Estabilidad, Consistencia y Convergencia de un Esquema Numérico (con aplicación la ecuación del calor)

Alberto Pérez-Cervera[1] and Joaquín Domínguez de Tena[1]
[1]*Universidad Complutense de Madrid, Dpto. de Matemática Aplicada, Madrid, España*

DETALLES PRELIMINARES

Sea $U(x,t)$ una función suficientemente continua. Sabemos entonces que podemos obtener una aproximación a la derivada

$$\begin{aligned}\partial_t U(x,t) &= \frac{U(x,t+\Delta t)-U(x,t)}{\Delta t} \; - \; \partial^2_{tt}U(x,t)\frac{\Delta t}{2} \; + \; \dots \\ &= \frac{U(x,t+\Delta t)-U(x,t)}{\Delta t} \; - \mathbf{T_{\Delta t}}(x,t)\end{aligned} \tag{1}$$

con $\mathbf{T_{\Delta t}} = \mathcal{O}(\Delta t)$ la diferencia entre la derivada y nuestra aproximación forward y por ello denotado error de truncamiento.

Así mismo, también conocemos la siguiente aproximación a la derivada segunda

$$\begin{aligned}\partial^2_{xx}U(x,t) &= \frac{U(x+\Delta x,t)-2U(x,t)+U(x-\Delta x,t)}{\Delta x^2} \; + \; \partial^4_{xx}U(x,t)\frac{\Delta x^2}{12} \; + \; \dots, \\ \partial^2_{xx}U(x,t) &= \frac{U(x+\Delta x,t)-2U(x,t)+U(x-\Delta x,t)}{\Delta x^2} + \mathbf{T_{\Delta x}}(x,t),\end{aligned} \tag{2}$$

donde, análogamente tenemos el error de truncamiento asociado $\mathbf{T_{\Delta x}} = \mathcal{O}(\Delta x^2)$.

Por tanto si consideramos la ecuación del calor

$$\partial_t U(x,t) - D\partial^2_{xx}U(x,t) = f(x,t) \tag{3}$$

y sustituimos en las derivadas en (3) por las discretizaciones en (1) y (2), obtenemos

$$\frac{U(x,t+\Delta t)-U(x,t)}{\Delta t} - D\frac{U(x+\Delta x,t)-2U(x,t)+U(x-\Delta x,t)}{\Delta x^2} = f(x,t) + \mathbf{T_h}(x,t) \tag{4}$$

donde $\mathbf{T_h}(x,t) = \mathbf{T_{\Delta t}}(x,t) - \mathbf{T_{\Delta x}}(x,t)$ es el error de truncamiento total. Nótese que el subindice h busca remarcar la dependencia de este en el tamaño de la discretizacion realizada.

Inspirados por (4), nosotros vamos a resolver

$$\frac{u_i^{j+1}-u_i^j}{\Delta t} - D\frac{u_{i+1}^j-2u_i^j+u_{i-1}^j}{\Delta x^2} = f_i^j \tag{5}$$

y pretender que con ello resolvemos (3).

El paso de (4) a (5), en el cual está implícita la asociación $U(x,t) \to u_i^j$, genera la **pregunta clave** de este curso: ¿cómo de parecidas son $U(x,t)$ (solución de (4)) y u_i^j (solución de (5))?

UNA VISIÓN CONCEPTUAL DE LA ESTABILIDAD Y LA CONSISTENCIA

Para responder a esta importante pregunta, reordenamos (4) y (5) de la siguiente forma, sin duda más familiar

$$\begin{aligned} U(x,t+\Delta t) &= U(x,t) + \frac{D\Delta t}{\Delta x^2}(U(x+\Delta x,t) - 2U(x,t) + U(x-\Delta x,t)) + f(x,t)\Delta t + \mathbf{T_h}(x,t)\Delta t, \\ u_i^{j+1} &= u_i^j + \frac{D\Delta t}{\Delta x^2}(u_{i+1}^j - 2u_i^j + u_{i-1}^j) + f(x,t)\Delta t \end{aligned} \tag{6}$$

Si definimos el error como

$$E_i^j = U(x,t) - u_i^j \tag{7}$$

podemos obtener el error de nuestro esquema restando ambas ecuaciones

$$E_i^{j+1} = E_i^j + \frac{D\Delta t}{\Delta x^2}(E_{i+1}^j - 2E_i^j + E_{i-1}^j) + \Delta t \mathbf{T}_i^j \tag{8}$$

Si reescribimos (6) de la siguiente forma

$$\begin{aligned} U(x,t+\Delta t) &= \mathcal{N}U(x,t) + f(x,t)\Delta t + \mathbf{T_h}(x,t)\Delta t, \\ u^{j+1} &= \mathcal{N}u^j + f^j\Delta t \end{aligned} \tag{9}$$

podemos obtener la siguiente relación de recurrencia para el error

$$E_{j+1} = \mathcal{N}E_j + \Delta t \mathbf{T}_j \tag{10}$$

donde, abusando de la notación, en este caso hemos bajado el índice temporal j (las razones para este cambio serán evidentes en breve). Para entender mejor esta recurrencia, desarrollémosla para un j cualquiera, por ejemplo $j = 3$

$$\begin{aligned} E_3 &= \mathcal{N}E_2 + \Delta t \mathbf{T}_2, \\ &= \mathcal{N}\big[\mathcal{N}E_1 + \Delta t \mathbf{T}_1\big] + \Delta t \mathbf{T}_2, \\ &= \mathcal{N}\left[\mathcal{N}\big[\mathcal{N}E_0 + \Delta t \mathbf{T}_0\big] + \Delta t \mathbf{T}_1\right] + \Delta t \mathbf{T}_2 = \Delta t\big(\mathcal{N}\left[\mathcal{N}\mathbf{T}_0 + \mathbf{T}_1\right] + \mathbf{T}_2\big), \end{aligned} \tag{11}$$

donde en la última igualdad hemos usado que $E_i^0 = 0 \ \ \forall i$, pues la condición inicial no genera error (es la misma para U y u). Por tanto, por inducción obtenemos que

$$\begin{aligned} E_{j+1} &= \sum_{n=1}^{j+1} \mathcal{N}^{j+1-n}\mathbf{T}^{n-1} \\ &= \Delta t(\mathbf{T}_j + \mathcal{N}\mathbf{T}_{j-1} + \cdots + \mathcal{N}^j\mathbf{T}_0) \end{aligned} \tag{12}$$

donde se puede la razón por la que bajamos el subíndice del tiempo en (10): para que no se confundiese con los exponentes $\mathcal{N}^j$.

La ecuación (12) pone de manifiesto un detalle conceptual muy relevante: nos muestra que el error E_j (es decir la diferencia entre la solución exacta $U(x,t)$ y la de el esquema u_i^j a tiempo $t = t_0 + j\Delta t$ depende de los errores de truncamiento $\mathbf{T}_i^j$ anteriores y del esquema $\mathcal{N}$ utilizado).

A continuación, llega el momento de introducir dos definiciones:

Consistencia: Diremos que un método es consistente si

$$\|\mathbf{T_h}(x,t)\| \to 0 \qquad \text{conforme} \quad h \to 0 \tag{13}$$

Estabilidad: Un esquema lineal de la forma $u^{j+1} = \mathcal{A}_h u_j + b_j$ es Lax-Richtmyer estable si para todo tiempo t_j existe una constante C positiva tal que

$$\|\mathcal{A}_h^j\| \leq C \tag{14}$$

notese que se cumple

$$\|A^j\| = \|AA^{j-1}\| \leq \|A\|\|A^{j-1}\| \leq \cdots \leq \|A\|^j \tag{15}$$

Para rematar, dejadme tomar normas en (12) y asumir que nuestro esquema es estable

$$\begin{aligned} \|E_{j+1}\| &\leq \Delta t\big(\|\mathbf{T}_j\| + \|\mathcal{N}\|\|\mathbf{T}_{j-1}\| + \cdots + \|\mathcal{N}^j\|\|\mathbf{T}_0\|\big), \\ &\leq \Delta t\|\mathbf{T}_j\| + \Delta tC\big(\|\mathbf{T}_{j-1}\| + \cdots + \|\mathbf{T}_0\|\big) \end{aligned} \tag{16}$$

donde en la segunda desigualdad hemos usado las propiedades (14) y (15).

Si finalmente llamamos $\mathbf{T}_{\text{max}}$ al máximo de todos los módulos de $\mathbf{T}_{\mathbf{h}}(x,t)$, entonces la norma infinita de $\mathbf{T}_j \leq \mathbf{T}_{\text{max}}$, y por tanto

$$\|E_{j+1}\| \leq \Delta t\mathbf{T}_{\text{max}} + \Delta tCj\mathbf{T}_{\text{max}} = \Delta t\mathbf{T}_{\text{max}} + TC\mathbf{T}_{\text{max}} \tag{17}$$

con $T = j\Delta t$ el tiempo final.

A continuación, veremos que, de forma general, todo esquema lineal

$$u_{n+1} = \mathcal{A}_h u_n + b_n \tag{18}$$

que sea consistente (13) y estable (14), cumplirá que

$$\|E_h\| \to 0 \qquad \text{conforme} \quad h \to 0 \tag{19}$$

con $E_h = U(x,t) - u_i^j$, el error definido en (7). Todo esquema que cumpla (19) será **convergente** [1].

Este resultado se conoce como el Teorema de Lax, que enunciamos a continuación:

Lax Equivalence Theorem: Sea un método lineal $u_{n+1} = \mathcal{A}_h u_n + b_n$. Dicho método sera convergente sii es consistente y estable.

Su prueba, se sigue trivialmente todo el trabajo realizado. Aunque por motivos docentes, hemos empezado trabajando con la ecuación del calor,al final los pasos realizados se basan en la linealidad de los operadores de diferencias finitas. De esta linealidad, se sigue la linealidad del sistema que obtenemos al discretizar. Por tanto, acabaremos con un error esquema numérico, cuyo error siga una expresión en la forma (7), va a ser algo bastante general. Despues, siguiendo los mismos cálculos, si nuestro sistema fuese estable llegaríamos a (20).

Finalmente si nuestro método es consistente

$$\|E_{j+1}\| \leq \Delta t\mathbf{T}_{\text{max}} + \Delta tCj\mathbf{T}_{\text{max}} = \Delta t\mathbf{T}_{\text{max}} + TC\mathbf{T}_{\text{max}} \to 0 \quad \text{as} \;\; h \to 0 \tag{20}$$

y con esto tendríamos la convergencia y concluiríamos la prueba del Teorema de Lax.

APLICACIONES PARA EL ESQUEMA EXPLICITO DE LA ECUACIÓN DEL CALOR

Que el esquema explicito para la ecuación del calor es consistente, se sigue trivialmente de las definiciones (1) y (2) de las derivadas utilizadas para discretizar.

Ver la estabilidad nos llevara un poco mas de trabajo. Hemos visto que

$$E_i^{j+1} = E_i^j + \lambda(E_{i+1}^j - 2E_i^j + E_{i-1}^j) + \Delta t \mathbf{T}_i^j = (1-2\lambda)E_i^j + \lambda(E_{i+1}^j + E_{i+1}^j) + \Delta t \mathbf{T}_i^j \tag{21}$$

donde hemos introducido $\lambda = \frac{D\Delta t}{\Delta x^2}$.

Si definimos $\varepsilon^j = \max |E_i^j|$ fijaros que

$$\varepsilon^{j+1} \leq (|1-2\lambda| + 2\lambda)\varepsilon^j + C(\Delta t(\Delta x^2 + \Delta t)) \tag{22}$$

Si, de forma parecida a (11), iteramos esta desigualdad desde usando $\varepsilon_0 = 0$, obtendríamos

$$\varepsilon^{j+1} \leq C(\Delta t(\Delta x^2 + \Delta t))[1 + \alpha_\lambda + \alpha_\lambda^2 + \dots] \tag{23}$$

con $\alpha_\lambda = (|1-2\lambda| + 2\lambda)$.

Ahora tenemos que distinguir dos casos. Si $\lambda \leq 1/2$, entonces $\alpha_\lambda = 1$

$$\varepsilon^{j+1} \leq C(T(\Delta x^2 + \Delta t)) \to 0 \qquad \text{as } \Delta x,\ \Delta t \to 0 \tag{24}$$

es decir para $\lambda \leq 1/2$ el esquema es convergente.

Por contra si $\lambda > 1/2 \to \alpha_\lambda = 4\lambda - 1$, y por tanto, se puede demostrar que ya no podemos asegurar convergencia.

La restricción $\lambda \leq 1/2$ para que el esquema explícito sea estable es conocida como condición CFL (por Courant-Friederich-Lax). Esta condición nos condena a usar un paso temporal pequeño. Por ejemplo, imaginad un problema en el cual $\Delta x = 0.1$ y $D = 1$. Si quisiéramos llegar hasta $T = 1$, la condición CFL nos dice que $\Delta t \leq 0.01/2 = 0.005$. Es decir, necesitaríamos al menos $N = 200$ pasos para llegar a la meta. Esto no hace más que empeorar cuanto más fina (espacialmente) sea la malla. Esta restricción en Δt, motivará la discusión de nuevos esquemas.

COMENTARIOS FINALES

Si bien estas notas pueden parecer poco formales, su objetivo es el de crear intuición sobre los objetos con los que se trabaja. El lector más formal, puede satisfacer su paladar recurriendo a la bibliografía:

---

[1] Nótese la belleza del resultado: (19) muestra que las soluciones del esquema (que son discretas) tienden a las de problema (que son continuas) conforme aumentamos la densidad del mallado.

[2] G. D. Smith, *Numerical solution of partial differential equations: finite difference methods* (Oxford university press, 1985).

[3] R. J. LeVeque, *Finite difference methods for ordinary and partial differential equations: steady-state and time-dependent problems* (SIAM, 2007).

[4] E. Zuazua, *Métodos numéricos de resolución de ecuaciones en derivadas parciales* (2009).