Métodos Theta

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

INTRODUCCIÓN

Seguimos pretendiendo resolver la la ecuación del calor

$$\partial_t U(x,t) - D\partial^2_{xx} U(x,t) = f(x,t). \tag{1}$$

Consideremos ya fijo el mallado que queremos utilizar.

Hasta ahora, hemos visto dos métodos: el explicito y el implícito. Más allá de sus ventajas e inconvenientes, quiero señalar una de las principales diferencias de construcción entre ambos métodos.

- El método explicito consideraba la solución en $U(x,t)$ y discretizaba hacia adelante en tiempo.
- El método implícito consideraba la solución en $U(x,t+\Delta t)$ y discretizaba hacia detrás en tiempo.

Aunque quizás la primera vez que nos situamos en $U(x,t+\Delta t)$ pudo sorprender (pues no es la elección más "natural"), espero que ahora esto nos prepare para la siguiente vuelta de tuerca: ¿qué pasa si considero la solución en un punto intermedio del mallado? (denotando este punto intermedio por $U(x,t+\theta\Delta t)$ donde, por supuesto $0 \leq \theta \leq 1$).

Es decir, voy a considerar

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

Lo primero es recordar que el mallado está fijo. Por lo tanto, como he elegido arrancar mi método desde un punto que no pertenece a mi mallado, lo primero va a ser reescribir las derivadas en (2) como una combinación lineal de las derivadas en los puntos t_0 y t_1 (estos si pertenecientes al mallado). Para ello, recurrimos a la magia de Taylor

$$U(x,t_\theta+h_t) = U(x,t_\theta) + \partial_t U(x,t)|_{t=t_\theta} h_t + \partial_{tt} U(x,t)|_{t=t_\theta}\frac{h_t^2}{2} + \partial_t^3 U(x,t)|_{t=t_\theta}\frac{h_t^3}{6} + ... \tag{3}$$

donde h_t es un parámetro que podemos elegir a conveniencia.

De hecho si vemos que

$$t_\theta = t_0 + \theta\Delta t = t_0 + \theta(t_1 - t_0) = t_0(1-\theta) + \theta t_1 \tag{4}$$

podemos ver que

$$\begin{aligned} t_\theta + h_t &= t_1 \qquad \text{si} \quad h_t = (1-\theta)\Delta t, \\ t_\theta + h_t &= t_0 \qquad \text{si} \quad h_t = -\theta\Delta t, \end{aligned} \tag{5}$$

y por tanto, si sustituimos las h_t de (5) en (3), vemos que

$$\begin{aligned} U(x,t_1) &= U(x,t_\theta) + \partial_t U(x,t)|_{t=t_\theta}(1-\theta)\Delta t + \partial_{tt} U(x,t)|_{t=t_\theta}\frac{(1-\theta)^2\Delta t^2}{2} + \partial_t^3 U(x,t)|_{t=t_\theta}\frac{(1-\theta)^3\Delta t^3}{6} + \dots, \\ U(x,t_0) &= U(x,t_\theta) - \partial_t U(x,t)|_{t=t_\theta}\theta\Delta t + \partial_{tt} U(x,t)|_{t=t_\theta}\frac{(\theta\Delta t)^2}{2} - \partial_t^3 U(x,t)|_{t=t_\theta}\frac{(\theta\Delta t)^3}{6} + \dots \end{aligned} \tag{6}$$

y finalmente si restamos

$$U(x,t_1) - U(x,t_0) = \partial_t U(x,t)|_{t=t_\theta}\Delta t + \partial_{tt} U(x,t)|_{t=t_\theta}\frac{(1-2\theta)\Delta t^2}{2} + \mathcal{O}(\Delta t^3) + \dots \tag{7}$$

es decir

$$\begin{aligned}\partial_t U(x,t)|_{t=t_\theta} &= \frac{U(x,t_1)-U(x,t_0)}{\Delta t} - \partial_{tt}U(x,t)|_{t=t_\theta}\frac{(1-2\theta)\Delta t}{2} - \mathcal{O}(\Delta t^2)\ldots, \\ &= \frac{U(x,t_1)-U(x,t_0)}{\Delta t} - \mathbf{T}_{\mathbf{\Delta_t}}(x,t)\end{aligned} \tag{8}$$

Por tanto, ya tendríamos la derivada temporal en (2). Para obtener la derivada espacial segunda, no tenemos más que proceder de forma similar. Por ejemplo consideremos (6) y multipliquemos cada ecuación por un factor "estratégico"

$$\begin{aligned}\theta U(x,t_1) &= \theta\Big(U(x,t_\theta) + \partial_t U(x,t)|_{t=t_\theta}(1-\theta)\Delta t + \partial_{tt}U(x,t)|_{t=t_\theta}\frac{(1-\theta)^2\Delta t^2}{2} + \partial_t^3 U(x,t)|_{t=t_\theta}\frac{(1-\theta)^3\Delta t^3}{6} + \ldots\Big), \\ (1-\theta)U(x,t_0) &= (1-\theta)\Big(U(x,t_\theta) - \partial_t U(x,t)|_{t=t_\theta}\theta\Delta t + \partial_{tt}U(x,t)|_{t=t_\theta}\frac{(\theta\Delta t)^2}{2} - \partial_t^3 U(x,t)|_{t=t_\theta}\frac{(\theta\Delta t)^3}{6} + \ldots\Big)\end{aligned} \tag{9}$$

y si sumamos ambas ecuaciones

$$\theta U(x,t_1) + (1-\theta)U(x,t_0) = U(x,t_\theta) + \mathcal{O}(\Delta t^2) \tag{10}$$

de donde obtenemos

$$\partial_{xx}^2 U(x,t_\theta) = \partial_{xx}^2\big[\theta U(x,t_1) + (1-\theta)U(x,t_0) - \mathcal{O}(\Delta t^2)\big] \tag{11}$$

por tanto ya tenemos las derivadas segundas en (2).

Finalmente, dejo como ejercicio comprobar que

$$f(x,t_\theta) = \theta f(x,t_1) + (1-\theta)f(x,t_0) + \mathcal{O}(\Delta t^2) \tag{12}$$

EL MÉTODO THETA

Por tanto, en lugar de resolver

$$\partial_t U(x,t)|_{t=t_\theta} - D\partial_{xx}^2 U(x,t_\theta) = f(x,t_\theta) \tag{13}$$

vamos a resolver

$$\frac{u_i^{j+1}-u_i^j}{\Delta t} = D\Big(\theta\frac{u_{i+1}^{j+1}-2u_i^{j+1}+u_{i-1}^{j+1}}{\Delta x^2} + (1-\theta)\frac{u_{i+1}^{j}-2u_i^{j}+u_{i-1}^{j}}{\Delta x^2}\Big) + \theta f_i^{j+1} + (1-\theta)f_i^j \tag{14}$$

llamado, por razones evidentes "theta method", y pretender que con ello resolvemos (13).

Llegados a este punto es necesario remarcar que puede haber quien piense que "para este viaje no hacían falta alforjas" ya que (14) no es más que la suma ponderada de los esquemas explícito e implícito anteriormente derivados.

¿Hemos ganado algo?

Pues si. Para ello, veamos cual es el truncamiento de nuestro método. Mientras que los truncamientos de (11) y (12), claramente son $\mathcal{O}(\Delta x^2)$, $\mathcal{O}(\Delta t^2)$ el truncamiento de (8) es un poco más interesante

$$\mathbf{T}_{\mathbf{\Delta_t}} = \partial_{tt}U(x,t)|_{t=t_\theta}\frac{(1-2\theta)\Delta t}{2} + \mathcal{O}(\Delta t^2)\ldots \tag{15}$$

de hecho, para la elección $\theta = 1/2$, el truncamiento de la derivada temporal ya no será $\mathcal{O}(\Delta t)$ sino $\mathcal{O}(\Delta t^2)$.

A la elección particular $\theta = 1/2$ se le conoce como método de Crank-Nicolson (CN). Nótese que al final el menor error del método de CN ($\theta = 1/2$) no sorprende tanto si nos damos cuenta de que, en verdad, $\theta = 1/2$ es equivalente a situarnos en el punto intermedio entre t_1 y t_0 y aproximar la derivada temporal usando diferencias centradas (cuyo error de truncamiento sabemos es $\mathcal{O}(\Delta t^2)$).

CONSISTENCIA, ESTABILIDAD Y CONVERGENCIA

Nuevamente, de todos los cálculos anteriormente realizados, se sigue que los métodos theta son consistentes

$$\mathbf{T_h} = \mathcal{O}(\Delta t, \Delta x^2) \quad \text{si } \theta \neq 1/2 \qquad \mathbf{T_h} = \mathcal{O}(\Delta t^2, \Delta x^2) \quad \text{si } \theta = 1/2 \tag{16}$$

Para discutir su estabilidad y convergencia, consideremos (13) y sustituyamos las derivadas por sus Taylors correspondientes en (8), (11) y (12)

$$\partial_t U(x,t) = D\big(\theta \partial^2_{xx} U(x,t_1) + (1-\theta)\partial^2_{xx} U(x,t_0)\big) + \theta f(x,t_1) + (1-\theta) f(x,t_0) + \mathbf{T_h} \tag{17}$$

donde $\mathbf{T_h}$ recoge todos los truncamientos anteriores. Si restamos esta ecuación al esquema theta anteriormente derivado en (14)

$$\frac{u_i^{j+1} - u_i^j}{\Delta t} = D\Big(\theta \frac{u_{i+1}^{j+1} - 2u_i^{j+1} + u_{i-1}^{j+1}}{\Delta x^2} + (1-\theta)\frac{u_{i+1}^{j} - 2u_i^{j} + u_{i-1}^{j}}{\Delta x^2}\Big) + \theta f_i^{j+1} + (1-\theta) f_i^j \tag{18}$$

obtenemos una ecuación para el error

$$\begin{aligned} E_i^{j+1} - E_i^j &= \lambda\Big(\theta(E_{i+1}^{j+1} - 2E_i^{j+1} + E_{i-1}^{j+1}) + (1-\theta)(E_{i+1}^{j} - 2E_i^{j} + E_{i-1}^{j})\Big) + \Delta t \mathbf{T_h}, \\ (1+2\theta\lambda)E_i^{j+1} &= \lambda\Big(\theta(E_{i+1}^{j+1} + E_{i-1}^{j+1}) + (1-\theta)(E_{i+1}^{j} + E_{i-1}^{j})\Big) + (1-2(1-\theta)\lambda)E_i^j + \Delta t \mathbf{T_h}, \\ &\leq \lambda\Big(\theta(E_{i+1}^{j+1} + E_{i-1}^{j+1}) + (1-\theta)(E_{i+1}^{j} + E_{i-1}^{j})\Big) + (1-2(1-\theta)\lambda)E_i^j + \Delta t C \mathbf{T_{max}}, \end{aligned} \tag{19}$$

ahora, si exigimos $(1-2(1-\theta)\lambda) > 0$, es decir, $\lambda(1-\theta) \leq 1/2$, podemos, llamando $\varepsilon^j = \max|\varepsilon_i^j|$

$$\begin{aligned} (1+2\theta\lambda)\varepsilon^{j+1} &\leq 2\lambda\theta\varepsilon^{j+1} + \varepsilon^j + \Delta t C \mathbf{T_{max}}, \\ \varepsilon^{j+1} &\leq \varepsilon^j + \Delta t C \mathbf{T_{max}} \end{aligned} \tag{20}$$

y por tanto el método es estable y convergente si $\lambda(1-\theta) \leq 1/2$. Nótese que esta condición recoge las dos previamente derivadas tanto para el esquema explícito ($\theta = 0$) como el implícito ($\theta = 1$).

UN EJEMPLO DE USO

Ahora que ya entendemos de donde vienen los theta métodos y su interés (podemos pensar que es una forma compacta de programar cualquier método anteriormente aprendido o de mejorar el error de nuestro numérico solo variando θ) vamos a discutir un ejemplo de uso.

Para ello, consideremos el theta método

$$\frac{u_i^{j+1}-u_i^j}{\Delta t} = D\Big(\theta\frac{u_{i+1}^{j+1}-2u_i^{j+1}+u_{i-1}^{j+1}}{\Delta x^2} + (1-\theta)\frac{u_{i+1}^{j}-2u_i^{j}+u_{i-1}^{j}}{\Delta x^2}\Big) + \theta f_i^{j+1} + (1-\theta)f_i^j \tag{21}$$

Supongamos condiciones de contorno conocidas

$$u_i^0 = g(x_i) \tag{22}$$

y las de borde también (por comodidad asumiremos Dirichlet e iguales a cero)

$$u_0^j = u_N^j = 0 \tag{23}$$

No tenemos más que separar (21) en conocido y por conocer. Trabajando de esta forma, llegamos a

$$(1+2\theta\lambda)u_i^{j+1} - \lambda\theta(u_{i+1}^{j+1}+u_{i-1}^{j+1}) = \lambda(1-\theta)(u_{i+1}^j+u_{i-1}^j) + u_i^j(1-2\lambda(1-\theta)) + \Delta t(\theta f_i^{j+1} + (1-\theta)f_i^j) \tag{24}$$

para j fijada, si vamos variando $i = 1, \ldots, N-1$ acabaremos dándonos cuenta que se genera el siguiente sistema

$$\begin{pmatrix} (1+2\theta\lambda) & -\lambda\theta & 0 & \dots & 0 \\ -\lambda\theta & (1+2\theta\lambda) & -\lambda\theta & \dots & 0 \\ 0 & -\lambda\theta & (1+2\theta\lambda) & \ddots & \vdots \\ \vdots & & \ddots & (1+2\theta\lambda) & -\lambda\theta \\ & & & -\lambda\theta & (1+2\theta\lambda) \end{pmatrix} \begin{pmatrix} u_1^{j+1} \\ u_2^{j+1} \\ \vdots \\ u_{N-2}^{j+1} \\ u_{N-1}^{j+1} \end{pmatrix} = \begin{pmatrix} b_1^j \\ b_2^j \\ \vdots \\ b_{N-2}^j \\ b_{N-1}^j \end{pmatrix} \tag{25}$$

donde los b_i^j, agrupan todos los términos conocidos

$$b_i^j = \lambda(1-\theta)(u_{i+1}^j+u_{i-1}^j) + u_i^j(1-2\lambda(1-\theta)) + \Delta t(\theta f_i^{j+1} + (1-\theta)f_i^j) \tag{26}$$

y por tanto, resolviendo el sistema lineal, tendríamos la solución a tiempo t_{j+1}. De forma iterativa, iríamos usando cada nueva solución a tiempo t para hallar la nueva solución a tiempo $t+\Delta t$. Nótese que aunque la matriz A es constante, los coeficientes b dependen de j, por lo que puede ser necesario tener que actualizar el vector b a cada iteración.

Finalmente, las mismas ideas se aplicarían en caso de diferentes condiciones de contorno. Pase lo que pase, solo tenemos que generar el sistema de ecuaciones separando en conocido, y por conocer.