EL METODO SEMIDISCRETO DE GALERKIN: ECUACIONES PARABOLICAS

por

Diego Bricio Hernandez

CIMAT, Apdo. Postal 402
36000-Guanajuato, Gto., Mexico 1987

0. INTRODUCCION

Durante el primer semestre de 1982, el autor y otros profesores de la División de Ciencias Básicas e Ingeniería de la UAM - Iztapalapa expusieron diversos temas dentro del marco de un Seminario de Elemento Finito. Dicho seminario estuvo basado principalmente en el libro de Schultz citado en la bibliografía; en particular, estas notas acusan una gran deuda con el capítulo 9 de dicha referencia.

Estas notas recogen algunas de las exposiciones dadas por el autor dentro de dicho seminario, e incluyen motivaciones y ejemplos que se espera hagan el material aquí presentado útil para quien desee asomarse a este tema. Los aspectos numéricos asociados han sido tratados por el autor con mayor amplitud en su libro de Análisis Numérico citado en la bibliografía.

El suscrito desea aprovechar esta oportunidad para agradecer a Hans Fetter Nathansky, Luis Mier y Terán y Luis Verde Star por sus muchos y útiles comentarios durante los trabajos de dicho seminario.

La Valenciana, Gto., 20.03.87

1. UN EJEMPLO INTRODUCTORIO

Consideremos el problema de valores iniciales y a la frontera

$$\frac{\partial u}{\partial t} = \frac{1}{2}\frac{\partial^2 u}{\partial x^2} \quad ; \quad t>0,\ 0<x<1 \qquad (1)'$$

$$u(t,0) = 0,\ u(t,1) = 0, \quad t>0 \qquad (1)''$$

$$u(0,x) = f(x) \quad ;0<x<1 \qquad (1)'''$$

con $f(0) = f(1) = 0$. La solución de este problema es bien conocida: por el método de separación de variables se llega fácilmente a que

$$u(t,x) = \sum_{k=1}^{\infty} A_k\, e^{-k^2\pi^2 t} \operatorname{sen} k\pi x \qquad (2)'$$

con

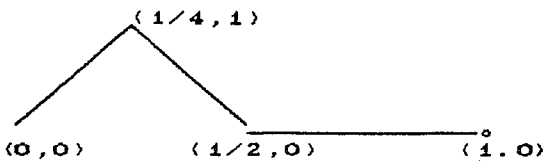
$$A_k = \sqrt{2}\int_0^1 f(x)\ \operatorname{sen} k\pi x dx \qquad (2)''$$

Esta solución define una trayectoria ω en el espacio de funciones continuas en $[0,1]$ que se anulan en los extremos del intervalo, con $\omega(0) = f$, $\omega(t)x = u(t,x)$.

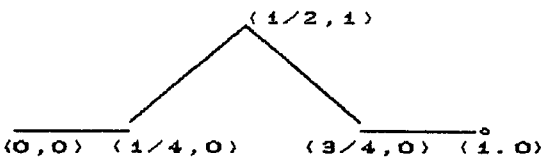
En este capítulo construiremos una familia de subespacios $\{C_h, h>0\}$ con los cuales aproximaremos al espacio C, y con ello también una familia $\{\omega_h, h>o\}$ de trayectorias, con $\omega_h \in C_h$, las que a su vez aproximarán a ω. Por ahora contentémonos con la aproximación correspondiente a $h = 1/4$.

Para ello, divídase el intervalo $[0,1]$ en cuatro partes iguales y considérense las funciones φ_1, φ_2, φ_3 cuyas gráficas son:

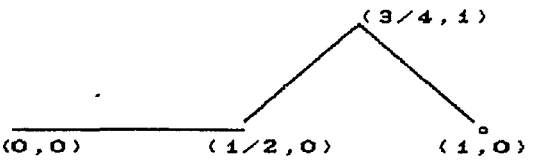
φ_1



φ_2



φ_3



Claramente φ_1, φ_2, $\varphi_3 \in C$. Además son linealmente independientes.

En efecto,

$$C_1\varphi_1 + C_2\varphi_2 + C_3\varphi_3 = 0$$

implica que

$$C_1\varphi_1(1/2) + C_2\varphi_2(1/2) + C_3\varphi_3\ (1/2) = 0$$

es decir, $C_2 = 0$, ya que $\varphi_2(1/2) = 1$, $\varphi_1(1/2) = \varphi_3(1/2) = 0$

Por otro lado, $(\varphi_1, \varphi_3) = 0$, con

$$(f,g) := \int_0^1 f(x)g(x)dx \qquad (3)$$

por lo que $C_1 = C_2 = 0$. Las funciones φ_1, φ_2 y φ_3 se llaman elementos finitos lineales.

Sea entonces $C_{1/4}$ el subespacio tridimensional de C generado por estos tres elementos. Una trayectoria $\omega_{1/4}:[0,\infty)\longrightarrow C_{1/4}$ que aproxime a ω se puede construir de la manera siguiente:

Poniendo

$$u_h(t,x) := \omega_h(t)x = \sum_{j=1}^{3} C_j(t)\varphi_j(x)$$

resulta que u_h deja un residuo en la ecuación diferencial de (1) dado por

$$\frac{\partial u_h}{\partial t} - \frac{1}{2}\frac{\partial^2 u_h}{\partial x^2} = \sum_{j=1}^{3}\{C_j(t)\varphi_j(x) - \tfrac{1}{2}C_j(t)\varphi_j''(x)\}$$

Este residuo no se anula, pero podemos imponer la condición de que sea ortogonal a todo el espacio $C_{1/4}$. Se obtiene:

$$\sum_{j=1}^{3} C_j(t))(\varphi_i,\varphi_j) = \frac{1}{2}\sum_{j=1}^{3} C_j(t)(\varphi_i,\varphi_j'') \qquad i = 1,2,3$$

donde el producto interior es (3). Si usamos integración por partes para f, g ∈ C dos veces diferenciables por tramos, resulta que

$$(f, g'') = -(f', g'),$$

de donde se obtiene el sistema lineal

$$M\dot{C} = AC \qquad (4)'$$

Aquí A y M son matrices simétricas, con

$$C_{ij} = \tfrac{1}{2}(\varphi_i', \varphi_j')$$

$$M_{ij} = (\varphi_i, \varphi_j)$$

es decir:

$$A = \begin{bmatrix} 1 & -\frac{1}{2} & 0 \\ -\frac{1}{2} & 1 & -\frac{1}{2} \\ 0 & -\frac{1}{2} & 1 \end{bmatrix} \qquad M = \frac{1}{24}\begin{bmatrix} 4 & -5 & 0 \\ -5 & 4 & 1 \\ 0 & 1 & 4 \end{bmatrix}$$

En cuanto a la condición inicial (2)''', u_h deja un residuo

$$u_h(0,x) - f(x) = \sum_{j=1}^{3} C_j(0)\varphi_j(x) - f(x)$$

Basta pedir que este residuo sea ortogonal a todo $C_{1/4}$ para llegar a que

$$MC(0) = b \qquad (4)'''$$

donde

$$b_i := (\varphi_i, f),$$

Tomando $f(x) = x(1-x)$, resulta que

$$b = \left[\frac{59}{384}, \frac{11}{96}, \frac{3955}{768} \right]$$

Así pues, encontramos la aproximación $\omega_{1/4}$ resolviendo el problema de valores iniciales

$$\begin{bmatrix} 4 & -5 & 0 \\ -5 & 4 & 1 \\ 0 & 1 & 4 \end{bmatrix} \begin{bmatrix} C_1 \\ C_2 \\ C_3 \end{bmatrix} = \begin{bmatrix} 24 & -12 & 0 \\ -12 & 24 & -12 \\ 0 & -12 & 24 \end{bmatrix} \begin{bmatrix} C_1 \\ C_2 \\ C_3 \end{bmatrix}$$

$$\begin{bmatrix} 4 & -5 & 0 \\ -5 & 4 & 1 \\ 0 & 1 & 4 \end{bmatrix} \begin{bmatrix} C_1(0) \\ C_2(0) \\ C_3(0) \end{bmatrix} = \frac{1}{768} \begin{bmatrix} 118 \\ 88 \\ 3955 \end{bmatrix}$$

cuya solución es única dado que M es no singular. Véase la sección 3.

Alternativamente, podríamos tomar los elementos finitos

$$\psi_j(x) = \sqrt{2}\, \text{sen}\, j\pi x \;, j = 1, 2, 3,$$

linealmente independientes al ser ortonormales con respecto al producto interior (3). Por ello mismo la matriz M en (4) se reduce a la identidad. Si tomamos, igual que antes, $f(x) = x(1-x)$

$$A = \frac{\pi^2}{2} \text{ diag } (-1, -4, -9), \quad C_i(0) = (f, \psi_i), \; i = 1, 2, 3$$

El sistema (4) es entonces

$$\dot{C}_i = -\frac{i^2\pi^2}{2} C_i \qquad C_i(0) = \frac{(-1)^i\sqrt{2}}{i\pi}\left\{-1 + \frac{1}{\pi i} - \frac{2}{\pi^2 i^2}\right\}$$

$$= A_i\sqrt{2}$$

por lo que

$$C_i(t) = e^{-\frac{\pi^2 i^2}{2}t} C_i(0) \qquad i = 1, 2, 3$$

y entonces la trayectoria aproximada en el subespacio tridimensional generado por φ_1, φ_2, φ_3 está dada por

$$\tilde{u}(t,x) = \sum_{k=1}^{3} A_k e^{-k^2\pi^2 t} \operatorname{sen} k\pi x$$

Nótese cómo esta vez se han obtenido los modos de decaímiento con precisión absoluta y cómo las matrices obtenidas eran todas negativas. Esto no debe sorprender, ya que elegimos nuestra base de elementos finitos de manera idónea, al tomar las primeras tres funciones propias del problema. Esto, por supuesto, no puede hacerse en general. Sin embargo, bases estándar como la formada por elementos finitos lineales y por otros esplaines de orden mayor tienen la ventaja de que siempre conducirán a matrices M y A tridiagonales, con buenas propiedades de condicionamiento. En la práctica habría que tener ciertamente más de tres de estos elementos, y entonces la tridiagonalidad de las matrices resultará de la mayor utilidad. Examinaremos estos conceptos en la sección 3, precedida por la exposición del método general que haremos en la sección 2. Finalmente se tratará la convergencia y estabilidad de los métodos numéricos propuestos en la sección 3 (sección 4) para finalizar con la posible extensión al caso casi-lineal en la sección 5.

2. EL METODO

Consideremos el problema de valores iniciales y a la frontera (PVIF)

$$\frac{\partial u}{\partial t} + Au = f(x); t>0, 0<x<1 \qquad (5)'$$

$$u(0,x) = \bar{u}(x); \quad 0<x<1 \qquad (5)''$$

Aquí, A es un operador diferencial lineal de segundo orden, definido en un subconjunto denso del espacio C_o de todas las funciones continuas en [0,1] que se anulan en los extremos del intervalo [0,1]: es decir, $u \in C_o$ si y sólo si u es continua y $u(0)=0$, $u(1)=0$. Supondremos que $\bar{u}$. $f \in C_o$.

En C_o definamos el producto interior (,) como en (3) y supongamos que A es autoadjunto con respecto a él. Entonces A debe tener la forma

$$Au(x) = -\frac{d}{dx}\left[p(x)\frac{du}{dx}\right] + g(x)u \qquad (6)$$

[Courant-Hilbert, vol. I, p. 279}. Por conveniencia en las demostraciones que siguen, supondremos que los coeficientes del operador satisfacen la condición

H_1: p, q son acotadas y continuas por tramos en [0,1]

En virtud de lo general de H_1 resulta necesario precisar el sentido en que se va a entender la validez de (5), ya que H permite coeficientes para los que el segundo miembro de (6) no está definido en todo [0,1]. Para ello, supongamos por un momento que p es continuamente diferenciable por tramos, igual que $u:[0,\infty)x(0,1) \longrightarrow \mathbb{R}$, con $u(t,0) = 0 = u(t,1)$.

La función u deja un residuo en la ecuación igual a

$$\frac{\partial u}{\partial t} + Au - f \ ,$$

el que será ortogonal a todo $v \in C_o$ si y sólo si

$$\left(\frac{\partial u}{\partial t}, v\right) + (Au,v) = (f,v)$$

Aplicando integración por partes al producto (Au,v), restringiéndonos a v continuamente diferenciable por tramos, resulta que

$$(Au,v) = \int_o^1 \left\{p(x)\left(\frac{du}{dx}\right)^2 + q(x)u^2\right\}dx$$

Esto a su vez sugiere definir, para p y q como en H_1, la forma bilineal simétrica y positiva definida a dada por

$$a(u,v) := \int_o^1 [p(x)\dot{u}(x)\dot{v}(x) + q(x)u(x)v(x)]dx \qquad (7)$$

y definida para $u,v \in PC_o^{1,2}$. Este espacio consta de todas las funciones reales f continuas en [0,1], en donde son continuamente diferenciables por segmentos y $\|f'\| < \infty$, además de que $f(0) = f(1) = 0$.

En virtud de lo anterior, resulta natural la siguiente

Definición

Una solución (generalizada) del problema (5), con A como en (6) y bajo H_1, es una trayectoria $t \mapsto u(t,.) \in PC_o^{1,2}$ tal que

i) $\frac{\partial u}{\partial t}(t,.) \in PC_o^{1,2}$ para cada $t>0$

ii) $u(0,.) = \bar{u}$

iii) $\left(\frac{\partial u}{\partial t}, v\right) + a(u(t,.),v) = (f,v) \quad \forall\, v \in PC_o^{1,2}$,

con $a : PC_o^{1,2} \times PC_o^{1,2} \to \mathbb{R}$ definida en (7).

Se dirá que la forma bilineal a (.,.) es coercitiva si existe una solución generalizada $u \in PC^{2,2} \cap PC_o^{1,2}$ para cualquier $f \in PC^{0,2}$ y además,

$$\|u''\|_2 \leq \Gamma \|f\|_2$$

para $\Gamma > 0$ independiente de f. Las funciones en $PC^{k,p}$ tienen k-1 derivadas continua en [0;1], su derivada de orden k es continua por segmentos ahí y $\|f^{(k)}\| < \infty$.

Por conveniencia, definamos

$$a[u] := a(u,u) \ , \ u \in PC_o^{1,2} \qquad (8)$$

y por razones meramente técnicas, introduzcamos

H_2: El operador A es elíptico; es decir, hay constantes positivas $\gamma < \mu$ tales que

$$\gamma \|u'\|^2 \leq a[u] \leq \mu \|u'\|^2$$

donde $\| \ \|$ es la norma imducida por (3).

En vista de H_2 se dirá que el operador $\frac{\partial}{\partial t} + A$ es parabólico.

Igual que en la introducción, sean $\varphi_1, \ldots, \varphi_n \in PC_o^{1,2}$, linealmente independientes, y sea S_N el subespacio de $PC_o^{1,2}$ generado por ellas. Por ejemplo, puede tomarse una base de elementos finitos lineales, o de interpolación de Hermite o de esplaines cúbicos o de cualquier otro tipo. Una aproximación a la solución generalizada de (5) será una trayectoria $t \mapsto u_N(t,.) \in S_N$ y por lo tanto de la forma

$$u_N(t,x) = \sum_{j=1}^{N} C_j(t)\varphi_j(x) \qquad (9)$$

con $C_1, \ldots, C_N$ continuamente diferenciables. Para determinar una tal aproximación, fijémonos en el residuo que deja (9)

$$\sum_{j=1}^{N} \left\{ \dot{C}_j(t)\varphi_j + C_j(t)A\varphi_j \right\} - f$$

al cual impondremos el requisito de que sea ortogonal a todo S_N. A sabiendas de que

$$(A\varphi_j, \varphi_i) = a(\varphi_i, \varphi_j)$$

obtenemos que

$$\sum_{j=1}^{N} \left\{ \dot{C}_j(t)(\varphi_i, \varphi_j) + C_j(t)a(\varphi_i, \varphi_j) \right\} = (f, \varphi_i) \quad i = 1, \ldots, N$$

En $t = 0$, u_N de (9) deja el residuo

$$\sum_{j=1}^{N} C_j(0)\ \varphi_j - \bar{u}$$

el cual da lugar a

$$\sum_{j=1}^{N} (\varphi_i, \varphi_j) C_j(0) = (\bar{u}, \varphi_i) \quad , \quad 1 = 1, \cdots, N$$

si va a ser ortogonal a todo S_N.

Definamos

$$a_{ij}^{(N)} := a(\varphi_i, \varphi_j) \quad , \quad m_{ij}^{(N)} := (\varphi_i, \varphi_j)$$

$$f_i^{(N)} := (f, \varphi_i) \quad , \quad \bar{u}_i^{(N)} := (\bar{u}, \varphi_i)$$

y sean A_N, M_N, b_N las matrices

$$A_N := \|a_{ij}^{(N)}\|, \quad M_N := \|m_{ij}^{(N)}\|, \quad f_N := \|f_i^{(N)}\|. \quad \bar{u}_N := \|\bar{u}_i^{(N)}\|$$

las condiciones que hemos impuesto a u_N de (9) determinan a $t \mapsto (C_1(t), \ldots, C_N(t))$ como solución del problema de valores iniciales (PVI)

$$M_N \dot{C} + A_N C = f_N \qquad (10)'$$

$$M_N C(0) = \bar{u}_N \qquad (10)''$$

Es claro que A_N y M_N son simétricas, cualquiera que sea la base elegida, simplemente porque $(.,.)$ y a son simétricas. Además la positividad definida de estas dos formas bilineales la heredan también las dos matrices. En particular, M_N es invertible, por lo que (10) tiene solución única [Coddington-Levinson,Cap. 3] dada por la fórmula de variación de parámetros

$$C(t) = e^{-tM_N^{-1}A_N} M_N^{-1} \bar{u}_N + \int_0^t e^{-(t-s)M_N^{-1}A_N} M_N^{-1} f_N ds \qquad (11)$$

Entonces la aproximación u_N está bien definida.

La fórmula anterior muestra varios aspectos. Primero, el mismo método puede aplicarse sin ninguna complicación adicional al caso en que f dependa de t en (5)': La solución (11) se cumple con $f_N(s)$ en lugar de f_N. También puede aplicarse el método con A dependiente de t en (5)': resulta $A_N(t)$ en lugar de A_N en (10), pero en este caso no se cuenta con una representación explícita como lo es (11). Sin embargo, queda todavía abierta la posibilidad de resolver (10) por algún método numérico. De hecho resulta más

práctico este enfoque aún en el caso en que A y f no dependan de t y a continuación nos referiremos a ello.

3. ASPECTOS COMPUTACIONALES

Es muy conveniente la simetría de las matrices en (10), por lo que trataremos de no perderla al despejar la derivada. Para simplificar la notación, escribamos A, M, f y $\bar{u}$ en lugar de A_N, M_N, f_N y $\bar{u}_N$ en (10). Por la positividad definida de M y puesto que es simétrica, tiene una raíz cuadrada simétrica positiva definida y sólo una [Mal'cev p. 122]. Si definimos

$$\gamma(t) := M^{1/2}C(t) ,$$

esta función vectorial satisface el PVI

$$\dot{\gamma}=-E\gamma+C, \gamma(0)=\gamma_0 \qquad (12)$$

donde hemos puesto

$$E := M^{1/2}AM^{-1/2}$$

$$C := M^{-1/2}f \quad , \quad \gamma_0 := M^{-1/2}\bar{u}$$

Nótese que E en (12) es positiva definida y simétrica.

Una solución aproximada de (12) será una sucesión

$$(0,\gamma_0), \quad (t_1,\gamma_1), \quad (t_2\gamma_2), \quad \ldots \qquad (13)$$

en $(0,\infty) \times \mathbb{R}^N$. La bondad de la aproximación se juzgará en términos de las propiedades de la sucesión de errores.

$$|\gamma(t_j) - \gamma_j| \ , \ j \geq 1 \qquad (14)$$

Si esta sucesión es acotada se dirá que el método usado para generar (13) es <u>estable</u> Si cada elemento de (14) tiende a cero siempre que

$$\sup_{j\geq 1} (t_{j+1} - t_j) \to 0$$

se dirá que el método en cuestión <u>converge</u>. Veamos ahora algunos métodos iterativos para generar (13), en todos los cuales se

tendrá

$$t_{j+1} = t_j + h, \quad t_o = 0$$

En el método de Euler reemplazamos $\gamma(t_j)$ por $(\gamma_{j+1} - \gamma_j)/h$ en (12), lo cual da lugar a

$$\gamma_{j+1} = (I - h\,E)\gamma_j + h\,C \tag{15}$$

Una variante de este método pide remplazar $\gamma(t_j)$ por $(\gamma_j - \gamma_{j-1})/h$ y da lugar a

$$(I + h\,E)\,\gamma_{j+1} = \gamma_j + h\,C$$

o sea

$$\gamma_{j+1} = (I + h\,E)^{-1}\,\gamma_j + h\,(I + h\,E)^{-1}C\,, \tag{16}$$

el método de Euler implícito. Es bien sabido que este segundo método es siempre estable en tanto que (15) requiere limitar h [Ferziger, Cap. 3]. En efecto, una ecuación en diferencias como

$$\gamma_{j+1} = P\,\gamma_j + q\,,\ \gamma_o \text{ dado}$$

tiene como solución a

$$\gamma_j = P^j\gamma_o + \left(\sum_{k=0}^{j-1} P^k\right)q$$

Una condición inicial errónea $\tilde{\gamma}_o$ produce una sucesión de errores dada por

$$\gamma_j - \tilde{\gamma}_j = P^j\,(\gamma_o - \tilde{\gamma}_o)$$

Así pues, los errores iniciales no se amplifican si y sólo si

$$|\lambda| \leq 1 \quad \forall \quad \lambda \in \text{Espec}(P) \tag{E}$$

y de hecho se amortiguan si y sólo si la desigualdad es estricta.

Para (15),

$$\text{Espec}(P) = \{1 - \lambda h: \ \lambda \in \text{Espec}(E)\}$$

y entonces se tendrá estabilidad si y sólo si

$$0 < h < \frac{2}{\lambda_{max}}, \text{ con } \lambda_{max} = \max \text{Espec}(E)$$

En cambio, para (16)

$$\text{Espec}(P) = \left\{ \frac{1}{1 + \lambda h} : \lambda \in \text{Espec}(E) \right\}$$

y (E) se cumple de manera automática para cualquier $h>0$. De hecho se tendrá la desigualdad estricta, todo esto en virtud de la positividad definida y simetría de E.

Estos dos métodos se pueden sumergir en una familia continua de métodos, indicada por $\alpha \in [0,1]$, si ponemos

$$\frac{\gamma_{j+1} - \gamma_j}{h} = E\left\{(1 - \alpha)\gamma_j + \alpha\, \gamma_{j+1}\right\} + C$$

Entonces (15) y (16) son los extremos inicial ($\alpha = 0$) y final ($\alpha = 1$) de esta familia. El punto medio ($\alpha = 1/2$) resulta de interés: es el método de Crank - Nicholson

$$(I + \frac{h}{2} E)\, \gamma_{j+1} = (I - \frac{h}{2} E)\, \gamma_j + h\, C\ ,$$

es decir,

$$\gamma_{j+1} = (I + \frac{h}{2} E)^{-1}(I - \frac{h}{2} E)\gamma_j + h\, (I + \frac{h}{2} E)^{-1} C \qquad (17)$$

Más adelante nos ocuparemos de la estabilidad de este modelo acoplado con el de Galerkin.

Por otro lado, de (12)

$$\gamma(t_{j+1}) = e^{-hE}\gamma(t_j) + \int_{t_j}^{t_{j+1}} e^{-(t_{j+1}-s)E} c\, ds$$

que, si aplicamos la regla de integración numérica

$$\int_a^b f(s)ds \simeq f(b)(b-a)$$

se reduce a

$$\gamma(t_{j+1}) \simeq e^{-hE}\gamma(t_j) + hC \qquad (19)$$

Comparando con (15) y recordando que

$$e^{-hE} = I - hE + o(h)$$

vemos que el método de Euler equivale a aproximar e^{-z} mediante el polinomio $1-z$. Análogamente se puede verificar que (16) se obtiene de (19) aproximando e^{-z} mediante $(1 + z)^{-1}$, en tanto que (17) proviene de usar $(1+\frac{1}{2}z)^{-1}(1-\frac{1}{2}z)$ en lugar de e^{-z}. Esto nos lleva al problema de obtener aproximaciones racionales de e^{-z} y con ello otros tantos métodos para generar (13). Podemos pensar en las aproximaciones de Padé, que son funciones racionales

$$R_{n,m}(z) = \frac{a_0+a_1z + \ldots + a_nz^n}{1 + b_1z + \ldots + b_nz^m}$$

cuyos n+m+1 coeficientes podemos determinar exigiendo que

$$e^{-z} - R_{n,m}(z) = O(z^{m+n+1})$$

-[Bender - Orzag, Cap. 8]. En otras palabras, la serie de $R_{n,m}(z)$ debe coincidir con

$$e^{-z} = \sum_{j=0}^{m+n} \frac{(-z)^j}{j!} + \ldots$$

coeficiente por coeficiente hasta el de z^{m+n}. Así pues, si

$$R_{n,m}(z) = \sum_{j=0}^{m+n} C_j z^j + \ldots$$

debe cumplirse que

$$j!C_j = (-1)^j \qquad j = 0, 1, \ldots, m + n$$

Aplicando este criterio se obtiene fácilmente que

$$R_{1,0}(z) = 1 - z \ , \ R_{0,1}(z) = \frac{1}{1+z} \ , \ R_{1,1}(z) = \frac{1-z/2}{1+z/2}$$

Podemos utilizar las diferentes aproximaciones $R_{n,m}(hE)$ a e^{-hE} y proponer otros tantos métodos si observamos que la solución de

$$\dot{\gamma} = -E\,\gamma + C \qquad \gamma(t_j) \text{ dado}$$

es, para $t = t_{j+1}$

$$\gamma(t_{j+1}) = E^{-1}C + e^{-hE}\,(\gamma(t_j) - E^{-1}C)$$

Esto sugiere proponer

$$\gamma_{j+1} = E^{-1}C + R_{n,m}(hE)\,(\gamma_j - E^{-1}C) \qquad (20)$$

como método iterativo para generar (13). Fácilmente se ve cómo (20) se especializa a (15), (16) y (17) para los valores adecuados de m y n.

Enseguida veremos cómo analizar la estabilidad de (20) acoplado con el método de Galerkin para (5).

4. ESTABILIDAD Y CONVERGENCIA

En lo que sigue, sea $\{\gamma_j(n,m)\}_j = 0, 1, \ldots$ la sucesión (13) obtenida al aplicar el método iterativo (20). Entonces, pongamos

$C^{n,m}(j)$: = vector de coeficientes de la aproximación a $u_N(t_j,.)$

por lo que

$$C^{n,m}(j) = M_N^{-1/2}\,\gamma_j(n,m)$$

y

$$u_N^{n,m}(t_j,.): = \sum_{i=1}^{N} C^{n,m}(j)\varphi_i$$

da la aproximación a $u_n(t,.)$.

En lo que sigue, supondremos que hemos normalizado la base $\varphi_1,\ldots,\varphi_n$ de tal manera que $\sqrt{N}\ \|\varphi_i\| = 1, i = 1,\ldots,N$. Sean $0<\lambda_{min}^{(N)}<\lambda_{max}^{(N)}$ los valores propios extremos de M_N, por lo que se tendrá que

$$\lambda_{min}^{(N)}\ |X|^2 \le X^T M_N X \le \lambda_{max}^{(N)}\ |X|^2\ ,\ X \in \mathbb{R}^N \qquad (21)'$$

donde $|X|: = \sqrt{X^T X}$. Nótese que

$$m_{ii}^{(N)} = \|\varphi_i\|^2 = \frac{1}{N}$$

por lo que

$$\text{tr } M_N = \sum \{\lambda:\lambda\in \text{Espec } M_N\} = 1$$

y entonces

$$0 < \lambda_{min}^{(N)} < \lambda_{max}^{(N)} < 1$$

Finalmente, pongamos

$$|A| := \sup_{x\neq 0} \frac{|Ax|}{|x|}$$

para toda matriz A de orden NxN, con lo cual se tendrá que

$$|Ax| \leq |A||x|$$

$$|AB| \leq |A||B|$$

Nótese que

$$|A|^2 = \sup \frac{x^T A^T A x}{x^T x} = \max \text{ Espec } A^T A$$

de tal manera que, si A es simétrica, entonces

$$|A| = \max \{|\lambda|: \lambda \in \text{Espec}(A)\} = :\rho(A) \qquad (22)$$

y |A| coincide con el radio espectral de A.

Finalmente, sea γ el coeficiente de elipticidad de A, es decir tal que

$$a[u] \geq \gamma \|u'\|^2$$

y pongamos $\text{cond}(M_N) := \lambda_{max}^{(N)} / \lambda_{min}^{(N)}$.

Se tiene entonces el siguiente

Teorema

Bajo H_1, H_2 y en la notación anterior, la aproximación que a la solución de (5) da el método de Galerkin acoplado con (20) satisface

$$\|u_N^{n,m}(t_j, .)\| \leq \alpha_j^{(N)} \|f\| + \beta_j^{(N)} \|\bar{u}\| \ , \ j = 0, 1, 2, \ldots$$

donde

$$\alpha_j^{(N)} := \frac{\text{cond } (M_N)}{\gamma \pi^2 \sqrt{\lambda_{min}^{(N)}}} \left[(1 + |R_{nm}^j (hE)| \right]$$

$$\beta_j^{(N)} := \frac{\text{cond } (M_N)}{\sqrt{\lambda_{min}^{(N)}}} |R_{nm}^j (hE)|$$

Demostración

De

$$\gamma_{j+1} - E^{-1}C = R_{n,m}(hE)[\gamma_j - E^{-1}C] \quad j = 0, 1, \ldots$$

se llega fácilmente a que

$$\gamma_j - E^{-1}C = R^{j}_{n,m}(hE)\ (\gamma_o - E^{-1}C)$$

de donde

$$|\gamma_j| \leq |E^{-1}C| + |R^{j}_{n,m}(hE)|\ |\gamma_o - E^{-1}C|$$

Por otro lado,

$$\|u_N^{n,m}(t_j,.)\|^2 = \int_o^1 \sum_{i,k} C_i^{n,m}(j) C_k^{n,m}(j) \varphi_i(x) \varphi_j(x) dx$$

$$= \sum_{i,k} C_i^{n,m}(j) C_k^{n,m}(j) (\varphi_i, \varphi_j)$$

$$= C^{n,m}(j)^T M_N C^{n,m}(j)$$

$$\leq \lambda_{max}^{(N)} |C^{n,m}(j)|^2$$

$$\leq \lambda_{max}^{(N)} |M_N^{-1/2}|\ |\gamma_j(n,m)|^2$$

y entonces

$$\|u_N^{n,m}(t_j,.)\| \leq \sqrt{\lambda_{max}^{(N)}}\ |M_N^{1/2}|\ \left\{ |E^{-1}C| + |R^{j}_{nm}(hE)|\,|\gamma_o - E^{-1}C| \right\}$$

De las definiciones de γ_o, E y C, se obtiene fácilmente que

$$|E^{-1}C| \leq |M_N^{1/2}|\ |A_N^{-1} f_N|$$

$$|\gamma_0 - E^{-1}C| \leq |M_N^{1/2}| \, |M_N^{-1}\bar{u} - A_N^{-1}f_N|$$

$$\leq |M_N^{1/2}| \left(|M_N^{-1}\bar{u}_N| + |A_N^{-1}f_N| \right)$$

Para estimar $|M_N^{-1}\bar{u}|$ y $|A_N^{-1}f_N|$, procedamos como sigue:

Nótese que

$$a[\varphi] \geq \gamma \, \|\varphi'\|^2 \geq \gamma\pi^2 \, \|\varphi\|^2$$

en virtud de la elipticidad de A; por la desigualdad de Rayleigh-Ritz,

$$\pi^2 \|\varphi\|^2 \leq \|\varphi'\|^2$$

(véase el Teorema 1.2 de [Schultz (1973)]),esto para $\varphi \in PC_0^{1,2}$ arbitraria.

Tomando

$$\varphi = \sum_{i=1}^{N} x_i \varphi_i$$

resulta que

$$x^T A_N x \geq \gamma\pi^2 x^T M_N x$$

Finalmente, de (21)

$$|x|^2 \geq \frac{1}{\sigma_N} \, x^T A_N x \quad , \quad x \in \mathbb{R}^N$$

con

$$\sigma_N := \gamma \, \pi^2 \lambda_{min}^{(N)}$$

Asi pues

$$\text{Espec } (A_N) \subset (\sigma_N, \infty)$$

$$\text{Espec } (A_N^{-2}) \subset (0, \sigma_N^{-2})$$

y entonces

$$|A_N^{-1} f_N|^2 = f_N \cdot A_N^{-2} f_N \leq \sigma_N^{-2} |f_N|^2$$

Análogamente, se obtiene que

$$|M_N^{-1} \bar{u}_N|^2 = \bar{u}_N \cdot M_N^{-2} \bar{u}_N \leq \gamma^2 \pi^4 \sigma_N^{-2} |\bar{u}_N|^2$$

Por otro lado,

$$|f_N|^2 = \sum_{i=1}^{N} (f, \varphi_i)^2 \leq \|f\|^2 \sum_{i=1}^{N} \|\varphi_i\|^2 = \|f\|^2$$

y análogamente para $\bar{u}_N$, por lo que

$$|f_N| \leq \|f\| \quad , \quad |\bar{u}_N| \leq \|\bar{u}\|$$

Por lo tanto,

$$|E^{-1} C| \leq |M_N^{1/2}| \sigma_N^{-1} \|f\|$$

$$|\gamma_o - E^{-1} C| \leq |M_N^{1/2}| \sigma_N^{-1} \left[\gamma \|\bar{u}| + \|f\| \right]$$

Basta observar que, de (22)

$$|M_N^{-1/2}| \, |M_N^{1/2}| = \sqrt{\text{cond}(M_N)} = \sqrt{\frac{\lambda_{max}^{(N)}}{\lambda_{min}^{(N)}}}$$

para llegar a que

$$\|u_N^{n,m}(t_j, \cdot)\| \leq$$

$$\leq \frac{\text{cond}(M_N)}{\gamma \pi^2 \sqrt{\lambda_{(min)}^{(N)}}} \left\{ \|f\| + |R_{nm}^{j}(hE)| \cdot \left[\|f\| + \gamma \pi^2 \|\bar{u}\| \right] \right\}$$

y con ello al resultado deseado. ∎

Recordemos que E es positiva-definida y simétrica, por lo que

$$\text{Espec}(E) \subset (0, \infty)$$

y sea

$$E = \sum\{\lambda P_\lambda : \lambda \in Espec(E)\}$$

su descomposición espectral. Entonces,

$$R_{nm}(hE) = \sum_{\lambda \in Espec(E)} R_{nm}(h\lambda) P_\lambda$$

y el radio espectral de $R_{nm}(hE)$ está dado por

$$\rho\left[R_{nm}(hE)\right] = \sup_{\lambda \in Espec(E)} R_{nm}(h\lambda)$$

Sea

$$\tau_{nm} := \sup \{t \geq 0 : R_{nm}(Z) \leq 1,\ 0 \leq Z \leq t\} \qquad (23)'$$

y tomemos

$$0 < h < \tau_{nm} / |\lambda(E)|_{min} \qquad (23)''$$

Entonces

$$R_{nm}(h|\lambda(E)|_{min}) \leq 1$$

y, <u>suponiendo que</u> R_{nm} <u>es decreciente en</u> Espec(E), se obtiene entonces que

$$\rho[R_{nm}(hE)] \leq R_{nm}\left(\frac{h}{|\lambda(E)|_{min}}\right)$$

Claramente se llega entonces al siguiente

<u>Corolario</u>

Sea h como en (23). Por otro lado, sean $\mathcal{B}$ una colección de bases finitas $b := \{\varphi_i\}_{i=1}^{\#b} \subset PC^{1,2}$ y $\Lambda > 0$ tales que

$$0 < \frac{cond(M_{\#b})}{\sqrt{\lambda_{min}^{(\#b)}}} \leq \Lambda \quad , \quad b \in \mathcal{B}$$

Entonces, si $b \in \mathcal{B}$, $N = \#b$,

$$\|u_N^{n,m}(t_j,.)\| \leq \Lambda \left(\|\bar{u}\| + \frac{2}{\gamma}\|f\|\right), \quad j \geq 0$$

Consideremos por un momento el caso $f = 0$ y supongamos que se usa una condición inicial $\tilde{u}$ en lugar de $\bar{u}$. Por linealidad, el error $u_N - \tilde{u}_N$ corresponde a la condición inicial $u - \tilde{u}$ y entonces

$$\sup_{j\geq 0} \|u_N^{n,m}(t_j,.) - \tilde{u}_N^{n,m}(t_j,.)\| \leq \Lambda \|\bar{u} - \tilde{u}\|$$

Análogamente, en el caso $\bar{u} = 0$, un error en el término forzante (si se usa $\tilde{f}$ en lugar de f) da lugar a un error $U_N - \tilde{u}_N$ en la solución. Por linealidad y aplicando el corolario anterior,

$$\sup_{j\geq 0} \|u_N^{n,m}(t_j,.) - \tilde{u}_N^{n,m}(t_j,.)\| \leq \frac{\Lambda}{\gamma}\|f - \tilde{f}\|$$

Así pues, los errores no se amplifican en ninguno de los dos casos, por lo que el método iterativo (20) es estable.

Supongamos ahora que $m \geq n$ en la aproximación de Padé $R_{nm}(z)$ a e^{-z} (es decir, el grado del numerador no excede al grado del denominador). Entonces,

$$R_{nm}(z) = \frac{a_0 + a_1 z + \ldots + a_m z^m}{1 + b_1 z + \ldots + b_m z^m}, \quad b_m \neq 0$$

y se tiene

$$\lim_{z \mapsto \infty} R_{nm}(z) = \frac{a_m}{b_m}$$

que puede ser cero (si $n < m$). Si este límite es menor o igual que 1, entonces $\tau_{nm} = \infty$ por lo que h no está restringida en ningún modo. Así pues

Corolario

Si $m \geq n$ y si $a_m/b_m \leq 1$ entonces (20) es incondicionalmente estable.

En particular, el método Galerkin - Euler implícito y el Galerkin-Crank - Nicholson son incondicionalmente estables. Investiguemos ahora la convergencia de (20), limitándonos al caso m = n = 1, el método de Galerkin - Crank - Nicholson.

Debemos estimar la diferencia $u - u_N^{1,1}$, con objeto de poder valernos de resultados previos acerca del método de Galerkin; sea $\varphi \mapsto \varphi_N$ la proyección ortogonal de $\varphi \in PC^{1,2}$ sobre el subespacio generado por $\varphi_1, \ldots, \varphi_N$, tomada con respecto al producto interior $(\varphi,\psi) \mapsto a(\varphi,\psi)$. De los teoremas 7.19, 7.20 y 7.21 de [Schultz (1973)], sabemos que (para A adecuada en cada caso)

$$\|u(t_j,.)-u_N^a(t_j,.)\| \leq Ah^m \quad (24)$$

donde

$$m = \begin{cases} 2 \text{ si } \{\varphi_i\} \text{ consta de elementos finitos lineales} \\ 4 \text{ si } \{\varphi_i\} \text{ consta de elementos de Hermite} \\ \quad \text{o esplaines cúbicos,} \end{cases}$$

y u_N^a es la aproximación respectiva, esto siempre y cuando la forma a(.,.) sea coercitiva.

Para j = 0, 1, 2, ... sean

$$\delta_j := u(t_j,.) - u_N^a(t_j,.)$$

$$\varepsilon_j := u_N^{1,1}(t_j,.) - u_N^a(t_j,.)$$

de tal manera que por la desigualdad del triángulo

$$\|u(t_j,.) - u_N^{1,1}(t_j,.)\| \leq \|\delta_j\| + \|\varepsilon_j\| \quad (25)$$

Tenemos el siguiente resultado, en el que

$$D_+ z_j := \frac{z_{j+1} - z_j}{h}.$$

Lema

Supongamos que (5) tiene una solución $u \in C^3([0,\infty)x[0,1])$.

Entonces,

$$D_+\|\varepsilon_j\|^2 \leq$$

$$\leq \quad \|D_+\delta_j\|^2 + \frac{h^2}{24} \sup_{t\leq(j+1)h} \|\frac{\partial^3 u}{\partial t^3}(t,.)\| \Big\}$$

Demostración

Por definición

$$D_+\delta_j = \frac{1}{h}\,[D_+u(t_j,.)] - D_+u_N^a(t_j,.)$$

y, por el teorema de Taylor

$$u(t_{j+1/2},.) = u(t_j,.) + \frac{\partial u}{\partial t}(t_j,.)\frac{h}{2} + \frac{h^2}{8}\frac{\partial u}{\partial t^2}(t_j,.) + \frac{h^3}{48}\frac{\partial^3 u}{\partial t^3}(t_j+\theta\frac{h}{2},.)$$

con análoga expresión en torno a t_{j+1}, $0 < \theta < 1$. De aquí resulta

$$D_+u(t_j,.) = \frac{1}{2}\left[\frac{\partial u}{\partial t}(t_j,.) + \frac{\partial u}{\partial t}(t_{j+1},.)\right] + \frac{h^2}{24}e_j$$

donde

$$e_j := \frac{\frac{\partial^3 u}{\partial t^3}(t_j+\theta_{+2}\frac{h}{2},.) + \frac{\partial^3 u}{\partial t^3}(t_{j+1}-\theta_{-2}\frac{h}{2},.)}{2}$$

Así pues, $\|e_j\| \leq \sup_{t\leq(j+1)h} \|\frac{\partial^3 u}{\partial^3 t}\|$, y

$$(D_+\delta_j,w) = \left[\frac{1}{2}\left[\frac{\partial u}{\partial t}(t_j,.) + \frac{\partial u}{\partial t}(t_{j+1},.)\right],\ w\right]$$

$$+ \frac{h^2}{24}(e_j,w) - (D_+u_N^a(t_j,.),w)$$

para cualquier w que sea combinación lineal de $\varphi_1, \ldots, \varphi_N$, es decir, que esté en el subespacio generado por ellas, al que denotamos por $[\varphi_1, \ldots, \varphi_N]$.

Por otro lado, como u es solución de (5)

$$\left(\frac{\partial u}{\partial t}(t_j,.),w\right) = (f,\ w) - a(u(t,.),w)$$

y, por definición del método de Crank - Nicholson acoplado con Galerkin,

$$(D_+u_N^{1,1}(t_j,.),\ w) + a\left(\frac{u_N^{1,1}(t_j,.) + u_N^{1,1}(t_{j+1},.)}{2},w\right) = (f,\ w)$$

Combinando estas últimas tres relaciones, con $t = t_j$ y $t = t_{j+1}$, se obtiene que

$$(D_+\delta_j,w) = a\left(\frac{u_N^{1,1}(t_j,.) + u_N^{1,1}(t_{j+1},.)}{2},\ w\right)$$

$$-a\left(\frac{u(t_j,.) + u(t_{j+1},.)}{2}\ w\right)+\frac{h^2}{24}(e_j,\ w)$$

$$+ (D_+u_N^{1,1}(t_j,.),w) - (D_+u_\alpha(t_j,.),\ w)$$

Dado que

$$a(u(t,.),w) = a(u_N^\alpha(t,.),w)$$

para cualquier $w \in [\varphi_1, \ldots, \varphi_N]$, resulta entonces que

$$(D_+\delta_j,w) = \frac{1}{2}a(\varepsilon_j + e_j,\ w)$$

$$+ (D_+\varepsilon_j,w) + \frac{h^2}{24}(\varepsilon_j,w)$$

Tomemos $w = \varepsilon_j+\varepsilon_{j+1}$ y recordemos la desigualdad

$$a,\ b \in \mathbb{R},\ \eta > 0 \quad \Rightarrow \quad ab \geq -\eta^{-1}a^2 - \frac{1}{4}\eta b^2 \qquad (26)$$

Fácilmente se verifica que

$$a[\varepsilon_j+\varepsilon_{j+1}] \geq \gamma\pi^2\|\varepsilon_j+\varepsilon_{j+1}\|^2$$

$$(D_+\varepsilon_j,\ \varepsilon_j+\varepsilon_{j+1}) = D_+\|\varepsilon_j\|^2$$

Para $\eta = \gamma\pi^2$,

$$(\varepsilon_j,\ \varepsilon_j + \varepsilon_{j+1}) \geq -\frac{1}{\gamma\pi^2}\|\varepsilon_j\|^2 - \frac{\gamma\pi^2}{4}\|\varepsilon_j+\varepsilon_{j+1}\|^2$$

Combinando estas últimas relaciones, se llega a que

$$D_+\|\varepsilon_j\|^2 \leq (D_+\delta_j,\ \varepsilon_j+\varepsilon_{j+1}) + \frac{(h^2/24)}{\gamma\pi^2}\|\varepsilon_j\|^2 - \frac{\gamma\pi^2}{4}\|\varepsilon_j+\varepsilon_{j+1}\|^2$$

Por otro lado, aplicando (26) con $\eta = \gamma\pi^2$, se verifica que

$$-(D_+\delta_j,\ \varepsilon_j+\varepsilon_{j+1}) \geq -\frac{1}{\gamma\pi^2}\|D_+\delta_j\|^2 - \frac{\gamma\pi^2}{4}\|\varepsilon_j+\varepsilon_j\|^2$$

por lo que

$$D_+\|\varepsilon_j\|^2 \leq \frac{1}{\gamma\pi^2}\left\{\|D_+\delta_j\|^2 + \frac{h^2}{24}\|e_j\|^2\right\}$$

y de ahí el resultado. ‖

Para probar (26), basta tomar $\lambda,\mu > 0$ con $\lambda\,\mu = 1$ en

$$(\lambda\ a + \mu\ b)^2 \geq 0$$

y luego restringirse al caso $\lambda = \sqrt{\frac{2}{\eta}} = \frac{1}{\mu}$.

Lema

Bajo las mismas hipótesis del lema anterior,

$$\|\varepsilon_j\|^2 - \|\varepsilon_o\|^2 \leq$$

$$\leq \frac{t_j}{\gamma\pi^2}\ \sup_{k\leq j}\left\{\|D_+\delta_k\|^2 + \sup_{t\leq(K+1)h}\left\|\frac{\partial^3 u}{\partial t^3}\right\|\right.$$

Demostración

Basta observar que

$$D_+\|\varepsilon_j\|^2 \leq \Delta_j := \frac{1}{\gamma\pi^2}\ \sup_{k\leq j}\left\{\|D_+\delta_h\| + \sup_{t\leq(K+1)\ h}\left\|\frac{\partial^3 u}{\partial t^3}\right\|\right\}$$

luego reescribirla en la forma

$$\|\varepsilon_{j+i}\|^2 - \|\varepsilon_j\|^2 \leq h\ \Delta_i$$

para $i \leq j-1$ y sumar, para tener que

$$\|\varepsilon_j\|^2 - \|\varepsilon_o\|^2 \leq h \sum_{i<j} \Delta_i \leq j\ h\ \Delta_j$$

y así lograr el resultado deseado. ‖

Combinando los dos lemas anteriores con (25), se obtiene que

$$\|u(t_j,.) - u_N^{1,1}(t_j,.)\| \leq \|u(t_j,.) - u_N^{\alpha}(t_j,.)\| \qquad (27)$$

$$+ \frac{t_j}{\gamma\pi^2} \sup_{k\leq j} \left(\|D_+(u(t_k,.) - u_N^{\alpha}(t_k))\|^2 + \frac{h^2}{24} \sup_{t\leq(k+1)h} \|\frac{\partial^3 u}{\partial t^3}\| \right)$$

$$+ \|u_N^{1,1}(0,\circ) - u_N^{\alpha}(0,\circ)\|$$

Finalmente, combinando (24) y (27) llegamos al importante resultado que damos a continuación, acerca de la convergencia del método de Crank - Nicholson - Galerkin para los diferentes elementos finitos.

Teorema. Si a(.,.) es coercitiva, entonces

$$\|u(t_j,.) - u^{1,1}(t_j,.)\| = 0(h^2 + k^m)$$

donde m = 2 si los elementos finitos son lineales, m = 4 si son parte de la base de Hermite o esplaines cúbicos. Aquí

$$k = \max_{1\leq i\leq N} \Delta x_i.$$

5. EXTENSIONES AL CASO SEMILINEAL

Sea un tubo esbelto de longitud ℓ, que supondremos lleno de una solución de concentración C. Supongamos que el soluto se descompone bajo condiciones isotérmicas según la cinética conocida

r(c) := tasa de desaparición de soluto por unidad de longitud por concepto de la reacción

Entonces

$$\frac{\partial C}{\partial t} = D\,\frac{\partial^2 C}{\partial z^2} - r(c);\quad t>0,\ 0<z<\ell$$

donde D es el coeficiente de difusión del soluto y c(t,z) := concentración de soluto en el punto z, t segundos después de iniciada la operación.

Supongamos que

$$C(0,z) = C_o(z)$$

Supongamos que el extremo izquierdo del tubo se mantiene artificialmente a una alta concentración C_v(constante) y que el tubo es suficientemente largo como para que los cambios de concentración no se dejen sentir en el extremo derecho. Entonces

$$C(t,0) = C_v \quad , \quad C(t,\ell) = C_o(\ell)$$

Sea λ la función lineal que satisface

$$\lambda(0) = C_v \quad , \quad \lambda(\ell) = C_o(\ell)$$

y pongamos

$$u(t,x) := C(t,\ell x) - \lambda(\ell x)$$

Entonces, n satisface el PVIF

$$\frac{\partial u}{\partial t} = \frac{D}{\ell^2}\frac{\partial^2 u}{\partial x^2} - r(\lambda(\ell x)+u) \quad ; \quad t>0,\ 0<X<\ell \qquad (28)'$$

$$u(t,0) = 0 \qquad u(t,\ell) = 0 \qquad (28)''$$

$$u(0,x) = \bar{u}(x) \qquad (28)'''$$

donde

$$\bar{u}(x) := C_o(\ell x) - \lambda(\ell x)$$

Este PVIF es caso particular de la siguiente situación:

"Sea A un operador diferencial lineal de segundo orden sobre $PC_o^{1,2}$ autoadjunto, como en (5), (6). Sea F un operador no lineal sobre el mismo espacio, dado por

$$F(u)(x) = f(x,u(x)) \qquad u \in PC_o^{1,2}"$$

"Encuentre una trayectoria $t \mapsto u(t,.) \in PC_o^{1,2}$ tal que

$$\frac{\partial u}{\partial t} + Au = F(u) \tag{29}$$

$$u(0,.) = \bar{u}$$

donde $\bar{u} \in PC_o^{1,2}$ está dado".

En efecto, en (28)

$$p(x) = \frac{D}{\ell^2} \quad , \quad q(x) = 0 \quad , \quad f(x,u) = -r(\lambda/\ell x)+u)$$

El PVIF recién enunciado se llamará <u>semilineal.</u>

Así como para el caso lineal, conviene precisar el sentido en que se entenderá la ecuación diferencial (29). Para ello y para cada $u \in PC_o^{1,2}$ sea R(u) la funcional lineal "residuo que deja u", definida sobre $PC_o^{1,2}$ mediante

$$R(u) = \left(\frac{\partial u}{\partial t},.\right) + a(u,.) - (F(u),.) \tag{30}$$

donde

$$a(u,v) = (Au,v)$$

igual que antes.

Entonces, u es solución de (29) si

$$R(u) = 0$$

Así pues, podemos dar la siguiente

Definición

Una solución (generalizada) del PVIF para (29), con A como en (6) y bajo H_1, es una trayectoria $t \mapsto u(t,.) \in PC_o^{1,2}$ tal que

(i) $\frac{\partial u}{\partial t}(t,.) \in PC_o^{1,2}$ para cada $t>0$

(ii) $u(0,.) = \bar{u}$

(iii) $\left(\frac{\partial u}{\partial t}, v\right) + a(u(t,.), v) = (F(u(t,.), v) \quad \forall v \in PC_o^{1,2}$

con $a : PC_o^{1,2} \times PC_o^{1,2} \mapsto \mathbb{R}$ definida en (7).

Igual que en el caso lineal, definamos aproximaciones a la solución del problema anterior de la manera siguiente. Sean $\varphi_1, \ldots, \varphi_N \in PC_o^{1,2}$ funciones linealmente independientes y sea S_N el subespacio que generan. Una trayectoria en S_n está dada por

$$t \mapsto \sum_{j=1}^{N} C_j(t)\, \varphi_j =: u_N(t,.) \qquad (31)$$

con $C_1, \ldots C_N$ continuamente diferenciables. Podamos calcular el residuo $R(u_N(t,.))$ que deja esta trayectoria en el instante t, simplemente de (30), mismo que será una funcional lineal definida en todo $PC^{1,2}$. Diremos que (31) es una solución aproximada del PVIF en cuestión si

$$R(u_N(t,.))\Big|_{S_N} = 0 \quad , \quad t>0 \qquad (32)$$

De manera más explícita, el residuo en el instante t está dado por

$$\sum_{j=1}^{N} \dot{C}_j(t)(\varphi_j,.) + \sum_{j=1}^{N} C_j(t)a(\varphi_j,.) -$$

$$- \left(F\left(\sum_{j=1}^{N} C_j(t)\varphi_j\right),. \right)$$

y la condición (32) se logra si

$$\sum_{j=1}^{N} (\varphi_i, \varphi_j)\, \dot{C}_j(t) + \sum_{j=1}^{N} a(\varphi_i, \varphi_j)\, C_j(t) =$$

$$= (F(\sum_{j=1}^{N} C_j(t)\varphi_j), \varphi_i) \qquad i = 1, \ldots N$$

Es decir, (31) será una solución aproximada de (29) si $t \mapsto C(t) := (C_n(t), \ldots C_N(t)$ es solución de

$$M_N \dot{C} + A_N C = f_N(C) \qquad (33)'$$

donde A y M son las mismas del caso lineal, en tanto que

$$f_N(x_1, \ldots, x_N) = \begin{pmatrix} \int_0^1 f(s, \sum_{j=1}^{N} x_j \varphi_j(s))\varphi_1(s)ds \\ \vdots \\ \int_0^1 f(s, \sum_{j=1}^{N} x_j \varphi_j(s))\varphi_N(s)ds \end{pmatrix} \qquad (33)''$$

En cuanto a la condición inicial, (31) deja un residuo

$$\sum_{j=1}^{N} C_j(0)\ (\varphi_j, .) - (\bar{u}, .)\ ,$$

el cual se anulará en S_N si y sólo si

$$M_N\ C(0) = \bar{u}_N \qquad (33)'''$$

con

$$\bar{u}_N := ((\bar{u}, \varphi_1), \ldots, (\bar{u}, \varphi_N)) \qquad (33)^{iv}$$

Así pues, basta resolver el PVI (33) para obtener la aproximación deseada. Para ello, observemos que el elemento ik del jacobiano de f_N es

$$\int_0^1 \frac{\partial f}{\partial u}(s, \sum_{j=1}^{N} x_j \varphi_j(s))\varphi_i(s)\varphi_k(s)ds$$

Para acotarlo, conviene introducir la siguiente hipótesis, adicional a H_1 y H_2

H_3. Tanto f como $\frac{\partial f}{\partial u}$ son continuas y $\left|\frac{\partial f}{\partial u}\right| \leq B$.

Entonces, el elemento ik del jacobiano de f_N está acotado por

$$B \int_0^1 |\varphi_i(s)\varphi_k(s)| ds$$

con lo que f_N será Lipschitz continua, de constante L, con

$$L := B \sup_{ik} |m_{ik}^{(N)}|$$

Dado que $M_N$es no singular, $M_N^{-1}f_N$también será Lipschitziana, por lo que (33) tiene una y sólo una solución [Coddington-Levinson, Cap. 1], y la aproximación está bien definida bajo las hipótesis H_1, H_2 y H_3. Veamos ahora cómo obtener la solución de (33) -o, al menos aproximaciones a ella- de manera iterativa.

Para ello, conviene definir

$$\gamma(t) := M_N^{-1/2} C(t) \qquad (34)'$$

$$g(\gamma) := M_N^{-1/2} f_N(M_N^{-1/2}\gamma) \qquad (34)''$$

$$\gamma_o := M_N^{1/2} \bar{u}_N \qquad (34)'''$$

y entonces se tendrá

$$\dot{\gamma} = -E\gamma + g(\gamma) \qquad \gamma(0) = \gamma_o \qquad (35)$$

con E igual que en la sección 3, positiva definida y simétrica. Nótese que (35) es no lineal y tiene solución única, además de que

$$g'(\gamma) = M^{-1/2} f_N'(M_N^{-1/2}\gamma) M_N^{-1/2}$$

Demos $\{t_j\}$ mediante

$$t_{j+1} = t_j + h \quad , \quad t_o = 0$$

Para generar $\{\gamma_j\}$ como en (13), integremos (35) de t_j a t, con lo que se obtiene

$$\gamma(t) = e^{-(t-t_j)E}\gamma(t_j)+\int_{t_j}^{t} e^{-(s-t_j)E} g(\gamma(s))ds$$

Elijamos una regla de integración numérica

$$\int_a^b f(s)ds \simeq (b-a)\{(1-\alpha)f(a) + \alpha f(b)\}$$

con $0 \leq \alpha \leq 1$, con lo que

$$\gamma(t_{j+1}) \simeq e^{-hE}\gamma(t_j) + h\{(1-\alpha)g(\gamma(t_j)) + \alpha e^{-hE} g(\gamma(t_{j+1})\}$$

Sea $\{\gamma_j\}$ la sucesión definida por

$$\gamma_o \text{ dada}$$

$$\gamma_{j+1} = \alpha h e^{-hE} g(\gamma_{j+1}) + (1-\alpha)hg(\gamma_j) + e^{-hE}\gamma_j$$

Más aún. sea $R_{nm}(Z)$ la correspondiente aproximación racional y propongamos el siguiente algoritmo:

$$\gamma_o \text{ dada} \qquad (36)'$$

$$\gamma_{j+1} := \alpha h R_{nm}(hE) g(\gamma_{j+1}) + (1-\alpha)hg(\gamma_j) + R_{nm}(hE)\gamma_j \qquad (36)''$$

Veamos bajo qué condiciones (36) está bien definido. Para ello, observemos que el jacobiano del segundo miembro de (36)'' con respecto a γ_{j+1} es

$$\alpha h R_{nm}(hE) M^{-1/2} f_N'(M_N^{-1/2}\gamma_{N+1}) M^{-1/2}$$

y su norma N satisface

$$N \leq \alpha\, h\, |R_{nm}(h\,E)|\,|M_N^{-1/2}|^2 \sup_{X\in\mathbb{R}^N} |f_N'(x)|$$

Para estimarla, conviene introducir la siguiente hipótesis adicional

H_4. Para $0 \leq x \leq 1$, $u \in \mathbb{R}$,

$$\frac{\partial f}{\partial u}(x,u) \leq \Lambda := \inf_{0 \neq \varphi \in PC^{1,2}} \frac{a[\varphi]}{\|\varphi\|}$$

Como consecuencia, el elemento ik de $f_N'(x)$ está acotado por

$$\Lambda m_{ik}^{(N)}$$

y entonces

$$|f_N'(x)y| \leq \Lambda |M_N y| \qquad \forall \nu \in \mathbb{R}^N$$

con lo cual

$$|f_N'(x)| \leq \Lambda \lambda_{max}^{(N)} \qquad \forall x \in \mathbb{R}^N$$

en la notación de la sección 3. Asimismo,

$$|M_N^{-1/2}|^2 \leq {}^{1}\!/\!\lambda_{min}^{(N)}$$

y, en consecuencia

$$N \leq \alpha h \operatorname{cond}(M_N) |R_{nm}(hE|$$

Si elegimos h como en (23), se tendrá que $|R_{nm}(hE)| \leq 1$ y entonces

$$N \leq \alpha \Lambda h \operatorname{cond}(M_N) \leq \Lambda h \operatorname{cond}(M_N)$$

Por el principio de contracción [Lyusternik-Sobolev, p. 27] (36)'' tiene una y sólo una solución siempre que, además de (23), h satisfaga la condición

$$h < \frac{\eta}{\lambda \operatorname{cond}(M_N)}, \text{ con } \eta < 1$$

Es decir, basta tomar

$$0 < h < \min \left\{ \frac{\tau_{nm}}{|\lambda(E)|_{min}}, \frac{\eta}{\Lambda \operatorname{cond}(M_N)} \right\} \qquad (37)$$

Hecho esto, la sucesión $\{\gamma_{j+1}^{(m)}\}$ dada por

$$\gamma_{j+1}^{(0)} \text{ arbitrario}$$

$$\gamma_{j+1}^{(m+1)} = \alpha h R_{nm}(hE) g(\gamma_{j+1}^{(m)}) + (1-\alpha) h g(\gamma_j) + R_{nm}(hE)\gamma_j$$

converge a γ_{j+1}. Podemos entonces sugerir el siguiente algoritmo del tipo predictor-corrector para resolver (35) en forma aproximada:

Algoritmo

Dado γ_0, genérese γ_{j+1} a partir de γ_j aplicando los siguientes pasos:

1. Predictor (Euler, $\alpha = 0$)

$$\gamma_{j+1}^{pred} := R_{nm}(hE)\gamma_j + hg(\gamma_j)$$

2. Corrector (De orden $p \geq 1$)

$$\gamma_{j+1}^{(0)} := \gamma_{j+1}^{pred}$$

$$\gamma_{j+1}^{(k)} := \alpha h R_{nm}(hE) g(\gamma_{j+1}^{(k-1)}) + (1-\alpha) h g(\gamma_j) + R_{nm}(hE)\gamma_j \ , \ k = 1, \ldots, p$$

3. Actualización

$$\gamma_{j+1} := \gamma_{j+1}^{(P)}$$

Queda todavía por establecer la convergencia y estabilidad de este método iterativo acoplado con el de Galerkin, siguiendo las mismas lineas de lo hecho en el caso lineal.

Bibliografía

Bender, C. M. y Orszag, S. A., "Advanced Mathematical Methods for Scientists and Engineers", McGraw Hill, Nueva York, 1978.

Coddington, E. A., y Levinson, N., "Theory of Ordinary Differential Equation", McGraw Hill, N. Y., 1955.

Courant, R., y Hilbert, D., "Methods of Mathematical Physics, Vol. I", Interscience, Nueva York, 1961.

Ferziger, J. H., "Numerical methods for engineering applications", J. Wiley & Sons, Nueva York, 1981.

Hernández, D. B., "Análisis Numérico", Departamento de Matemáticas-CINVESTAV (Serie Roja), Cd. de México, 1983.

Liusternik, L. y Sobolev, V., "Elements of Functional Analysis", Ungar, Nueva York, 1961.

Schultz, M. H., "Spline Analysis", Prentice-Hall, Inc., Englewood Cliffs, N. J., 1973.