el paso, moderado.

1.3.1. Métodos de pasos múltiples basados en integración numérica. Supongamos que $\boldsymbol{y} = \boldsymbol{y}(x)$ es una solución exacta de $\boldsymbol{y}' = \boldsymbol{f}(x, \boldsymbol{y})$. Entonces podemos escribir

$$\begin{aligned}\boldsymbol{y}(x_{i+1}) &= \boldsymbol{y}(x_{i-j}) + \int_{x_{i-j}}^{x_{i+1}} \boldsymbol{y}'(\tau)\,\mathrm{d}\tau \\ &= \boldsymbol{y}(x_{i-j}) + \int_{x_{i-j}}^{x_{i+1}} \boldsymbol{f}\big(\tau, \boldsymbol{y}(\tau)\big)\,\mathrm{d}\tau, \quad j \in \mathbb{N} \cup \{0\}.\end{aligned}$$

La función $\boldsymbol{f}(\tau, \boldsymbol{y}(\tau))$ bajo la integral es remplazada por el polinomio de interpolación del grado máximo p definido por los puntos

$$\big(x_i, \boldsymbol{f}\big(x_i, \boldsymbol{y}_i^h\big)\big), \ldots, \big(x_{i-p}, \boldsymbol{f}\big(x_{i-p}, \boldsymbol{y}_{i-p}^h\big)\big),$$

luego calculamos la integral exacta. Esto resulta en el método

$$\boldsymbol{y}_{i+1}^h = \boldsymbol{y}_{i-j}^h + \int_{x_{i-j}}^{x_{i+1}} \left(\sum_{k=0}^{p} \boldsymbol{f}\big(x_{i-k}, \boldsymbol{y}_{i-k}^h\big) \prod_{\substack{l=0 \\ l\neq k}}^{p} \frac{\tau - x_{i-l}}{x_{i-k} - x_{i-l}} \right) \mathrm{d}\tau; \tag{1.35}$$

si los nodos son equidistantes, podemos calcular a priori las constantes

$$\beta_{j,p,k} := \int_{x_{i-j}}^{x_{i+1}} \prod_{\substack{l=0 \\ l\neq k}}^{p} \frac{\tau - x_{i-l}}{x_{i-k} - x_{i-l}} \, \mathrm{d}\tau. \tag{1.36}$$

Así, el método de paso múltiples es dado por

$$\boldsymbol{y}_{i+1}^h = \boldsymbol{y}_{i-j}^h + \sum_{k=0}^{p} \beta_{j,p,k} \boldsymbol{f}\big(x_{i-k}, \boldsymbol{y}_{i-k}^h\big).$$

Para $x_{i+1} - x_i = h$ para todo i y $p = j = 0$, resulta el método de Euler explícito; para $j = 0$, p arbitrario obtenemos los *métodos de Adams-Bashforth*. A continuación comunicamos los valores de algunos de los coeficientes (1.36) para valores moderados de p:

$\frac{1}{h}\beta_{0,p,k}$	$k=0$	$k=1$	$k=2$	$k=3$	orden
$p=0$	1				1
$p=1$	$\frac{3}{2}$	$-\frac{1}{2}$			2
$p=2$	$\frac{23}{12}$	$-\frac{16}{12}$	$\frac{5}{12}$		3
$p=3$	$\frac{55}{24}$	$-\frac{59}{24}$	$\frac{37}{24}$	$-\frac{9}{24}$	4

.

En la práctica, no se usan mucho los métodos de Adams-Bashforth solos, a pesar de simplicidad y del alto orden que se puede lograr (por ejemplo, para $p = 3$ podemos alcanzar el orden 4, usando solamente una evaluación de $\boldsymbol{f}$ por paso). El problems consiste en el pequeño dominio de establidad de estos métodos; este aspecto se discutirá mas detalladamente más adelante.

Una alternativa (a los métodos explícitos) podemos generar definiendo métodos implícitos a través de la inclusión del punto $(x_{i+1}, \boldsymbol{f}(x_{i+1}, \boldsymbol{y}_{i+1}))$ a la interpolación de $\boldsymbol{f}(\tau, \boldsymbol{y}(\tau))$. El resultado es un método implícito que determina $\boldsymbol{y}_{i+1}^h$ mediante un sistema de ecuaciones habitualmente no lineal. En este caso, la ecuación que remplaza (1.35) es

$$\boldsymbol{y}_{i+1}^h = \boldsymbol{y}_{i-j}^h + \int_{x_{i-j}}^{x_{i+1}} \left(\sum_{k=-1}^{p} \boldsymbol{f}\big(x_{i-k}, \boldsymbol{y}_{i-k}^h\big) \prod_{\substack{l=-1 \\ l\neq k}}^{p} \frac{\tau - x_{i-l}}{x_{i-k} - x_{i-l}} \right) \mathrm{d}\tau. \tag{1.37}$$

Para $x_{i+1} - x_i = h$, $j = 0$ y todo $p = -1, 0, 1, \ldots$, obtenemos los *métodos de Adams-Moulton*. Podemos calcular a priori los coeficientes

$$\tilde{\beta}_{j,p,k} := \int_{x_{i-j}}^{x_{i+1}} \prod_{\substack{l=-1 \\ l\neq k}}^{p} \frac{\tau - x_{i-l}}{x_{i-k} - x_{i-l}} \, \mathrm{d}\tau; \tag{1.38}$$

así, el método final es de la forma

$$\boldsymbol{y}^h_{i+1} = \boldsymbol{y}^h_{i-j} + \sum_{k=-1}^{p} \tilde\beta_{j,p,k} \boldsymbol{f}\big(x_{i-k}, \boldsymbol{y}^h_{i-k}\big).$$

Para $j = 0$ y valores de p moderados resultan los siguientes valores:

$\frac{1}{h}\tilde\beta_{0,p,k}$	$k=-1$	$k=0$	$k=1$	$k=2$	orden
$p=-1$	1				1
$p=0$	$\frac{1}{2}$	$\frac{1}{2}$			2
$p=1$	$\frac{5}{12}$	$\frac{8}{12}$	$-\frac{1}{12}$		3
$p=2$	$\frac{9}{24}$	$\frac{19}{24}$	$-\frac{5}{24}$	$\frac{1}{24}$	4

.

Los métodos de Adams-Moulton incluyen para $p = -1$ el *método de Euler implícito*

$$\boldsymbol{y}^h_{i+1} = \boldsymbol{y}^h_i + h\boldsymbol{f}\big(x_{i+1}, \boldsymbol{y}^h_{i+1}\big) \tag{1.39}$$

y la *regla trapezoidal*

$$\boldsymbol{y}^h_{i+1} = \boldsymbol{y}^h_i + \frac{h}{2}\Big(\boldsymbol{f}\big(x_i, \boldsymbol{y}^h_i\big) + \boldsymbol{f}\big(x_{i+1}, \boldsymbol{y}^h_{i+1}\big)\Big). \tag{1.40}$$

Los métodos implícitos se usan poco en la práctica. Pero si usamos un método de Adams-Moulton y remplazamos el valor $\boldsymbol{f}(x_{i+1}, \boldsymbol{y}^h_{i+1})$ por una aproximación $\boldsymbol{f}(x_{i+1}, \tilde{\boldsymbol{y}}^h_{i+1})$, donde $\tilde{\boldsymbol{y}}^h_{i+1}$ es generado por un método explícito, obtenemos un método del tipo predictor-corrector. Por ejemplo, combinando el método de Adams-Moulton con 3 pasos y orden 4 con un método de Adams-Bashforth con 3 pasos y del orden 3 (como predictor), resulta un método explícito de 3 pasos y del orden 4 que requiere sólo dos evaluaciones de $\boldsymbol{f}$ por paso, y que es esencialmente equivalente en sus características al método de Runge-Kutta clásico del orden 4, y que requiere 4 evaluaciones de $\boldsymbol{f}$ por paso. Explícitamente, este método predictor-corrector es dado por

$$\tilde{\boldsymbol{y}}^h_{i+1} = \boldsymbol{y}^h_i + \frac{h}{12}\Big(23\boldsymbol{f}\big(x_i, \boldsymbol{y}^h_i\big) - 16\boldsymbol{f}\big(x_{i-1}, \boldsymbol{y}^h_{i-1}\big) + 5\boldsymbol{f}\big(x_{i-2}, \boldsymbol{y}^h_{i-2}\big)\Big), \tag{1.41}$$

$$\boldsymbol{y}^h_{i+1} = \boldsymbol{y}^h_i + \frac{h}{24}\Big(9\boldsymbol{f}\big(x_{i+1}, \tilde{\boldsymbol{y}}^h_{i+1}\big) + 19\boldsymbol{f}\big(x_i, \boldsymbol{y}^h_i\big) - 5\boldsymbol{f}\big(x_{i-1}, \boldsymbol{y}^h_{i-1}\big) + \boldsymbol{f}\big(x_{i-2}, \boldsymbol{y}^h_{i-2}\big)\Big). \tag{1.42}$$

1.3.2. Métodos implícitos de pasos múltiples basados en derivación numérica. Usamos la identidad

$$\boldsymbol{y}'(x_{i+1}) = \boldsymbol{f}\big(x_{i+1}, \boldsymbol{y}(x_{i+1})\big)$$

y aproximamos $\boldsymbol{y}(x)$ por el polinomio de interpolación $\boldsymbol{P}(x; i)$ del grado máximo $k+1$ para los datos

$$\big(x_{i+1}, \boldsymbol{y}^h_{i+1}\big), \dots, \big(x_{i-k}, \boldsymbol{y}^h_{i-k}\big), \quad k \geqslant 0, \tag{1.43}$$

donde el valor $\boldsymbol{y}^h_{i+1}$ aún es desconocido, y definimos $\boldsymbol{y}^h_{i+1}$ por la ecuación

$$\frac{\mathrm{d}}{\mathrm{d}x}\boldsymbol{P}(x; i)\bigg|_{x=x_{i+1}} = \boldsymbol{f}\big(x_{i+1}, \boldsymbol{y}^h_{i+1}\big).$$

Usando para la representación del polinomio la fórmula de Newton,

$$\boldsymbol{P}(x;i) = \sum_{j=0}^{k+1} \boldsymbol{y}^h_{[x_{i+1},\ldots,x_{i+1-j}]} \prod_{l=0}^{j-1}(x - x_{i+1-l}),$$

donde $\boldsymbol{y}^h_{[x_{i+1},\ldots,x_{i+1-j}]}$ es una j-ésima diferencia dividida con respecto a los datos (1.43), tenemos

$$\sum_{j=1}^{k+1} \boldsymbol{y}^h_{[x_{i+1},\ldots,x_{i+1-j}]} \prod_{l=1}^{j-1}(x_{i+1} - x_{i+1-l}) = \boldsymbol{f}\big(x_{i+1}, \boldsymbol{y}^h_{i+1}\big),$$

es decir, un sistema de ecuaciones para determinar $\boldsymbol{y}^h_{i+1}$.

Para el caso equidistante podemos calcular explícitamente los coeficientes de estas fórmulas. Tal procedimiento entrega las fórmulas

$$\boldsymbol{y}^h_{i+1} = \sum_{l=0}^{k} \alpha_{kl} \boldsymbol{y}^h_{i-l} + h\beta_0 \boldsymbol{f}\big(x_{i+1}, \boldsymbol{y}^h_{i+1}\big) \tag{1.44}$$

con los coeficientes

k	0	1	2	3	4	5
β_0	1	$\frac{2}{3}$	$\frac{6}{11}$	$\frac{12}{25}$	$\frac{60}{137}$	$\frac{60}{147}$
α_{k0}	1	$\frac{4}{3}$	$\frac{18}{11}$	$\frac{48}{25}$	$\frac{300}{137}$	$\frac{360}{147}$
α_{k1}		$-\frac{1}{3}$	$-\frac{9}{11}$	$-\frac{36}{25}$	$-\frac{300}{137}$	$-\frac{450}{147}$
α_{k2}			$\frac{2}{11}$	$\frac{16}{25}$	$\frac{200}{137}$	$\frac{400}{147}$
α_{k3}				$-\frac{3}{25}$	$-\frac{75}{137}$	$-\frac{225}{147}$
α_{k4}					$\frac{12}{137}$	$\frac{72}{147}$
α_{k5}						$-\frac{10}{147}$

.

Veremos que estas fórmulas son interesantes solamente para $k \leqslant 5$. Para $k = 0$, obtenemos nuevamente el método de Euler implícito. Las fórmulas (1.44) son conocidas como *fórmulas BDF* (backward differentiation formula) o *método de Gear*.

1.3.3. Breve teoría de métodos de pasos múltiples. En lo siguiente, consideraremos sólo *métodos lineales de k pasos*, que pueden ser representados como

$$\sum_{i=0}^{k} \alpha_i \boldsymbol{y}^h_{m+i} = h \sum_{i=0}^{k} \beta_i \boldsymbol{f}\big(x_{m+i}, \boldsymbol{y}^h_{m+i}\big), \quad m = 0, \ldots, N_h - k, \tag{1.45}$$

donde $h = x_{j+1} - x_j$ para todo j. Eso incluye los métodos de Adams-Bashforth, Adams-Moulton y BDF, pero *no* incluye el método predictor-corrector.

Las propiedades de (1.45) pueden ser caracterizadas por los polinomios

$$\varrho(\xi) := \sum_{i=0}^{k} \alpha_i \xi^i, \quad \sigma(\xi) := \sum_{i=0}^{k} \beta_i \xi^i.$$

Los conceptos del error de truncación local y del error de discretización global son introducidos en forma análoga a los métodos de paso simple, es decir, definimos

$$\boldsymbol{\tau}\big(x, \boldsymbol{y}(x); h\big) := \frac{1}{h}\left(\sum_{i=0}^{k} \alpha_i \boldsymbol{y}(x+ih) - h\sum_{i=0}^{k} \beta_i \boldsymbol{f}\big(x+ih, \boldsymbol{y}(x+ih)\big)\right),$$
$$\boldsymbol{\varepsilon}(x; h) := \boldsymbol{y}(x) - \boldsymbol{y}^h_{N_h}, \quad \text{donde } N_h h = x - x_0.$$

El método se llama *consistente al menos del orden p* si

$$\boldsymbol{\tau}\big(x, \boldsymbol{y}(x); h\big) = \mathcal{O}(h^p) \quad \text{para } h \to 0,$$

y *convergente del orden p* si

$$\boldsymbol{\varepsilon}(x; h) = \mathcal{O}(h^p) \quad \text{para } h \to 0.$$

Teorema 1.3. *El método de k pasos* (1.45) *es consistente al menos del orden p si*

$$\varrho_1(1) = 0, \quad \varrho_{j+1}(1) = j\sigma_j(1), \quad j = 1, \dots, p,$$

donde los polinomios ϱ_j y σ_j son definidos por la recursión

$$\varrho_1(\xi) \equiv \varrho(\xi), \quad \sigma_1(\xi) \equiv \sigma(\xi), \quad \varrho_{j+1}(\xi) \equiv \xi\varrho_j'(\xi), \quad \sigma_{j+1}(\xi) \equiv \xi\sigma_j'(\xi).$$

El método es convergente del orden p si además tenemos:

1. *Todos los ceros de ϱ estan localizados en el interior o en el borde del círculo unitario, y los ceros del borde son simples, es decir,*
$$\big(\varrho(\xi) = 0 \Rightarrow |\xi| \leqslant 1\big) \wedge \big(\varrho(\xi) = 0 \wedge |\xi| = 1 \Rightarrow \varrho'(\xi) \neq 0\big). \tag{1.46}$$
2. *Los errores iniciales son del orden p:*
$$\boldsymbol{y}^h_i - \boldsymbol{y}(x_i) = \mathcal{O}(h^p), \quad i = 0, \dots, k-1.$$

Demostración. Ver, por ejemplo, Deuflhard & Bornemann. ∎

Las raices diferentes de 1 se llaman *raices parasitarias*. Ellas afectan decisivamente el uso práctico del método. La condición (1.46) es conocida como *condición de ceros* o *condición de estabilidad asintótica*. Veremos en una tarea que existen métodos muy razonables que *no* satisfacen la condición de estabilidad asintótica. Los métodos de Adams-Moulton y Adams-Bashforth satisfacen la condición (Tarea), pro las fórmulas (1.44) no satisfacen (1.46) para $k \geqslant 6$.

Uno podria tratar de maximizar el orden de un método eligiendo los coeficientes α_j y β_j de forma óptima para un valor dado de k. Este procedimiento permite alcanzar un orden de consistencia $2k$. Sin embargo, los métodos que resultan son inutiles para las aplicaciones debido al siguiente teorema.

Teorema 1.4 (Primera cota de orden de Dahlquist)**.** *El orden máximo de un método estable y consistente de k pasos es*

$$p = \begin{cases} k+1 & \text{para } k \text{ impar,} \\ k+2 & \text{para } k \text{ par.} \end{cases}$$

Demostración. Ver Hairer, Nørsett & Wanner, 1993. ∎

El orden $k+1$ ya se alcanza por las fórmulas de k pasos del tipo predictor-corrector usando las fórmulas de Adams-Bashforth y Adams-Moulton. Lo único que aún aparece interesante son las fórmulas del orden $k+2$ para k par. La discusión de un ejemplo, el método de Milne-Simpson, en el Ejemplo 1.5, require la presentación del siguiente teorema. (Sin embargo, el ejemplo ilustra el poco interés práctico para este tipo de métodos.)

Teorema 1.5. *Consideremos la siguiente ecuación de diferencias para una sucesión* $\{\eta_i\}_{i\in\mathbb{N}_0}$*:*

$$\sum_{i=0}^{k} \gamma_i \eta_{i+m} = 0, \quad m \in \mathbb{N}_0, \quad \gamma_0\gamma_k \neq 0. \tag{1.47}$$

Entonces la solución de (1.47) *está dada por*

$$\eta_j = \sum_{l=1}^{s} \sum_{i=1}^{v_l} C_{li} j^{i-1} \xi_l^j, \quad j > 0,$$

donde $\xi_1, \ldots, \xi_s$, $\xi_i \neq \xi_j$ *para* $i \neq j$, *son los ceros del polinomio*

$$P(\xi) := \sum_{i=0}^{k} \gamma_i \xi^i$$

con las multiplicidades $v_1, \ldots, v_s$, *o sea*

$$\sum_{l=1}^{s} v_l = k, \quad v_l \in \mathbb{N},$$

y las constantes C_{li} *son determinadas unicamente por los valores iniciales* $\eta_0, \ldots, \eta_{k-1}$.

Ejemplo 1.5. *El* método de Milne-Simpson

$$y_{i+1}^h = y_{i-1}^h + \frac{h}{3}\Big(f\big(x_{i+1}, y_{i+1}^h\big) + 4f\big(x_i, y_i^h\big) + f\big(x_{i-1}, y_{i-1}^h\big)\Big)$$

(considerado aquí para $n = 1$*) es un método de dos pasos del orden* 4. *Lo aplicamos al problema* $y' = \lambda y$, $y(0) = 1$ *con la solución* $y(x) = \mathrm{e}^{\lambda x}$. *Resulta la ecuación de diferencias lineal*

$$(1-q)y_{i+1}^h - 4qy_i^h - (1+q)y_{i-1}^h = 0, \quad y_0^h = 1, \quad q = \frac{h\lambda}{3}. \tag{1.48}$$

Aplicando el Teorema 1.5 a la ecuación (1.48), *llegamos a la representación*

$$y_j^h = C\mu_1^j + (1-C)\mu_2^j,$$

donde $C = C(y_1^h)$ *y*

$$\mu_1 = \frac{1}{1-q}\left(2q + \sqrt{1+3q^2}\right), \quad \mu_2 = \frac{1}{1-q}\left(2q - \sqrt{1+3q^2}\right).$$

Para $\lambda < 0$ *tenemos*

$$\mu_1 = \frac{1}{1+|q|}\left(\sqrt{1+3q^2} - 2|q|\right) = 1 - \mathcal{O}(h),$$
$$\mu_2 = \frac{1}{1+|q|}\left(-\sqrt{1+3q^2} - 2|q|\right) = -1 - \mathcal{O}(h) < -1.$$

Entonces para $C \neq 1$ *la solución siempre tiene una contribución oscilatoria* μ_2 *que crece con* $x = jh$. *(El efecto de error de redondeo causa que en la práctica, siempre* $C \neq 1$.*)*

Para $y_1^h - \mathrm{e}^{\lambda h} = \mathcal{O}(h^4)$, $j \leqslant N$ *y* $Nh = x$ *fijado esta contribución es sólo del orden* $\mathcal{O}(h^4)$ *cuando* $h \to 0$, *o sea el método es convergente del orden* 4. *Pero para la computación práctica uno no puede elegir* h *arbitrariamente pequeño, y la contribución oscilatoria es muy molestosa.*

Capítulo 2

Problemas de valores iniciales para ecuaciones diferenciales ordinarias (Parte II)

2.1. Ecuaciones rígidas (problemas *stiff*)

Desde el estudio de problemas de interpolación, aproximación y cuadratura ya sabemos que el error cometido por esos métodos depende de forma decisiva de una de las derivadas más altas de la función aproximada o del integrando. Obviamente, se puede suponer que lo mismo es válido para la solución del problema de valores iniciales de una ecuación diferencial ordinaria. Pero aquí, también es importante la regularidad (suavidad) de la entera variedad de soluciones en la vecindad de la solución exacta del problema. Esta observación tiene consecuencias importantes para la aplicación de métodos numericos. Discutiremos primero un ejemplo.

Ejemplo 2.1. *Se considera el problema de valores iniciales*

$$y' = \lambda(y - \mathrm{e}^{-x}) - \mathrm{e}^{-x}, \quad y(0) = 1 + \varepsilon, \quad \lambda \in \mathbb{R}. \tag{2.1}$$

La ecuación diferencial ordinaria tiene la solución general

$$y(x) = \alpha \mathrm{e}^{\lambda x} + \mathrm{e}^{-x}, \quad \alpha \in \mathbb{R}.$$

Usando el valor inicial prescrito, tenemos

$$1 + \varepsilon = \alpha + 1,$$

es decir, $\alpha = \varepsilon$, y la solución del problema (2.1) es

$$y(x) = \varepsilon \mathrm{e}^{\lambda x} + \mathrm{e}^{-x}.$$

Para $\varepsilon = 0$ y $x \geqslant 0$, la solución y todas sus derivadas son en valor ebsoluto menores que 1, cualquier que sea el valor de λ. En general, tenemos

$$y^{(p)}(x) := \frac{\mathrm{d}^p y}{\mathrm{d}x^p} = \varepsilon \lambda^p \mathrm{e}^{\lambda x} + (-1)^p \mathrm{e}^{-x},$$

es decir,

$$\left| y^{(p)}(x) \right| \geqslant |\lambda|^p \left(\frac{\varepsilon}{\mathrm{e}} - \frac{1}{|\lambda|^p} \right) \quad \text{para } x \in \left[0, -\frac{1}{\lambda} \right], \ \lambda < 0.$$

Por ejemplo, para $\varepsilon \neq 0$ y $\lambda = -1000$ las derivadas asumen ordenes de tamaño gigantescos en una pequeña vecindad de $x = 0$. Lo mismo es válido si para cualquier x_0 prescribimos el valor inicial

$$y(x_0) = \mathrm{e}^{-x_0} + \varepsilon.$$

x	$h = 0{,}1$	$h = 0{,}01$	$h = 0{,}002$	$h = 0{,}001$	$h = 0{,}0005$
0,1	0,9	$1{,}74 \times 10^5$	0,905	0,904837	0,904837
0,5	$-4{,}6 \times 10^6$	$\lvert\cdot\rvert > 10^{38}$	0,607	0,6065304	0,606530
1,0	$4{,}38 \times 10^{16}$	$\lvert\cdot\rvert > 10^{38}$	0,368	0,367879	0,367879

CUADRO 2.1. Ejemplo 2.1: Valores de la solución aproximada de (2.1) con $\varepsilon = 0$, generados por el método de Euler explícito (2.2) con diferentes valores de h.

x	$h = 0{,}1$	$h = 0{,}01$	$h = 0{,}001$	$h = 0{,}0005$
0,1	0,904837	0,904884	0,904838	0,904838
0,5	0,696531	0,606562	0,606531	0,606531
1,0	0,367879	0,367899	0,367880	0,367880

CUADRO 2.2. Ejemplo 2.1: Valores de la solución aproximada de (2.1) con $\varepsilon = 0$, generados por el método de Euler implícito (2.3) con diferentes valores de h.

En lo siguiente, consideremos $\lambda = -1000$ y $\varepsilon = 0$. En este caso, la solución consiste sólo en la componente suave e^{-x}. Ya para $x > 1/10$, cualquier solución es prácticamente idéntica a la componente suave, si ε es del orden de tamaño 1, dado que $\exp(-100) \approx 3{,}7 \times 10^{-44}$. Entonces la solución exacta y la componente suave disminuyen exponencialmente con x.

Aplicaremos algunos métodos de discretización a este problema. El método de Euler explícito entrega la fórmula

$$\begin{aligned} y_{i+1}^h &= y_i^h + h\big[-1000\big(y_i^h - \exp(-ih)\big) - \exp(-ih)\big] \\ &= (1 - 1000h)y_i^h + 999h\exp(-ih), \quad i = 0, 1, 2, \ldots, \\ y_0^h &= 1. \end{aligned} \tag{2.2}$$

Este método entrega los valores aproximados del Cuadro 2.1, donde las 6 primeras cifras decimales de la solución exacta coinciden con el resultado numérico para $h = 0{,}0005$. Obviamente, sólo para $h \leqslant 0{,}002$ obtenemos soluciones aceptables. (Esto proviene del requerimiento de estabilidad $|1 - 1000h| \leqslant 1$ para este problema; este criterio se establecerá más adelante.)

Aplicamos ahora el método de Euler implícito. A pesar de ser implícito, en el caso de la ecuación (2.1), que es lineal con respecto a y, obtenemos una fórmula de iteración explícita:

$$\begin{aligned} & y_{i+1}^h = y_i^h + h\big[-1000y_{i+1}^h + 1000\exp\big(-(i+1)h\big) - \exp\big(-(i+1)h\big)\big] \\ \Leftrightarrow \quad & (1 + 1000h)y_{i+1}^h = y_i^h + 999h\exp\big(-(i+1)h\big) \\ \Leftrightarrow \quad & y_{i+1}^h = \frac{1}{1+1000h}y_i^h + \frac{999h}{1+1000h}\exp\big(-(i+1)h\big), \quad i = 0, 1, 2, \ldots, \\ & y_0^h = 1. \end{aligned} \tag{2.3}$$

Este método genera los valores numéricos indicados el Cuadro 2.2. Obviamente, los malos resultados obtenidos por (2.2) no son una consecuencia del bajo orden de consistencia del

método, ya que el orden de (2.3) *también es solamente uno. Los mismos efectos pueden ser observados con cualquier método de Runge-Kutta explícito (Tarea).*

Los fenómenos observados en el ejemplo son relacionados con la presencia de soluciones de la ecuación diferencial ordinaria con derivadas grandes en la cercanía de la solución exacta del problema de valores iniciales dado. La solución exacta también asume valores en este rango de pendientes fuertes, y los métodos explícitos producen un efecto oscilatorio (si h no es suficientemente pequeño) que nos aleja rapidamente de la solución exacta. (Por otro lado, el problema de valores iniciales

$$y' = -y, \quad y(0) = 1$$

con la misma solución exacta se podría aproximar facilmente por el método de Euler explícito, usando $h = 0{,}1$ y calculando hasta $x = 1$.) Las ecuaciones diferenciales ordinarias que presentan este problema se llaman *rígidas* (problemas "stiff").

El prototipo de una ecuación rígida es el siguiente sistema lineal homogéneo:

$$\begin{aligned} &\boldsymbol{y}' = \boldsymbol{A}\boldsymbol{y}, \quad \boldsymbol{y}(0) = \boldsymbol{y}_0, \quad \boldsymbol{A} \in \mathbb{R}^{n\times n} \text{ diagonalizable,} \\ &\sigma(\boldsymbol{A}) = \{\lambda_1, \ldots, \lambda_n\}, \quad \operatorname{Re}\lambda_1 \leqslant \ldots \leqslant \operatorname{Re}\lambda_n \leqslant 0, \quad -\operatorname{Re}\lambda_1 \gg 1. \end{aligned} \tag{2.4}$$

Aquí uno podria elegir $S := -\operatorname{Re}\lambda_1$ como medida de la rigidez. En el caso de un sistema más general $\boldsymbol{y}' = \boldsymbol{f}(x, \boldsymbol{y})$ podriamos definir

$$S := -\inf_{\substack{(x,\boldsymbol{z}),(x,\boldsymbol{y})\in\mathcal{D}_{\mathrm{f}} \\ \boldsymbol{y}\neq\boldsymbol{z}}} \frac{(\boldsymbol{f}(x,\boldsymbol{y}) - \boldsymbol{f}(x,\boldsymbol{z}))^{\mathrm{T}}(\boldsymbol{y}-\boldsymbol{z})}{(\boldsymbol{y}-\boldsymbol{z})^{\mathrm{T}}(\boldsymbol{y}-\boldsymbol{z})}.$$

2.2. Estabilidad de métodos de discretización para problemas de valores inicales de ecuaciones diferenciales ordinarias

2.2.1. Estabilidad lineal. Analizaremos ahora el comportamiento de de métodos de discretización aplicados al problema test

$$y' = \lambda y, \quad y(0) = y_0; \quad \operatorname{Re}\lambda < 0. \tag{2.5}$$

Un sistema del tipo (2.4) puede ser transformado a un sistema de n ecuaciones deacopladas del tipo (2.5), donde λ representa uno de los valores propios de $\boldsymbol{A}$. Por lo tanto, las siguientes consideraciones serán relevantes no sólo para ecuaciones escalares, sino que también para sistemas del tipo (2.4).

Definición 2.1. *Se dice que un método de paso simple o de pasos múltiples* pertenece a la clase CA (de coeficientes analíticos) *si su aplicación al problema* (2.5) *lleva a una ecuación de diferencias*

$$\sum_{j=0}^{k} g_j(h\lambda) y^h_{m+j} = 0, \quad m = 0, 1, \ldots, N-k, \tag{2.6}$$

donde las funciones $g_0, \dots, g_k$, $k \geqslant 1$, *son analíticas (en el sentido de funciones complejas),* $g_k(0) \neq 0$, *y el polinomio*

$$z \mapsto \sum_{j=0}^{k} g_j(0) z^j$$

satisface la condición de ceros (1.46).

Ejemplo 2.2. *El método de Runge-Kutta clásico (Ejemplo 1.2) corresponde a*

$$k = 1, \quad g_1(z) \equiv 1, \quad g_0(z) = -\left(1 + z + \frac{z^2}{2} + \frac{z^3}{6} + \frac{z^4}{24}\right),$$

el método de Euler implícito (1.39) *a*

$$k = 1, \quad g_1(z) = 1 - z, \quad g_0(z) \equiv -1,$$

el método trapezoidal (1.40) *a*

$$k = 1, \quad g_1(z) = 1 - \frac{z}{2}, \quad g_0(z) = -\left(1 + \frac{z}{2}\right)$$

y el método predictor-corrector (1.41), (1.42) *a*

$$k = 3, \quad g_3(z) \equiv 1, \quad g_2(z) = -1 - \frac{28}{24} z - \frac{9}{24} \cdot \frac{23}{12} z^2,$$
$$g_1(z) = \frac{5}{24} z + \frac{9}{24} \cdot \frac{16}{12} z^2, \quad g_0(z) = -\frac{1}{24} z + \frac{9}{24} \cdot \frac{5}{12} z^2.$$

Usando el Teorema 1.5 podemos determinar la solución de (2.6) (y entonces el crecimiento de los valores discretizados y_i^h) directamente del polinomio (llamado *polinomio de estabilidad*)

$$P(z; q) := \sum_{j=0}^{k} g_j(q) z^j, \quad q = h\lambda. \tag{2.7}$$

Nuestra discusión aclara que es deseable que el polinomio P satisfaga la condición de ceros (1.46) para un dominio de valores

$$\{q \in \mathbb{C} \mid \operatorname{Re} q < 0\}$$

lo más grande posible. El interior del dominio donde se cumple (1.46) se llama *dominio de estabilidad.*

Definición 2.2. *Un método de paso simple o de pasos múltiples de la case (CA) se llama* absolutamente estable *en un dominio* $G \subset \mathbb{C}$ *si sus funciones coeficientes* g_k *son analíticas en* G, $g_k(q) \neq 0$ *para todo* $q \in G$, *y si los zeros de* $P(z; q)$ *de* (2.7) *tienen un valor absoluto menor que uno:*

$$\forall q \in G : \ P(z; q) = \sum_{j=0}^{k} g_j(q) z^j = 0 \Longrightarrow |z| < 1.$$

El método se llama A-estable *si es absolutamente estable en* G *con*

$$\left\{z \mid \operatorname{Re} z < 0\right\} \subset G,$$

$A(\alpha)$-estable si es absolutamente estable en G con

$$\big\{z \mid \arg(-z) \in (-\alpha, \alpha)\big\} \subset G \quad (\alpha > 0),$$

y A_0-estable si es absolutamente estable en G con

$$\big\{z \mid \operatorname{Re} z < 0,\ \operatorname{Im} z = 0\big\} \subset G.$$

Si el método es A-estable y adicionalmente

$$\limsup_{\operatorname{Re} q \to -\infty} \max\big\{|z| \mid P(z;q) = 0\big\} < 1,$$

*el método se llama L-*estable.

Un método A-estable permite la integración estable (no: exacta) de $y' = \lambda y$, $y(0) = y_0$ con un tamaño de paso h arbitrario. Podemos esperarnos que un tal método también permite la integración estable y exacta de sistemas rígidos si el tamaño de paso es determinado según los componentes suaves y lentamente decrecientes, una vez que los componentes rapidamente decrecientes de la solución ya no son importantes. Ningun método explícito es A-estable.

Ejemplo 2.3. *El método de Euler explícito es absolutamente estable precisamente en*

$$\big\{z \in \mathbb{C} \mid |z + 1| < 1\big\},$$

y no puede ser usado, por ejemplo, para ecuaciones diferenciales ordinarias de oscilación, dado que este método siempre amplifica las amplitudes de la solución por un factor $(1 + hc)$ por paso, donde C es una constante que no depende de h. El método de Euler implícito es absolutamente estable precisamente en

$$\big\{z \in \mathbb{C} \mid |z - 1| > 1\big\},$$

este método también es A-estable y L-estable, pero no es apto para ecuaciones de oscilación ya que amortigua la amplitud de la solución discretizada por un factor $1/(1 + hC)$ en cada paso.

La regla trapezoidal (1.40) *es absolutamente estable estrictamente para*

$$\left|\frac{1 + \frac{z}{2}}{1 - \frac{z}{2}}\right| < 1,$$

es decir, es absolutamente estable estrictamente para

$$z \in \mathbb{C}_- := \{z \in \mathbb{C} \mid \operatorname{Re} z < 0\}.$$

Es apta para ecuaciones de oscilación (por ejemplo, $y' = \omega \mathrm{i} y$, $\mathrm{i} = \sqrt{-1}$) porque

$$\operatorname{Re} z = 0 \Longrightarrow \left|\frac{1 + \frac{z}{2}}{1 - \frac{z}{2}}\right| = 1.$$

k	1	2	3	4	5	6
α	$90°$	$90°$	$88°02'$	$73°21'$	$51°50'$	$17°50'$

CUADRO 2.3. Los ángulos α de $A(\alpha)$-estabilidad de los métodos BDF de k pasos.

$m = p$	1	2	3	4
γ	-2	-2	$-2{,}51$	$-2{,}78$

CUADRO 2.4. Los intervalos reales $(\gamma, 0)$ de estabilidad de algunos métodos de Runge-Kutta con $m = p$ pasos.

k	1	2	3	4	5
γ	$-\infty$	-6	-3	$-\frac{90}{49}$	$-\frac{45}{38}$

CUADRO 2.5. Los intervalos reales $(\gamma, 0)$ de estabilidad de los métodos de Adams-Moulton con k pasos.

Sin embargo, la regla trapezoidal no es L-estable. Eso se nota de forma desgradable en el caso de soluciones decrecientes porque a la solución discretizada se agregan oscilaciones pequeñas (pero acotadas). Para la solución discretizada de (2.5)*, el método trapezoidal entrega*

$$y_i^h = \left(\frac{1+\frac{h\lambda}{2}}{1-\frac{h\lambda}{2}}\right)^i y_0 = (-1)^i \left(\frac{h\lambda+2}{h\lambda-2}\right)^i y_0 = (-1)^i \left(1+\frac{4}{h\lambda-2}\right)^i y_0,$$

o sea, para $h\lambda < -2$ obtenemos una solución amortiguada, pero oscilatoria de la solución exacta monótona.

Los métodos BDF de k pasos son $A(\alpha)$-estables, con los valores $\alpha = \alpha(k)$ dados en el Cuadro 2.3. En particular, el método con $k = 2$ es A-estable (e incluso L-estable). Con este método (el método BDF con $k = 2$) y la regla trapezoidal ya conocemos dos métodos lineales de pasos múltiples A-estables.

Lamentablemente, no existen métodos de pasos múltiples lineales y A-estables del orden mayor que 2.

Teorema 2.1 (Segunda cota de orden de Dahlquist)**.** *Cada método consistente, A-estable, lineal y de pasos múltiples es del orden de consistencia $\leqslant 2$.*

El dominio de estabilidad de los métodos de Adams-Moulton de $k \geqslant 2$ pasos, de los métodos de Adams-Bashforth de $k \geqslant 1$ pasos, del método predictor-corrector y de los métodos explícitos de Runge-Kutta es relativamente pequeño. Para todos estos métodos, la condición

k	1	2	3	4	5
γ	-2	-1	$-\frac{6}{11}$	$-\frac{3}{10}$	$-\frac{90}{551}$

CUADRO 2.6. Los intervalos reales $(\gamma, 0)$ de estabilidad de los métodos de Adams-Bashforth con k pasos.

de estabilidad es

$$\left|\frac{h}{\lambda}\right| < C,$$

donde C es una constante del orden de magnitud 1. Los intervalos reales $(\gamma, 0)$ de la estabilidad absoluta son dados por el Cuadro 2.4 para los métodos de Runge-Kutta explícitos de m pasos (lo que incluye el método clásico con $m = 4$, por el Cuadro 2.5 para los métodos de Adams-Moulton de k pasos y por el Cuadro 2.6 para los métodos de Adams-Bashforth de k pasos.

De los métodos discutidos hasta ahora, sólo el método trapezoidal, el método del punto medio implícito, y los métodos BDF con $k \leqslant 6$ pasos son aptos para ecuaciones muy rígidas, estos últimos con la restricción que no deben aparecer componentes de soluciones con $\arg(-\lambda) > \alpha$, con α del Cuadro 2.3.

2.2.2. Estabilidad no lineal. Los resultados de la Sección 2.2.1 son insatisfactorios por que permiten solamente conclusiones muy tentativas respecto al comportamento de métodos de disretización aplicados a problemas no lineales. Vemos que sistemas lineales con coeficientes variables ya presentan problemas particulares.

Ejemplo 2.4. *Consideremos el problema de valores iniciales*

$$y' = \lambda(x)y, \quad y(0) = y_0; \quad \lambda : \mathbb{R}^+ \to \mathbb{R}^-, \quad \lambda \in C^1(\mathbb{R}^+).$$

Aquí el método de Euler implícito es estable sin restricción para $h > 0$. El método trapezoidal genera la recursión

$$y^h_{i+1} = \frac{2 + \lambda(x_i)h}{2 - \lambda(x_{i+1})h} y^h_i,$$

y la condición

$$\forall i \in \mathbb{N}_0 : \quad \left|y^h_{i+1}\right| \leqslant \left|y^h_i\right|$$

nos entrega para

$$\sup_{x \in \mathbb{R}^+} \lambda'(x) = \beta > 0$$

la restricción

$$h \leqslant \frac{2}{\sqrt{\beta}}.$$

El método del punto medio implícito,

$$y_{i+1}^h = y_i^h + hf\left(\frac{x_i + x_{i+1}}{2}, \frac{y_i^h + y_{i+1}^h}{2}\right),$$

es equivalente al método trapezoidal para una ecuación lineal con coeficientes constantes. Pero aquí el método del punto medio implícito asume la forma

$$y_{i+1}^h = \frac{2 + h\lambda\left(\dfrac{x_i + x_{i+1}}{2}\right)}{2 - h\lambda\left(\dfrac{x_i + x_{i+1}}{2}\right)} y_i^h,$$

o sea el método es estable sin restricciones.

Más encima, para sistemas lineales con coeficientes variables ya no podemos concluir que la solución es estable cuando los valores propios de

$$\partial_2 \boldsymbol{f}(x, \boldsymbol{y})\big|_{\boldsymbol{y}=\boldsymbol{y}(x)} = \boldsymbol{A}(x)$$

son suficientemente pequeños, como non muestra el siguiente ejemplo.

Ejemplo 2.5. *Consideremos el problema de valores iniciales*

$$\boldsymbol{y}' = \frac{1}{\varepsilon}\boldsymbol{U}^{\mathrm{T}}(x)\begin{bmatrix} -1 & \eta \\ 0 & -1 \end{bmatrix}\boldsymbol{U}(x)\boldsymbol{y} \equiv \boldsymbol{A}(x)\boldsymbol{y}, \quad \boldsymbol{y}(0) = \boldsymbol{y}_0,$$

con un parámetro $\varepsilon > 0$ y la matriz

$$\boldsymbol{U}(x) = \begin{bmatrix} \cos(\alpha x) & \sin(\alpha x) \\ -\sin(\alpha x) & \cos(\alpha x) \end{bmatrix}$$

Aquí los valores propios de $\boldsymbol{A}$ son $\lambda_1 = \lambda_2 = -1/\varepsilon$. Usando la transformación

$$\boldsymbol{v}(x) := \boldsymbol{U}(x)\boldsymbol{y}(x),$$

tenemos $\|\boldsymbol{y}(x)\|_2 = \|\boldsymbol{v}(x)\|_2$, y $\boldsymbol{v}$ es solución del problema de valores inicales

$$\boldsymbol{v}' = \begin{bmatrix} -\dfrac{1}{\varepsilon} & \dfrac{\eta}{\varepsilon} + \alpha \\ -\alpha & -\dfrac{1}{\varepsilon} \end{bmatrix}\boldsymbol{v}, \quad \boldsymbol{v}(0) = \boldsymbol{y}_0$$

es decir,

$$\boldsymbol{v}(x) = \boldsymbol{v}_1 \exp(\kappa_1 x) + \boldsymbol{v}_2 \exp(\kappa_2 x), \quad \kappa_{1,2} = -\frac{1}{\varepsilon} \pm \sqrt{-\alpha\left(\frac{\eta}{\varepsilon} + \alpha\right)},$$

donde los vectores $\boldsymbol{v}_1$ y $\boldsymbol{v}_2$ dependen de $\boldsymbol{y}_0$. Vemos que para $0 < \varepsilon < 1$, $\alpha = -1$ y $\eta > 2$ existen soluciones que crecen exponencialmente.

Existen varias extensiones de la teoría lineal de estabilidad para métodos de discretización a sistemas generales (esto es tema de investigación corriente). Para el siguiente tipo de ecuaciones ya existen resultados.

Definición 2.3. *Sea* $\boldsymbol{f} : \mathbb{R}^+ \times \mathbb{R}^n \to \mathbb{R}^n$*, y para un producto escalar* $\langle \cdot, \cdot \rangle$*, supongamos que*

$$\forall x \in \mathbb{R}^+ : \ \forall \boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^n : \quad \big\langle \boldsymbol{f}(x, \boldsymbol{u}) - \boldsymbol{f}(x, \boldsymbol{v}), \boldsymbol{u} - \boldsymbol{v} \big\rangle \leqslant m \langle \boldsymbol{u} - \boldsymbol{v}, \boldsymbol{u} - \boldsymbol{v} \rangle. \tag{2.8}$$

Si $m \leqslant 0$*, entonces la ecuación diferencial ordinaria* $\boldsymbol{y}' = \boldsymbol{f}(x, \boldsymbol{y})$ *se llama* disipativa.

Ejemplo 2.6. *Consideremos el problema*

$$\dot{\boldsymbol{x}}(t) = -\nabla h\big(\boldsymbol{x}(t)\big), \quad t > 0; \quad \boldsymbol{x}(0) = \boldsymbol{x}_0, \tag{2.9}$$

donde $h : \mathbb{R}^n \to \mathbb{R}$ *sea una función dos veces continuamente diferenciable y uniformemente convexa, es decir,*

$$\exists \gamma > 0 : \ \forall \boldsymbol{y}, \boldsymbol{z} \in \mathbb{R}^n : \quad \gamma \boldsymbol{z}^{\mathrm{T}} \boldsymbol{z} \leqslant \boldsymbol{z}^{\mathrm{T}} \nabla^2 h(\boldsymbol{y}) \boldsymbol{z}.$$

Con la notación de la Definición 2.3, tenemos

$$\boldsymbol{f} = -\nabla h$$

(esta función no depende explícitamente de t*), y usando el producto escalar definido por* $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \boldsymbol{x}^{\mathrm{T}} \boldsymbol{y}$*, tenemos*

$$\begin{aligned} \big(\boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{f}(\boldsymbol{y})\big)^{\mathrm{T}} (\boldsymbol{x} - \boldsymbol{y}) &= (\boldsymbol{y} - \boldsymbol{x})^{\mathrm{T}} \left(\int_0^1 \nabla^2 h\big(\boldsymbol{x} + \tau(\boldsymbol{y} - \boldsymbol{x})\big) \, \mathrm{d}\tau \right) (\boldsymbol{x} - \boldsymbol{y}) \\ &\leqslant -\gamma (\boldsymbol{x} - \boldsymbol{y})^{\mathrm{T}} (\boldsymbol{x} - \boldsymbol{y}), \end{aligned}$$

o sea, $m = -\gamma < 0$ *en* (2.8)*.*

La ecuación diferencial ordinaria (2.9) *tiene como única solución estacionaria el único mínimo* $\boldsymbol{x}^*$ *de* h*, y para* $t \to \infty$ *la solución de cualquier problema de valores iniciales de* $\dot{\boldsymbol{x}} = -\nabla h(\boldsymbol{x})$ *converge a* $\boldsymbol{x}^*$*. Entonces, se puede aproximar* $\boldsymbol{x}^*$ *por la solución numérica de* (2.9)*, y como uno quiere hacer* $t \to \infty$*, se prefiere usar u tamaño de paso lo más grande posible, sin destruir la convergencia del método a la solución estacionaria. La ecuación* (2.9) *puede ser extraordinariamente rígida, lo que ocurre cuando* $\nabla^2 h$ *es una matriz mal acondicionada.*

Hay que tomar en cuenta que (2.8) no utiliza $\|\partial_2 \boldsymbol{f}\|$, es decir, la clase de ecuaciones diferenciales disipativas incluye problemas arbitrariamente rígidos.

Definición 2.4. *Un método de discretización para problemas de valores iniciales de ecuaciones diferenciales ordinarias se llama* optimalmente B-convergente del orden p sobre la clase de todos los problemas de valores iniciales disipativos*, si*

$$\big\| \boldsymbol{y}(x_i) - \boldsymbol{y}_i^h \big\| \leqslant C(x_i) h^p \quad \text{para } h < \bar{h} \text{ y } i \in \mathbb{N}_0,$$

donde $C(x_i)$ *depende solamente de*

$$\text{máx}\big\{ \big\| \boldsymbol{y}^{(j)}(x) \big\| \mid x_0 \leqslant x \leqslant x_i, \, 1 \leqslant j \leqslant \bar{p} \big\}$$

para un cierto $\bar{p}$*, suponiendo que la verdadera solución* $\boldsymbol{y}(x)$ *del problema es suficientemente diferenciable, para una función* $\boldsymbol{f}$ *que satisface* (2.8) *y* $m \leqslant 0$*.*

La gran ventaja de los métodos optimalmente B-convergentes consiste en que la restricción del tamaño de paso no depende de la rigidez del sistema, y que el error global de la discretización no depende tampoco de la rigidez del sistema. Al contrario de eso, los métodos explícitos de Runge-Kutta, por ejemplo, poseen un error $h^p C(x_i)$, donde $C(x_i)$ tambien depende de $\|\partial_2 \boldsymbol{f}\|$. En general, el valor de $\bar{h}$ en la Definción 2.4 depende de m.

Teorema 2.2. *El método de Euler implícito es optimalmente B-convergente del orden 1, y la regla trapezoidal y la regla implícita del punto medio son optimalmente B-convergentes del orden 2, en cada caso sobre la clase de los problemas de valores iniciales disipativos.*

2.3. Métodos de Runge-Kutta implícitos y métodos de Rosenbrock

Los métodos de Runge-Kutta generales son caracterizados por el diagrama de Butcher (1.20). Recordamos que un esquema de Runge-Kutta es implícito si existe un coeficiente $\beta_{ij} \neq 0$ con $j \geqslant i$. Debido al alto esfuerzo computacional, la aplicación de tales métodos para problemas no rígidos no sería muy económica; sin embargo, para problemas rígidos estos métodos pueden ser muy ventajosos, dado que incluyen métodos optimalmente B-convergentes de orden arbitrariamente alto.

Primero demostramos que los métodos de Runge-Kutta implícitos pertenecen a la clase CA, ver Definición 2.1. Recordamos que

$$\boldsymbol{k}_j = \boldsymbol{f}\left(x + h\alpha_j, \boldsymbol{y}_i^h + h\sum_{l=1}^{m} \beta_{jl}\boldsymbol{k}_l\right), \quad j = 1, \ldots, m,$$

lo que implica que para $\boldsymbol{f}(x, \boldsymbol{y}) = \lambda \boldsymbol{y}$ y $n = 1$

$$(\boldsymbol{I} - \lambda h \boldsymbol{B})\boldsymbol{k} = \lambda y_i^h \boldsymbol{e}, \quad \boldsymbol{k} = \begin{pmatrix} k_1 \\ \vdots \\ k_m \end{pmatrix}, \quad \boldsymbol{B} = (\beta_{jl})_{j,l=1,\ldots,m}, \quad \boldsymbol{e} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}.$$

Para h suficientemente pequenño, $\boldsymbol{I} - \lambda h \boldsymbol{B}$ siempre es invertible y por lo tanto, $\boldsymbol{k}$ es bien definido. Según la regla de Cramer tenemos

$$k_j = \lambda y_j^h R_j(\lambda h), \quad j = 1, \ldots, m,$$

donde R_j es una función racional de λh:

$$R_j(z) = \frac{P_{j,m-1}(z)}{\det(\boldsymbol{I} - z\boldsymbol{B})} = \frac{P_{j,m-1}(z)}{\prod_{i=1}^{m}(1 - z\mu_i)}, \tag{2.10}$$

donde $\mu_1, \ldots, \mu_m$ son los valores propios de $\boldsymbol{B}$ y $P_{j,m-1}$ es un polinomio del grado máximo $m-1$; el grado máximo del polinomio del denominador de (2.10) es m. Puesto que

$$y_{i+1}^h = y_i^h + h\lambda \sum_{j=1}^{m} \gamma_j k_j,$$

obtenemos que

$$y_{i+1}^h = y_i^h + h\lambda y_i^h \sum_{j=1}^{m} \gamma_j R_j(\lambda h) = \tilde{R}_m(\lambda h) y_i^h,$$

donde $\tilde{R}_m$ es una función racional (cuociente de dos polinomios del grado máximo m).

Ejemplo 2.7. *Consideremos el método con $m = 2$ dado por*

$$\begin{array}{c|cc} \frac{1}{4} & \frac{1}{4} & 0 \\ \frac{3}{4} & \frac{1}{4} & \frac{1}{2} \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array}. \tag{2.11}$$

Poniendo $f(x, y) = \lambda y$ obtenemos

$$k_1 = \lambda\left(y_i^h + \frac{h}{4}k_1\right), \quad k_2 = \lambda\left(y_i^h + h\left(\frac{1}{4}k_1 + \frac{1}{2}k_2\right)\right),$$

es decir,

$$k_1 = \frac{\lambda y_i^h}{1 - \frac{h\lambda}{4}}, \quad k_2 = \frac{\lambda y_i^h}{1 - \frac{h\lambda}{4}} + \frac{\frac{\lambda h}{2}\lambda y_i^h}{\left(1 - \frac{h\lambda}{4}\right)^2},$$

lo que significa que

$$\begin{aligned} y_{i+1}^h &= y_i^h \frac{\left(1 - \frac{h\lambda}{4}\right)^2 + h\lambda\left(1 - \frac{h\lambda}{4}\right) + \left(\frac{h\lambda}{2}\right)^2}{\left(1 - \frac{h\lambda}{4}\right)^2} \\ &= y_i^h\left(1 + \frac{h\lambda}{(1 - h\lambda/4)^2}\right). \end{aligned}$$

Dado que

$$\left|1 + \frac{h\lambda}{(1 - h\lambda/4)^2}\right| = \frac{\left(1 + \frac{(\operatorname{Re}\lambda)h}{4}\right)^2 + \frac{(\operatorname{Im}\lambda)^2 h^2}{16}}{\left(1 - \frac{(\operatorname{Re}\lambda)h}{4}\right)^2 + \frac{(\operatorname{Im}\lambda)^2 h^2}{16}},$$

este método es A-estable. Es consistente del orden 2. También es optimalmente B-convergente del orden 2.

Según la Definición 2.2, la propiedades de estabilidad lineal de un método de Runge-Kutta implícito depende de para cuales $z \in \mathbb{C}$ se cumple

$$\left|\tilde{R}_m(z)\right| < 1.$$

En esta clase de métodos existen métodos A-estables de orden arbitrariamente alto.

Si elegimos como α_i los nodos de Gauss para la función de peso 1 sobre $[0, 1]$, y como γ_i los pesos de Gauss asociados y como β_{ij} (los pesos de la cuadratura interior) de tal forma

que

$$\sum_{j=1}^{m} \beta_{ij} g(\alpha_j) = \int_0^{\alpha_i} g(x)\,\mathrm{d}x \quad \text{para } g \in \Pi_{m-1},\ i = 1, \dots, m,$$

obtenemos los *métodos de Gauss-Runge-Kutta*. Por ejemplo, para $m = 2$ obtenemos el siguiente esquema.

$$\begin{array}{c|cc} \frac{3-\sqrt{3}}{6} & \frac{1}{4} & \frac{3-2\sqrt{3}}{12} \\ \frac{3+\sqrt{3}}{6} & \frac{3+2\sqrt{3}}{12} & \frac{1}{4} \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array}.$$

Teorema 2.3.

(a) *Los métodos de Gauss-Runge-Kutta de m pasos son consistentes del orden $2m$.*
(b) *Los métodos de Gauss-Runge-Kutta de m pasos son A-estables y optimalmente B-convergentes del orden m sobre la clase de los problemas disipativos.*

Demostración. Para las demostraciones de (a) y (b), ver (por ejemplo) Grigorieff (1972) y Burrage y Hundsdorfer (1987), respectivamente. ■

En particular, el orden de la B-convergencia puede ser significativamente menor que el orden de consistencia clásico. Esto es debido al requerimiento de la B-convergencia optimal que el residuo (término de error) debe ser independiente de la rigidez del sistema.

Se puede demostrar que para una función $\boldsymbol{f}$ disipativa, las pendientes $\boldsymbol{k}_j$ de un método de Gauss-Runge-Kutta son determinados únicamente de la ecuación del método, independientemente de la rigidez del sistema y para $h < \bar{h}$, para un $\bar{h} > 0$ fijo. Sin embargo, la implementación de un tal método es bastante dificil, dado que, por ejemplo, en cada paso de integración para un sistema de ecuaciones diferenciales del orden n hay que resolver un sistema no lineal del orden nm.

En la práctica se presenta una situación más simple si $\beta_{ij} = 0$ para $j > i$, lo que corresponde a los *métodos de Runge-Kutta diagonalmente implícitos*. Un ejemplo es el método (2.11) del Ejemplo 2.7. Desafortunadamente, la mayoría de estos métodos son solamente optimalmente B-convergentes de primer orden.

A veces se presentan problemas donde solamente $\|\partial_2 \boldsymbol{f}\|$ es grande, mientras que el orden de magnitud de todas las demás derivadas de $\boldsymbol{f}$ es pequeño comparado con $\|\partial_2 \boldsymbol{f}\|$, por ejemplo,

$$\boldsymbol{y}' = \boldsymbol{A}\boldsymbol{z} + \boldsymbol{g}(x, \boldsymbol{y}),$$

con una función $\boldsymbol{g}$ para la cual todas las derivadas parciales tienen una norma del orden de magnitud $\mathcal{O}(1)$, mientras que los valores propios son dispersos en

$$\{z \in \mathbb{C} \mid \operatorname{Im} z < 0\}.$$

En estos casos, podemos aplicar exitosamente los *métodos de Rosenbrock* y sus modificaciones.

Podemos derivar los métodos de Rosenbrock de la siguiente forma. Partimos de las ecuaciones de un método de Runge-Kutta diagonalmente implícito:

$$\begin{aligned}
\boldsymbol{k}_1 &= \boldsymbol{f}\big(x+\alpha_1 h, \boldsymbol{y}^h + \beta_{11} h\boldsymbol{k}_1\big),\\
\boldsymbol{k}_2 &= \boldsymbol{f}\big(x+\alpha_2 h, \boldsymbol{y}^h + \beta_{21} h\boldsymbol{k}_1 + \beta_{22} h\boldsymbol{k}_2\big),\\
&\vdots\\
\boldsymbol{k}_m &= \boldsymbol{f}\big(x+\alpha_m h, \boldsymbol{y}^h + \beta_{m1} h\boldsymbol{k}_1 + \ldots + \beta_{mm} h\boldsymbol{k}_m\big).
\end{aligned}$$

Estas ecuaciones se resuelven sucesivamente de forma iterativa, ejecutando un paso del método de Newton con

$$\boldsymbol{k}_i^{(0)} := 0$$

como dato inicial. Esto entrega las ecuaciones explícitas

$$\begin{aligned}
&\left[\boldsymbol{I} - h\beta_{ii}\partial_2\boldsymbol{f}\left(x+\alpha_i h, \boldsymbol{y}^h + h\sum_{j=1}^{i-1}\beta_{ij}\boldsymbol{k}_j\right)\right]\boldsymbol{k}_i\\
&= \boldsymbol{f}\left(x+\alpha_i h, \boldsymbol{y}^h + h\sum_{j=1}^{i-1}\beta_{ij}\boldsymbol{k}_j\right), \quad i=1,\ldots,m.
\end{aligned}$$

Aquí hay que resolver sucesivamente m sistemas lineales. La desventaja aun consiste en el hecho de que hay que calcular m matrices de Jacobi $\partial_2\boldsymbol{f}$ y ejecutar m descomposiciones triangulares. Con una matriz fija

$$\boldsymbol{I} - h\beta\partial_2\boldsymbol{f}(x,\boldsymbol{y})$$

en al lado izquierdo y una corrección correspondiente al lado derecho obtenemos la fórmula de planteo modificada

$$\begin{aligned}
&\big(\boldsymbol{I} - h\beta\partial_2\boldsymbol{f}(x,\boldsymbol{y}^h)\big)\boldsymbol{k}_i\\
&= \boldsymbol{f}\left(x+\alpha_i h, \boldsymbol{y}^h + h\sum_{j=1}^{i-1}\beta_{ij}\boldsymbol{k}_j\right) + h\partial_2\boldsymbol{f}(x,\boldsymbol{y}^h)\sum_{j=1}^{i-1}\sigma_{ij}\boldsymbol{k}_j, \quad i=1,\ldots,m.
\end{aligned}$$

Estas fórmulas son las *fórmulas de Rosenbrock-Wanner*. En esta clase de métodos hay muchos ejemplos de métodos A-estables o $A(\alpha)$-estables.

Ejemplo 2.8 (Kaps & Rentrop, 1979)**.** *Los siguientes coeficientes definen un método Rosenbrock-Wanner A-estable de orden 4:*

$\beta = 0{,}395,$

β_{ij}	$j=1$	$j=2$	$j=3$
$i=2$	0,438		
$i=3$	0,796920457938	0,073079542062	
$i=4$	0	0	0

,

σ_{ij}	$j=1$	$j=2$	$j=3$
$i=2$	$-0{,}767672395484$		
$i=3$	$-0{,}851675323742$	0,522967289188	
$i=4$	0,288463109545	0,08802142734	$-0{,}337389840627$

,

	$i=1$	$i=2$	$i=3$	$i=4$
γ_i	0,199293275702	0,482645235674	0,0680614886256	0,25

.

Capítulo 3

Problemas de valores de frontera para ecuaciones diferenciales ordinarias

3.1. Introducción

En el caso escalar, se busca una solución $y = y(x)$ da le ecuación diferencial ordinaria

$$F\big(x, y, y', \ldots, y^{(n)}\big) = 0 \tag{3.1}$$

sujeta a las condiciones de frontera

$$R_j(y) = R_j\big[y(a), y'(a), \ldots, y^{(n-1)}(a); y(b), y'(b), \ldots, y^{(n-1)}(b)\big] = \alpha_j, \quad j = 1, \ldots, n, \tag{3.2}$$

donde $x = a$ y $x = b$ son puntos en $\mathbb{R}$ y $\alpha_j \in \mathbb{R}$. En general, es dificil obtener resultados acerca de la existencia y la unicidad de soluciones de estos problemas de valores de frontera. Esencialmente, existen sólo resultados para el problema lineal

$$\begin{aligned} (Ly)(x) &\equiv \sum_{i=0}^{n} f_i(x) y^{(i)} = g(x), \quad x \in (a, b), \quad f_n \not\equiv 0 \text{ sobre } (a, b), \\ R_j[y] &= \sum_{k=0}^{n-1} \big(\alpha_{j,k+1} y^{(k)}(a) + \beta_{j,k+1} y^{(k)}(b)\big) = \alpha_j, \quad j = 1, \ldots, n, \end{aligned} \tag{3.3}$$

donde $\alpha_{j,k+1}$ y $\beta_{j,k+1}$ son constantes. Para $g \equiv 0$, la ecuación diferencial ordinaria se llama *homogénea*; para $\alpha_j \equiv 0$, las condiciones de frontera se llaman homogéneas. Si $g \equiv 0$ y $\alpha_j \equiv 0$, el problema de valores de frontera se llama homogéneo. En general, se supone que por lo menos $f_0, \ldots, f_n \in C^0(a, b)$.

El problema de valores de frontera lineal de segundo orden es

$$\begin{aligned} (Ly)(x) &\equiv f_2(x) y''(x) + f_1(x) y'(x) + f_0(x) y(x) = g(x), \\ R_1[y] &= \alpha_{11} y(a) + \beta_{11} y(b) + \alpha_{12} y'(a) + \beta_{12} y'(b) = \alpha_1, \\ R_2[y] &= \alpha_{21} y(a) + \beta_{21} y(b) + \alpha_{22} y'(a) + \beta_{22} y'(b) = \alpha_2. \end{aligned}$$

En muchos casos,

$$\begin{aligned} R_1[y] &= \alpha_{11} y(a) + \alpha_{12} y'(a) = \alpha_1, \\ R_2[y] &= \beta_{21} y(b) + \beta_{22} y'(b) = \alpha_2. \end{aligned}$$

Más generalmente, las condiciones de frontera se llaman *separadas* si

$$\begin{aligned} &\beta_{j,k+1} = 0, \quad j = 1, \ldots, m, \quad k = 0, \ldots, n-1, \quad m < n; \\ &\alpha_{j,k+1} = 0, \quad j = m+1, \ldots, n; \quad k = 0, \ldots, n-1. \end{aligned}$$

Las condiciones de frontera se llaman *linealmente independientes* si el rango de la siguiente matriz es n:

$$\boldsymbol{A} = \begin{bmatrix} \alpha_{11} & \dots & \alpha_{1n} & \beta_{11} & \dots & \beta_{1n} \\ \vdots & & \vdots & \vdots & & \vdots \\ \alpha_{n1} & \dots & \alpha_{nn} & \beta_{n1} & \dots & \beta_{nn} \end{bmatrix}$$

Frecuentamente, el problema de valores de frontera (3.3) puede ser reducido facilmente a un problema con condiciones homogéneas, donde $\alpha_j = 0$ para $j = 1, \dots, n$. Esto ocurre si existe un polinomio $Q \in \Pi_{n-1}$ tal que

$$R_j[Q] = \alpha_j, \quad j = 1, \dots, n.$$

En este caso definimos

$$z(x) := y(x) - Q(x), \quad f := LQ.$$

Según (3.3), obtenemos entonces el problema de valores de frontera

$$\begin{aligned} (Lz)(x) = (Ly - LQ)(x) = g(x) - f(x) = h(x), \\ R_j[z] = R_j[y] - R_j[Q] = \alpha_j - \alpha_j = 0, \quad j = 1, \dots, n. \end{aligned}$$

3.1.1. Ecuaciones diferenciales ordinarias autoadjuntas. Para un operador diferencial L dado por (3.3) definimos el *operador adjunto* L^* a través de

$$L^* y \equiv \sum_{j=0}^{n} (-1)^j \frac{\mathrm{d}^j}{\mathrm{d}x^j}\big(f_j(x) y\big).$$

El operador se llama *autoadjunto* si

$$Ly \equiv L^* y, \quad y \in C^n(a, b). \tag{3.4}$$

Obviamente, un operador diferencial autoadjunto es de orden par. Una ecuación $Ly = g(x)$ con un operador diferencial L autoadjunto se llama *ecuación autoadjunta*. Para $n = 2$, la condición (3.4) exige que

$$\begin{aligned} f_0 y + f_1 y' + f_2 y'' &\equiv f_0 y - (f_1 y)' + (f_2 y)'' \\ &\equiv f_0 y - f_1' y - f_1 y' + f_2'' y + 2 f_2' y' + f_2 y'' \end{aligned}$$

donde omitimos el argumento "(x)". Esto implica que

$$2(f_1 - f_2')y' - (f_2'' - f_1')y \equiv 0, \quad y \in C^2(a, b),$$

lo cual es válido si y sólo si

$$f_1(x) \equiv f_2'(x).$$

Usando $f_2(x) = -p(x)$ y $f_0(x) = q(x)$, podemos escribir una ecuación diferencial ordinaria de segundo orden autoadjunta como

$$Ly = -\frac{\mathrm{d}}{\mathrm{d}x}\left(p(x)\frac{\mathrm{d}y}{\mathrm{d}x}\right) + q(x)y = g. \tag{3.5}$$

(Un aspecto que se discutirá más adelante es que (3.5) también es la ecuación de Euler del problema variacional

$$I[y] := \frac{1}{2}\int_a^b \big(p(x)(y')^2 + q(x)y^2 - 2g(x)y\big)\,\mathrm{d}x \stackrel{!}{=} \text{mín}\,.)$$

Teorema 3.1. *Cada ecuación lineal de segundo orden*

$$f_2(x)y'' + f_1(x)y' + f_0(x)y - h(x) = 0, \quad f_2(x) \neq 0 \quad \text{para todo } x \in (a,b) \tag{3.6}$$

puede ser transformada a una ecuación autoadjunta de segundo orden.

Demostración. Multiplicamos (3.6) por la función

$$-p(x) = \exp\left(\int_{x_0}^{x} \frac{f_1(\xi)}{f_2(\xi)}\mathrm{d}\xi\right), \quad x_0, x \in (a,b),$$

donde $x_0 \in (a,b)$ puede ser elegido libremente. El resultado es

$$-p(x)f_2(x)y'' - p(x)f_1(x)y' - p(x)f_0(x)y + p(x)h(x) = 0. \tag{3.7}$$

La función $p(x)$ satisface

$$-p(x)f_1(x) = -p(x)\frac{f_1(x)}{f_2(x)}f_2(x) = -p'(x)f_2(x),$$

es decir, dividiendo (3.7) por $f_2(x)$ y definiendo

$$q(x) := -p(x)\frac{f_0(x)}{f_2(x)}, \quad g(x) := -p(x)\frac{h(x)}{f_2(x)},$$

obtenemos la ecucación autoadjunta

$$-\frac{\mathrm{d}}{\mathrm{d}x}\left(p(x)\frac{\mathrm{d}y}{\mathrm{d}x}\right) + q(x)y - g(x) = 0.$$

■

Para el tratamiento de problemas de valores de frontera de ecuaciones diferenciales ordinarias de segundo orden podríamos limitarnos a ecuaciones autoadjuntas. Pero, si la ecuación del problema dado no es autoadjunta, la reducción al tipo autoadjunto según la demostración del Teorema 3.1 frecuentamente entrega una ecuación bastante complicada. Por ello es más adecuado resolver el problema en la forma originalmente dada, siempre que existe un método adecuado.

3.1.2. Problemas de valores de frontera para sistemas de ecuaciones diferenciales ordinarias de primer orden. Similarmente a lo que hemos visto para problema de valores iniciales, podemos reducir problemas de valores de frontera de una ecuación diferencial ordinaria del orden n a sistemas de ecuaciones diferenciales de primer orden. Para ilustrar eso, supongamos que la ecuación (3.1) está dada en la forma explícita

$$y^{(n)} = f\big(x, y, y', \ldots, y^{(n-1)}\big),$$

y definimos

$$y^{(j)} =: y_{j+1}, \quad j = 0, \ldots, n-1.$$

Entonces resulta el sistema de primer orden

$$\begin{aligned} y_1' &= y_2, \\ &\vdots \\ y_{n-1}' &= y_n, \\ y_n' &= f(x, y_1, \dots, y_n) \end{aligned} \tag{3.8}$$

con las condiciones de frontera

$$R_j[y_1, \dots, y_n] = R_j\big[y_1(a), \dots, y_n(a); y_1(b), \dots, y_n(b)\big] = \alpha_j, \quad j = 1, \dots, n. \tag{3.9}$$

En particular, si (3.1), (3.2) es un problema de valores de frontera lineal, también el problema (3.8), (3.9) es lineal.

En general, el problema de valores de frontera de un sistema de ecuaciones diferenciales ordinarias de primer orden es

$$\begin{aligned} &y_i' = f_i(x, y_1, \dots, y_n), \quad i = 1, \dots, n, \\ &R_j\big[y_1(a), \dots, y_n(a); y_1(b), \dots, y_n(b)\big] = \alpha_j, \quad j = 1, \dots, n. \end{aligned} \tag{3.10}$$

Definiendo

$$\boldsymbol{y} := \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad \boldsymbol{f}(x, \boldsymbol{y}) := \begin{pmatrix} f_1(x, \boldsymbol{y}) \\ \vdots \\ f_n(x, \boldsymbol{y}) \end{pmatrix}, \quad \boldsymbol{R} := \begin{pmatrix} R_1 \\ \vdots \\ R_n \end{pmatrix}, \quad \boldsymbol{\alpha} := \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{pmatrix},$$

podemos escribir (3.10) como

$$\boldsymbol{y}' = \boldsymbol{f}(x, \boldsymbol{y}), \quad \boldsymbol{R}\big[\boldsymbol{y}(a); \boldsymbol{y}(b)\big] = \boldsymbol{\alpha}. \tag{3.11}$$

El problema (3.11) se llama *lineal* si posee la siguiente forma, donde $\boldsymbol{A}(x)$, $\boldsymbol{B}_1$ y $\boldsymbol{B}_2$ son matrices:

$$\boldsymbol{y}' = \boldsymbol{A}(x)\boldsymbol{y} + \boldsymbol{g}(x), \quad \boldsymbol{B}_1\boldsymbol{y}(a) + \boldsymbol{B}_2\boldsymbol{y}(b) = \boldsymbol{\alpha}.$$

3.2. Métodos de diferencias finitas

3.2.1. Método de diferencias finitas para un problema de valores de frontera lineal. En lo siguiente, consideramos el problema de valores de frontera

$$-y'' + p(x)y' + q(x)y - g(x) = 0, \quad x \in (a, b), \tag{3.12}$$

$$\alpha_{11}y(a) + \alpha_{12}y'(a) = \alpha_1, \tag{3.13}$$

$$\beta_{21}y(b) + \beta_{22}y'(b) = \alpha_2. \tag{3.14}$$

Para discretizarlo, subdividimos el intervalo $[a, b]$ en N subintervalos del tamaño h poniendo

$$x_0 := a; \quad x_i = a + ih, \quad i = 0, \dots, N; \quad x_N = a + Nh = b.$$

En el punto x_i se aproxima la solución del problema de valores de frontera (3.12)–(3.14) remplazando el cuociente diferencial por un cuociente de diferencias. Un tal método se llama *método de diferencias finitas*.

Si $y \in C^4$ es la solución de (3.12)–(3.14), entonces en el punto $x = x_i$ tenemos que

$$-\frac{y(x_{i-1}) - 2y(x_i) + y(x_{i+1})}{h^2} + p(x_i)\frac{y(x_{i+1}) - y(x_{i-1})}{2h} + q(x_i)y(x_i) - g(x_i) = \mathcal{O}(h^2), \tag{3.15}$$

$$\alpha_{11}y(a) + \alpha_{12}\frac{y(x_1) - y(a)}{h} = \alpha_1 + \mathcal{O}(h),$$
$$\beta_{21}y(b) + \beta_{22}\frac{y(b) - y(x_{N-1})}{h} = \alpha_2 + \mathcal{O}(h).$$

Despreciando los términos $\mathcal{O}(h^2)$ y $\mathcal{O}(h)$ y remplazando $y(x_i)$ por y_i^h, obtenemos el sistema de ecuaciones lineales

$$\begin{gathered}(-\alpha_{12} + h\alpha_{11})y_0^h + \alpha_{12}y_1^h = h\alpha_1,\\ -\left(1 + \frac{h}{2}p(x_i)\right)y_{i-1}^h + \big(2 + h^2q(x_i)\big)y_i^h - \left(1 - \frac{h}{2}p(x_i)\right)y_{i+1}^h = h^2g(x_i), \quad i = 1, \dots, N-1,\\ -\beta_{22}y_{N-1}^h + (\beta_{22} + h\beta_{21})y_N^h = h\alpha_2.\end{gathered} \tag{3.16}$$

Para un valor de h fijo y definiendo

$$\varphi_i := 1 + \frac{h}{2}p(x_i), \quad \psi_i := 2 + h^2q(x_i), \quad \bar{\varphi}_i := 1 - \frac{h}{2}p(x_i), \quad i = 1, \dots, N-1,$$

podemos escribir (3.16) como el sistema

$$\boldsymbol{A}(h)\boldsymbol{y}^h = \boldsymbol{b}(h), \tag{3.17}$$

donde $\boldsymbol{A}(h) \in \mathbb{R}^{(N+1)\times(N+1)}$ es la matriz tridiagonal

$$\boldsymbol{A}(h) = \begin{bmatrix} -\alpha_{12} + h\alpha_{11} & \alpha_{12} & 0 & \dots & 0 \\ -\varphi_1 & \psi_1 & -\bar{\varphi}_1 & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & -\varphi_{N-1} & \psi_{N-1} & -\bar{\varphi}_{N-1} \\ 0 & \dots & 0 & -\beta_{22} & \beta_{22} + h\beta_{21} \end{bmatrix}$$

y definimos los vectores

$$\boldsymbol{y}^h := \begin{pmatrix} y_0^h \\ \vdots \\ y_N^h \end{pmatrix}, \quad \boldsymbol{b}(h) := \begin{pmatrix} h\alpha_1 \\ h^2g(x_1) \\ \vdots \\ h^2g(x_{N-1}) \\ h\alpha_2 \end{pmatrix}.$$

Teorema 3.2. *Bajo las siguientes condiciones existe una constante $h_0 > 0$ tal que el sistema (3.17) tiene una solución única $\boldsymbol{y}^h$:*

1. $\alpha_{11} > 0$, $\alpha_{12} \leqslant 0$, $\beta_{21} > 0$ *y* $\beta_{22} \geqslant 0$,
2. $q(x) \geqslant 0$ *y* $h_0|p(x)| < 2$ *para* $x \in [a, b]$.

Para la demostración del Teorema 3.2, usaremos los siguientes conceptos ya introducidos en el Análisis Numérico II.

Definición 3.1. *Una matriz* $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ *se llama* M-matriz *si* $\alpha_{ij} \leqslant 0$ *para* $i \neq j$ *y* $\boldsymbol{A}^{-1}$ *existe y* $\boldsymbol{A}^{-1} \geqslant 0$.

Teorema 3.3. *Sea* $\boldsymbol{A}$ *estrictamente o irreduciblemente diagonaldominante con* $\alpha_{ii} > 0$ *para* $i = 1, \ldots, n$ *y* $\alpha_{ij} \leqslant 0$ *para* $i \neq j$ *(es decir,* $\boldsymbol{A}$ *es una* L-matriz*). En este caso,* $\boldsymbol{A}$ *es una M-matriz.*

Demostración del Teorema 3.2. Observamos primero que $\boldsymbol{A}(h)$ es una L-matriz. Si $\alpha_{12} < 0$ y $\beta_{22} > 0$, entonces $\boldsymbol{A}(h)$ es irreduciblemente diagonaldominante, y por lo tanto una M-matriz. Por otro lado, consideremos el caso $\alpha_{12} = 0$ o $\beta_{22} = 0$. Si por ejemplo $\alpha_{12} = 0$, tenemos

$$y_0^h = \frac{\alpha_1}{\alpha_{11}} = y(a).$$

Si $\beta_{22} > 0$ en este caso, después de remplazar $y_0^h = y(a)$, $0 < h \leqslant h_0$ el sistema (3.17) se reduce a un sistema de N ecuaciones en las desconocidas $\boldsymbol{y}^h = (y_1^h, \ldots, y_N^h)^{\mathrm{T}}$, cuya matriz nuevamente es una M-matriz. Si además $\beta_{22} = 0$, tenemos que

$$y_N^h = \frac{\alpha_2}{\beta_{21}} = y(b),$$

y obtenemos un sistema lineal de $N-1$ ecuaciones para $\boldsymbol{y}^h = (y_1^h, \ldots, y_{N-1}^h)^{\mathrm{T}}$, cuya matriz es una L-matriz irreduciblemente diagonaldominante, es decir, una M-matriz. ■

Para la solución del sistema lineal (3.17), podríamos utilizar el algoritmo de Gauss (o el algoritmo de Thomas). Sin embargo, para valores de N muy grandes, es decir, para h muy pequeño, este algoritmo es poco apropiado puesto que $\boldsymbol{A}(h)$ es casi singular. La mala condición de $\boldsymbol{A}(h)$ en este caso tendrá como consecuencia que el resultado será substancialmente falsificado por errores de redondeo.

Los métodos numéricos iterativos para sistemas lineales han sido desarrollados para el tratamiento de aquellas matrices que provienen de discretizaciones de problemas de ecuaciones diferenciales. La convergencia del método SOR fue demostrada para sistemas lineales con una M-matriz y un parámetro de relajación $0 < \omega \leqslant 1$. La velocidad de convergencia también depende del número de condición de $\boldsymbol{A}(h)$. Por otro lado, hay que considerar que es suficiente resolver el sistema (3.17) solamente aproximadamente, dado que ya se cometió un error al remplazar la ecuación diferencial por su discretización, y la solución exacta de (3.17) es solamente una aproximación a la verdadera solución de la ecuación diferencial.

Para el caso $\alpha_{12} = \beta_{22} = 0$, podemos ilustrar que $\boldsymbol{A}(h)$ es casi singular. Como $h = (b-a)/N \approx 0$ si N es muy grande, la matriz $\boldsymbol{A}(h)$ aproxima

$$\boldsymbol{A} = \begin{bmatrix} 2 & -1 & 0 & \ldots & 0 \\ -1 & 2 & -1 & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & -1 & 2 & -1 \\ 0 & \cdots & 0 & -1 & 2 \end{bmatrix}.$$

Esta matriz tiene los valores propios

$$\lambda_j = 2\left[1 - \cos\left(\frac{j\pi}{N}\right)\right], \quad j = 1, \ldots, N-1.$$

Es decir, el valor propio mínimo, λ_1, satisface $\lambda_1 < 10^{-3}$ para $N = 100$ y $\lambda_1 < 10^{-5}$ para $N = 1000$, mientras que $\lambda_{N-1} \approx 4$.

Una matriz simétrica resulta para $p(x) \equiv 0$, $\alpha_{12} = -\beta_{22} = -1$ o $\alpha_{12} = \beta_{22} = 0$. En este caso, el problema de valores de frontera (3.12)–(3.14) asume la forma

$$\begin{aligned} &-y'' + q(x)y - g(x) = 0, \quad x \in (a,b), \\ &\alpha_{11}y(a) - y'(a) = \alpha_1, \quad \beta_{21}y(b) + y'(b) = \alpha_2 \\ \text{o alternativamente} \quad &\alpha_{11}y(a) = \alpha_1, \quad \beta_{21}y(b) = \alpha_2 \quad (\alpha_{11}, \beta_{21} \neq 0). \end{aligned}$$

Para una ecuación autoadjunta *siempre* podemos hallar una aproximación de diferencias tal que la matriz del sistema lineal que resulta no sólo es simétrica, sino que tambíen definida positiva.

3.2.2. Método de diferencias finitas para un problema de valores de frontera no lineal. Los métodos de diferencias finitas también pueden ser usados para la solución numérica de problemas no lineales. A modo de ejemplo, estudiamos el problema

$$-y'' + f(x, y, y') = 0, \quad a < x < b; \quad y(a) = \alpha, \quad y(b) = \beta.$$

Se supone que la ecuación es no lineal, es decir, f es una función no lineal de y o de y'. Se supone además que f es diferenciable con respecto a y e y', y que $|f_{y'}| \leqslant M$. La discretización entrega en este caso el sistema no lineal

$$-\frac{1}{h^2}\left(y_{i-1}^h - 2y_i^h + y_{i+1}^h\right) + f\left(x_i^h, y_i^h, \frac{y_{i+1}^h - y_{i-1}^h}{2h}\right) = 0, \quad i = 1, \ldots, N-1.$$

Multiplicando por h^2 y definiendo $\boldsymbol{y}^h = (y_1^h, \ldots, y_{N-1}^h)^{\mathrm{T}}$, obtenemos

$$t_i(\boldsymbol{y}^h) := -y_{i-1}^h + 2y_i^h - y_{i+1}^h + h^2 f\left(x_i^h, y_i^h, \frac{y_{i+1}^h - y_{i-1}^h}{2h}\right) = 0, \quad i = 1, \ldots, N-1, \quad (3.18)$$

es decir

$$\boldsymbol{T}(\boldsymbol{y}^h) = 0, \quad \boldsymbol{T}(\boldsymbol{y}^h) := \left(t_1(\boldsymbol{y}^h), \ldots, t_{N-1}(\boldsymbol{y}^h)\right)^{\mathrm{T}}.$$

La matriz funcional

$$\boldsymbol{T}'(\boldsymbol{z}) := \left(\frac{\partial t_i}{\partial z_j}\right)_{i,j=1,\ldots,N-1}(\boldsymbol{z})$$

puede ser evaluada facilmente: usando

$$f_y\left(x_i, z_i, \frac{z_{i+1} - z_{i-1}}{2h}\right) =: f_y^{(i)}, \quad f_{y'}\left(x_i, z_i, \frac{z_{i+1} - z_{i-1}}{2h}\right) =: f_{y'}^{(i)},$$

tenemos

$$\frac{\partial t_i}{\partial z_j} = 0, \quad j \notin \{i-1, i, i+1\},$$

$$\frac{\partial t_i}{\partial z_{i-1}} = -1 - \frac{h}{2} f_{y'}^{(i)}, \quad \frac{\partial t_i}{\partial z_i} = 2 + h^2 f_y^{(i)}, \quad \frac{\partial t_i}{\partial z_{i+1}} = -1 + \frac{h}{2} f_{y'}^{(i)}.$$

Además, tenemos $z_0 = \alpha$ y $z_N = \beta$ en (3.18). Definiendo

$$c_i := -1 - \frac{h}{2} f_{y'}^{(i)}, \quad d_i := 2 + h^2 f_y^{(i)}, \quad e_i := -1 + \frac{h}{2} f_{y'}^{(i)}, \quad i = 1, \ldots, N-1,$$

podemos escribir

$$\boldsymbol{T}'(\boldsymbol{z}) = \begin{bmatrix} d_1 & e_1 & 0 & \ldots & 0 \\ c_2 & d_2 & e_2 & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & c_{N-2} & d_{N-2} & e_{N-2} \\ 0 & \ldots & 0 & c_{N-1} & d_{N-1} \end{bmatrix}.$$

El sistema $\boldsymbol{T}(\boldsymbol{y}^h) = 0$ debe ser solucionado iterativamente, por ejemplo usando el método SOR-Newton.

El método SOR-Newton es apto para resolver numéricamente el sistema no lineal

$$\boldsymbol{F}(\boldsymbol{x}) = 0, \quad \boldsymbol{F}(\boldsymbol{x}) = \begin{pmatrix} f_1(\boldsymbol{x}) \\ \vdots \\ f_n(\boldsymbol{x}) \end{pmatrix}, \quad \boldsymbol{F} : \mathbb{R}^n \supset \mathcal{B} \to \mathbb{R}^n, \quad \boldsymbol{F} \in \big(C^2(\mathcal{B})\big)^n,$$

con la solución exacta $\boldsymbol{x}^* \in \mathcal{B}$. El método es definido por la recursión

$$x_i^{(k+1)} = x_i^{(k)} - \omega \frac{f_i(\boldsymbol{x}^{(k),i})}{\partial_i f_i(\boldsymbol{x}^{(k),i})}, \quad i = 1, \ldots, n,$$

donde definimos

$$\boldsymbol{x}^{(k),i} := \big(x_1^{(k+1)}, \ldots, x_{i-1}^{(k+1)}, x_i^{(k)}, \ldots, x_n^{(k)}\big)^{\mathrm{T}}, \quad i = 1, \ldots, n.$$

Teorema 3.4. *Existe una vecindad $\mathcal{B}_0$ de $\boldsymbol{x}^*$, $\mathcal{B}_0 \subset \mathcal{B}$, tal que para todo $\boldsymbol{x}^{(0)} \in \mathcal{B}_0$ el método SOR-Newton converge si*

(a) $0 < \omega \leqslant 1$ *y* $\boldsymbol{F}'(\boldsymbol{x}^*)$ *es estrictamente o irreduciblemente diagonal dominante o*
(b) $0 < \omega \leqslant 2$ *y* $\boldsymbol{F}'(\boldsymbol{x}^*)$ *es simétrica y definida positiva o*
(c) $0 < \omega \leqslant 1$ *y* $\boldsymbol{F}'(\boldsymbol{x}^*)$ *es una M-matriz.*

En nuestro caso, tenemos el siguiente teorema.

Teorema 3.5. *Sea $f_y \geqslant 0$ y h tan pequeño que $hM < 2$. Entonces para todo $\boldsymbol{z} \in \mathbb{R}^{N-1}$, $\boldsymbol{T}'(\boldsymbol{z})$ es una matriz irreduciblemente diagonal dominante.*

Demostración. Dado que $|f_{y'}| \leqslant M$, tenemos

$$-1 + \frac{h}{2} f_{y'}^{(i)} < 0, \quad -1 - \frac{h}{2} f_{y'}^{(i)} < 0, \quad 2 + h^2 f_y^{(i)} \geqslant 2,$$

y luego

$$\left|1 + \frac{h}{2} f_{y'}^{(i)}\right| + \left|1 - \frac{h}{2} f_{y'}^{(i)}\right| = 2 \leqslant 2 + h^2 f_y^{(i)}, \quad i = 2, \ldots, N-2.$$

Finalmente, tenemos

$$\left|1+\frac{h}{2}f_{y'}^{(j)}\right| < 2 \leqslant 2+h^2 f_y^{(j)}, \quad j \in \{1, N-1\}.$$

Concluimos que $\boldsymbol{T}'(\boldsymbol{z})$ es diagonal dominante para todo $\boldsymbol{z} \in \mathbb{R}^{N-1}$, y la diagonaldominancia es estricta en la primera y la última fila. Además, las matriz es irreducible, lo que concluye la demostración del teorema. ■

3.2.3. Convergencia del método para problemas lineales. Recordamos que un método se llama *del orden* p, $p > 0$, si

$$y_i^h - y(x) = \mathcal{O}(h^p), \quad h \to 0, \quad x = a + ih \in [a, b].$$

Por simplicidad, consideremos ahora el problema de valores de frontera

$$-y'' + q(x)y - g(x) = 0, \quad a < x < b; \quad y(a) = \alpha, \quad y(b) = \beta. \tag{3.19}$$

Suponiendo que la solución satisface $y \in C^4[a, b]$, tenemos (3.15) con $p(x) \equiv 0$, y para la computación de los valores aproximados obtenemos el sistema (3.17), en nuestro caso

$$\boldsymbol{A}(h)\boldsymbol{y}^h = \boldsymbol{b}(h)$$

con la matriz

$$\boldsymbol{A}(h) = \begin{bmatrix} 2+h^2q(x_1) & -1 & 0 & \cdots & 0 \\ -1 & 2+h^2q(x_2) & -1 & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & -1 & 2+h^2q(x_{N-2}) & -1 \\ 0 & \cdots & 0 & -1 & 2+h^2q(x_{N-1}) \end{bmatrix} \tag{3.20}$$

y el vector

$$\boldsymbol{b}(h) = \begin{pmatrix} \alpha + h^2 g(x_1) \\ h^2 g(x_2) \\ \vdots \\ h^2 g(x_{N-2}) \\ \beta + h^2 g(x_{N-1}) \end{pmatrix}. \tag{3.21}$$

Sean $y(x_i)$ los valores de la solución exacta de (3.19) en $x_i = a + ih$, $i = 0, \ldots, N$, con

$$y(x_0) = y(a) = \alpha, \quad y(x_N) = y(b) = \beta, \quad \boldsymbol{y}(h) = \big(y(x_1), \ldots, y(x_{N-1})\big)^{\mathrm{T}}.$$

Entonces en virtud de (3.15) sabemos que

$$\boldsymbol{A}(h)\boldsymbol{y}(h) = \boldsymbol{b}(h) + \mathcal{O}(h^4),$$

donde $\mathcal{O}(h^4)$ denota un vector de $\mathbb{R}^{N-1}$. Entonces, la cantidad $\boldsymbol{\varepsilon}_h := \boldsymbol{y}^h - \boldsymbol{y}(h)$ satisface

$$\boldsymbol{\varepsilon}_h = \boldsymbol{A}(h)^{-1}\mathcal{O}(h^4). \tag{3.22}$$

Las componentes de $\mathcal{O}(h^4)$ son la cuarta derivada de la solución exacta y, evaluada en puntos intermedios y multiplicada por $h^4/12$. La cuarta derivada $y^{(4)}(x)$ está acotada con respecto a $\|\cdot\|_\infty$ según hipótesis. Entonces, existe una constante K tal que

$$\|\boldsymbol{\varepsilon}_h\|_\infty \leqslant K\big\|\boldsymbol{A}(h)^{-1}\big\|_\infty h^4.$$

Obviamente, la dificultad es ¿cómo estimar $\|\boldsymbol{A}(h)^{-1}\|_\infty$? Sean $\boldsymbol{G} = (g_{ij})$ y $\boldsymbol{H} = (h_{ij})$ dos matrices de $\mathbb{R}^{n\times m}$, con $g_{ij} \leqslant h_{ij}$ para todo $1 \leqslant i \leqslant n$ y $1 \leqslant j \leqslant m$. En este situación escribimos $\boldsymbol{G} \leqslant \boldsymbol{H}$. Si una matriz $\boldsymbol{B}$ tiene sólo elementos no negativos, entonces escribimos $\boldsymbol{B} \geqslant 0$, donde "0" representa la 0-matriz. Según (3.20),

$$\boldsymbol{A}(h) = \tilde{\boldsymbol{A}} + h\boldsymbol{Q}, \quad \tilde{\boldsymbol{A}} = \begin{bmatrix} 2 & -1 & 0 & \cdots & 0 \\ -1 & 2 & -1 & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & -1 & 2 & -1 \\ 0 & \cdots & 0 & -1 & 2 \end{bmatrix}, \quad \boldsymbol{Q} = \operatorname{diag}\big(q(x_1), \ldots, q(x_{N-1})\big).$$

Teorema 3.6. *Sean $q(x_i) \geqslant 0$ para $i = 1, \ldots, N-1$. Entonces*

$$\boldsymbol{A}(h)^{-1} \leqslant \tilde{\boldsymbol{A}}^{-1}. \tag{3.23}$$

Demostración. Dado que $q(x_i) \geqslant 0$, las matrices $\boldsymbol{A}(h)$ y $\tilde{\boldsymbol{A}}$ son M-matrices irreduciblemente diagonaldominantes y simétricas. Entonces

$$\boldsymbol{A}(h)^{-1} \geqslant 0, \quad \tilde{\boldsymbol{A}}^{-1} \geqslant 0. \tag{3.24}$$

Ahora podemos escribir $\tilde{\boldsymbol{A}} = \boldsymbol{D} - \boldsymbol{L} - \boldsymbol{U}$ con

$$\boldsymbol{D} = \operatorname{diag}(2, \ldots, 2), \quad \boldsymbol{L} = \begin{bmatrix} 0 & \cdots & \cdots & \cdots & 0 \\ 1 & \ddots & & & \vdots \\ 0 & \ddots & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & 1 & 0 \end{bmatrix}, \quad \boldsymbol{U} = \boldsymbol{L}^{\mathrm{T}}.$$

Entonces

$$\boldsymbol{A}(h) = \boldsymbol{D} - \boldsymbol{L} - \boldsymbol{U} + h^2\boldsymbol{Q},$$

y poniento

$$\bar{\boldsymbol{D}}(h) := \boldsymbol{D} + h^2\boldsymbol{Q}$$

obtenemos

$$\boldsymbol{A}(h) = \bar{\boldsymbol{D}}(h) - \boldsymbol{L} - \boldsymbol{U}.$$

Además,

$$\boldsymbol{D}^{-1}\tilde{\boldsymbol{A}} = \boldsymbol{I} - \boldsymbol{D}^{-1}(\boldsymbol{L} + \boldsymbol{U}), \quad \bar{\boldsymbol{D}}(h)^{-1}\boldsymbol{A}(h) = \boldsymbol{I} - \bar{\boldsymbol{D}}(h)^{-1}(\boldsymbol{L} + \boldsymbol{U}).$$

Obviamente, $\boldsymbol{D} \leqslant \bar{\boldsymbol{D}}(h)$, $\boldsymbol{D}^{-1} \geqslant \bar{\boldsymbol{D}}(h)^{-1}$, y entonces $-\boldsymbol{D}^{-1} \leqslant -\bar{\boldsymbol{D}}(h)^{-1}$. Usando (3.24), obtenemos

$$\begin{aligned}\big(\boldsymbol{A}(h)^{-1}\boldsymbol{D}(h)\big)\big(\boldsymbol{I} - \bar{\boldsymbol{D}}(h)^{-1}(\boldsymbol{L}+\boldsymbol{U})\big) &\leqslant \big(\boldsymbol{A}(h)^{-1}\bar{\boldsymbol{D}}(h)\big)\big(\boldsymbol{I} - \bar{\boldsymbol{D}}(h)^{-1}(\boldsymbol{L}+\boldsymbol{U})\big)\\ &= \big(\bar{\boldsymbol{D}}(h)^{-1}\boldsymbol{A}(h)\big)^{-1}\big(\boldsymbol{I} - \bar{\boldsymbol{D}}(h)^{-1}(\boldsymbol{L}+\boldsymbol{U})\big)\\ &= \boldsymbol{I}\\ &= (\boldsymbol{D}^{-1}\tilde{\boldsymbol{A}})^{-1}\big(\boldsymbol{I} - \boldsymbol{D}^{-1}(\boldsymbol{L}+\boldsymbol{U})\big)\\ &\leqslant (\tilde{\boldsymbol{A}}^{-1}\boldsymbol{D})\big(\boldsymbol{I} - \bar{\boldsymbol{D}}(h)^{-1}(\boldsymbol{L}+\boldsymbol{U})\big).\end{aligned}$$

Dado que $\bar{\boldsymbol{D}}(h) > 0$ y $\boldsymbol{D} > 0$, podemos concluir que como

$$\boldsymbol{A}(h) = \bar{\boldsymbol{D}}(h) - \boldsymbol{L} - \boldsymbol{U}$$

es una M-matriz, también

$$\boldsymbol{D}\cdot\bar{\boldsymbol{D}}(h)^{-1}\boldsymbol{A}(h) = \boldsymbol{D}\big(\boldsymbol{I} - \bar{\boldsymbol{D}}(h)^{-1}(\boldsymbol{L}+\boldsymbol{U})\big)$$

es una M-matriz, y por lo tanto

$$\Big(\boldsymbol{D}\big(\boldsymbol{I} - \bar{\boldsymbol{D}}(h)^{-1}(\boldsymbol{L}+\boldsymbol{U})\big)\Big)^{-1} \geqslant 0.$$

Dado que $\tilde{\boldsymbol{A}}$ es una M-matriz, concluimos que (3.23) es válido. ■

Teorema 3.7. *Sea $y^{(4)} \in C[a,b]$ y $|y^{(4)}(x)| \leqslant M$ para $x \in [a,b]$. Entonces existe una constante $L \geqslant 0$ tal que*

$$\big|y_i^h - y(x_i)\big| \leqslant Li(N-i)h^4 = L(x_i - a)(b - x_i)h^2.$$

Demostración. Sea $\boldsymbol{e} = (1,\ldots,1)^{\mathrm{T}} \in \mathbb{R}^{N-1}$ y para $\boldsymbol{u} \in \mathbb{R}^{N-1}$ escribimos

$$|\boldsymbol{u}| := (|u_1|,\ldots,|u_{N-1}|)^{\mathrm{T}}.$$

Entonces, debido a $\boldsymbol{A}^{-1}(h) \geqslant 0$, (3.22) y (3.23) implican que

$$|\boldsymbol{\varepsilon}_h| = \frac{h^4}{12}M\boldsymbol{A}(h)^{-1}\boldsymbol{e} \leqslant \frac{h^4}{12}M\tilde{\boldsymbol{A}}^{-1}\boldsymbol{e}. \tag{3.25}$$

Hay que calcular $\tilde{\boldsymbol{A}}^{-1}\boldsymbol{e}$.

Ahora, los valores de la solución aproximada coinciden con la solución exacta si la solución es un polinomio del grado 2. La solución única de

$$-y'' = 1, \quad y(a) = y(b) = 0$$

es

$$y = -\frac{1}{2}(x-a)(x-b). \tag{3.26}$$

Dado que $q(x) \equiv 0$, $g(x) \equiv 1$, $\alpha = \beta = 0$, según (3.20) y (3.21) obtenemos en este caso $\boldsymbol{A}(h) = \tilde{\boldsymbol{A}}$ y $\boldsymbol{b}(h) = h^2\boldsymbol{e}$. Entonces, con

$$\tilde{\boldsymbol{y}}(h) = \left(-\frac{1}{2}(x_1-a)(x_1-b),\ldots,-\frac{1}{2}(x_{N-1}-a)(x_{N-1}-b)\right)^{\mathrm{T}}$$

tenemos

$$\tilde{\boldsymbol{A}}\boldsymbol{y}^h = h^2\boldsymbol{e}, \quad \boldsymbol{y}^h = h^2\tilde{\boldsymbol{A}}^{-1}\boldsymbol{e} = \tilde{\boldsymbol{y}}(h). \tag{3.27}$$

Usando (3.26), tenemos

$$y_i^h = y(x_i) = -\frac{1}{2}(x_i - x_0)(x_i - x_N) = \frac{1}{2}hi(N-i).$$

Con $L = M/24$ obtenemos de (3.25) y (3.27)

$$|\boldsymbol{\varepsilon}_h| \leqslant \frac{1}{12}h^4 M\frac{1}{h^2}\tilde{\boldsymbol{y}}(h) = 2Lh^2\tilde{\boldsymbol{y}}(h), \tag{3.28}$$

es decir

$$\left|y_i^h - y(x_i)\right| \leqslant Li(N-i)h^4 = L(x_i - a)(b - x_i)h^2.$$

■

3.3. Métodos de disparo

La solucion de un problema de valores de frontera puede ser reducida a la solución de un número de problemas de valores iniciales. Esto es la idea básica de los métodos de disparo simple y múltiple. Vamos a estudiar primero la situación en el caso lineal.

3.3.1. Métodos de disparo para problemas lineales. Comenzando con métodos de disparo simple, estudiamos primero el problema

$$\begin{aligned} &L\boldsymbol{y} \equiv \boldsymbol{y}' - \boldsymbol{A}(x)\boldsymbol{y} = \boldsymbol{g}(x), \quad a < x < b, \\ &R\boldsymbol{y} \equiv \boldsymbol{B}_1\boldsymbol{y}(a) + \boldsymbol{B}_2\boldsymbol{y}(b) = \boldsymbol{r}. \end{aligned} \tag{3.29}$$

Aquí $\boldsymbol{A}(x)$ es una matriz $n \times n$ cuyos elementos son funciones continuas de x en $[a, b]$, y las matrices $\boldsymbol{B}_1$ y $\boldsymbol{B}_2$ son constantes. Cualquier problema de valores de frontera de mayor orden lineal puede ser reducido a este tipo.

El concepto del método de disparo es el siguiente. Para un vector de parámetros $\boldsymbol{s} = (s_1, \ldots, s_n)^{\mathrm{T}}$, calculamos la solución $\boldsymbol{z} = \boldsymbol{z}(x; \boldsymbol{s})$ del problema

$$L\boldsymbol{z} = \boldsymbol{g}, \quad \boldsymbol{z}(a) = \boldsymbol{s}; \tag{3.30}$$

luego tratamos de determinar las componentes de $\boldsymbol{s}$ de tal forma que $\boldsymbol{z}$ también satisface las condiciones de frontera, es decir

$$R\boldsymbol{z} = \boldsymbol{B}_1\boldsymbol{z}(a; \boldsymbol{s}) + \boldsymbol{B}_2\boldsymbol{z}(b; \boldsymbol{s}) = \boldsymbol{r}.$$

Eligiendo $\boldsymbol{s}$ de forma correcta, "disparamos" a las condiciones de frontera.

Para el problema (3.29), el método de disparo consiste en los siguientes pasos.

1. Calcular $\boldsymbol{z}^0(x)$ de tal forma que

$$L\boldsymbol{z}^0 = \boldsymbol{g}, \quad \boldsymbol{z}^0(a) = 0. \tag{3.31}$$

2. Calcular un sistema fundamental $\boldsymbol{z}^i(x)$, $i = 1, \ldots, n$, de soluciones del problema homogéneo $L\boldsymbol{z} = 0$, con la condición inicial $\boldsymbol{z}^i(a) = \boldsymbol{e}_i$, donde $\boldsymbol{e}_i$ es el i-ésimo vector unitario. Definiendo la matriz
$$\boldsymbol{Z}(x) := \begin{bmatrix} \boldsymbol{z}^1(x) & \cdots & \boldsymbol{z}^n(x) \end{bmatrix},$$
determinamos
$$\boldsymbol{z}(x; \boldsymbol{s}) = \boldsymbol{z}^0(x) + \boldsymbol{Z}(x)\boldsymbol{s} = \boldsymbol{z}^0(x) + \sum_{i=1}^{n} s_i \boldsymbol{z}^i(x). \tag{3.32}$$
3. Suponiendo que $\boldsymbol{B}_1 + \boldsymbol{B}_2\boldsymbol{Z}(b)$ es no singular, calculamos $\boldsymbol{s}$ como solución del sistema lineal
$$\big(\boldsymbol{B}_1 + \boldsymbol{B}_2\boldsymbol{Z}(b)\big)\boldsymbol{s} = \big(\boldsymbol{r} - \boldsymbol{B}_2\boldsymbol{z}^0(b)\big). \tag{3.33}$$

Para el vector $\boldsymbol{s}$ que resulta de (3.33), la solución $\boldsymbol{z}(x; \boldsymbol{s})$ de (3.32) es la solución $\boldsymbol{z}^*(x)$ del problema de valores de frontera (3.29). Para verificar eso, notamos primero que (3.32) es la solución del problema de valores iniciales (3.30), ya que
$$L\boldsymbol{z} = L\boldsymbol{z}^0 + \sum_{i=1}^{n} s_i L\boldsymbol{z}^i = \boldsymbol{g} + 0 = \boldsymbol{g},$$
$$\boldsymbol{z}(a; \boldsymbol{s}) = \boldsymbol{z}^0(a) + \sum_{i=1}^{n} s_i \boldsymbol{z}^i(a) = 0 + \sum_{i=1}^{n} s_i \boldsymbol{e}_i = \boldsymbol{s}.$$
Luego, del requerimiento $R\boldsymbol{z} = \boldsymbol{r}$ obtenemos
$$\begin{aligned} \boldsymbol{r} &= R(\boldsymbol{z}^0 + \boldsymbol{Z}\boldsymbol{s}) \\ &= \boldsymbol{B}_1\big(\boldsymbol{z}^0(a) + \boldsymbol{Z}(a)\boldsymbol{s}\big) + \boldsymbol{B}_2\big(\boldsymbol{z}^0(b) + \boldsymbol{Z}(b)\boldsymbol{s}\big) \\ &= \big(\boldsymbol{B}_1 + \boldsymbol{B}_2\boldsymbol{Z}(b)\big)\boldsymbol{s} + \boldsymbol{B}_2\boldsymbol{z}^0(b), \end{aligned}$$
lo que es precisamente (3.33).

Efectivamente, hemos reducido la solución del problema de valores de frontera (3.29) a la solución de $n+1$ problemas de valores iniciales: un problema de valores iniciales es (3.31); los n demás problemas son
$$L\boldsymbol{z}^i = 0, \quad \boldsymbol{z}^i(a) = \boldsymbol{e}_i, \quad i = 1, \ldots, n, \tag{3.34}$$
cuyas soluciones forman el sistema fundamental $\boldsymbol{Z}(x)$.

Ejemplo 3.1. *Consideremos el problema de valores de frontera*
$$\begin{aligned} y'' - y &= x^2 - 2, \quad 0 < x < 1, \\ y(0) - y(1) &= 0, \\ y'(0) - y'(1) &= 1. \end{aligned}$$
Este problema de valores de frontera es equivalente al problema para un sistema de primer orden
$$\boldsymbol{y}' - \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \boldsymbol{y} = \begin{pmatrix} 0 \\ x^2 - 2 \end{pmatrix}, \quad 0 < x < 1,$$

$$\boldsymbol{y}(0) - \boldsymbol{y}(1) = \begin{pmatrix} 0 \\ 1 \end{pmatrix},$$

es decir, en este caso tenemos

$$\boldsymbol{A} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \quad \boldsymbol{B}_1 = -\boldsymbol{B}_2 = \boldsymbol{I}.$$

Un planteo polinomial entrega

$$\boldsymbol{z}^0(x) = \begin{pmatrix} -x^2 \\ -2x \end{pmatrix}, \quad \text{donde } \boldsymbol{z}(0) = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

Los valores propios de $\boldsymbol{A}$ *son* $\lambda_1 = 1$ *y* $\lambda_2 = -1$*, con los vectores propios correspondientes* $(1,1)^{\mathrm{T}}$ *y* $(1,-1)^{\mathrm{T}}$*. Entonces, la solución general del sistema* $\boldsymbol{y}' - \boldsymbol{A}\boldsymbol{y} = 0$ *es*

$$\boldsymbol{y} = \begin{pmatrix} C_1 \mathrm{e}^x + C_2 \mathrm{e}^{-x} \\ C_1 \mathrm{e}^x - C_2 \mathrm{e}^{-x} \end{pmatrix}.$$

Para obtener el sistema fundamental, es decir los vectores $\boldsymbol{z}^1$ *y* $\boldsymbol{z}^2$*, exigimos que*

$$\boldsymbol{z}^1(0) = \begin{pmatrix} C_{11} + C_{12} \\ C_{11} - C_{12} \end{pmatrix} = \boldsymbol{e}_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad \boldsymbol{z}^2(0) = \begin{pmatrix} C_{21} + C_{22} \\ C_{21} - C_{22} \end{pmatrix} = \boldsymbol{e}_2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix};$$

lo que entrega $C_{11} = C_{12} = \frac{1}{2}$ *y* $C_{21} = -C_{22} = \frac{1}{2}$*, entonces*

$$\boldsymbol{z}^1(x) = \frac{1}{2}\begin{pmatrix} \mathrm{e}^x + \mathrm{e}^{-x} \\ \mathrm{e}^x - \mathrm{e}^{-x} \end{pmatrix} = \begin{pmatrix} \cosh x \\ \sinh x \end{pmatrix}, \quad \boldsymbol{z}^2(x) = \frac{1}{2}\begin{pmatrix} \mathrm{e}^x - \mathrm{e}^{-x} \\ \mathrm{e}^x + \mathrm{e}^{-x} \end{pmatrix} = \begin{pmatrix} \sinh x \\ \cosh x \end{pmatrix}.$$

Luego calculamos que

$$\boldsymbol{Z}(x) = \begin{bmatrix} \cosh x & \sinh x \\ \sinh x & \cosh x \end{bmatrix}, \quad \det \boldsymbol{Z}(x) = 1, \quad \boldsymbol{B}_1 + \boldsymbol{B}_2 \boldsymbol{Z}(1) = \begin{bmatrix} 1 - \cosh 1 & -\sinh 1 \\ -\sinh 1 & 1 - \cosh 1 \end{bmatrix}.$$

Entonces, para $\alpha := \cosh 1$ *y* $\beta := \sinh 1$ *tenemos que resolver el sistema* (3.33)*, es decir*

$$\begin{aligned} (1-\alpha)s_1 - \beta s_2 &= -1, \\ -\beta s_1 + (1-\alpha)s_2 &= -1, \end{aligned}$$

con la solución

$$s_1^* = s_2^* = \frac{\alpha - \beta + 1}{2\beta} =: \gamma = 0{,}58198.$$

Entonces, la solución deseada del problema de valores de frontera es

$$\boldsymbol{z}(x; \boldsymbol{s}^*) = \boldsymbol{z}^*(x) = \begin{pmatrix} -x^2 + \gamma(\cosh x + \sinh x) \\ -2x + \gamma(\sinh x + \cosh x) \end{pmatrix}.$$

Esta solución también satisface las condiciones de frontera $\boldsymbol{z}^*(0) - \boldsymbol{z}^*(1) = (0,1)^{\mathrm{T}}$.

3.3.2. Método de disparo numérico para problemas lineales. En general, los $n+1$ problemas de valores iniciales no pueden ser resueltos exactamente; hay que utilizar métodos numéricos. Se sugiere el siguiente procedimiento. Usando un método de paso simple (o de paso múltiple), se determinan soluciones aproximadas, es decir, vectores

$$\boldsymbol{z}_j^{h,i} = \big(z_{1,j}^{h,i}, \ldots, z_{n,j}^{h,i}\big)^{\mathrm{T}}, \quad j = 0, \ldots, N, \quad i = 0, \ldots, n$$

con los valores iniciales

$$\boldsymbol{z}_0^{h,0} = 0; \quad \boldsymbol{z}_0^{h,i} = \boldsymbol{e}_i, \quad i = 1, \ldots, n.$$

Aquí los vectores $\boldsymbol{z}_j^{h,0}$ son los valores aproximados de la solución de (3.31), evaluada en los puntos de malla x_j, $j = 0, \ldots, N$, y los vectores $\boldsymbol{z}_j^{h,i}$, $i = 1, \ldots, n$, corresponden a los problemas de valores iniciales homogéneos (3.34). Como en el paso 2 mencionado arriba, formamos en cada punto de malla x_j las matrices

$$\boldsymbol{Z}_j^h = \begin{bmatrix} \boldsymbol{z}_j^{h,1} & \cdots & \boldsymbol{z}_j^{h,n} \end{bmatrix}, \quad j = 0, \ldots, N,$$

es decir, las matrices cuyas columnas son los vectores $\boldsymbol{z}_j^{h,i}$. El algoritmo que resulta de estas consideraciones es el siguiente.

1. Usando el método de paso simple o de pasos múltiples, determinamos los vectores $\boldsymbol{z}_j^{h,0}$, donce $\boldsymbol{z}_0^{h,0} = 0$ para $j = 0, \ldots, N$.
2. Para $\boldsymbol{z}_0^{h,i} = \boldsymbol{e}_i$ se calculan (mediante el método de paso simple o de pasos múltiples) los vectores $\boldsymbol{z}_j^{h,i}$, $i = 1, \ldots, n$, $j = 1, \ldots, N$. Ponemos

$$\boldsymbol{z}_j^h = \boldsymbol{z}_j^{h,0} + \boldsymbol{Z}_j^h \boldsymbol{s}^h = \boldsymbol{z}_j^{h,0} + \sum_{i=1}^{n} s_i^h \boldsymbol{z}_j^{h,i}. \tag{3.35}$$

3. Si $\boldsymbol{B}_1 + \boldsymbol{B}_2 \boldsymbol{Z}_N^h$ es regular, resolvemos el sistema

$$(\boldsymbol{B}_1 + \boldsymbol{B}_2 \boldsymbol{Z}_N^h)\boldsymbol{s}^{h,*} = \boldsymbol{r} - \boldsymbol{B}_2 \boldsymbol{z}_N^{h,0}. \tag{3.36}$$

   Para el vector $\boldsymbol{s}^{h,*}$ que resulta de (3.36), (3.35) es, en los puntos de malla $x_0, \ldots, x_N$, una solución aproximada $\boldsymbol{z}_j^{h,*}$ de la solución exacta $\boldsymbol{z}^*(x_j)$ del problema de valores de frontera (3.29).

Efectivamente, la función de malla determinada así satisface las condiciones de frontera

$$\boldsymbol{B}_1 \boldsymbol{z}_0^{h,*} + \boldsymbol{B}_2 \boldsymbol{z}_N^{h,*} = \boldsymbol{r};$$

en virtud de

$$\boldsymbol{r} = \boldsymbol{B}_1\big(\boldsymbol{z}_0^{h,0} + \boldsymbol{Z}_0^h \boldsymbol{s}^h\big) + \boldsymbol{B}_2\big(\boldsymbol{z}_N^{h,0} + \boldsymbol{Z}_N^h \boldsymbol{s}^h\big),$$

y dado que $\boldsymbol{z}_0^{h,0} = 0$ y $\boldsymbol{Z}_0^h = \boldsymbol{I}$, tambien tenemos (3.36). Si el método de paso simple o de pasos múltiples es del orden p, tenemos que

$$\boldsymbol{z}^*(x_j) - \boldsymbol{z}_k^{h,*} = \mathcal{O}(h^p), \quad j = 0, \ldots, N;$$

entonces el esquema es del mismo orden.

Ejemplo 3.2. *Seguimos considerando el problema introducido en el Ejemplo 3.1. Usando el método de Euler explícito ($p = 1$), obtenemos los siguientes valores numéricos para $h = 0{,}125$ ($N = 8$).*

j	1	2	3	4	5	6	7	8
$z_{1,j}^{h,0}$	0	−0,03125	−0,06226	−0,15528	−0,27832	−0,43113	−0,61344	−0,82494
$z_{2,j}^{h,0}$	−0,25	−0,49805	−0,74414	−0,98434	−1,22250	−1,45846	−1,69204	−1,92302
$z_{1,j}^{h,1} = z_{2,j}^{h,2}$	1	1,01563	1,04688	1,09400	1,15748	1,23805	1,33671	1,45471
$z_{2,j}^{h,1} = z_{1,j}^{h,2}$	0,125	0,25	0,37695	0,50781	0,64456	0,78925	0,94401	1,11110
$z_{1,j}^{h,*}$	0,69331	0,92313	0,78253	0,78259	0,76584	0,73397	0,68892	0,63288
$z_{2,j}^{h,*}$	0,31256	0,15118	0,00054	−0,13406	−0,25499	−0,36042	−0,44836	−0,51673

CUADRO 3.1. Ejemplo 3.1: Solución de un problema de valores de frontera mediante el método de disparo: valores numéricos.

1. *Usando*

$$\boldsymbol{z}_{j+1}^{h,0} = \boldsymbol{z}_j^{h,0} + 0{,}125 \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \boldsymbol{z}_j^{h,0} + 0{,}125 \begin{pmatrix} 0 \\ (0{,}125j)^2 - 2 \end{pmatrix}, \quad j = 0, \ldots, 7,$$
$$\boldsymbol{z}_0^{h,0} = 0$$

obtenemos los valores de las primeras dos filas del Cuadro 3.1.
2. *Luego calculamos*

$$\boldsymbol{z}_{j+1}^{h,i} = \boldsymbol{z}_j^{h,i} + 0{,}125 \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \boldsymbol{z}_j^{h,i}, \quad i = 1, 2;$$
$$\boldsymbol{z}_0^{h,1} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad \boldsymbol{z}_0^{h,2} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}.$$

Si $\boldsymbol{z}_k^{h,1} = (z_{1,k}^{h,1}, z_{2,k}^{h,1})^{\mathrm{T}}$*, esto implica que* $\boldsymbol{z}_k^{h,2} = (z_{2,k}^{h,1}, z_{1,k}^{h,1})^{\mathrm{T}}$*. La tercera y la cuarta fila del Cuadro 3.1 muestran los valores numéricos.*
3. *Hay que resolver el sistema de ecuaciones lineales*

$$\big(\boldsymbol{B}_1 + \boldsymbol{B}_2 \boldsymbol{Z}_8^h\big) \boldsymbol{s}^{h,*} = \boldsymbol{r} - \boldsymbol{B}_2 \boldsymbol{z}_8^{h,0},$$

es decir

$$\left(\boldsymbol{I} - \begin{bmatrix} 1{,}45471 & 1{,}11110 \\ 1{,}11110 & 1{,}45471 \end{bmatrix} \right) \begin{pmatrix} s_1^{h,*} \\ s_2^{h,*} \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix} + \begin{pmatrix} -0{,}82494 \\ -1{,}92302 \end{pmatrix}.$$

El resultado es $s_1^{h,*} = 0{,}63288$*,* $s_2^{h,*} = 0{,}48345$*. Las soluciones* (3.32) *son*

$$\boldsymbol{z}_j^{h,*} = \boldsymbol{z}_j^{h,0} + s_1^{h,*} \boldsymbol{z}_j^{h,1} + s_2^{h,*} \boldsymbol{z}_j^{h,2}, \quad \boldsymbol{z}_0^{h,1} = \boldsymbol{s}^{h,*} = \begin{pmatrix} 0{,}63288 \\ 0{,}48345 \end{pmatrix}.$$

Las dos últimas filas del Cuadro 3.1 muestran los valores numéricos. Vemos que la condición de borde en $x = 1$ *es satisfecha aproximadamente ya que*

$$\boldsymbol{z}_0^{h,*} - \boldsymbol{z}_8^{h,*} = \begin{pmatrix} 0 \\ 1{,}00018 \end{pmatrix}.$$

3.3.3. Métodos de disparo para problemas de valores de frontera no lineales. El método de disparo ya discutido de la reducción de un problema de valores de frontera a la solución de $n+1$ problemas de valores iniciales es simple por dos motivos. Por un lado, para un problema lineal la existencia de la solución del problema de valores iniciales

$$\boldsymbol{y}' - \boldsymbol{A}(x)\boldsymbol{y} = \boldsymbol{g}(x), \quad \boldsymbol{y}(a) = \boldsymbol{s}$$

para todo $\boldsymbol{s} \in \mathbb{R}^n$ y $x \in [a,b]$ está asegurada siempre que $\boldsymbol{A} \in C[a,b]$. Por otro lado, existe un sistema fundamental $\boldsymbol{Z}(x)$ que describe la solución general del problema de valores iniciales homogéneo. Ambas propiedades no son válidas para un problema de valores de frontera no lineal. Vamos a describir simplemente el método de disparo que puede ser aplicado en el caso no lineal. El problema está dado por

$$\begin{aligned} &\boldsymbol{y}' = \boldsymbol{F}(x,\boldsymbol{y}), \quad \boldsymbol{R}\big(\boldsymbol{y}(a),\boldsymbol{y}(b)\big) = 0, \\ &\boldsymbol{F} : [a,b] \times \mathbb{R}^n \to \mathbb{R}^n, \quad \boldsymbol{R} : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n. \end{aligned}$$

El problema de valores iniciales asociado es

$$\boldsymbol{y}' = \boldsymbol{F}(x,\boldsymbol{y}), \quad \boldsymbol{y}(a) = \boldsymbol{s}. \tag{3.37}$$

Se supone que la solución de (3.37) existe en el intervalo $[a,b]$ completo; si $\boldsymbol{F}$ depende de forma diferenciable de $\boldsymbol{y}$, esa solución existe en un intervalo suficientemente pequeño $[a, a+\delta]$. En lo siguiente, se supone que la función $\boldsymbol{F}$ es diferenciable. Entonces, la solución de (3.37) también es una función de $\boldsymbol{s}$, y depende de forma diferenciable de $\boldsymbol{s}$:

$$\boldsymbol{y} = \boldsymbol{y}(x;\boldsymbol{s}).$$

La ecuación se escribe entonces como

$$\boldsymbol{y}'(x;\boldsymbol{s}) = \boldsymbol{F}\big(x,\boldsymbol{y}(x;\boldsymbol{s})\big); \quad \boldsymbol{y}(a;\boldsymbol{s}) = \boldsymbol{s}.$$

Ahora tenemos que resolver el sistema no lineal

$$\boldsymbol{R}\big(\boldsymbol{y}(a;\boldsymbol{s}),\boldsymbol{y}(b;\boldsymbol{s})\big) = \boldsymbol{R}\big(\boldsymbol{s},\boldsymbol{y}(b;\boldsymbol{s})\big) = 0$$

con respecto a $\boldsymbol{s}$, donde $\boldsymbol{y}(b;\boldsymbol{s})$ es la solución de (3.37) evaluada en $x = b$. En general, este sistema no lineal puede ser resuelto solamente por el método de Newton-Raphson o alguna de sus variantes. Para discutir eso, definimos

$$\boldsymbol{G}(s) := \boldsymbol{R}\big(\boldsymbol{s};\boldsymbol{y}(b;\boldsymbol{s})\big).$$

Ahora tenemos que determinar la matriz $\nabla_{\boldsymbol{s}}\boldsymbol{G}(\boldsymbol{s})$ para poder ejecutar el método de Newton-Raphson. Teóricamente podemos cacular esa matriz a través de un problema de valores iniciales. Sin embargo, es más simple (y el método preferido en las aplicaciones) aproximar la matriz por diferencias finitas, por ejemplo

$$\nabla_{\boldsymbol{s}}\boldsymbol{G}(\boldsymbol{s}) \approx \frac{1}{\tau}\left[\boldsymbol{G}(\boldsymbol{s}+\tau\boldsymbol{e}_1) - \boldsymbol{G}(\boldsymbol{s}) \quad \cdots \quad \boldsymbol{G}(\boldsymbol{s}+\tau\boldsymbol{e}_n) - \boldsymbol{G}(\boldsymbol{s})\right].$$

Para calcular $\boldsymbol{G}(\boldsymbol{s})$ y la aproximación para $\nabla_{\boldsymbol{s}}\boldsymbol{G}(\boldsymbol{s})$ se deben resolver $n+1$ problemas de valores iniciales (3.37) con los valores iniciales $\boldsymbol{s}, \boldsymbol{s}+\tau\boldsymbol{e}_1, \ldots, \boldsymbol{s}+\tau\boldsymbol{e}_n$.

Un paso con el método de Newton-Raphson (amortiguado) entrega un valor mejorado de $\boldsymbol{s}$ con el cual se repite el procdimiento, etc. En la mayoría de las aplicaciones, no se conoce un valor inicial $\boldsymbol{s}^{(0)}$ muy bueno para buscar $\boldsymbol{s}^*$ con $\boldsymbol{G}(\boldsymbol{s}^*) = 0$. Además, aparece el problema

de que para algún valor de k, $\boldsymbol{y}(x;\boldsymbol{s}^{(k)})$ puede no existir en el intervalo completo $[a,b]$, pero sólo en un sub-intervalo.

Ejemplo 3.3. *Consideremos el problema*

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

con la solución exacta $y(x)=\ln(s^*x+1)+1$, $s^*=\mathrm{e}^{-5}-1$, *y*

$$y'(x)=\frac{s^*}{s^*x+1},$$

entonces, $y'(0)=s^*$. *El problema de valores iniciales*

$$y''+(y')^2=0,\quad y(0)=1,\quad y'(0)=s$$

tiene la solución $y(x;s)=1+\ln(sx+1)$, *entonces*

$$\begin{aligned} y(1;-0{,}99)&=-3{,}6052,\\ y(1;s^*=-0{,}993262053\dots)&=-4,\\ y(1;-0{,}999)&=-5{,}9078, \end{aligned}$$

mientras que para $s\leqslant -1$, *la solución del problema de valores iniciales existe solamente en en intervalo* $[0,-1/s]$.

3.3.4. Métodos de disparos múltiples. Como para cada $x_i\in[a,b]$, la solución del problema de valores iniciales

$$\boldsymbol{y}'=\boldsymbol{F}(x,\boldsymbol{y}),\quad \boldsymbol{y}(x_i)=\boldsymbol{s}^i$$

existe en un intervalo $x_i-\delta_i<x<x_i+\delta_i$, podemos tratar de evitar el problema de no existencia de una solución global mediante la subdivisión en problemas parciales. Para una partición

$$a=x_0<x_1<x_2<\cdots<x_m=b$$

y valores iniciales apropiados $\boldsymbol{s}^i\in\mathbb{R}^n$ en las posiciones x_i se definen soluciones parciales $\boldsymbol{y}_{[i]}(x;\boldsymbol{s}^i)$ a través de

$$\begin{aligned} \boldsymbol{y}'_{[i]}(x;\boldsymbol{s}^i)&=\boldsymbol{F}\big(x,\boldsymbol{y}_{[i]}(x;\boldsymbol{s}^i)\big),\quad x_i\leqslant x\leqslant x_{i+1},\\ \boldsymbol{y}_{[i]}(x_i,\boldsymbol{s}^i)&=\boldsymbol{s}^i,\quad i=0,\dots,m-1, \end{aligned}\tag{3.38}$$

donde $\boldsymbol{y}_{[0]}(a;\boldsymbol{s}^0)=\boldsymbol{y}(a)$. Las soluciones parciales de (3.38) forman una función continua si

$$\boldsymbol{y}_{[i]}(x_{i+1};\boldsymbol{s}^i)=\boldsymbol{s}^{i+1},\quad i=0,\dots,m-1,\tag{3.39}$$

donde $\boldsymbol{s}^m=\boldsymbol{y}(b)$, y la solución así compuesta es una solución del problema de valores de frontera si

$$\boldsymbol{R}(\boldsymbol{s}^0,\boldsymbol{s}^m)=0.\tag{3.40}$$

En total, tenemos $m+1$ ecuaciones de n componentes para las $m+1$ desconocidas de n componentes $\boldsymbol{s}^0,\dots,\boldsymbol{s}^m$, cuya solución simultánea entrega la solución del problema de valores de frontera. El sistema no lineal definido por (3.39) y (3.40) se resuelve por el método de Newton-Raphson amortiguado. Este procedimiento se llama *método de disparos múltiples*.

Capítulo 4

Problemas de valores de frontera para ecuaciones diferenciales parciales elípticas

4.1. Clasificación

Una ecuación diferencial parcial de segundo orden en n variables independientes $x_1, \ldots, x_n$ para una función buscada $u = u(x_1, \ldots, x_n)$ es la ecuación

$$F\big(x_1, \ldots, x_n, u, u_{x_1}, \ldots, u_{x_n}, u_{x_1x_1}, u_{x_1x_2}, \ldots, u_{x_nx_n}\big) = 0, \quad n \geqslant 2.$$

Las ecuaciones de segundo orden que aparecen en las aplicaciones casi siempre son cuasi-lineales, semi-lineales o lineales y pueden ser representadas en la forma

$$Lu \equiv \sum_{i,k=1}^{n} A_{ik} u_{x_i x_k} = f. \tag{4.1}$$

Definición 4.1. *Una ecuación diferencial parcial de la forma* (4.1) *se llama*

- cuasi-lineal*, si por lo menos uno de los coeficientes A_{ik} es una función de por lo menos una de las variables $u, u_{x_1}, \ldots, u_{x_n}$,*
- semi-lineal*, si las funciones A_{ik} son a lo más funciones de $x_1, \ldots, x_n$, pero f depende de forma no lineal de por lo menos una de las variables $u, u_{x_1}, \ldots, u_{x_n}$,*
- lineal*, si las funciones A_{ik} son a lo más funciones de $x_1, \ldots, x_n$, y*

$$f = \sum_{i=1}^{n} A_i u_{x_i} + Au + B,$$

donde $A_1, \ldots, A_n$, A y B pueden ser funciones de las variables independientes $x_1, \ldots, x_n$.

Para la clasificación según tipo, definimos la forma cuadrática

$$Q = \sum_{i,k=1}^{n} A_{ik} p_i p_k \tag{4.2}$$

con las variables $p_1, \ldots, p_n$. Definiendo la matriz $\boldsymbol{A} := (A_{ik})$ y el vector $\boldsymbol{p} := (p_1, \ldots, p_n)^{\mathrm{T}}$, podemos escribir (4.2) como

$$Q = \boldsymbol{p}^{\mathrm{T}} \boldsymbol{A} \boldsymbol{p}.$$

Se supone que $\boldsymbol{A}$ es simétrica (si no lo es, remplazamos $\boldsymbol{A}$ por $\bar{\boldsymbol{A}} = \frac{1}{2}(\boldsymbol{A} + \boldsymbol{A}^{\mathrm{T}})$). Supongamos que en un dominio $B \subset \mathbb{R}^n$ existe una solución $u(\boldsymbol{x}) = u(x_1, \ldots, x_n)$. En el caso cuasi-lineal, se supone que la solución está incertada en los A_{ik}, entonces en cada caso $\boldsymbol{A}$ depende sólo de $\boldsymbol{x} = (x_1, \ldots, x_n)^{\mathrm{T}}$. Debido a la simetría de $\boldsymbol{A}$, existe una matriz ortonormal $\boldsymbol{T}$ tal que

$$\boldsymbol{T}^{\mathrm{T}} \boldsymbol{A} \boldsymbol{T} = \boldsymbol{B},$$

donde $\boldsymbol{B} = \operatorname{diag}(B_1, \ldots, B_n)$, y $B_1, \ldots, B_n$ son los valores propios de $\boldsymbol{A}$. Definiendo $\boldsymbol{q} = \boldsymbol{T}^{\mathrm{T}}\boldsymbol{p}$, tenemos

$$Q = \boldsymbol{p}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{p} = \boldsymbol{q}^{\mathrm{T}}\boldsymbol{B}\boldsymbol{q} = \sum_{i=1}^{n} B_i q_i^2.$$

Se llama *índice inercial* τ al número de los B_i negativos y *defecto* δ al número de los $B_i = 0$ de la forma cuadrática Q.

Definición 4.2. *En $\boldsymbol{x} \in B \subset \mathbb{R}^n$, la ecuación diferencial parcial* (4.1) *se llama*

- hiperbólica*, si allí $\delta = 0$ y $\tau = 1$ o $\tau = n - 1$,*
- parabólica*, si allí $\delta > 0$,*
- elíptica*, si allí $\delta = 0$ y $\tau = 0$ o $\tau = n$, y*
- ultrahiperbólica*, si allí $\delta = 0$ y $1 < \tau < n - 1$ (obviamente, esto puede ocurrir sólo si $n \geqslant 4$).*

La clasificación es del tipo geométrico para ecuaciones diferenciales parciales de segundo orden, dado que

$$\sum_{i=1}^{n} B_i x_i^2 = c, \quad c > 0$$

es un hiperboloide, un paraboloide o un elipsoide en los respectivos casos (a), (b) y (c) de la Definición 4.2. Por otro lado, la clasificación puede ser realizada sólo en un punto $\boldsymbol{x}$, y entonces es de naturaleza local. Si todos los coeficientes A_{ik} son constantes, la clasificación es global. De hecho, el tipo de una ecuación cuasi-lineal no sólo depende de $\boldsymbol{x} \in B \subset \mathbb{R}^n$, sino que también del valor de la solución. Por ejemplo, la ecuación

$$u_{x_1x_1} + uu_{x_2x_2} = 0$$

es hiperbólica, parabólica o elíptica en un punto $\boldsymbol{x} = (x_1, x_2)$ dependiendo de si $u(\boldsymbol{x}) < 0$, $u(\boldsymbol{x}) = 0$ o $u(\boldsymbol{x}) > 0$ en este punto.

En lo siguiente, nos restringimos al caso $n = 2$, y ponemos $x = x_1$ e $y = x_2$. Consideramos la ecuación

$$Lu \equiv au_{xx} + 2bu_{xy} + cu_{yy} = f, \tag{4.3}$$

es decir, consideramos la matriz

$$\boldsymbol{A} = \begin{bmatrix} a & b \\ b & c \end{bmatrix},$$

cuyos valores propios son

$$\lambda_{1,2} = \frac{a+c}{2} \pm \frac{1}{2}\sqrt{(a+c)^2 - 4(ac - b^2)}.$$

Obviamente,

$$\begin{aligned} &\operatorname{sgn}\lambda_1 = -\operatorname{sgn}\lambda_2 && \text{si } ac - b^2 < 0, \\ &\lambda_1 = a + c, \quad \lambda_2 = 0 && \text{si } ac - b^2 = 0, \\ &\operatorname{sgn}\lambda_1 = \operatorname{sgn}\lambda_2 && \text{si } ac - b^2 > 0. \end{aligned}$$

Lema 4.1. *En un punto* $(x, y) \in \mathbb{R}^2$ *fijo, la ecuación diferencial parcial* (4.3) *es del tipo* $\left\{ \begin{array}{c} \textit{hiperbólico} \\ \textit{parabólico} \\ \textit{elíptico} \end{array} \right\}$ *si en este punto,* $\left\{ \begin{array}{l} ac - b^2 < 0 \\ ac - b^2 = 0 \\ ac - b^2 > 0 \end{array} \right\}$.

El tipo de las condiciones de borde adecuadas para (4.3) depende intrinsicamente del tipo de la ecuación.

4.2. Problemas de valores de frontera para ecuaciones elípticas

Para ecuaciones hiperbólicas, los problemas de valores iniciales, y para ecuaciones parabólicas, los problemas de valores iniciales y de frontera son bien puestos en general. (También hay situaciones donde los problemas "vice versa" son bien puestos.) Para las ecuaciones elípticas, los problemas de valores de frontera en general son bien puestos.

Se considera un dominio $G \subset \mathbb{R}^2$ abierto y acotado, cuya frontera ∂G es una curva diferenciable, es decir, existe el vector normal $\boldsymbol{\nu}$. Si α, β y γ son funciones continuas dadas sobre $\bar{G}$, podemos identificar los siguientes problemas de valores de frontera, que son los más importantes: el primer problema de valores de frontera

$$\begin{aligned} Lu &\equiv f \quad \text{para } (x, y) \in G, \\ u(x, y) &= \gamma(x, y) \quad \text{para } (x, y) \in \partial G, \end{aligned}$$

el segundo problema de valores de frontera

$$\begin{aligned} Lu &\equiv f \quad \text{para } (x, y) \in G, \\ \frac{\partial u}{\partial \boldsymbol{\nu}}(x, y) &= \gamma(x, y) \quad \text{para } (x, y) \in \partial G, \end{aligned}$$

y el tercer problema de valores de frontera

$$\begin{aligned} Lu &\equiv f \quad \text{para } (x, y) \in G, \\ \alpha(x, y)u(x, y) - \beta(x, y)\frac{\partial u}{\partial \boldsymbol{\nu}}(x, y) &= \gamma(x, y) \quad \text{para } (x, y) \in \partial G, \end{aligned}$$

donde definimos

$$\frac{\partial u}{\partial \boldsymbol{\nu}} = \nu_1 \frac{\partial u}{\partial x} + \nu_2 \frac{\partial u}{\partial y} \quad \text{si } \boldsymbol{\nu} = \begin{pmatrix} \nu_1 \\ \nu_2 \end{pmatrix}.$$

En lo siguiente, siempre se supone que existe una solución del problema de valores de frontera considerado.

4.3. Problemas de valores de frontera y problemas variacionales

Consideremos la ecuación

$$Lu \equiv -a_{11}u_{xx} - 2a_{12}u_{xy} - a_{22}u_{yy} - a_1u_x - a_2u_y + au = f,$$

donde a_{ik}, a_i, a y f $(i, k = 1, 2)$ son funciones de (x, y). Si $a_{ik} \in C^2(G)$ y $a_i \in C^1(G)$, $i, k = 1, 2$, el operador

$$L^*u \equiv -(a_{11}u)_{xx} - 2(a_{12}u)_{xy} - (a_{22}u)_{yy} + (a_1u)_x + (a_2u)_y + au$$

se llama *operador adjunto* de L. Si $L^*u = Lu$ para toda función $u \in C^2(G)$, el operador L s llama *autoadjunto sobre* G; en este caso, la ecuación diferencial $Lu = f$ se llama *ecuación diferencial autoadjunta*. Facilmente podemos concluir que un operador autoadjunto siempre es de la forma

$$Lu = -(a_{11}u_x)_x - (a_{12}u_x)_y - (a_{12}u_y)_x - (a_{22}u_y)_y + au. \tag{4.4}$$

En particular, para $a_{11} \equiv a_{22} \equiv 1$ y $a_{12} \equiv a \equiv 0$ obtenemos el *operador de Laplace*

$$Lu = -\Delta_2 u \equiv -u_{xx} - u_{yy}.$$

Como para las ecuaciones diferenciales ordinarias, existe una conexión entre los problemas de valores de frontera para ecuaciones autoadjuntas y problemas variacionales. Para explicar eso, consideremos el problema

$$Lu = f \quad \text{en } G, \quad u = 0 \quad \text{en } \partial G, \tag{4.5}$$

donde L es el operador autoadjunto (4.4). El dominio de L es el conjunto de todas las funciones definidas sobre $\bar{G} = G \cup \partial G$ y dos veces continuamente diferenciables sobre G que desaparecen sobre ∂G, es decir, este dominio es

$$\mathcal{D} := \big\{ v \in C^0(\bar{G}) \cap C^2(G) \mid v = 0 \text{ en } \partial G \big\}.$$

Así, el problema (4.5) puede ser escrito de la siguiente forma: se busca una solución de

$$Lv = f, \quad v \in \mathcal{D}. \tag{4.6}$$

Luego, sea $L^2(G)$ el espacio de las funciones cuadraticamente integrables sobre G, para el cual definimos el producto escalar

$$(v, w) := \int_G v(x, y)w(x, y)\,\mathrm{d}x\,\mathrm{d}y, \quad v, w \in L^2(G),$$

y la norma asociada

$$\|v\|_2 := (v, v)^{1/2}.$$

Respecto a la ecuación $Lu = f$, se supone que

$$\begin{aligned} &a_{ik} \in C^2(\bar{G}), \quad i, k = 1, 2, \quad a, f \in C^0(\bar{G}), \quad a(x, y) \geqslant 0, \quad (x, y) \in \bar{G}, \\ &\sum_{i,k=1}^{2} a_{ik}(x, y)\xi_i\xi_k \geqslant \alpha \sum_{i=1}^{2} \xi_i^2, \quad (x, y) \in \bar{G}, \quad \xi_1, \xi_2 \in \mathbb{R}, \end{aligned} \tag{4.7}$$

donde $\alpha > 0$ es independiente de ξ_1 y ξ_2. Se sabe que bajo las hipótesis (4.7),

$$\forall v, w \in \mathcal{D}: \quad (v, Lw) = (Lv, w), \quad (Lv, v) > 0 \quad (v \neq 0).$$

Teorema 4.1. *La función $u \in \mathcal{D}$ es una solución del problema de valores de frontera* (4.6) *con el operador elíptico autoadjunto L si y sólo si bajo las hipótesis* (4.7) *y definiendo*

$$I[v| := (v, Lv) - 2(v, f),$$

tenemos

$$I[u] = \min_{v \in \mathcal{D}} I[v].$$

Según este teorema, podemos resolver el problema (4.5), (4.6) resolviendo el problema variacional

$$(v, Lv) - 2(v, f) = \int_G \big(a_{11}v_x^2 + 2a_{12}v_x v_y + a_{22}v_y^2 + av^2 - 2vf\big)\,\mathrm{d}x\,\mathrm{d}y \xrightarrow{!} \text{mín}, \quad v \in \mathcal{D}. \tag{4.8}$$

Obviamente, no podemos resolver el problema variacional de forma exacta, sino que solamente de forma aproximada. Observamos que la integral en (4.8) no es definida solamente para funciones $v, w \in \mathcal{D}$, sino que para una mayor clase de funciones.

Definición 4.3. *Una función* v pertenece al espacio $V(G)$ *si* $v \in C^0(\bar{G})$, v *es diferenciable por trozos con respecto a* x *e* y *sobre* $\bar{G}$, *y* $v_x, v_y \in L^2(G)$, *es decir,*

$$\|v\|_{V(G)} := \left(\int_G \Big(\big(v(x,y)\big)^2 + \big(v_x(x,y)\big)^2 + \big(v_y(x,y)\big)^2\Big)\mathrm{d}x\,\mathrm{d}y\right)^{1/2} < \infty.$$

Se confirma facilmente que $\|\cdot\|_{V(G)}$ efectivamente es una norma. Ahora definimos

$$D := \big\{v \in V(G) \mid v = 0 \text{ en } \partial G\big\},$$

y para $v, w \in D$ la forma bilineal simétrica

$$[v, w] := \int_G \big(a_{11}v_x w_x + a_{12}(v_x w_y + v_y w_x) + a_{22}v_y w_y + avw\big)\,\mathrm{d}x\,\mathrm{d}y.$$

Obviamente, $\mathcal{D} \subset D$ y $[v, w] = (Lv, w)$ para $v, w \in \mathcal{D}$.

Teorema 4.2. *Sea* $u \in \mathcal{D}$ *la solución del problema de valores de frontera* (4.6). *Entonces*

$$\forall v \in D: \quad I[u] \leqslant I[v],$$

donde

$$I[v] := [v, v] - 2(v, f) \quad \text{para } v \in D.$$

4.4. Métodos de diferencias

Consideremos ahora el problema modelo

$$\begin{aligned} -\Delta u = -u_{xx} - u_{yy} = f(x,y), \quad (x,y) \in G := (0,1)^2, \\ u(x,y) = 0, \quad (x,y) \in \partial G, \end{aligned} \tag{4.9}$$

donde se supone que $f \in C^0(\bar{G})$. Sobre el cuadrado $\bar{G} = G \cup \partial G$ se define una malla con $\Delta x = \Delta y = h$, donde G_h denota la totalidad de los puntos interiores y ∂G_h la de los puntos de frontera. Se supone que $u = u(x, y)$ es una solución de la ecuación diferencial en (4.9), la cual no necesariamente debe desaparacer en ∂G, y sea $u \in C^4(\bar{G})$. En este caso,

$$\begin{aligned} u_{xx}(x_i, y_k) &= \frac{u(x_{i+1}, y_k) - 2u(x_i, y_k) + u(x_{i-1}, y_k)}{h^2} + \varepsilon_{ik}(h), \\ u_{yy}(x_i, y_k) &= \frac{u(x_i, y_{k+1}) - 2u(x_i, y_k) + u(x_i, y_{k-1})}{h^2} + \eta_{ik}(h), \end{aligned} \tag{4.10}$$

donde

$$\begin{aligned}\varepsilon_{ik}(h) &= \frac{h^2}{12}\frac{\partial^4 u}{\partial x^4}(x_i+\vartheta_1 h, y_k), \quad -1 \leqslant \vartheta_1 \leqslant 1;\\ \eta_{ik}(h) &= \frac{h^2}{12}\frac{\partial^4 u}{\partial y^4}(x_i, y_k+\vartheta_2 h), \quad -1 \leqslant \vartheta_2 \leqslant 1.\end{aligned} \tag{4.11}$$

Insertando (4.10) y (4.11) en (4.9), obtenemos

$$\begin{aligned}&(-\Delta_2 u)(x_i,y_k) - f(x_i,y_k)\\ &= \frac{1}{h^2}\big(4u(x_i,y_k) - u(x_{i-1},y_k) - u(x_{i+1},y_k) - u(x_i,y_{k-1}) - u(x_i,y_{k+1})\big)\\ &\quad - f(x_i,y_k) - \varepsilon_{ik}(h) - \eta_{ik}(h)\\ &= 0.\end{aligned} \tag{4.12}$$

Despreciando el término $\varepsilon_{ik}(h)+\eta_{ik}(h) = \mathcal{O}(h^2)$, obtenemos el sistema lineal

$$\begin{aligned}&-(L_h \boldsymbol{u}^h)_{ik} = \frac{1}{h^2}\big(4u^h_{ik} - u^h_{i-1,k} - u^h_{i+1,k} - u^h_{i,k-1} - u^h_{i,k+1}\big) = f(x_i,y_k),\\ &i,k = 1,\ldots,N_h-1.\end{aligned} \tag{4.13}$$

Aquí u^h_{ik} son valores de la función de malla $\boldsymbol{u}^h$, la cual podemos representar como un vector con las componentes u^h_{ik}, $i,k = 1,\ldots,N_h-1$. Se presenta el problema de la enumeración apropiada de los u^h_{ik}. Por motivos que se excplicarán más abajo, definimos

$$\boldsymbol{u}^h := \big(u^h_{11}, u^h_{21}, \ldots, u^h_{l-1,1}, u^h_{l-2,2}, \ldots, u^h_{1,l-1}, \ldots, u^h_{N_h-1,N_h-1}\big)^{\mathrm{T}},$$

es decir, después de u^h_{11} siguen sucesivamente los u^h_{ik} con $i+k = 3,4,\ldots,2N_h-2$, donde dentro del bloque con $i+k=l$ el ordenamiento es

$$u^h_{l-1,1}, u^h_{l-2,2}, \ldots, u^h_{1,l-1}, \quad l = 3,4,\ldots,2N_h-2.$$

Después de multiplicar con h^2, el sistema (4.13) asume la forma

$$\boldsymbol{A}(h)\boldsymbol{u}^h = \boldsymbol{b}(h). \tag{4.14}$$

Teorema 4.3. *La matriz $\boldsymbol{A}(h)$ es una M-matriz irreduciblemente diagonaldominante y simétrica, y $\boldsymbol{A}(h)$ es una matriz tridiagonal por bloques de la forma*

$$\boldsymbol{A}(h) = \begin{bmatrix} \boldsymbol{D}_1 & \boldsymbol{H}_1 & & & \\ \boldsymbol{H}_1 & \boldsymbol{D}_2 & \boldsymbol{H}_2 & & \\ & \ddots & \ddots & \ddots & \\ & & \boldsymbol{H}_{s-2} & \boldsymbol{D}_{s-1} & \boldsymbol{H}_{s-1} \\ & & & \boldsymbol{H}_{s-1} & \boldsymbol{D}_s \end{bmatrix},$$

donde los $\boldsymbol{D}_i$ son matrices diagonales (de diferentes tamaños) de la forma

$$\boldsymbol{D}_i = \operatorname{diag}(4,\ldots,4), \quad i = 1,\ldots,s;$$

además,

$$\boldsymbol{b}(h) = h^2 \begin{pmatrix} f(h,h) \\ f(2h,h) \\ \vdots \\ f((N_h-1)h,(N_h-1)h) \end{pmatrix}.$$

Como $\boldsymbol{A}(h)$ es definida positiva, el sistema (4.14) puede ser resuelto usando el método SOR con $0 < \omega \leqslant 2$. Dado que además $\boldsymbol{A}(h)$ es ordenada consistentemente, existe un parámetro óptimo de relajación ω_{opt} que asegura la velocidad de convergencia óptima del método SOR. Con nuestra enumeración de las componentes de $\boldsymbol{u}^h$ hemos entonces asegurado que la matriz es ordenada consistentemente y admite la existencia de ω_{opt}. En la mayoría de casos, el sistema (4.14) es esparso, pero de gran tamaño. Por otro lado, se puede demostrar que el número de condición es

$$\text{cond}_{\|\cdot\|_2}(\boldsymbol{A}(h)) = \|\boldsymbol{A}(h)\|_2 \|\boldsymbol{A}(h)^{-1}\|_2 = \mathcal{O}(h^{-2}) = \mathcal{O}(N_h^2),$$

es decir, el sistema es mal acondicionado para $h \to 0$.

4.5. Convergencia del método de diferencias

Sea $u = u(x,y)$ la solución del problema (4.9), y

$$\boldsymbol{u}(h) = \big(u(h,h), u(2h,h), u(h,2h), \ldots, u((N_h-1)h,(N_h-1)h)\big)^{\mathrm{T}},$$

donde respetamos la enumeración ya establecida, y el vector de errores

$$\boldsymbol{\varepsilon}^h := \boldsymbol{u}^h - \boldsymbol{u}(h).$$

Usando (4.12) y

$$\varepsilon_{ik}(h) + \eta_{ik}(h) = \mathcal{O}(h^2), \quad i,k = 1,\ldots,N_h-1,$$

tenemos (análogamente a (4.14))

$$\boldsymbol{A}(h)\boldsymbol{u}(h) = \boldsymbol{b}(h) + h^2\mathcal{O}(h^2), \tag{4.15}$$

lo que representa el sistema (4.12) multiplicado por h^2. Restando (4.15) de (4.14) obtenemos

$$\boldsymbol{A}(h)\boldsymbol{\varepsilon}^h = h^2\mathcal{O}(h^2),$$

o sea

$$\boldsymbol{\varepsilon}^h = h^2\boldsymbol{A}(h)^{-1}\mathcal{O}(h^2). \tag{4.16}$$

Aquí, $\mathcal{O}(h^2)$ es un vector de $(N_h-1)^2$ componentes, cada una acotada por Ch^2. Supongamos que

$$\big\|\boldsymbol{A}(h)^{-1}\big\|_\infty \leqslant Kh^{-2} \tag{4.17}$$

con una constante K independiente de h. En este caso, (4.16) implicaría

$$\|\boldsymbol{\varepsilon}^h\|_\infty \leqslant Ch^2K = Mh^2$$

con una constante M independiente de h. Para $h \to 0$, tendríamos

$$\big|u_{ik}^h - u(x_i,y_k)\big| = \big|u_{ik}^h - u(ih,kh)\big| \leqslant Mh^2, \quad i,k = 1,\ldots,N_h-1,$$

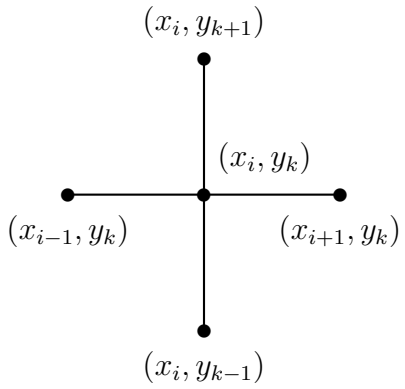
donde la constante M es independiente de h, y el método sería de segundo orden. La propiedad (4.17) efectivamente se cumple (literatura especializada).

Teorema 4.4. *Supongamos que el problema de valores de frontera* (4.9) *posee una solución* $u \in C^4(\bar{G})$. *Entonces el método de diferencias finitas dado por* (4.13) *es convergente del orden 2, es decir para* $h \to 0$ *tenemos*

$$u_{ik}^h - u(x_i, y_k) = \mathcal{O}(h^2), \quad i, k = 1, \ldots, N_h - 1.$$

4.6. Dominios con frontera curvada

Ahora consideramos el caso donde G es un dominio abierto, acotado y convexo, con una frontera continua ∂G. Sobre G podemos definir una malla rectangular con Δx y Δy como largo de los lados de cada rectángulo. Por simplicidad, usaremos una malla cuadrática G_h con $\Delta x = \Delta y = h$. Ahora los puntos del borde de la malla, ∂G_h, no van en general pertenecer a ∂G. Entonces tenemos que construir el conjunto de los puntos ∂G_h, que forman el *borde numérico*. Para tal efecto, denominamos como *estrella* o *molécula* el siguiente arreglo de puntos:



El conjunto de todos los puntos de la malla que son puntos centrales de estrellas que enteramente pertenecen a $\bar{G}$ son la *malla* G_h. El conjunto ∂G_h de los puntos de borde está formado por todos los puntos de la malla que pertenecen a estrellas completamente contenidas en $\bar{G}$, pero que no son puntos centrales de tales estrellas.

Ahora podemos resolver numéricamente el problema (4.9) con el mismo método numérico usado anteriormente, donde exigimos que

$$u_{rs}^h = 0 \quad \text{para } (x_r, y_s) \in \partial G_h.$$

Obviamente, estos valores de frontera causan un error si $(x_r, y_s) \notin \partial G$. Podemos ver facilmente que estos errores son proporcionales a h, es decir, los valores exactos de borde en los puntos de ∂G_h son del orden de magnitud $\mathcal{O}(h)$.

Podemos construir valores de frontera más exactos por interpolación lineal, ver Figura 4.1. Consideremos los puntos co-lineales $(x_{i-1}, y_k) \in G_h$, $(x_i, y_k) \in \partial G_h$ y $(x, y_k) \in \partial G$ para $x > x_i$. Con $x - x_i =: \delta < h$ y usando $u(x, y_k) = 0$ obtenemos, usando interpolación lineal,

$$u(x_i, y_k) - \frac{\delta}{h+\delta} u(x_{i-1}, y_k) = \frac{h}{h+\delta} u(x, y_k) + \mathcal{O}(h^2) = \mathcal{O}(h^2). \tag{4.18}$$

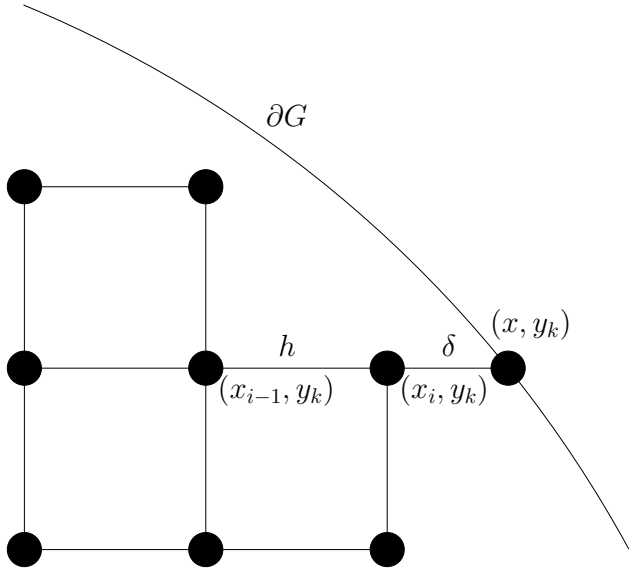


FIGURA 4.1. Derivación de valores de borde por interpolación.

Despreciando el término $\mathcal{O}(h^2)$, obtenemos la ecuación aproximada

$$u^h_{ik} - \frac{\delta}{h+\delta} u^h_{i-1,k} = 0, \quad (x_i, y_k) \in \partial G_h.$$

Ahora, el borde ∂G no necesariamente debe ser localizado como en la Figura 4.1. Pero vemos facilmente que la interpolación siempre entrega una de las ecuaciones aproximadas

$$u^h_{ik} - \frac{\delta_{ik}}{h+\delta_{ik}} u^h_{i\pm1,k} = 0, \quad u^h_{ik} - \frac{\delta_{ik}}{h+\delta_{ik}} u^h_{i,k\pm1} = 0, \quad (x_i, y_k),\, (x_{i\pm1}, y_k),\, (x_i, y_{k\pm1}) \in G_h, \tag{4.19}$$

donde $\delta_{ik} \in (0, h)$ es la distancia del punto $(x_i, y_k) \in \partial G_h$ de aquel punto de ∂G que está situado sobre la recta que pasa por (x_i, y_k) y $(x_{i\pm1}, y_k)$ o (x_i, y_k) y $(x_i, y_{k\pm1})$, respectivamente.

En (4.19), ambas cantidades u^h_{ik} y $u^h_{i\pm1,k}$, $u^h_{i,k\pm1}$ son desconocidas, es decir, para cada punto de ∂G_h obtenemos una ecuación adicional. Si G_h y ∂G_h contienen exactamente M_h y $\tilde{M}_h$ puntos respectivamente, tenemos que resolver ahora un sistema de $M_h + \tilde{M}_h$ ecuaciones. El esfuerzo adicional nos asegura valores numéricos de frontera de mayor precisión; debido a (4.18) su error es proporcional a h^2. Si agregamos (4.19), la matriz del sistema de ecuaciones lineales que resulta en general ya no es simétrica, pero se puede demostrar que todavía es una M-matriz.

Ejemplo 4.1. *Consideremos el dominio G dibujado en la Figura 4.2. Para cada punto $(x_i, y_k) \in G_h$ (puntos marcados por $\bullet$) usamos la ecuación* (4.13) *(después de la multiplicación con h^2), y para cada punto $(x_i, y_k) \in \partial G_h$ (puntos marcados por $\circ$), usamos una versión de* (4.19). *Entonces, la discretización entrega aquí la matriz $\boldsymbol{A}(h)$ y el vector correspondiente*

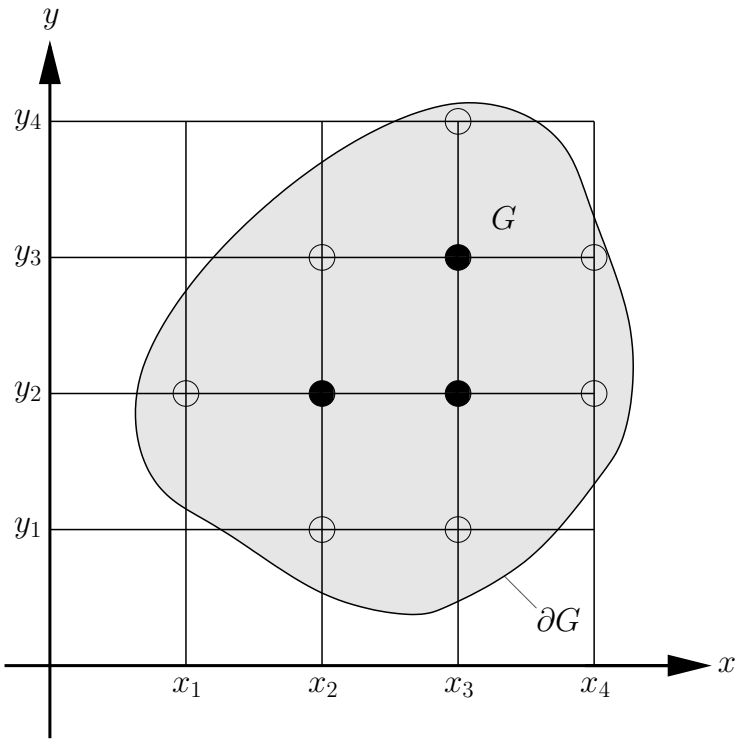


FIGURA 4.2. Ejemplo 4.1: Dominio G y puntos de malla en G_h ($\bullet$) y en ∂G_h ($\circ$).

$\boldsymbol{u}^h$ *dados por*

$$\boldsymbol{A}(h) = \begin{bmatrix} 1 & 0 & 0 & -\frac{\delta_{21}}{h+\delta_{21}} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & -\frac{\delta_{12}}{h+\delta_{12}} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & -\frac{\delta_{31}}{h+\delta_{31}} & 0 & 0 & 0 & 0 & 0 \\ -1 & -1 & 0 & 4 & -1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & -1 & 4 & 0 & -1 & -1 & 0 & 0 \\ 0 & 0 & 0 & -\frac{\delta_{23}}{h+\delta_{23}} & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -\frac{\delta_{42}}{h+\delta_{42}} & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & -1 & 0 & 4 & -1 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & -\frac{\delta_{43}}{h+\delta_{43}} & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & -\frac{\delta_{34}}{h+\delta_{34}} & 0 & 1 \end{bmatrix}, \quad \boldsymbol{u}^h = \begin{pmatrix} u_{21}^h \\ u_{12}^h \\ u_{31}^h \\ u_{22}^h \\ u_{32}^h \\ u_{23}^h \\ u_{42}^h \\ u_{33}^h \\ u_{43}^h \\ u_{34}^h \end{pmatrix}.$$

Dado que $0 \leqslant \delta_{ik} < h$, *la matriz* $\boldsymbol{A}(h)$ *es una L-matriz debilmente diagonaldominante con dominancia diagonal fuerte, por ejemplo, en la fila 1. La matriz es irreducible y por lo tanto, una M-matriz.*

Capítulo 5

Problemas de valores iniciales y de frontera para EDPs hiperbólicas y parabólicas

Este capítulo trata de la solución numérica de aquellos tipos de ecuaciones de derivadas parciales de segundo orden que normalmente describen procesos que dependen del tiempo: las ecuaciones hiperbólicas describen procesos de radiación (por ejemplo, la propagación de ondas), mientras que las ecuaciones parabólicas describen procesos de difusión (por ejemplo, la conducción del calor). Antes de discutir métodos numéricos para estos tipos de ecuaciones vamos a introducir el concepto importante de las *características*. Sin embargo, en lo siguiente nos limitaremos a sistemas de dos ecuaciones escalares de primer orden y a ecuaciones escalares de segundo orden, y a ecuaciones con solamente dos variables independientes.

5.1. Teoría de las características

5.1.1. Ecuaciones cuasi-lineales escalares de segundo orden. Sea $G \subset \mathbb{R}^2$ un dominio y $u(x, y) : \mathbb{R}^2 \supset G \to \mathbb{R}$ una solución de la siguiente ecuación cuasi-lineal definida en G:

$$au_{xx} + bu_{xy} + cu_{yy} = f, \tag{5.1}$$

donde $a, b, c, f : \ G \times \mathbb{R}^3 \to \mathbb{R}$ son funciones suficientemente suaves de x, y, u y ∇u. Además, se supone que $a \neq 0$ en $G \times \mathbb{R}^3$. Sea $\Gamma \subset G$ una curva suave, por ejemplo representada por una parametrización

$$\big(x(\sigma), y(\sigma)\big) : \ \mathbb{R} \supset [0, 1] \to \Gamma \subset G \subset \mathbb{R}^2.$$

Sin restricción esencial supondremos que siempre $\mathrm{d}x(\sigma)/\mathrm{d}\sigma \neq 0$, es decir, en ningún punto Γ posee una tangente vertical, asi que están bien definidos los conceptos de los puntos "arriba" y "debajo" de Γ.

Para un punto arbitrario $P \in \Gamma$ nos preguntamos si el comportamento de la solución de (5.1) en una vecindad de P *arriba* de Γ (o tambíen *debajo* de Γ) está unicamente definido por el conocimiento de la solución y de sus primeras derivadas en una vecindad de P *a lo largo de* Γ. En otras palabras, ¿se pueden determinar las segundas derivadas u_{xx}, u_{xy} y u_{yy} en $P \in \Gamma$ de forma única si conocemos u, u_x y u_y en $P \in \Gamma$?

Formando para las funciones u_x y u_y las derivadas tangenciales $\mathrm{d}(u_x)$ y $\mathrm{d}(u_y)$ en $P \in \Gamma$, es decir, las derivadas en la dirección de la curva Γ, obtenemos las siguientes ecuaciones diferenciales:

$$\begin{aligned} \mathrm{d}(u_x) &= u_{xx}\,\mathrm{d}x + u_{xy}\,\mathrm{d}y, \\ \mathrm{d}(u_y) &= u_{xy}\,\mathrm{d}x + u_{yy}\,\mathrm{d}y. \end{aligned} \tag{5.2}$$

Combinando (5.1) y (5.2) obtenemos el sistema de ecuaciones

$$\begin{bmatrix} a & b & c \\ \mathrm{d}x & \mathrm{d}y & 0 \\ 0 & \mathrm{d}x & \mathrm{d}y \end{bmatrix} \begin{pmatrix} u_{xx} \\ u_{xy} \\ u_{yy} \end{pmatrix} = \begin{pmatrix} f \\ \mathrm{d}(u_x) \\ \mathrm{d}(u_y) \end{pmatrix}. \tag{5.3}$$

Claramente, la pregunta formulada aquí debe ser contestada por "no" si y sólo si el sistema (5.3) no posee una solución única. Esto sucede si y sólo si el determinante de la matriz en (5.3) desaparece, es decir, si

$$a\left(\frac{\mathrm{d}y}{\mathrm{d}x}\right)^2 - b\frac{\mathrm{d}y}{\mathrm{d}x} + c = 0, \quad \text{con } a \neq 0 \text{ en } P, \quad \mathrm{d}x \neq 0. \tag{5.4}$$

La ecuación (5.4) se llama *ecuación característica* de (5.1) y es una ecuación cuadrática en $\mathrm{d}y/\mathrm{d}x(P)$. Las soluciones de (5.4) se llaman *direcciones características* de (5.1) en el punto P. Podemos definir ahora:

Definición 5.1. *En el punto $P \in G$, la ecuación diferencial* (5.1) *se llama*

a) elíptica*, si* (5.4) *no posee soluciones reales, es decir, si $b^2 - 4ac < 0$,*
b) hiperbólica*, si* (5.4) *posee dos soluciones reales, es decir, si $b^2 - 4ac > 0$,*
b) parabólica*, si* (5.4) *posee exactamente una solución real, es decir, si $b^2 - 4ac = 0$.*

5.1.2. Sistemas cuasi-lineales de primer orden. En lo siguiente, consideraremos en $G \subset \mathbb{R}^2$ un sistema cuasi-lineal de primer orden ($a_1c_2 - a_2c_1 \neq 0$):

$$\begin{aligned} a_1u_x + b_1u_y + c_1v_x + d_1v_y &= f_1, \\ a_2u_x + b_2u_y + c_2v_x + d_2v_y &= f_2, \end{aligned} \tag{5.5}$$

donde los coeficientes a_i, b_i, c_i y d_i ($i = 1, 2$) son funciones de x, y, u y v. Se supone que los valores u, v de la solución están dados sobre $\Gamma \subset G$. Fijamos un punto $P \in \Gamma$ y preguntamos si las primeras derivadas u_x, u_y, v_x y v_y pueden ser determinadas de forma única en el punto P. Formando las derivadas de u y v en $P \in \Gamma$ en la dirección de Γ obtenemos

$$\begin{aligned} \mathrm{d}u &= u_x\,\mathrm{d}x + u_y\,\mathrm{d}y, \\ \mathrm{d}v &= v_x\,\mathrm{d}x + v_y\,\mathrm{d}y. \end{aligned} \tag{5.6}$$

Combinando (5.5) y (5.6) obtenemos el sistema lineal

$$\begin{bmatrix} a_1 & b_1 & c_1 & d_1 \\ a_2 & b_2 & c_2 & d_2 \\ \mathrm{d}x & \mathrm{d}y & 0 & 0 \\ 0 & 0 & \mathrm{d}x & \mathrm{d}y \end{bmatrix} \begin{pmatrix} u_x \\ u_y \\ v_x \\ v_y \end{pmatrix} = \begin{pmatrix} f_1 \\ f_2 \\ \mathrm{d}u \\ \mathrm{d}v \end{pmatrix},$$

cuyo determinante desaparece si y sólo si las primeras derivadas de la solución no son unicamente determinadas en P. Esto entrega la ecuación

$$\begin{aligned} (a_1c_2 - a_2c_1)\left(\frac{\mathrm{d}y}{\mathrm{d}x}\right)^2 - (a_1d_2 - a_2d_1 + b_1c_2 - b_2c_1)\frac{\mathrm{d}y}{\mathrm{d}x} \\ + b_1d_2 - b_2d_1 = 0 \quad \text{con } a_1c_2 - a_2c_1 \neq 0. \end{aligned} \tag{5.7}$$

La ecuación (5.7) es la *ecuación característica* asociada al sistema (5.6), y sus soluciones son las *direcciones características* en el punto P. El discriminante de (5.7) es

$$D(P) = (a_1d_2 - a_2d_1 + b_1c_2 - b_2c_1)^2 - 4(a_1c_2 - a_2c_1)(b_1d_2 - b_2d_1). \tag{5.8}$$

Definición 5.2. *En el punto $P \in G$ el sistema* (5.5) *se llama*

a) elíptico *si $D(P) < 0$, es decir, en P no existe nunguna dirección característica (real),*
b) hiperbólico *si $D(P) > 0$, es decir, en P existen dos direcciones características diferentes,*
b) parabólico *si $D(P) = 0$, es decir, en P existe exactamente una dirección característica.*

Si transformamos una ecuación

$$az_{xx} + bz_{xy} + cz_{yy} = f$$

a un sistema de primer orden del tipo (5.5) mediante la transformación

$$u := z_x, \quad v := z_y,$$

por supuesto las Definiciones 5.1 y 5.2 deben ser equivalentes. Pero es facil verificar que efectivamente son equivalentes: las ecuaciones características (5.4) y (5.7) son idénticas, y por lo tanto el sistema posee las mismas direcciones características que la ecuación diferencial escalr de segundo orden.

Mencionamos que las mismas consideraciones pueden ser aplicadas a ecuaciones escalares de primer orden del tipo

$$au_x + bu_y = f, \quad a \neq 0. \tag{5.9}$$

Derivando en $P \in \Gamma$ a lo largo de Γ obtenemos

$$\mathrm{d}u = u_x \,\mathrm{d}x + u_y \,\mathrm{d}y, \tag{5.10}$$

y combinando (5.9) y (5.10) llegamos a la ecuación característica

$$\frac{\mathrm{d}y}{\mathrm{d}x} = \frac{b}{a},$$

es decir, la ecuación diferencial posee en cada punto sólo una dirección característica, lo es el resultado esperado ya que las ecuaciones del tipo (5.9) describen fenómenos de convección.

5.1.3. Características de ecuaciones hiperbólicas. Queremos dedicarnos ahora al caso hiperbólico, es decir al caso en el cual en cada punto del dominio de definición existen dos direcciones características diferentes. Si los coeficientes en (5.1) y (5.5) son continous, entonces la solución de las ecuaciones características (5.4) y (5.7), respectivamente, entrega dos campos de direcciones continuos sobre el dominio considerado. Estos campos de direcciones definen dos familias de curvas, las llamadas *características* de las ecuaciones diferenciales respectivas (5.1) y (5.5). La suavidad de los coeficientes en (5.1) y (5.5) se refleja en la suavidad de las características. Las características son portadores de información de la solución; en otras palabras, a lo largo de las características se realizan fenómenos físicos de propagación. Volviendo al origen de nuestras consideraciones, podemos formular el siguiente teorema.

Teorema 5.1. *Una solución de la ecuación diferencial hiperbólica* (5.1) *(respectivamente, del sistema hiperbólico* (5.5)*) dada sobre* $\Gamma \subset G$ *(en el caso de* (5.1)*, se supone que también las primeras derivadas están dadas sobre* Γ*) está determinada localmente y unicamente mas allá de* Γ *si y sólo si en ningún punto de* Γ*, la curva* Γ *coincide con una de las direcciones características de* (5.1) *(respectivamente,* (5.5)*); en otra palabras, si y sólo si intersecta las características bajo un ángulo positivo.*

En particular, este teorema implica que si los coeficientes de la ecuación diferencial son suficientemente suaves, un problema de valores iniciales hiperbólico posee una solución unicamente determinada si los valores iniciales estan dados sobre una curva *no* característica.

Consideremos ahora problemas de valores iniciales de ecuaciones diferenciales hiperbólicas de segundo orden y de sistemas hiperbólicos de primer orden. Sea Γ una curva suave en el plano (x, y), y supongamos que en cada punto Γ posee una pendiente finita, o sea, existe una parametrización $(x(\sigma), y(\sigma))$ de Γ con $\mathrm{d}x \neq 0$ sobre la totalidad de Γ. Además, sea Γ una curva *no característica* de las ecuaciones diferenciales correspondientes.

Ahora estamos buscando soluciones de los problemas

$$\begin{aligned} a_1 u_x + b_1 u_y + c_1 v_x + d_1 v_y &= f_1 \quad \text{para } (x,y) \in G \text{ arriba de } \Gamma, \\ a_2 u_x + b_2 u_y + c_2 v_x + d_2 v_y &= f_2 \quad \text{para } (x,y) \in G \text{ arriba de } \Gamma, \\ u(x,y) &= u_0(x,y) \quad \text{para } (x,y) \in \Gamma, \\ u_y(x,y) &= u_1(x,y) \quad \text{para } (x,y) \in \Gamma, \end{aligned} \tag{5.11}$$

o alternativamente,

$$\begin{aligned} a u_{xx} + b u_{xy} + c u_{yy} &= f \quad \text{para } (x,y) \in G \text{ arriba de } \Gamma, \\ u(x,y) &= u_0(x,y) \quad \text{para } (x,y) \in \Gamma, \\ u_y(x,y) &= u_1(x,y) \quad \text{para } (x,y) \in \Gamma. \end{aligned} \tag{5.12}$$

Comentamos que en (5.12) al lugar del dato u_y se puede especificar alguna otra derivada direccional de u sobre Γ, siempre que esta derivada direccional no coincida con la derivada en la dirección de Γ, dado que esta derivada ya esta automaticamente dada a través del dato u sobre Γ.

Los problemas de valores iniciales del tipo (5.11) y (5.12) conducen a soluciones unicamente determinadas. Aplicando la teoría de las características, analizaremos solamente el problema (5.12); las consideraciones para (5.11) son análogas.

Las soluciones de la ecuación cuadrática entregan las direcciones características de (5.12); aquí obtenemos

$$\begin{aligned} \left(\frac{\mathrm{d}y}{\mathrm{d}x}\right)_1 &= \alpha = \frac{1}{2a}\left(b + \sqrt{b^2 - 4ac}\right), \\ \left(\frac{\mathrm{d}y}{\mathrm{d}x}\right)_2 &= \beta = \frac{1}{2a}\left(b - \sqrt{b^2 - 4ac}\right), \quad a \neq 0. \end{aligned} \tag{5.13}$$

A lo largo de las curvas características el determinante del sistema de ecuaciones (5.3) desaparece, por lo tanto en este caso (5.3) posee soluciones en este caso solamente si no crece

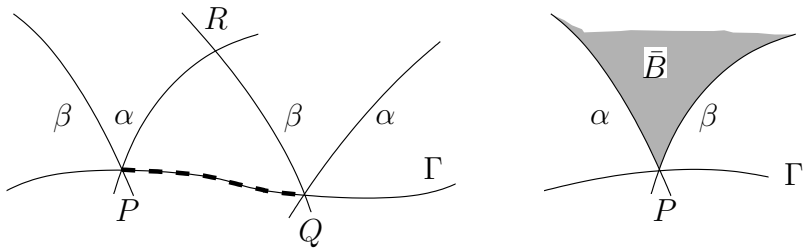


FIGURA 5.1. El intervalo de dependencia $[P, Q]$ del punto R (izquierda) y el dominio $\bar{B}$ de influencia del punto $P \in \Gamma$ (derecha).

el rango se la matriz extendida en (5.3), es decir, por ejemplo se debe satisfacer

$$\begin{vmatrix} a & f & c \\ \mathrm{d}x & \mathrm{d}(u_x) & 0 \\ 0 & \mathrm{d}(u_y) & \mathrm{d}y \end{vmatrix} = 0,$$

o sea,

$$a\,\mathrm{d}(u_x)\,\mathrm{d}y - f\,\mathrm{d}x\,\mathrm{d}y + c\,\mathrm{d}(u_y)\,\mathrm{d}x = 0;$$

por lo tanto, utilizando (5.13) obtenemos las siguientes ecuaciones ($a \neq 0$:)

$$\begin{aligned} a\alpha\,\mathrm{d}(u_x) + c\,\mathrm{d}(u_y) - f\,\mathrm{d}y = 0, \\ a\beta\,\mathrm{d}(u_x) + c\,\mathrm{d}(u_y) - f\,\mathrm{d}y = 0. \end{aligned} \tag{5.14}$$

Estas ecuaciones describen condiciones que la solución de (5.12) debe satisfacer *a lo largo de las características* α, β. Hay que tomar en cuenta que las derivadas $\mathrm{d}(u_x)$, $\mathrm{d}(u_y)$ y $\mathrm{d}y$ se refieren a las direcciones características α y β, respectivamente. Las ecuaciones (5.14) dan origen a un método de construcción de soluciones, ver Sección 5.2.

El comportamiento de la solución $u(x, y)$ de (5.12) depende solamente de los datos iniciales especificados en el segmento $[P, Q] \subset \Gamma$, es decir una modificación de los datos iniciales en $\Gamma \backslash [P, Q]$ (*fuera* de $[P, Q]$) *no* afecta el valor de la solución $u(R)$. El segmento $[P.Q]$ se llama *intervalo de dependencia* del punto R (ver Figura 5.1 (izquierda)).

Por otro lado, sea $P \in \Gamma$ un punto arbitrario $B \subset G$ el domino arriba de Γ localizado entre las dos características que pasan por P. Su clausura $\bar{B}$ es el conjunto de aquellos puntos en los cuales el valor de la solución $u(x, y)$ de (5.12) depende de los valores iniciales puestos en P; en otras palabras, el valor de la solución en algún punto fuera de $\bar{B}$ *no* es afecto por una modificación de las condiciones iniciales en $P \in \Gamma$ (y en una vecindad $\mathcal{U}(P) \subset \Gamma$ suficientemente pequeña). El dominio B se llama *dominio de influencia* del punto $P \in \Gamma$ (ver Figura 5.1 (derecha)).

Finalmente, comentamos que si a las ecuaciones (5.13) y (5.14) se agrega la ecuación diferencial

$$\mathrm{d}u = u_x\,\mathrm{d}x + u_y\,\mathrm{d}y,$$

donde hay que tomar las derivadas en una de las dos direcciones características, entonces la solución de (5.12) junto con los datos iniciales está determinada únicamente si suponemos que $a \neq 0$ y $c \neq 0$ en $G \times \mathbb{R}^3$.

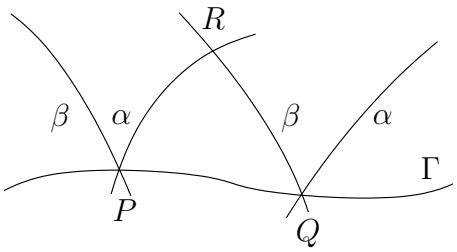


FIGURA 5.2. La curva inicial $\Gamma \ni P, Q$ y la interesección de las características.

5.2. Métodos de características numéricos

Utilizando el concepto de características podemos calcular soluciones de problemas de valores iniciales de ecuaciones hiperbólicas por un método que usa una malla característica compuesta por los puntos de intersección de líneas características. Este método origina en un trabajo de Massau (1899).

5.2.1. Método de características aproximado. Queremos estudiar este método para eole ejemplo de un problema de valores iniciales del tipo (5.12). Supongamos que la curva inicial Γ que aparece en (5.12) sea no característica. Entonces, según la teoría desarrollada en la Sección 5.1, el sistema que debe ser aproximado numéricamente es

$$\left(\frac{\mathrm{d}y}{\mathrm{d}x}\right)_1 = \alpha = \frac{1}{2a}\left(b + \sqrt{b^2 - 4ac}\right), \tag{5.15}$$

$$\left(\frac{\mathrm{d}y}{\mathrm{d}x}\right)_2 = \beta = \frac{1}{2a}\left(b - \sqrt{b^2 - 4ac}\right), \tag{5.16}$$

$$a\alpha\,\mathrm{d}(u_x) + c\,\mathrm{d}(u_y) - f\,\mathrm{d}y = 0, \tag{5.17}$$

$$a\beta\,\mathrm{d}(u_x) + c\,\mathrm{d}(u_y) - f\,\mathrm{d}y = 0, \tag{5.18}$$

$$\mathrm{d}u - u_x\,\mathrm{d}x - u_y\,\mathrm{d}y = 0, \tag{5.19}$$

donde los coeficientes

$$a, b, c, f : G \times \mathbb{R}^3 \to \mathbb{R}$$

son funciones dadas de x, y, u y ∇u. Las derivadas que aparecen en (5.17) y (5.18) se toman en las direcciones características α y β, respectivamente, mientras que las derivadas en (5.19) se toman en *una* de esas direcciones (α *o* β). Además, se supone que

$$b^2 - 4ac > 0 \quad \text{(hiperbolicidad)}, \quad ac \neq 0 \quad \text{en } G \times \mathbb{R}^3. \tag{5.20}$$

El caso $c = 0$ requiere un análisis separado.

Las ecuaciones (5.15)–(5.19) junto con los valores iniciales en Γ determinan la solución de (5.12) de manera única y por lo tanto son equivalentes a (5.12).

Ahora la curva Γ se particiona por un número de puntos, y sean P y Q puntos de partición vecinos, ver Figura 5.2. Sea R el punto de intersección de la α-característica por P con la β-característica por Q. Aproximando las derivadas en (5.15)–(5.19) por cuocientes

de diferencias y formando promedios para obtener los valores de promedio de los valores de las funciones restantes, obtenemos

$$y(R) - y(P) - \frac{1}{2}\big[\alpha(R) + \alpha(P)\big]\big[x(R) - x(P)\big] = 0, \tag{5.21}$$

$$y(R) - y(Q) - \frac{1}{2}\big[\beta(R) + \beta(Q)\big]\big[x(R) - x(Q)\big] = 0, \tag{5.22}$$

$$\begin{aligned} &\frac{1}{2}\big[a(R)\alpha(R) + a(P)\alpha(P)\big]\big[u_x(R) - u_x(P)\big] \\ +\frac{1}{2}\big[c(R) + c(P)\big]\big[u_y(R) - u_y(P)\big] &- \frac{1}{2}\big[f(R) + f(P)\big]\big[y(R) - y(P)\big] = 0, \end{aligned} \tag{5.23}$$

$$\begin{aligned} &\frac{1}{2}\big[a(R)\beta(R) + a(Q)\beta(Q)\big]\big[u_x(R) - u_x(Q)\big] \\ +\frac{1}{2}\big[c(R) + c(Q)\big]\big[u_y(R) - u_y(Q)\big] &- \frac{1}{2}\big[f(R) + f(Q)\big]\big[y(R) - y(Q)\big] = 0, \end{aligned} \tag{5.24}$$

$$\begin{aligned} u(R) - u(P) &- \frac{1}{2}\big[u_x(R) + u_x(P)\big]\big[x(R) - x(P)\big] \\ &- \frac{1}{2}\big[u_y(R) + u_y(P)\big]\big[y(R) - y(P)\big] = 0. \end{aligned} \tag{5.25}$$

Al lugar de (5.25) podriamos tambien considerar una aproximación en la dirección β:

$$\begin{aligned} u(R) - u(Q) &- \frac{1}{2}\big[u_x(R) + u_x(Q)\big]\big[x(R) - x(Q)\big] \\ &- \frac{1}{2}\big[u_y(R) + u_y(Q)\big]\big[y(R) - y(Q)\big] = 0. \end{aligned}$$

Las ecuaciones de diferencias (5.21)–(5.25) entregan un método que es consistente de segundo orden con (5.15)–(5.19) si todos los tamaños de paso se eligen del mismo orden de magnitud. En virtud de los datos iniciales puestos en Γ, todos los valores de funciones en los puntos P y Q son conocidos (Tarea). Las cantidades que hay que determinar son las cinco desconocidas

$$x(R), y(R), u(R), u_x(R), u_y(R). \tag{5.26}$$

Entonces, tenemos que resolver un sistema de cinco ecuaciones no lineales en cinco desconocidas.

Aplicando el método (5.21)–(5.25) a una familia de puntos $P_1, \ldots, P_N \in \Gamma$ obtenemos los datos (5.26) en una sucesión de puntos arriba de Γ, y desde allí se puede seguir con el método (ver Figura 5.3). Así, el dominio de computación queda acotado por las dos características límites por P_1 y P_N. El dominio incluido por estas dos características se llama *dominio de determinación* de la solución de (5.12) asociado con sel segmento $[P_1, P_N] \subset \Gamma$.

Se puede demostrar que el método de características es convergente de segundo orden.

5.2.2. Método predictor-corrector. Discutiremos ahora la solución del sistema de ecuaciones no lineales, proponiendo el *método predictor-corrector*. Para la computación de la primera solución aproximada (del predictor) utilizamos fórmulas explícitas al lugar de las cuatro ecuaciones implícitas (5.21)–(5.24). Luego, después de la solución de estas ecuaciones, se calcula la primera aproximación de $u(R)$ desde la quinta ecuación (5.25). Con la ayuda

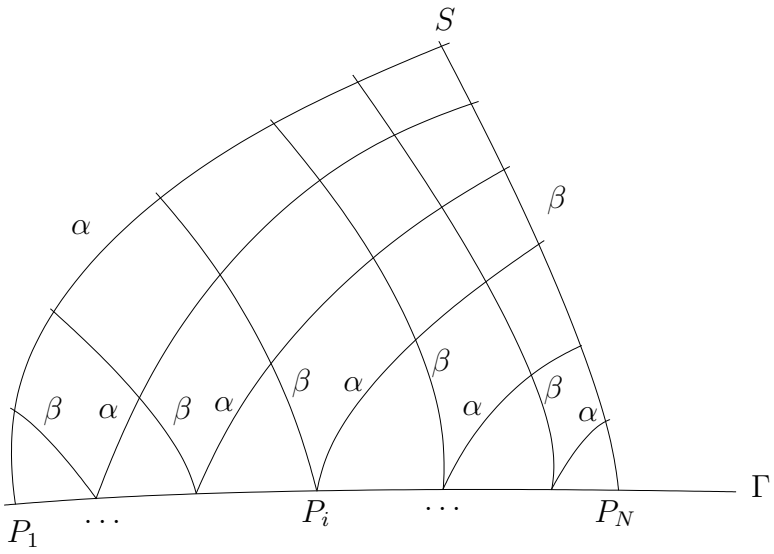


FIGURA 5.3. Método de características.

de estas primeras soluciones aproximadas y de las ecuaciones (5.21)–(5.25) podemos calcular segundas soluciones aproximadas (el corrector). Las fórmulas son las siguientes.

1. Definimos las abreviaturas

$$\begin{aligned} \alpha^{(0)} &:= \alpha(P), & c_p^{(0)} &:= c(P), \\ \beta^{(0)} &:= \beta(Q), & c_q^{(0)} &:= c(Q), \\ f_p^{(0)} &:= f(P), & \phi_p^{(0)} &:= a(P)\alpha(P), \\ f_q^{(0)} &:= f(Q), & \phi_q^{(0)} &:= a(Q)\beta(Q), \end{aligned}$$

y para $\nu = 1, \ldots, K-1$:

$$\begin{aligned} \alpha^{(\nu)} &:= \frac{1}{2}\big(\alpha^{(\nu)}(R) + \alpha(P)\big), & c_p^{(\nu)} &:= \frac{1}{2}\big(c^{(\nu)}(R) + c(P)\big), \\ \beta^{(\nu)} &:= \frac{1}{2}\big(\beta^{(\nu)}(R) + \beta(Q)\big), & c_q^{(\nu)} &:= \frac{1}{2}\big(c^{(\nu)}(R) + c(Q)\big), \\ f_p^{(\nu)} &:= \frac{1}{2}\big(f^{(\nu)}(R) + f(P)\big), & \phi_p^{(\nu)} &:= \frac{1}{2}\big(a^{(\nu)}(R)\alpha^{(\nu)}(R) + a(P)\alpha(P)\big), \\ f_q^{(\nu)} &:= \frac{1}{2}\big(f^{(\nu)}(R) + f(Q)\big), & \phi_q^{(\nu)} &:= \frac{1}{2}\big(a^{(\nu)}(R)\beta^{(\nu)}(R) + a(Q)\beta(Q)\big), \end{aligned}$$

y para $\nu = 0, \ldots, K-1$:

$$\begin{aligned} \psi_1^{(\nu)} &:= \alpha^{(\nu)}x(P) - y(P), & \psi_3^{(\nu)} &:= \phi_p^{(\nu)}u_x(P) + c_p^{(\nu)}u_y(P) - f_p^{(\nu)}y(P), \\ \psi_2^{(\nu)} &:= \beta^{(\nu)}x(Q) - y(Q), & \psi_4^{(\nu)} &:= \phi_q^{(\nu)}u_x(Q) + c_q^{(\nu)}u_y(Q) - f_q^{(\nu)}y(Q), \end{aligned}$$

y para $\nu = 1, \ldots, K-1$:

$$u_x^{(\nu)} := \frac{1}{2}\left(u_x^{(\nu)}(R) + u_x(P)\right), \qquad \delta x^{(\nu)} := x^{(\nu)}(R) - x(P),$$
$$u_y^{(\nu)} := \frac{1}{2}\left(u_y^{(\nu)}(R) + u_y(P)\right), \qquad \delta y^{(\nu)} := y^{(\nu)}(R) - y(P).$$

2. Luego, para $\nu = 0, 1, \ldots, K-1$ se resuelven las ecuaciones

$$\begin{bmatrix} \alpha^{(\nu)} & -1 & 0 & 0 \\ \beta^{(\nu)} & -1 & 0 & 0 \\ 0 & -f_p^{(\nu)} & \phi_p^{(\nu)} & c_p^{(\nu)} \\ 0 & -f_q^{(\nu)} & \phi_q^{(\nu)} & c_q^{(\nu)} \end{bmatrix} \begin{pmatrix} x^{(\nu+1)}(R) \\ y^{(\nu+1)}(R) \\ u_x^{(\nu+1)}(R) \\ u_y^{(\nu+1)}(R) \end{pmatrix} = \begin{pmatrix} \psi_1^{(\nu)} \\ \psi_2^{(\nu)} \\ \psi_3^{(\nu)} \\ \psi_4^{(\nu)} \end{pmatrix}, \tag{5.27}$$
$$u^{(\nu+1)}(R) = u(P) + u_x^{(\nu+1)}\delta x^{(\nu+1)} + u_y^{(\nu+1)}\delta y^{(\nu+1)}.$$

   Los coeficientes que aparecen en la matrix y en el vector del lado derecho de (5.27) son funciones de a, b, c y f y por lo tanto funciones de x, y, u y ∇u en los puntos P, Q y $R^{(\nu)}$. Se calculan utilizando la solución determinada en el paso anterior (con el índice ν).
3. Ponemos

$$x(R) := x^{(K)}(R), \quad y(R) := y^{(K)}(R),$$
$$u_x(R) := u_x^{(K)}(R), \quad u_y(R) := u_y^{(K)}(R), \quad u(R) = u^{(K)}(R).$$

La selección de $K \in \mathbb{Z}$ depende de la precisión con la cual queremos resolver el sistema no linear. Para $K = 2$ se calcula solamente una solución corrector.

La ecuación (5.27) implica que las primeras dos ecuaciones pueden ser resueltas independientemente de las demás ecuaciones. Estas ecuaciones sirven para la computación de las características. Geometricamente, (5.21) y (5.22) significan que las curvas características son remplazadas por segmentos de parábolas. Si la ecuación (5.12) es lineal o semilineal, se puede anticipar la computación completa de las características, dado que en este caso la solución de las primeras dos ecuaciones no es necesaria la computación de los valores $u(R)$, $u_x(R)$ y $u_y(R)$. Frecuentamente, en este caso también es posible integrar directamente, por métodos elementales, las ecuaciones diferenciales ordinarias (5.15) y (5.16), de manera que no es necesario utilizar las aproximaciones por diferencias (5.21) y (5.22). En el caso lineal, además, no se requiere aplicar el método predictor-corrector, dado que solamente hay resolver un sistema de ecuaciones lineales para cada punto de la malla.

Es facil ver que las condiciones (5.20) implican que las matrices que aparecen en (5.27) son no singulares, es decir, existe una solución única. El método no funciona bien para valores pequeños de $b^2 - 4ac$, es decir, si las dos familias características casí coinciden, en otras palabras, si la malla caracteristíca es muy degenerada. En el caso límite ($b^2 - 4a = 0$), es decir, en el caso parabólico, ya no se puede ejecutar el método de características.

En aquellos puntos en los cuales c desaparece (siempre suponiendo que $a \neq 0$) la ecuación (5.18) ya no aparece (puesto que $\beta = 0$ y $\mathrm{d}y = 0$). Para la discretización esto implica que

(5.27) se pone singular, y ya no podemos determinar u_y. Sin embargo podemos utilizar el método en el caso lineal (Tarea).

Ejemplo 5.1. *Consideremos dos ejemplos elementales de ecuaciones diferenciales lineales que permiten la computación de la malla característica "a priori".*

1. *Sean $a, b, c \in \mathbb{R}$ constantes. Las ecuaciones* (5.15) *y* (5.16) *inmediatamente implican que $\alpha, \beta \in \mathbb{R}$ son constantes, es decir, las características son dos familias paralelas de rectas.*
2. *Sean $a \equiv x^2$, $b \equiv 0$, $c \equiv -1$, $G := \{(x, y) : x > 0\} \subset \mathbb{R}^2$, y $\Gamma := \{(x, y) : x > 0,\ y = 0\} \subset G$. Las ecuaciones* (5.15) *y* (5.16) *implican*

$$\left(\frac{\mathrm{d}y}{\mathrm{d}x}\right)_{1,2} = \pm\frac{1}{x}, \quad x > 0.$$

Integrando obtenemos para $(x, y) \in G$ las siguientes ecuaciones de las características:

$$y_1(x) = \log x + d, \quad y_2(x) = -\log x + d, \quad d \in \mathbb{R}.$$

5.3. Métodos de diferencias finitas para problemas hiperbólicos

A parte de los métodos de características existen los métodos de diferencias finitas. Estos métodos usan una malla rectangular, paralela a los ejes, y en los puntos de la malla de aproximan las derivadas por cuocientes de diferencias. Obviamente, debido a la regularidad de la malla este método posee grandes ventajas comparado con los métodos de características, pero hay que tomar en cuenta que posiblemente la malla rectangular no refleja adecuadamente la naturaleza del problema matemático, dado que esta malla se impone artificialmente, y siempre hay que considerar las características. Queremos estudiar este problema para el siguiente problema de valores iniciales de la ecuación de la onda:

$$\begin{aligned} u_{tt}(x,t) &= c^2 u_{xx}(x,t), \quad x \in \mathbb{R}, \quad t > 0, \\ u(x,0) &= f(x), \quad x \in \mathbb{R}, \\ u_t(x,0) &= g(x), \quad x \in \mathbb{R}. \end{aligned} \tag{5.28}$$

5.3.1. La fórmula de d'Alembert. Antes de discutir métodos de diferencias finitas para (5.28) demostraremos que (5.28) puede ser resuelto explicitamente. De hecho, usando la transformación de variables

$$\xi = x + ct, \quad \eta = x - ct, \quad \Phi(\xi, \eta) = u(x, t) \tag{5.29}$$

y observando que

$$\frac{\partial}{\partial x} = \frac{\partial}{\partial \xi} + \frac{\partial}{\partial \eta}, \quad \frac{\partial}{\partial t} = c\left(\frac{\partial}{\partial \xi} - \frac{\partial}{\partial \eta}\right), \quad c > 0,$$

obtenemos de (5.28) la ecucación diferencial transformada

$$4c^2 \frac{\partial^2 \Phi}{\partial \xi \partial \eta} = 0,$$

cuyas soluciones son faciles de determinar:

$$\Phi(\xi, \eta) = P(\xi) + Q(\eta),$$

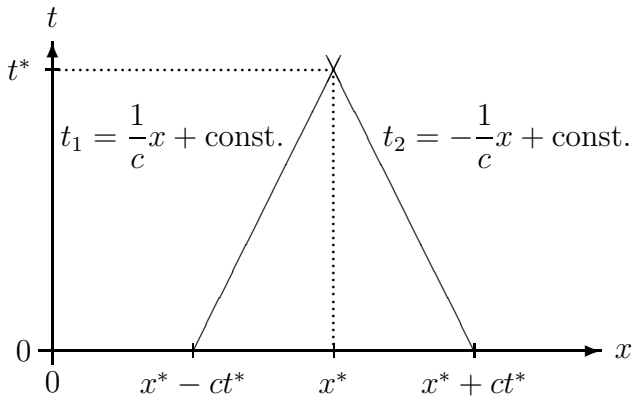


FIGURA 5.4. Características (5.31) y el intervalo de dependencia $[x^*-ct^*, x^*+ct^*]$ de la ecuación de la onda (5.28).

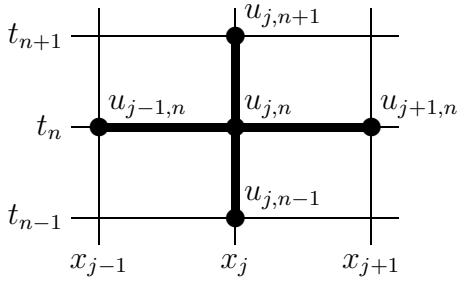


FIGURA 5.5. Esquema de 5 puntos del método (5.32).

con lo cual obtenemos en virtud de (5.29)

$$u(x,t) = P(x+ct) + Q(x-ct).$$

Aquí, P y Q son funciones arbitrarias y dos veces continuamente diferenciables. Utilizando esta solución general, obtenemos de las condiciones iniciales de (5.28) la siguiente expresión, conocida como la *fórmula de d'Alembert*:

$$u(x,t) = \frac{1}{2}\big(f(x+ct) + f(x-ct)\big) + \frac{1}{2c}\int_{x-ct}^{x+ct} g(\xi)\,\mathrm{d}\xi. \tag{5.30}$$

Calculando las características de la ecuación de la onda (5.28), obtenemos

$$\frac{\mathrm{d}t}{\mathrm{d}x} = \pm\frac{1}{c},$$

es decir,

$$t_1 = \frac{1}{c}x + \text{const.}, \quad t_2 = -\frac{1}{c}x + \text{const.} \tag{5.31}$$

La fórmula (5.30) implica directamente que el valor de la solución en el punto (x^*, t^*) depende precisamente de los valores iniciales en el intervalo $[x^* - ct^*, x^* + ct^*]$ del eje x. Esto es precisamente el intervalo de dependencia del punto (x^*, t^*), ver Figura 5.4.

5.3.2. Métodos explícitos para la ecuación de la onda. Consideraremos ahora aproximaciones de la ecuación de la onda por diferencias finitas. Para tal efecto utilizamos la malla

$$(x_j, t_n) = (j\Delta x, n\Delta t), \quad j \in \mathbb{Z}, \quad n \in \mathbb{N}_0,$$

y los siguientes cuocientes de diferencias de segundo orden para aproximar las derivadas en (5.28):

$$\begin{aligned} &u_{j,n+1} = 2(1 - c^2\lambda^2)u_{jn} + c^2\lambda^2(u_{j+1,n} + u_{j-1,n}) - u_{j,n-1}, \\ &j \in \mathbb{Z}, \quad n \in \mathbb{N}, \quad \lambda := \frac{\Delta t}{\Delta x}, \end{aligned} \tag{5.32}$$

donde $u_{jn} \approx u(\Delta x, t_n)$. Este método conecta 5 puntos de la malla según el esquema de la Figura 5.5. Para la computación de la capa de t número $n+1$ necesitamos los valores de las funciones de las dos capas que están debajo, por lo tanto tenemos que resolver el problema de calcular los valores numéricos en la primera capa t_1, es dedir, en los puntos de malla $(x_j, \Delta t)$.

Dado que el método (5.32) es de segundo orden de consistencia en Δt y Δx, queremos mantener el segundo orden de consistencia también al calcular la primera capa de tiempo. En virtud de (5.28) obtenemos mediante un desarrollo en serie de Taylor:

$$u(x, \Delta t) = f(x) + \Delta t g(x) + \frac{\Delta t^2}{2} c^2 f''(x) + \mathcal{O}(\Delta t^3), \tag{5.33}$$

lo que corresponde a la aproximación

$$u_{j1} = f_j + \Delta t g_j + \frac{c^2\lambda^2}{2}(f_{j-1} - 2f_j + f_{j+1}), \quad j \in \mathbb{Z}. \tag{5.34}$$

Por otro lado, extendiendo el desarrollo en serie de Taylor en (5.33) podemos lograr aproximaciones mas exactas; por ejemplo, (5.28) implica que

$$u(x, \Delta t) = f(x) + \Delta t g(x) + \frac{\Delta t^2}{2} c^2 f''(x) + \frac{\Delta t^3}{6} c^2 g''(x) + \mathcal{O}(\Delta t^4),$$

lo que motiva la aproximación de tercer orden

$$u_{j1} = f_j + \Delta t g_j + \frac{c^2\lambda^2}{2}(f_{j-1} - 2f_j + f_{j+1}) + \Delta t \frac{c^2\lambda^2}{6}(g_{j-1} - 2g_j + g_{j+1}), \quad j \in \mathbb{Z}. \tag{5.35}$$

Por supuesto, si podemos calcular las derivadas de f y g en forma explícita, no es necesario utilizar las aproximaciones por diferencias en (5.34) y (5.35).

5.3.3. La condición de Courant-Friedrichs-Lewy (CFL). La convergencia del método (5.32) o de un método equivalente a (5.32) se estudiará en la Sección 5.5 en el marco de una teoría general de convergencia para problemas de valores iniciales lineales. Aquí queremos estudiar desde un punto de vista puramente geométrico bajo qué condiciones (5.32) *no* puede converger a la solución de (5.28) para $\Delta x, \Delta t \to 0$.

La ecuación (5.32) implica que el valor de la solución aproximada en el punto (x^*, t^*) depende precisamente de los datos iniciales en los puntos de malla en el intervalo

$$I := \left[x^* - \frac{\Delta x}{\Delta t} t^*, x^* + \frac{\Delta x}{\Delta t} t^* \right]$$

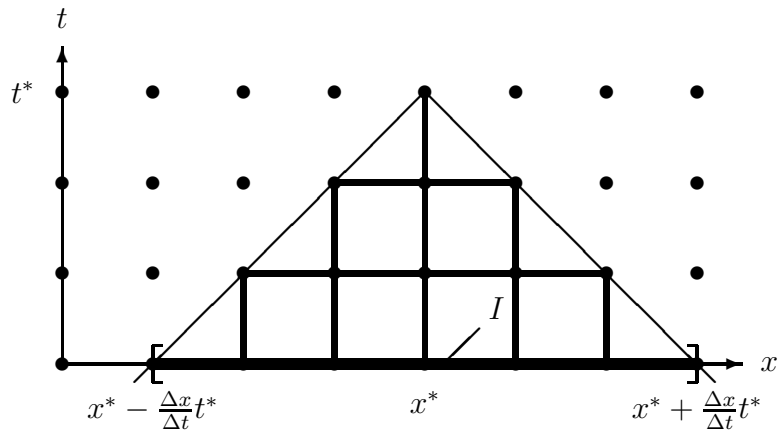


FIGURA 5.6. El intervalo I de dependencia numérica del punto de malla (x^*, t^*).

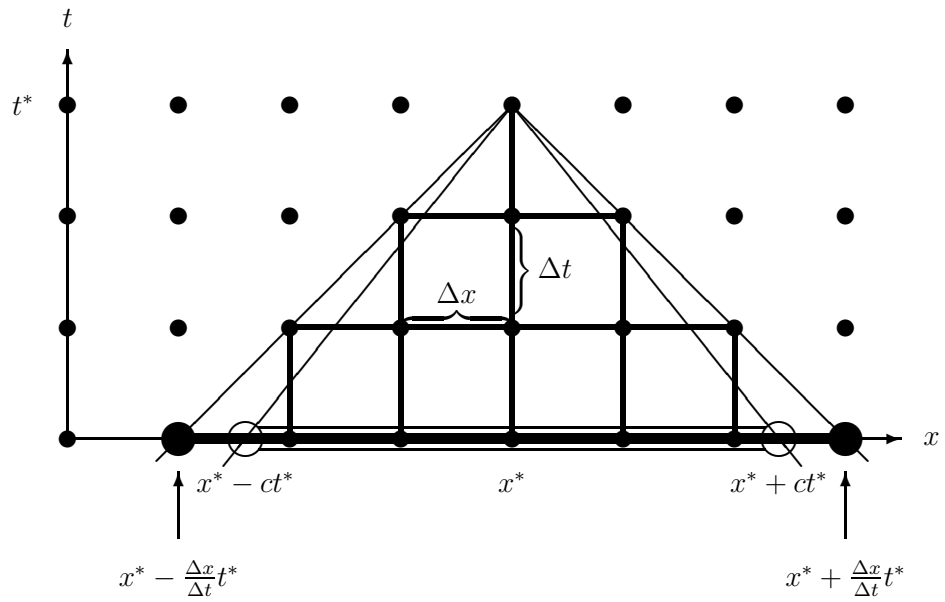


FIGURA 5.7. Satisfacción de la condición CFL para $\Delta t/\Delta x \leqslant 1/c$. Los subintervalos del eje x limitados por círculos grandes abiertos ($\circ$) y cerrados ($\bullet$) son los intervalos de dependencia analítica y numérica del punto (x^*, t^*), respectivamente.

del eje x. El intervalo I se llama *intervalo de dependencia numérica* del punto de malla (x^*, t^*), ver Figura 5.6.

Ahora, si

$$c\lambda = \frac{c\Delta t}{\Delta x} \leqslant 1,$$

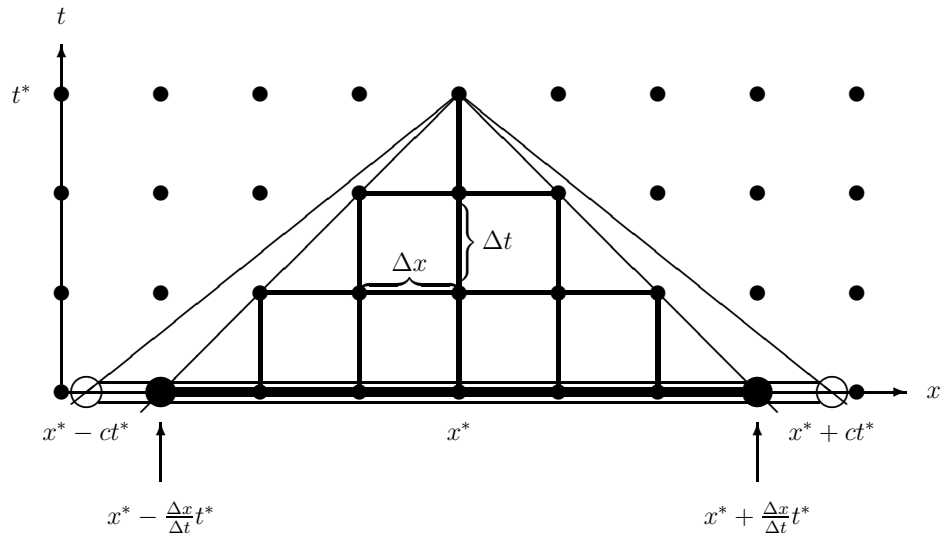


FIGURA 5.8. Violación de la condición CFL para $\Delta t/\Delta x > 1/c$. Los subintervalos del eje x limitados por círculos grandes abiertos ($\circ$) y cerrados ($\bullet$) son los intervalos de dependencia analítica y numérica del punto (x^*, t^*), respectivamente.

entonces

$$\frac{\Delta t}{\Delta x} \leqslant \frac{1}{c}$$

y el intervalo de dependencia numérica de (x^*, t^*) *incluye* el intervalo de dependencia analítica $[x^* - ct^*, x^* + ct^*]$ de este punto, ver Figura 5.7. Por otro lado, si $\lambda > 1/c$ y dejamos Δx y Δt tender a cero, entonces llegamos a un intervalo de dependencia de (x^*, t^*) que intrínsicamente está contenido en el intervalo de dependencia analítica, ver Figura 5.8. Por lo tanto, en este caso el método (5.32) no puede converger. Tenemos el siguiente teorema.

Teorema 5.2 (Condición de Courant, Friedrichs y Lewy (CFL))**.** *Una condición necesaria para que la solución de* (5.32) *converge en* (x^*, t^*) *hacia la solución de* (5.28) *para* $\Delta x, \Delta t \to 0$ *y datos iniciales* $f(x)$ *y* $g(x)$ *arbitrariamente suaves es la condición que el intervalo de dependencia numérica de* (x^*, t^*) *incluya el intervalo de dependencia analítica de este punto.*

Este teorema es válido para problemas de valores iniciales de ecuaciones hiperbólicas en general, y para todos métodos de diferencias finitas.

El método (5.32) es un método de dos pasos explícito. Queremos indicar dos métodos implícitos para la ecuación de la onda. La computación de la primera capa de tiempo se realiza como en (5.32) o (5.34).

Utilizando la notación

$$\delta^2 \phi_j := \phi_{j-1} - 2\phi_j + \phi_{j+1}$$

podemos aproximar (5.28) por

$$u_{j,n+1} - 2u_{jn} + u_{j,n-1} = \frac{c^2\lambda^2}{2}(\delta^2 u_{j,n+1} + \delta^2 u_{j,n-1}),$$

o equivalentemente,

$$\begin{aligned} &-\frac{c^2\lambda^2}{2}u_{j-1,n+1} + (1+c^2\lambda^2)u_{j,n+1} - \frac{c^2\lambda^2}{2}u_{j+1,n+1} \\ &= 2u_{jn} + \frac{c^2\lambda^2}{2}u_{j-1,n-1} - (1+c^2\lambda^2)u_{j,n-1} + \frac{c^2\lambda^2}{2}u_{j+1,n-1}. \end{aligned} \tag{5.36}$$

Una sóla aplicación del esquema (5.36) conecta 7 puntos de la malla. La evaluación numérica de (5.36) se realiza para cada capa de t mediante la solución de un sistema de ecuaciones lineales cuya matriz es una M-matriz simétrica, por lo tanto la solución de (5.36) es única y no pone problemas.

Otra posibilidad de aproximación es la fórmula

$$u_{j,n+1} - 2u_{jn} + u_{j,n-1} = \frac{c^2\lambda^2}{4}(\delta^2 u_{j,n+1} + 2\delta^2 u_{jn} + \delta^2 u_{j,n-1}),$$

o equivalentemente,

$$\begin{aligned} &-\frac{c^2\lambda^2}{4}u_{j-1,n+1} + \left(1+\frac{c^2\lambda^2}{2}\right)u_{j,n+1} - \frac{c^2\lambda^2}{4}u_{j+1,n+1} \\ &= \frac{c^2\lambda^2}{2}u_{j-1,n} + (2-c^2\lambda^2)u_{jn} + \frac{c^2\lambda^2}{4}u_{j+1,n} \\ &\quad + \frac{c^2\lambda^2}{4}u_{j-1,n-1} - \left(1+\frac{c^2\lambda^2}{2}\right)u_{j,n-1} + \frac{c^2\lambda^2}{4}u_{j+1,n-1}. \end{aligned} \tag{5.37}$$

El esquema (5.37) conecta 9 puntos. Igualmente, tal como para el método (5.36), la solución de (5.37) requiere la solución de un sistema de ecuaciones lineales.

Los métodos (5.36) y (5.37) son consistentes de segundo orden en Δt y Δx, lo que se puede demostrar facilmente calculando el error de truncación local (Tarea). También se puede demostrar la convergencia de estos métodos para $\Delta x, \Delta t \to 0$ y $\lambda > 0$ arbitrario (ver Richtmyer & Morton 1967).

5.3.4. Ecuación de la onda con datos iniciales y de frontera. Hasta ahora solamente hemos considerado el problema de valores iniciales para la ecuación de la onda. Sin embargo, frecuentamente esta ecuación aparece con datos inicales *y de frontera*, por ejemplo en el siguiente problema de valores iniciales y de frontera:

$$\begin{aligned} u_{tt}(x,t) &= c^2 u_{xx}(x,t), \quad t \geqslant 0, \quad 0 \leqslant x \leqslant L, \\ u(x,0) &= f(x), \quad 0 \leqslant x \leqslant L, \\ u_t(x,0) &= g(x), \quad 0 \leqslant x \leqslant L, \\ u(0,t) &= 0, \quad t \geqslant 0, \\ u(L,t) &= 0, \quad t \geqslant 0. \end{aligned} \tag{5.38}$$

Por supuesto, los datos iniciales y de frontera deben satisfacer ciertas condiciones de compatibilidad. Utilizando un planteo de separación también podemos resolver exactamente el problema (5.38). Por otro lado, en otras situaciones se presentan condiciones de periodicidad, por ejemplo como

$$\begin{aligned} u_{tt}(x,t) &= c^2 u_{xx}(x,t), \quad t \geqslant 0, \quad x \in \mathbb{R}, \\ u(x,0) &= f(x), \quad x \in \mathbb{R}, \\ u_t(x,0) &= g(x), \quad x \in \mathbb{R}, \\ u(x,t) &= u(x+L,t), \quad t \geqslant 0, \quad x \in \mathbb{R}. \end{aligned} \tag{5.39}$$

El tratamiento numérico de (5.38) y (5.39) puede ser realizado mediante el método (5.32), (5.36) o (5.37). Debido a la periodicidad, el problema (5.39) debe ser resuelto sólo en el intervalo $0 \leqslant x \leqslant L$, y para la computación de cada capa de t se pueden utilizar los puntos de malla en $[0, L]$ a la izquierda y a la derecha.

5.3.5. Métodos de diferencias finitas para sistemas hiperbólicos de primer orden. Ya mencionamos que las ecuaciones diferenciales de segundo orden pueden ser transformadas en sistemas equivalentes de primer orden. Por ejemplo, definamos

$$v(x,t) := u(x,t), \quad w(x,t) := \frac{1}{c}\int_0^x g(\xi)\,\mathrm{d}\xi + c\int_0^t u_x(x,\tau)\,\mathrm{d}\tau, \quad c > 0.$$

Ahora un cálculo directo, con la ayuda de (5.28), entrega las siguientes identidades:

$$\begin{aligned} v_t = u_t &= \int_0^t u_{\tau\tau}\,\mathrm{d}\tau + g(x) = c^2\int_0^t u_{xx}\,\mathrm{d}\tau + g(x) \\ &= c\left(c\int_0^t u_{xx}(x,\tau)\,\mathrm{d}\tau + \frac{g(x)}{c}\right) = cw_x, \\ w_t &= cu_x = cv_x, \\ v(x,0) &= u(x,0) = f(x), \\ w(x,0) &= \frac{1}{c}\int_0^x g(\xi)\,\mathrm{d}\xi =: G(x). \end{aligned}$$

Entonces, podemos escribir el problema de valores iniciales como

$$\begin{pmatrix} v_t(x,t) \\ w_t(x,t) \end{pmatrix} = c\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}\begin{pmatrix} v_x(x,t) \\ w_x(x,t) \end{pmatrix}, \quad \begin{pmatrix} v(x,0) \\ w(x,0) \end{pmatrix} = \begin{pmatrix} f(x) \\ G(x) \end{pmatrix}. \tag{5.40}$$

Sustituyendo

$$v := u_t, \quad w := cu_x$$

obtenemos el mismo sistema de ecuaciones, pero con la condición inicial

$$\begin{pmatrix} v(x,0) \\ w(x,0) \end{pmatrix} = \begin{pmatrix} g(x) \\ cf'(x) \end{pmatrix}.$$

Para el tratamiento numérico por métodos de diferencias finitas la ventaja de sistemas de primer orden (comparado con ecuaciones escalares de segundo orden) es la posibilidad de poder aplicar métodos de paso simple, es decir podemos considerar métodos de diferencias

finitas que conectan solamente dos capas de t. Por supuesto, también aquí hay que considerar el comportamiento de las características. Se puede verificar que las características del sistema (5.40) son idénticas a las características de la ecuación de la onda (5.28).

Consideraremos ahora problemas de valores iniciales hiperbólicos que poseen la forma un poco más general

$$\begin{aligned} \boldsymbol{u}_t &= \boldsymbol{A}\boldsymbol{u}_x, \quad x \in \mathbb{R}, \quad t \geqslant 0, \\ \boldsymbol{u}(x,0) &= \boldsymbol{\phi}(x), \quad x \in \mathbb{R}, \end{aligned} \tag{5.41}$$

donde

$$\boldsymbol{u}(x,t) = \begin{pmatrix} v(x,t) \\ w(x,t) \end{pmatrix}, \quad \boldsymbol{\phi}(x) = \begin{pmatrix} f(x) \\ g(x) \end{pmatrix}, \quad \boldsymbol{A} = \begin{bmatrix} a & b \\ b & c \end{bmatrix}, \quad a, b, c \in \mathbb{R},$$

y donde se supone que $\boldsymbol{A}$ es regular y

$$(a+c)^2 - 4(ac - b^2) > 0. \tag{5.42}$$

En virtud de la Definición 5.2 y (5.8), (5.42) implica la hiperbolicidad del sistema (5.41). Utilizando (5.7) se puede demostrar facilmente que los valores propios de $-\boldsymbol{A}^{-1}$ son las direcciones características de (5.41). La hipótesis de simetría de $\boldsymbol{A}$ no es una restricción esencial, dado que en muchas aplicaciones prácticas $\boldsymbol{A}$ es efectivamente simétrica.

Sea $\lambda := \Delta t/\Delta x$. Se puede definir el método de diferencias

$$\begin{aligned} \begin{pmatrix} v_{j,n+1} \\ w_{j,n+1} \end{pmatrix} &= \begin{pmatrix} v_{jn} \\ w_{jn} \end{pmatrix} + \frac{\lambda}{2}\boldsymbol{A} \begin{pmatrix} v_{j+1,n} - v_{j-1,n} \\ w_{j+1,n} - w_{j-1,n} \end{pmatrix}, \quad j \in \mathbb{Z}, \quad n \in \mathbb{N}_0; \\ \begin{pmatrix} v_{j,0} \\ w_{j,0} \end{pmatrix} &= \begin{pmatrix} f_j \\ g_j \end{pmatrix}, \quad j \in \mathbb{Z}. \end{aligned} \tag{5.43}$$

Sin embargo, en el Ejemplo 5.2 en la Sección 5.5 veremos que este método posee propiedades de convergencia muy desventajosas y por lo tanto es inutil para la práctica.

Otro método (mucho mejor, como veremos en el Ejemplo 5.3 en la Sección 5.5) es

$$\begin{pmatrix} v_{j,n+1} \\ w_{j,n+1} \end{pmatrix} = \frac{1}{2}\begin{pmatrix} v_{j+1,n} + v_{j-1,n} \\ w_{j+1,n} + w_{j-1,n} \end{pmatrix} + \frac{\lambda}{2}\boldsymbol{A} \begin{pmatrix} v_{j+1,n} - v_{j-1,n} \\ w_{j+1,n} - w_{j-1,n} \end{pmatrix}, \quad j \in \mathbb{Z}, \quad n \in \mathbb{N}_0. \tag{5.44}$$

Los métodos (5.43) y (5.44) son métodos de paso simple que son consistentes de segundo orden en Δx y de primer orden en Δt.

Otro método consiste en calcular los valores de v en los puntos $(j\Delta x, n\Delta t)$ y los valores de w en los puntos intermedios $((j+1/2)\Delta x, n\Delta t)$. Esto entrega, por ejemplo,

$$\begin{pmatrix} v_{j,n+1} \\ w_{j-1/2,n+1} \end{pmatrix} = \begin{pmatrix} v_{jn} \\ w_{j-1/2,n} \end{pmatrix} + \lambda\boldsymbol{A} \begin{pmatrix} v_{j,n+1} - v_{j-1,n+1} \\ w_{j+1/2,n} - w_{j-1/2,n} \end{pmatrix}, \quad j \in \mathbb{Z}, \quad n \in \mathbb{N}_0.$$

Este es un método implícito. En el caso de que $\boldsymbol{A}$ posee la forma especial dada por (5.40), obtenemos el método

$$\begin{aligned} v_{j,n+1} &= v_{jn} + c\lambda\left(w_{j+1/2,n} - w_{j-1/2,n}\right), \\ w_{j-1/2,n+1} &= w_{j-1/2,n} + c\lambda\left(v_{j,n+1} - v_{j-1,n+1}\right). \end{aligned} \tag{5.45}$$

Este método es facil de resolver, dado que podemos primero calcular los valores $v_{j,n+1}$ desde la primera ecuación (explícita), luego utilizando los valores $v_{j,n+1}$ podemos calcular todos los valores $w_{j-1/2,n+1}$ usando la segunda ecuación.

Otro método para el sistema (5.40) es

$$\begin{aligned} v_{j,n+1} &= v_{jn} + \frac{c\lambda}{2}\left(w_{j+1/2,n} - w_{j-1/2,n} + w_{j+1/2,n+1} - w_{j-1/2,n+1}\right), \\ w_{j-1/2,n+1} &= w_{j-1/2,n} + \frac{c\lambda}{2}\left(v_{j,n+1} - v_{j-1,n+1} + v_{jn} - v_{j-1,n}\right). \end{aligned} \tag{5.46}$$

Los sistemas de ecuaciones de diferencias (5.45) y (5.46) son equivalentes a las ecuaciones (5.32) y (5.37), respectivamente, si identificamos v_{jn} con $(1/\Delta t)(u_{jn} - u_{j,n-1})$ y $w_{j-1/2,n}$ con $(c/\Delta x)(u_{jn} - u_{j-1,n})$. Esta identificación es razonable en virtud de la substitución $v = u_t$ y $w = cu_x$.

5.4. Métodos de diferencias finitas para problemas parabólicos

5.4.1. Solución exacta y características de la ecuación del calor. Para las ecuaciones diferenciales parabólicas tenemos una sola dirección característica en cada punto, y por lo tanto no podemos aplicar los métodos característicos descritos en Sección 5.2.

Aquí consideraremos métodos de diferencias finitas para el siguiente problema de valores iniciales para la ecuación del calor:

$$\begin{aligned} u_t &= u_{xx}, \quad t > 0, \quad x \in \mathbb{R}, \\ u(x,0) &= f(x), \quad x \in \mathbb{R}. \end{aligned} \tag{5.47}$$

Aqui se supone que $f \in C(\mathbb{R})$ es una función acotada. La solución explícita de (5.47) está dada por

$$u(x,t) = \frac{1}{\sqrt{4\pi t}} \int_{\mathbb{R}} \exp\left(-\frac{(\xi - x)^2}{4t}\right) f(\xi)\, \mathrm{d}\xi, \tag{5.48}$$

lo que es fácil de verificar tomando las derivadas bajo la integral. Por otro lado, en virtud de

$$\int_{\mathbb{R}} \exp(-y^2)\, \mathrm{d}y = \sqrt{\pi}$$

podemos escribir la solución (5.48) en la forma

$$u(x,t) = f(x) + \frac{1}{\sqrt{4\pi t}} \int_{\mathbb{R}} \exp\left(-\frac{(\xi - x)^2}{4t}\right) \big(f(\xi) - f(x)\big)\, \mathrm{d}\xi, \quad t > 0. \tag{5.49}$$

Esto implica que

$$\lim_{t \to 0} \frac{1}{\sqrt{4\pi t}} \exp\left(-\frac{(\xi - x)^2}{4t}\right) = 0 \quad \text{para } \xi \neq x,$$

además el integrando en (5.49) desaparece para $\xi = x$. Por lo tanto, las funciones (5.49) y (5.48) satisfacen la condición inicial en (5.47).

A través de la sustitución

$$v := u, \quad w := u_x$$

podemos transformar (5.47) al siguiente problema de valores iniciales:

$$\begin{aligned} v_x &= w, \\ v_t - w_x &= 0, \quad t > 0,\ x \in \mathbb{R}; \\ v(x,0) &= f(x), \\ w(x,0) &= f'(x), \quad x \in \mathbb{R}. \end{aligned} \tag{5.50}$$

La ecuación característica (5.7) para el sistema (5.50) entrega la ecuación $(\mathrm{d}t/\mathrm{d}x)^2 = 0$, es decir

$$t = \text{const.}$$

Entonces, las características de (5.47) es la familia de rectas paralelas al eje x. Además, (5.48) implica que el comportamiento de la solución $u(x,t)$ en un punto (x^*,t^*) depende de la función inicial $f(x)$ en la totalidad de la recta $t = 0$, es decir, el intervalo de dependencia de cada punto (x^*,t^*), $t^* > 0$, es el eje x entero. Además, la condición inicial in (5.47) es dada a lo largo de una característica, al contrario de los problemas de valores iniciales hiperbólicos.

Normalmente, la ecuación del calor aparece en la combinación con datos iniciales y de la frontera, por ejemplo como

$$\begin{aligned} u_t &= u_{xx}, \quad 0 \leqslant x \leqslant 1, \quad t > 0, \\ u(x,0) &= f(x), \quad 0 \leqslant x \leqslant 1, \\ u(0,t) &= g(t), \quad t > 0, \\ u(1,t) &= \tilde{g}(t), \quad t > 0. \end{aligned} \tag{5.51}$$

5.4.2. Métodos de paso simple para la ecuación del calor. Para el tratamiento numérico de (5.51) utilizamos la malla de puntos

$$(x_j, t_n) = (j\Delta x, n\Delta t), \quad j = 0, \dots, N+1, \quad n \in \mathbb{N}_0, \quad (N+1)\Delta x = 1,$$

sobre la cual definimos la aproximación por diferencias

$$\begin{aligned} \frac{1}{\Delta t}\left(u_{j,n+1} - u_{jn}\right) = \frac{1}{\Delta x^2}\big\{&(1-\alpha)(u_{j-1,n} - 2u_{jn} + u_{j+1,n}) \\ &+ \alpha(u_{j-1,n+1} - 2u_{j,n+1} + u_{j+1,n+1})\big\}. \end{aligned}$$

Definiendo

$$\lambda := \frac{\Delta t}{\Delta x^2}, \quad \delta^2\phi_j := \phi_{j-1} - 2\phi_j + \phi_{j+1}$$

y utilizando las condiciones iniciales y de frontera adicionales de (5.51), obtenemos los siguientes sistemas de ecuaciones:

$$\begin{aligned} u_{j,n+1} - \alpha\lambda\delta^2 u_{j,n+1} &= u_{jn} + (1-\alpha)\lambda\delta^2 u_{jn}, \quad j = 1, \dots, N, \quad n \geqslant 0, \\ u_{j0} &= f_j, \quad j = 0, \dots, N+1, \\ u_{0n} &= g_n, \quad u_{N+1,n} = \tilde{g}_n, \quad n \geqslant 0. \end{aligned} \tag{5.52}$$

El método es explícito para $\alpha = 0$ e implícito para $\alpha > 0$. Se trata de un método de paso simple que conecta la nueva "capa" de tiempo $(n+1)\Delta t$ con el peso α y la capa antigua con el peso $1-\alpha$. Podemos reescribir (5.52) en la forma

$$(\boldsymbol{I} + \alpha\lambda\boldsymbol{B})\boldsymbol{u}_{n+1} = \big(\boldsymbol{I} - (1-\alpha)\lambda\boldsymbol{B}\big)\boldsymbol{u}_n + \lambda\boldsymbol{g}_n, \quad n \in \mathbb{N}_0, \tag{5.53}$$

con $\boldsymbol{u}_0 := (f_1, \ldots, f_N)^{\mathrm{T}} \in \mathbb{R}^N$, $\boldsymbol{u}_n := (u_{1n}, \ldots, u_{Nn})^{\mathrm{T}} \in \mathbb{R}^N$, y

$$\boldsymbol{B} = \begin{bmatrix} 2 & -1 & 0 & \cdots & 0 \\ -1 & 2 & -1 & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & -1 & 2 & -1 \\ 0 & \cdots & 0 & -1 & 2 \end{bmatrix} \in \mathbb{R}^{N\times N}, \quad \boldsymbol{g}_n = \begin{pmatrix} g_n + \alpha(g_{n+1} - g_n) \\ 0 \\ \vdots \\ 0 \\ \tilde{g}_n + \alpha(\tilde{g}_{n+1} - \tilde{g}_n) \end{pmatrix} \in \mathbb{R}^N.$$

Dado que $\boldsymbol{I} + \alpha\lambda\boldsymbol{B}$ es una M-matriz tridiagonal, la solución de (5.53) no presenta problemas.

Para verificar la consistencia del método queremos calcular el error de truncación de (5.52). Supongamos que la solución de (5.51) es suficientemente suave. Puesto que

$$u_{tt} = u_{xxxx},$$

un desarrollo en serie de Taylor entrega

$$u(x_j, t_{n+1}) - u(x_j, t_n) = \Delta t \frac{\partial^2 u}{\partial x^2}(x_j, t_n) + \frac{\Delta t^2}{2}\frac{\partial^4 u}{\partial x^4}(x_j, t_n) + \mathcal{O}(\Delta t^3), \tag{5.54}$$

$$\frac{\partial^2 u}{\partial x^2}u(x_j, t_n) = \frac{1}{\Delta x^2}\delta^2 u(x_j, t_n) - \frac{\Delta x^2}{12}\frac{\partial^4 u}{\partial x^4}(x_j, t_n) + \mathcal{O}(\Delta x^4). \tag{5.55}$$

Insertando (5.55) en (5.54) obtenemos

$$\begin{aligned} u(x_j, t_{n+1}) - u(x_j, t_n) = \lambda\delta^2 u(x_j, t_n) + \Delta t\left(\frac{\Delta t}{2} - \frac{\Delta x^2}{12}\right)\frac{\partial^4 u}{\partial x^4}(x_j, t_n) \\ + \Delta t\big(\mathcal{O}(\Delta x^4) + \mathcal{O}(\Delta t^2)\big). \end{aligned}$$

Analogamante, desarrollando por (x_j, t_{n+1}) obtenemos

$$\begin{aligned} u(x_j, t_{n+1}) - u(x_j, t_n) = \lambda\delta^2 u(x_j, t_{n+1}) + \Delta t\left(-\frac{\Delta t}{2} - \frac{\Delta x^2}{12}\right)\frac{\partial^4 u}{\partial x^4}(x_j, t_{n+1}) \\ + \Delta t\big(\mathcal{O}(\Delta x^4) + \mathcal{O}(\Delta t^2)\big). \end{aligned}$$

Multiplicando la primera ecuación por $(1-\alpha)$, la segunda por α, y sumando los resultados obtenemos

$$u(x_j, t_{n+1}) - u(x_j, t_n) = \lambda\big((1-\alpha)\delta^2 u(x_j, t_n) + \alpha\delta^2 u(x_j, t_{n+1})\big) + \Delta t\tau_{jn},$$

donde

$$\tau_{jn} = \begin{cases} \mathcal{O}(\Delta t) + \mathcal{O}(\Delta x^2) & \text{para } 0 \leqslant \alpha \leqslant 1,\ \alpha \neq 1/2, \\ \mathcal{O}(\Delta t^2) + \mathcal{O}(\Delta x^4) & \text{para } \alpha = 0 \text{ y } \lambda = 1/6, \\ \mathcal{O}(\Delta t^2) + \mathcal{O}(\Delta x^2) & \text{para } \alpha = 1/2. \end{cases} \tag{5.56}$$

El resultado (5.56) implica que (5.52) es consistente a (5.51); en particuar, para el caso explícito ($\alpha = 0$) la selección $\lambda = 1/6$ entrega un orden de consistencia mayor que para $\lambda \neq 1/6$. Este resultado interesante fur observado por Milne en 1953. El método para $\alpha = 1/2$ se llama *método de Crank-Nicolson* (1947) y se usa frecuentemente debido a su alto orden de consistencia. Obviamente, un método explícito posee la ventaja de que no es necesario resolver sistemas de ecuaciones diferenciales. Conoceremos sus desventajas ahora al investigar las propiedades de convergencia.

En lo siguiente, no consideraremos errores de redondeo o de ingreso de datos. Sea

$$z_{jn} := u_{jn} - u(x_j, t_n), \quad j = 0, \ldots, N+1, \quad n = 0, \ldots, M$$

el error entre la solución exacta y la solución aproximada. Supongamos que las computaciones se realizan hasta un tiempo finito y fijo $T = M\Delta t$. Definamos

$$\boldsymbol{z}_n := (z_{1n}, \ldots, z_{Nn})^{\mathrm{T}} \in \mathbb{R}^N, \quad n = 0, \ldots, M$$

y consideremos que

$$u_{jn} = u(x_j, t_n) \quad \text{para } n = 0,\ j = 0 \text{ y } j = N+1$$

y por lo tanto

$$z_{0n} = z_{N+1,n} = z_{j0} = 0, \quad j = 0, \ldots, N+1, \quad n = 0, \ldots, M. \tag{5.57}$$

Entonces, restando (5.52) de (5.56), obtenemos en virtud de (5.57)

$$\begin{aligned} \boldsymbol{z}_{n+1} &= \boldsymbol{C}\boldsymbol{z}_n + \Delta t \boldsymbol{\sigma}_n, \quad n = 0, \ldots, M-1, \\ \boldsymbol{z}_0 &= 0, \end{aligned} \tag{5.58}$$

donde definimos

$$\begin{aligned} \boldsymbol{C} &:= (\boldsymbol{I} + \alpha\lambda\boldsymbol{B})^{-1}\big(\boldsymbol{I} - (1-\alpha)\lambda\boldsymbol{B}\big) \in \mathbb{R}^{N\times N}, \\ \boldsymbol{\sigma}_n &:= (\boldsymbol{I} + \alpha\lambda\boldsymbol{B})^{-1}\boldsymbol{\tau}_n \in \mathbb{R}^N, \\ \boldsymbol{\tau}_n &:= (\tau_{1n}, \ldots, \tau_{Nn})^{\mathrm{T}} \in \mathbb{R}^N. \end{aligned}$$

Para vectores $\boldsymbol{y} = (y_1, \ldots, y_N)^{\mathrm{T}} \in \mathbb{R}^N$ definimos la norma Euclidiana

$$\|\boldsymbol{y}\|_2 := \left(\frac{1}{N}\sum_{i=1}^{N} y_i^2\right)^{1/2};$$

entonces la norma matricial inducida es la norma espectral. Supongamos ahora que $\lambda := \Delta t/\Delta x^2$ satisface la condición

$$\lambda \begin{cases} \leqslant \dfrac{1}{2(1-2\alpha)} & \text{si } 0 \leqslant \alpha < \dfrac{1}{2}, \\ \text{arbitrario} & \text{si } \dfrac{1}{2} \leqslant \alpha \leqslant 1. \end{cases} \tag{5.59}$$

Entonces la simetría de $\boldsymbol{C}$ entrega después de un pequeño cálculo

$$\|\boldsymbol{C}\|_2 = r_\sigma(\boldsymbol{C}) < 1,$$

puesto que los valores propios $\gamma_1, \ldots, \gamma_N$ de $\boldsymbol{B}$ satisfacen

$$\gamma_j = 4\sin^2\left(\frac{j\pi}{2(N+1)}\right), \quad j = 1, \ldots, N.$$

Además, sabemos que

$$\left\|(\boldsymbol{I} + \alpha\lambda\boldsymbol{B})^{-1}\right\|_2 < 1,$$

y por lo tanto

$$\|\boldsymbol{\sigma}_n\|_2 \leqslant \|\boldsymbol{\tau}_n\|_2.$$

Entonces, tomando la norma en (5.58) obtenemos

$$\begin{aligned} \|\boldsymbol{z}_{n+1}\|_2 &\leqslant \|\boldsymbol{z}_n\|_2 + \Delta t\|\boldsymbol{\tau}_n\|_2, \quad n = 0, \ldots, M-1 = \frac{T-\Delta t}{\Delta t}, \\ \|\boldsymbol{z}_0\|_2 &= 0. \end{aligned} \tag{5.60}$$

Puesto que

$$\|\boldsymbol{\tau}_n\|_2 \leqslant \max_{1\leqslant j\leqslant N} |\tau_{jn}|,$$

se tiene que (5.56) también es válido para $\|\boldsymbol{\tau}_n\|_2$, $n = 0, \ldots, M$. Definiendo

$$\tau := \max_{0\leqslant n\leqslant M} \|\boldsymbol{\tau}_n\|_2,$$

obtenemos de (5.56) y (5.60)

$$\|\boldsymbol{z}_n\|_2 \leqslant T\tau = \begin{cases} \mathcal{O}(\Delta t) + \mathcal{O}(\Delta x^2) & \text{para } 0 \leqslant \alpha \leqslant 1 \\ \mathcal{O}(\Delta t^2) + \mathcal{O}(\Delta x^4) & \text{para } \alpha = 0 \text{ y } \lambda = 1/6, \\ \mathcal{O}(\Delta t^2) + \mathcal{O}(\Delta x^2) & \text{para } \alpha = 1/2, \end{cases} \quad n = 0, \ldots, M. \tag{5.61}$$

Resumimos los resultados de nuestro análisis en el siguiente teorema.

Teorema 5.3. *Bajo la condición* (5.59) *la solución del método de diferencias* (5.52) *converge para* $\Delta x, \Delta t \to 0$ *a la solución de* (5.51) *en la norma Euclidiana, y del orden indicado por* (5.61). *Aqui se supone que la solución de* (5.51) *es suficientemente suave.*

Comentamos que (5.59) implica que para $\alpha \geqslant 1/2$ no es necesario imponer restricciones a $\Delta t/\Delta x^2$.

En la Sección 5.5 veremos que (5.59) esencialmente también es necesario para la convergencia.

5.4.3. Métodos de dos pasos para la ecuación del calor. Enseguida discutiremos dos métodos de dos pasos para el tratamiento numérico de (5.47) y (5.51), respectivamente. Desde un punto de vista naivo, se podria considerar el siguiente método

$$u_{j,n+1} = u_{j,n-1} + 2\lambda\delta^2 u_{jn}, \quad j = 1, \ldots, N, \quad n \geqslant 1. \tag{5.62}$$

En la Sección 5.5 veremos que este método es enteramente inutil, puesto que no converge para ningún valor de λ, tan pequeño que sea.

Por otro lado, sí se recomienda el uso del *método de Du Fort y Frankel* (1953):

$$u_{j,n+1} = u_{j,n-1} + 2\lambda(u_{j+1,n} - u_{j,n+1} - u_{j,n-1} + u_{j-1,n}), \quad j = 1, \ldots, N, \quad n \geqslant 1. \tag{5.63}$$

Veremos en la Sección (5.5) que el método (5.63) tiene muy buenas propiedades de convergencia. Puesto que, además, el método es explícito, requiere sólo poco esfuerzo computacional. Sin embargo, queremos destacr una particularidad de la consistencia de (5.63). Algunos desarrollos en serie de Taylor de la solución exacta entregan las siguientes identidades:

$$\begin{aligned}
\frac{1}{2\Delta t}\big(u(x_j,t_{n+1})-u(x_j,t_{n-1})\big) &= \frac{\partial u}{\partial t}(x_j,t_n)+\frac{\Delta t^2}{6}\frac{\partial^3 u}{\partial t^3}(x_j,t_n)+\mathcal{O}(\Delta t^4),\\
\frac{1}{\Delta x^2}\delta^2 u(x_j,t_n) &= \frac{\partial^2 u}{\partial x^2}(x_j,t_n)+\frac{\Delta x^2}{12}\frac{\partial^4 u}{\partial x^4}(x_j,t_n)+\mathcal{O}(\Delta x^4),\\
&\frac{1}{\Delta x^2}\big(u(x_{j+1},t_n)-u(x_j,t_{n+1})-u(x_j,t_{n-1})+u(x_{j-1},t_n)\big)\\
&= \frac{1}{\Delta x^2}\delta^2 u(x_j,t_n)-\frac{\Delta t^2}{\Delta x^2}\frac{\partial^2 u}{\partial t^2}(x_j,t_n)+\mathcal{O}\left(\frac{\Delta t^4}{\Delta x^2}\right).
\end{aligned}$$

Debido a $u_t = u_{xx}$ la combinación de esas tres ecuaciones lleva a

$$\begin{aligned}
&\frac{u(x_j,t_{n+1})-u(x_j,t_{n-1})}{2\Delta t}-\frac{u(x_{j+1},t_n)-u(x_j,t_{n+1})-u(x_j,t_{n-1})+u(x_{j-1},t_n)}{\Delta x^2}\\
&=\frac{\Delta t^2}{\Delta x^2}\frac{\partial^2 u}{\partial t^2}(x_j,t_n)+\frac{\Delta t^2}{6}\frac{\partial^3 u}{\partial t^3}(x_j,t_n)-\frac{\Delta x^2}{12}\frac{\partial^4 u}{\partial x^4}(x_j,t_n)\\
&+\mathcal{O}\left(\frac{\Delta t^4}{\Delta x^2}\right)+\mathcal{O}(\Delta x^4)+\mathcal{O}(\Delta t^4).
\end{aligned}$$

Aqui vemos que una condición necesaria para la consistencia de (5.63) con la ecuación del calor $u_t = u_{xx}$ es

$$\lim_{\Delta x,\Delta t\to 0}\frac{\Delta t}{\Delta x}=0,$$

ya que si existiera un número $\beta > 0$ tal que

$$\lim_{\Delta x,\Delta t\to 0}\frac{\Delta t}{\Delta x}=\beta,$$

entonces (5.63) seria una aproximación consistente con la ecuación *hiperbólica*

$$u_t - u_{xx} + \beta^2 u_{tt} = 0.$$

5.5. La teoría de Lax y Richtmyer

5.5.1. El teorema de equivalencia de Lax. Esta sección resume la teoría de estabilidad para métodos de diferencias finitas de problemas de valores iniciales publicada en 1956 por P.D. Lax y R.D. Richtmyer. La teoría es similar a la teoría de Dahlquist para problemas de valores iniciales de ecuaciones diferenciales ordinarias. Aquí tambien obtenemos un teorema de equivalencia que nos permite estudiar el comportamiento de métodos de diferencias para una gran clase de problemas de valores iniciales, y queremos aplicar esta teoría a los métodos estudiados en las últimas dos secciones. La transformación de Fourier sera una herramienta útil para lograr esto.

Sea $I \subset \mathbb{R}$ un intervalo y X un espacio de Banach de funciones vectoriales definidas sobre I:

$$X = \Big\{ \boldsymbol{u} = \big(u^{(1)}, \ldots, u^{(p)}\big)^{\mathrm{T}} \,:\, \mathbb{R} \supset I \to \mathbb{C}^p \Big\}.$$

Sea $\|\cdot\|$ la norma de X. Sea $0 \leqslant t \leqslant T$ el intervalo de un parámetro, y sea $A : X \supset D(A) \to X$ un operador lineal diferencial con el dominio $D(A) \subset X$. Podemos escribir A en forma matricial de la siguiente manera:

$$A = (a_{ik}), \quad i,k = 1, \ldots, p, \quad a_{ik} = \sum_{\nu=0}^{s} a_{ik}^{(\nu)}(x) \frac{\partial^\nu}{\partial x^\nu}, \quad a_{ik}^{(\nu)} : \mathbb{R} \supset I \to \mathbb{C},$$

es decir, A *no* debe depender del parámetro t.

Se considera ahora el problema de valores iniciales

$$\frac{d}{dt}\boldsymbol{u}(t) = A\boldsymbol{u}(t), \quad 0 < t \leqslant T; \quad \boldsymbol{u}(0) = \boldsymbol{u}_0 \in X. \tag{5.64}$$

Una solución de (5.64) es una familia de elementos de X, $\boldsymbol{u}(t) : \mathbb{R} \supset [0,T] \to X$, que depende de t de forma diferenciable, y que satisface (5.64). Ahora, sea $D \subset D(A) \subset X$ el conjunto de aquellas funciones $\boldsymbol{u}_0 \in X$ a las cuales pertenece una solución *clásica* (o *intrínsica*) de (5.64). A través de (5.64), se define para cada $t \in [0,T]$ un operador lineal

$$\mathcal{E}_0(t) : X \supset D \to X, \quad \mathcal{E}_0(t)\boldsymbol{u}_0 = \boldsymbol{u}(t), \quad \boldsymbol{u}_0 \in D.$$

El operador $\mathcal{E}_0(t)$ se llama *operador de solución* asociado al problema (5.64). Se define lo siguiente.

Definición 5.3. *El problema de valores iniciales* (5.64) *se llama* correctamente puesto *si*

a) *El dominio $D \subset X$ de $\mathcal{E}_0(t)$ es denso en X.*
b) *La familia de operadores $\mathcal{E}_0(t)$, $t \in [0,T]$, es uniformemente acotada, es decir,*

$$\exists K_{\mathcal{E}} > 0 : \ \forall t \in [0,T] : \quad \|\mathcal{E}_0(t)\| \leqslant K_{\mathcal{E}}.$$

La segunda condición (b) dice que cada solución clásica $\boldsymbol{u}(t)$ de (5.64) depende continuamente de la función inicial $\boldsymbol{u}_0 \in D \subset X$. La primera condición (a) dice que cualquier función inicial $\boldsymbol{u}_0 \in X$ puede ser aproximada a precisión arbitraria por una función inicial que pertenece a D. Además, $\mathcal{E}_0(t)$ posee una extensión lineal y continua bien definida $\mathcal{E}(t) : X \to X$ definida sobre todo el espacio X con

$$\|\mathcal{E}(t)\| = \|\mathcal{E}_0(t)\| \quad \text{para } t \in [0,T].$$

Nos referimos a $\mathcal{E}(t)$ como *operador de solución generalizado* y a las funciones

$$\boldsymbol{u}(t) = \mathcal{E}(t)\boldsymbol{u}_0, \quad \boldsymbol{u}_0 \in X,$$

como *soluciones generalizadas* de (5.64). Se puede verificar facilmente que

$$\mathcal{E}(s+t) = \mathcal{E}(s)\mathcal{E}(t) \quad \text{para } s,t \geqslant 0. \tag{5.65}$$

Comentamos que los problemas de valores iniciales estudiados en las dos secciones anteriores son correctamente puestos.

Ahora consideremos aproximaciones por diferencias finitas al problema (5.64). Sean Δx y Δt tamaños de pasos escogidos de tal manera que Δx depende de alguna manera de Δt:

$$\Delta x = \psi(\Delta t), \quad \text{donde} \quad \lim_{\Delta t \to 0} \psi(\Delta t) = 0.$$

Además, sea $T : X \to X'$ el siguiente operador de translación:

$$T\boldsymbol{u}(x) = \boldsymbol{u}(x + \Delta x) \quad \forall \boldsymbol{u} \in X. \tag{5.66}$$

En los ejemplos que se considerarán más adelante, el espacio de Banach X' siempre es idéntico a X. En virtud de (5.66) es obvio como hay que definir las potencias T^ν, $\nu \in \mathbb{N}$.

Una aproximación por diferencias finitas que conecta dos capas de t (es decir, un método de paso simple) puede ser escrita en la forma

$$\boldsymbol{B}_1(\Delta t)\boldsymbol{u}_{n+1} = -\boldsymbol{B}_0(\Delta t)\boldsymbol{u}_n, \quad n = 0, \ldots, N-1 = \frac{T - \Delta t}{\Delta t},$$

donde

$$\boldsymbol{B}_\varrho(\Delta t) = \sum_{|\nu| < \infty} \boldsymbol{B}_\nu^{(\varrho)} T^\nu, \quad \boldsymbol{B}_\nu^{(\varrho)} \in \mathbb{C}^{p \times p}, \quad \varrho = 0, 1. \tag{5.67}$$

Suponiendo que el operador $\boldsymbol{B}_1(\Delta t)$ es invertible, podemos definir

$$\boldsymbol{C}(\Delta t) := -\boldsymbol{B}_1^{-1}(\Delta t)\boldsymbol{B}_0(\Delta t)$$

y escribir la aproximación por diferencias finitas en la forma

$$\boldsymbol{u}_{n+1} = \boldsymbol{C}(\Delta t)\boldsymbol{u}_n, \quad n = 0, \ldots, N-1. \tag{5.68}$$

Definición 5.4. *La aproximación por diferencias finitas* (5.68) *se llama* consistente *a* (5.64) *si para cada solución intrínsica* $\boldsymbol{u}(t)$, $t \in [0, T]$, *de una clase* $U \subset X$ *de funciones cuyas funciones iniciales* $\boldsymbol{u}(0) = \boldsymbol{u}_0$ *forman un conjunto denso en* X, *se satisface la relación*

$$\lim_{\Delta t \to 0} \frac{1}{\Delta t} \big\| \boldsymbol{u}(t + \Delta t) - \boldsymbol{C}(\Delta t)\boldsymbol{u}(t) \big\| = 0 \quad \text{uniformemente para } t \in [0, T]. \tag{5.69}$$

La expresión

$$\frac{1}{\Delta t} \big\| \boldsymbol{u}(t + \Delta t) - \boldsymbol{C}(\Delta t)\boldsymbol{u}(t) \big\|$$

se llama error de truncación *de la aproximación* (5.68).

Definición 5.5. *La aproximación por diferencias finitas* (5.68) *se llama* convergente *si para cada sucesión* $\{\Delta_j t\}_{j \in \mathbb{N}}$ *tal que* $\Delta_j t \to 0$ *para* $j \to \infty$ *y cada sucesión* $\{n_j\}_{j \in \mathbb{N}}$, $n_j \in \mathbb{N}$ *tal que* $n_j \Delta_j t \to t$ *para* $j \to \infty$ *con* $t \in [0, T]$ *arbitrario, y cualquier función inicial* $\boldsymbol{u}_0 \in X$, *tenemos para la solución* $\boldsymbol{u}(t) \in X$ *(posiblemente se trata de una solución generalizada) de* (5.64) *correspondiente:*

$$\lim_{j \to \infty} \big\| \boldsymbol{C}(\Delta_j t)^{n_j} \boldsymbol{u}_0 - \boldsymbol{u}(t) \big\| = 0. \tag{5.70}$$

Definición 5.6. *La familia de operadores de diferencias*

$$\boldsymbol{C}(\Delta t) : X \to X, \quad \Delta t > 0$$

se llama estable *si existen constantes* $\tau > 0$ *y* $K > 0$ *tales que*

$$\big\|\boldsymbol{C}(\Delta t)^n\big\| \leqslant K \quad \text{para } \Delta t \in (0, \tau],\ n\Delta t \in [0, T]. \tag{5.71}$$

Lema 5.1. *Si para cada* $\boldsymbol{u} \in X$ *existe una constante* $K(\boldsymbol{u}) > 0$ *tal que*

$$\big\|\boldsymbol{C}^n(\Delta t)\boldsymbol{u}\big\| \leqslant K(\boldsymbol{u}) \quad \text{para } \Delta t \in (0, \tau],\ n\Delta t \in [0, T],$$

entonces existe una constante K *tal que se satisface la condición de estabilidad* (5.71).

Demostración. Ver Richtmyer & Morton (1967), §2.5. ■

Teorema 5.4 (Teorema de equivalencia de Lax)**.** *Sea el problema de valores iniciales* (5.64) *correctamente puesto, y sea* (5.68) *consistente a* (5.64)*. Entonces la estabilidad de los operadores de diferencias* $\boldsymbol{C}(\Delta t)$ *en* (5.68) *es equivalente con la convergencia de* (5.68)*.*

Demostración.

1. Primero demostramos que la convergencia implica la estabilidad. Para tal efecto, notamos primero que para $\Delta t > 0$, los operadores $\boldsymbol{C}(\Delta t) : X \to X$ son lineales y continuos. Supongamos ahora que (5.71) *no* es válido. Entonces, existen una función $\boldsymbol{u}_0 \in X$ y sucesiones $\{\Delta_j t\}_{j\in\mathbb{N}}$ y $\{n_j\}_{j\in\mathbb{N}}$, $n_j \in \mathbb{N}$, $j \in \mathbb{N}$ y $n_j\Delta_j t \in [0, T]$ tales que
$$\big\|\boldsymbol{C}(\Delta_j t)^{n_j}\boldsymbol{u}_0\big\| \to \infty \quad \text{para } j \to \infty. \tag{5.72}$$
Como consecuencia, se debe satisfacer
$$\Delta_j t \overset{j\to\infty}{\longrightarrow} 0,$$
puesto que sino la sucesión $\{n_j\}_{j\in\mathbb{N}}$ sería acotada y en virtud de la dependencia continua de $\boldsymbol{C}(\Delta t)$ de $\Delta t > 0$, (5.72) no podria ser válido. Entonces existe una subsucesión $\{j_\nu\}_{\nu\in\mathbb{N}}$ tal que $n_{j_\nu}\Delta_{j_\nu} t \to t_0$ para $\nu \to \infty$ y algún $t_0 \in [0, T]$. En virtud de (5.72) tenemos que la condición de convergencia (5.70) no puede estar satisfecha, puesto que debido a la hipótesis de la correctitud del problema (5.64), el valor
$$\big\|\mathcal{E}(t_0)\boldsymbol{u}_0\big\| = \big\|\boldsymbol{u}(t_0)\big\|$$
es finito. Esto concluye la demostración de esta dirección.

   Se nota que hasta ahora no hemos utilizado la condición de consistencia (5.69). Efectivamente existen métodos de diferencias inconsistentes que convergen.
2. Ahora demostramos que la estabilidad implica la convergencia. Sea $D \subset X$ un subconjunto denso en X, tal que para cada $\boldsymbol{u}_0 \in D$ la familia de funciones $\boldsymbol{u}(t) = \mathcal{E}(t)\boldsymbol{u}_0$ entrega una solución intrínsica de (5.64), y la condición de consistencia (5.69) se satisface. Sean $\{n_j\}_{j\in\mathbb{N}}$ y $\{\Delta_j t\}_{j\in\mathbb{N}}$ sucesiones tales que $n_j\Delta_j t \to t \in [0, T]$ para $j \to \infty$. Además, sea ψ_j la diferencia entre la solución numérica y la solución exacta en el punto $n_j\Delta_j t$, es decir,
$$\begin{aligned}\psi_j &= \big(\boldsymbol{C}(\Delta_j t)^{n_j} - \mathcal{E}(n_j\Delta_j t)\big)\boldsymbol{u}_0 \\ &= \sum_{k=0}^{n_j-1} \boldsymbol{C}(\Delta_j t)^k\big(\boldsymbol{C}(\Delta_j t) - \mathcal{E}(\Delta_j t)\big) \cdot \mathcal{E}\big((n_j - 1 - k)\Delta_j t\big)\boldsymbol{u}_0;\end{aligned} \tag{5.73}$$

facilmente podemos confirmar que la segunda ecuación es correcta. Debido a la condición de estabilidad (5.71) sabemos que

$$\big\|\boldsymbol{C}(\Delta_j t)^k\big\| \leqslant K \quad \text{para } k = 0, \ldots, n_j - 1.$$

Además, la condición de consistencia (5.69) implica que para cada $\varepsilon > 0$ existe un número $\delta > 0$ tal que para $\Delta_j t < \delta$ y $k = 0, \ldots, n_j - 1$ se tiene que

$$\big\|\big(\boldsymbol{C}(\Delta_j t) - \mathcal{E}(\Delta_j t)\big)\boldsymbol{u}\big((n_j - 1 - k)\Delta_j t\big)\big\| \leqslant \varepsilon \Delta_j t.$$

Utilizando la desigualdad triangular, obtenemos ahora de (5.73):

$$\|\psi_j\| \leqslant K\varepsilon n_j \Delta_j t \leqslant KT\varepsilon \quad \text{para } \Delta_j t < \delta. \tag{5.74}$$

Dado que ε fue elegido arbitrariamente pequeño, (5.74) implica que

$$\lim_{j\to\infty} \|\psi_j\| = 0, \tag{5.75}$$

es decir, la condición de convergencia (5.70) está demostrada para cada $\boldsymbol{u}_0 \in D$ si el operador $\mathcal{E}(n_j\Delta_j t)$ puede ser remplazado por el operador $\mathcal{E}(t)$. Definiendo

$$t' := \min\{t, n_j\Delta_j t\}, \quad s := |t - n_j\Delta_j t|,$$

obtenemos de (5.65)

$$\mathcal{E}(n_j\Delta_j t) - \mathcal{E}(t) = \pm\big(\mathcal{E}(s) - \boldsymbol{I}\big),$$

donde "+" y "−" son válidos para $t < n_j\Delta_j t$ y $t \geqslant n_j\Delta_j t$, respectivamente. Ahora concluimos que

$$\big\|\big(\mathcal{E}(n_j\Delta_j t) - \mathcal{E}(t)\big)\boldsymbol{u}_0\big\| \leqslant K_{\mathcal{E}}\big\|\big(\mathcal{E}(s) - \boldsymbol{I}\big)\boldsymbol{u}_0\big\|,$$

donde $K_{\mathcal{E}}$ es la constante de la Definición 5.3. Puesto que $s \to 0$ cuando $n_j\Delta_j t \to t$ y $\mathcal{E}(0) = \boldsymbol{I}$, esta última desigualdad nos permite concluir que

$$\big\|\big(\mathcal{E}(n_j\Delta_j t) - \mathcal{E}(t)\big)\boldsymbol{u}_0\big\| \to 0 \quad \text{para } n_j\Delta_j t \to t \text{ y } \boldsymbol{u}_0 \in D. \tag{5.76}$$

Ahora, tomando en cuenta que

$$\big\|\big(\boldsymbol{C}(\Delta_j t)^{n_j} - \mathcal{E}(t)\big)\boldsymbol{u}_0\big\| \leqslant \|\psi_j\| + \big\|\big(\mathcal{E}(n_j\Delta_j t) - \mathcal{E}(t)\big)\boldsymbol{u}_0\big\|,$$

podemos deducir de (5.75) y (5.76) la condición de convergencia (5.70) para todo $\boldsymbol{u}_0 \in D$. Ahora sea $\boldsymbol{u} \in X$ arbitrario y $\{\boldsymbol{u}_\nu\}_{\nu\in\mathbb{N}} \subset D$ una sucesión que converge a $\boldsymbol{u}$. En este caso,

$$\begin{aligned}\big(\boldsymbol{C}(\Delta_j t)^{n_j} - \mathcal{E}(t)\big)\boldsymbol{u} = \big(\boldsymbol{C}(\Delta_j t)^{n_j} - \mathcal{E}(t)\big)\boldsymbol{u}_\nu + \boldsymbol{C}(\Delta_j t)^{n_j}(\boldsymbol{u} - \boldsymbol{u}_\nu)\\ - \mathcal{E}(t)(\boldsymbol{u} - \boldsymbol{u}_\nu).\end{aligned}$$

Obviamente, para j y ν suficientemente grande podemos achicar arbitrariamente los tres términos del lado derecho. Entonces, hemos establecido (5.70) para funcions $\boldsymbol{u} \in X$ arbitrarias, lo que concluye la demostración del teorema.

■

Según el Teorema 5.4, para la investigación de la convergencia de métodos de diferencias consistentes para problemas de valores iniciales correctamente puestos es suficiente verificar la estabilidad (o la inestabilidad) de la familia de operadores $\boldsymbol{C}(\Delta t)$ para $\Delta t > 0$. Antes de examinar algunos métodos especificos, demostraremos como podemos convertir un método de pasos múltiples en un método de paso simple.

5.5.2. Conversión de un método de pasos múltiples en un método de paso simple. En general, un método de diferencias finitas es de la forma

$$\boldsymbol{B}_q(\Delta t)\boldsymbol{u}_{n+q} + \ldots + \boldsymbol{B}_1(\Delta t)\boldsymbol{u}_{n+1} + \boldsymbol{B}_0(\Delta t)\boldsymbol{u}_n = 0. \tag{5.77}$$

Aquí se supone que $\boldsymbol{u}_n, \ldots, \boldsymbol{u}_{n+q-1}$ son conocidas, y que hay que calcular $\boldsymbol{u}_{n+q}$ desde (5.77). Para $q = 1$, nos encontramos en el caso especial (5.64). Para admitir una solución única, la inversa $(\boldsymbol{B}_q(\Delta t))^{-1}$ debe existir; ahora definimos

$$\boldsymbol{C}_\varrho(\Delta t) := -\big(\boldsymbol{B}_q(\Delta t)\big)^{-1}\boldsymbol{B}_\varrho(\Delta t), \quad \varrho = 0, \ldots, q-1,$$

es decir podemos escribir (5.77) como

$$\boldsymbol{u}_{n+q} = \boldsymbol{C}_{q-1}(\Delta t)\boldsymbol{u}_{n+q-1} + \ldots + \boldsymbol{C}_0(\Delta t)\boldsymbol{u}_n. \tag{5.78}$$

Ahora consideramos

$$\tilde{\boldsymbol{u}}_n := (\boldsymbol{u}_{n+q-1}, \ldots, \boldsymbol{u}_n)^{\mathrm{T}}, \quad n \in \mathbb{N}_0, \quad \boldsymbol{u}_\varrho \in X \quad \text{para } \varrho = n, \ldots, n+q-1$$

como elementos del espacio de Banach

$$\tilde{X} := X^q = \underbrace{X \times \cdots \times X}_{q}.$$

Aqui definimos la norma de $\tilde{X}$ de la siguiente manera: si $\|\boldsymbol{u}_\varrho\|$, $\varrho = n, \ldots, n+q-1$ es la norma de X, entonces el espacio $\tilde{X}$ está equipado con la norma

$$\|\tilde{\boldsymbol{u}}_n\| := \left(\sum_{\varrho=n}^{n+q-1} \|\boldsymbol{u}_\varrho\|^2\right)^{1/2}.$$

Ahora definimos los siguientes operadores de diferencias sobre $\tilde{X}$:

$$\tilde{\boldsymbol{C}}(\Delta t) := \begin{bmatrix} \boldsymbol{C}_{q-1}(\Delta t) & \cdots & \cdots & \cdots & \boldsymbol{C}_0(\Delta t) \\ \boldsymbol{I} & 0 & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & \boldsymbol{I} & 0 \end{bmatrix}, \quad \Delta t > 0.$$

El sistema de ecuaciones de diferencias (5.78) es equivalente a

$$\tilde{\boldsymbol{u}}_{n+1} = \tilde{\boldsymbol{C}}(\Delta t)\tilde{\boldsymbol{u}}_n.$$

Una pequeña modificación (ver §7.3 en Richtmyer y Morton, 1967) de la demostración del Teorema 5.4 demuestra que la estabilidad de la familia de operadores $\tilde{\boldsymbol{C}}(\Delta t) : \tilde{X} \to \tilde{X}$ es equivalente con la convergencia del método de q pasos (5.77) o (5.78).

5.5.3. Transformación de Fourier de métodos de diferencias. Consideremos ahora la aplicaciones del Teorema de Equivalencia de Lax. El análisis se realizará en la norma L^2, ya que en este caso los espacios de Banach X son espacios de Hilbert, y podemos utilizar la herramienta del análisis de Fourier.

Sea $I = [0, L] \subset \mathbb{R}$ un intervalo fijo, y consideremos funciones cuadraticamente integrables del período L:

$$\boldsymbol{u}(x) := \big(u^{(1)}(x), \dots, u^{(p)}(x)\big)^{\mathrm{T}} : \mathbb{R} \to \mathbb{C}^p, \quad \text{donde } \boldsymbol{u}(x) = \boldsymbol{u}(x + L) \text{ para } x \in \mathbb{R}.$$

Además, definimos la forma bilineal

$$(\boldsymbol{u}, \boldsymbol{v}) := \int_0^L \sum_{\nu=1}^{p} u^{(\nu)}(x)\bar{v}^{(\nu)}(x)\,\mathrm{d}x$$

y la norma

$$\|\boldsymbol{u}\| := (\boldsymbol{u}, \boldsymbol{u})^{1/2}.$$

Ahora, el espacio de Hilbert es

$$X := \big\{\boldsymbol{u} : \mathbb{R} \to \mathbb{C}^p \,:\, \boldsymbol{u}(x) = \boldsymbol{u}(x + L) \text{ para } x \in \mathbb{R};\ (\boldsymbol{u}, \boldsymbol{u}) < \infty\big\}. \tag{5.79}$$

Cada función $\boldsymbol{u} \in X$ puede ser representada como serie de Fourier en exactamente una manera. Por ejemplo, si definimos

$$\mathcal{L} := \{k = 2\pi m/L \,:\, m \in \mathbb{Z}\},$$

podemos escribir

$$\boldsymbol{u}(x) = \frac{1}{\sqrt{L}} \sum_{k\in\mathcal{L}} \boldsymbol{c}_k \mathrm{e}^{\mathrm{i}kx}, \quad x \in \mathbb{R}; \quad \boldsymbol{c}_k = \frac{1}{\sqrt{L}} \int_0^L \mathrm{e}^{-\mathrm{i}kx} u(x)\,\mathrm{d}x \in \mathbb{C}^p, \quad k \in \mathcal{L}. \tag{5.80}$$

Definimos

$$\boldsymbol{c}_k := \big(c_k^{(1)}, \dots, c_k^{(p)}\big)^{\mathrm{T}} \in \mathbb{C}^p$$

y la sucesión $\boldsymbol{c} = \{\boldsymbol{c}_k\}_{k\in\mathcal{L}}$. Entre dos sucesiones $\boldsymbol{c}$ y $\boldsymbol{d}$ de este tipo definimos la forma bilineal

$$(\boldsymbol{c}, \boldsymbol{d}) := \sum_{\nu=1}^{p} \sum_{k\in\mathcal{L}} c_k^{(\nu)} \bar{d}_k^{(\nu)}$$

y la norma

$$\|\boldsymbol{c}\| := (\boldsymbol{c}, \boldsymbol{c})^{1/2}.$$

Ahora sea V el espacio de Hilbert

$$V := \big\{\boldsymbol{c} = \{\boldsymbol{c}_k\}_{k\in\mathcal{L}} \,:\, \boldsymbol{c}_k \in \mathbb{C}^p,\ (\boldsymbol{c}, \boldsymbol{c}) < \infty\big\}. \tag{5.81}$$

Con cada $\boldsymbol{u} \in X$ asociamos de manera única el sistema $\boldsymbol{c} \in V$ de sus coeficientes de Fourier:

$$\phi : X \ni \boldsymbol{u} \mapsto \boldsymbol{c} \in V.$$

Es fácil ver que la aplicación ϕ es biyectiva y lineal. Además, en virtud de (5.80), sabemos que

$$\begin{aligned}\|\boldsymbol{u}\|^2 &= \int_0^L \sum_{\nu=1}^{p} u^{(\nu)}(x)\bar{u}^{(\nu)}(x)\,\mathrm{d}x \\ &= \frac{1}{L}\sum_{\nu=1}^{p}\sum_{k,l\in\mathcal{L}} c_k^{(\nu)}\bar{c}_l^{(\nu)} \int_0^L \exp\bigl(\mathrm{i}(k-l)x\bigr)\,\mathrm{d}x \\ &= \sum_{\nu=1}^{p}\sum_{k\in\mathcal{L}}\bigl|c_k^{(\nu)}\bigr|^2 = (\boldsymbol{c},\boldsymbol{c}) = \|\boldsymbol{c}\|^2.\end{aligned}$$

Acabamos de demostrar el siguiente resultado.

Lema 5.2. *Sean X y V los espacios de Hilbert definidos por* (5.79) *y* (5.81)*, respectivamente. Entonces la aplicación $\phi : X \to V$ mapea X a V de manera isomorfa e isométrica (es decir, se conserva la norma).*

Sea $\Delta x > 0$ un tamaño de paso y $T : X \to X$ el operador de translación definido en (5.66). Ahora queremos ver de cuál forma es el análogo isométrico de T, es decir, el operador

$$\phi T \phi^{-1} : V \to V.$$

Para $\boldsymbol{u} \in X$ y utilizando (5.66) y (5.80), obtenemos

$$\begin{aligned}\boldsymbol{v}(x) = (T\boldsymbol{u})(x) &= T\left(\frac{1}{\sqrt{L}}\sum_{k\in\mathcal{L}} \boldsymbol{c}_k \mathrm{e}^{\mathrm{i}kx}\right) \\ &= \frac{1}{\sqrt{L}}\sum_{k\in\mathcal{L}} \boldsymbol{c}_k T\boldsymbol{e}^{\mathrm{i}kx} \\ &= \frac{1}{\sqrt{L}}\sum_{k\in\mathcal{L}} \bigl(\boldsymbol{c}_k \mathrm{e}^{\mathrm{i}k\Delta x}\bigr)\mathrm{e}^{\mathrm{i}kx} \\ &= \frac{1}{\sqrt{L}}\sum_{k\in\mathcal{L}} \boldsymbol{d}_k \mathrm{e}^{\mathrm{i}kx},\end{aligned}$$

es decir, los coeficientes de Fourier de $\boldsymbol{v}(x)$ son

$$\boldsymbol{d}_k = \boldsymbol{c}_k \mathrm{e}^{\mathrm{i}k\Delta x} \quad \text{para } k \in \mathcal{L},$$

o sea, a la translación $T : X \to X$ por Δx (ver (5.66)) corresponde, en el espacio V de los coeficientes de Fourier, la multiplicación del coeficiente número k ($k \in \mathcal{L}$) por $\mathrm{e}^{\mathrm{i}k\Delta x}$.

5.5.4. Matrices de amplificación. Ahora supongamos que el operador A de (5.64) posee coeficientes constantes. En este caso, la aplicación ϕ mapea los operadores de diferencias

$$\boldsymbol{B}_\varrho(\Delta t) = \sum_{|\nu|<\infty} \boldsymbol{B}_\nu^{(\varrho)} T^\nu : X \to X, \quad \varrho = 0, 1$$

ya definido en (5.67) a la siguiente familia de matrices en el espacio V:

$$\boldsymbol{H}_\varrho(\Delta t,k):=\sum_{|\nu|<\infty}\mathrm{e}^{\mathrm{i}\nu k\Delta x}\boldsymbol{B}_\nu^{(\varrho)}\in\mathbb{C}^{p\times p},\quad \varrho=0,1,\quad k\in\mathcal{L}.$$

Definiendo

$$\boldsymbol{G}(\Delta t,k):=\boldsymbol{H}_1^{-1}(\Delta t,k)\boldsymbol{H}_0(\Delta t,k),\quad k\in\mathcal{L},$$

obtenemos como análogo isométrico de (5.68) la familia de sistemas lineales

$$\boldsymbol{c}_{k,n+1}=\boldsymbol{G}(\Delta t,k)\boldsymbol{c}_{k,n},\quad k\in\mathcal{L},\quad n=0,\dots,N-1,$$

donde $\boldsymbol{c}_{k,n},\boldsymbol{c}_{k,n+1}\in\mathbb{C}^p$ y $\boldsymbol{G}(\Delta t,k)\in\mathbb{C}^{p\times p}$. Formalmente, podemos escribir

$$\big\{\boldsymbol{G}(\Delta t,k)\big\}_{k\in\mathcal{L}}=\phi\boldsymbol{C}(\Delta t)\phi^{-1}.$$

Definición 5.7. *Las matrices $\boldsymbol{G}(\Delta t,k)\in\mathbb{C}^{p\times p}$, $k\in\mathcal{L}$, se llaman* matrices de amplificación *asociadas al operador de diferencias $\boldsymbol{C}(\Delta t)$.*

5.5.5. Teorema de Lax y Richtmyer y condición de estabilidad de von Neumann. Debido a la isometría de los espacios X y V podemos demostrar el siguiente teorema.

Teorema 5.5 (Lax y Richtmyer, 1956)**.** *Los operadores de diferencias $\boldsymbol{C}(\Delta t)$ que aparecen en* (5.68) *son estables (ver Definición 5.6) si y sólo si existen constantes $\tau>0$ y $K>0$ tales que*

$$\big\|\boldsymbol{G}^n(\Delta t,k)\big\|_2\leqslant K\quad\text{para }\Delta t\in(0,\tau],\ n\Delta t\in[0,T],\ k\in\mathcal{L}.\tag{5.82}$$

Aquí las matrices $\boldsymbol{G}(\Delta t,k)\in\mathbb{C}^{p\times p}$ son las matrices de amplificación asociadas a $\boldsymbol{C}(\Delta t)$, y $\|\cdot\|_2$ es la norma espectral.

Demostración.

a) Escogemos un elemento arbitrario $k\in\mathcal{L}$ y $\boldsymbol{c}_k\in\mathbb{C}^p$ tal que

$$\|\boldsymbol{c}_k\|_2=1,\quad \big\|\boldsymbol{G}^n(\Delta t,k)\boldsymbol{c}_k\big\|_2=\big\|\boldsymbol{G}^n(\Delta t,k)\big\|_2$$

para algún $n\in\mathbb{N}$. Luego definimos

$$\boldsymbol{u}(x):=\frac{1}{\sqrt{L}}\boldsymbol{c}_k\mathrm{e}^{\mathrm{i}kx}\in X.$$

Sabemos que $\|\boldsymbol{u}\|=1$, y en virtud del análisis anterior,

$$\big\|\boldsymbol{G}^n(\Delta t,k)\big\|_2=\big\|\boldsymbol{G}^n(\Delta t,k)\boldsymbol{c}_k\big\|_2=\big\|\boldsymbol{C}^n(\Delta t)\boldsymbol{u}\big\|\leqslant\big\|\boldsymbol{C}^n(\Delta t)\big\|,$$

por lo tanto obtenemos que

$$\sup_{k\in\mathcal{L}}\big\|\boldsymbol{G}^n(\Delta t,k)\big\|_2\leqslant\big\|\boldsymbol{C}^n(\Delta t)\big\|.$$

b) Sea $\boldsymbol{u}\in X$ tal que $\|\boldsymbol{u}\|=1$. Si $\{\boldsymbol{c}_k\}_{k\in\mathcal{L}}$ son los coeficientes de Fourier de $\boldsymbol{u}$, entonces sabemos que

$$\sum_{k\in\mathcal{L}}\|\boldsymbol{c}_k\|_2^2=\|\boldsymbol{u}\|^2=1,$$

y luego

$$\begin{aligned}\left\|\boldsymbol{C}^n(\Delta t)\boldsymbol{u}\right\|^2 &= \sum_{k\in\mathcal{L}}\left\|\boldsymbol{G}^n(\Delta t,k)\boldsymbol{c}_k\right\|_2^2\\ &\leqslant \sup_{k\in\mathcal{L}}\left\|\boldsymbol{G}^n(\Delta t,k)\right\|_2^2\sum_{k\in\mathcal{L}}\|\boldsymbol{c}_k\|_2^2\\ &= \sup_{k\in\mathcal{L}}\left\|\boldsymbol{G}^n(\Delta t,k)\right\|_2^2.\end{aligned}$$

Dado que $\boldsymbol{u}\in X$ con $\|\boldsymbol{u}\|=1$ fue arbitrario, podemos concluir que

$$\left\|\boldsymbol{C}^n(\Delta t)\right\| \leqslant \sup_{k\in\mathcal{L}}\left\|\boldsymbol{G}^n(\Delta t,k)\right\|_2.$$

Esto concluye la demostración del Teorema 5.5.

■

En virtud de los Teoremas 5.4 y 5.5, la convergencia del método de diferencias (5.68) es equivalente con la validez de (5.82). Veremos que esta condición resulta muy útil.

El siguiente teorema entrega un criterio de estabilidad necesario muy importante.

Teorema 5.6 (Condición de estabilidad de von Neumann)**.** *Una condición necesaria para la convergencia del método de diferencias* (5.68) *es la condición*

$$r_\sigma\big(\boldsymbol{G}(\Delta t,k)\big) \leqslant 1+\mathcal{O}(\Delta t) \quad \text{para } \Delta t\in(0,\tau],\ k\in\mathcal{L}. \tag{5.83}$$

Demostración. Sea (5.68) convergente. En virtud de los Teoremas 5.4 y 5.5 obtenemos la validez de (5.82). Sin pérdida de generalidad sea $K>1$, donde K es la constante que aparece en (5.82). Puesto que $r_\sigma^n(\boldsymbol{G}) = r_\sigma(\boldsymbol{G}^n) \leqslant \|\boldsymbol{G}^n\|_2$ tenemos que $r_\sigma(\boldsymbol{G}) \leqslant K^{1/n}$, y en particular

$$r_\sigma(\boldsymbol{G}) \leqslant K^{\Delta t/T} = (K^{1/T})^{\Delta t} =: f(\Delta t).$$

La función $f:[0,\tau]\to\mathbb{R}$ es estrictamente isotónica y convexa con $f(0)=1$. Entonces existe una constante $c>0$ tal que $f(\Delta t)\leqslant 1+c\Delta t$ para $\Delta t\in[0,\tau]$. ■

Comentamos que en muchos casos la condición (5.83) también es suficiente para la convergencia, por ejemplo si las matrices de amplificación $\boldsymbol{G}(\Delta t,k)$ son normales. En este caso se tiene $r_\sigma(\boldsymbol{G}) = \|\boldsymbol{G}\|_2$. Trivialmente, esto se cumple para $p=1$.

5.5.6. Análisis de estabilidad de algunos métodos de diferencias. En lo siguiente, analizaremos las propiedades de estabilidad y convergencia de los métodos de diferencias presentados en las Secciones 5.3 y 5.4. Aqui vamos a verificar la validez de (5.82) para cada uno se los métodos. Todos los métodos de diferencias examinados son consistentes con los problemas de valores iniciales correspondientes.

Ejemplo 5.2. *Se examina el método* (5.43) *de la Sección 5.3, el cual puede ser escrito en la forma*

$$\begin{pmatrix} v_{j,n+1}\\ w_{j,n+1}\end{pmatrix} = \left\{\boldsymbol{I}T^0 + \frac{\lambda}{2}\boldsymbol{A}(T^1-T^{-1})\right\}\begin{pmatrix} v_{j,n}\\ w_{j,n}\end{pmatrix}, \quad \lambda=\frac{\Delta t}{\Delta x}.$$

El operador $\{\dots\}$ *corresponde a* $\boldsymbol{C}(\Delta t)$ *en* (5.68)*. Definiendo* $\phi := k\Delta x$ *y aplicando la substitución* $T^\nu \mapsto \mathrm{e}^{\mathrm{i}\nu\phi}$ *obtenemos las matrices de amplificación*

$$\boldsymbol{G}(\Delta t, k) = \boldsymbol{I} + \mathrm{i}\lambda \sin\phi \boldsymbol{A} \in \mathbb{C}^{2\times 2}, \quad k \in \mathcal{L}, \quad \Delta t > 0. \tag{5.84}$$

Dado que según hipótesis, la matriz $\boldsymbol{A} \in \mathbb{R}^{2\times 2}$ *es simétrica, podemos chequear facilmente que* $\boldsymbol{G}\boldsymbol{G}^* = \boldsymbol{G}^*\boldsymbol{G}$*, lo que es equivalente con que la matriz* $\boldsymbol{G}$ *en* (5.84) *es normal, y la condición de von Neumann* (5.83) *es equivalente con la estabilidad* (5.82)*. Obtenemos el radio espectral*

$$r_\sigma(\boldsymbol{G}) = \left(1 + \lambda^2 \sin^2\phi r_\sigma^2(\boldsymbol{A})\right)^{1/2}.$$

En virtud de $r_\sigma(\boldsymbol{A}) > 0$*, esto implica que la condición de von Neumann* (5.83) *está satisfecha para todo* $\phi \in [0, 2\pi]$ *si y sólo si*

$$\exists \beta \geqslant 0 : \quad \frac{\Delta t}{\Delta x^2} \leqslant \beta. \tag{5.85}$$

Por lo tanto, (5.85) *es equivalente con la convergencia del método. Dado que la relación* (5.85) *es muy deventajosa, no se puede recomendar este método.*

Ejemplo 5.3. *Se examina el método* (5.44) *de la Sección 5.3, el cual puede ser escrito en la forma*

$$\begin{pmatrix} v_{j,n+1} \\ w_{j,n+1} \end{pmatrix} = \left\{ \frac{1}{2}\boldsymbol{I}(T + T^{-1}) + \frac{\lambda}{2}\boldsymbol{A}(T^1 - T^{-1}) \right\} \begin{pmatrix} v_{j,n} \\ w_{j,n} \end{pmatrix}, \quad \lambda = \frac{\Delta t}{\Delta x}.$$

Definiendo $\phi := k\Delta x$ *obtenemos las matrices de amplificación correpondientes*

$$\boldsymbol{G}(\Delta t, k) = \cos\phi \boldsymbol{I} + \mathrm{i}\lambda \sin\phi \boldsymbol{A} \in \mathbb{C}^{2\times 2}, \quad k \in \mathcal{L}, \quad \Delta t > 0.$$

Tal como en el Ejemplo 5.2 vemos que $\boldsymbol{G}$ *es normal. Si* $\mu_1, \mu_2 \in \mathbb{R}$ *son los valores propios de* $\boldsymbol{A}$*, entonces los valores propios* $\sigma_1, \sigma_2 \in \mathbb{C}$ *de* $\boldsymbol{G}$ *satisfacen*

$$|\sigma_j|^2 = \cos^2\phi + \lambda^2\mu_j^2 \sin^2\phi, \quad j = 1, 2,$$

lo que implica que la condición

$$\lambda \leqslant \frac{1}{r_\sigma(\boldsymbol{A})} \tag{5.86}$$

es equivalente con (5.83) *y por lo tanto con la convergencia del método, puesto* $\lambda > 0$ *se considera fijo. Para cualquier constante* λ *que no satisface* (5.86) *el método ya no converge.*

Ejemplo 5.4. *Se examina el método* (5.45) *de la Sección 5.3, el cual entrega la matrices de amplificación*

$$\boldsymbol{G}(\Delta t, k) = \begin{bmatrix} 1 & \mathrm{i}a \\ \mathrm{i}a & 1 - a^2 \end{bmatrix}, \quad a := 2c\lambda \sin\frac{\phi}{2}, \quad \phi := k\Delta x. \tag{5.87}$$

Aquí $\boldsymbol{G}$ no *es normal. La matriz* $\boldsymbol{G}$ *posee los valores propios*

$$\sigma_{1,2} = 1 - \frac{1}{2}a^2 \pm \mathrm{i}\frac{a}{2}\sqrt{4 - a^2}. \tag{5.88}$$

Según el Teorema de Schur existe una matrix unitaria $\boldsymbol{U}$ *tal que*

$$\boldsymbol{G} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^*, \quad \text{con } \boldsymbol{U}^* = \boldsymbol{U}^{-1} \in \mathbb{C}^{2\times 2},$$

donde $\mathbf{\Lambda}$ *es una matriz triangular superior. Aquí obtenemos*

$$\mathbf{\Lambda} = \begin{bmatrix} \sigma_1 & b \\ 0 & \sigma_2 \end{bmatrix}, \quad \boldsymbol{U} = \frac{1}{2}\sqrt{2}\begin{bmatrix} 1 & \mathrm{i}\dfrac{1-\bar{\sigma}_1}{a} \\ \mathrm{i}\dfrac{1-\sigma_1}{a} & 1 \end{bmatrix}, \quad b := a\left(-a\sqrt{4-a^2} + \mathrm{i}\left(a^2 - \frac{1}{2}\right)\right).$$

En virtud de $\|\boldsymbol{U}\|_2 = \|\boldsymbol{U}^*\|_2 = 1$ *sabemos que* $\|\boldsymbol{G}^n\|_2 = \|\mathbf{\Lambda}^n\|_2$ *para* $n \in \mathbb{N}$*, es decir podemos utilizar las matrices* $\mathbf{\Lambda} = \mathbf{\Lambda}(\Delta t, k)$ *para verificar* (5.82)*. Dado que para cada par de normas matriciales* $\|\cdot\|$ *y* $\|\cdot\|'$ *existen constantes* $m > 0$ *y* $M > 0$ *tales que*

$$\forall n \in \mathbb{N} : \forall \boldsymbol{A} \in \mathbb{C}^{n\times n} : \quad m\|\boldsymbol{A}\|' \leqslant \|\boldsymbol{A}\| \leqslant M\|\boldsymbol{A}\|',$$

sabemos que (5.82) *se satisface si y sólo si todos los elementos de* $\mathbf{\Lambda}^n$*,* $n \in \mathbb{N}_0$ *son uniformemente acotados. En nuestro caso,*

$$\mathbf{\Lambda}^n = \begin{bmatrix} \sigma_1^n & bp_n \\ 0 & \sigma_2^n \end{bmatrix}, \quad p_n = \sum_{\nu=0}^{n-1} \sigma_1^\nu \sigma_2^{n-\nu-1}, \quad n \in \mathbb{N}. \tag{5.89}$$

Supongamos ahora que

$$\lambda < \frac{1}{c}.$$

En este caso, en virtud de (5.87)*,*

$$a^2 < 4, \quad |\sigma_1| = |\sigma_2| = 1.$$

Además, sabemos que

$$|\sigma_1 - \sigma_2| \geqslant \delta|a|, \quad \delta := 2\sqrt{1-c^2\lambda^2} > 0.$$

Utilizando (5.89) *sabemos que*

$$p_n = \frac{\sigma_1^n - \sigma_2^n}{\sigma_1 - \sigma_2} \quad \text{si } a \neq 0.$$

Ademas chequeamos que $|b| \leqslant 4|a|$*, por lo tanto*

$$|bp_n| \leqslant \frac{8}{\delta}, \quad n \in \mathbb{N},$$

lo que significa que (5.82) *se satisface para* $\mathbf{\Lambda}(\Delta t, k)$*, y el método converge.*

Supongamos ahora que

$$\lambda \geqslant \frac{1}{c}.$$

En el caso $\lambda > 1/c$*, para algunos* $\phi = k\Delta x$ *la condición de von Neumann* (5.83) *ya no se cumple, y el método ya no converge. Para* $\lambda = 1/c$ *tenemos que* $a^2 = 4$ *para ciertos* ϕ *y debido a* (5.88) $\sigma_1 = \sigma_2 = -1$*. Utilizando* (5.89) *se tiene en este caso*

$$\mathbf{\Lambda}^n = (-1)^n \begin{bmatrix} 1 & -nb \\ 0 & 1 \end{bmatrix}, \quad n \in \mathbb{N}.$$

Puesto que $b \neq 0$ *(acordándonos que* $\boldsymbol{G}$ *no es normal) vemos que* (5.82) *no está satisfecha y por lo tanto el método no converge.*

Ejemplo 5.5. *Consideremos el método* (5.46) *de la Sección 5.3. Las matrices de amplificación para este método son*

$$\boldsymbol{G}(\Delta t,k)=\frac{1}{1+\dfrac{a^2}{4}}\begin{bmatrix}1-\dfrac{a^2}{4} & \mathrm{i}a\\ \mathrm{i}a & 1-\dfrac{a^2}{4}\end{bmatrix},\quad a:=2c\lambda\sin\frac{\phi}{2}.$$

Aquí se verifica de inmediato que $\boldsymbol{G}\boldsymbol{G}^*=\boldsymbol{G}^*\boldsymbol{G}=\boldsymbol{I}$, *lo que significa* $r_\sigma(\boldsymbol{G})=\|\boldsymbol{G}\|_2=1$ *para todo* $a\in\mathbb{R}$. *Entonces, la condición* (5.82) *está satisfecha para todo* $\lambda>0$, *lo que implica la convergencia del método para* todo *valor de* $\lambda=\Delta t/\Delta x>0$. *Tales métodos de diferencias se llaman* incondicionalmente estables.

Ejemplo 5.6. *Consideremos el método* (5.52) *de la Sección 5.4. Este método puede ser escrito como*

$$\begin{aligned}&\big\{\boldsymbol{I}-\alpha\lambda(T^{-1}-2I+T)\big\}\boldsymbol{u}_{j,n+1}\\&=\big\{\boldsymbol{I}+(1-\alpha)\lambda(T^{-1}-2I+T)\big\}\boldsymbol{u}_{j,n},\quad \alpha\in[0,1],\quad \lambda=\frac{\Delta t}{\Delta x^2}.\end{aligned}$$

Para las matrices de amplificación ($p=1$) *obtenemos*

$$G(\Delta t,k)=\frac{1-2(1-\alpha)\lambda(1-\cos\phi)}{1+2\alpha\lambda(1-\cos\phi)}\in\mathbb{R},\quad \phi=k\Delta x.$$

Ahora se verifica facilmente que la condición (5.59) *es equivalente con*

$$\big|G(\Delta t,k)\big|\leqslant 1\quad\text{para }\Delta t>0,\ k\in\mathcal{L}.$$

Pero esto significa la satisfacción de (5.82), *y el método converge. El método aún converge si la cota para* λ *en* (5.59) *se excede en una cantidad* $\mathcal{O}(\Delta t)$, *ya que la condición* (5.83) *aún está satisfecha en este caso- Sin embargo, para cualquier* λ *fijo tal que*

$$\lambda>\frac{1}{2(1-2\alpha)},\quad \alpha<\frac{1}{2},$$

el método diverge.

Ejemplo 5.7. *Consideremos el método* (5.62) *de la Sección 5.4. Primero transformamos este método de dos pasos en un método de paso simple. La substitución* $\boldsymbol{v}_{jn}:=\boldsymbol{u}_{j,n-1}$ *entrega*

$$\begin{pmatrix}\boldsymbol{u}_{j,n+1}\\ \boldsymbol{v}_{j,n+1}\end{pmatrix}=\begin{bmatrix}2\lambda(T^{-1}-2\boldsymbol{I}+T) & \boldsymbol{I}\\ \boldsymbol{I} & 0\end{bmatrix}\begin{pmatrix}\boldsymbol{u}_{jn}\\ \boldsymbol{v}_{jn}\end{pmatrix},\quad \lambda=\frac{\Delta t}{\Delta x^2}.$$

Las matrices de amplificación son

$$\boldsymbol{G}(\Delta t,k)=\begin{bmatrix}-4\lambda(1-\cos\phi) & 1\\ 1 & 0\end{bmatrix},\quad \phi=k\Delta x.\tag{5.90}$$

Para $\cos\phi\neq 1$, *los valores propios* σ_1 *y* σ_2 *de* (5.90) *satisfacen*

$$\sigma_1\sigma_2=-1,\quad \sigma_1+\sigma_2<0,$$

por lo tanto $r_\sigma(\boldsymbol{G})>1$ *para* $\cos\phi\neq 1$ *y* cada *valor fijo* $\lambda>0$. *Entonces, no está satisfecha la condición de Neumann* (5.83), *y vemos que el método es inutil.*

Ejemplo 5.8. *Consideremos el método* (5.63) *de la Sección 5.4. Tal como en el ejemplo anterior, transformamos el método a un método de paso simple. Aquí obtenemos las matrices de amplificación*

$$\boldsymbol{G}(\Delta t, k) = \begin{bmatrix} \dfrac{4\lambda\cos\phi}{1+2\lambda} & \dfrac{1-2\lambda}{1+2\lambda} \\ 1 & 0 \end{bmatrix}, \quad \lambda = \frac{\Delta t}{\Delta x^2}, \quad \phi = k\Delta x.$$

Los valores propios de (5.66) *son*

$$\sigma_{1,2} = \frac{2\lambda\cos\phi \pm \sqrt{1-4\lambda^2\sin^2\phi}}{1+2\lambda}, \tag{5.91}$$

lo que implica que $|\sigma_1| \leqslant 1$, $|\sigma_2| \leqslant 1$ *para todo* $\lambda > 0$ *y* $\phi \in \mathbb{R}$.

Podemos proceder como en el Ejemplo 5.4 analizando las potencias de la matriz triangular $\boldsymbol{\Lambda}(\Delta, k)$ *(ver* (5.89)*). Para* $\lambda \leqslant 1/2$ *podemos concluir (como en el Ejemplo 5.4) que* (5.82) *es válido. Nos queda para analizar el caso* $\lambda > 1/2$. *Aqui distinguimos dos casos de los valores propios.*

a) *Si* $\sigma_1, \sigma_2 \in \mathbb{R}$, *un pequeño cálculo (utilizando* (5.91)*) verifica que*

$$\min\big\{|\sigma_1|, |\sigma_2|\big\} \leqslant \frac{2\lambda}{1+2\lambda} < 1,$$

lo que implica

$$|p_n| \leqslant 1 + 2\lambda, \quad n \in \mathbb{N}.$$

Esto nos permite concluir que (5.82) *queda válido para* $\boldsymbol{\Lambda}(\Delta t, k)$.

b) *Si* $\sigma_1, \sigma_2 \notin \mathbb{R}$, *podemos calcular (utilizando* (5.91)*) que*

$$|\sigma_1^2| = |\sigma_2^2| = \frac{4\lambda^2 - 1}{4\lambda^2 + 4\lambda + 1} =: \gamma^2 < 1.$$

Esto inmediatamente entrega que

$$|p_n| \leqslant n\gamma^{n-1}, \quad n \in \mathbb{N},$$

lo que significa que también en este caso las potencias en (5.89) *estan uniformemente acotadas.*

Cocluimos que el método converge para todo $\lambda > 0$, *es decir, el método es incondicionalmente estable.*

Los resultados de convergencia son válidos en la norma L^2, dado que esta norma aparece en la transformación de Fourier. También existen análises de convergencia en otras normas. Por supuesto, toda la teoría puede ser extendidad al caso de varias variables espaciales.

Capítulo 6

Introducción al Método de Elementos Finitos

6.1. Problemas de valores de frontera de ecuaciones diferenciales ordinarias

Ya comentamos anteriormente que la ecuación diferencial ordinaria de segundo orden autoadjunta

$$Ly \equiv -\frac{\mathrm{d}}{\mathrm{d}x}\left(p(x)\frac{\mathrm{d}y}{\mathrm{d}x}\right) + q(x)y = g(x) \tag{6.1}$$

es la ecuación diferencial de Euler del problema variacional

$$I[u] = \frac{1}{2}\int_a^b \left(p(x)(u')^2 + q(x)u^2 - 2g(x)u\right)\mathrm{d}x \stackrel{!}{=} \text{mín}\,. \tag{6.2}$$

Cada función $y = y(x)$, $y \in C^2[a,b]$, que minimiza $I[u]$, es decir, que satisface

$$\forall u \in C^2[a,b]: \quad I[y] \leqslant I[u],$$

necesariamente es una solución de la ecuación diferencial ordinaria (6.1). Por otro lado, bajo ciertas hipótesis una solución $y = y(x)$, $y \in C^2[a,b]$, de (6.1) también es una solución del problema variacional $I[u] \stackrel{!}{=} \text{mín}$. Es la equivalencia entre ambos problemas que forma el fundamento de los métodos que vamos a discutir ahora, y que se llaman *métodos variacionales*. Basicamente la idea de estos métodos es la solución aproximada del problema variacional, pero donde uno no se limita a funciones en $C^2[a,b]$. La solución aproximada del problema variacional también es considerada una solución de la ecuación diferencial.

Consideremos la ecuación (6.1) junto con las condiciones de borde

$$y(a) = \alpha, \quad y(b) = \beta. \tag{6.3}$$

Las condiciones se convierten en condiciones homogéneas si definimos

$$z(x) := y(x) - Q(x), \quad Q(x) := \alpha\frac{b-x}{b-a} - \beta\frac{a-x}{b-a},$$

de manera que

$$Q(a) = \alpha, \quad Q(b) = \beta, \quad \frac{dQ}{\mathrm{d}x} = -\frac{\alpha-\beta}{b-a}, \quad LQ = \frac{\alpha-\beta}{b-a}p'(x) + q(x)Q(x) =: f(x),$$

y definiendo $h(x) := g(x) - f(x)$, podemos escribir el problema de valores de frontera como

$$(Lz)(x) = h(x), \quad z(a) = z(b) = 0.$$

Entonces, en general podemos limitarnos a estudiar el problema de valores de frontera semihomogéneo

$$Ly \equiv -\frac{\mathrm{d}}{\mathrm{d}x}\left(p(x)\frac{\mathrm{d}y}{\mathrm{d}x}\right) + q(x)y = g(x), \quad y(a) = y(b) = 0. \tag{6.4}$$

Aquí se supone que

$$p \in C^1[a,b], \quad q, g \in C^1[a,b], \quad p(x) \geqslant p_0 > 0, \quad q(x) \geqslant 0, \quad x \in [a,b]. \tag{6.5}$$

Se puede demostrar que en este caso, el problema (6.4) posee una única solución $y \in C^2[a,b]$.

El dominio $\mathcal{D}$ del operador diferencial L es el conjunto de todas las funciones en $C^2[a.b]$ que desaparecen en a y b:

$$\mathcal{D} := \big\{u \in C^2[a,b] \mid u(a) = u(b) = 0\big\}.$$

El problema de valores de frontera es equivalente al problema de encontrar una solución de

$$Lu = g, \quad u \in \mathcal{D}. \tag{6.6}$$

Bajo las hipótesis (6.5) existe una única solución de este problema.

Ahora definimos el producto escalar

$$(u,v) := \int_a^b u(x)v(x)\,\mathrm{d}x, \quad u, v \in L^2[a,b]$$

y la norma

$$\|u\|_2 := (u,u)^{1/2}.$$

Un operador T con el dominio (de definición) D_T y la propiedad

$$\forall u, v \in D_T: \quad (u, Tv) = (Tu, v)$$

se llama *simétrico* o *autoadjunto* sobre D_T. Además, este operador se llama *definido positivo* si

$$\forall u \in D_T: \quad u \not\equiv 0 \Longrightarrow (Tu, u) > 0.$$

Teorema 6.1. *El operador L es simétrico y definido positivo sobre $\mathcal{D}$.*

Demostración. Dado que $Lv = -(pv')' + qv$, obtenemos

$$\begin{aligned}(u, Lv) &= \int_a^b u(x)\Big(-\big(p(x)v'(x)\big)' + q(x)v(x)\Big)\,\mathrm{d}x \\ &= -u(x)p(x)v'(x)\Big|_a^b + \int_a^b \big(p(x)u'(x)v'(x) + q(x)u(x)v(x)\big)\,\mathrm{d}x.\end{aligned} \tag{6.7}$$

El primero término en el lado derecho de (6.7) desaparece ya que $u \in \mathcal{D}$, y el segundo es simétrico con respecto a u y v, es decir,

$$(u, Lv) = (v, Lu) = (Lu, v).$$

Entonces, L es simétrico. Ahora, en virtud de (6.2) resulta para $u|_{[a,b]} \not\equiv 0$

$$(Lu, u) = \int_a^b \Big(p(x)\big(u'(x)\big)^2 + q(x)\big(u(x)\big)^2\Big)\,\mathrm{d}x$$

$$\geqslant p_0 \int_a^b \big(u'(x)\big)^2 \,\mathrm{d}x + \int_a^b q(x)\big(u(x)\big)^2 \,\mathrm{d}x$$
$$\geqslant \frac{p_0}{(b-a)^2} \int_a^b \big(u(x)\big)^2 \,\mathrm{d}x > 0.$$

Aquí usamos la desigualdad de Cauchy-Schwarz para integrales

$$\big(u(x)\big)^2 = \left(\int_a^x 1 \cdot u'(\xi)\,\mathrm{d}\xi\right)^2 \leqslant \int_a^x 1^2 \,\mathrm{d}\xi \int_a^x \big(u'(\xi)\big)^2 \,\mathrm{d}\xi \leqslant (b-a)\int_a^b \big(u'(\xi)\big)^2 \,\mathrm{d}\xi,$$

la cual a su vez implica

$$\int_a^b \big(u(x)\big)^2 \,\mathrm{d}x \leqslant \int_a^b \left((b-a)\int_a^b \big(u'(\xi)\big)^2 \,\mathrm{d}\xi\right)\mathrm{d}x = (b-a)^2 \int_a^b \big(u'(\xi)\big)^2 \,\mathrm{d}\xi.$$

■

Ahora, (6.7) implica

$$\forall u, v \in \mathcal{D}: \quad (u, Lv) = (Lu, v) = \int_a^b \big(p(x)u'(x)v'(x) + q(x)u(x)v(x)\big)\,\mathrm{d}x.$$

Pero la integral no existe sólo para $u, v \in \mathcal{D}$, sino que también (por ejemplo) para funciones diferenciables por trozos cuyas primeras derivadas son cuadraticamente integrables. En general, consideraremos el espacio de funciones $V^r[a,b)$, donde $w \in V^r(a,b)$ si y sólo si se cumplen las siguientes condiciones:

a) $w \in C^{r-1}[a,b]$.
b) Con la excepción de un número finito o a lo más contable de puntos, $w^{(r-1)}$ es diferenciable en $[a,b]$.
c) La función $w^{(r)}(x)$ es cuadraticamente integrable sobre $[a,b]$, donde $r \in \mathbb{N} \cup \{0\}$.
d) Además, se requiere que

$$\|w\|_{V^r(a,b)} := \left(\int_a^b \sum_{\varrho=0}^{r} \big|w^{(\varrho)}(x)\big|^2 \,\mathrm{d}x\right)^{1/2} < \infty. \tag{6.8}$$

Obviamente, para $r = 0$, (a) y (b) no hacen sentido, y (c) exige que

$$\|w\|_{V^0(a,b)} = \left(\int_a^b |w(x)|^2 \,\mathrm{d}x\right)^{1/2} = \|w\|_2 < \infty,$$

es decir, $V^0(a,b) = L^2(a,b)$.

Ejemplo 6.1. *Subdividimos el intervalo $[a,b]$ en n sub-intervalos del tamaño h y definimos los nodos $x_i = a + ih$ para $i = 0, \ldots, n$, $a + nh = b$. En este caso, los trazados poligonales*

$$s_\Delta(x) := \frac{f_{i+1} - f_i}{x_{i+1} - x_i}(x - x_i) + f_i = \frac{f_{i+1} - f_i}{h}(x - x_i) + f_i, \quad x \in [x_i, x_{i+1}]$$

son elementos de $V^1(a,b)$*, puesto que* $s_\Delta \in C^0[a,b]$ *y* $s_\Delta(x)$ *es diferenciable en todas partes con la excepción de los nodos* $x_1, \ldots, x_{n-1}$*. Finalmente, para* $f \in C^1[a,b]$ *resulta*

$$s'_\Delta(x) = \frac{f_{i+1} - f_i}{h} = f'(\xi_i), \quad \xi_i \in [x_i, x_{i+1}],$$

lo que nos permite concluir que

$$\int_a^b \Big(\big(s_\Delta(x)\big)^2 + \big(s'_\Delta(x)\big)^2 \Big)\, \mathrm{d}x = \sum_{i=0}^{n-1} \int_{x_i}^{x_{i+1}} \Big(\big(s_\Delta(x)\big)^2 + \big(s'_\Delta(x)\big)^2 \Big)\, \mathrm{d}x < \infty.$$

Ahora sea

$$D := \big\{ w \in V^1(a,b) \mid w(a) = w(b) = 0 \big\},$$

y definimos la forma bilineal simétrica

$$[u,v] := \int_a^b \big(p(x)u'(x)v'(x) + q(x)u(x)v(x) \big)\, \mathrm{d}x, \quad u, v \in D. \tag{6.9}$$

Obviamente, $\mathcal{D} \subset D$. Para $u, v \in \mathcal{D} \subset D$ tenemos $[u,v] = (Lu, v)$. Si $g(x)$ denota la parte derecha de la ecuación diferencial (6.1), la siguiente expresión es bien definida para cada $u \in D$:

$$\begin{aligned} I[u] &= [u,u] - 2(u,g) \\ &= \int_a^b \Big(p(x)\big(u'(x)\big)^2 + q(x)\big(u(x)\big)^2 - 2u(x)g(x) \Big)\, \mathrm{d}x. \end{aligned} \tag{6.10}$$

Ahora demostramos que $I[u]$ asume su mínimo para la solución del problema de valores de frontera.

Teorema 6.2. *Sea* y *la solución de* (6.1) *o* (6.3)*, respectivamente. Entonces*

$$\forall u \in D, u \neq y: \quad I[y] < I[u].$$

Demostración. Dado que $Ly = g$ e $y \in \mathcal{D}$, sabemos que

$$(u,g) = (u, Ly) = [u,y],$$

y según (6.10),

$$I[y] = [y,y] - 2(y,g) = (y,Ly) - 2(y,Ly) = -(y,Ly) = -[y,y].$$

La simetría de $[\cdot,\cdot]$ implica que

$$[u-y, u-y] = [u,u] - 2[u,y] + [y,y].$$

Entonces podemos expresar $I[u]$ como

$$\begin{aligned} I[u] &= [u,u] - 2(u,g) \\ &= [u,u] - 2(u,Ly) \\ &= [u,u] - 2[u,y] + [y,y] - [y,y] \\ &= [u-y, u-y] - [y,y]. \end{aligned}$$

Pero en virtud de (6.9),

$$\forall u \in D, u \neq y: \quad [u-y, u-y] > 0,$$

tal que

$$I[u] = [u-y, u-y] - [y,y] > -[y,y] = I[y].$$

■

Si tenemos solamente $p(x), q(x) \geqslant 0$, entonces $(Lu, u) \geqslant 0$ para $u \in \mathcal{D}$ y $[u,u] \geqslant 0$ para $u \in D$. Si (6.1) o (6.3) tiene la solución y,

$$\forall u \in \mathcal{D}: \quad I[y] \leqslant I[u],$$

pero la función y posiblemente no es única.

Según el Teorema 6.2, el problema de valores de frontera (6.6) es resuelto si se ha encontrado la función y para la cual $I[u]$ asume su mínimo. En el *método de Ritz* (1908), se minimiza $I[u]$ de forma aproximada. Se considera un sub-espacio M-dimesnsional $D_M \subset D$, por ejemplo generado por las funciones $\varphi_1, \ldots, \varphi_M$. Puesto que $D_M \subset D$, estas funciones deben satisfacer $\varphi_j(a) = \varphi_j(b) = 0$, y toda función $v \in D_M$ puede ser representada como combinación lineal

$$v = \sum_{\mu=1}^{M} \alpha_\mu \varphi_\mu, \quad \alpha_1, \ldots, \alpha_M \in \mathbb{R}.$$

Esta función v satisface

$$\begin{aligned} I[v] &= [v,v] - 2(v,g) \\ &= \left[\sum_{\mu=1}^{M} \alpha_\mu \varphi_\mu, \sum_{\nu=1}^{M} \alpha_\nu \varphi_\nu\right] - 2\left(\sum_{\mu=1}^{M} \alpha_\mu \varphi_\mu, g\right) \\ &= \sum_{\mu,\nu=1}^{M} \alpha_\mu \alpha_\nu [\varphi_\mu, \varphi_\nu] - 2\sum_{\mu=1}^{M} \alpha_\mu (\varphi_\mu, g) \\ &=: \Phi(\alpha_1, \ldots, \alpha_M). \end{aligned} \tag{6.11}$$

Los números $\alpha_1, \ldots, \alpha_M$ se determinan de tal forma que $\Phi(\alpha_1, \ldots, \alpha_M)$ asume su mínimo. Una condición necesario para que eso ocurra es la condición

$$\frac{1}{2}\frac{\partial \Phi(\alpha_1, \ldots, \alpha_M)}{\partial \alpha_i} = \frac{1}{2}\sum_{\mu,\nu=1}^{M} [\varphi_\mu, \varphi_\nu]\left(\frac{\partial \alpha_\mu}{\partial \alpha_i}\alpha_\nu + \alpha_\mu \frac{\partial \alpha_\nu}{\partial \alpha_i}\right) - (\varphi_i, g) = 0.$$

Dado que

$$\frac{\partial \alpha_k}{\partial \alpha_i} = \delta_{ik} = \begin{cases} 1 & \text{si } i = k, \\ 0 & \text{sino,} \end{cases}$$

podemos escribir

$$\frac{1}{2}\sum_{\nu=1}^{M} [\varphi_i, \varphi_\nu]\alpha_\nu + \frac{1}{2}\sum_{\mu=1}^{M} [\varphi_\mu, \varphi_i]\alpha_\mu = \sum_{j=1}^{M} [\varphi_i, \varphi_j]\alpha_j,$$

por lo tanto los coeficientes $\alpha_1, \ldots, \alpha_M$ buscados deben satisfacer el sistema lineal

$$\frac{1}{2}\frac{\partial \Phi(\alpha_1, \ldots, \alpha_M)}{\partial \alpha_i} = \sum_{j=1}^{M} [\varphi_i, \varphi_j]\alpha_j - (\varphi_i, g) = 0, \quad i = 1, \ldots, M.$$

Definiendo

$$a_{ij} := [\varphi_i, \varphi_j], \quad b_i := (\varphi_i, g), \tag{6.12}$$

$$\boldsymbol{A} := \begin{bmatrix} a_{11} & \cdots & a_{1M} \\ \vdots & & \vdots \\ a_{M1} & \cdots & a_{MM} \end{bmatrix}, \quad \boldsymbol{b} := \begin{pmatrix} b_1 \\ \vdots \\ b_M \end{pmatrix}, \quad \boldsymbol{\alpha} := \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_M \end{pmatrix},$$

obtenemos el sistema

$$\boldsymbol{A}\boldsymbol{\alpha} = \boldsymbol{b}. \tag{6.13}$$

La matriz $\boldsymbol{A}$ es simétrica. Para ver que también es definida positiva, supongamos que $\boldsymbol{k} = (k_1, \ldots, k_M)^{\mathrm{T}} \neq 0$. Entonces

$$\sum_{i=1}^{M} k_i \varphi_i =: w \in D_M, \quad w \not\equiv 0,$$

y calculamos que

$$0 < [w, w] = \sum_{i,j=1}^{M} [\varphi_i, \varphi_j] k_i k_j = \sum_{i,j=1}^{M} a_{ij} k_i k_j = \boldsymbol{k}^{\mathrm{T}} \boldsymbol{A} \boldsymbol{k},$$

es decir, $\boldsymbol{A}$ es definida positiva; por ello concluimos que el sistema lineal posee una solución única $\boldsymbol{\alpha}^* = (\alpha_1^*, \ldots, \alpha_M^*)^{\mathrm{T}}$. Ahora, según (6.11) y (6.12),

$$\Phi(\boldsymbol{\alpha}) = \boldsymbol{\alpha}^{\mathrm{T}} \boldsymbol{A} \boldsymbol{\alpha} - 2\boldsymbol{\alpha}^{\mathrm{T}} \boldsymbol{b}, \tag{6.14}$$

y por lo tanto

$$\begin{aligned} (\boldsymbol{\alpha} - \boldsymbol{\alpha}^*)^{\mathrm{T}} \boldsymbol{A} (\boldsymbol{\alpha} - \boldsymbol{\alpha}^*) &= \boldsymbol{\alpha}^{\mathrm{T}} \boldsymbol{A} \boldsymbol{\alpha} - 2\boldsymbol{\alpha}^{\mathrm{T}} \boldsymbol{A} \boldsymbol{\alpha}^* + (\boldsymbol{\alpha}^*)^{\mathrm{T}} \boldsymbol{A} \boldsymbol{\alpha}^* \\ &= \boldsymbol{\alpha}^{\mathrm{T}} \boldsymbol{A} \boldsymbol{\alpha} - 2\boldsymbol{\alpha}^{\mathrm{T}} \boldsymbol{b} + (\boldsymbol{\alpha}^*)^{\mathrm{T}} \boldsymbol{A} \boldsymbol{\alpha}^* \\ &= \Phi(\boldsymbol{\alpha}) + (\boldsymbol{\alpha}^*)^{\mathrm{T}} \boldsymbol{A} \boldsymbol{\alpha}^*. \end{aligned} \tag{6.15}$$

De (6.14) obtenemos con $\boldsymbol{\alpha} = \boldsymbol{\alpha}^*$ y $\boldsymbol{A}\boldsymbol{\alpha}^* = \boldsymbol{b}$:

$$\Phi(\boldsymbol{\alpha}^*) = (\boldsymbol{\alpha}^*)^{\mathrm{T}} \boldsymbol{A} \boldsymbol{\alpha}^* - 2(\boldsymbol{\alpha}^*)^{\mathrm{T}} \boldsymbol{A} \boldsymbol{\alpha}^* = -(\boldsymbol{\alpha}^*)^{\mathrm{T}} \boldsymbol{A} \boldsymbol{\alpha}^*.$$

Por otro lado, como $\boldsymbol{A}$ es definida positiva, para $\boldsymbol{\alpha} \neq \boldsymbol{\alpha}^*$ sabemos que

$$(\boldsymbol{\alpha} - \boldsymbol{\alpha}^*)^{\mathrm{T}} \boldsymbol{A} (\boldsymbol{\alpha} - \boldsymbol{\alpha}^*) > 0.$$

Finalmente, en virtud de lo anterior, (6.15) implica que

$$0 < \Phi(\boldsymbol{\alpha}) - \Phi(\boldsymbol{\alpha}^*), \quad \boldsymbol{\alpha} \neq \boldsymbol{\alpha}^*.$$

Eso significa que la función buscada de D_M, para la cual $I[v]$ asume su mínimo sobre D_M, es

$$v^*(x) = \sum_{\mu=1}^{M} \alpha_\mu^* \varphi_\mu(x),$$

donde

$$\min_{v \in D_M} I[v] = I[v^*].$$

Si por casualidad entre las funciones $\varphi_1, \ldots, \varphi_M$ se encuentra la solución del problema de valores de frontera, o sea si $y \equiv \varphi_k$ para algún índice k, la solución del sistema lineal (6.13) es

$$\alpha_i = \begin{cases} 0 & \text{para } i = 1, \ldots, M,\ i \neq k, \\ 1 & \text{para } i = k, \end{cases}$$

es decir, $v^*(x) = \varphi_k(x) = y(x)$. Esto es una consecuencia directa del Teorema 6.2, tomando en cuenta que

$$I[v] = \Phi(\boldsymbol{\alpha}) \geqslant \Phi(\boldsymbol{\alpha}^*) = I[v^*] \geqslant I[y].$$

Podemos resumir que el método de Ritz consiste en la computación del vector $\boldsymbol{\alpha}^*$ del sistema $\boldsymbol{A}\boldsymbol{\alpha} = \boldsymbol{b}$, donde $\boldsymbol{A} = ([\varphi_i, \varphi_j])_{i,j=1,\ldots,M}$, $b_i = (\varphi_i, g)$. La combinación

$$v^* = \sum_{\mu=1}^{M} \alpha_\mu^* \varphi_\mu$$

es la función deseada, la cual minimiza el funcional $I[u]$ sobre D.

Para discutir algunos aspectos prácticos, recordamos que las componentes de $\boldsymbol{A}$ y $\boldsymbol{b}$ son dadas por

$$a_{ij} = [\varphi_i, \varphi_j] = \int_a^b \big(p(x)\varphi_i'(x)\varphi_j'(x) + q(x)\varphi_i(x)\varphi_j(x)\big)\,\mathrm{d}x,$$
$$b_i = (\varphi_i, g) = \int_a^b \varphi_i(x) g(x)\,\mathrm{d}x.$$

El problema de si estas integrales pueden ser calculadas exactamente depende de las funciones de base $\varphi_1, \ldots, \varphi_M$. En general, tendremos que utilizar un método de cuadratura (integración numérica), lo cual causa un error adicional. La selección de las funciones $\varphi_1, \ldots, \varphi_M$ frecuentamente es una tarea bastante dificil, la cual requiere de cierta experiencia. Depende del problema puesto y de la precisión deseada. Por ejemplo, se usarán funciones de base periódicas si se sabe que la solución del problema de valores de frontera es periódica. Por otro lado, cuando no hay ninguna información acerca de la solución, frecuentamente se usan polinomios, por ejemplo polinomios ortogonales. Las dificultades mencionadas aquí se pueden evitar parcialmente si usamos polinomios definidos por trozos, por ejemplo funciones *spline*. Estas consideraciones nos llevan al *método de elementos finitos*.

Ejemplo 6.2. *Consideremos el problema de valores de frontera*

$$-y'' = \sin x, \quad y(0) = y(\pi) = 0,$$

el cual puede ser resuelto exactamente. Aquí $p \equiv 1$, $q \equiv 0$, *y* $g(x) = \sin x$.

a) *Sea* $M = 1$ *y* $\varphi_1(x) = x(\pi - x)$, $\varphi_1'(x) = \pi - 2x$. *En este caso,*

$$\begin{aligned} a_{11} &= \int_0^\pi \big(\varphi_1'(x)\big)^2 \,\mathrm{d}x = \int_0^\pi (\pi^2 - 4\pi x + 4x^2)\,\mathrm{d}x = \frac{\pi^3}{3}, \\ b_1 &= \int_0^\pi \varphi_1(x) g(x)\,\mathrm{d}x = \int_0^\pi x(\pi - x)\sin x\,\mathrm{d}x \\ &= \pi \int_0^\pi x \sin x\,\mathrm{d}x - \int_0^\pi x^2 \sin x\,\mathrm{d}x \\ &= \pi \Big[\sin x - x\cos x\Big]_0^\pi - \Big[2x\sin x + (2 - x^2)\cos x\Big]_0^\pi \\ &= \pi^2 + (2 - \pi^2) + 2 = 4. \end{aligned}$$

Entonces, $\alpha_1 = 12/\pi^3$ *y*

$$v^*(x) = \alpha_1 \varphi_1(x) = \frac{12}{\pi^3} x(\pi - x) \approx 0{,}387 x(\pi - x).$$

La función $v^*(x)$ *asume su máximo en* $x = \pi/2$ *con*

$$v * \left(\frac{\pi}{2}\right) = \frac{3}{\pi} \approx 0{,}955.$$

b) *Sea* $M = 2$, $\varphi_1(x) = \sin x$, $\varphi_1'(x) = \cos x$, $\varphi_2(x) = \sin(2x)$, $\varphi_2'(x) = 2\cos(2x)$. *Tenemos*

$$\begin{aligned} a_{11} &= \int_0^\pi \big(\varphi_1'(x)\big)^2 \,\mathrm{d}x = \int_0^\pi \cos^2 x\,\mathrm{d}x = \frac{\pi}{2}, \\ a_{12} &= \int_0^\pi \varphi_1'(x)\varphi_2'(x)\,\mathrm{d}x = 2\int_0^\pi \cos x \cos(2x)\,\mathrm{d}x = 0, \\ a_{22} &= \int_0^\pi \big(\varphi_2'(x)\big)^2 \,\mathrm{d}x = 4\int_0^\pi \cos^2(2x)\,\mathrm{d}x = 2\pi, \\ b_1 &= \int_0^\pi \varphi_1(x) g(x)\,\mathrm{d}x = \int_0^\pi \sin^2 x\,\mathrm{d}x = \frac{\pi}{2}, \\ b_2 &= \int_0^\pi \varphi_2(x) g(x)\,\mathrm{d}x = \int_0^\pi \sin(2x)\sin x\,\mathrm{d}x = 0. \end{aligned}$$

Entonces

$$\boldsymbol{A} = \begin{bmatrix} \pi/2 & 0 \\ 0 & 2\pi \end{bmatrix}; \quad \alpha_1 = 1, \quad \alpha_2 = 0; \quad v^*(x) = \sin x,$$

lo que es la solución exacta del problema.

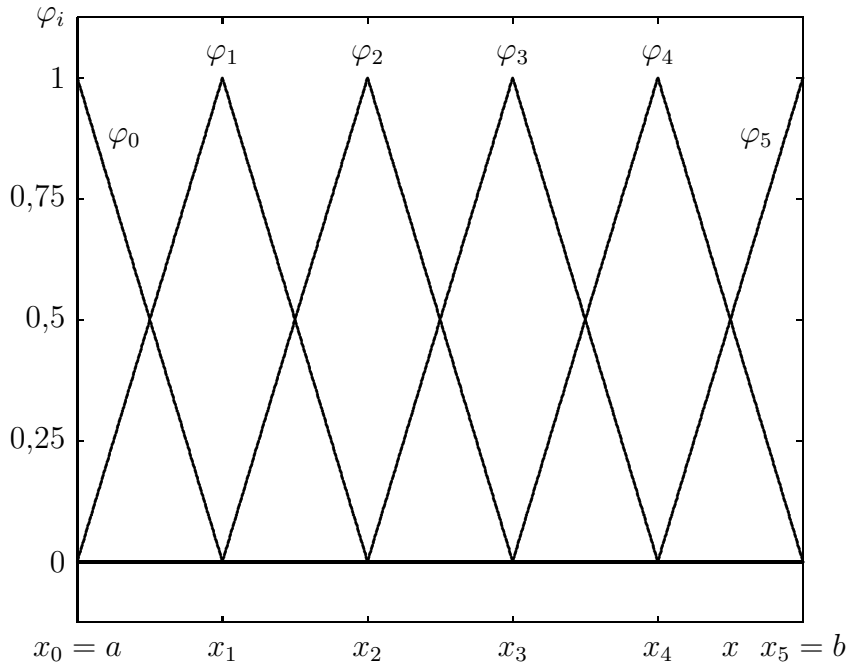


FIGURA 6.1. Las funciones φ_i, $i = 0, \ldots, n$, dadas por (6.17) para $n = 5$.

6.2. Elementos finitos para problemas de valores de frontera de ecuaciones diferenciales ordinarias

6.2.1. Funciones de planteo lineales por trozos. Ahora seguimos desarrollando el método de Ritz en una versión donde cada función φ_i es diferente de cero sólo en un (pequeño) subintervalo de $[a, b]$. Como para la interpolación por splines, subdividimos el intervalo en n partes de igual tamaño h introduciendo los nodos

$$x_i = a + ih, \quad i = 0, \ldots, n; \quad a + nh = b. \tag{6.16}$$

Consideremos primero funciones de planteo lineales por trozos y continuas s_h. El espacio S_h de estas funciones tiene como base las funciones $\varphi_0, \ldots, \varphi_n$ definidas por

$$\varphi_i(x) = \begin{cases} \dfrac{x-a}{h} - i + 1 & \text{si } x_{i-1} \leqslant x \leqslant x_i, \\ -\dfrac{x-a}{h} + i + 1 & \text{si } x_i \leqslant x \leqslant x_{i+1}, \\ 0 & \text{sino,} \end{cases} \quad i = 1, \ldots, n-1,$$

$$\varphi_0(x) = \begin{cases} -\dfrac{x-a}{h} + 1 & \text{si } x_0 \leqslant x \leqslant x_1, \\ 0 & \text{sino,} \end{cases} \tag{6.17}$$

$$\varphi_n(x) = \begin{cases} \dfrac{x-a}{h} - n + 1 & \text{si } x_{n-1} \leqslant x \leqslant x_n, \\ 0 & \text{sino.} \end{cases}$$

Cada función de S_h tiene la representación

$$s_h(x) = \sum_{i=0}^{n} \alpha_i \varphi_i(x).$$

Para la aproximación de la solución de (6.4), usaremos solamente aquellas funciones en S_h que satisfacen las condiciones de borde $s_h(a) = s_h(b) = 0$, es decir, para las cuales

$$s_h(a) = \sum_{i=0}^{n} \alpha_i \varphi_i(a) = \alpha_0 \varphi_0(a) = 0, \quad s_h(b) = \sum_{i=0}^{n} \alpha_i \varphi_i(b) = \alpha_n \varphi_n(b) = 0.$$

Dado que $\varphi_0(a) = \varphi_n(b) = 1$, este requerimiento implica que $\alpha_0 = \alpha_n = 0$, o sea podemos considerar el espacio de las funciones $s_h^0(x)$ que pueden ser representadas como

$$s_h^0(x) = \sum_{i=1}^{n-1} \alpha_i \varphi_i(x). \tag{6.18}$$

Cada función s_h^0 es diferenciable en $[a, b]$ en todas partes, con la excepción de los nodos $x_1, \ldots . x_{n-1}$, e integrable cuadraticamente junto con su derivada. Si llamamos $P_h^1(a, b)$ al espacio de todas las funciones (6.18), observamos que $P_h^1(a, b) \subset D$, o sea las funciones (6.18) sirven de planteo para el método de Ritz.

Ya sabemos que los elementos de la matriz $\boldsymbol{A}$ se calculan de la siguiente forma:

$$\begin{aligned} a_{ij} = [\varphi_i, \varphi_j] &= \int_a^b \big(p(x)\varphi_i'(x)\varphi_j'(x) + q(x)\varphi_i(x)\varphi_j(x)\big)\,\mathrm{d}x \\ &= \sum_{k=1}^{n} \int_{x_{k-1}}^{x_k} \big(p(x)\varphi_i'(x)\varphi_j'(x) + q(x)\varphi_i(x)\varphi_j(x)\big)\,\mathrm{d}x. \end{aligned}$$

Dado que

$$\varphi_j(x) = 0 \quad \text{para } j \notin \{i-1, i, i+1\} \text{ y } x \in [x_{i-1}, x_{i+1}],$$

sabemos que $[\varphi_i, \varphi_j] = 0$ para $|i - j| \geqslant 2$, es decir,

$$\boldsymbol{A} = \begin{bmatrix} a_{11} & a_{21} & 0 & \cdots & 0 \\ a_{21} & a_{22} & a_{23} & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & a_{n-1,n-2} & a_{n-1,n-1} & a_{n,n-1} \\ 0 & \cdots & 0 & a_{n,n-1} & a_{nn} \end{bmatrix}, \quad b_i = (\varphi_i, g) = \int_{x_{i-1}}^{x_{i+1}} \varphi_i(x) g(x)\,\mathrm{d}x.$$

Ahora, para $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_{n-1})^{\mathrm{T}}$ y $\boldsymbol{b} = (b_1, \ldots, b_{n-1})^{\mathrm{T}}$ determinamos $\boldsymbol{\alpha}^* = (\alpha_1^*, \ldots, \alpha_{n-1}^*)^{\mathrm{T}}$ como solución del sistema lineal

$$\boldsymbol{A}\boldsymbol{\alpha} = \boldsymbol{b}. \tag{6.19}$$

La solución del sistema (6.19) entrega la solución aproximada del problema de valores de frontera (6.4):

$$s_h^*(x) = \sum_{i=1}^{n-1} \alpha_i^* \varphi_i(x).$$

Ya demostramos que la matriz $\boldsymbol{A}$ siempre es simétrica y definida positiva. Para que una función s_h^* sea una aproximación suficientemente exacta de la verdadera solución y, el tamaño de paso h debe ser muy pequeño. Normalmente, (6.19) es un sistema de gran tamaño, con una matriz que siempre es simétrica, definida positiva y tridiagonal, por lo tanto, el sistema (6.19) puede ser resuelto por muchos métodos iterativos.

Ejemplo 6.3. *Consideremos nuevamente el ejemplo de la ecuación diferencial ordinaria* $-y'' = \sin x$ *con las condiciones de borde* $y(0) = y(\pi) = 0$. *Para* $i = 1, \ldots, n$, *obtenemos*

$$a_{i,i-1} = \int_0^\pi \varphi_i'(x)\varphi_{i-1}'(x)\,\mathrm{d}x = \int_{x_{i-1}}^{x_i} \varphi_i'(x)\varphi_{i-1}'(x)\,\mathrm{d}x,$$

dado que fuera de $[x_{i-1}, x_i]$, *por lo menos una de las funciones o* φ_{i-1} *o* φ_i *desaparece. Entonces*

$$\begin{aligned} a_{i,i-1} &= \int_{x_{i-1}}^{x_i} \frac{1}{h}\left(-\frac{1}{h}\right)\mathrm{d}x = -\frac{1}{h}, \\ a_{ii} &= \int_0^\pi \big(\varphi_i'(x)\big)^2\,\mathrm{d}x = \int_{x_{i-1}}^{x_{i+1}} \big(\varphi_i'(x)\big)^2\,\mathrm{d}x = \frac{2}{h}, \\ a_{i,i+1} &= a_{i+1,i} = -\frac{1}{h} \end{aligned}$$

para $i = 1, \ldots, n-1$, *es decir, obtenemos la matriz definida positiva*

$$\boldsymbol{A} = \frac{1}{h}\begin{bmatrix} 2 & -1 & 0 & \cdots & 0 \\ -1 & 2 & -1 & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & -1 & 2 & -1 \\ 0 & \cdots & 0 & -1 & 2 \end{bmatrix}$$

6.2.2. Funciones de planteo cúbicas por trozos. Para poder aplicar el método de Ritz con funciones de planteo cúbicas, definimos la siguiente base de funciones sobre $[a, b]$, donde

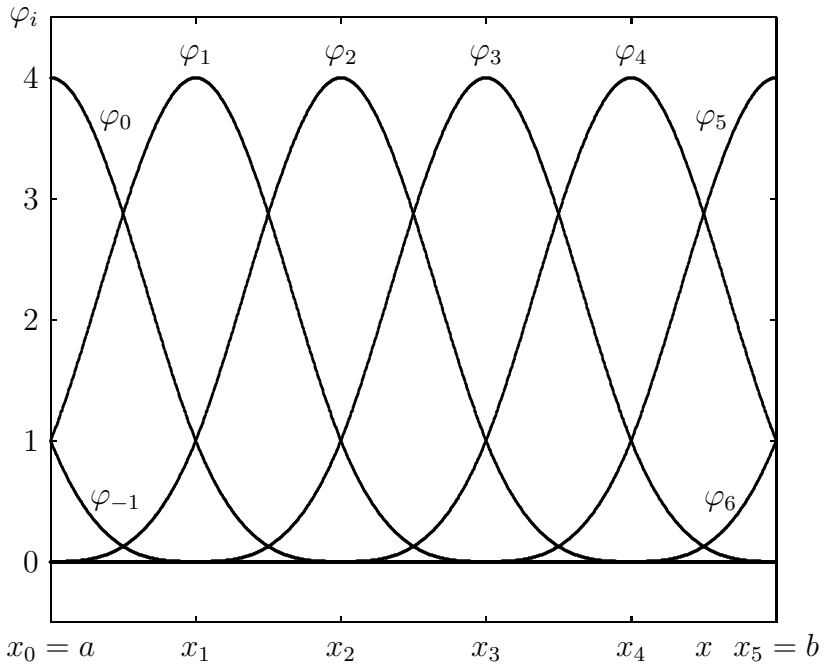


FIGURA 6.2. Las funciones φ_i, $i = -1, \ldots, n+1$, dadas por (6.20) para $n = 5$.

a los nodos ya definidos (6.16) agregamos $x_{-k} = a - kh$ y $x_{n+k} = b + kh$ para $k = 1, 2, 3$:

$$\varphi_i(x) = \begin{cases} \left(-\dfrac{a-x}{h} - i + 2\right)^3 & \text{para } x \in [x_{i-2}, x_{i-1}] \cap [a, b], \\ 1 + 3\left(-\dfrac{a-x}{h} - i + 1\right) + 3\left(-\dfrac{a-x}{h} - i + 1\right)^2 - 3\left(-\dfrac{a-x}{h} - i + 1\right)^3 & \\ & \text{para } x \in [x_{i-1}, x_i] \cap [a, b], \\ 1 + 3\left(-\dfrac{a-x}{h} + i + 1\right) + 3\left(\dfrac{a-x}{h} + i + 1\right)^2 - 3\left(\dfrac{a-x}{h} + i + 1\right)^3 & \\ & \text{para } x \in [x_i, x_{i+1}] \cap [a, b], \\ \left(\dfrac{a-x}{h} + i + 2\right)^3 & \text{para } x \in [x_{i+1}, x_{i+2}] \cap [a, b], \\ 0 & \text{sino,} \end{cases}$$

$$i = -1, 0, \ldots, n, n+1. \tag{6.20}$$

A modo de ejemplo, la Figura 6.2 muestra las funciones φ_i, $i = -1, \ldots, n+1$, dadas por (6.20) para $n = 5$.

Cada función spline es de la forma

$$s_h(x) = \sum_{i=-1}^{n+1} \alpha_i \varphi_i(x). \tag{6.21}$$

Sin embargo, para la aproximación usaremos sólo aquellas funciones spline que satisfacen las condiciones de borde $s_h(a) = s_h(b) = 0$. Sea $P_h^3(a,b)$ el espacio de estas funciones. Pero $\varphi_{-1}(a) = \varphi_1(a) = \varphi_{n-1}(b) = \varphi_{n+1}(b) = 1$ y $\varphi_0(a) = \varphi_n(b) = 4$, es decir, las funciones (6.21) no van a satisfacer en general las condiciones de borde, por lo tanto definimos las nuevas funciones de base

$$\psi_i(x) := \begin{cases} \varphi_i(x) & \text{para } i = 2, \ldots, n-2, \\ \varphi_0(x) - 4\varphi_{-1}(x) & \text{para } i = 0, \\ \varphi_0(x) - 4\varphi_1(x) & \text{para } i = 1, \\ \varphi_n(x) - 4\varphi_{n-1}(x) & \text{para } i = n-1, \\ \varphi_n(x) - 4\varphi_{n+1}(x) & \text{para } i = n. \end{cases}$$

Obviamente, estas funciones satisfacen $\psi_0(a) = \psi_1(a) = \psi_{n-1}(b) = \psi_n(b) = 0$; entonces, $\psi_i(a) = \psi_i(b) = 0$ para $i = 0, \ldots, n$. Se puede demostrar que estas funciones son linealmente independientes. Concluimos que cada función $s_h^0 \in P_h^3(a,b)$ posee una representación

$$s_h^0(x) = \sum_{i=0}^{n} \alpha_i \psi_i(x), \quad s_h^0(a) = s_h^0(b) = 0.$$

Las funciones ψ_i forman una base del espacio $P_h^3(a,b)$, el cual a su vez es un subespacio de $\mathcal{D}$. Por lo tanto, podemos usar las funciones ψ_i como funciones de planteo para el método de Ritz. Los elementos de la matriz $\boldsymbol{A}$ ahora son

$$\begin{aligned} a_{ij} = [\psi_i, \psi_j] &= \int_a^b \big(p(x)\psi_i'(x)\psi_j'(x) + q(x)\psi_i(x)\psi_j(x)\big)\,\mathrm{d}x \\ &= \sum_{k=1}^{n} \int_{x_{k-1}}^{x_k} \big(p(x)\psi_i'(x)\psi_j'(x) + q(x)\psi_i(x)\psi_j(x)\big)\,\mathrm{d}x. \end{aligned}$$

Obviamente, $[\psi_i, \psi_j] = 0$ para $|i-j| \geqslant 4$. Ahora la matrix $\boldsymbol{A}$ es una matriz de banda, que en su i-ésima fila tiene sólo los elementos $a_{i,i-3}, a_{i,i-2}, \ldots, a_{i,i+3}$ diferentes de cero. La matriz es simétrica y definida positiva.

6.2.3. Estudio del error y extensiones. Brevemente vamos a discutir el error cometido por el método de Ritz. Para el espacio D ya definimos la norma $\|\cdot\|_{V^1(a,b)}$ en (6.8). Si y es la solución exacta del problema de valores de frontera (6.4) y $u^* \in D$ es una solución aproximada, nso interesa una cota del error $\|u^* - y\|_{V^1(a,b)}$. Para el planteo con funciones cúbicas por trozos, también es posible estimar

$$\|u^* - y\|_\infty := \max_{x\in[a,b]} \big|u^*(x) - y(x)\big|.$$

Nos vamos a referir al siguiente lema sin desmostración.

Lema 6.1. *Bajo las hipótesis de regularidad* (6.5)*, existen constantes* $0 < \gamma_{2,1} < \Gamma_{2,1}$ *y* $0 < \gamma_\infty < \Gamma_\infty$ *tales que*

$$\begin{aligned} &\forall u \in V^2(a,b): \quad \gamma_\infty \|u\|_\infty^2 \leqslant [u,u] \leqslant \Gamma_\infty \|u\|_\infty^2, \\ &\forall u \in V^1(a,b): \quad \gamma_{2,1} \|u\|_{V^1(a,b)}^2 \leqslant [u,u] \leqslant \Gamma_{2,1} \|u\|_{V^1(a,b)}^2. \end{aligned}$$

Teorema 6.3. *Sea* y *la solución exacta del problema de valores de frontera* (6.4) *y* $v^* \in D_M$ *la aproximación obtenida por el método de Ritz. Entronces, para cada* $v \in D_M$ *tenemos las desigualdades*

$$\begin{aligned} \|v^* - y\|_{V^1(a,b)} &\leqslant \left(\frac{\Gamma_{2,1}}{\gamma_{2,1}}\right)^{1/2} \|v - y\|_{V^1(a,b)} \quad \text{si } D_M \subset V^1(a,b), \\ \|v^* - y\|_\infty &\leqslant \left(\frac{\Gamma_\infty}{\gamma_\infty}\right)^{1/2} \|v' - y'\|_\infty \quad \text{si } D_M \subset V^2(a,b). \end{aligned} \tag{6.22}$$

Demostración. Ya establecimos $[v-y, v-y] = I[v] + [y,y]$ para toda función $v \in V^1(a,b)$, es decir, tambien para todo $v \in D_M \subset D = V^1(a,b)$. Por otro lado, de acuerdo con $\min_{v \in D_M} I[v] = I[v^*]$,

$$\min_{v \in D_M} [v-y, v-y] = \min_{v \in D_M} I[v] + [y,y] = I[v^*] + [y,y] = [v^*-y, v^*-y].$$

Entonces,

$$\forall v \in D_M: \quad [v-y, v-y] \geqslant [v^*-y, v^*-y].$$

Dado que $v - y \in V^1(a,b)$ para $D_M \subset V^1(a,b)$ y $(v-y) \in V^2(a,b)$ si $D_M \subset V^2(a,b)$, Teorema 6.3 sigue en virtud del Lemma 6.1. ∎

Usaremos el teorema para estimar el error cometido por el método de elementos finitos. Para ello aprovechamos que (6.22) es válido para cada función $v \in D_M$. Primero, sea $D_M = P_h^1(a,b)$, es decir, el espacio de las funciones continuas y lineales por trozos con $s_h(a) = s_h(b) = 0$. Además, sea v_h la única función de $P_h^1(a,b)$ que satisface $v_h(x_i) = y(x_i)$ para $i = 0, \ldots, n$:

$$v_h(x) = \frac{y(x_{i+1}) - y(x_i)}{h}(x - x_i) + y(x_i), \quad x \in [x_i, x_{i+1}], \quad i = 0, \ldots, n-1.$$

Entonces, para cada $y \in C^2[a,b]$ tenemos las cotas

$$\|v_h' - y'\|_\infty \leqslant 2Lh, \quad \|v_h - y\|_\infty \leqslant Lh^2$$

con una constante L. Insertando esta expresión en (6.22), obtenemos

$$\begin{aligned} \|v^* - y\|_{V^1(a,b)} &\leqslant (\Gamma_{2,1}/\gamma_{2,1})^{1/2} \|v_h - y\|_{V^1(a,b)} \\ &\leqslant (\Gamma_{2,1}/\gamma_{2,1})^{1/2} L(b-a)h(4+h^2)^{1/2}. \end{aligned}$$

Esto demuestra que el error medido en la norma $\|\cdot\|_{V^1(a,b)}$ es por lo menos proporcional a h, o sea el método converge para $h \to 0$, $n \to \infty$, y $nh = b - a$ del primer orden. El método parece menos exacto que el método de diferencias finitas. Pero hay que tomar en cuenta que

en $\|\cdot\|_{V^1(a,b)}$ también se mide el error de la aproximación de la derivada. Por ejemplo, se puede demostrar que

$$\|v^* - y\|_{V^0(a,b)} = \mathcal{O}(h^2).$$

Ahora elegimos el espacio $P_h^3(a,b) \subset V^2(a,b)$ de las funciones spline cúblicas, y sea $v_h(x)$ el spline que satisface las condiciones

$$v_h(x_i) = y(x_i), \quad i = 0, \ldots, n; \tag{6.23}$$

$$v_h'(a) = y'(a), \quad v_h'(b) = y'(b). \tag{6.24}$$

A través de las condiciones (6.23) y (6.24), la función v_h está determinada unicamente. Según resultados de la aproximación de una función por una función spline,

$$\|v_h' - y'\|_\infty \leqslant M_1 N L h^3.$$

Usando (6.22), obtenemos

$$\max_{x\in[a,b]} \big|v^*(x) - y(x)\big| = \|v^* - y\|_\infty \leqslant M_1 N L \sqrt{\frac{\Gamma_\infty}{\gamma_\infty}} h^3 = \mathcal{O}(h^3). \tag{6.25}$$

Eso significa que el error es del orden a lo menos $\mathcal{O}(h^3)$, y el método es más exacto que él de diferencias finitas; sin embargo, aquí la cota (6.25) aún puede ser mejorada. Por otro lado, un método de diferencias finitas es mucho más facil de implementar.

En general, para el método de Ritz podriamos usar funciones de planteo φ_i polinomiales por trozos de alto orden para incrementar el orden de convergencia. Sin embargo, tal medida no vale el esfuerzo computacional adicional. El método de elementos finitos puede ser aplicado a la solución numérica de cualquier problema lineal de valores de frontera de una ecuación diferencial ordinaria de segundo orden.

6.3. El método de Ritz y elementos finitos para problemas de valores de ffontera de ecuaciones diferenciales parciales elípticas

Según el Teorema 4.1, ya sabemos que el problema

$$Lv = f, \quad v \in \mathcal{D}, \tag{6.26}$$

$$L^*u \equiv -(a_{11}u)_{xx} - 2(a_{12}u)_{xy} - (a_{22}u)_{yy} + (a_1u)_x + (a_2u)_y + au, \tag{6.27}$$

$$\mathcal{D} := \big\{v \in C^0(\bar{G}) \cap C^2(G) \mid v = 0 \text{ en } \partial G\big\} \tag{6.28}$$

es solucionado si encontramos una función $u = u(x,y)$ tal que

$$I[u] = \min_{v\in\mathcal{D}} = \min_{v\in\mathcal{D}} \big\{(v, Lv) - 2(v,f)\big\}, \tag{6.29}$$

donde

$$(v, Lv) - 2(v,f) = \int_G \big(a_{11}v_x^2 + 2a_{12}v_xv_y + a_{22}v_y^2 + av^2 - 2vf\big)\,\mathrm{d}x\,\mathrm{d}y.$$

En el método de Ritz, $I[v]$ se minimiza aproximadamente. Este método es muy similar al método ya discutido para la solución de problemas de valores de frontera para ecuaciones diferenciales ordinarias. Se elige un sub-espacio $D_M \subset D$ M-dimensional, donde

$$D = \big\{v \in V(G) \mid v = 0 \,\text{en}\, \partial G\big\},$$

y se considera como base un sistema de funciones $\varphi_1, \ldots, \varphi_M$ linealmente independientes. Cada función $v \in D_M$ es de la forma

$$v = \sum_{\mu=1}^{M} \alpha_\mu \varphi_\mu, \quad \alpha_1, \ldots, \alpha_M \in \mathbb{R}. \tag{6.30}$$

Esta función v satisface

$$\begin{aligned} I[v] &= [v,v] - 2(v,f) \\ &= \left[\sum_{\mu=1}^{M} \alpha_\mu \varphi_\mu, \sum_{\nu=1}^{M} \alpha_\nu \varphi_\nu\right] - 2\left(\sum_{\mu=1}^{M} \alpha_\mu \varphi_\mu, f\right) \\ &= \sum_{\mu,\nu=1}^{M} \alpha_\mu \alpha_\nu [\varphi_\mu, \varphi_\nu] - 2\sum_{\mu=1}^{M} \alpha_\mu (\varphi_\mu, f) \\ &=: \Phi(\alpha_1, \ldots, \alpha_M). \end{aligned}$$

Como en el caso de ecuaciones diferenciales ordinarias, el requerimiento $I[v] \to$ mín lleva al sistema de ecuaciones lineales

$$\frac{1}{2}\frac{\partial \Phi(\alpha_1, \ldots, \alpha_M)}{\partial \alpha_i} = \sum_{j=1}^{M} [\varphi_i, \varphi_j]\alpha_j - (\varphi_i, f) = 0, \quad i = 1, \ldots, M. \tag{6.31}$$

Definiendo

$$\boldsymbol{S} = ([\varphi_i, \varphi_j]), \quad (\varphi_i, f) = b_i, \quad \boldsymbol{b} = (b_1, \ldots, b_M)^{\mathrm{T}}, \quad \boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_M)^{\mathrm{T}},$$

podemos escribir el sistema (6.31) en la forma

$$\boldsymbol{S}\boldsymbol{\alpha} = \boldsymbol{b}. \tag{6.32}$$

La matriz $\boldsymbol{S}$ es simétrica y definida positiva, es decir (6.32) tiene una solución única $\boldsymbol{\alpha}^* = (\alpha_1^*, \ldots, \alpha_M^*)^{\mathrm{T}}$. Tal como en el caso de ecuaciones diferenciales ordinarias, se demuestra que $\Phi(\boldsymbol{\alpha}^*) \leqslant \Phi(\boldsymbol{\alpha})$ y entonces

$$I[v^*] = \min_{v \in D_M} I[v], \quad v^*(x,y) = \sum_{\mu=1}^{M} \alpha_\mu^* \varphi_\mu(x,y).$$

La selección de las funciones φ_μ, es decir, del sub-espacio D_M, es dificil; además, hay que determinar los valores de las integrales numericamente. Una parte de las dificultades puede ser evitada si subdividimos $\bar{G}$ an áreas pequeños y usamos sólo funciones de base φ_μ donde cada una es diferente de cero sólo sobre pocos subdominios.

Consideremos una subdivisión de $\bar{G}$ en pequeños triángulos, es decir, una *triangulación*. Supongamos que G es un dominio acotado, abierto y convexo con una frontera ∂G continua,

la cual puede ser aproximada por un trazado poligonal ∂G_h, el cual enteramente pertenece a G y cuyos vértices pertenecen a ∂G. Este trazado poligonal es el borde de un dominio poligonal G_h.

Ahora triangulamos el dominio $\bar{G}$ en la siguiente forma, que permite representar $\bar{G}_h$ como

$$\bar{G}_h = \bigcup_{i=1}^{N} D_i,$$

con triángulos $D_1, \ldots, D_N$, donde dos triángulos o no tienen ningún punto en común, o coinciden en exactamente un lado que pertenece a ambos triángulos, o coinciden en exactamente un vértice. De este modo, se evitan los llamados "nodos colgantes".

Sea h_i la distancia máxima de dos vértices de D_i y

$$h := \max_{1 \leqslant j \leqslant N} h_j.$$

En este caso, ∂G_h y $\bar{G}_h$ dependerán de h, y obviamente, ∂G es mejor aproximado por ∂G_h cuando h es pequeño.

Los vértices de los triángulos D_i se llaman *nodos*, el conjunto de los nodos de llama R, el conjunto de los nodos en ∂G_h se llama ∂R. Bajo nustras hipótesis, ∂G_h puede ser contruido de tal forma que los nodos en ∂G_h también pertenece a ∂G, es decir, exigimos que $\partial R \subset \partial G$. En general, una modificación del parámetro h implicará un nuevo borde ∂G_h. Supongamos que R y ∂R contienen M y $\tilde{M}$ nodos, respectivamente, los que enumeramos en un cierto orden; sean (x_j, y_j), $j = 1, \ldots, M$ y $(\tilde{x}_k, \tilde{y}_k)$, $k = 1, \ldots, \tilde{M}$ estos puntos. Definimos las funciones base φ_i, $i = 1, \ldots, M$, de la siguiente forma:

1. Para $i = 1, \ldots, M$, $\varphi_i \in C^0(\bar{G}_h)$.
2. Sobre cada triángulo D_l, $l = 1, \ldots, N$, cada función φ_i es un polinomio de grado 1 en x e y.
3. Para $i, j = 1, \ldots, M$, $\varphi_i(x_j, y_j) = \delta_{ij}$.
4. Para $i = 1, \ldots, M$ y $k = 1, \ldots, \tilde{M}$, $\varphi_i(\tilde{x}_k, \tilde{y}_k) = 0$.

Estos requerimientos determinan las funciones $\varphi_1, \ldots, \varphi_M$ en forma única. En (x_i, y_i), la función φ_i tiene el valore 1, y en todos los demás nodos el valor cero.

Las funciones φ_i son elementos del espacio $V(G_h)$, o sea podemos formar las formas bilineales $[\varphi_i, \varphi_j]$ remplazando G por G_h, dado que las funciones φ_i son diferenciables con respecto a x e y sobre cada triángulo D_l, $l = 1, \ldots, N$. La integral puede ser representada como

$$\int_{G_h} = \sum_{i=1}^{N} \int_{D_i}.$$

Las funciones $\varphi_1, \ldots, \varphi_M$ son linealmente independientes sobre $\bar{G}_h$. (Si no fuera así, debería existir una ecuación

$$\sum_{i=1}^{M} \alpha_i \varphi_i(x, y) = 0, \quad (x, y) \in \bar{G}_h,$$

donde no todos los coeficientes α_i desaparecen. Pero

$$\sum_{i=1}^{M} \alpha_i \varphi_i(x_j, y_j) = \sum_{i=1}^{M} \alpha_i \delta_{ij} = \alpha_j = 0, \quad j = 1, \ldots, M,$$

así que todos los coeficientes deben desaparecer.) Entonces, las funciones φ_i generan como funciones base un espacio D_M; especificamente, D_M es el espacio de aquellas funciones continuas definidas sobre G_h que son afinamente lineales sobre cada triángulo D_l, y que desaparecen sobre ∂G_h.

Ahora queremos aproximar la solución del problema variacional (6.29) por una función de D_M; esta función será también una aproximación del problema de valores de frontera (6.26). El planteo (6.30) nos lleva a la solución aproximada

$$v^*(x, y) = \sum_{i=1}^{M} \alpha_i^* \varphi_i(x, y),$$

donde

$$v^*(x_j, y_j) = \sum_{i=1}^{M} \alpha_i^* \varphi_i(x_j, y_j) = \sum_{i=1}^{M} \alpha_i^* \delta_{ij} = \alpha_j^*,$$

es decir

$$v^*(x, y) = \sum_{i=1}^{M} v^*(x_i, y_i) \varphi_i(x, y).$$

La matriz $\boldsymbol{S}$ del sistema lineal (6.32) tiene muy pocos elementos diferentos de cero, dado que

$$[\varphi_i, \varphi_j] = \int_{G_h} \Big\{ a_{11}(\varphi_i)_x(\varphi_j)_y + a_{12}\big((\varphi_i)_x(\varphi_j)_y + (\varphi_i)_y(\varphi_j)_x\big) + a_{22}(\varphi_i)_y(\varphi_j)_y + a\varphi_i\varphi_j \Big\} \, \mathrm{d}x \, \mathrm{d}y$$

y $\varphi_i \varphi_j = 0$ si (x_i, y_i) y (x_j, y_j) no son puntos vecinos. Si existen exactamente k_i triángulos con el vértice $(x_i, y_i$, la matriz $\boldsymbol{S}$ tendrá en su i-ésima fila exactamente $k_i + 1$ elementos diferentes de cero. En el caso discutido, la matriz $\boldsymbol{S}$ es simétrica y definida positiva, y por lo tanto apta para la solución numérica del sistema (6.32) por el método SOR, de gradientes conjugados, o un método multi-mallas.

Si la triangulación no es uniforme, la computación de los coeficientes del sistema lineal (6.32) puede ser bastante costosa; pero normalmente existe software que se encarga automaticamente de esta tarea. En general, se trata de trinagular un dominio de tal forma que los triángulos D_i no diferen demasiado en forma y área. Pero para un dominio G con una frontera curvada habrá que elegir triángulos diferentes cerca de la frontera. La gran libertad que existe en la selección de los triángulos constituye una gran ventaja de los métodos de elementos finitos comparado con el método de diferencias finitas.

En el interior de G, en muchos casos facilmente de puede generar una triangulación donde x_i e y_i son múltiples enteros de h, por ejemplo $x_i = m_i h$, $y = n_i h$. En este caso (ver

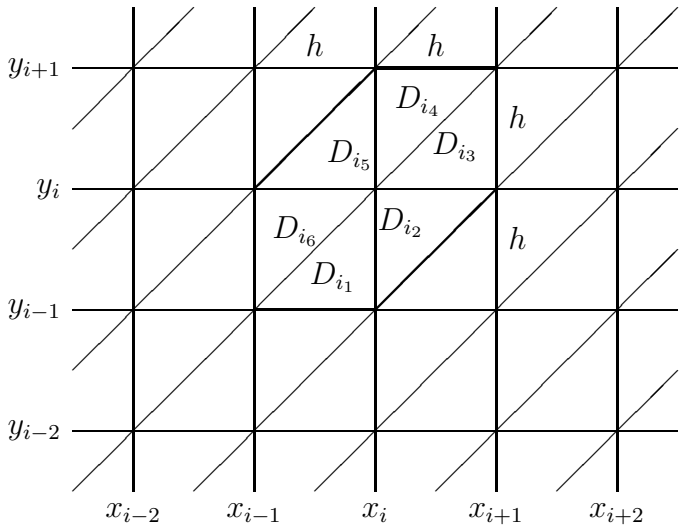


FIGURA 6.3. La definción de la función base φ_i según (6.33).

Figura 6.3),

$$\varphi_i(x,y) = \begin{cases} 1 - n_i + \dfrac{y}{h} & \text{en } D_{i_1}, \\ 1 + m_i - n_i - \dfrac{x-y}{h} & \text{en } D_{i_2}, \\ 1 + m_i - \dfrac{x}{h} & \text{en } D_{i_3}, \\ 1 + n_i - \dfrac{y}{h} & \text{en } D_{i_4}, \\ 1 - m_i + n_i + \dfrac{x-y}{h} & \text{en } D_{i_5}, \\ 1 - m_i + \dfrac{x}{h} & \text{en } D_{i_6}, \\ 0 & \text{en todas otras partes.} \end{cases} \tag{6.33}$$

Para la computación de $[\varphi_i, \varphi_j]$ y (φ_i, f) habrá que usar una fórmula de cubatura si a_{ik}, a y f son funciones complicadas.

Para la discusión de algunos aspectos prácticos, consideremos el problema de valores de frontera

$$\begin{aligned} Lu \equiv -u_{xx} - u_{yy} &= f \quad \text{en } G, \\ u &= 0 \quad \text{en } \partial G \end{aligned}$$

con el problema variacional asociado

$$I[v] = \int_G (v_x^2 + v_y^2 - 2vf)\,\mathrm{d}x\,\mathrm{d}y \stackrel{!}{=} \text{mín}, \quad v = 0 \text{ en } \partial G. \tag{6.34}$$

Supongamos que

$$\bar{G} = \bigcup_{i=1}^{N} D_i,$$

es decir, G es la unión de un número finito N de triángulos, entonces (6.34) implica que

$$I[v] = \sum_{i=1}^{N} \int_{D_i} (v_x^2 + v_y^2 - 2vf)\,\mathrm{d}x\,\mathrm{d}y \stackrel{!}{=} \text{mín}\,. \tag{6.35}$$

Sobre cada triángulo D_i, v es una función afina, es decir, existen constantes a_i, b_i y c_i tales que

$$v(x,y) = a_i + b_i x + c_i y, \quad v_x = b_i, \quad v_y = c_i, \quad (x,y) \in D_i.$$

Los vértices de D_i sean (x_{ik}, y_{ik}) para $k = 1, \ldots, 3$. Si v_{ik} es el valor de la aproximación $v = v(x,y)$ en (x_{ik}, y_{ik}), podemos escribir

$$v(x_{ik}, y_{ik}) = v_{ik} = a_i + b_i x_{ik} + c_i y_{ik}, \quad k = 1, \ldots, 3,$$

y de este sistema de ecuaciones podemos determinar de forma única a_i, b_i y c_i como funciones de x_{ik}, y_{ik} y v_{ik}. Por lo tanto, en nuestro caso podemos escribir (6.35) como

$$I[v] = \sum_{i=1}^{N} \int_{D_i} \big(b_i^2 + c_i^2 - 2f(x,y)(a_i + b_i x + c_i y)\big)\,\mathrm{d}x\,\mathrm{d}y \stackrel{!}{=} \text{mín}\,. \tag{6.36}$$

Si denominamos por v_j el valor de $v(x,y)$ en el nodo (x_j, y_j), $j = 1, \ldots, M$, podemos escribir obviamente

$$I[v] = \Phi(v_1, \ldots, v_M).$$

El valor v_j, $1 \leqslant j \leqslant M$, aparece entre los valores v_{ik}, $i = 1, \ldots, N$, $k = 1, \ldots, 3$, L_j veces si (x_j, v_j) es vértice de L_j triángulos. Si definimos $\boldsymbol{v}^h := (v_1, \ldots, v_M)^{\mathrm{T}}$, obtenemos con la matriz $\boldsymbol{S} \in \mathbb{R}^{M \times M}$ y el vector $\boldsymbol{b} \in \mathbb{R}^M$ la identidad

$$\Phi(v_1, \ldots, v_M) = (\boldsymbol{v}^h)^{\mathrm{T}} \boldsymbol{S} \boldsymbol{v}^h - 2(\boldsymbol{v}^h)^{\mathrm{T}} \boldsymbol{b}.$$

El requerimiento $\Phi(v_1, \ldots, v_m) \stackrel{!}{=} \text{mín}$ lleva al sistema de ecuaciones

$$\boldsymbol{S} \boldsymbol{v}^h = \boldsymbol{b}.$$

La computación de $\boldsymbol{S}$ y $\boldsymbol{b}$ es (por lo menos) un poco complicada, puesto que hay que integrar sobre cada triángulo separadamente. Esta complicación puede ser evitada mediante la interoducción de un *triángulo de referencia*. A través de una transformación de coordenadas biyectiva, cada triángulo D_i puede ser transformado al triángulo de referencia con los vértices $(0,0)$, $(1,0)$ y $(0,1)$. La transformación es

$$\begin{aligned} x &= \alpha_i(\xi, \eta) = x_{i1} + (x_{i2} - x_{i1})\xi + (x_{i3} - x_{i1})\eta, \\ y &= \beta_i(\xi, \eta) = y_{i1} + (y_{i2} - y_{i1})\xi + (y_{i3} - y_{i1})\eta. \end{aligned} \tag{6.37}$$

Denominamos el triángulo de referencia por D_0, definiendo

$$v(x,y) = v\big(\alpha_i(\xi,\eta), \beta_i(\xi,\eta)\big) = w_i(\xi,\eta), \quad (\xi,\eta) \in D_0.$$

En este caso, $w_i(\xi,\eta)$ está definido sobre D_0 y

$$w_i(\xi,\eta) = c_{i0} + c_{i1}\xi + c_{i2}\eta.$$

Los coeficientes c_{ij} pueden ser calculados facilmente; de

$$\begin{aligned} w_i(0,0) &= c_{i0},\\ w_i(1,0) &= c_{i0} + c_{i1},\\ w_i(0,1) &= c_{i0} + c_{i2} \end{aligned}$$

obtenemos

$$\begin{aligned} c_{i0} &= w_i(0,0),\\ c_{i1} &= w_i(1,0) - w_i(0,0),\\ c_{i2} &= w_i(0,1) - w_i(0,0) \end{aligned}$$

y por lo tanto

$$w_i(\xi,\eta) = (1-\xi-\eta)w_i(0,0) + \xi w_i(1,0) + \eta w_i(0,1).$$

Para la formulación de las ecuaciones de los elementos finitos, podríamos partir de (6.36), pero consideramos otro planteo. Escribimos (6.35) como

$$\sum_{i=1}^{N} \int_{D_i} F(x,y)\,\mathrm{d}x\,\mathrm{d}y \stackrel{!}{=} \text{mín}, \quad F(x,y) := v_x^2(x,y) + v_y^2(x,y) - 2v(x,y)f(x,y),$$

y tomamos en cuenta que

$$\int_{D_i} F(x,y)\,\mathrm{d}x\,\mathrm{d}y = \int_{D_0} F\big(\alpha_i(\xi,\eta),\beta_i(\xi,\eta)\big)\Delta_i(\xi,\eta)\,\mathrm{d}\xi\,\mathrm{d}\eta$$

con el determinante funcional

$$\Delta_i(\xi,\eta) = \begin{vmatrix} (\alpha_i)_\xi(\xi,\eta) & (\beta_i)_\xi(\xi,\eta) \\ (\alpha_i)_\eta(\xi,\eta) & (\beta_i)_\eta(\xi,\eta) \end{vmatrix} = \begin{vmatrix} x_\xi & y_\xi \\ x_\eta & y_\eta \end{vmatrix}.$$

Las derivadas de α_i y β_i pueden ser calculadas inmediatamente; obtenemos

$$\Delta_i(\xi,\eta) = (x_{i2}-x_{i1})(y_{i3}-y_{i1}) - (x_{i3}-x_{i1})(y_{i2}-y_{i1}),$$

es decir, para cada i, Δ_i es constante: $\Delta_i(\xi,\eta) = \Delta_i$. La transformación (6.37) puede ser invertida facilmente, y nos entrega

$$\begin{aligned} \xi = \gamma_i(x,y) &= \frac{1}{\Delta_i}\big((x-x_{i1})(y_{i3}-y_{i1}) - (y-y_{i1})(x_{i3}-x_{i1})\big),\\ \eta = \delta_i(x,y) &= \frac{1}{\Delta_i}\big(-(x-x_{i1})(y_{i2}-y_{i1}) + (y-y_{i1})(x_{i2}-x_{i1})\big). \end{aligned} \tag{6.38}$$

Dado que $v(x,y) = w_i(\xi,\eta)$ para $(x,y) \in D_i$, tenemos

$$v_x = (w_i)_\xi \xi_x + (w_i)_\eta \eta_x, \quad v_y = (w_i)_\xi \xi_y + (w_i)_\eta \eta_y \quad \text{en } D_i.$$

Usando (6.38), podemos escribir

$$\xi_x = \frac{y_{i3}-y_{i1}}{\Delta_i}, \quad \eta_x = -\frac{y_{i2}-y_{i1}}{\Delta_i}, \quad \xi_y = -\frac{x_{i3}-x_{i1}}{\Delta_i}, \quad \eta_y = \frac{x_{i2}-x_{i1}}{\Delta_i},$$

y finalmente

$$\begin{aligned} I_i[v] &:= \int_{D_i} \big(v_x^2(x,y) + v_y^2(x,y) - 2v(x,y)f(x,y)\big)\,\mathrm{d}x\,\mathrm{d}y \\ &= \Delta_i \int_{D_0} \Bigg\{ \left[(w_i)_\xi \frac{y_{i3}-y_{i1}}{\Delta_i} - (w_i)_\eta \frac{y_{i2}-y_{i1}}{\Delta_i}\right]^2 \\ &\qquad + \left[-(w_i)_\xi \frac{x_{i3}-x_{i1}}{\Delta_i} + (w_i)_\eta \frac{x_{i2}-x_{i1}}{\Delta_i}\right]^2 - 2w_i h_i(\xi,\eta)\Bigg\}\mathrm{d}\xi\,\mathrm{d}\eta, \end{aligned} \tag{6.39}$$

donde definimos $h_i(\xi,\eta) := f(\alpha_i(\xi,\eta),\beta_i(\xi,\eta))$. En virtud de

$$\begin{aligned} (w_i)_\xi &= c_{i1} = w_i(1,0) - w_i(0,0) = v(x_{i2},y_{i2}) - v(x_{i1},y_{i1}) = v_{i2} - v_{i1}, \\ (w_i)_\eta &= c_{i2} = w_i(0,1) - w_i(0,0) = v(x_{i3},y_{i3}) - v(x_{i1},y_{i1}) = v_{i3} - v_{i1}, \\ w_i &= c_{i0} + c_{i1}\xi + c_{i2}\eta = v_{i1} + (v_{i2}-v_{i1})\xi + (v_{i3}-v_{i1})\eta \end{aligned}$$

obtenemos de (6.39)

$$\begin{aligned} I_i[v] =& \frac{1}{\Delta_i} \int_{D_0} \Big\{ \big[(v_{i2}-v_{i1})(y_{i3}-y_{i1}) - (v_{i3}-v_{i1})(y_{i2}-y_{i1})\big]^2 \\ &+ \big[(v_{i3}-v_{i1})(x_{i2}-x_{i1}) - (v_{i2}-v_{i1})(x_{i3}-x_{i1})\big]^2 \\ &- 2\Delta_i^2 \big[v_{i1} + (v_{i2}-v_{i1})\xi + (v_{i3}-v_{i1})\eta\big] h_i(\xi,\eta)\Big\}\,\mathrm{d}\xi\,\mathrm{d}\eta. \end{aligned} \tag{6.40}$$

Del requerimiento

$$\sum_{i=1}^{N} I_i[v] \stackrel{!}{=} \text{mín}$$

obtenemos finalmente los valores deseados de v en los nodos. Formulamos el sistema de ecuaciones correspondiente. Primero calcuamos $I_i[v]$, en general por integración numérica. Las primeras dos líneas de (6.40) llevan a la expresión

$$(\boldsymbol{v}^i)^{\mathrm{T}}\big(\bar{\boldsymbol{S}}_1^{(i)} + \bar{\boldsymbol{S}}_2^{(i)}\big)\boldsymbol{v}^i = (\boldsymbol{v}^i)^{\mathrm{T}}\bar{\boldsymbol{S}}^{(i)}\boldsymbol{v}^i = \sum_{l,m=1}^{3} \bar{s}_{lm}^{(i)} v_{il} v_{im}, \quad \boldsymbol{v}^i = \begin{pmatrix} v_{i1} \\ v_{i2} \\ v_{i3} \end{pmatrix}.$$

A modo de ejemplo, calculamos $(\boldsymbol{v}^i)^{\mathrm{T}}\bar{\boldsymbol{S}}_1^{(i)}\boldsymbol{v}^i$ explícitamente. Obtenemos

$$\begin{aligned} (\boldsymbol{v}^i)^{\mathrm{T}}\bar{\boldsymbol{S}}_1^{(i)}\boldsymbol{v}^i =& (y_{i2}-y_{i3})^2 v_{i1}^2 + (y_{i3}-y_{i1})^2 v_{i2}^2 + (y_{i2}-y_{i1})^2 v_{i3}^2 - 2(y_{i3}-y_{i1})(y_{i3}-y_{i2})v_{i1}v_{i2} \\ &- 2(y_{i2}-y_{i1})(y_{i2}-y_{i3})v_{i1}v_{i3} - 2(y_{i3}-y_{i1})(y_{i2}-y_{i1})v_{i2}v_{i3}. \end{aligned}$$

Entonces, la matriz simétrica $\bar{\boldsymbol{S}}_1^{(i)}$ es

$$\bar{\boldsymbol{S}}_1^{(i)} = \begin{bmatrix} (y_{i2}-y_{i3})^2 & -(y_{i3}-y_{i1})(y_{i3}-y_{i2}) & -(y_{i2}-y_{i1})(y_{i2}-y_{i3}) \\ -(y_{i3}-y_{i1})(y_{i3}-y_{i2}) & (y_{i3}-y_{i1})^2 & -(y_{i3}-y_{i1})(y_{i2}-y_{i1}) \\ -(y_{i2}-y_{i1})(y_{i2}-y_{i3}) & -(y_{i3}-y_{i1})(y_{i2}-y_{i1}) & (y_{i2}-y_{i1})^2 \end{bmatrix}.$$

La matriz $\bar{\boldsymbol{S}}_2^{(i)}$ puede ser calculada analogamente. Definiendo $\boldsymbol{S}^{(i)} := \Delta_i^{-1}\bar{\boldsymbol{S}}^{(i)}$ y tomando en cuenta que

$$\int_{D_0} \mathrm{d}\xi\,\mathrm{d}\eta = \frac{1}{2},$$

obtenemos

$$I_i[v] = \frac{1}{2}(\boldsymbol{v}_i)^{\mathrm{T}}\boldsymbol{S}^{(i)}\boldsymbol{v}^i - 2\Delta_i \int_{D_0} \big(v_{i1} + (v_{i2} - v_{i1})\xi + (v_{i3} - v_{i1}\eta\big)h_i(\xi,\eta)\,\mathrm{d}\xi\,\mathrm{d}\eta.$$

La integral puede ser aproximada por la fórmula de cuadratura de Gauss más simple,

$$\int_{D_0} F(\xi,\eta)\,d\xi\,d\eta \approx \frac{1}{2}F\left(\frac{1}{3},\frac{1}{3}\right).$$

Después de una pequeña compuatción llegamos a

$$I_i[v] \approx \tilde{I}_i[v] := \frac{1}{2}(\boldsymbol{v}_i)^{\mathrm{T}}\boldsymbol{S}^{(i)}\boldsymbol{v}^i - \frac{\Delta_i}{3}h_i\left(\frac{1}{3},\frac{1}{3}\right)(v_{i1} + v_{i2} + v_{i3}).$$

Con $\boldsymbol{S}^{(i)} = (s_{lm}^{(i)})$, $l,m = 1,\ldots,3$ tenemos

$$\begin{aligned} I[v] \approx \tilde{I}[v] = \sum_{i=1}^{N}\tilde{I}_i[v] &= \frac{1}{2}\sum_{i=1}^{N}\left((\boldsymbol{v}^i)^{\mathrm{T}}\boldsymbol{S}^{(i)}\boldsymbol{v}^i - 2(\boldsymbol{b}^i)^{\mathrm{T}}\boldsymbol{v}^i\right) \\ &= \frac{1}{2}\sum_{i=1}^{N}\left(\sum_{l,m=1}^{3} s_{lm}^{(i)}v_{il}v_{im} - \frac{2\Delta_i}{3}h_i\left(\frac{1}{3},\frac{1}{3}\right)(v_{i1} + v_{i2} + v_{i3})\right). \end{aligned} \tag{6.41}$$

Ahoram, la tarea de minimizar $\tilde{I}[v]$ entrega el sistema de ecuaciones

$$\frac{\partial\tilde{I}[v]}{\partial v_k} = 0, \quad k = 1,\ldots,M, \tag{6.42}$$

donde $v_k = v(x_k, y_k)$. Estas variables también se llaman *variables de nodos*. Definiendo

$$\frac{\partial v_{ij}}{\partial v_k} = \delta_k^{(ij)} = \begin{cases} 0 & \text{si } v_{ij} \neq v_k, \\ 1 & \text{si } v_{ij} = v_k, \end{cases}$$

podemos escribir (6.42) como

$$\frac{\partial\tilde{I}[v]}{\partial v_k} = \frac{1}{2}\sum_{i=1}^{N}\left(\sum_{l,m=1}^{3} s_{lm}^{(i)}\big(\delta_k^{(li)}v_{im} + \delta_k^{(mi)}v_{il}\big) - \frac{2\Delta_i}{3}h_i\left(\frac{1}{3},\frac{1}{3}\right)\big(\delta_k^{(1i)} + \delta_k^{(2i)} + \delta_k^{(3i)}\big)\right) = 0.$$

Entonces, el sistema de ecuacions de los elementos finitos es

$$\frac{1}{2}\sum_{i=1}^{N}\sum_{l,m=1}^{3} s_{lm}^{(i)}\big(\delta_k^{(li)}v_{im} + \delta_k^{(mi)}v_{il}\big) = \sum_{i=1}^{N} b_k^{(i)}, \quad k = 1,\ldots,M, \tag{6.43}$$

donde

$$b_k^{(i)} = \frac{2\Delta_i}{3}h_i\left(\frac{1}{3},\frac{1}{3}\right)\big(\delta_k^{(1i)} + \delta_k^{(2i)} + \delta_k^{(3i)}\big). \tag{6.44}$$

Este sistema tiene la forma $\boldsymbol{S}\boldsymbol{v}^h = \boldsymbol{b}$, $\boldsymbol{v}^h = (v_1, \ldots, v_M)^{\mathrm{T}}$. La parte derecha (6.44) puede ser calculada facilmente, pero es costoso determinar los elementos de $\boldsymbol{S}$ desde (6.43). Para eso, volvemos nuevamente a (6.41) y fijamos la numeración de las variables de nodo v_k, $k = 1, \ldots, M$. Las identificamos con aquellos v_{ij}, $i = 1, \ldots, N$, $j = 1, \ldots, 3$, que no corresponden a valores de frontera. Cada v_{ij} es igual a una variable de nodos v_k. (Por supuesto, una variable v_k puede corresponder a varios v_{ij}.) Puesto que $v_{ij} = 0$ en los punto de frontera, podemos escribir (6.41) como

$$\tilde{I}[v] = \frac{1}{2} \sum_{k,l=1}^{M} s_{kl} v_k v_l - \sum_{k=1}^{M} b_k v_k, \quad b_k = \sum_{i=1}^{N} b_k^{(i)}.$$

Obviamente, podemos facilmente determinar la matriz $\boldsymbol{S} = (s_{kl})$. Sus elementos son sumas de elementos de las matrices $\boldsymbol{S}^{(i)}$.