Práctica 2 de MNEDP

Diego Rodríguez Atencia

8 de noviembre de 2020

1. Introducción

En las siguientes páginas, realizaremos un análisis de los métodos propuestos para aproximar numéricamente la ecuación planteada:

$$u_t + au_x = 0$$

Esta se conoce como la ecuación de convección, útil en numerosos problemas donde tenemos masas de objetos que se mueven a cierta velocidad, como por ejemplo al meteorología, la hidráulica o la termodinámica. Aunque no sea especialmente complicada, es el entorno ideal para probar nuevos métodos numéricos para ecuaciones hiperbólicas, sin necesidad de aumentar mucho la comlejidad de los esquemas.
Utilizaremos métodos de aproximación basados en las diferencias finitas, y luego analizaremos su **estabilidad** y **consistencia** para demostrar su **convergencia**. Además, daremos una medida del **error de fase**, gracias al análisis de Fourier, para poder evaluar el nivel de difusión y achatamiento que sufren las soluciones numéricas. Finalmente, explicaremos cómo se relaciona cada uno con el **principio del máximo**, una cualidad deseable en este tipo de métodos.

2. La solución de la ecuación

Primero, hallemos la solución, para observar la estabilidad y consistencia del esquema.
La siguiente familia de funciones reciben el nombre de características, y vienen dadas por la siguiente ecuación:

$$\frac{dx}{dt} = a(x,t)$$

Podemos demostrar que u es constante para todas las características:

$$\frac{du(x(t),t)}{dt} = \frac{\partial u}{\partial t} + \frac{\partial u}{\partial x}\frac{dx}{dt} = 0$$

Como en este ejercicio $a(x,t)$ está definida en un compacto y es continua (debido a que el numerador nunca se anula en el espacio donde está definida, y está formada por la composición de funciones continuas), las soluciones de la ecuación no se cruzan, ya que $a(x,t)$ es Lipschtitz. Como $a(x,t)$Sea

$$u(x,0) = u^0(x)$$

El $a(x,t)$ que nos dan es el siguiente:

$$a(x,t) := \frac{1+x^2}{1+2xt+2x^2+x^4}$$

Para la solución exacta usamos el dato de antes sobre las características: Sabemos que para $(x_j, 0)$, sólamente puede pasar una característica por el punto. Como la $u(x(t),t)$ es entonces constante para todo punto de la característica, tenemos que $u(x(t),t) = u^0(x_j)$. Para hallar la solución para un tiempo t, sólamente tenemos que desplazar la condición inicial a donde sea necesario, recorriendo las características. Resolviendo la ecuación de las características, y sustituyendo en la ecuación inicial, tenemos:

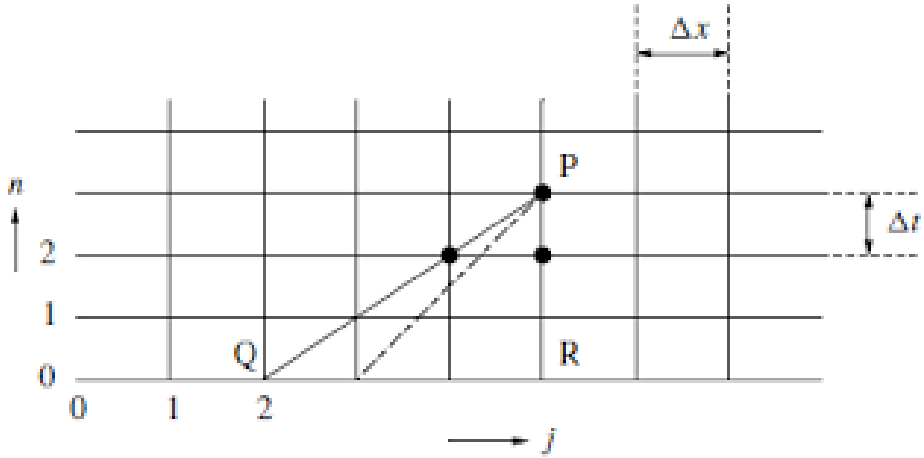
$$u(x,t) = u^0(x - ta(x,0)) = u^0(x - \frac{t}{1+x^2})$$

donde hemos sustituido la x por la solución de la ecuación característica con la condición inicial $x(0) = x$, donde la primera x es la función característica y la seguna es el nombre que le hemos puesto al parámetro de la función u.
donde $x^* = x - \frac{t}{1+x^2}$, que nos confirma la naturaleza de la ecuación. La solución no hace más que desplazarse por el eje x.
La solución de nuestra ecuación es:

$$u(x,t) = u^0(x - \frac{t}{1+x^2})$$

3. La condión CFL

Courant, Friedichs y Lewy, formularon una condición conocida como la condición CFL para la convegencia de una aproximación por diferencias finitas, en función del dominio de dependencia. El dominio de dependencia se trata de la región que abarca el "grid"para el cual la diferencia nos dá valores.



Com podemos observar en la imagen, el dominio de dependencia de este modelo (exactamente el modelo upwind para $a > 0$, como veremos más adelate) en particular estaría comprendido entre los puntos P, Q y R, ya que para para P necesiamos los dos puntos inmediatamente inferiores según el dibujo, y así recursivamente. Entonces, la condición CFL afirma que para que un método de aproximación por diferencias finitas sea convergente, es necesario que el dominio de dependencia de la solución de la ecuación en derivadas parciales debe estar contenido en el dominio de dependencia del esquema numérico. En otro caso, el método diverge.

Por tanto la condición CFL para la ecuación es que la velocidad de la solución no supere a la "velocidad del modelo.º la velocidad con la que el esquema barre el grid. Esto es:

$$\frac{|a(x_j, t_n)|\Delta t}{\Delta x} \leq 1$$

Para este experimento, es fácil obsevar que $a(x, t)$, está acotado superiormente por 1 e inferiormente por 0, de forma que por parte de la condición vista arriba, no hace falta que nos preocupemos (dado que sabemos que $\frac{dt}{dx} = 1$.)
Está demostrado que la condición CFL es necesaria pero no suficiente para la convergencia de un modelo. Un ejemplo puede ser el siguiente esquema, cuya ecuación es la siguiente:

$$\frac{U_j^{n+1} - U_j^n}{\Delta t} + a(x_j, t_n)\frac{U_{j+1}^n - U_{j-1}^n}{2\Delta x} = 0$$

Aunque este esquema cumpla la condición CFL, en un análisis de estabilidad más profundo descubriríamos que es inestable incondicionalmente.

4. El método upwind

El método upwind es un primer acercamiento a esta ecuación. Dada la siguiente regla de actualización:

$$\frac{U_j^{n+1} - U_j^n}{\Delta t} + a(x_j, t_n)\frac{U_j^n - U_{j-1}^n}{\Delta x} = 0$$

Tras unos experimentos iniciales, está claro que este método no converge cuando $a < 0$, precisamente porque entonces ya no se está cumpliendo la condición CFL. Para lidiar con este problema, haremos una ligera modificación con el fin de contener el dominio de dependencia de la solución en el dominio de dependencia del esquema. Esto puede hacerse dando dos modelos diferentes en función del signo de $a(x,t)$:

$$a(x_j, t_n) > 0 : \frac{U_j^{n+1} - U_j^n}{\Delta t} + a(x_j, t_n)\frac{U_j^n - U_{j-1}^n}{\Delta x} = 0$$
$$a(x_j, t_n) < 0 : \frac{U_j^{n+1} - U_j^n}{\Delta t} + a(x_j, t_n)\frac{U_{j+1}^n - U_j^n}{\Delta x} = 0$$

4.1. Análisis del error de truncación

Si u es suficientemente suave, podemos estimar el orden de truncación así:

$$T_j^n := \frac{u_j^{n+1} - u_j^n}{\Delta t} + a_j^n \frac{u_j^n - u_{j-1}^n}{\Delta x}$$

$$\approx [u_t + \frac{1}{2}\Delta t + ...]_j^n + [a(u_x - \frac{1}{2}\Delta x u_{xx} + ...)]_j^n$$

$$= \frac{1}{2}(\Delta t u_{tt} - a\Delta x u_{xx}) + ...$$

De donde deducimos que el orden del error de truncación es $O(\Delta t, \Delta x)$. De este modo, sabemos que $T_j^n \to 0$ cuando $\Delta t \to 0$. Con esto se demuestra la **consistencia** del esquema.

4.2. Análisis de Fourier

El análisis de Fourier para la estabilidad consistirá en sustituir los elementos de la ecuación por los nodos de Fourier, para despejar λ en función de $k, \Delta x, \Delta t, a$.
Sustituyendo $U_j^n = (\lambda)^n e^{ik(j\Delta x)}$, nos queda la siguiente expresión:

$$(\lambda)^{n+1} e^{ik(j\Delta x)} = (\lambda)^n e^{ik(j\Delta x)} + \frac{a\Delta t}{\Delta x}((\lambda)^n e^{ik((j+1)\Delta x)} - (\lambda)^n e^{ik(j\Delta x)}) \implies \lambda = 1 - \frac{a\Delta t}{\Delta x}(1 - e^{-ik\Delta x})$$

(Recordemos que hemos definido $\nu := \frac{a\Delta t}{\Delta x}$). Elevando al cuadrado la expresión:

$$\begin{aligned} |\lambda(k)|^2 &= [(1-\nu) + \nu\cos k\Delta x]^2 + [\nu \sin k\Delta x]^2 \\ &= (1-\nu)^2 + \nu^2 + 2\nu(1-\nu)\cos k\Delta x \\ &= 1 - 2\nu(1-\nu)(1-\cos k\Delta x) \end{aligned}$$

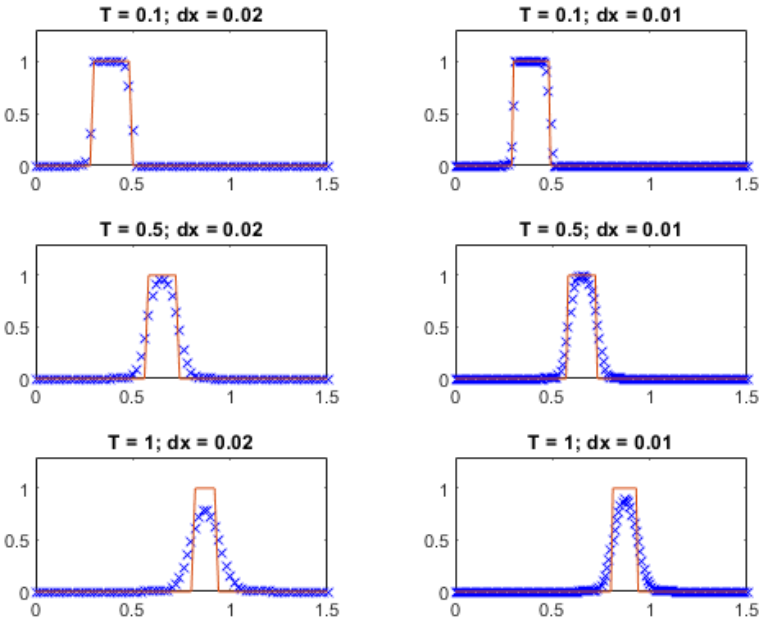
De donde deducimos que:

$$|\lambda^2| = 1 - 4\nu(1-\nu)sin^2(\frac{1}{2}k\Delta x)$$

Dado que el esquema es estable cuando $|\lambda| \leq 1$, deducimos que el esquema será estable siempre y cuando $0 \leq \nu \leq 1$. Es decir:

$$0 \leq \frac{a(x_j, t_n)\Delta t}{\Delta x} \leq 1$$

Lo que nos dice que en este caso, el cumplimiento de la condición CFL es equivalente a la estabilidad del esquema.
Podemos observar que para nuestro experimento, como $a(x,t)$ está acotado entre 0 y 1, y el cociente $\frac{dt}{dx}$ es 1, podemos afirmar que el esquema es **estable** para este experimento.
A continuación se incluyen los dibujos correspondientes a los experimentos:



Donde a la izquierda tenemos los experimentos para el paso en tiempo $dx = 0{,}02$, y a la derecha $dx = 0{,}01$. Luego en cada columna variamos el tiempo.
Es posible apreciar que este esquema cumple el principio del máximo, ya que el máximo se alcanza en la solución inicial y el mínimo se alcanza en la condición de contorno.

4.3. El principio del máximo

Presentemos el principio del máximo. Se dice que el esquema cumple el principio del máximo si:

$$U_{min} \leq U_j^n \leq U_{max}$$

Donde:

$$U_{min} = min\{U_0^m, 0 \leq m \leq n; U_j^0, 0 \leq j \leq J; U_J^m, 0 \leq m \leq n\}$$

$$U_{max} = max\{U_0^m, 0 \leq m \leq n; U_j^0, 0 \leq j \leq J; U_J^m, 0 \leq m \leq n\}$$

Claramente, el principio del máximo es una condición deseable, que nos garantiza que el modelo se comporte bien. Para demostrar que el principio del máximo se cumple en el esquema upwind, escribamos la ecuación del esquema de la siguiente forma:

$$U_j^{n+1} = (1 - \nu)U_j^n + \nu U_{j-1}^n$$

Veamos que los dos coeficientes son positivos. Para empezar, en nuestro experimento a siempre era positivo, lo que implica que $\nu = \frac{a\Delta t}{\Delta x} = a > 0 \implies \nu \geq 0$. Asimismo, $1 - \nu$ es positivo también, debido a que en este caso se cumplía la condición CFL, así que:

$$\nu \leq \frac{|a|\Delta t}{\Delta x} \leq 1 \implies 1 - \nu \geq 0$$

Ahora llega la parte de la demostración. Supongamos que el máximo se alcanza en un punto intermedio (lo que significaría que no se cumple el principio del máximo), digamos U_j^{n+1}. Sea $U^* = max\{U_j^n, U_{j-1}^n\}$. Entonces, como los coeficientes no son negativos, tenemos que $U_j^{n+1} \leq U^*$. Sin embargo, como asumimos que U_j^{n+1} era el máximo, tenemos que $U_j^{n+1} = U^*$, entonces el máximo debe hallarse tras una secuencia de puntos, en la condición inicial.Para $a < 0$ es exactamente el mismo proceso, variando en algunos puntos. Hemos probado por lo tanto que para el esquema upwind se cumple el **principio del máximo**.

4.4. Análisis del error del modelo

Antes, en el análisis de Fourier, obtuvimos la siguiente fórmula:

$$\lambda(u) = 1 - \nu(1 - e^{-ik\Delta x})$$

La fase es, entonces:

$$\lambda(u) = (1 - \nu + \nu cos(k\Delta x) - i\nu sin(k\Delta x) \implies arg\lambda = -tan^{-1}(\frac{\nu sin(k\Delta x)}{(1-\nu) + \nu cos(k\Delta x)}$$

Conviene introducir un lema para desarrollar la expresión anterior:
Si q posee una cierta expansión en potencias:

$$q \approx c_1 p + c_2 p^2 + c_3 p^3 + c_4 p^4 + ... \; p \to 0 \implies tan^{-1}(q) \approx c_1 p + c_2 p^2 + (c_3 - \frac{1}{3}c_1)p^3 + (c_4 - c_1^2 c_2)p^4 + ...$$

Ahora podemos entonces expandir el argumento para obervar el orden de convergencia hasta $e^{-ika\Delta t}$, donde este término es el término complejo de la solución real por el método de Fourier:
Denotemos $\xi = k\Delta x$

$$\begin{aligned} arg\lambda &\approx -tan^{-1}(\nu(\xi - \frac{1}{6}\xi^3 + ...)(1 - \frac{\nu}{2}\xi^2 + ...)^{-1}) = -tan^{-1}(\nu\xi - \frac{1}{6}\nu(1-3\nu)\xi^3 + ...) \\ &= -\nu\xi(1 - \frac{1}{6}(1-\nu)(1-2\nu)\xi^2 + ...) \end{aligned}$$

En el último paso hemos aplicado el lema a nuestro desarrollo con $p = \xi, c_1 = -\nu, c_2 = 0, c_3 = \frac{1}{6}\nu(1-3\nu)$
Observemos que el desarrollo de $arg\lambda$ constituye una aproximación a $-ak\Delta t$. Recordemos que $\nu\xi = a\frac{\Delta t}{\Delta x}k\Delta x$. Dado que:

$$|\lambda|^2 = 1 - 4\nu(1-\nu)sin^2\frac{k\Delta x}{2} = 1 - 4\nu(1-\nu)(\frac{\xi^2}{4} + ...)$$

Por lo tanto, el error en $|\lambda|^2$ se comporta como ξ^2. Por otra parte, para el error de fase relativo obtenemos:

$$\frac{arg\lambda - (-ak\Delta t)}{-ak\Delta t} = -\frac{1}{6}(1-\nu)(1-2\nu)\xi^2 + ... = O(\xi^2)$$

4.5. Conclusión

Hemos podido obsevar que en nuestro experimento, bajo las condiciones descritas, el esquema es **estable** y **consistente**, lo que quiere decir que el método es convergente. Por otra parte, hemos probado una característica muy buena en un esquema, que es la del **principio del máximo**. Sin embargo, a pesar de tener consigo todas estas buenas características, cabe destacar que este método posee sólamente un orden de convergencia lineal con respecto al paso en tiempo, lo que hace que se necesiten usar muchos pasos para llegar a una solución aceptable en muchos casos. Podemos observar también un claro achatamiento y una clara dispersión de la gráfica, junto con un "smoothing"de la forma original.

5. El método Lax-Wendroff

El método Lax-Wendroff se trata del siguiente paso lógico al Upwind. Si en el Upwind estábamos usando una interpolación lineal para aproximar la solución con dos puntos, en este esquema usaremos tres puntos, y consecuentemente una interpolación cuadrática. Para los siguientes experimentos, dado que $a(x,t)$ no es constante, haremos el siguiente desarrollo:

$$\begin{aligned} & u_t(x,t) = -a(x,t)u_x(x,t) \\ & = u_{tt} = -a_t u_x - a u_{xt} \\ & (u_{tx} = u_{xt} = -(au_x)_x \\ & \implies u_{tt} = -a_t u_x + a(au_x)_x \end{aligned}$$

Tengamos en cuenta de que una forma de aproximar u_{tt} es la siguiente:

$$u(x,t)_{tt}\Delta t^2 + O(\Delta t^3) = u(x,t+\Delta t) - u(x,t) - u_t \Delta t$$

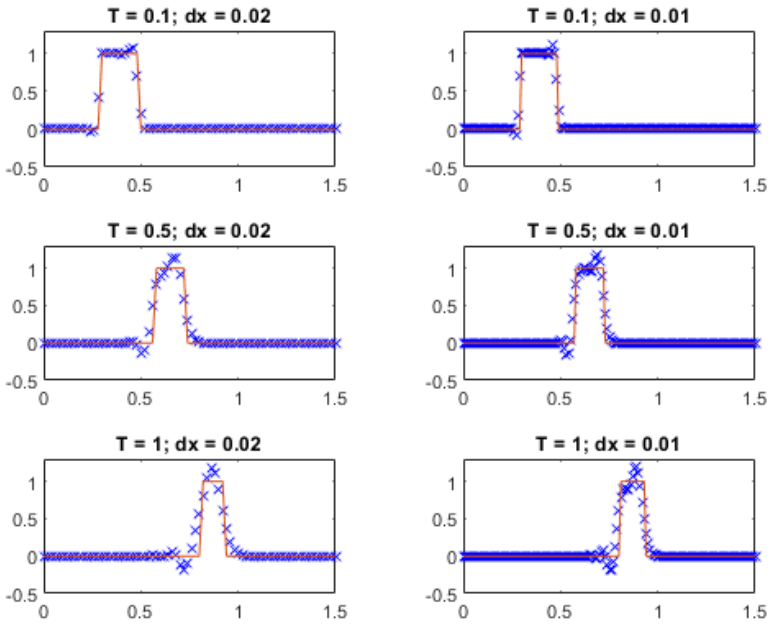
De este modo, sustituimos todas las derivadas por sus respectivas aproximaciones por diferencias. Después de despejar U_j^{n+1} en función de $U_k^n k = j-1, j, j+1$ nos queda la siguiente expresión para el método:

$$U_j^{n+1} = U_j^n - \nu\Delta_{0x}U_j^n + \frac{1}{2}(\Delta t)^2(-(a_t)_j^n \frac{\Delta_{0x}U_j^n}{\Delta x} + a_j^n \frac{\delta_x(a_j^n \delta_x U_j^n)}{(\Delta x)^2})$$

Como el dominio de dependencia de este esquema es simétrico, no tenemos que enunciar dos modelos diferentes dependiendo de a, así que el hecho de que se cumpla la condición CFL sólo dependerá del tamaño a. Según nuestro análisis de la estabilidad sustituyendo los elementos del aecuación por nodos de Fourier, resulta que otra vez, la condición CFL vuelve a corresponder unívocamente con la estabilidad, ya que sustituyendo $\lambda^n e^{ikj\Delta x}$ por U_j^n en cada caso, y despejando λ, resulta que:

$$|\lambda| \leq 1 \iff |\nu| \leq 1 \tag{1}$$

Donde $\nu = \frac{a\Delta t}{\Delta x}$. Con el análisis del error de truncación correspondiente, deducimos que este método es de orden $O((\Delta x)^2)$
Seguidamente se adjuntan las imágenes de los experimentos realizados:



Puede apreciarse que este esquema no cumple el princpio del máximo, ya que hay soluciones intermedias más grandes que la solución inicial o más pequeñas que las condiciones de contorno.

5.1. Análisis del error de truncación

Veamos el error de truncación que posee el método.Hay que dividir por Δt^2 el término de la izquierda, ya que estamos aproximando u_{tt}.Recordemos que

$O(\Delta t) = O(\Delta x)$ por la ecuación y por que $a(x,t)$ es acotada.

$$T_j^n = \frac{u(x,t+\Delta t) - u(x,t) + \Delta t(a(x,t)(u(x+\Delta x,t) - u(x-\Delta x,t))}{\Delta t^2}$$
$$- \frac{1}{2}(-a_t(x,t)\frac{u(x+\Delta x,t) - u(x-\Delta x,t)}{\Delta x} + a(x,t)\frac{\delta_x a(x,t)\delta_x u(x,t)}{\Delta x^2}) \implies$$
$$T_j^n = \frac{u_t(x,t)\Delta t + u_{tt}(x,t)\frac{\Delta t^2}{2} + \Delta t(a(x,t)(u_x(x,t)\Delta x + O(\Delta x^3))}{\Delta t^2}$$
$$- (-a_t(x,t)\frac{u_x(x,t)\Delta x + O(\Delta x^3)}{\Delta x} + a(x,t)\frac{\delta_x a(x,t)(u_x(x,t)\Delta x + O(\Delta x^3))}{\Delta x^2}) \implies$$
$$T_j^n = \frac{u_{tt}(x,t)\frac{\Delta t^2}{2} + O(\Delta t^3) + u_t(x,t)\Delta t - u_t(x,t)\Delta t + O(\Delta t^4)}{\Delta t^2}$$
$$- (-a_t u_x(x,t) + O(\Delta x^2) + a(x,t)\frac{(a(x,t)u_x(x,t))_x \Delta x^2 + O(\Delta x^3)}{\Delta x^2}) \implies$$
$$T_j^n = u_{tt} + O(\Delta t) - (-a_t u_x + O(\Delta x^2) + a(au_x)_x + O(\Delta x)) \implies$$
$$T_j^n = O(\Delta t) + O(\Delta x)$$

Con lo cual, el orden de convergencia del error de truncación resulta ser $O(\Delta t)$ en tiempo y $O(\Delta x)$ en espacio. De este modo, también se ifiere que el error de truncación tiende a 0 cuando $\Delta t \to 0$. Esto quiere decir que el método es **consistente**

5.2. Análisis de Fourier

Podemos reescribir el esquema LW de la siguiente manera (simplificándolo para que a sea constante):

$$U_j^{n+1} = U_j^n - \nu\Delta_{0x}U_j^n + \frac{1}{2}\nu^2\delta_x^2 U_j^n$$

El análisis de Fourier aplicado a esta forma resulta en la siguiente expresión:

$$\lambda = 1 - \nu[\frac{e^{ik\Delta x} - e^{-ik\Delta x}}{2}] + \frac{1}{2}\nu^2[e^{ik\Delta x} - 2 + e^{-ik\Delta x}]\lambda = 1 - i\nu\sin k\Delta x - 2\nu^2\sin^2\frac{k\Delta x}{2}$$

Tras unas manipulaciones llegamos a que:

$$|\lambda|^2 = 1 - 4\nu^2(1-\nu^2)\sin^4\frac{k\Delta x}{2}$$

Por lo tanto, llegamos a la conlusión de que el esquema es estable sí y solo sí $|\nu| \leq 1$, por lo que el cumplimiento del la condición CFL vuelve a implicar **estabilidad**

5.3. principio del máximo

Como hemos hecho con el esquema upwind, primero vamos a expresar el método por la ecuación correspondiente, y lo ampliaremos:

Recordemos primero que significában todos esos términos:

$$\Delta_o x U_j^n = (U_{j+1}^n - U_{j-1}^n)/2$$

$$\delta_x U_j^n = (U_{j+\frac{1}{2}}^n - U_{j-\frac{1}{2}}^n)$$

$$U_j^{n+1} = U_j^n - \nu\Delta_{0x}U_j^n + \frac{1}{2}(\Delta t)^2(-(a_t)_j^n\frac{\Delta_{0x}U_j^n}{\Delta x} + a_j^n\frac{\delta_x(a_j^n\delta_x U_j^n)}{(\Delta x)^2})$$

Que es equivalente a la siguiente ecuación:

$$U_j^{n+1} = \frac{1}{2}\nu(1-\nu)U_{j-1}^n + (1-\nu^2)U_j^n + \frac{1}{2}\nu(1-\nu)U_{j+1}^n$$

Como tenemos que cumplir la condición CFL para tener la posibilidad de convergencia, resulta que el térino que acompaña a U_{j+1}^n es negativo, ya que $\nu > 0$. Por tanto, U_j^{n+1} está dado por por una media ponderada de tres valores de el tiempo anterior, pero como dos de ellos son positivos y uno es negativo, puede existir una solución numérica que tenga oscilaciones que provocen un máximo o mínimo interno. Podeos obervar también en los dibujos que el esquema LW no cumple con el principio del máximo, ya que cuenta con soluciones intermedias dotadas de un máximo o un mínimo (si se cummpliese, el máximo y el mínimo se deberían alcanzar o bien en la solución inicial o bien en las condiciones de contorno). Por estos motivos, el esquema LW **no cumple con el principio del máximo**

5.4. Análisis del error del modelo

Recordemos la siguiente fórmula explicada con anterioridad:

$$\lambda = 1 - i\nu sin(k\Delta x) - 2\nu^2 sin^2(\frac{k\Delta x}{2})$$

Claramente, el argumento del número complejo λ

$$\begin{aligned} arg\lambda &= -tan^{-1}[\frac{\nu \sin k\Delta x}{1 - 2\nu^2 sin^2(\frac{k\Delta x}{2}}] \\ &= -tan^{-1}[\nu(\xi - \frac{\xi^3}{6} + ...)(1 - 2\nu^2\frac{\xi^2}{4} + ...] \\ &= -tan^{-1}[\nu\xi - \frac{1}{6}\nu(1-3\nu^2)\xi^3 + ...] \end{aligned}$$

Aplicando el lema que vimos con anterioridad, podemos comprobar que:

$$arg\lambda = -\nu\xi(1 - \frac{1}{6}(1-\nu^2)\xi^3 + ...$$

Con lo que podemos concluir que el error para $|\lambda|^2$ es $O(\xi^4)$, y el error relativo en fase será $O(\xi^2)$, después de verlo en la siguiente fórmula:

$$\frac{arg\lambda - truephase}{truephase} = -\frac{1}{6}(1-\nu^2)\xi^2 + ... = O(\xi^2)$$

5.5. conclusión

Dado que hemos podido ver que el método LW está dotado de **consistencia** y **estabilidad**, podemos afirmar que el método es, en definitiva, **convergente**. Otro factor a tener en cuenta, en torno a la comparación entre modelos, es en relación a la dispersión y achatamiento que sufren estos, a medida que aumentamos el tiempo final en el que evaluamos el método. Como hemos visto que el método LW tiene mayor convergencia de fase, esto quiere decir que en general **sufre de menos achatamiento y dispersión que el método upwind**, algo que podemos contrastar empíricamente con los experimentos. Otro factor a destacar, desfavorable para el método LW, es el hecho de que **no cumple con el principio del máximo**, lo que finalmente nos deja sin poder ofrecer una cota en la que se encuentre toda la solución numérica.