Métodos para Resolución Numérica de PDE's

Versión revisada

Notas de lectura
curso P.Dmitruk - A.Sztrajman

E.F. Lavia

Versión 1.2 - Sep.2017

Publicado digitalmente por E.F. Lavia, Buenos Aires.

Ediciones domésticas

E.F. Lavia, Métodos para resolución numérica de PDE's.
Creado en la República Argentina.
10 9 8 7 6 5 4 3 2 1

Índice general

Prefacio

Estas notas de lectura son básicamente una transcripción ligeramente ampliada del curso *Temas de Dinámica de Fluidos*, dictado por el Dr. Pablo Dmitruk durante el segundo cuatrimestre de 2013 en la Facultad de Ciencias Exactas y Naturales de la Universidad de Buenos Aires.

Es una materia dictada por el Dpto. de Física optativa para varias carreras, y para el doctorado en Física donde adopta el nombre "Métodos numéricos para ecuaciones en derivadas parciales". Según se puede leer en la descripción de la misma:

> *En esta materia se ven los principales métodos (diferencias finitas, elementos finitos, volúmenes finitos y pseudoespectrales) para la resolución numérica de ecuaciones diferenciales en derivadas parciales. Se describe la implementación práctica de los métodos a distintos tipos de ecuaciones diferenciales en diferentes áreas de la física con aplicaciones. Se incluyen, mediante ejemplos, nociones de programación para la implementación de los métodos. La materia es optativa para la Lic. en Cs. Físicas, para la Lic. en Cs. de la Atmósfera y en Oceanografía, y para el doctorado en Física (Facultad de Cs. Exactas y Naturales). En Física el puntaje máximo otorgado es 5 ptos.*

La razón más poderosa que justifica y explica la génesis de estas notas es LaTeX. No hay suficientes palabras para agradecerle a Donald Knuth el haber hecho que sus obsesiones estilísticas cristalizaran en TeX. También por supuesto a Leslie Lamport, y a la larga lista de personas que hicieron realidad este entorno de composición.

La función de estas notas, en cuanto a contenido y pertinencia como material de estudio y aprendizaje de la materia, puedo juzgarla como de complementaria. Al ser meras notas de curso, sufren de rispideces e inconexidades que solamente son pulidas y ligadas, respectivamente, cuando el material es sometido al paulatino proceso de corrección y ampliación. Dicho proceso necesita la ayuda de otros docentes, alumnos, especialistas e interesados. Ese proceso no ha tenido lugar con este material.

Entonces, en pocas palabras, digamos que es un material complementario a la cursada de la materia citada y que no reemplaza, ni busca hacerlo, al proceso de transmisición docente-alumno que está implicado en la asistencia a una clase.

No quiero dejar pasar la ocasión para agradecer a mis compañeros de cursada, el colombiano y Fede, con quienes discutí problemas computacionales de la materia en el bar del Pabellon I y a Alejandro Sztrajman, docente con un profundo amor por la programación y un contagioso entusiasmo por este tipo de emprendimientos. Finalmente, destacar que Laya Marconi tuvo, antes que yo, la iniciativa de pasar las notas de esta materia cuando se dictó en 2011, ¡habiendo aprendido LaTeXese mismo año! Esa iniciativa, además de valerosa, muestra que el trabajo sistemático es garantía para la epopeya.

Seguramente sobreviven en el texto errores; tipográficos y de mi entendimiento limitado de los temas. Exonero al Dr. Dmitruk de los mismos con estas líneas pero invito a los lectores a que los descubran y me contacten para corregirlos.

E.F. Lavia
Buenos Aires, Febrero de 2014.

Sobre la edición revisada

Me consta que la primera edición de esta obra recibió, al menos, dos pasajes del formato electrónico al papel lo cual representa mucho más de lo que hubiera imaginado al comenzar su redacción. Esta actual revisión se debe al prolijo escrutinio que realizara Adrián R. Lopez, con quien estoy desde este momento en deuda.

Buenos Aires, Septiembre de 2014.

Correcciones menores

Mientras utilizaba parte del material abarcado por estas notas encontré algunos *typos* y errores que corrijo pese a que seguramente sobreviven en el texto otros.

Buenos Aires, Septiembre de 2017.

Capítulo 1

Definiciones preliminares

§ 1. Discretización

Desde cierto punto de vista la descripción de la realidad que hace la física conlleva a dos tipos de modelos,

- De partículas; descripción discreta. Su evolución en el tiempo suele conducir a ecuaciones diferenciales ordinarias.

- De medios continuo; descripción continua. Su evolución temporal suele llevar a una ecuación en derivadas parciales.

No obstante ello, la solución numérica de los problemas del continuo implica también discretizar de manera que siempre trabajamos en el mundo discreto.

Algunas ecuaciones en derivadas parciales (PDE's de aquí en más[1]) son la de continuidad,

$$\frac{\partial \rho}{\partial t} = \boldsymbol{\nabla} \cdot \rho \mathbf{v}$$

la ecuación de ondas

$$\frac{\partial^2 \psi}{\partial t^2} = v^2 \frac{\partial^2 \psi}{\partial x^2}$$

o la ecuación de Schrödinger,

$$\left[-\frac{\hbar^2}{2m} \nabla^2 + V(x) \right] \psi = i\hbar \frac{\partial \psi}{\partial t}$$

[1] Esta abreviatura viene de las siglas en inglés; partial differential equations.

En todos estos casos la función es al menos de dos variables $\rho(x,t), \psi(x,t)$. La idea de este curso es entonces la resolución numérica de este tipo de ecuaciones.

El problema general que estaremos interesados en resolver en este curso es algo como

$$\frac{\partial \psi}{\partial t} = f\left(\frac{\partial \psi}{\partial x}, \frac{\partial^2 \psi}{\partial x^2}, ...\right), \tag{1.1}$$

y empezaremos descomponiendo (discretizando) el miembro derecho, el espacio, y luego el izquierdo, la parte temporal.

Capítulo 2

Diferencias finitas

§ 2. Derivadas discretas

Nos ocuparemos inicialmente del método de las diferencias finitas. Para ello supongamos una función

$$f = f(x) \qquad 0 \leq x \leq L$$

y comenzaremos discretizando el espacio. Resolveremos (1.1) con $\psi = f$ para cierto conjunto $\{x_j\}$ en el cual se dividió el espacio. Pese a que la función $f(x)$ es continua trabajaremos con los valores en los puntos x_j, que son los puntos de grilla o malla[1]. Consideremos una grilla equiespaciada (ver Figura 2.1), llamada grilla uniforme, a través de

$$x_j = j\Delta x \qquad j = 0, 1, ..., N$$

$$f(x_j) = f_j \qquad \{x_j\} \to f_j$$

La resolución de (1.1) comienza escribiendo las derivadas de f en forma discreta, como diferencias de valores de la función f.

La aproximación más sencilla es que la derivada sea la pendiente de la recta tangente, que es lo que se hace en análisis matemático pero sin tomar el límite de $\Delta x \to 0$. Así tendremos

$$\left(\frac{\partial f}{\partial x}\right)_j = \frac{f_{j+1} - f_j}{\Delta x} \qquad [\textit{forward scheme}]$$

[1] En inglés *grid*.

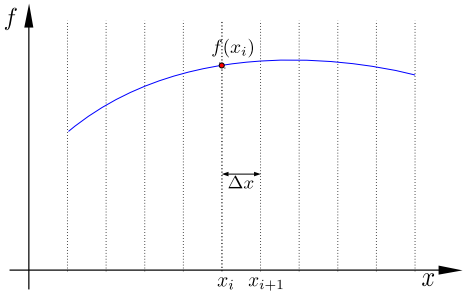


Figura 2.1 Grilla equiespaciada.

$$\left(\frac{\partial f}{\partial x}\right)_j = \frac{f_j - f_{j-1}}{\Delta x} \qquad [backward\ scheme]$$

Por supuesto, si Δx no es constante se tiene $\Delta x = x_{j+1} - x_j$ o $\Delta x = x_j - x_{j-1}$ en cada caso, respectivamente. Un tercer esquema es el centrado

$$\left(\frac{\partial f}{\partial x}\right)_j = \frac{f_{j+1} - f_{j-1}}{2\Delta x} \qquad [centered\ scheme]$$

Estos esquemas surgen de considerar las aproximaciones de Taylor de f en $x_{j+1} = x_j + \Delta x$,

$$f_{j+1} \equiv f(x_{j+1}) \approx f(x_j) + \Delta x \left(\frac{\partial f}{\partial x}\right)_j + \frac{1}{2}\Delta x^2 \left(\frac{\partial^2 f}{\partial x^2}\right)_j + \frac{1}{3!}\Delta x^3 \left(\frac{\partial^3 f}{\partial x^3}\right)_j + \frac{1}{4!}\Delta x^4 \left(\frac{\partial^3 f}{\partial x^3}\right)_j + \ldots \quad (2.1)$$

y la correspondiente en $x_{j-1} = x_j - \Delta x$

$$f_{j-1} \equiv f(x_{j-1}) \approx f(x_j) - \Delta x \left(\frac{\partial f}{\partial x}\right)_j + \frac{1}{2}\Delta x^2 \left(\frac{\partial^2 f}{\partial x^2}\right)_j - \frac{1}{3!}\Delta x^3 \left(\frac{\partial^3 f}{\partial x^3}\right)_j + \frac{1}{4!}\Delta x^4 \left(\frac{\partial^3 f}{\partial x^3}\right)_j + \ldots \quad (2.2)$$

Luego, desde (2.1) ya se puede intuir cómo sale el asunto, en efecto despejando en esa ecuación directamente

$$\frac{f(x_{j+1}) - f(x_j)}{\Delta x} \approx \left(\frac{\partial f}{\partial x}\right)_j + \frac{1}{2}\Delta x \left(\frac{\partial^2 f}{\partial x^2}\right)_j + \frac{1}{3!}\Delta x^2 \left(\frac{\partial^3 f}{\partial x^3}\right)_j + \ldots$$

y quedándonos hasta orden uno

$$\left(\frac{\partial f}{\partial x}\right)_j = \frac{f(x_{j+1}) - f(x_j)}{\Delta x} - \frac{1}{2}\Delta x \left(\frac{\partial^2 f}{\partial x^2}\right)_j$$

y se ve que el error va como Δx si suponemos que

$$\left(\frac{\partial^2 f}{\partial x^2}\right)_j \leq M,$$

la derivada segunda está acotada. Decimos que el error de truncamiento es de orden Δx.

Para el otro esquema, utilizando (2.2), se tiene equivalentemente

$$\left(\frac{\partial f}{\partial x}\right)_j = \frac{f(x_j) - f(x_{j-1})}{\Delta x} + \frac{1}{2}\Delta x \left(\frac{\partial^2 f}{\partial x^2}\right)_j$$

mientras que para el esquema centrado

$$\left(\frac{\partial f}{\partial x}\right)_j = \frac{f(x_{j+1}) - f(x_{j-1})}{\Delta x} - \frac{1}{3!}\Delta x^2 \left(\frac{\partial^3 f}{\partial x^3}\right)_j$$

es notorio que el error de truncamiento es menor (resultó de $\mathcal{O}\left(\Delta x^2\right)$) en general salvo alguna particularidad de $\partial_x^3 f$. Se ve esto gráficamente en la Figura 2.2 por la situación de la pendiente.

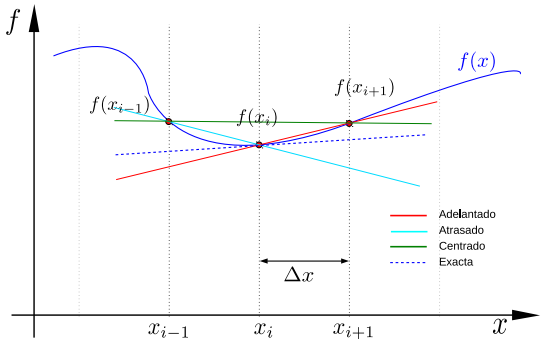


Figura 2.2 Los esquemas básicos para la derivada.

Puede utilizarse la misma inteligencia para hallar la derivada segunda. Haciendo la suma de las ecuaciones (2.1) y (2.2) resulta

$$f_{j+1} + f_{j-1} = 2f_j + \Delta x^2 \left(\frac{\partial^2 f}{\partial x^2}\right)_j + \frac{2}{4!}\Delta x^4 \left(\frac{\partial^4 f}{\partial x^4}\right)_j + \ldots$$

es decir,

$$\left(\frac{\partial^2 f}{\partial x^2}\right)_j = \frac{f_{j+1} + f_{j-1} - 2f_j}{\Delta x^2} - \frac{2}{4!}\Delta x^2 \left(\frac{\partial^4 f}{\partial x^4}\right)_j - ...,$$

una aproximación de orden $\mathcal{O}\left(\Delta x^2\right)$ centrada.

No se pueden obtener derivadas de orden más alto a partir de este esquema con dos vecinos y una recta.

Sumando y restando expresiones de Taylor para f de x_{j+1} y de x_{j-1} podemos hallar las expresiones más sencillas (que utilizan hasta tres puntos). Para una aproximación de orden mayor o derivada de orden superior necesito más puntos de grilla a considerar.

Tomo más vecinos y los relaciono con una ley no lineal (cuadrática, por ejemplo). Veamos una manera de generalizar este procedimiento.

§ 3. Aproximación general para la derivada

Todo este proceso puede hacerse de manera más sistemática. Podemos plantear, si queremos una derivada centrada, que la derivada enésima ($\partial^n f/\partial x^n$) sea una combinación lineal del tipo

$$Af_{j+2} + Bf_{j+1} + Cf_j + Df_{j-1} + Ef_{j-2} = \Delta x^n \left.\frac{\partial^n f}{\partial x^n}\right|_j, \tag{3.1}$$

donde las incógnitas son los coeficientes A, B, C, D, E.

Si tomamos cinco puntos para la expansión, como en (3.1), tomaremos en los diferentes polinomios de Taylor f_k los cinco primeros términos (que llegan hasta orden cuatro) y los retendremos en el miembro izquierdo. Los subsiguientes términos serán los correspondientes al error, el cual irá en principio como Δx^5 y se pasará al miembro derecho quedando expresado globalmente como $\mathcal{O}\left(\Delta x^5\right)$.

Entonces podemos escribir

$$Af_{j+2} + Bf_{j+1} + Cf_j + Df_{j-1} + Ef_{j-2} = \Delta x^n \left.\frac{\partial^n f}{\partial x^n}\right|_j + \mathcal{O}(\Delta x^5)$$

donde insistimos en que el miembro izquierdo tiene derivadas de f de a lo sumo orden cuatro. Las derivadas de orden subsiguiente están, como dijimos, en el término $\mathcal{O}\left(\Delta x^5\right)$. Si ahora consideramos cada uno de los polinomios f_k y

agrupamos extrayendo como factor común las derivadas resulta

$$(A+B+C+D+E)\,f_j+(2A+B-D-2E)\,\Delta x\,\left.\frac{\partial f}{\partial x}\right|_j+\left(2A+\frac{B}{2}+\frac{D}{2}+2E\right)\Delta x^2\,\left.\frac{\partial^2 f}{\partial x^2}\right|_j+\left(\frac{4A}{3}+\frac{B}{6}-\frac{D}{6}-\frac{4E}{3}\right)\Delta x^3\,\left.\frac{\partial^3 f}{\partial x^3}\right|_j+\left(\frac{2A}{3}+\frac{B}{24}+\frac{D}{24}+\frac{2E}{3}\right)\Delta x^4\,\left.\frac{\partial^4 f}{\partial x^4}\right|_j=\Delta x^n\,\left.\frac{\partial^n f}{\partial x^n}\right|_j+\mathcal{O}(\Delta x^5)$$

Para fijar ideas supongamos que queremos calcular la derivada segunda, entonces $n=2$ y despreciando el término $\mathcal{O}\left(\Delta x^5\right)$ resulta

$$(A+B+C+D+E)\,f_j+(2A+B-D-2E)\,\Delta x\,\left.\frac{\partial f}{\partial x}\right|_j+\left(2A+\frac{B}{2}+\frac{D}{2}+2E-1\right)\Delta x^2\,\left.\frac{\partial^2 f}{\partial x^2}\right|_j+\left(\frac{4A}{3}+\frac{B}{6}-\frac{D}{6}-\frac{4E}{3}\right)\Delta x^3\,\left.\frac{\partial^3 f}{\partial x^3}\right|_j+\left(\frac{2A}{3}+\frac{B}{24}+\frac{D}{24}+\frac{2E}{3}\right)\Delta x^4\,\left.\frac{\partial^4 f}{\partial x^4}\right|_j=0$$

pero como tienen que ser nulos cada uno de los paréntesis por separado ello es equivalente al siguiente sistema matricial

$$\begin{pmatrix}1 & 1 & 1 & 1 & 1\\ 2 & 1 & 0 & -1 & -2\\ 2 & 1/2 & 0 & 1/2 & 2\\ 4/3 & 1/6 & 0 & -1/6 & -4/3\\ 2/3 & 1/24 & 0 & 1/24 & 2/3\end{pmatrix}\begin{pmatrix}A\\B\\C\\D\\E\end{pmatrix}=\begin{pmatrix}0\\0\\1\\0\\0\end{pmatrix}\tag{3.2}$$

cuya solución es

$$A=-\frac{1}{12},\quad B=\frac{4}{3},\quad C=-\frac{5}{2},\quad D=\frac{4}{3},\quad E=-\frac{1}{12}.$$

Si volvemos al término que agrupaba las derivadas subsiguientes y escribimos explícitamente las derivadas de orden cinco con sus coeficientes resulta

$$\mathcal{O}\left(\Delta x^5\right)=-\left(A\frac{2^5}{5!}+\frac{B}{5!}-\frac{D}{5!}-E\frac{2^5}{5!}\right)\Delta x^5\,\left.\frac{\partial^5 f}{\partial x^5}\right|_j+\mathcal{O}\left(\Delta x^6\right)$$

y dada la solución recién calculada para los coeficientes podemos verificar que la misma hace nulo el paréntesis de modo que en realidad los términos que constituyen el error van como Δx^6.

Entonces

$$-\frac{1}{12}f_{j+2}+\frac{4}{3}f_{j+1}-\frac{5}{2}f_j+\frac{4}{3}f_{j-1}-\frac{1}{12}f_{j-2}=\Delta x^2\left.\frac{\partial^2 f}{\partial x^2}\right|_j+\mathcal{O}(\Delta x^6)$$

y hemos obtenido una expresión para la derivada segunda con error de orden $\mathcal{O}\left(\Delta x^4\right)$, es decir:

$$\left.\frac{\partial^2 f}{\partial x^2}\right|_j=\frac{-f_{j+2}+16f_{j+1}-30f_j+16f_{j-1}-f_{j-2}}{12\Delta x^2}+\mathcal{O}(\Delta x^4).$$

Evidentemente la elección de n afecta la posición del 1 en el vector de constantes de la derecha del sistema (3.2) y también el orden de precisión. Para cinco puntos, como hemos tomado, podremos en principio tener fórmulas de derivada primera con error como Δx^4, derivada segunda con error Δx^3 y finalmente derivada cuarta con error Δx. Todo ello puede modificarse y aumentar el orden de precisión si se diera el caso, como sucede en este ejemplo, de que dentro del término de error se anule el orden siguiente.

Asimismo, si lo que se quiere es una derivada no centrada tendremos que partir de otra combinación lineal diferente; una que tenga en cuenta los puntos hacia una u otra dirección de x_j.

§ 3.1. Generalización de la fórmula para derivadas

Una forma de generalizar la obtención de un esquema de diferencias finitas para la derivada de orden m es a través de

$$\left.\frac{\partial^m f}{\partial x^m}\right|_j=\sum_{i=j-q}^{j+p}\gamma_i f_i \tag{3.3}$$

donde necesito $p+q\geq m$ que es la cantidad de puntos a utilizar[2] y donde los coeficientes γ_i surgen de resolver el sistema

$$\sum_{i=j-q}^{j+p}\frac{1}{n!}\gamma_i(x_i-x_j)^n=c_n\qquad\text{con}\quad c_n=\begin{cases}0 & \text{si } n\neq m\\ 1 & \text{si } n=m\end{cases}$$

Haciendo este proceso y si el paso de discretización es constante, es decir si se cumple que

$$x_i-x_j=\Delta x$$

[2]Entonces p y q son valores constantes que dicen cuántos puntos vamos a utilizar a izquierda y a derecha, respectivamente. Si $p=q$ el esquema es, evidentemente, centrado.

para todo par de puntos i, j consecutivos obtenemos una expresión para la derivada de

$$\mathcal{O}\left(\Delta x^{p+q-m+1}\right),$$

o bien de

$$\mathcal{O}\left(\Delta x^{p+q-m+2}\right)$$

si el esquema es centrado.

EJEMPLO 2.1. Ejemplos de derivadas
Veamos algún ejemplo de la fórmula general (3.3) para $m = 1$, $q = 1$ y $p = 0$. Eso lleva a

$$\left(\frac{\partial f}{\partial x}\right)_j = \gamma_{j-1} f_{j-1} + \gamma_j f_j$$

donde los coeficientes se extraen usando los valores de c_n

$$\left.\begin{array}{ll} n=0 & 0 = \gamma_{j-1} + \gamma_j \\ n=1 & 1 = -\Delta x \gamma_{j-1} \end{array}\right\}$$

que tiene solución

$$\gamma_{j-1} = -\frac{1}{\Delta x} \qquad \gamma_j = \frac{1}{\Delta x},$$

lo cual conduce a nuestra expresión para la derivada primera bajo el esquema que llamamos *backwar* (atrasado)

$$\left(\frac{\partial f}{\partial x}\right)_j = \frac{f_j - f_{j-1}}{\Delta x}, \quad \mathcal{O}\left(\Delta x\right).$$

Este esquema atrasado tiene aplicación directa en el caso de contornos con nodos en una pared (si $j = N$ es el último nodo, en la pared, calculamos la derivada allí utilizando el valor f_N y f_{N-1}).
Ahora calcularemos una aproximación para la derivada con dos puntos hacia atrás ($q = 2, p = 0$)

$$\left(\frac{\partial f}{\partial x}\right)_j = \gamma_{j-2} f_{j-2} + \gamma_{j-1} f_{j-1} + \gamma_j f_j$$

$$\left.\begin{array}{ll} n=0 & 0 = \gamma_{j-2} + \gamma_{j-1} + \gamma_j \\ n=1 & 1 = -2\Delta x \gamma_{j-2} - \Delta x \gamma_{j-1} \\ n=2 & 1 = 2\Delta x^2 \gamma_{j-2} - \frac{\Delta x^2}{2} \gamma_{j-1} \end{array}\right\}$$

luego

$$\gamma_{j-2} = \frac{1}{2\Delta x} \qquad \gamma_{j-1} = -\frac{2}{\Delta x} \qquad \gamma_j = \frac{3}{2\Delta x},$$

de manera que finalmente

$$\left(\frac{\partial f}{\partial x}\right)_j = \frac{f_{j-2} - 4f_{j-1} + 3f_j}{2\Delta x}.$$

§ 3.2. Cálculo de derivadas con polinomios

Otra forma de obtener aproximaciones en diferencias finitas, debida a Fornberg, está basada en el uso de polinomios interpoladores que pasen por tantos

puntos $f(x_k)$ como sea el orden que se requiere. Es decir que para dos puntos, polinomio de orden uno, utilizaremos una recta, para tres puntos una parábola y así sucesivamente.

La aproximación es considerar que la derivada de la función f será la derivada de ese polinomio $P(x)$. Para conseguir un polinomio que pase por los puntos en cuestión se suele utilizar el interpolador de Lagrange. Sea por ejemplo el polinomio de orden dos, $P_2(x)$ que pasa por x_k con $k = j-1, j, j+1$, es decir

$$P_2(x) = \frac{f_{j+1}(x-x_{j-1})(x-x_j)}{(x_{j+1}-x_{j-1})(x_{j+1}-x_j)} + \\ \frac{f_j(x-x_{j-1})(x-x_{j+1})}{(x_j-x_{j-1})(x_j-x_{j+1})} + \frac{f_{j-1}(x-x_j)(x-x_{j+1})}{(x_{j-1}-x_j)(x_{j-1}-x_{j+1})},$$

donde se puede ver fácilmente por sustitución que verifica

$$P_2(x_j) = f_j \qquad P_2(x_{j-1}) = f_{j-1} \qquad P_2(x_{j+1}) = f_{j+1}.$$

Aproximamos entonces

$$\left(\frac{\partial f}{\partial x}\right)_j \quad \text{por} \quad \left(\frac{\partial P_2}{\partial x}\right)_j$$

y es

$$\left(\frac{\partial P_2}{\partial x}\right)_j = \frac{f_{j+1}\Delta x}{2\Delta x^2} + \frac{f_j(-\Delta x)}{-\Delta x^2} + \frac{f_j \Delta x}{-\Delta x^2} + \frac{f_{j-1}(-\Delta x)}{2\Delta x^2}$$

$$\left(\frac{\partial P_2}{\partial x}\right)_j = \frac{f_{j+1}-f_{j-1}}{2\Delta x}, \qquad \mathcal{O}\left(\Delta x^2\right)$$

de modo que resultó ser la diferencia centrada de orden dos. Parece ser un método más directo de obtener las derivadas antes de utilizar la expresión general y resolver el sistema de ecuaciones.

Si hacemos la derivada dos veces, usando los puntos $j-2, j-1, j$ resulta

$$\left(\frac{\partial^2 P_2}{\partial x^2}\right)_j = \frac{f_{j-2} - 2f_{j-1} + f_j}{\Delta x^2}.$$

Las tablas que se muestran a continuación muestran los coeficientes que acompañan a las derivadas primera y segunda, centradas y no centradas.

§ 4. Análisis espectral

La idea es desarrollar en Fourier la función $f(x)$ y ver cuánto difiere su derivada de la aproximación, según el modo Fourier. Supongo

$$f(x) = A\, \mathrm{e}^{ikx2\pi/L}, \quad k \in \mathbb{Z},$$

Cuadro 2.1 Coeficientes para derivadas centradas

Derivadas	Coeficientes						
$\Delta x\, \partial_x f_j$	$j-3$	$j-2$	$j-1$	j	$j+1$	$j+2$	$j+3$
$\mathcal{O}\left(\Delta x^2\right)$			-1/2	0	1/2		
$\mathcal{O}\left(\Delta x^4\right)$		1/12	-2/3	0	2/3	-1/12	
$\mathcal{O}\left(\Delta x^6\right)$	-1/60	3/20	-3/4	0	3/4	-3/20	1/60

Cuadro 2.2 Coeficientes para derivadas no centradas

Derivadas	Coeficientes			
$\Delta x\, \partial_x f_j$	j	$j \pm 1$	$j \pm 2$	$j \pm 3$
$\mathcal{O}\left(\Delta x\right)$	$\mp$ 1	$\pm$ 1		
$\mathcal{O}\left(\Delta x^2\right)$	$\mp$ 3/2	$\pm$ 2	$\mp$ 1/2	
$\mathcal{O}\left(\Delta x^3\right)$	$\mp$ 11/6	$\pm$ 3	$\mp$ 3/2	$\pm$ 1/3

luego

$$\frac{\partial f(x)}{\partial x} = ik\frac{2\pi}{L}f(x). \tag{4.1}$$

Sea ahora que usamos diferencias finitas para evaluar la derivada de f y comparar con (4.1) que es exacto. La derivada discreta en el esquema centrado era

$$\left(\frac{\partial f}{\partial x}\right)_j = \frac{f_{j+1} - f_{j-1}}{2\Delta x} \tag{4.2}$$

entonces

$$\left(\frac{\partial f}{\partial x}\right)_j = \frac{A\,\mathrm{e}^{ikx_{j+1}2\pi/L} - A\,\mathrm{e}^{ikx_{j-1}2\pi/L}}{2\Delta x}$$

pero si el espaciado de la grilla es uniforme puedo escribir

$$\left(\frac{\partial f}{\partial x}\right)_j = \frac{A\,\mathrm{e}^{ikx_j 2\pi/L}(\,\mathrm{e}^{ik\Delta x 2\pi/L} - \,\mathrm{e}^{-ik\Delta x 2\pi/L})}{2\Delta x}$$

$$\left(\frac{\partial f}{\partial x}\right)_j = \frac{A\,i\,\mathrm{e}^{ikx_j 2\pi/L}[\mathrm{sen}(k\Delta x 2\pi/L)]}{\Delta x}$$

de modo que finalmente

$$\left(\frac{\partial f}{\partial x}\right)_j = i\frac{\mathrm{sen}(k\Delta x 2\pi/L)}{\Delta x}f_j,$$

Cuadro 2.3 Coeficientes para derivadas segundas centradas

Derivadas	Coeficientes						
$\Delta x^2\, \partial_x^2 f_j$	$j-3$	$j-2$	$j-1$	j	$j+1$	$j+2$	$j+3$
$\mathcal{O}\left(\Delta x^2\right)$			1	-2	1		
$\mathcal{O}\left(\Delta x^4\right)$		-1/12	4/3	-5/2	4/3	-1/12	
$\mathcal{O}\left(\Delta x^6\right)$	-1/90	-3/20	3/2	-49/18	3/2	-3/20	1/90

Cuadro 2.4 Coeficientes para derivadas segundas no centradas

Derivadas	Coeficientes			
$\Delta x^2\, \partial_x^2 f_j$	j	$j\pm1$	$j\pm2$	$j\pm3$
$\mathcal{O}\left(\Delta x\right)$	1	-2	1	
$\mathcal{O}\left(\Delta x^2\right)$	2	-5	4	-1

y entonces vemos que esta expresión difiere de (4.1) pero coinciden en el límite $\Delta x \to 0$, pues $\text{sen}(\beta x)/x \to \beta$ con $x \to 0$.

Podemos decir entonces que

$$\left(\frac{\partial f}{\partial x}\right)_j = ik\frac{2\pi}{L} f_j = i\omega_k f_j \quad \text{con} \quad \omega_k = k\frac{2\pi}{L}$$

es exacto, mientras que

$$\left(\frac{\partial f}{\partial x}\right)_j = i\omega'_k f_j \quad \text{con} \quad \omega'_k = \frac{\text{sen}(k\Delta x 2\pi/L)}{\Delta x}$$

es la aproximación. Estas funciones $\omega(k)$ son las relaciones de dispersión de la derivada con respecto al índice del modo k.

En la Figura 4.3 se comparan las dispersiones de la derivada exacta y de la aproximación centrada (4.2). A los efectos del cálculo de los gráficos mostrados se consideró la discretización

$$0 < x < L \qquad x_j = j\Delta x$$

donde si tomamos $\Delta x = 0{,}1$ y $L = 100$ el número N de modos Fourier verificará

$$N\Delta x = L \qquad j = 0, 1, ..., N$$

y por ende el gráfico de $\omega(k)$ irá hasta $N/2 = 500$.

La relación de dispersión del método de diferencias finitas, para el esquema centrado, es parecida a la relación exacta sólo para k chicos. El problema entonces

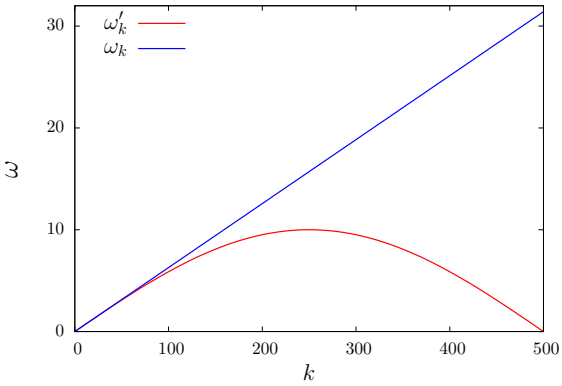


Figura 4.3 Relaciones de dispersión para la derivada exacta (ω_k) y aproximada (ω'_k) según un esquema centrado.

es querer ajustar una función que oscila de manera demencial con un método basado en pocos puntos.

Si aproximamos el seno

$$\text{sen}\left(k\Delta x\frac{2\pi}{L}\right) \approx k\Delta x\frac{2\pi}{L} - \frac{1}{3!}\left(k\Delta x\frac{2\pi}{L}\right)^3 + \mathcal{O}\left(\left(k\Delta x\frac{2\pi}{L}\right)^5\right)$$

$$\omega'_k \approx k\frac{2\pi}{L}\left[1 - \frac{1}{3!}(k\Delta x 2\pi/L)^2 + \mathcal{O}\left((k\Delta x 2\pi/L)^4\right)\right]$$

$$\omega'_k \approx \omega_k\left[1 - \frac{1}{3!}(k\Delta x 2\pi/L)^2 + \mathcal{O}\left((k\Delta x 2\pi/L)^4\right)\right]$$

luego el error en ω'_k va como

$$(k\Delta x 2\pi/L)^2.$$

Una manera de mejorar diferencias finitas es aproximar mejor a ω_k con un *compact* FDM (S. Lele, *Compact Finite Difference Schemes With Spectral-like Resolution*, J C Phys **103**, 16-42, 1992).

§ 5. Consistencia de los métodos

Supongamos una PDE, como ser la ecuación de difusión

$$\frac{\partial f}{\partial t} = \nu\frac{\partial^2 f}{\partial x^2},$$

donde el miembro derecho, la parte espacial, se hace de la manera vista por diferencias finitas y el miembro izquierdo también se discretiza[3] utilizando un índice temporal n. Es decir que hacemos

$$\frac{\partial^2 f}{\partial x^2} \xrightarrow{\text{centered}} \frac{f_{j+1}^n + f_{j-1}^n - 2f_j^n}{\Delta x^2}$$

$$\frac{\partial f}{\partial t} \xrightarrow{\text{forward}} \frac{f_j^{n+1} - f_j^n}{\Delta t},$$

estando la discretización definida por

$$x_j = j\Delta x \qquad j = 0, 1, ..., N$$

$$t_n = n\Delta t \qquad n = 0, 1, ..., M$$

$$f(x_j, t_n) = f_j^n$$

Luego reemplazamos las discretizaciones de las derivadas por las derivadas continuas en la ecuación de difusión de manera que

$$\frac{f_j^{n+1} - f_j^n}{\Delta t} = \nu \frac{f_{j+1}^n + f_{j-1}^n - 2f_j^n}{\Delta x^2}$$

es una *nueva* ecuación para la cual f es solución (es la ecuación "aproximada", para distinguirla de la ecuación original). Sea ahora F solución exacta de la ecuación diferencial de manera que se cumple

$$\frac{\partial F}{\partial t} = \nu \frac{\partial^2 F}{\partial x^2}.$$

No necesariamente esta F verifica la ecuación aproximada, la diferencia es el llamado error de discretización (o truncamiento). Si denominamos a ese error $D(F)$ será

$$D(F) = \frac{F(x_j, t_{n+1}) - F(x_j, t_n)}{\Delta t} - \nu \frac{F(x_{j+1}, t_n) + F(x_{j-1}, t_n) - 2F(x_j, t_n)}{\Delta x^2}$$

y puedo hacer aproximaciones de Taylor dentro de esta expresión

$$\begin{aligned} D(F) = \frac{1}{\Delta t} &\left[F(x_j, t_n) + \Delta t \partial_t F_j^n + (1/2)\Delta t^2 \partial_t^2 F_j^n + ... - F(x_j, t_n)\right] \\ &- \frac{\nu}{\Delta x^2} \left[F(x_j, t_n) + \Delta x \partial_x F_j^n + (1/2)\Delta x^2 \partial_x^2 F_j^n + ... \right. \\ &\left. + F(x_j, t_n) - \Delta x \partial_x F_j^n + (1/2)\Delta x^2 \partial_x^2 F_j^n - ... - 2F(x_j, t_n)\right] \end{aligned}$$

[3] En el capítulo siguiente veremos de manera detallada la discretización temporal. La introducimos aquí de manera algo forzada y sucinta para motivar el concepto de consistencia de un método.

y luego de trabajar un poco, cancelando y simplificando, arribar a

$$D(F) = \left[\partial_t F_j^n + (1/2)\Delta t \partial_t^2 F_j^n + \mathcal{O}\left(\Delta t^2\right)\right] + \left[-\nu \partial_x^2 F_j^n + \mathcal{O}\left(\Delta x^2\right)\right]$$

y dado que F es solución de la ecuación exacta se tiene

$$D(F) = (1/2)\Delta t \partial_t^2 F_j^n + \mathcal{O}\left(\Delta t^2\right) + \mathcal{O}\left(\Delta x^2\right) = \mathcal{O}\left(\Delta t\right) + \mathcal{O}\left(\Delta x^2\right).$$

Entonces es claro que se verifica

$$D(F) \to 0 \quad \text{si} \begin{cases} \Delta t \to 0 \\ \Delta x \to 0, \end{cases}$$

lo que significa que el error de discretización tiende a cero cuando Δt y Δx tienden a cero de manera independiente. En este caso se dice que el esquema es consistente.

Por supuesto, lo que se quiere es que

$$f_j^n = f(x_j, t_n) \to F(x_j, t_n),$$

que la solución f de la ecuación aproximada tienda a la solución F de la ecuación exacta para los puntos de grilla. Esto significa que el esquema es convergente.

Si el método es consistente y estable se da que es convergente. La sóla consistencia del método no basta para garantizar la convergencia. Esto es el teorema de Lax, que puede escribirse brevemente como

$$\text{Consistente} + \text{Estable} \Rightarrow \text{Convergente}$$

El aspecto de estabilidad de los métodos será abordado también en el próximo capítulo.

Capítulo 3

Esquemas temporales

§ 6. Métodos no iterativos

El tiempo en general siempre avanza y en la práctica más que derivar tenemos que integrar[1], es decir hacer una cuadratura. Supongamos $q = q(t)$ y

$$\frac{dq}{dt} = f(q,t) \tag{6.1}$$

donde $q(0)$ evoluciona por integración hacia $q(t)$. Esta es una ecuación diferencial ordinaria (ODE). Asimismo una PDE, luego de discretizada la parte espacial nos queda un sistema de ODE's para cada valor espacial. Veamos un par de ejemplos.

EJEMPLO 3.1. Oscilador armónico
El oscilador armónico se puede plantear

$$m\frac{dv}{dt} = -kx \qquad v = \frac{dx}{dt},$$

donde debemos notar que no estamos utilizando la ecuación de segundo orden para x sino el par de ecuaciones de primer orden; de esta manera estamos en la forma (6.1). Si

$$\mathbf{q} = (x, v)$$

puedo definir el sistema

$$\frac{d\mathbf{q}}{dt} = L\mathbf{q}$$

siendo

$$L = \begin{pmatrix} 0 & 1 \\ -k/m & 0 \end{pmatrix}.$$

Otra forma es utilizar un formalismo complejo,

$$q = x - i\frac{v}{\omega}, \quad \omega = \sqrt{\frac{k}{m}}$$

[1] Integramos porque en general se puede despejar dq/dt y la ecuación pasa a ser una ODE.

donde la ecuación a resolver es

$$\frac{dq}{dt} = i\omega q$$

Ambos tratamientos son casos particulares de (6.1).

EJEMPLO 3.2. Ecuación de difusión
Otro ejemplo podría ser la ecuación de difusión,

$$\frac{\partial \phi}{\partial t} = \nu \frac{\partial^2 \phi}{\partial x^2},$$

con $\phi = \phi(x,t)$. Se puede pensar como

$$q \equiv \phi(x,t)$$

$$f(q,t) = \nu \frac{\partial^2 q}{\partial x^2}$$

Si tratamos a las derivadas como formas discretas

$$\phi_i = \phi(x_i,t) = q_i \qquad x_i = i\Delta x \quad i = 0,1,...,N$$

$$\mathbf{q} = \{\phi_i\} = \{q_i\} \qquad i = 0,1,...,N$$

vemos que el $\mathbf{q}$ dará origen a una matriz donde habrá relaciones entre los diversos q_i. Si consideramos una expansión en funciones base $\{\psi_j\}$ tendremos un sistema para los coeficientes de la expansión. Pero eso, eso es otra historia que llegará oportunamente.

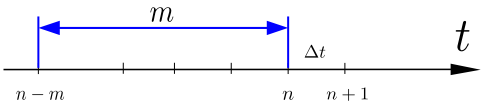
Volviendo al problema general (6.1) discretizaremos el tiempo

$$t \to n\Delta t \quad n = 0,1,...,N$$

de forma que integrando (6.1) será

$$q\Big|_{(n-m)\Delta t}^{(n+1)\Delta t} = \int_{(n-m)\Delta t}^{(n+1)\Delta t} f(q(t),t)dt$$

siendo m un número fijo. El siguente esquema muestra los límites en la recta real.



Utilizaremos la nomenclatura

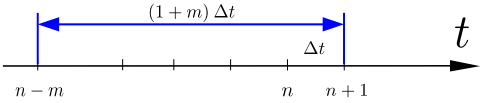
$$q(n\Delta t) \equiv q^n \qquad q^{n+1} \equiv q((n+1)\Delta t)$$

$$f^n \equiv f(q^n, n\Delta t) \qquad f^{n+1} \equiv f(q^{n+1}, (n+1)\Delta t)$$

Ahora aproximaremos el miembro derecho de (6.1) con una sumatoria de valores ponderados. El intervalo de integración es

$$(n+1)\Delta t - (n-m)\Delta t = (1+m)\Delta t$$

según se ilustra en la figura



de manera que la versión discreta de la integral (6.1) será

$$\frac{q^{n+1}-q^{n-m}}{(1+m)\Delta t}=\beta f^{n+1}+\alpha_n f^n+\alpha_{n-1}f^{n-1}+...+\alpha_{n-l}f^{n-l} \qquad (6.2)$$

donde el parámetro l dice cuantos puntos se tomarán y puede ser igual o diferente que m.

El problema está determinado por los coeficientes $\beta, \alpha_n, ..., \alpha_{n-l}$.

El coeficiente β está asociado a f^{n+1}. Como f^{n+1} depende de q^{n+1}, que es la incógnita a despejar en función de los q^n, para calcularla será necesario poder extraerla desde f^{n+1} y pasarla al miembro izquierdo. En problemas lineales suele ser fácil. Cuando $\beta \neq 0$ se dice que el método es implícito porque hay que invertir el miembro derecho para poder despejar q^{n+1}.

La ecuación en diferencias (6.2) se usa para aplicarle la solución exacta del problema y plantear una aproximación de Taylor

$$\frac{1}{(1+m)\Delta t}\left[q+q'\Delta t+q''\frac{\Delta t^2}{2}+...-\left(q-q'm\Delta t+q''\frac{(m\Delta t)^2}{2}+...\right)\right]=$$
$$\beta\left[f+f'\Delta t+f''\frac{\Delta t^2}{2}+...\right]+\alpha_n f+$$
$$\alpha_{n-1}\left[f-f'\Delta t+f''\frac{\Delta t^2}{2}+...\right]+...+\alpha_{n-l}\left[...\right]+\varepsilon$$

donde hemos introducido todos los términos de orden superior de las diferentes expansiones en el término general ε.

Agrupando ahora términos según potencias de Δt pasamos todo del lado izquierdo y resulta

$$q'(1-\beta-\alpha_n-...-\alpha_{n-l})+$$
$$q''\Delta t\left[\frac{1}{2}\left(\frac{1-m^2}{1+m}\right)-\beta+\alpha_{n-1}+2\alpha_{n-2}+3\alpha_{n-3}+...+l\alpha_{n-l}\right]$$
$$+q'''\frac{\Delta t^2}{2}\left[\frac{1}{3}\left(\frac{1+m^3}{1+m}\right)-\beta+...\right]=\varepsilon.$$

Entonces desearíamos que

$$\varepsilon\to 0 \quad \text{si} \quad \Delta t\to 0$$

y vemos que el primer miembro no lo cumple. Para ser consistente necesitamos además que

$$1 - \beta - \alpha_n - ... - \alpha_{n-l} = 0$$

es decir que

$$\beta + \alpha_n + \alpha_{n-1} + ... + \alpha_{n-l} = 1. \tag{6.3}$$

Esta condición (6.3) nos asegura un método de $\mathcal{O}(\Delta t)$; si además el término en $q''\Delta t$ es nulo el método será de orden $\mathcal{O}\left(\Delta t^2\right)$.

Eligiendo los α_i, β puedo hacer que $\varepsilon = \mathcal{O}\left(\Delta t^{l+2}\right)$. Si el método es explícito ($\beta = 0$) en general puedo ajustar el error como $\varepsilon = \mathcal{O}\left(\Delta t^{l+1}\right)$, no obstante subsiste el m que puede servir para anular algún término.

§ 6.1. Algunos métodos usuales

Tomando $m = 0$ y $l = 0$ resulta

$$\frac{q^{n+1} - q^n}{\Delta t} = \alpha_n f^n$$

y según la condición (6.3) se requiere $\alpha_n = 1$ para que el método sea de $\mathcal{O}(\Delta t)$. Es el método de Euler adelantado

$$\frac{q^{n+1} - q^n}{\Delta t} = f^n \qquad \varepsilon = q''\frac{\Delta t}{2}, \quad \mathcal{O}(\Delta t)$$

Tomando $m = 0$ y $l > 0$ definimos los métodos de Adams-Bashforth, explícitos ($\beta = 0$). Sea $l = 1$

$$\frac{q^{n+1} - q^n}{\Delta t} = \alpha_n f^n + \alpha_{n-1} f^{n-1}$$

y tendremos

$$1 = \alpha_n + \alpha_{n-1}$$

con

$$\varepsilon = q''\Delta t[\alpha_{n-1} + 1/2] + \mathcal{O}\left(\Delta t^2\right)$$

y puedo tomar $\alpha_n = -1/2$ y entonces resulta $\alpha_{n-1} = 3/2$ de modo que tenemos un método de orden $\mathcal{O}\left(\Delta t^2\right)$

$$\frac{q^{n+1} - q^n}{\Delta t} = \frac{3}{2} f^n - \frac{1}{2} f^{n-1} \qquad \mathcal{O}\left(\Delta t^2\right).$$

El cuadro 3.1 resume los diferentes coeficientes para los métodos de Adams-Bashforth.

La desventaja es el uso de memoria porque hay que guardar valores anteriores, pero en compensación evalúa pocas veces f respecto de otros métodos.

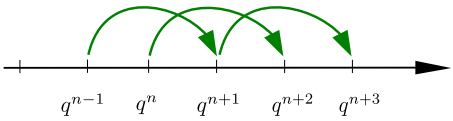
Cuadro 3.1 Coeficientes para Adams-Bashforth

l	α_n	α_{n-1}	α_{n-2}	α_{n-3}	ε
1	3/2	-1/2			$\mathcal{O}\left(\Delta t^2\right)$
2	23/12	-4/3	5/12		$\mathcal{O}\left(\Delta t^3\right)$
3	55/24	-59/24	39/24	-9/24	

Tomando $m = 1$ y $l = 0$ arribamos al método *leapfrog* (salto de rana)

$$\frac{q^{n+1} - q^{n-1}}{2\Delta t} = \alpha_n f^n$$

cuya idea pictórica aparece a continuación



y explica el porqué de su nombre. El método resulta de $\mathcal{O}\left(\Delta t^2\right)$ porque el m forzado a 1 anula el término de Δt. Estos métodos son problemáticos para soluciones oscilatorias.

Tomemos ahora $\beta \neq 0$, $m = 0$ y $l = 0$

$$\frac{q^{n+1} - q^n}{\Delta t} = \beta f^{n+1} + \alpha_n f^n$$

y la ecuación de consistencia

$$\beta + \alpha_n = 1$$

y esto da origen a dos métodos.

Si $\beta = 1, \alpha_n = 0$ tenemos $\mathcal{O}\left(\Delta t\right)$ y es el método de Euler atrasado

$$\frac{q^{n+1} - q^n}{\Delta t} = f^{n+1}.$$

Si en cambio es $\beta = 1/2, \alpha_n = 1/2$ será $\mathcal{O}\left(\Delta t^2\right)$ y es el llamado método trapezoidal

$$\frac{q^{n+1} - q^n}{\Delta t} = \frac{1}{2}(f^{n+1} + f^n)$$

Tomando $\beta \neq 0, m = 0$ y $l > 0$ estamos en presencia de los métodos de Adams-Moulton. El cuadro 3.2 muestra los coeficientes respectivos.

Cuadro 3.2 Coeficientes para Adams-Moulton

l	β	α_n	α_{n-1}	α_{n-2}	ε
1	5/12	8/12	-1/12		$\mathcal{O}\left(\Delta t^3\right)$
2	9/21	19/24	-5/24	1/24	$\mathcal{O}\left(\Delta t^4\right)$

§ 7. Métodos iterativos

Un método iterativo se consigue cuando en un esquema implícito calculamos el valor f^{n+1} utilizando el q^{n+1} dado por un método explícito.

Supongamos el atrasado

$$\frac{q^{n+1}-q^n}{\Delta t}=f^{n+1}$$

donde

$$f^{n+1}=f(q^{n+1},(n+1)\Delta t)$$

$$f^{n+1*}=f(q^{n+1*},(n+1)\Delta t)$$

y donde q^{n+1*} lo extraigo de un método explícito. Este paso previo es el llamado paso predictor, porque q^{n+1*} es en realidad una predicción. Podríamos extraerlo del esquema explícito de Euler (adelantado),

$$\frac{q^{n+1*}-q^n}{\Delta t}=f^n$$

con

$$f^n=f(q^n,(n)\Delta t)$$

y luego realizar un paso corrector

$$q^{n+1}=q^n+\Delta t f(q^{n+1*},(n+1)\Delta t)$$

y esto me lleva a un método de $\mathcal{O}\left(\Delta t\right)$ que se llama Euler atrasado o Matsuno.

Podemos hacer esto mismo para el método trapezoidal,

$$\frac{q^{n+1}-q^n}{\Delta t}=\frac{f^{n+1}+f^n}{2}$$

y los pasos serían, el predictor

$$q^{n+1*}=q^n+\Delta t f^n$$

y el corrector

$$q^{n+1}=q^n+\frac{\Delta t}{2}(f^{n+1*}+f^n)$$

esto resulta en un método de orden $\mathcal{O}(t^2)$ que es el método de Heun.

Una alternativa al método de Matsuno es hacer un semipaso $\Delta t/2$

$$q^{n+1/2*} = q^n + \frac{\Delta t}{2} f(q^n, n\Delta t)$$

y el corrector

$$q^{n+1} = q^n + \frac{\Delta t}{2}(f^{n+1/2*}, (n+1)/2\Delta t)$$

de modo que resulta el método de Runge-Kutta de orden $\mathcal{O}(t^2)$. Si f es lineal ambos métodos coinciden.

Un Runge-Kutta de orden cuatro (un método que está implementado en los *solvers* de muchos paquetes numéricos) hace una aproximación

$$q^{n+1} = q^n + \frac{\Delta t}{6}(K_1 + 2K_2 + 2K_3 + K_4),$$

en donde

$$\begin{aligned} K_1 &= f(q^n, n\Delta t) \\ K_2 &= f(q^n + K_1 \frac{\Delta t}{2}, (n+1/2)\Delta t) \\ K_3 &= f(q^n + K_2 \frac{\Delta t}{2}, (n+1/2)\Delta t) \\ K_4 &= f(q^n + K_3 \Delta t, (n+1)\Delta t) \end{aligned}$$

Estos métodos también se llaman métodos *multipaso* ó *predictor-corrector*.

§ 8. Error por pasos

En resumen hemos visto que el problema general se puede escribir como

$$\frac{dq}{dt}(t) = f(q(t), t) \qquad t = n\Delta t \qquad n = 0, 1, ...$$

y mencionamos que del error de discretización (o truncamiento) generalmente expresábamos su orden, es decir

$$D = \varepsilon = \mathcal{O}\left(\Delta t^{\alpha}\right) \qquad \text{con } \alpha \geq 1$$

y tambien se dijo que si

$$D \to 0 \quad \text{con} \quad \Delta t \to 0$$

entonces el método era consistente.

También tenemos un error entre pasos asociado a la evolución

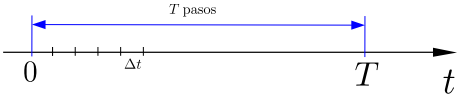
$$q^n \longrightarrow q^{n+1},$$

y ese error será, como comprobaremos inmediatamente,

$$\mathcal{O}(\Delta t^{\alpha+1}),$$

es decir, de un orden mayor que el del método.

Así, por ejemplo, para RK2 tendremos un error por paso $\mathcal{O}\left(\Delta t^3\right)$. Para un transcurso de $t = 0$ a $t = T$ finito (como se ve en la ilustración siguiente)



el orden del error global E_G para N pasos[2], donde

$$N = \frac{T}{\Delta t},$$

puesto que Δt lo consideramos constante, será entonces el orden del error por paso multiplicado por el número de pasos, es decir

$$E_G = \mathcal{O}\left(\Delta t^3\right) N = \mathcal{O}\left(\Delta t^2\right)$$

de manera que este error global tiene justamente el orden del método. Es más, el orden del método es el orden del error global ($E_G = D$).

§ 9. Estabilidad de los métodos

Ahora estaremos interesados en comparar la solución numérica dada por el método con la exacta. Consideraremos el error de discretización

$$D = |q^n - q(n\Delta t)| \longrightarrow 0 \quad \text{con} \quad n \to \infty$$

donde q^n es la solución numérica y $q(n\Delta t)$ la exacta, y su convergencia, es decir que se verifique el límite.

Un criterio de estabilidad es pedir que

$$|q^n - q(n\Delta t)| \leq M \quad \text{con} \quad n \to \infty$$

es decir que el error se halle acotado a partir de cierto momento. Pero esto es difícil porque en general no es conocida la solución exacta. Podemos, no obstante, ver que la solución se halla acotada cuando tenemos alguna *constraint* en el sistema físico que se está representando. La energía en un fluido podría ser tal ejemplo.

[2]El error global es el error acumulado en una evolución dada.

Von Neumann tiene, cuando no, un criterio para ver la estabilidad asociado a examinar el comportamiento de un modo genérico de Fourier. Defino para ello un factor de amplificación

$$\lambda \equiv \frac{q^{n+1}}{q^n}, \qquad \lambda \in \mathbb{C}$$

de modo que haciendo recurrencia se tiene

$$q^{n+1} = \lambda q^n = ... = \lambda^{n+1} q^0,$$

y esto vale para una ecuación lineal (λ es el mismo para todos los pasos).

Si se cumple que

$$|\lambda| \leq 1 \tag{9.1}$$

la solución se mantiene estable. En el caso de un sistema lineal la condición (9.1) es condición suficiente para la estabilidad.

Por otro lado, aún con $|\lambda| > 1$ la solución puede ser útil porque de alguna manera conozco cuanto se aparta de la solución exacta. El factor λ depende del método, del Δt y de los parámetros de la ecuación que estamos intentando resolver.

Para ecuaciones no lineales lo que puede hacerse es considerar $f(q,t) \propto q$, es decir linealizar f y evaluar la estabilidad lineal. Si la ecuación linealizada no es estable tampoco lo será la ecuación original, no obstante si resultara ser estable eso no nos garantiza la estabilidad de la versión no lineal. En suma, la linealización sirve para descartar algunos casos.

EJEMPLO 3.3. Ecuación advectiva lineal
La ecuación advectiva

$$\frac{\partial u}{\partial t} + v\frac{\partial u}{\partial x} = 0,$$

con v constante es el caso lineal de la ecuación advectiva no lineal (donde $v = u$), la cual surge en fluidos al aproximar u en torno a un valor $u_0 = v$.
El examen de estabilidad de esta ecuación para un método particular puede arrojar luz sobre el comportamiento de la versión no lineal.

§ 9.1. Análisis de estabilidad para el oscilador armónico

Consideraremos ahora la ecuación del oscilador armónico en el formalismo complejo ya introducido. Escrita de esa manera es un caso de $f(q,t) \propto q$, es decir de problema lineal.

La ecuación a resolver es

$$\frac{dq}{dt} = i\omega q$$

donde $q = x - iv/\omega$ y la frecuencia es $\omega = \sqrt{k/m}$.

La solución exacta, sabemos, es

$$q(t) = q(0)\mathrm{e}^{i\omega t}$$

que para el tiempo discreto es

$$q(n\Delta t) = q(0)\mathrm{e}^{i\omega n\Delta t}$$

y para el instante de tiempo $n+1$ siguiente es

$$q((n+1)\Delta t) = q(0)\mathrm{e}^{i\omega(n+1)\Delta t} = q(0)\mathrm{e}^{i\omega n\Delta t}\mathrm{e}^{i\omega\Delta t}$$

o bien

$$q((n+1)\Delta t) = q(n\Delta t)\mathrm{e}^{i\omega\Delta t},$$

luego se ve claramente que

$$\lambda_e = \mathrm{e}^{i\omega\Delta t}$$

que es el factor de amplificación λ de la solución exacta. Este factor es un número complejo en general siendo en este caso de módulo $|\lambda_e| = 1$ y fase $\theta_e = \omega\Delta t$.

Supongamos hacer lo mismo pero para los métodos numéricos. Consideremos primeramente el método de Euler (adelantado)

$$q^{n+1} = q + \Delta t f(q^n, n\Delta t)$$

que para este caso resulta

$$q^{n+1} = q^n + i\omega\Delta t q^n = q^n(1 + i\omega\Delta t)$$

y entonces

$$\lambda = 1 + i\omega\Delta t$$

de manera que

$$|\lambda| = \sqrt{1 + (\omega\Delta t)^2}$$

Vemos que $\lambda \to \lambda_e$ con $\Delta t \to 0$ pero si Δt no es cero el módulo siempre es mayor a la unidad; el método es incondicionalmente inestable. Decimos que es incondicionalmente inestable porque no existe condición sobre los parámetros de la discretización que permita establecer una región de estabilidad.

La inestabilidad está cuantificada por el producto $\omega\Delta t$. Luego, el método de Euler no es aconsejable para soluciones oscilatorias y peor aún cuando las frecuencias son altas.

La fase θ está asociada a

$$\tan\,\theta = \frac{\mathcal{I}(\lambda)}{\mathcal{R}(\lambda)},$$

que para nuestro ejemplo significa

$$\theta = \operatorname{atan}(\omega \Delta t),$$

y donde vemos que con $\Delta t \to 0$, se da $\theta \to \theta_e$. La fase nos informa sobre las oscilaciones de la solución en torno al valor hacia el cual converge la misma.

Apliquemos a la misma ecuación el método trapezoidal

$$q^{n+1} = q^n + i\omega \frac{\Delta t}{2}(q^{n+1} + q^n)$$

$$q^{n+1}\left(1 - i\omega \frac{\Delta t}{2}\right) = q^n \left(1 + i\omega \frac{\Delta t}{2}\right)$$

entonces

$$\lambda = \frac{1 + i\omega \frac{\Delta t}{2}}{1 - i\omega \frac{\Delta t}{2}},$$

desde donde inmediatamente vemos

$$|\lambda| = \sqrt{\frac{1 + \omega^2 \frac{\Delta t^2}{4}}{1 + \omega^2 \frac{\Delta t^2}{4}}} = 1$$

de forma que este sí es un buen método para una ecuación con soluciones oscilatorias. Es independiente de Δt y de ω. Es un método incondicionalmente estable, y además no introduce amortiguamiento en la solución numérica. Sin embargo, tiene error de fase

$$\theta = \operatorname{atan}\left(\frac{\omega \Delta t}{1 - \omega^2 \Delta t^2/4}\right)$$

y se ve que con $\Delta t \to 0$ es $\theta \to 0$.

Supongamos ahora el esquema atrasado (implícito)

$$q^{n+1} = q^n + \Delta t f^{n+1}$$

que para nuestra ecuación es

$$q^{n+1} = q^n + i\omega \Delta t q^{n+1}$$

entonces

$$\lambda = \frac{1}{1 - i\omega \Delta t} = \frac{1 + i\omega \Delta t}{1 + \omega^2 \Delta t^2},$$

y su módulo es

$$|\lambda| = \frac{\sqrt{1 + \omega^2 \Delta t^2}}{1 + \omega^2 \Delta t^2} = \frac{1}{\sqrt{1 + \omega^2 \Delta t^2}},$$

mientras la fase da

$$\theta = \omega\Delta t.$$

Como se cumple $|\lambda| < 1$ el método es incondicionalmente estable, pero introduce una amortiguación dada por $(1 + \omega^2\Delta t^2)^{-1/2}$. Con altas frecuencias la solución está demasiado amortiguada.

El amortiguamiento no es necesariamente perjudicial puesto que bajo ciertas circunstancias puede ser necesario filtrar oscilaciones espúreas y se consideran amortiguaciones de manera explícita para ciertas frecuencias de modo que la solución no oscile.

Supongamos ahora el esquema de Matsuno, los pasos predictor y corrector serán, respectivamente,

$$q^{n+1*} = q^n + i\omega\Delta t q^n$$

$$q^{n+1} = q^n + \Delta t i\omega(q^n + i\omega\Delta t q^n) = q^n(1 + i\omega\Delta t[1 + i\omega\Delta t])$$

luego

$$\lambda = (1 + i\omega\Delta t[1 + i\omega\Delta t])$$

y el módulo cuadrado

$$|\lambda|^2 = (1 - \omega^2\Delta t^2)^2 + \omega^2\Delta t^2 = 1 - \omega^2\Delta t^2 + \omega^4\Delta t^4$$

En la Figura 9.1 vemos un gráfico de $|\lambda|^2$ en función de $\omega\Delta t$. Podemos inferir desde su comportamiento el de $|\lambda|$, que será

$$\text{si} \quad \omega\Delta t = 1 \quad \longrightarrow |\lambda| = 1$$

$$\text{si} \quad \omega\Delta t > 1 \quad \longrightarrow |\lambda| > 1$$

$$\text{si} \quad \omega\Delta t < 1 \quad \longrightarrow |\lambda| < 1$$

Luego, el método es condicionalmente estable si se verifica

$$\omega\Delta t \leq 1 \quad \text{o bien} \quad \Delta t \leq 1/\omega.$$

En este caso existe una condición para la estabilidad en función de los parámetros de la ecuación y de su discretización (Δt y ω); hablamos por ello de estabilidad condicional.

Supongamos el método Runge-Kutta de orden 2, que era un Matsuno con un medio paso.

$$q^{(n+1/2)*} = q^n + i\omega\frac{\Delta t}{2}q^n$$

$$q^{n+1} = q^n + \Delta t i\omega(q^n + i\omega\frac{\Delta t}{2}q^n) = q^n\left(1 + i\omega\Delta t\left[1 + i\omega\frac{\Delta t}{2}\right]\right)$$

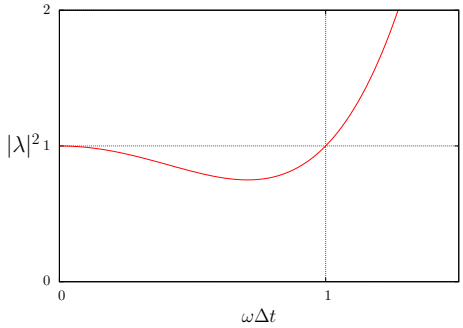


Figura 9.1 Comportamiento del $|\lambda|^2$ para el método Matsuno en la ecuación del oscilador armónico. Evidentemente $|\lambda| < 1$ en la misma región en la cual $|\lambda|^2 < 1$.

de modo que resulta

$$\lambda = (1 + i\omega\Delta t - \frac{\omega^2\Delta t^2}{2}),$$

y

$$|\lambda|^2 = \left(1 - \frac{\omega^2\Delta t^2}{2}\right)^2 + \omega^2\Delta t^2 = 1 + \omega^4\Delta t^4$$

donde es fácil ver que está cantidad será mayor a la unidad siempre. Es incondicionalmente inestable. Sin embargo el crecimiento es con Δt^4 y si el Δt es chico este crecimiento es realmente lento.

Supongamos finalmente el método Leap-frog

$$q^{n+1} - q^{n-1} = 2\Delta t i\omega q^n \tag{9.2}$$

En este método tropezamos con el problema de que se requiere para $n = 0$ el q^{-1} que no puede venir del mismo método. Necesito una condición inicial física o computacional. Puedo arrancar a partir de $n = 1$ utilizando q^0 calculado por otro método.

Supongamos el caso especial de que $\omega = 0$, entonces $q = cte$ debiera ser solución. Pero si resolvemos numéricamente y proponemos $q^0 \neq q^1$ serían

$$q^2 - q^0 = 0$$

$$q^3 - q^1 = 0$$

$$q^4 - q^2 = q^4 - q^2 = 0$$

y siguiendo puede verse que

$$q^n = \begin{cases} q^0 & \text{con } n \text{ par} \\ q^1 & \text{con } n \text{ impar} \end{cases}$$

De esta forma, aún para el caso trivial de frecuencia nula obtenemos oscilaciones espúreas. Es esperable entonces que para una solución oscilatoria tengamos otras oscilaciones provenientes de la inestabilidad del método.

Veamos ahora qué sucede cuando tenemos una frecuencia $\omega \neq 0$. Dado que definíamos el factor de amplificación λ a través de

$$\lambda = \frac{q^{n+1}}{q^n},$$

eso implica que

$$q^n = \lambda q^{n-1}$$

y correspondientemente podemos transformar (9.2) en

$$\lambda^2 q^{n-1} - q^{n-1} - \lambda 2\Delta t i\omega q^{n-1} = 0,$$

ecuación que requiere, para verificarse,

$$\lambda^2 - \lambda 2\Delta t i\omega - 1 = 0$$

y esto viene a significar que tiene las dos raíces siguientes

$$\lambda = \begin{cases} \lambda_1 = i\Omega + \sqrt{1-\Omega^2} \\ \lambda_2 = i\Omega - \sqrt{1-\Omega^2} \end{cases}$$

donde $\Omega = \omega\Delta t$. Cuando $\Delta t \to 0$ vemos que uno de los autovalores tiende a 1 mientras que el otro tiende a -1. Eso hará oscilar la solución entre valores. Estudiemos ahora la estabilidad. Evidentemente si

$$|\Omega| \leq 1$$

la solución es estable, mientras que si

$$|\Omega| > 1$$

será inestable. Esto nos lleva a la siguiente condición de estabilidad,

$$\omega\Delta t \leq 1 \qquad \Rightarrow \quad \Delta t \leq 1/\omega,$$

que es la misma que obtuviéramos para el esquema de Matsuno.

Esto significa que debemos establecer un paso de tiempo Δt que sea menor al período natural de la oscilación más rápida del sistema.

§ 9.2. Estabilidad para la ecuación de decaimiento

Vamos a considerar ahora

$$\frac{dq}{dt} = -\kappa q \quad \text{con } \kappa > 0$$

que es la ecuación de decaimiento exponencial. Esta ecuación surge también bajo otros ámbitos; así, por ejemplo, de la ecuación diferencial en derivadas parciales

$$\frac{d\phi}{dt}(x,t) = \nu \frac{\partial^2 \phi}{\partial x^2}(x,t),$$

que es la ecuación de difusión, al proponer una onda plana unidimensional

$$\phi(x,t) = A(t)\mathrm{e}^{ikx}$$

resulta

$$\frac{\partial A}{\partial t} = -\nu k^2 A,$$

ecuación que tiene solución exponencial decreciente

$$A(t) = A(0)\mathrm{e}^{-\nu k^2 t}.$$

Volviendo ahora a la ecuación de decaimiento podemos escribir entonces la solución exacta como

$$q(n\Delta t) = q(0)\mathrm{e}^{-\kappa n \Delta t}$$
$$q((n+1)\Delta t) = q(0)\mathrm{e}^{-\kappa(n+1)\Delta t} = q(n\Delta t)\mathrm{e}^{-\kappa \Delta t}$$

de manera que finalmente

$$\lambda = \mathrm{e}^{-\kappa \Delta t} \quad \Longrightarrow \quad |\lambda| < 1.$$

Supongamos que resolvemos con el método de Euler,

$$q^{n+1} = q^n - \kappa \Delta t q^n = q^n(1 - \kappa \Delta t).$$

Luego, la estabilidad implica

$$|1 - \kappa \Delta t| < 1$$

o bien

$$-1 < 1 - \kappa \Delta t < 1$$
$$1 > -1 + \kappa \Delta t > -1$$
$$2 > \kappa \Delta t > 0 \quad \Rightarrow \quad \Delta t < 2/\kappa.$$

Nos aseguramos que el método es condicionalmente estable si elegimos Δt menor al tiempo natural de decaimiento del sistema.

Digamos también, sin mostrarlo explícitamente, que el esquema atrasado, implícito, es incondicionalmente estable y que el trapezoidal también lo es. Los esquemas Matsuno y RK2 son condicionalmente estables, siendo la condición la misma que hallamos para Euler.

§ 9.3. Estabilidad de ecuaciones no lineales

Recordemos que según el teorema de Lax tenemos

$$[\,\text{Estabilidad} \quad + \quad \text{Consistencia}\,] \quad \Longrightarrow \quad \text{Convergencia}$$

si el problema es lineal.

Generalizando un poco para

$$\frac{\partial q}{\partial t} = f(q,t),$$

si el problema es lineal se estudiará el factor de amplificación λ pero si el problema es no-lineal lo que puede hacerse, como ya dijéramos, es estudiar la estabilidad del error en el paso $n+1$,

$$q^{n+1} \longrightarrow q^{n+1} + \epsilon^{n+1} = f(q^n + \epsilon^n) = f(q^n) + \frac{\partial f}{\partial q}\epsilon^n$$

y entonces

$$\epsilon^{n+1} = \frac{\partial f}{\partial q}\epsilon^n$$

de modo que se analiza la estabilidad del problema linealizado. Si $\mathbf{q}$ es un vector tendré un vector $\boldsymbol{\varepsilon}$ de errores

$$\boldsymbol{\varepsilon}^{\,n+1} = \frac{\partial f}{\partial \mathbf{q}}\boldsymbol{\varepsilon}^{\,n}$$

y si se diagonaliza la matriz $\partial_{\mathbf{q}} f$ se deberán analizar los autovalores pidiéndose como condición de estabilidad que el módulo de los mismos sea menor igual a uno.

Capítulo 4

Ecuaciones parabólicas

§ 10. Resolución de una PDE

Volviendo ahora a las PDE, nos ocuparemos precisamente de las parabólicas, cuyo ejemplo usual es la ecuación de difusión

$$\frac{\partial u}{\partial t} = \nu \frac{\partial^2 u}{\partial x^2} \quad \text{con} \quad \nu = cte.$$

siendo $u = u(x,t)$. Necesito condiciones iniciales

$$u(x, t=0) = a(x)$$

donde a es una función conocida. Asimismo necesito también condiciones de contorno,

$$u(x=0,t) = c_1(t)$$
$$u(x=L,t) = c_2(t)$$

donde $c_i(t)$ son valores conocidos. Estamos resolviendo, por supuesto, para $0 \leq x \leq L$.

Este problema aparece muchas veces en el modelado físico de una ley de conservación. Pensemos en una magnitud E tal que

$$E = \int_V \varepsilon \, dV$$

donde ε es una densidad volumétrica de E y entonces la variación de E dentro de un volumen V se debe al flujo $\mathbf{q}$ que abandona dicho volumen, es decir

$$\frac{\partial E}{\partial t} = \frac{\partial}{\partial t} \int_V \varepsilon \, dV = -\int_S \mathbf{q} \cdot d\mathbf{S}$$

pero aplicando el teorema de Gauss e introduciendo la derivada temporal dentro de la integral, lo cual puede hacerse porque consideramos el volumen fijo, resulta en

$$\frac{\partial \varepsilon}{\partial t} + \boldsymbol{\nabla} \cdot \mathbf{q} = 0$$

Si en un gas se da que $\varepsilon \propto T$ y el flujo de energía es

$$\mathbf{q} \propto -\boldsymbol{\nabla} T$$

entonces

$$\frac{\partial T}{\partial t} = \kappa \nabla^2 T$$

que en una dimensión espacial resulta

$$\frac{\partial T}{\partial t} = \kappa \frac{\partial^2 T}{\partial x^2}.$$

La ecuación de difusión lleva, intuitivamente a cosas como lo que se ve en la Figura 10.1.

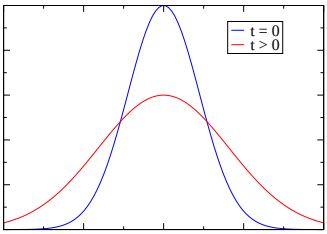


Figura 10.1 Comportamiento cualitativo de una solución que difunde con el paso del tiempo t.

Suponiendo contornos fijos

$$\left.\begin{matrix} c_1(t) \\ c_2(t) \end{matrix}\right\} = 0$$

la solución será

$$u(x,t) = \sum_{m=1}^{\infty} \hat{a}_m(t)\, sen\left(\frac{m\pi x}{L}\right)$$

donde el índice m refiere al modo. Los modos de m alto son los que oscilan más rápidamente. Los coeficientes serán

$$\hat{a}_m(t) = \hat{a}_m(0)\, \mathrm{e}^{-m^2\pi^2\nu t/L^2}$$

siendo el valor del coeficiente al instante inicial

$$\hat{a}_m(0) = \frac{2}{\pi} \int_0^L a(x)\, sen\left(\frac{m\pi x}{L}\right) dx. \tag{10.1}$$

Cada modo decae en el tiempo de acuerdo al tiempo de difusión

$$\tau = \frac{L^2}{\nu m^2 \pi^2},$$

de forma que los modos rápidos son los que decaen primero.

Volvamos a la solución numérica de la ecuación de difusión. Para ello introducimos una grilla

$$x_j = j\Delta x \qquad \text{con } j = 1, 2, ..., N$$

y discretizamos la parte espacial de las derivadas,

$$\left.\frac{\partial^2 u}{\partial x^2}\right|_j = \frac{u_{j+1} - 2u_j + u_{j-1}}{\Delta x^2}.$$

Entonces los contornos son

$$u_0 = c_1 \qquad u_N = c_2$$

Planteamos una solución

$$\mathbf{U}(t) = (u_1(t), u_2(t), ..., u_{N-1}(t))$$

notando que los contornos están fijos, es decir que no varían con el tiempo. Esto conduce a un sistema matricial

$$\frac{\partial \mathbf{U}}{\partial t} = L\mathbf{U} + \mathbf{C}$$

siendo $\mathbf{C}$ un vector de condiciones de contorno,

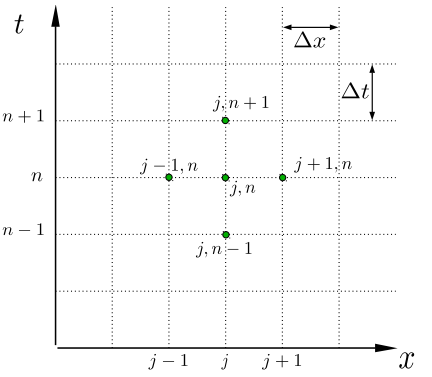
$$\mathbf{C} = (c_1, 0, ..., 0, c_2)\frac{\nu}{\Delta x^2}$$

y

$$L = \begin{pmatrix} -2 & 1 & 0 & ... & 0 \\ 1 & -2 & 1 & ... & 0 \\ 0 & 1 & -2 & ... & 0 \\ ... & ... & ... & ... & 1 \\ 0 & ... & ..... & 1 & -2 \end{pmatrix} \frac{\nu}{\Delta x^2}.$$

Ahora el problema pasó de ser una ecuación PDE a ser un sistema de ODE's acopladas. En la práctica se hace de forma conjunta la parte espacial y la parte

temporal. Si la discretización temporal es de acuerdo a $n\Delta t$ y la espacial de acuerdo a $j\Delta x$ podemos acomodar la evolución de la solución en la cuadrícula que aparece bajo estas líneas, donde se muestran t versus x.



Si se aplica un método de Euler en el tiempo para cada j y una diferencia centrada en el espacio para cada n resulta

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} = \nu \frac{u_{j+1}^n - 2u_j^n + u_{j-1}^n}{\Delta x^2}$$

$$u_j^{n+1} = u_j^n + \frac{\nu \Delta t}{\Delta x^2} \left(u_{j+1}^n - 2u_j^n + u_{j-1}^n \right)$$

En § 5 se analizó la consistencia de este método tomando una solución F de la ecuación exacta, reemplazándola en la ecuación en diferencias y arribando al error de discretización $D(F)$, que para este caso era

$$D = \mathcal{O}(\Delta t) + \mathcal{O}(\Delta x^2)$$

de modo que

$$D \to 0 \qquad \text{si} \begin{cases} \Delta t \to 0 \\ \Delta x \to 0 \end{cases}$$

y vemos que el método es consistente.

Revisemos ahora el asunto de la estabilidad. Suponemos un error (perturbación) que tenga forma de modo Fourier

$$E_j^n = E^n \mathrm{e}^{ikx_j}$$

con un k arbitrario. Como el esquema también vale para el error se tendrá

$$E^{n+1} \mathrm{e}^{ikx_j} = E^n \mathrm{e}^{ikx_j} + \Gamma \left(E^n \mathrm{e}^{ik(x_j + \Delta x)} + E^n \mathrm{e}^{ik(x_j - \Delta x)} - 2E^n \mathrm{e}^{ikx_j} \right)$$

donde $\Gamma = \nu\Delta t/\Delta x^2$

$$E^{n+1}\mathrm{e}^{ikx_j} = E^n\mathrm{e}^{ikx_j}\left[1 + \Gamma\left(\mathrm{e}^{ik\Delta x} + \mathrm{e}^{-ik\Delta x} - 2\right)\right]$$

y entonces

$$\lambda = \frac{E^{n+1}}{E^n} = 1 + \Gamma\left(\mathrm{e}^{ik\Delta x} + \mathrm{e}^{-ik\Delta x} - 2\right)$$

$$\lambda = \frac{E^{n+1}}{E^n} = 1 + 2\Gamma\left(\cos(k\Delta x) - 1\right) = 1 - 4\Gamma\,\mathrm{sen}^2(k\Delta x/2),$$

donde para la última igualdad hemos utilizado una identidad trigonométrica usual. Si ahora pedimos que $|\lambda| \leq 1$ será

$$-1 \leq 1 - 4\Gamma\,\mathrm{sen}^2(k\Delta x/2) \leq 1$$

$$1 \geq 4\Gamma\,\mathrm{sen}^2(k\Delta x/2) - 1 \geq -1$$

$$2 \geq 4\Gamma\,\mathrm{sen}^2(k\Delta x/2) \geq 0$$

Asumiendo que siempre $\mathrm{sen}^2(\beta) \leq 1$ deberemos garantizar la estabilidad con

$$4\Gamma \leq 2$$

lo que lleva a

$$\Delta t \leq \Delta x^2/2\nu. \tag{10.2}$$

Vemos que resulta un vínculo entre la discretización temporal y la espacial. El intervalo temporal dado por (10.2) es una especie de tiempo de decaimiento numérico (asociado con la grilla) o tiempo medio de difusión espacial. El intervalo temporal debe ser más corto que dicho tiempo para que el esquema sea estable.

En conclusión, dado que la ecuación es lineal y de acuerdo al teorema de Lax, como la solución es consistente y estable el método es convergente.

§ 11. Condición CFL

En los métodos numéricos existe un concepto que se llama la condición CFL (Courant-Friedrichs-Lewy) [1928]. Estos investigadores desarrollaron el concepto de rango de influencia de una ecuación diferencial.

La idea del rango de influencia es, considerando la solución de una ODE en un punto (x, t), ¿de qué puntos a $t = 0$ depende la misma? O dicho de otra manera, el rango de influencia son los puntos x del dominio, a $t = 0$, que son necesarios conocer para obtener la solución en x, t.

En la ecuación parabólica la solución exacta en (x, t) depende de todos los puntos a $t = 0$. Esto puede verse por la integral de Fourier (10.1); la misma es un reflejo de que $u(x, t)$ depende de $x \in (0, L)$.

Pero también existe un rango de influencia numérico. Si resolvemos numéricamente un problema, como el de la ecuación de difusión unidimensional, asociamos según hemos visto una grilla con los puntos en x, t. Dada la solución numérica u de la ecuación en un punto x_j, t_n es claro que la misma depende de los valores anteriores, los cuales también dependen de los anteriores a ellos.

El rango numérico dependerá del método en cuestión. Para Euler vemos que se forma un patrón, pues

$$u_j^{n+1} = u_j^n + \frac{\nu \Delta t}{\Delta x^2} \left(u_{j+1}^n - 2u_j^n + u_{j-1}^n \right)$$

y entonces la solución u en $j, n+1$ depende de la solución en n y en los puntos $j-1, j$ y $j+1$. Asimismo, cada uno de estos últimos depende del valor temporalmente anterior (en $n-1$) y evaluado en los índices espaciales correspondientes.

La Figura 11.2 ilustra de manera pictórica un rango de influencia para un punto genérico de la solución en $j+1, n$. La doble flecha verde es el rango de influencia para dicha solución, que como vemos no abarca todo el dominio posible.

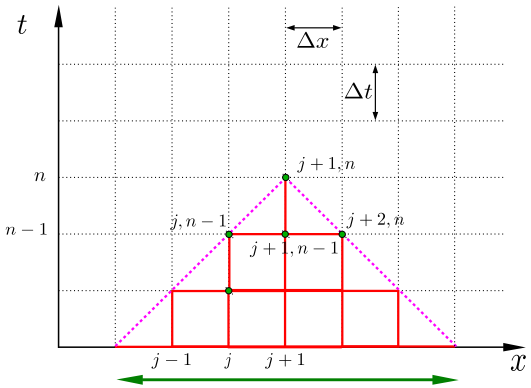


Figura 11.2 Rango de influencia para una ecuación diferencial (doble flecha verde) para un punto $j+1, n$.

Si reducimos de igual manera el paso espacial y el temporal vemos que no cambia el intervalo de influencia. Se logra tender el rango teórico (todo el dominio de la ecuación) si achico el Δt asociado de acuerdo a

$$\Delta t \leq \frac{\Delta x^2}{2\nu}$$

Esto es estar de acuerdo con la condición CFL.

§ 12. Esquema implícito de Crank-Nicolson (1950)

Este método permite escapar del problema del rango de influencia. Se cambiará el método para la parte temporal; usaremos el método trapezoidal. De esta forma la discretización de la ecuación de difusión es ahora

$$u_j^{n+1} = u_j^n + \frac{\nu\Delta t}{2\Delta x^2}\left[(u_{j+1}^{n+1} - 2u_j^{n+1} + u_{j-1}^{n+1}) + (u_{j+1}^n - 2u_j^n + u_{j-1}^n)\right]. \quad (12.1)$$

Esto se lleva a un sistema matricial, como se vio en la sección § 10 para el método de Euler, definiendo

$$\mathbf{U} = (u_1, u_2, ..., u_{N-1})$$

con $u_0 = c_1$ y $u_N = c_2$, y entonces se tendrá

$$\mathbf{U}^{n+1} - \mathbf{U}^n = \frac{1}{2}L(\mathbf{U}^n + \mathbf{U}^{n+1}) + \mathbf{C}$$

siendo ahora la matriz definida por

$$L = \begin{pmatrix} -2 & 1 & 0 & ... & 0 \\ 1 & -2 & 1 & ... & 0 \\ 0 & 1 & -2 & ... & 0 \\ ... & ... & ... & ... & 1 \\ 0 & ... & ..... & 1 & -2 \end{pmatrix} \frac{\nu\Delta t}{\Delta x^2}.$$

y el vector de condiciones por

$$\mathbf{C} = (u_0, 0, ..., 0, u_N)\frac{\nu\Delta t}{\Delta x^2}.$$

Amasando, algebraicamente, este sistema se llega a

$$(\mathbb{1} - L/2)\mathbf{U}^{n+1} = (\mathbb{1} + L/2)\mathbf{U}^n + \mathbf{C},$$

que puede simplificarse algo considerando $\Gamma \equiv \nu\Delta t/\Delta x^2$ y una nueva matriz

$$L' = \mathbb{1} - L/2 = \begin{pmatrix} 1+\Gamma & -\Gamma/2 & 0 & ... & 0 \\ -\Gamma/2 & 1+\Gamma & -\Gamma/2 & ... & 0 \\ 0 & -\Gamma/2 & 1+\Gamma & ... & 0 \\ ... & ... & ... & ... & -\Gamma/2 \\ 0 & ... & ..... & -\Gamma/2 & 1+\Gamma \end{pmatrix}$$

tal que la resolución del sistema es

$$\mathbf{U}^{n+1} = L'^{-1}(\mathbb{1} + L/2)\mathbf{U}^n + L'^{-1}\mathbf{C}.$$

Haciendo aproximaciones de Taylor en tiempo y espacio para (12.1) se puede ver que resulta

$$D(F) \to 0 \qquad \text{si} \begin{cases} \Delta x \to 0 \\ \Delta t \to 0 \end{cases}$$

siendo F solución exacta. Entonces el método trapezoidal para la ecuación de difusión es consistente.

La estabilidad se plantea analizando el comportamiento de un modo Fourier

$$E_j^n = E^n \mathrm{e}^{ikx_j},$$

el cual una vez reemplazado dentro del esquema

$$E^{n+1}\mathrm{e}^{ikx_j} = E^n \mathrm{e}^{ikx_j} + \frac{\Gamma}{2}\left[\left(E^{n+1}\mathrm{e}^{ik(x_j+\Delta x)} + E^{n+1}\mathrm{e}^{ik(x_j-\Delta x)} - 2E^{n+1}\mathrm{e}^{ikx_j}\right)\right.$$
$$\left. + \left(E^n \mathrm{e}^{ik(x_j+\Delta x)} + E^n \mathrm{e}^{ik(x_j-\Delta x)} - 2E^n \mathrm{e}^{ikx_j}\right)\right] \quad (12.2)$$

y haciendo la penosa álgebra, conduce a

$$E^{n+1}(1 - \Gamma[\cos(k\Delta x) - 1]) = E^n(1 + \Gamma[\cos(k\Delta x) - 1]),$$

de manera que

$$\lambda = \frac{1 - 2\Gamma sen^2(k\Delta x/2)}{1 + 2\Gamma sen^2(k\Delta x/2)}$$

y puede comprobarse que

$$|\lambda| \leq 1 \qquad \forall t$$

de modo que resulta incondicionalmente estable.

Si examinamos el rango de influencia vemos que el valor en un punto está determinado por todos los puntos. En efecto, de inspeccionar (12.1) notamos que el valor de la solución u en el punto $j, n+1$ depende de valores anteriores temporales $(j, n), (j-1, n)$ y $(j+1, n)$ pero también de valores en el mismo instante de tiempo $(j, n+1), (j-1, n+1)$ y $(j+1, n+1)$. Estos últimos también dependen de valores en el tiempo $n+1$ pero en puntos espaciales más alejados $j-2, j+2$ de manera que el rango de influencia resulta ser todo el dominio, como en el caso exacto.

§ 12.1. Solución del método de Crank-Nicolson

Se puede encarar usando recurrencia por un método algorítmico TDMA. Un sistema tridiagonal se puede escribir siempre como

$$\alpha_j u_{j+1}^{n+1} + \beta_j u_j^{n+1} + \gamma_j u_{j-1}^{n+1} = \omega_j \quad (12.3)$$

que sale del problema con $j = 1, ..., N-1$ y específicamente para $\alpha_j = -\Gamma/2, \beta_j = 1 - \Gamma, \gamma_j = -\Gamma/2$. El borde cumple

$$u_0 = c_1 \qquad u_N = c_2$$

de manera que

$$\beta_0 = 1 \qquad \alpha_0 = \gamma_0 = 0$$
$$\beta_N = 1 \qquad \alpha_N = \gamma_N = 0$$
$$\omega_0 = u_0 \qquad \omega_N = u_N$$

Planteamos la siguiente recurrencia

$$u_{j+1}^{n+1} = x_j u_j^{n+1} + y_j \tag{12.4}$$

y la fuerzo a valer con (12.3) de modo que

$$\alpha_j(x_j u_j^{n+1} + y_j) + \beta_j u_j^{n+1} + \gamma_j u_{j-1}^{n+1} = \omega_j$$

y entonces

$$u_j^{n+1} = \frac{-\gamma_j}{\alpha_j x_j + \beta_j} u_{j-1}^{n+1} + \frac{\omega_j - \alpha_j y_j}{\alpha_j x_j + \beta_j} \tag{12.5}$$

y ahora identificamos en (12.4) quienes son x_j y_j, entonces

$$\frac{-\gamma_j}{\alpha_j x_j + \beta_j} = x_{j-1}$$

$$\frac{\omega_j - \alpha_j y_j}{\alpha_j x_j + \beta_j} = y_{j-1}$$

Tenemos una relación de recurrencia para x_j, y_j. Hay que hacer la recurrencia arrancando con $j = N - 1$ y elijo

$$u_N^{n+1} = c_2 = x_{N-1} u_{N-1}^{n+1} + y_{N-1}$$

la cual verifica

$$x_{n-1} = 0 \qquad y_{N-1} = c_2$$

y entonces calculo los x_j, y_j *bajando* de $N - 1$ a 0 usando (12.5). Una vez en posesión de los x_j, y_j calculo $u_1^{n+1}, ..., u_{N-1}^{N+1}$ con la recurrencia (12.4) desde 1 hasta $N - 1$.

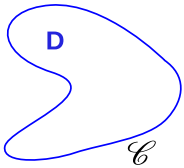
$$u_1^{n+1} = x_0 u_0^{n+1} + y_0 \tag{12.6}$$
$$u_2^{n+1} = x_1 u_1^{n+1} + y_1 \tag{12.7}$$
$$... \quad = \quad ... \tag{12.8}$$

§ 13. Ecuación parabólica en 2D

Estaremos intentando resolver la versión dos dimensional de la ecuación de difusión, esto es

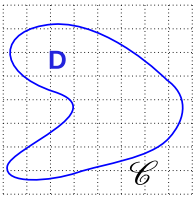
$$\frac{\partial u}{\partial t} = \nu \left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} \right)$$

para el dominio D con frontera $\mathscr{C}$ y sujeta a las condiciones detalladas a continuación



$$u(x, y, t = 0) = a(x, y)$$
$$\left.\begin{array}{l} u([x, y]\big|_{\mathscr{C}}, t) \\ \partial_n u \left([x, y]\big|_{\mathscr{C}}, t\right) \end{array}\right\} \text{datos (CC)}$$

Se grillará al dominio, y discretizaremos de acuerdo a



$$x_i = i\Delta x$$
$$x_j = j\Delta y$$
$$i, j = 1, 2, ..., N-1$$
(si fuera cuadrado)

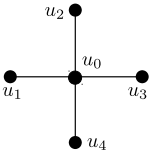
Un índice controlará diferencias en $\hat{x}$ y el otro en $\hat{y}$. Así, utilizando un esquema temporal de Euler y una derivada centrada, tendremos

$$u_{i,j}^{n+1} = u_{i,j}^n + \Gamma_x \left(u_{i+1,i}^n - 2u_{i,j}^n + u_{i-1,j}^n \right) + \Gamma_y \left(u_{i,j+1}^n - 2u_{i,j}^n + u_{i,j-1}^n \right)$$

con

$$\Gamma_x = \frac{\nu \Delta t}{\Delta x^2} \qquad \Gamma_y = \frac{\nu \Delta t}{\Delta y^2}.$$

En el caso sencillo simétrico, $\Delta x = \Delta y$ y si etiquetamos los vecinos que intervienen en el cálculo de un elemento u_0 en el tiempo $n+1$ (ver Figura bajo estas líneas) resultan



$$u_0^{n+1} = (1 - 4\Gamma)u_0 + \left(\sum_{k=1}^{4} u_k \right)$$

Este método es consistente y estable si

$$\Gamma_x, \Gamma_y \leq 1/4$$

Es un método explícito. Con la misma técnica podemos extender el método de Crank-Nicolson, aunque más engorrosamente, como

$$u_{i,j}^{n+1} = u_{i,j}^{n} + \frac{\Gamma_x}{2}\left[(u_{i+1,j}^{n+1} - 2u_{i,j}^{n+1} + u_{i-1,j}^{n+1}) + (u_{i+1,j}^{n} - 2u_{i,j}^{n} + u_{i-1,j}^{n})\right] + \frac{\Gamma_y}{2}\left[(u_{i,j+1}^{n+1} - 2u_{i,j}^{n+1} + u_{i,j-1}^{n+1}) + (u_{i,j+1}^{n} - 2u_{i,j}^{n} + u_{i,j-1}^{n})\right]$$

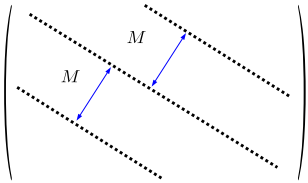
Esto se puede pensar como una matriz, considerando

$$\mathbf{U} = U_{iM+j} \equiv U_{i,j} \qquad i,j = 1,2,...,N-1,$$

es decir que escribimos una matriz como un vector y tenemos así todos los U en los puntos de grilla. El sistema resulta

$$A\mathbf{U}^{n+1} = B\mathbf{U}^{n} + \mathbf{C}$$

donde la matriz A tiene la forma esquematizada debajo



Como vemos la matriz ya no es tridiagonal, sino que es más difícil de invertir (es una matriz *sparse* o rala). Ya un sistema de 100×100 lleva 10000 incógnitas y define una matriz de 10000×10000. Para evitar esto se recurre a otros métodos.

§ 13.1. Método de dirección alternada

Se hace explícito en una variable (por ejemplo, y) e implícito en la otra (x) escribiendo primeramente

$$\frac{U_{i,j}^{n+1/2} - U_{i,j}^{n}}{\Delta t/2} = \nu\left(\left.\frac{\partial^2 U}{\partial x^2}\right|_{\mathrm{DF}}^{n+1/2} + \left.\frac{\partial^2 U}{\partial y^2}\right|_{\mathrm{DF}}^{n}\right),$$

donde el subíndice DF refiere a la expresión de la derivada segunda de acuerdo a diferencias finitas. Ahora hacemos el otro semipaso invirtiendo el orden

$$\frac{U_{i,j}^{n+1} - U_{i,j}^{n+1/2}}{\Delta t/2} = \nu\left(\left.\frac{\partial^2 U}{\partial x^2}\right|_{\mathrm{DF}}^{n+1/2} + \left.\frac{\partial^2 U}{\partial y^2}\right|_{\mathrm{DF}}^{n+1}\right),$$

Puede probarse que este método resulta consistente y estable de manera incondicional. Un sistema de 100×100 se transforma en 100 sistemas de 100×100 para cada ecuación.

§ 13.2. Método de splitting

Consiste en hacer primero una derivada en x y luego una en y

$$\frac{\bar{U}_{i,j}^{n+1} - U_{i,j}^{n}}{\Delta t} = \nu \left. \frac{\partial^2 U}{\partial x^2} \right|^{n+1}$$

$$\frac{U_{i,j}^{n+1} - \bar{U}_{i,j}^{n+1}}{\Delta t} = \nu \left. \frac{\partial^2 U}{\partial y^2} \right|^{n+1}$$

Este esquema también es consistente y estable (incondicionalmente).

Capítulo 5

Ecuaciones advectivas

§ 14. Ecuación de advección lineal

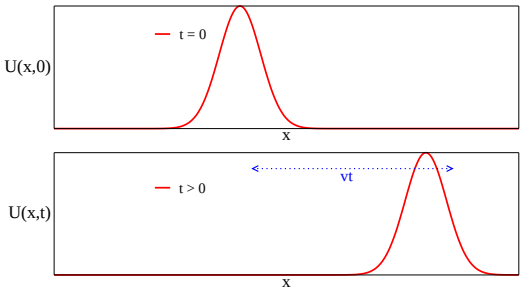
Consideraremos

$$\frac{\partial u}{\partial t} + v\frac{\partial u}{\partial x} = 0 \tag{14.1}$$

donde $u = u(x,t)$ y v es una constante, sujeta a condiciones iniciales

$$u(x,0) = a(x).$$

Físicamente la solución es la propagación de una forma sin perturbación. Se ilustra este hecho en la figura bajo estas líneas



Soluciones analíticas de esta ecuación son

$$u(x,t) = a(x - vt).$$

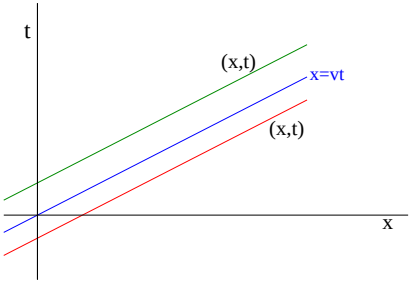
Las condiciones de contorno, por la derivada primera, se reducen a un valor

$$u(x=0,t) = g(t) \quad t > 0,$$

siendo g una función conocida

$$u(x,t) = \begin{cases} a(x-vt) & \text{si } x - vt \geq 0 \\ g(t - x/v) & \text{si } x - vt < 0 \end{cases}$$

Esto nos lleva a las curvas características, que se muestran debajo. La idea es que cualquier punto (x,t) evoluciona, por la ecuación, según las rectas de pendiente v.



La ecuación advectiva contiene oscilaciones, si propongo

$$u(x,t) = A(t)\, \mathrm{e}^{ikx}$$

y le pido que cumpla (14.1) entonces debe valer

$$\frac{dA}{dt} = -kvA$$

de modo que

$$A \propto \mathrm{e}^{-kvt},$$

lo que significa que el factor temporal difunde en el tiempo.

Este caso es muy sencillo y posee solución analítica, razón por la cual se lo suele utilizar como *benchmark* de métodos. También puede ser útil en problemas que tienen una parte advectiva.

Para resolver numéricamente la (14.1) podemos emplear un esquema de Euler en el tiempo y diferencias finitas de orden dos en la variable espacial x, discretizando de acuerdo a

$$x \to j\Delta x \qquad t \to n\Delta t$$

lo cual lleva a

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} = -v\frac{u_{j+1}^n - u_{j-1}^n}{2\Delta x}$$

$$u_j^{n+1} = u_j^n - \frac{v\Delta t}{2\Delta x}(u_{j+1}^n - u_{j-1}^n).$$

Para examinar la estabilidad consideramos un modo Fourier

$$E^n \mathrm{e}^{ikx_j}$$

y lo introducimos en la ecuación

$$E^{n+1}\mathrm{e}^{ikx_j} = E^n\mathrm{e}^{ikx_j} - \frac{v\Delta t}{2\Delta x}(E^n\mathrm{e}^{ikx_j}\mathrm{e}^{ik\Delta x} - E^n\mathrm{e}^{ikx_j}\mathrm{e}^{-ik\Delta x})$$

no siendo difícil llegar a

$$E^{n+1} = E^n\left[1 - i\frac{v\Delta t}{\Delta x}(\operatorname{sen} k\Delta x)\right]$$

de modo que

$$\lambda = 1 - i\alpha$$

$$|\lambda|^2 = 1 + \alpha^2,$$

lo cual significa que el método es incondicionalmente inestable. No es un buen método hacer Euler en el tiempo para esta ecuación; en cambio en la ecuación parabólica podíamos elegir el Δt de acuerdo a ciertas *constraints* y el método era estable. Esto no invalida la utilización de este esquema, pero dado que los resultados se arruinan al pasar el tiempo debe emplearse con cautela.

Es posible estabilizar este método si realizamos el siguiente reemplazo

$$u_j^n \to \frac{1}{2}(u_{j+1}^n + u_{j-1}^n)$$

que transforma la solución a

$$u_j^{n+1} = \frac{1}{2}(u_{j+1}^n + u_{j-1}^n) - \frac{v\Delta t}{2\Delta x}(u_{j+1}^n - u_{j-1}^n)$$

resultando en un método que continúa siendo consistente pero ahora es estable. Este reemplazo se debe a Lax.

Para comprobar la estabilidad hacemos el reemplazo del error

$$E^{n+1} = E^n\left[\cos(k\Delta x)1 - i\frac{v\Delta t}{2\Delta x}(\operatorname{sen}(k\Delta x))\right]$$

y vemos que el nuevo factor de amplificación tendrá

$$|\lambda|^2 = \cos(k\Delta x)^2 + \left(\frac{v\Delta t}{2\Delta x}\right)^2 \operatorname{sen}(k\Delta x)^2$$

$$|\lambda|^2 = 1 - \operatorname{sen}(k\Delta x)^2 \left[1 - \left(\frac{v\Delta t}{\Delta x}\right)^2\right] \tag{14.2}$$

Si ahora pedimos que sea

$$|\lambda|^2 \leq 1$$

se tiene

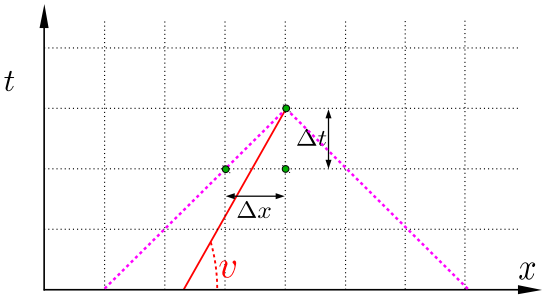
$$0 \leq \operatorname{sen}(k\Delta x)^2 \left[1 - \left(\frac{v\Delta t}{\Delta x}\right)^2\right]$$

lo cual conduce directamente a

$$\frac{v\Delta t}{\Delta x} \leq 1.$$

Hemos llegado a comprobar que necesitamos la condición CFL para la estabilidad. Se dice también, equivalentemente, que la velocidad de la grilla debe ser mayor a la velocidad física de la propagación. Esto garantiza la convergencia para problemas lineales.

Cuando miramos el rango de influencia, comprendemos que la condición CFL significa justamente que la pendiente de la velocidad v real de la propagación debe ser menor justamente a la velocidad de grilla $\Delta x/\Delta t$, según vemos pictóricamente debajo



El punto de solución exacta en $t = 0$ está dentro del rango de influencia. Recordemos que usando el método de las características la solución la obtenemos propagando desde el punto en $t = 0$.

Además, entre más se parece $\Delta x/\Delta t$ a v más preciso llegamos a la solución analítica.

§ 14.1. Difusión numérica en la ecuación advectiva lineal

Trabajemos ahora en la expresión de Lax sumando y restando u_j^n y dividiendo sobre Δt,

$$\frac{u_j^{n+1}-u_j^n}{\Delta t}=\frac{1}{2\Delta t}(u_{j+1}^n+u_{j-1}^n-2u_j^n)-\frac{v}{2\Delta x}(u_{j+1}^n-u_{j-1}^n),$$

y ahora sumemos y restemos en el miembro izquierdo u_j^{n-1}

$$\frac{u_j^{n+1}-u_j^n}{\Delta t}=\frac{u_j^{n+1}-u_j^{n-1}+u_j^{n-1}-u_j^n}{\Delta t}$$

y separando del siguiente modo

$$\frac{u_j^{n+1}-u_j^n}{\Delta t}=\frac{u_j^{n+1}-u_j^{n-1}}{2\Delta t}+\frac{u_j^{n+1}+u_j^{n-1}-2u_j^n}{2\Delta t},$$

llegamos a que la versión continua del miembro izquierdo es

$$\frac{\partial u}{\partial t}+\frac{\Delta t}{2}\frac{\partial^2 u}{\partial t^2}.$$

Juntando todo resulta en la siguiente ecuación

$$\frac{\partial u}{\partial t}+\frac{\Delta t}{2}\frac{\partial^2 u}{\partial t^2}=-v\frac{\partial u}{\partial x}+\frac{\Delta x^2}{2\Delta t}\frac{\partial^2 u}{\partial x^2}$$

$$\frac{\partial u}{\partial t}+v\frac{\partial u}{\partial x}+\frac{\Delta t}{2}\frac{\partial^2 u}{\partial t^2}-\frac{\Delta x^2}{2\Delta t}\frac{\partial^2 u}{\partial x^2}=0. \tag{14.3}$$

Estos últimos términos extra, que no estaban originalmente en (14.1) constituyen una difusión que emerge de la solución numérica, pues

$$\frac{\partial u}{\partial t}=-v\frac{\partial u}{\partial x}$$

$$\frac{\partial^2 u}{\partial t^2}=-v\frac{\partial^2 u}{\partial t\partial x}=-v\frac{\partial}{\partial x}\left(-v\frac{\partial u}{\partial x}\right)$$

entonces

$$\frac{\partial^2 u}{\partial t^2}=v^2\frac{\partial^2 u}{\partial x^2}$$

y reemplazando en (14.3)

$$\frac{\partial u}{\partial t}+v\frac{\partial u}{\partial x}+\frac{v^2\Delta t}{2}\frac{\partial^2 u}{\partial x^2}-\frac{\Delta x^2}{2\Delta t}\frac{\partial^2 u}{\partial x^2}=0$$

$$\frac{\partial u}{\partial t} + v\frac{\partial u}{\partial x} = \left[\frac{\Delta x^2}{2\Delta t} - \frac{v^2\Delta t}{2}\right]\frac{\partial^2 u}{\partial x^2}$$

que tiene la forma

$$\frac{\partial u}{\partial t} + v\frac{\partial u}{\partial x} = \kappa\frac{\partial^2 u}{\partial x^2}$$

Si $\partial_x u = 0$ resulta en una ecuación de difusión con constante κ. El método introduce una difusión numérica asociada a

$$\kappa \equiv \frac{\Delta x^2}{2\Delta t} - \frac{v^2\Delta t}{2}.$$

Esta difusión "achata" la solución y la hace, como vimos, estable pero a costo de que se altera la solución física. Necesitamos $\kappa > 0$ para la estabilización,

$$\frac{\Delta x^2}{\Delta t} > v^2\Delta t \qquad \Rightarrow \qquad \frac{\Delta x}{\Delta t} > v,$$

lo cual lleva a la condición CFL. Si $\kappa < 0$ el esquema no es estable. Se acumula el error y en algún momento explota.

§ 15. Dispersión de los métodos

Para el método de Lax se tenía el factor de amplificación dado por (14.2) cuya estabilidad dependía del cumplimiento de la condición CFL. Analicemos ahora la dependencia con la frecuencia de dicho factor que se halla en el seno que multiplicaba al corchete. Si

$$k\Delta x \ll 1$$

se tendrá

$$|\lambda|^2 \approx 1 - (k\Delta x)^2\left[1 - \left(\frac{v\Delta t}{\Delta x}\right)^2\right]$$

de forma que los modos con k pequeño se ven levemente amortiguados en comparación a aquellos que tienen k grande (siempre dentro de la aproximación $k\Delta x \ll 1$).

Un esquema como el de Lax filtra las oscilaciones de alta frecuencia. Veamos la relación de dispersión de estos métodos con la ecuación advectiva.

Proponemos para ello un modo Fourier en espacio y tiempo,

$$u = u\,\mathrm{e}^{i(\omega t - kx)}$$

el cual es solución exacta de la ecuación si se cumple que

$$v = \omega/k$$

que es también la velocidad de fase de la onda.

Como el $\omega \in \mathbb{R}$ no hay amortiguamiento pues se da que $k \in \mathbb{R}$ también.

En el esquema de Lax el modo Fourier resulta

$$u\,\mathrm{e}^{i(\omega(t^n+\Delta t)-kx_j)} = \frac{1}{2}u\,\mathrm{e}^{i(\omega t^n - kx_j)}(\mathrm{e}^{-ik\Delta x} + \mathrm{e}^{ik\Delta x}) - \frac{v\Delta t}{2\Delta x}u\,\mathrm{e}^{i(\omega t^n - kx_j)}(\mathrm{e}^{-ik\Delta x} - \mathrm{e}^{ik\Delta x})$$

$$\mathrm{e}^{i\omega\Delta t} = \frac{1}{2}(\mathrm{e}^{-ik\Delta x} + \mathrm{e}^{ik\Delta x}) - \frac{v\Delta t}{2\Delta x}(\mathrm{e}^{-ik\Delta x} - \mathrm{e}^{ik\Delta x})$$

$$\mathrm{e}^{i\omega\Delta t} = \cos(k\Delta x) + i\frac{v\Delta t}{\Delta x}(\mathrm{sen}(k\Delta x))$$

Si pensamos en una frecuencia compleja $\omega \in \mathbb{C}$ de la forma

$$\omega = \Omega + i\Gamma,$$

entonces

$$\mathrm{e}^{i\Omega\Delta t}\,\mathrm{e}^{-\Gamma\Delta t} = \cos(k\Delta x) + i\frac{v\Delta t}{\Delta x}(\mathrm{sen}(k\Delta x))$$

e igualando partes real e imaginaria por separado se tiene

$$\cos(\Omega\Delta t)\,\mathrm{e}^{-\Gamma\Delta t} = \cos(k\Delta x)$$

$$\mathrm{sen}(\Omega\Delta t)\,\mathrm{e}^{-\Gamma\Delta t} = \frac{v\Delta t}{\Delta x}(\mathrm{sen}(k\Delta x))$$

luego

$$\tan(\Omega\Delta t) = \frac{v\Delta t}{\Delta x}\tan(k\Delta x)$$

$$\mathrm{e}^{-2\Gamma\Delta t} = \cos(k\Delta x)^2 + \left(\frac{v\Delta t}{\Delta x}\right)(\mathrm{sen}(k\Delta x))^2$$

Si se diera $v\Delta t/\Delta x = 1$ (velocidad de la onda igual a la numérica), en ese caso sería $\Gamma = 0$, $\Omega = vk$ y es una solución exacta.

En cualquier otro caso hay que resolver numéricamente. La Figura 15.1 exhibe las curvas $\Omega(k)$ para tres valores diferentes del parámetro $v\Delta t/Dx$, y la Figura 15.2 hace lo propio para $\Gamma(k)$ para idénticos valores del parámetro.

El amortiguamiento es máximo para el valor $\pi/2\Delta x$.

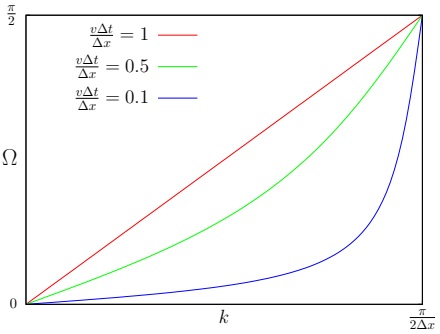


Figura 15.1 Relaciones de dispersión $\Omega(k)$.

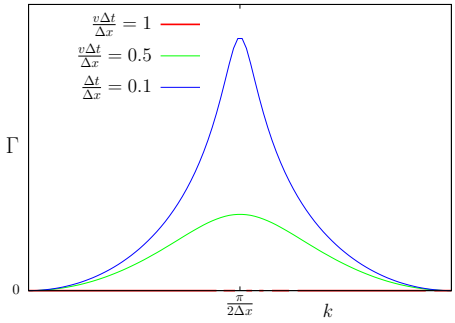


Figura 15.2 Relaciones de dispersión $\Gamma(k)$.

§ 16. Otros métodos para la ecuación de advección

También se puede utilizar leapfrog para la parte temporal, lo cual lleva la ecuación (14.1) al siguiente esquema

$$\frac{u_j^{n+1} - u_j^{n-1}}{2\Delta t} + v\frac{u_{j+1}^n - u_{j-1}^n}{2\Delta x} = 0$$

que es de orden dos en Δx y Δt. La estabilidad se comprueba utilizando un modo

$$E^n \mathrm{e}^{ikx_j}$$

para lo cual se obtiene

$$\frac{E^{n+1} - E^{n-1}}{2\Delta t} + v\frac{E^n \mathrm{e}^{ik\Delta x} - E^n \mathrm{e}^{-ik\Delta x}}{2\Delta x} = 0$$

$$\frac{\lambda - \lambda^{-1}}{2\Delta t} + v\frac{\mathrm{e}^{ik\Delta x} - \mathrm{e}^{-ik\Delta x}}{2\Delta x} = 0$$

$$\frac{\lambda - \lambda^{-1}}{2\Delta t} + v\frac{i\,\mathrm{sen}(k\Delta x)}{\Delta x} = 0$$

$$\lambda^2 - 1 + \frac{v\Delta t}{\Delta x} 2\lambda i\,\mathrm{sen}(k\Delta x) = 0$$

resultando una cuadrática para λ. Si tomamos

$$\alpha \equiv \frac{v\Delta t}{\Delta x}\,\mathrm{sen}(k\Delta x)$$

las raíces son

$$\lambda_{\pm} = i\alpha \pm \sqrt{1-\alpha^2}$$

y puede verse que si $\alpha < 1$ se cumple que $|\lambda| = 1$ y el método es condicionalmente estable. No obstante este método requiere, computacionalmente, más memoria que Lax porque es necesario guardar un paso de tiempo más en el cálculo.

Otro método conocido es Lax-Wendroff, que implica un medio paso extra. Este medio paso consiste en

$$u_{j+1/2}^{n+1/2} = \frac{1}{2}(u_j^n + u_{j+1}^n) - \frac{v\Delta t}{2\Delta x}(u_{j+1}^n - u_j^n)$$

siendo el paso completo

$$u_j^{n+1} = u_j^n - \frac{v\Delta t}{\Delta x}(u_{j+1/2}^{n+1/2} - u_{j-1/2}^{n+1/2}).$$

Este método también resulta estable con la misma condición

$$v\Delta t/\Delta x \leq 1.$$

Finalmente otro método conocido como "aguas arriba" (*upwind*), implícito, parte de la premisa de Euler para el tiempo pero con una derivada espacial implícita; se hace en el paso sisguiente. El método resulta de orden uno.

La ecuación discretizada resulta en

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} + v\frac{u_j^{n+1} - u_{j-1}^{n+1}}{\Delta x} = 0,$$

y es incondicionalmente estable. Por ello suele tener buena fama.

La idea pictórica es que el método *construye* desde el contorno, como se ve en la Figura 16.3.

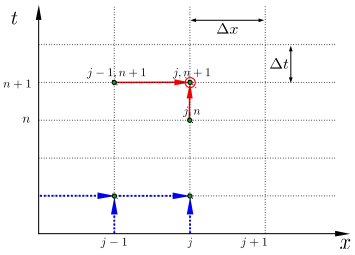


Figura 16.3 Esquema método *upwind*.

EJEMPLO 5.1. Estabilidad esquema upwind
Examínese la estabilidad del esquema *upwind* para la ecuación de advección lineal. Dado que este esquema es

$$\frac{u_j^{n+1}-u_j^n}{\Delta t}+v\frac{u_j^{n+1}-u_{j-1}^{n+1}}{\Delta x}=0$$

podemos examinar el comportamiento de un modo Fourier $E^n \mathrm{e}^{ikx_j}$ en la expresión anterior. Haciendo el reemplazo y multiplicando por Δt se llega a

$$E^{n+1}-E^n+\frac{v\Delta t}{\Delta x}E^{n+1}\left(\mathrm{e}^{ik\Delta x}-\mathrm{e}^{ik\Delta x}\right)=0$$

y dividiendo todo por E^n y usando la identidad de Euler y la equivalencia usual $\Gamma=\frac{v\Delta t}{\Delta x}$,

$$\lambda-1+\lambda 2i\Gamma(sen(k\Delta x))=0$$

o bien

$$\lambda=\frac{1}{1-2i\Gamma\,\mathrm{sen}(k\Delta x)}$$

y entonces

$$|\lambda|=\frac{1}{\sqrt{1+4\Gamma^2\,\mathrm{sen}^2(k\Delta x)}}$$

el cual es un número siempre menor a la unidad en razón de que el denominador siempre es mayor que la unidad. De esta forma vemos que el esquema resulta incondicionalmente estable.

§ 17. Ecuación de advección 2D

Para dos dimensiones espaciales (x, y) es

$$\frac{\partial u}{\partial t}+v_x\frac{\partial u}{\partial x}+v_y\frac{\partial u}{\partial y}=0 \tag{17.1}$$

donde $u(x, y, t)$ es la solución y $\mathbf{v}=(v_x, v_y)$ es constante.

El esquema de Lax resulta ahora

$$\frac{u_{i,j}^{n+1}-(1/4)\left(u_{i+1,j}^n+u_{i-1,j}^n+u_{i,j+1}^n+u_{i,j-1}^n\right)}{\Delta t} + v_x\frac{u_{i+1,j}^n-u_{i-1,j}^n}{2\Delta x}+v_y\frac{u_{i,j+1}^n-u_{i,j-1}^n}{2\Delta y}=0 \tag{17.2}$$

El análisis de estabilidad se hace con un modo 2D como

$$E^n \mathrm{e}^{i(k_x x_i + k_y y_j)}$$

el cual al reemplazar en (17.2) verifica la condición $|\lambda| \leq 1$ si cumple

$$\left(\frac{v_x \Delta t}{\Delta x}\right)^2 + \left(\frac{v_y \Delta t}{\Delta y}\right)^2 \leq 1,$$

o bien

$$\Delta t \leq \frac{1}{\sqrt{2}} \frac{1}{\sqrt{(\frac{v_x}{\Delta x})^2 + (\frac{v_y}{\Delta y})^2}}.$$

Si se da el caso simétrico $\Delta x = \Delta y$ se obtiene

$$\Delta t \leq \frac{1}{\sqrt{2}} \frac{\Delta x}{\sqrt{v_x^2 + v_y^2}} = \frac{\Delta x}{\sqrt{2}\, v},$$

que es la condición CFL.

Se ve claramente que si no se conoce en principio v es problemático el método porque la estabilidad justamente depende de la velocidad v. Debería poder, al menos, ser acotada.

Capítulo 6

Ecuaciones elípticas

§ 18. Ecuación de Poisson

Veremos el ejemplo canónico de las ecuaciones elípticas, que es la ecuación de Poisson,

$$\frac{d^2\phi}{dx^2} = -\rho(x) \quad \text{en} \quad [0, L]$$

donde ρ describe las fuentes. Las ecuaciones elípticas son problemas de contorno, son *boundary value problems*.

Las condiciones de contorno suelen ser de dos tipos,

$$\phi(0), \phi(L) \qquad [\text{ Dirichlet }]$$

$$\left.\frac{\partial\phi}{\partial x}\right|_0, \left.\frac{\partial\phi}{\partial x}\right|_L \qquad [\text{ Von Neumann }]$$

Discretizaremos, como es costumbre,

$$x_j \to j\Delta x$$

$$\phi_j \to \phi(x_j)$$

$$\rho_j \to \rho(x_j)$$

y la derivada segunda con la aproximación de orden $\mathcal{O}\left(\Delta x^2\right)$

$$\frac{d^2\phi_j}{dx^2} = \frac{\phi_{j+1} - 2\phi_j + \phi_{j-1}}{\Delta x^2}$$

y entonces

$$\phi_{j+1} - 2\phi_j + \phi_{j-1} = -\rho_j \Delta x^2$$

siendo

$$\boldsymbol{\phi} = (\phi_1, \phi_2, ..., \phi_{N-1})$$

incógnitas y ϕ_0, ϕ_N datos. Entonces podemos escribir el sistema como

$$A\boldsymbol{\phi} = \boldsymbol{\omega}$$

siendo

$$A = \begin{pmatrix} -2 & 1 & 0 & 0 & ... & 0 \\ 1 & -2 & 1 & 0 & ... & 0 \\ 0 & 1 & -2 & 1 & & 0 \\ ... & & & & & ... \\ 0 & ... & & 1 & -2 & 1 \\ 0 & ... & & 0 & 1 & -2 \end{pmatrix}$$

$$\boldsymbol{\omega} = (-\phi_0 - \rho_1 \Delta x^2, -\rho_2 \Delta x^2, ..., -\phi_N - \rho_{N-1} \Delta x^2)$$

y hay que invertir la matriz para obtener la solución. Es una matriz tridiagonal que se puede resolver por Crank-Nicolson y el algoritmo TDMA (recursivo). El número de operaciones necesario para la resolución es $\mathcal{O}(N-1) \approx \mathcal{O}(N)$.

Este procedimiento sirve para ecuaciones de la forma

$$f(x)\frac{d^2\phi}{dx^2} + g(x)\frac{d\phi}{dx} + h(x)\phi = \omega(x)$$

donde la matriz A será más complicada en general y las condiciones de contorno pueden mezclar los dos tipos vistos,

$$CC = \begin{cases} a \left.\frac{\partial \phi}{\partial x}\right|_0 + b\,\phi_0 = C \\ \\ a' \left.\frac{\partial \phi}{\partial x}\right|_L + b'\phi_L = C \end{cases}$$

Para dos dimensiones la ecuación de Poisson será

$$\nabla^2 \phi = -\rho(x, y) \tag{18.1}$$

y en un dominio rectangular

$$0 \le x \le L_x \qquad 0 \le y \le L_y$$

con condiciones de contorno

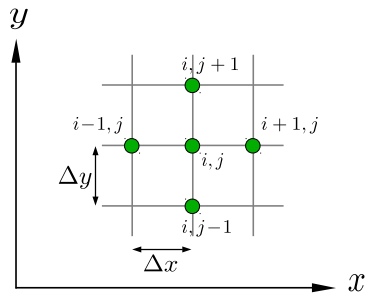
$$\phi(0, y) = dato \qquad \phi(x, 0) = dato$$

$$\phi(L_x, y) = dato \qquad \phi(x, L_y) = dato$$

y discretización dada por

$$y_j = j\Delta y \qquad x_i = i\Delta x$$

$$j = 0, 1, ..., M \qquad i = 0, 1, ..., N$$



Si el borde es arbitrario (irregular) podemos afinar allí la grilla o aproximar esta irregularidad por *cuadraditos* del mismo tamaño. Finalmente la discretización de (18.1) es

$$\frac{\phi_{i+1,j} - 2\phi_{ij} + \phi_{i-1,j}}{\Delta x^2} + \frac{\phi_{i,j+1} - 2\phi_{i,j} + \phi_{i,j-1}}{\Delta y^2} = -\rho_{ij}$$

lo cual amasando (y concediendo la igualdad de espaciados: $\Delta \equiv \Delta x = \Delta y$) resulta en

$$\phi_{i+1,j} + \phi_{i-1,j} + \phi_{i,j+1} + \phi_{i,j-1} - 4\phi_{ij} = -\Delta^2 \rho_{ij}$$

Si introducimos este resultado en una matriz ordenando por filas o por columnas

$$\phi_{i,j} \to \phi_{(i-1)\times(M-1)+j},$$

o bien

$$\phi_{i,j} \to \phi_{(j-1)\times(N-1)+i}$$

con

$$i = 1, 2, ..., N-1 \qquad j = 1, 2, ..., M-1$$

Todo esto conduce a un sistema de la forma

$$A\boldsymbol{\phi} = \boldsymbol{\omega}$$

donde $\boldsymbol{\phi} \in (N-1)(M-1)$ y A es una matrix rala, de bandas.

$$A = \begin{pmatrix} -4 & 1 & 0 & ... & 1 & \\ 1 & -4 & 1 & 0 & ... & 1 \\ 0 & 1 & -4 & 1 & 0 & \\ 0 & 0 & 1 & ... & & \\ ... & & & & & \\ 0 & & & & & -4 \end{pmatrix}.$$

§ 19. Sistemas matriciales

Comenzamos aquí una digresión, más o menos larga, referida a la resolución de sistemas matriciales del tipo que nos hemos encontrado ya en varias ocasiones como resultado de la resolución de un sistema numérico. Es decir partimos de

$$A\mathbf{u} = \boldsymbol{\omega},$$

o bien

$$a_{11}u_1 + a_{12}u_2 + ... + a_{1N}u_N = \omega_1$$

$$a_{21}u_1 + a_{22}u_2 + ... + a_{2N}u_N = \omega_2$$

$$.....$$

$$a_{N1}u_1 + a_{N2}u_2 + ... + a_{NN}u_N = \omega_N,$$

que es un sistema de n ecuaciones con n incógnitas. La forma directa de invertir la matriz es por determinantes

$$A^{-1}_{ij} = -i^{1+i}\frac{|A|^{ij}}{|A|}$$

donde

$$|A|^{ij} \equiv \det(\mathscr{A})$$

y $\mathscr{A}$ es la matriz que resulta de extraer la fila i, columna j. $|A|$ es el determinante de A. Esto tiene un número de operaciones que es $\mathcal{O}(n!)$ lo cual es obviamente inmenso. Este método es, por dicha causa, impráctico.

El siguiente método es eliminación gaussiana. Las incógnitas u_1, u_2 son eliminadas sucesivamente hasta que sobrevive solamente una única variable. Si multiplicamos la segunda ecuación por a_{11}/a_{21} y le resto la ecuación primera.

$$(a_{22}\frac{a_{11}}{a_{21}}u_2 - a_{12}u_2) + ... = \omega_2\frac{a_{11}}{a_{21}} - \omega_1$$

y se sigue hasta que resulte una única variable u_n.

Finalmente tenemos un sistema triangular (método LU). Este método lleva una matriz A a expresarla como un producto

$$A = LU$$

donde es L de *lower* (tiene elementos no nulos en la diagonal y por debajo) y U de *upper* (tiene elementos no nulos en la diagonal y por encima). Es decir

$$A = \begin{pmatrix} a & 0 & & ... & 0 \\ b & c & 0 & ... & 0 \\ ... & & & & \\ & & & & z \end{pmatrix} \begin{pmatrix} a & b & ... & & \\ 0 & c & ... & ... & \\ 0 & ... & & & \\ 0 & ... & & 0 & z \end{pmatrix}$$

Este método tiene un número de operaciones de $\mathcal{O}\left((2/3)n^3\right)$. Es lo mejor que podemos hacer para obtener solución exacta de la ecuación matricial.

Otros métodos son aproximados (solución iterativa).

§ 19.1. Métodos iterativos

Se aplican para la solución de sistemas

$$A\mathbf{u} = \mathbf{d}$$

y son como anticipamos, aproximados. Veremos tres de ellos, a saber: Jacobi, Gauss-Seidel y SOR (*Sucessive Over Relaxation*).

En la ecuación de Poisson para dos dimensiones

$$\nabla^2\phi = -\rho(x, y)$$

había resultado la discretización

$$\phi_{i+1,j} + \phi_{i-1,j} + \phi_{i,j+1} + \phi_{i,j-1} - 4\phi_{ij} = -\Delta^2\rho_{ij}$$

para $\Delta \equiv \Delta x = \Delta y$.

Podíamos despejar el valor ϕ_{ij} como

$$\phi_{i,j} = \frac{1}{4}(\phi_{i+1,j} + \phi_{i-1,j} + \phi_{i,j+1} + \phi_{i,j-1}) + \frac{\Delta^2}{4}\rho_{ij}$$

y si pensamos que los valores en el lado derecho son conocidos en un paso n se puede construir la siguiente iteración

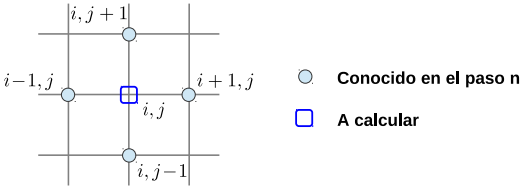
$$\phi^{n+1}_{i,j} = \frac{1}{4}(\phi^n_{i+1,j} + \phi^n_{i-1,j} + \phi^n_{i,j+1} + \phi^n_{i,j-1}) + \frac{\Delta^2}{4}\rho_{ij}$$

donde ρ_{ij} son datos del problema.

Se esperaría que con un número grande de pasos n se dé el límite

$$|\phi_{i,j}^{n+1} - \phi_{i,j}^{n}| \to 0,$$

es decir, que la solución converja. Este es el método de Jacobi para la ecuación de Poisson. Esquemáticamente



Existe una equivalencia con la evolución temporal del problema

$$\frac{\partial \phi}{\partial t} = \nu(\nabla^2 \phi + \rho)$$

donde se ha alcanzado el régimen estacionario. Con las discretizaciones consabidas resulta

$$\frac{\phi_{ij}^{n+1} - \phi_{ij}^{n}}{\Delta t} = \frac{\nu}{\Delta^2}(\phi_{i+1,j}^{n} + \phi_{i-1,j}^{n} + \phi_{i,j+1}^{n} + \phi_{i,j-1}^{n} - 4\phi_{ij}^{n}) + \nu\rho_{ij}^{n}$$

$$\phi_{ij}^{n+1} = \frac{\nu\Delta t}{\Delta^2}(\phi_{i+1,j}^{n} + \phi_{i-1,j}^{n} + \phi_{i,j+1}^{n} + \phi_{i,j-1}^{n} - 4\phi_{ij}^{n}) + \nu\Delta t\rho_{ij}^{n} + \phi_{ij}^{n}$$

$$\phi_{ij}^{n+1} = \frac{\nu\Delta t}{\Delta^2}(\phi_{i+1,j}^{n} + \phi_{i-1,j}^{n} + \phi_{i,j+1}^{n} + \phi_{i,j-1}^{n}) + \left(1 - 4\frac{\nu\Delta t}{\Delta^2}\right)\phi_{ij}^{n} + \nu\Delta t\rho_{ij}^{n},$$

y esto garantiza la convergencia. O sea, hacemos una evolución temporal porque vemos que es lo mismo y en base a ella deducimos que converge.

Un detalle de programación es que necesito la solución en el paso n para obtener la solución en el paso $n+1$. La iteración doble conviene hacerla de acuerdo a cómo se guardan las matrices en la máquina.

En el problema $A\mathbf{u} = \mathbf{d}$ en forma indicial

$$\sum_{j=1}^{n} a_{ij}u_j = d_i \qquad j = 1, 2, ..., m$$

$$u_i = \frac{1}{a_{ii}}[d_i - \sum_{\substack{j=1 \\ j\neq i}}^{m} a_{ij}u_j]$$

El método de Jacobi es, entonces, decir que

$$u_i^{n+1} = \frac{1}{a_{ii}}[d_i^n - \sum_{\substack{j=1 \\ j\neq i}}^{m} a_{ij}u_j^n].$$

Otra forma de trabajar es dividir la matriz en tres

$$A = D + L + U$$

entonces el método de Jacobi se escribe como siendo

$$A\mathbf{u} = \mathbf{d}$$

$$(D + L + U)\mathbf{u} = \mathbf{d}$$

$$D\mathbf{u} = \mathbf{d} - (L + U)\mathbf{u} \quad \Rightarrow \quad D\mathbf{u}^{n+1} = \mathbf{d}^n - (L + U)\mathbf{u}^n$$

Se define el residuo R^n en la iteración n como

$$R^n = A\mathbf{u}^n - \mathbf{d}^n$$

si ahora llamo $\Delta\mathbf{u}^n \equiv \mathbf{u}^{n+1} - \mathbf{u}^n$ será

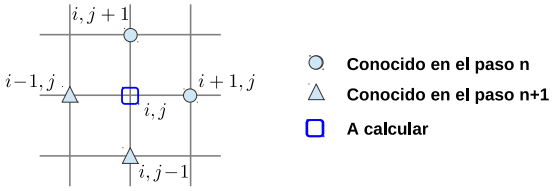
$$D\Delta\mathbf{u}^n = D\mathbf{u}^{n+1} - D\mathbf{u}^n = \mathbf{d}^n - (L + U)\mathbf{u}^n - D\mathbf{u}^n$$

$$D\Delta\mathbf{u}^n = \mathbf{d}^n - (L + U + D)\mathbf{u}^n = -R^n,$$

entonces se dice que el método de Jacobi *conduce* $\Delta\mathbf{u}^n$ a cero a través de D.

§ 19.2. Método de Gauss-Seidel

Es una mezcla. Uso valores ya iterados para evaluar el próximo punto. Utilizo $\phi_{i-1,j}^{n+1}, \phi_{i,j-1}^{n+1}$ para hallar $\phi_{i,j}^n$ en los puntos faltantes



Si lo vemos matricialmente será

$$u_i = \frac{1}{a_{ii}}[d_i - \sum_{\substack{j=1 \\ j\neq i}}^{m} a_{ij}u_j]$$

$$u_i^{n+1} = \frac{1}{a_{ii}}[d_i^n - \sum_{j=1}^{i-1} a_{ij}u_j^n - \sum_{j=i+1}^{m} a_{ij}u_j^n]$$

mientras que en términos de las otras matrices será

$$D\mathbf{u}^{n+1} + L\mathbf{u}^{n+1} = \mathbf{d}^n - U\mathbf{u}^n$$

$$(D+L)\mathbf{u}^{n+1} = \mathbf{d}^n - U\mathbf{u}^n$$

y también cambia el residuo y las diferencias entre pasos,

$$(D+L)\mathbf{u}^n = \mathbf{d}^n - U\mathbf{u}^n - (D+L)\mathbf{u}^n$$

$$(D+L)\Delta\mathbf{u}^n = -R^n$$

entonces el método de Gauss-Seidel conduce Δu^n a cero a través de $D+L$, que es llamado el prefactor.

§ 19.3. Métodos SOR

Si $\mathbf{u}^{n+1}$ es el valor obtenido con el esquema iterativo básico Jacobi o Gauss-Seidel el valor final se obtiene con

$$\mathbf{u}^{n+1} = \omega\mathbf{u}^{n+1} + (1-\omega)\mathbf{u}^n$$

donde ω es un coeficiente de sobre-relajación.

Para el caso de la ecuación de Poisson, Jacobi junto a sobre-relajación resultan en

$$\phi_{ij}^{n+1} = \frac{\omega}{4}(\phi_{i+1,j}^n + \phi_{i-1,j}^n + \phi_{i,j+1}^n + \phi_{i,j-1}^n + \Delta^2\rho_{ij}^n) + (1-\omega)\,\phi_{ij}^n.$$

Para la ecuación general

$$A\mathbf{u} = \mathbf{d}$$

con $A = D + L + U$

$$D\mathbf{u}^{n+1} = \omega\mathbf{d}^n - \omega(L+U)\mathbf{u}^n + (1-\omega)D\mathbf{u}^n$$

$$D\mathbf{u}^{n+1} = \omega(\mathbf{d}^n - A\mathbf{u}^n) + D\mathbf{u}^n$$

luego

$$D\mathbf{u}^{n+1} - D\mathbf{u}^n = D\Delta\mathbf{u}^n = \omega(\mathbf{d}^n - A\mathbf{u}^n) = -\omega R^n$$

y vemos que conduce $\Delta\mathbf{u}$ a cero ponderado con el ω.

Para Gauss-Seidel y sobre-relajación (que es el método SOR por antonomasia), y en la ecuación de Poisson, será

$$\bar{\phi}_{ij}^{n+1} = \frac{1}{4}(\phi_{i+1,j}^n + \phi_{i-1,j}^{n+1} + \phi_{i,j+1}^n + \phi_{i,j-1}^{n+1}) + \frac{\Delta^2}{4}\rho_{ij}^n$$

$$\phi_{ij}^{n+1} = \omega\phi_{ij}^{n+1} + (1-\omega)\,\phi_{ij}^{n}.$$

Para la ecuación general resultará

$$D\mathbf{u}^{n+1} + \omega L\mathbf{u}^{n+1} = \omega\mathbf{d}^n - \omega U\mathbf{u}^n + (1-\omega)D\mathbf{u}^n$$

$$(D+\omega L)\Delta\mathbf{u}^n = -\omega R^n$$

y conduce $\Delta\mathbf{u}$ a cero de acuerdo a la relación mostrada. El asunto es, entonces, cómo elegir sabiamente ω.

§ 19.4. Convergencia de los métodos iterativos

Introduciremos siempre un error

$$R^n = A\mathbf{u}^n - \mathbf{d},$$

pero supongamos la solución exacta del sistema $A\mathbf{u} = \mathbf{d}$ que sea

$$\mathbf{u}_e \;:\quad A\mathbf{u}_e = \mathbf{d}$$

entonces definimos el error en la iteración n como

$$\boldsymbol{\varepsilon}^n = \mathbf{u}^n - \mathbf{u}_e$$

$$R^n = A\mathbf{u}^n - \mathbf{d} - [A\mathbf{u}_e - \mathbf{d}],$$

siendo el último corchete nulo por definición,

$$R^n = A(\mathbf{u}^n - \mathbf{u}_e) = A\boldsymbol{\varepsilon}^n$$

luego si el $R^n \to 0$ entonces $\boldsymbol{\varepsilon}^n \to 0$, y se puede ver que es un sí y sólo sí.

Supongamos una subdivisión arbitraria de

$$A = P + S$$

y diseñemos un esquema iterativo para resolver $A\mathbf{u} = \mathbf{d}$, que será

$$(P+S)\mathbf{u} = \mathbf{d}$$

$$P\mathbf{u} = \mathbf{d} - S\mathbf{u}$$

y la iteración se hace sobre

$$P\mathbf{u}^{n+1} = \mathbf{d} - S\mathbf{u}^n$$

luego

$$P\Delta\mathbf{u}^n = -R^n$$

con

$$R^n = A\mathbf{u}^n - \mathbf{d}.$$

Entonces P es la matriz de convergencia o precondicionador. Se elige parecida a A pero en lo posible mucho más fácil de diagonalizar. En el caso de Jacobi es sencillamente la diagonal.

Notemos que

$$P\boldsymbol{\varepsilon}^{n+1} = P\mathbf{u}^{n+1} - P\mathbf{u}_e$$

$$P\boldsymbol{\varepsilon}^{n+1} = \mathbf{d} - S\mathbf{u}^n - P\mathbf{u}_e,$$

y utilizando la solución exacta

$$(P+S)\mathbf{u}_e = \mathbf{d}$$

$$P\mathbf{u}_e = \mathbf{d} - S\mathbf{u}_e$$

entonces

$$P\boldsymbol{\varepsilon}^{n+1} = \mathbf{d} - S\mathbf{u}^n - \mathbf{d} + S\mathbf{u}_e$$

$$P\boldsymbol{\varepsilon}^{n+1} = S(-\mathbf{u}^n + \mathbf{u}_e)$$

$$\boldsymbol{\varepsilon}^{n+1} = -P^{-1}S\boldsymbol{\varepsilon}^n$$

y la matriz $P^{-1}S$ es una amplificadora del error

$$\boldsymbol{\varepsilon}^{n+1} = -P^{-1}(A-P)\boldsymbol{\varepsilon}^n$$

$$\boldsymbol{\varepsilon}^{n+1} = (-P^{-1}A + \mathbb{1})\boldsymbol{\varepsilon}^n$$

$$\boldsymbol{\varepsilon}^{n+1} = G\boldsymbol{\varepsilon}^n$$

y G es la llamada matriz de amplificación. Finalmente

$$\boldsymbol{\varepsilon}^{n+1} = G^n\boldsymbol{\varepsilon}^1.$$

Si quiero que $|\boldsymbol{\varepsilon}^{n+1}| \to 0$ con $n \to \infty$ necesitaré que

$$\text{máx}\{|\lambda_j(G)|\} \leq 1$$

donde los $\lambda_j(G)$ son los autovalores de G.

Se define el radio espectral

$$\sigma(G) \equiv \text{máx}\{|\lambda_j(G)|\} < 1$$

Entonces para Jacobi,

$$D\mathbf{u}^{n+1} = \mathbf{d} - (L+U)\mathbf{u}^n$$

$$P = D \quad S = L + U$$
$$G = 1 - D^{-1}(D + L + U)$$
$$G = -D^{-1}(L + U)$$

mientras que para Gauss-Seidel

$$(D + L)\mathbf{u}^{n+1} = \mathbf{d} - U\mathbf{u}^n$$

$$P = D + L \quad S = U$$
$$G = 1 - (D + L)^{-1}(D + L + U)$$
$$G = -(D + L)^{-1}U$$

Se puede ver que ambos métodos convergen si A es dominante diagonal.
La tasa de reducción del error se define como

$$\Delta\varepsilon_n = \left(\frac{|\boldsymbol{\varepsilon}^n|}{|\boldsymbol{\varepsilon}^1|}\right)^{1/n}$$

y como

$$\varepsilon^n = G^n\boldsymbol{\varepsilon}^1$$
$$\Delta\boldsymbol{\varepsilon}_n \leq ||G^n||^{1/n}$$
$$\xrightarrow[n\to\infty]{} \sigma(G)$$

La tasa de convergencia es

$$s = \log\sigma(G)$$

o bien

$$s = |\log\sigma(G)|$$

Entonces, si quiero reducir el error 10 veces en un número n de iteraciones tal que

$$\underbrace{(1/10)^{1/n}}_{A} \geq \overbrace{\sigma(G)}^{B} \geq \Delta\varepsilon_n$$

se logra asegurando la parte A y queda demostrado B por el resultado

$$\log((1/10)^{1/n}) \geq \log(\sigma(G))$$
$$\frac{1}{n}\log((1/10)) \geq \log(\sigma(G))$$
$$-\frac{1}{n} \geq \log(\sigma(G))$$

$$n \geq \frac{1}{-\log(\sigma(G))}.$$

Ahora queremos aplicar estos resultados a los métodos vistos.

Para la ecuación de Poisson en la resolución por el método de Jacobi bajo condiciones de contorno periódicas la matriz lucía

$$A = \begin{pmatrix} -4 & 1 & 0 & ... & 1 & 0 & 0 \\ 1 & -4 & 1 & ... & 0 & 1 & 0 \\ 0 & 1 & -4 & ... & 0 & 0 & 1 \\ ... & ... & ... & ... & ... & & \\ ... & ... & ... & ... & 1 & -4 & 1 \\ 0 & ... & ..... & & & 1 & -4 \end{pmatrix}$$

y se puede ver que

$$\phi_{ij}^{lm} = \text{sen}\left(\frac{\pi l i}{M}\right) \text{sen}\left(\frac{\pi m j}{M}\right)$$

es autofunción de A con $i, j = 1, 2, ..., M$ y con $l, m = 1, 2, ..., M$. El par (i, j) varía con la posición en grilla. Los (l, m) identifican función y pueden ser cualesquiera. El autovalor será

$$\Omega_{lm} = 4\left(\text{sen}^2\left(\frac{\pi l}{2M}\right) + \text{sen}^2\left(\frac{\pi m}{2M}\right)\right)$$

y como $G = \mathbb{1} - D^{-1}A$ se sigue que

$$\lambda = 1 - \frac{\Omega_{lm}}{d}$$

son los autovalores de G siendo d el elemento diagonal. Resultan

$$\lambda = 1 - \left(\text{sen}^2\left(\frac{\pi l}{2M}\right) + \text{sen}^2\left(\frac{\pi m}{2M}\right)\right)$$

$$\lambda = \frac{1}{2}\left(\cos\left(\frac{\pi l}{M}\right) + \cos\left(\frac{\pi m}{M}\right)\right)$$

y el autovalor más grande para todo $l = m$ que ocurre $(1, 2, ..., M-1)$ se da con $l = m = 1$ de manera que

$$\sigma = \cos\left(\frac{\pi}{M}\right) < 1$$

Para $M \gg 1$ podemos hacer una expansión de Taylor y tendremos

$$\sigma = 1 - \frac{\pi^2}{2M^2}$$

Como ejemplo, para $M = 100$ es

$$\sigma = 0{,}9995 \quad \log(\sigma) = -0{,}0002$$

y necesito aproximadamente 4664 iteraciones para reducir el error en un factor de 10.

Computacionalmente para $M = 100$ son 100×100 cálculos en una iteración de modo que el número de operaciones total es de $\approx 05 \cdot 10^3 \cdot 10^2 \cdot 10^2 = 5 \cdot 10^7$. Si hubiésemos hecho eliminación gaussiana por Choleski, el número de operaciones es $\approx n^3/3 = (M^2)^3/3 = 10^{12}/3 \approx 3,4 \cdot 10^{12}$, el cual es mucho mayor.

Para Gauss-Seidel esto baja un factor de dos. Se puede ver que

$$\lambda = \frac{1}{4}\left(\cos\left(\frac{\pi l}{M}\right) + \cos\left(\frac{\pi m}{M}\right)\right)^2$$

ocurriendo el valor más chico para $l = m = 1$

$$\sigma = \cos^2\left(\frac{\pi}{M}\right),$$

y pudiéndose escribir para $M \gg 1$

$$\sigma = 1 - \frac{\pi^2}{M^2},$$

lo cual implica un aumento leve de la tasa de convergencia. Para $M = 100$ tengo un número de operaciones de aproximadamente 2300 (la mitad).

En los métodos SOR la tasa se incrementa exponencialmente

$$(D + \omega L)\Delta \mathbf{u}^n = -\omega R^n$$

entonces

$$G = (D + \omega L)^{-1}[(1 - \omega)D - \omega U]$$

con $\boldsymbol{\varepsilon}^{n+1} = G\boldsymbol{\varepsilon}^n$ y $\boldsymbol{\varepsilon}^n = \mathbf{u}^n - \mathbf{u}_e$, pero como

$$\det(G) = \prod_{j=1}^{n} \lambda_j$$

si pido

$$\det(G)^n < 1$$

lo satisfago con

$$\sigma(G) < 1$$

y vale que

$$\det(G)^n = (1 - \omega)^n$$

de modo que finalmente

$$0 < \omega < 2.$$

Existe un ω óptimo que minimiza el radio espectral. Dicho valor es

$$\omega_{opt} = \frac{2}{1 - \sqrt{1 - \sigma_j^2}}$$

donde σ_j es el radio espectral del método de Jacobi sin sobre-relajación. Entonces con ese ω_{opt} resulta

$$\sigma_{SOR} = \omega_{opt} - 1$$

Para el problema de Poisson

$$\sigma_j = \cos\left(\frac{\pi}{M}\right)$$

$$\omega_{opt} = \frac{2}{1 + \operatorname{sen}(\pi/M)}$$

y con $M \gg 1$

$$\omega_{opt} \approx \frac{2}{1 + [\pi/M - \pi^2/M^2]} \approx 2(1 - \pi/M + \pi^2/M^2)$$

$$\sigma_{SOR} = 1 - \frac{2\pi}{M} + \mathcal{O}\left(1/M^2\right)$$

y esto es mejor que con Jacobi y Gauss-Seidel. Para $M = 100$ se tiene un número de iteraciones de aproximadamente 35. Como vemos es una gran mejora.

Capítulo 7

Métodos Espectrales y Pseudoespectrales

§ 20. Métodos espectrales

El problema modelo es

$$\mathscr{L}(U) = f(x) \tag{20.1}$$

donde $U = U(x)$, $\mathscr{L}$ es un operador con derivadas y $f(x)$ son datos. La resolución de (20.1) se hará en el dominio $a \leq x \leq b$ sujeta a las condiciones de contorno $U(a) = U_a$ y $U(b) = U_b$.

La función incógnita U se expande en términos de funciones base,

$$U(x) = \sum_{n=0}^{N} c_n \phi_n(x).$$

Si se supiera que la función $U(x)$ es periódica es conveniente la utilización de funciones trigonométricas

$$\phi_n(x) = \mathrm{e}^{inx}$$

y entonces cualquier función periódica, por Fourier, se puede escribir

$$U(x) = \sum_{n=-\infty}^{\infty} c_n \, \mathrm{e}^{\, in(2\pi/L)x}$$

de manera que la solución $U^N(x)$ no es otra cosa que un truncamiento de la serie de Fourier correspondiente.

Reemplazaremos la aproximación en la ecuación diferencial y pediremos que se minimice el residuo

$$R^N(x) \equiv \mathscr{L}(U) - f(x) \tag{20.2}$$

La minimización de (20.2) definirá el tipo de método. No obstante primeramente debe hacerse la selección de las funciones de base que será de acuerdo a las condiciones de contorno del problema.

Para condiciones de contorno periódicas es conveniente una base de funciones trigonométricas

$$\phi_n(x) = \mathrm{e}^{in(2\pi/L)x}.$$

Otras opciones son utilizar polinomios de Chebyshev, o funciones que solamente sean no nulas en un pequeño dominio (no nulas en secciones). Esto último es la base del método de elementos finitos (FEM, por sus siglas en inglés[1])

Ahora tendríamos que calcular los coeficientes c_n. La complejidad pasa desde las funciones a los coeficientes. Pidiendo la minimización del residuo definimos, como se dijo, diferentes métodos, uno particular es el de Galerkin,

$$\int_a^b R^N(x)\phi_n(x)W(x)dx = 0 \qquad n = 0, 1, 2, ..., N \tag{20.3}$$

donde $W(x)$ son funciones de peso. Matemáticamente esto es pedir la anulación de los productos internos del residuo con cada función de la base, en el espacio de funciones $\{\phi_n\}$. Es decir que (20.3) equivale a

$$\left\langle R^N(x), \phi_n(x) \right\rangle = 0 \tag{20.4}$$

Para el espacio de funciones e^{inx} la normalización es trivialmente $W = 1$. Otra manera de explicitar con palabras (del álgebra) la condición (20.4) es decir que respecto del subespacio dado por $\{\phi_n\}$ el residuo R^N es ortogonal.

De esta condición se obtiene un sistema de $N + 1$ ecuaciones para las $N + 1$ incógnitas c_n. Notemos que no hemos introducido hasta el momento ninguna grilla. Hagámoslo a continuación.

Introduzcamos una grilla de puntos

$$x_0, x_1, ..., x_N$$

que serán $N + 1$ puntos en el intervalo $[a, b]$. Asociemos valores $U_0, U_1, ..., U_N$ a la función incógnita en los puntos de grilla. Si pedimos que

$$R^N(x_j) = 0 \qquad j = 0, 1, 2, ..., N \tag{20.5}$$

obtenemos $N + 1$ ecuaciones para las $N + 1$ incógnitas representadas por los c_n. Ahora

$$U^N(x_j) = \sum_{n=0}^{N} c_n \phi_n(x_j)$$

[1]Esta sigla viene en efecto de *Finite Element Method*.

y esto implica que, si la base es de funciones trigonométricas, las incógnitas c_n serán la transformada discreta de Fourier (DFT) del conjunto $U^N(x_j)$ y de igual modo, a menos de un factor, estas últimas son la transformada discreta de Fourier de las primeras. Es decir que existe una transformación que me lleva de un conjunto al otro. Pictóricamente

$$\{c_n\} \Longleftrightarrow \{U_j^N\}$$

y podemos plantear el problema en cualquier conjunto de incógnitas. Esto es, en breve, el método de colocación o pseudoespectral.

La elección de la base se hace teniendo en cuenta quién es $\mathscr{L}$; si hay derivadas la base tiene que ser derivable fácilmente. Puede darse el caso, incluso, de que la derivada pueda escribirse en función de la misma base, es decir

$$\frac{d}{dx}\left(\sum_{n=0}^{N} c_n\phi_n(x)\right) = \sum_{n=0}^{N} c_n \frac{d}{dx}\phi_n(x) = \sum_{n=0}^{N} d_n\phi_n(x),$$

en cuyo caso existiría alguna relación de recurrencia entre derivada y función,

$$\frac{d\phi_n}{dx} \propto \phi_n$$

La idea saliente aquí es que en este caso la derivada de la función en un punto depende del valor de la función en **todos** los puntos. Esta es una característica notable de los métodos espectrales: la derivada necesita el valor de la función en todos los puntos del dominio. Luego, la transformación

$$\{U_j^N\} \Longrightarrow \{c_n\}$$

requiere todos los elementos.

Estos métodos son entonces globales en lugar de locales como lo es diferencias finitas o los métodos de elementos finitos (de los cuales haremos una descripción somera en el capítulo siguiente).

En diferencias finitas si hacíamos una aproximación centrada teníamos un error de orden $\mathcal{O}\left(\Delta x^2\right)$. Para un método espectral tendremos, si son $N+1$ puntos, un error de orden $\mathcal{O}\left(\Delta x^N\right)$. Si considero un paso de discretización constante $\Delta x \propto 1/N$ se tendrá un $\mathcal{O}\left(1/N^2\right)$ para un método de diferencias finitas pero $\mathcal{O}\left(1/N^N\right)$ para un método pseudoespectral.

El error se hace tan pequeño que podemos decir[2] que la precisión es infinita. La comparación entre métodos depende, igualmente, de varios factores. Entre ellos la necesidad en pseudoespectral de transformadas de Fourier rápidas. Digamos que para problemas con condiciones de contorno periódicas y fluidos bajo

[2] Y a los acerrimos defensores de estos métodos les gusta decir.

turbulencia (con número de Reynolds $R \gg 1$) estos métodos funcionan muy bien.

Un par de referencias de las cuales pueden leerse estos temas son

- Fornberg, *A Practical Guide to Pseudoespectral Methods.*
- Boyd, *Chebyshev and Fourier Spectral Methods.*

§ 21. Método de Galerkin. Procedimiento

El problema se puede escribir

$$\mathscr{L}(U) = f(x) \qquad a \leq x \leq b$$

y quiero la solución

$$U(x) = U^N(x) = \sum_{n=0}^{N} c_n \phi_n(x)$$

siendo el residuo

$$R^N = \mathscr{L}(U) - f(x)$$

y exigiéndole ahora que

$$\left\langle \phi_n, R^N \right\rangle = 0 \qquad n = 0, 1, ..., N$$

El producto escalar para espacios de funciones se define

$$\langle f, g \rangle = \int_a^b f g^* W(x) dx.$$

Si $\mathscr{L}$ es lineal el sistema puede ponerse en forma matricial. Planteando

$$\left\langle \, \phi_n, \mathscr{L}(U^N) - f(x) \, \right\rangle = 0$$

podemos operar

$$\left\langle \, \phi_n, \mathscr{L}(U^N) \, \right\rangle = \langle \phi_n, f(x) \rangle$$

$$\left\langle \, \phi_n, \sum_{m=0}^{N} c_m \mathscr{L}(\phi_m) \, \right\rangle = \langle \phi_n, f(x) \rangle$$

$$\sum_{m=0}^{N} c_m \left\langle \, \phi_n, \mathscr{L}(\phi_m) \, \right\rangle = \langle \phi_n, f(x) \rangle$$

$$\sum_{m=0}^{N} c_m P_{nm} = f_n$$

o matricialmente

$$P\,\mathbf{c} = \mathbf{f}$$

de donde pueden extraerse las incógnitas invirtiendo la matriz P.

$$\mathbf{c} = P^{-1}\,\mathbf{f}$$

EJEMPLO 7.1. Ejemplo método de Galerkin

Consideremos la ecuación

$$\frac{\partial^2 U}{\partial x^2} - \frac{U}{2} = -\frac{3}{2}\cos(x) - \frac{9}{2}\cos(2x),$$

de modo que

$$\mathscr{L} \equiv \frac{\partial^2}{\partial x^2} - \frac{1}{2}$$

y solicitamos además

$$U(x) = U(x + 2\pi).$$

Esta ecuación tiene solución exacta analítica

$$u(x) = \cos(x) + \cos(2x),$$

pero supongamos que expandimos *a lo* Galerkin

$$U^{(2)}(x) = C_0 + C_1\cos(x) + C_2\cos(2x),$$

siendo la base

$$\phi_n = \cos(nx).$$

Entonces

$$U^N(x) = \sum_{n=0}^{N} C_n\cos(nx) \qquad N = 2$$

y hay que evaluar

$$\left\langle \phi_m, \mathscr{L}(U^N) \right\rangle = \langle \phi_m, f(x)\rangle \qquad m = 0, 1, 2.$$

Primeramente es

$$\langle \phi_m, f(x)\rangle = \int_0^{2\pi} \cos(mx)\left(-\frac{3}{2}\cos(x) - \frac{9}{2}\cos(2x)\right) dx$$

y también

$$\frac{\partial^2 U^N}{\partial x^2} = -C_1\cos(x) - 4C_2\cos(2x)$$

$$\left\langle \phi_m, \mathscr{L}(U^N) \right\rangle = \int_0^{2\pi} \cos(mx)\Big[-C_1\cos(x) - 4C_2\cos(2x) - \frac{1}{2}(C_0 + C_1\cos(x) + C_2\cos(2x))\Big]\, dx,$$

y recordamos que

$$\int_0^{2\pi} \cos(mx)\cos(nx)dx = \begin{cases} \pi\delta_{mn} \\ 2\pi \quad \text{si } m = n = 0 \end{cases}$$

de modo que las integrales resultan

$$\langle \phi_0, f(x)\rangle = 0 \quad \langle \phi_1, f(x)\rangle = -\frac{3}{2}\pi \quad \langle \phi_2, f(x)\rangle = -\frac{3}{2}\pi$$

$$\left\langle \phi_0, \mathscr{L}(U^N) \right\rangle = -\pi C_0 \qquad \left\langle \phi_1, \mathscr{L}(U^N) \right\rangle = -\frac{3}{2}\pi C_1$$
$$\left\langle \phi_2, \mathscr{L}(U^N) \right\rangle = -\frac{3}{2}\pi C_2.$$

Juntando todo se ve que la solución es

$$U = \cos(x) + \cos(2x)$$

que coincide con la analítica.

EJEMPLO 7.2. Segundo ejemplo método de Galerkin
Intentaremos resolver

$$\frac{d^2U}{dx^2} + (\cos(x) + \cos(x)^2)U = \mathrm{e}^{-1+\cos(x)}$$

en

$$0 \le x \le 2\pi$$

sujeta a condiciones de borde periódicas

$$U(x) = U(x + 2\pi)$$

siendo

$$\mathscr{L} \equiv \left(\frac{d^2}{dx^2} + [\cos(x) + cos^2(x)] \right).$$

Propongo para la solución un desarrollo que extenderé hasta $n = 3$,

$$U(x) = U^3(x) = c_0 + c_1\cos(x) + c_2\cos(2x) + c_3\cos(3x)$$

siendo la solución exacta

$$\mathrm{e}^{-1+\cos(x)}.$$

Si hacemos las cuentas

$$\frac{d^2U^{(3)}}{dx^2}(x) = -c_1\cos(x) - 4c_2\cos(2x) - 9c_3\cos(3x)$$

y las integrales

$$\int_0^{2\pi} \cos(mx)\left[\frac{d^2U^{(3)}}{dx^2}(x) + (\cos(x) + \cos(x)^2)U^{(3)}\right]dx$$

$$\int_0^{2\pi} \cos(mx)\cos(nx)dx = \begin{cases} \pi\delta_{mn} \\ 2\pi \quad \text{si} \quad m = n = 0 \end{cases}$$

$$\int_0^{2\pi} \cos(mx)\cos(x)^2 dx = \begin{cases} \pi\delta_{m2} \\ \pi \quad \text{si} \quad m = 0 \end{cases}$$

$$\int_0^{2\pi} \cos(mx)\mathrm{e}^{-1+\cos(x)}dx = \begin{cases} \mathrm{e}^{-1}\mathfrak{I}_0(1) \quad m = 0 \\ 2\mathrm{e}^{-1}\mathfrak{I}_m(1) \quad m \neq 0 \end{cases}$$

donde $\mathfrak{I}_\nu$ son las funciones de Bessel. Todo esto desemboca en el siguiente sistema

$$\begin{pmatrix} 1/2 & 1/2 & 1/4 & 0 \\ 1 & -1/4 & 1/2 & 1/4 \\ 1/2 & 1/2 & -7/2 & 1/2 \\ 0 & 1/4 & 1/2 & -17/2 \end{pmatrix} \begin{pmatrix} c_0 \\ c_1 \\ c_2 \\ c_3 \end{pmatrix} = \begin{pmatrix} f_0 \\ f_1 \\ f_2 \\ f_3 \end{pmatrix}$$

§ 22. Problemas con dependencia temporal

Sea

$$U = U(x,t)$$

y el problema

$$\frac{\partial U}{\partial t} = \mathscr{L}(U) + f(x,t)$$

siendo esta la forma general del problema PDE planteado en un dominio con condiciones de contorno dadas por

$$U(x, t=0) = g(x)$$

$$U(x=a) = U_a \quad U(x=b) = U_b \quad \text{con} \;\; a \leq x \leq b,$$

las cuales serán datos.

El método de Galerkin propondrá ahora una expansión

$$U(x,t) = U(x,t)^N = \sum_{n=0}^{N} C_n(t)\phi_n(x)$$

y definimos nuevamente al residuo como

$$R^N(x) \equiv \frac{\partial U^N}{\partial t} - \mathscr{L}(U^N) - f(x,t)$$

pidiéndole que

$$\left\langle \phi_m, R^N \right\rangle = 0 \qquad m = 0, 1, ..., N$$

y hacemos la cuenta

$$\sum_{n=0}^{N} \left[\frac{dC_n}{dt} \langle \phi_m, \phi_n \rangle \right] - \left\langle \phi_m, \mathscr{L}(U^N) \right\rangle = \langle \phi_m, f(x,t) \rangle$$

resultando de aquí un sistema de $N+1$ ecuaciones diferenciales ordinarias (ODE) acopladas para los $C_n(t)$ con $n = 0, 1, ..., N$.

Si quisiera resolver esto debo integrar en el tiempo

$$\frac{dC_n}{dt} = F(C_0, C_1, ..., C_N) \qquad \text{para} \quad n = 0, 1, ..., N$$

sujeto a las condiciones iniciales $C_n(t=0)$ que deben extraerse desde los datos

$$g(x) = U(x, t=0).$$

Si $\mathscr{L}$ es lineal

$$\sum_{n=0}^{N}\left[\frac{dC_n}{dt}\langle\phi_m,\phi_n\rangle - C_n\langle\phi_m,\mathscr{L}(\phi_n)\rangle\right] = \langle\phi_m, f(x,t)\rangle$$

que matricialmente puede escribirse

$$H\frac{d\mathbf{c}}{dt} - P\,\mathbf{c} = \mathbf{f}$$

extrayendo las condiciones iniciales desde

$$H\mathbf{c}(0) = \mathbf{g}.$$

EJEMPLO 7.3. Ecuación de difusión

Como ejemplo podemos ver la ecuación de difusión

$$\frac{\partial U}{\partial t} = \frac{\partial^2 U}{\partial x^2} \qquad 0 \le x \le 2\pi$$

sujeta a las condiciones

$$U(x,0) = g(x)$$
$$U(x=0,t) = U(x=\pi,t) = 0.$$

Usando la base

$$\phi_n = \operatorname{sen}(nx)$$

tendremos

$$U^N(x,t) = \sum_{n=1}^{N} C_n(t)\operatorname{sen}(nx)$$

y consecuentemente

$$\frac{\partial U^N}{\partial t} = \sum_{n=1}^{N}\frac{\partial C_n(t)}{\partial t}\operatorname{sen}(nx)$$

$$\frac{\partial^2 U^N}{\partial x^2} = \mathscr{L}(U^N) = \sum_{n=1}^{N} C_n(t)\mathscr{L}(\operatorname{sen}(nx)) = -\sum_{n=1}^{N} C_n(t)n^2\operatorname{sen}(nx)$$

y como los productos escalares son

$$\langle\phi_m,\phi_n\rangle = \int_0^{2\pi}\operatorname{sen}(mx)\operatorname{sen}(nx)dx = \frac{\pi}{2}\delta_{mn}$$

resulta

$$\sum_{n=1}^{N}\left[\frac{\partial C_n}{\partial t}\left(\frac{\pi}{2}\delta_{mn}\right) + n^2\frac{\pi}{2}\delta_{mn}C_n\right] = 0 \qquad m = 1,2,...,N$$

y vemos que el sistema estará desacoplado por las δ_{mn} de manera que arribamos a

$$\frac{\partial C_m}{\partial t} + m^2 C_m = 0 \qquad m = 1,...,N$$

con solución

$$C_m(t) = C_m(0)\,\mathrm{e}^{-m^2 t} \qquad m = 1,...,N$$

donde $C_m(0)$ es la condición inicial que saldrá desde

$$\sum_{n=1}^{N} \frac{\pi}{2} \delta_{mn} C_m(0) = g_m$$

$$\int_0^{\pi} g(x)\ \text{sen}(mx) dx$$

$$C_m(0) = \frac{2}{\pi} \int_0^{\pi} g(x)\ \text{sen}(mx) dx$$

$$U^N(x,t) = \sum_{n=1}^{N} C_n(0)\ \text{e}^{n^2 t}\, \text{sen}(nx)$$

La solución exacta sería la misma pero con $N \to \infty$. En este caso Galerkin da la solución exacta a menos de un truncamiento; claramente el error va como

$$\mathcal{O} \propto \text{e}^{-N^2}$$

y desde aquí vemos que el ajuste es muy bueno.

§ 23. Métodos pseudoespectrales

En este caso, a diferencia de Galerkin, necesito definir una grilla espacial. Discretizamos del siguiente modo

$$x_i = \frac{\pi i}{N+1} \quad \Delta x = \frac{\pi}{N+1} \quad x_0 = 0 \quad x_{N+1} = \pi \qquad i = 0, 1, ..., N+1$$

con las condiciones

$$U_0 = 0, \quad U_{N+1} = 0, \quad U(x,t) \equiv U_i$$

Planteo nuevamente un truncamiento de la serie

$$U^N(x,t) = \sum_{n=1}^{N} C_n(t)\ \phi_n(x)$$

siendo el conjunto de funciones $\phi_n(x) = \text{sen}(nx)$. Las incógnitas pueden despejarse en función de los $C_n(t)$ o de los puntos $U(x_i, t)$, encontrándose ambos relacionados a través de

$$U(x_i,t) = \sum_{n=1}^{N} C_n(t)\ \phi_n(x) = \sum_{n=1}^{N} C_n(t)\ \text{sen}\left(\frac{n\pi i}{N+1}\right)$$

Esto no es otra cosa que la transformada discreta de Fourier (DFT, según sus siglas en inglés), donde se ve que las $C_n(t)$ son

$$C_n(t) = \frac{2}{N+1} \sum_{n=1}^{N} U(x_i,t)\ \text{sen}\left(\frac{n\pi i}{N+1}\right) \tag{23.1}$$

siendo $2/(N+1)$ un factor de normalización. Esto vale para cada $n = 0, 1, ..., N$. La inversión de la matriz puede hacerse aplicando la identidad

$$\sum_{n=1}^{N} \operatorname{sen}\left(\frac{n\pi i}{N+1}\right) \operatorname{sen}\left(\frac{n\pi k}{N+1}\right) = \frac{N+1}{2}\delta_{ik} \tag{23.2}$$

pero podemos pasarla a una integral

$$\int_0^{\pi} \operatorname{sen}(xi) \operatorname{sen}(kx) dx = \frac{\pi}{2}\delta_{ik}$$

donde se ha considerado

$$x = \frac{n\pi}{N+1} \quad dx = \frac{\pi}{N+1}.$$

Queríamos ver que

$$R^N(x_i) = 0 \tag{23.3}$$

donde

$$R^N = \frac{\partial U^N}{\partial t} - \frac{\partial^2 U^N}{\partial x^2}$$

entonces eso lleva a que

$$R^N = \sum_{n=1}^{N} \frac{dC_n}{dt} \operatorname{sen}(nx) + \sum_{n=1}^{N} n^2 C_n \operatorname{sen}(nx) = \sum_{n=1}^{N} \left[\frac{dC_n}{dt} + n^2 C_n\right] \operatorname{sen}(nx) = 0$$

y si pedimos que sea nulo en todos los puntos de grilla como vale $\forall x_i$ debe ser nulo el corchete,

$$\frac{dC_n}{dt} + n^2 C_n = 0,$$

de manera que la solución es

$$c_n = C_n(0)\, \mathrm{e}^{-n^2 t}.$$

El método de colocación pretende trabajar con los valores de grilla. Quiero expresar la condición (23.3) en términos de U^N y no de sus coeficientes.

$$U(x_i, t) = \sum_{n=1}^{N} C_n(t)\ \operatorname{sen}\left(\frac{n\pi i}{N+1}\right)$$

siendo las C_n dadas por (23.1). Si introducimos toda esta información en una única ecuación resultará que la condición (23.3) se transforma en

$$\sum_{n=1}^{N} \frac{2}{N+1} \sum_{i=1}^{N} \left[\frac{\partial U}{\partial t}(x_i, t) + n^2 U(x_i, t)\right] \operatorname{sen}\left(\frac{n\pi i}{N+1}\right) \operatorname{sen}\left(\frac{n\pi k}{N+1}\right) = 0.$$

Trabajemos esta expresión un poco, comenzando por distribuir la suma e intercambiar las sumatorias,

$$\sum_{i=1}^{N} \frac{\partial U}{\partial t}(x_i,t)\left[\frac{2}{N+1}\sum_{n=1}^{N}\operatorname{sen}\left(\frac{n\pi i}{N+1}\right)\operatorname{sen}\left(\frac{n\pi k}{N+1}\right)\right]+ \sum_{i=1}^{N}\left[\frac{2}{N+1}\sum_{n=1}^{N} n^2\operatorname{sen}\left(\frac{n\pi i}{N+1}\right)\operatorname{sen}\left(\frac{n\pi k}{N+1}\right)\right]U(x_i,t)=0. \quad (23.4)$$

Si aplicamos la identidad (23.2) a esta expresión, y definimos

$$-D_{ki} \equiv \frac{2}{N+1}\sum_{n=1}^{N} n^2\operatorname{sen}\left(\frac{n\pi i}{N+1}\right)\operatorname{sen}\left(\frac{n\pi k}{N+1}\right)$$

entonces resulta

$$\sum_{i=1}^{N}\frac{\partial U}{\partial t}(x_i,t)\delta_{ik} - \sum_{i=1}^{N} D_{ki}U(x_i,t)=0,$$

o bien

$$\frac{\partial U}{\partial t}(x_k,t) = \sum_{i=1}^{N} D_{ki}U(x_i,t).$$

Esto es, en suma, el método de colocación. Insistir en trabajar con los valores de grilla en lugar de los coeficientes de la base.

§ 24. Ecuación de Burgers

Es la ecuación unidimensional

$$\frac{\partial U}{\partial t} + U\frac{\partial U}{\partial x} = \nu\frac{\partial^2 U}{\partial x^2} \quad (24.1)$$

donde el miembro derecho representa difusión, como hemos visto oportunamente y el segundo término del miembro izquierdo es una advección no-lineal.

Hasta el momento habíamos trabajado con advección lineal, es decir, términos de la forma

$$c\frac{\partial U}{\partial x} \quad c = \text{cte.}$$

La ecuación (24.1) no conserva la energía cuando le aplicacamos diferencias finitas.

Consideraremos condiciones de contorno periódicas y utilizaremos el método pseudoespectral.

$$0 \le x \le 2\pi \qquad U(x) = U(x+2\pi)$$

$$U(x,0) = g(x) \qquad \text{condición inicial}$$

Recordemos que en fluidos teníamos la ecuación de Navier-Stokes,

$$\frac{\partial \mathbf{U}}{\partial t} + (\mathbf{U}\cdot\nabla)\mathbf{U} = -\frac{1}{\rho}\nabla p + \nu\frac{\partial^2 \mathbf{U}}{\partial x^2},$$

en la cual las no-linealidades se originan por el término advectivo. Ese término genera además interacción entre las diversas no-linealidades. Sea, por ejemplo,

$$U(x, t=0) = \operatorname{sen}(x) + \operatorname{sen}(2x)$$

entonces

$$U\frac{\partial U}{\partial t} = (sen(x) + \operatorname{sen}(2x))(\cos(x) + 2\cos(2x))$$

$$U\frac{\partial U}{\partial t} = \operatorname{sen}(x)\cos(x) + \operatorname{sen}(2x)\cos(x) + 2\operatorname{sen}(x)\cos(2x) + 2\operatorname{sen}(2x)\cos(2x)$$

el término no-lineal genera un término en la base que no existía cuando arrancamos el problema.

La difusión de los términos afecta a los que tienen mayor frecuencia.

Plantearemos para resolver por Galerkin una suma

$$U^N(x,t) = \sum_{k=-N/2}^{N/2-1} U_k(t)\, \mathrm{e}^{ikx} \qquad k = -N/2, ..., N/2-1$$

y pediremos

$$\langle \phi_k, R^N \rangle = 0$$

siendo

$$R^N = \frac{\partial U^N}{\partial t} + U^N\frac{\partial U^N}{\partial x} - \nu\frac{\partial^2 U^N}{\partial x^2}.$$

De esta manera tenemos

$$\int_0^{2\pi} \left[\frac{\partial U^N}{\partial t} + U^N\frac{\partial U^N}{\partial x} - \nu\frac{\partial^2 U^N}{\partial x^2}\right] \mathrm{e}^{-ikx} dx = 0$$

Veamos quién es quién en esta expresión,

$$\frac{\partial U^N}{\partial t} = \sum_{k=-N/2}^{N/2-1} \frac{dU_m^N}{dt}\, \mathrm{e}^{imx}$$

$$\frac{\partial U^N}{\partial x} = \sum_{k=-N/2}^{N/2-1} im\, U_m\, \mathrm{e}^{imx}$$

$$\frac{\partial^2 U^N}{\partial t^2} = -\sum_{k=-N/2}^{N/2-1} m^2 U_m \, \mathrm{e}^{imx}$$

Entonces

$$U^N \frac{\partial U^N}{\partial x} = \left(\sum_{n=-N/2}^{N/2-1} U_n \, \mathrm{e}^{inx} \right) \left(\sum_{k=-N/2}^{N/2-1} im \, U_m \, \mathrm{e}^{imx} \right)$$

$$U^N \frac{\partial U^N}{\partial x} = \sum_{n=-N/2}^{N/2-1} \sum_{k=-N/2}^{N/2-1} im \, U_n U_m \, \mathrm{e}^{i(n+m)x}$$

y si usamos la identidad

$$\int_0^{2\pi} \mathrm{e}^{i(m-k)x} dx = 2\pi \delta_{mk}$$

se arriba fácilmente a

$$\frac{dU_k}{dt} + \nu k^2 U_k + \sum_{\substack{n,m=-N/2 \\ m+n=k}}^{N/2-1} im \, U_n U_m = 0$$

la cual es una ecuación para $U_k(t)$ con el acoplamiento dado por la $\sum_{m,n}$ de los modos.

Este sistema se puede resolver por los métodos usuales. No obstante hay que resolverlo para cada

$$k = -\frac{N}{2}, \frac{N}{2} + 1, ..., \frac{N}{2} - 1$$

y esto implica que el número de operaciones involucradas en la $\sum$ es $\mathcal{O}\left(N^2\right)$ que sería el mismo orden de hacer diferencias finitas con derivadas en todos los puntos.

Por supuesto, si la aproximación de diferencias finitas que se considera en la discretización de (24.1) contiene más puntos (otras aproximaciones para la derivada primera y la segunda) el problema se hace *cumbersome*. Para que Galerkin sea razonable N no debe ser muy grande.

Una alternativa es utilizar un método pseudoespectral o de colocación. Para ello nos definimos una grilla

$$x_j = \frac{2\pi j}{N} \quad \text{con } j = 0, 1, ..., N-1$$

y proponemos que la serie truncada se evalúe en los puntos de grilla

$$U^N(x_j,t) = \sum_{k=-N/2}^{N/2-1} U_k(t)\, \mathrm{e}^{ikx} = \sum_{k=-N/2}^{N/2-1} U_k(t)\, \mathrm{e}^{ik\frac{2\pi}{N}j}$$

donde esta última igualdad significa que U^N es una transformada discreta de Fourier y entonces

$$U_k(t) = \frac{1}{N} \sum_{j=0}^{N-1} U^N(x_j,t)\, \mathrm{e}^{-ik\frac{2\pi}{N}j},$$

relaciones que pueden escribirse matricialmente de acuerdo a

$$T_{kj} = \mathrm{e}^{ik\frac{2\pi}{N}j} \qquad T^{-1}_{kj} = \mathrm{e}^{ik\frac{2\pi}{N}j}.$$

Se puede ir de un espacio a otro con estas relaciones de transformación

$$U^N(x_j,t) \quad \Longleftrightarrow \quad U_k(t) \tag{24.2}$$

$$j = 0, 1, ..., N \qquad\qquad k = -\frac{N}{2}, ..., \frac{N}{2} - 1.$$

La idea, y justificación, de estos pasajes es que hay operaciones que se realizan más rápido (a menor costo) en un espacio que en otro.

La operación matemática de la DFT puede hacerse aún más rápido con una FFT (*Fast Fourier Transform*). La DFT implica N^2 operaciones para pasar entre espacios en la transformación (24.2) porque para cada k hay que realizar la suma que da el U_k. La operación de FFT consiste en ir subdividiendo el problema en un esquema tipo *divide &conquer*, lo cual es parte del algoritmo, y arribar así a una problema de orden $N\log(N)$. Aunque no parezca, esta diferencia es mucho ahorro.

Esquemáticamente esto opera así; comenzando con alguna condición inicial

$$U(x, t = 0) = g(x),$$

se obtienen $U(x_j, t = 0)$ y se transforman con la DFT a $U_k(0)$, calculándose las derivadas espaciales en el espacio de Fourier, lo cual es multiplicar (como en mecánica cuántica)

$$\frac{\partial}{\partial x} \longrightarrow ik$$

$$\frac{\partial^2}{\partial x^2} \longrightarrow -k^2$$

$$U^N(x,t) = \sum U_k(t)\, \mathrm{e}^{ikx}$$

$$\frac{\partial U^N}{\partial x}(x,t) = \sum ikU_k(t)\, \mathrm{e}^{ikx}$$

Luego, la transformación de (24.1) dejará

$$\frac{dU_k}{dt} + U_k(ikU_k) = -\nu k^2 U_k$$

El termino $(U\partial_x U)_k$ es equivalente a una doble suma. Si la función es

$$U(x_j) = \sum_{m=-N/2}^{N/2-1} U_m\, \mathrm{e}^{imx_j},$$

y la derivada

$$\frac{\partial U}{\partial x}(x_j) = \sum_{m=-N/2}^{N/2-1} imU_m\, \mathrm{e}^{imx_j},$$

entonces

$$w(x_j) \equiv U(x_j)\frac{\partial U}{\partial x}(x_j) = \sum_{m,n=-N/2}^{N/2-1} inU_mU_n\, \mathrm{e}^{i(m+n)x_j}.$$

Si aplicamos la DFT al $w(x_j)$ obtenemos

$$w_k = \frac{1}{N}\sum_{j=0}^{N-1} w(x_j)\, \mathrm{e}^{-ikx_j}$$

o bien

$$w_k = \frac{1}{N}\sum_{j=0}^{N-1}\sum_{m,n=-N/2}^{N/2-1} inU_mU_n\, \mathrm{e}^{i(m+n-k)x_j}$$

Tenemos una identidad que nos dice

$$\frac{1}{N}\sum_{j=0}^{N-1} \mathrm{e}^{ipx_j} = \delta(p, qN) \quad \text{con} \quad q = 0, \pm 1, \pm 2, ...$$

es decir que vale la unidad si q es múltiplo. Usándola

$$w_k = \sum_{m,n=-N/2}^{N/2-1} inU_mU_n\, \delta(p, qN)$$

$$p = m + n - q = qN = \begin{cases} 0 \\ \pm N \end{cases}$$

$$k = -\frac{N}{2}, ..., \frac{N}{2} - 1$$

Podría ser

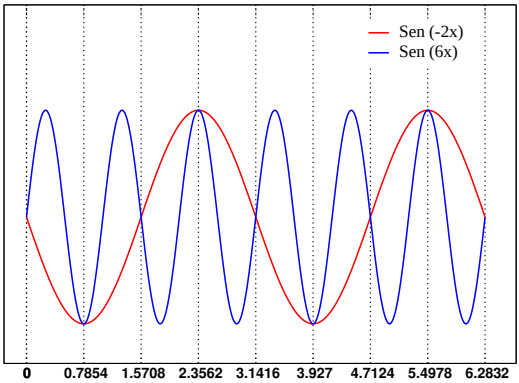
$$m = \frac{N}{2} - 1, \quad n = \frac{N}{2} - 1, \quad k = -2$$

y es $m+n-k = N$. Son tres casos en los cuales la doble suma resulta multiplicada por algo no nulo y diferente de uno.

$$w_k = \sum_{\substack{m,n=-N/2 \\ m+n=k}}^{N/2-1} inU_m U_n + \sum_{\substack{m,n=-N/2 \\ m+n=k\pm N}}^{N/2-1} inU_m U_n$$

y vemos que el primer término es el que obteníamos por Galerkin pero el segundo no es necesariamente cero; son acoples que no deberían estar. Es *aliasing*. Es la ambigüedad existente entre ondas que no se muestrean bien.

Supongamos $k = 6$, vemos que para un $k = -2$ en Fourier dos funciones, como por ejemplo sen$(6x)$ y sen$(-2x)$, coinciden totalmente si utilizo $6 - (-2) = 8$ puntos. El diagrama bajo estas líneas ilustra la situación



Si muestreamos ambas funciones solamente en los puntos indicados $(0, 0{,}7854, 1{,}5708, ...)$ no podemos distinguirlas. En resumen, tenemos que en discretos puntos ambas coinciden porque

$$k_1 - k_2 = N.$$

Para remover el aliasing puede usarse la llamada regla de los 2/3, que es en realidad un filtro. Si utilizo M modos para representar una función y lo elijo

como

$$M \geq \frac{3}{2}N$$

y hago que los modos con $|k| \geq \frac{N}{2}$ sean nulos entonces

$$|k| \geq \frac{M}{3}$$

y eso anulará el término de aliasing (mato los modos más rápidos). La suma será no nula si existen

$$m + n - k = \pm M$$

pero anulando los U_n, U_m asociados por la restricción resulta

$$-M < -\frac{3}{2}N + 1 \leq m + n - k \leq \frac{3}{2}N - 2 < M$$

$$|m + n - k| < M$$

En la práctica si trabajo con M modos filtro

$$|k| > M/3$$

en cada iteración temporal.

En resumen la ida y vuelta al espacio de Fourier genera espúreamente un término de aliasing. Se remueve filtrando pero con la consiguiente pérdida de resolución en los modos altos.

§ 25. Algunos aspectos más de la ecuación de Burgers

Habíamos escrito dicha ecuación en una dimensión como

$$\frac{\partial U}{\partial t} + U\frac{\partial U}{\partial x} = \nu\frac{\partial^2 U}{\partial x^2} \tag{25.1}$$

con $U(x) = U(x + 2\pi)$ y $0 < x < L$. Podemos definir también una energía por unidad de masa,

$$E = \frac{1}{2L}\int_0^L U^2 dx.$$

A través de la mecánica de los fluidos sabíamos que

$$E = \frac{1}{2}\int_0^L \rho u^2 dx.$$

donde

$$\rho = M/L$$

es una densidad lineal de masa.

Para examinar la variación de la energía en el tiempo podemos multiplicar U por la ecuación de Burgers (25.1) de manera que

$$\frac{1}{2}\frac{\partial U^2}{\partial t} + \frac{1}{3}\frac{\partial U^3}{\partial x} = U\nu\frac{\partial^2 U}{\partial x^2}$$

si escribimos el miembro derecho como

$$U\nu\frac{\partial^2 U}{\partial x^2} = \nu\frac{\partial}{\partial x}\left(U\frac{\partial U}{\partial x}\right) - \nu\left(\frac{\partial U}{\partial x}\right)^2$$

podemos integrar en ambos miembros obteniendo

$$\frac{dE}{dt} + \overbrace{\frac{1}{3L}U^3\Big|_0^L} = \overbrace{\frac{1}{L}U\frac{\partial U}{\partial x}\Big|_0^L} - \frac{\nu}{L}\int_0^L \left(\frac{\partial U}{\partial x}\right)^2 dx \tag{25.2}$$

donde los dos términos señalados con llaves pueden eliminarse porque tanto la función como su derivada serán periódicas. Si no lo fueran, esta condición se puede forzar pidiendo que sea nula la función en los extremos. Finalmente

$$\frac{dE}{dt} = -2\nu\Omega$$

donde se define

$$\Omega \equiv \frac{1}{2L}\int_0^L \left(\frac{\partial U}{\partial x}\right)^2 dx$$

como la *enstropía*. En fluidos tridimensionales quedarían en lugar de $\partial_x xU$ la vorticidad.

La viscosidad ν introduce disipación y pérdidas. Vemos que $E \to 0$ si $t \to \infty$ con $\nu \neq 0$. Por otro lado, si $\nu = 0$ debería conservarse la energía.

En general en diferencias finitas por más que $\nu = 0$ la discretización introduce una viscosidad numérica. El método espectral, al contrario, conserva la energía. Veamos cómo se materializa el cumplimiento del balance de energía a través de la identidad de Parseval.

Partiendo desde la expresión de la energía

$$E = \frac{1}{4\pi}\int_0^{2\pi} U^2 dx \quad \xrightarrow{\text{discreto}} \quad \frac{1}{2N}\sum_j U(x_j,t)^2,$$

si tomamos $x_j = (j-1)\frac{2\pi}{N}$ será

$$E = \frac{1}{2N}\sum_j U(x_j,t)^2 = \frac{1}{2}\sum_k |U_k(t)|^2$$

donde esta última identidad es Parseval.

Deberíamos ver que $\sum_k |U_k|^2$ cumple el balance de energia.

$$\frac{\partial U_k(t)}{\partial t} + \left[U\frac{\partial U}{\partial x}\right]_k = -\nu k^2 U_k \tag{25.3}$$

para cada valor de $k = -N/2, ..., N/2 - 1$. Escribimos

$$|U_k|^2 = U_k U_k^* = U_k U_{(-k)}$$

siendo lo último válido si U es real[3]. Es una redundancia de información, o al menos puede pensarse así.

Pero debemos derivar $|U_k|^2$ para lo cual podemos utilizar (25.3) pero conjugada.

$$\frac{\partial U_{-k(t)}}{\partial t} + \left[U\frac{\partial U}{\partial x}\right]_{-k} = -\nu k^2 U_{-k} \tag{25.4}$$

y hacemos entonces

$$U_{-k} \times (25.3) + U_k \times (25.4) = -2\nu k^2 U_{-k} U_k$$

$$U_{-k}\frac{\partial U_k}{\partial t} + U_k\frac{\partial U_{-k}}{\partial t} + U_{-k}\left[U\frac{\partial U}{\partial x}\right]_k + U_k\left[U\frac{\partial U}{\partial x}\right]_{-k} = -2\nu k^2 |U_k|^2$$

aún podemos aplicar cosmética para agrupar

$$\frac{\partial}{\partial t}\left(U_{-k}U_k\right) + U_{-k}\left[U\frac{\partial U}{\partial x}\right]_k + U_k\left[U\frac{\partial U}{\partial x}\right]_{-k} = -2\nu k^2 |U_k|^2$$

Luego, si aplicamos a cada miembro $1/2\sum_k$ será

$$\frac{\partial E}{\partial t} + \frac{1}{2}\sum_k \left(U_{-k}\left[U\frac{\partial U}{\partial x}\right]_k + U_k\left[U\frac{\partial U}{\partial x}\right]_{-k}\right) = -\nu\sum_k k^2|U_k|^2 = -2\nu\Omega$$

y necesitamos que dentro del miembro izquierdo el término en la $\sum_j$ sea nulo para que se mantenga la energía.

$$\sum_k U_{-k}\left[U\frac{\partial U}{\partial x}\right]_k = \sum_k U_{-k}\sum_{m+n=k} imU_mU_n = \sum_{m+n-k=0} imU_mU_nU_{-k}$$

y hacemos un cambio de variables tal que $-k \to k$ y entonces

$$\sum_{m+n+k=0} imU_mU_nU_k$$

[3]Si $U \in \mathbb{R}$ se da que conjugar es considerar el modo negativo.

Dada la simetria de la expresión podemos hacer sin peligro

$$\frac{1}{3}\sum_{m+n+k=0} imU_mU_nU_k + \frac{1}{3}\sum_{m+n+k=0} imU_mU_nU_k + \frac{1}{3}\sum_{m+n+k=0} imU_mU_nU_k$$

y cambiando índices

$$\frac{1}{3}\sum_{m+n+k=0} imU_mU_nU_k + \frac{1}{3}\sum_{m+n+k=0} ikU_mU_nU_k + \frac{1}{3}\sum_{m+n+k=0} inU_mU_nU_k$$

$$\frac{1}{3}\sum_{m+n+k=0} i(m+k+n)U_mU_nU_k$$

pero justamente como $m+n+k=0$ se tiene que toda la sumatoria es nula y verifica

$$\sum_k U_{-k}\left[U\frac{\partial U}{\partial x}\right]_k = 0.$$

Hemos confirmado entonces que

$$\frac{\partial E}{\partial t} = -2\nu\Omega$$

donde

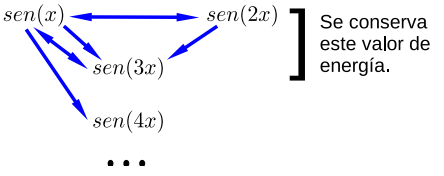
$$E = \frac{1}{2}\sum |U_k|^2$$

$$\Omega = \frac{1}{2}\sum k^2|U_k|^2.$$

Igualmente faltaria la integración temporal donde podrían introducirse pérdidas numéricas. Estas cantidades también se conservan en 2D. En tres dimensiones aparece la helicidad que también se conserva.

El kid de la demostración es la tríada de índices m, n, k que muestra el hecho de que el acople de dos modos forma un tercero tal que su número de onda está relacionado con los números de los otros dos por la suma.

La energía es un *invariante robusto* porque el método hace que los modos conserven la energía. Pese a que se generan otros modos por interacción, la energía se reparte de manera acorde (ver esquema debajo).



Recordemos no obstante que estamos con $\nu = 0$. Si hubiera viscosidad la pérdida de energía quedaría enmascarada.

§ 26. Estabilidad Burgers con RK2 en el tiempo

Veamos como aproximación burda el caso lineal (ideal, $\nu = 0$)

$$\frac{\partial U}{\partial t} + V\frac{\partial U}{\partial x} = 0 \qquad V \text{ cte.}$$

La versión espectral será

$$\frac{\partial U_k}{\partial t} + ikVU_k = 0$$

y con RK2 será

$$U_k^{n+1/2} = U_k^n - \frac{\Delta t}{2} ikVU_k^n$$

$$U_k^{n+1} = U_k^n - \Delta t\, ikVU_k^{n+1/2}$$

Planteamos cómo crece el error ε^n en el tiempo

$$U_k \to U_k + \varepsilon$$

$$\varepsilon^{n+1/2} = \varepsilon^n - \frac{\Delta t}{2} ikV\varepsilon^n$$

$$\varepsilon^{n+1} = \varepsilon^n - \Delta t ikV\varepsilon^{n+1/2}$$

entonces

$$\varepsilon^{n+1} = \varepsilon^n - \Delta t\, ikV(1 - \frac{i\Delta tkV}{2})\varepsilon^n$$

$$\varepsilon^{n+1} = \lambda\varepsilon^n$$

con

$$\lambda = 1 - i\Delta tkV - \frac{i\Delta t^2k^2V^2}{2}$$

$$|\lambda|^2 = \left(1 - \frac{i\Delta t^2k^2V^2}{2}\right)^2 + \Delta t^2k^2V^2$$

$$|\lambda|^2 = 1 + \frac{\Delta t^4k^4V^4}{4}$$

El método no es estable pero la tasa de crecimiento del error puede ser muy pequeña si el factor Δt^4 domina sobre k^4.

Si queremos ver cual es la amplificación del error a lo largo de un número muy grande de pasos

$$|\lambda| = \sqrt{|\lambda|^2} \approx 1 + \frac{\Delta t^4k^4V^4}{8}$$

y luego

$$|\lambda| = 1 + \alpha\Delta t$$

si $\alpha \equiv \Delta t^3 k^4 v^4/8$. Esto ilustra que

$$|\varepsilon^{n+1}| = |\lambda||\varepsilon^n| = (1 + \alpha\Delta t)|\varepsilon^n|$$

$$|\varepsilon^n| \approx (1 + \alpha\Delta t)^n |\varepsilon^0|$$

Si integro la ecuación hasta un tiempo $T = n\Delta t$, con $\Delta t = T/n$

$$|\varepsilon^n| \approx \left(1 + \frac{\alpha T}{n}\right)^n |\varepsilon^0|$$

si $n \to \infty$

$$|\varepsilon^n| \approx \mathrm{e}^{\alpha T} |\varepsilon^0|$$

entonces evidentemente α sintoniza el crecimiento del error. La cota máxima es en el k mayor, que era

$$k_M = \frac{N}{2}$$

entonces

$$\alpha = \frac{\Delta t^3 k^4 v^4}{168}$$

y esto es una especie de compromiso (como CFL). Si se requiere una tasa fija de error debo controlar Δt y N para que la ley se mantenga y esto lleva a que necesitemos

$$\Delta t \approx \frac{1}{N^{4/3} V^{4/3}}$$

que es un CFL un poco peor que el de diferencias finitas.

Existe un inconveniente extra y es que hemos linealizado; la V es una velocidad promedio del sistema. En realidad si la velocidad aumenta la tasa del error crece.

Si el problema es no-lineal este es un análisis somero. Pero si hubiera disipación la misma ayuda a que la solución se mantenga estable.

EJEMPLO 7.4. Ecuación de vorticidad

Consideramos un fluido plano 2D, lo cual es bastante aplicable a ciencias de la atmósfera. Las suposiciones serán

$$\mathbf{U} = (u_x, u_y)$$

fluido imcompresible

$$u_x = u_x(x, y, t) \quad u_y = u_y(x, y, t)$$

Las ecuaciones de Navier-Stokes son

$$\frac{\partial \mathbf{U}}{\partial t} + (\mathbf{U} \cdot \nabla)\mathbf{U} = -\nabla(p) + \nu \nabla^2 \mathbf{U}$$

siendo ν la viscosidad cinemática. También usaremos la vorticidad

$$\boldsymbol{\omega} = \nabla \times \mathbf{U}.$$

Por ser imcompresible podemos hallar una función corriente $\psi(x,y,t)$ *buena* tal que

$$\mathbf{U} = \left(\frac{\partial\psi}{\partial x}, -\frac{\partial\psi}{\partial y}\right)$$

con

$$\nabla\cdot\mathbf{U} = 0.$$

Además

$$\boldsymbol{\omega} = \nabla\times\mathbf{U} = \begin{pmatrix}\hat{x} & \hat{y} & \hat{z}\\ \partial_x & \partial_y & 0\\ u_x & u_y & 0\end{pmatrix} = (0,0,\partial_x u_y - \partial_y u_x)$$

$$\boldsymbol{\omega} = -\nabla^2\psi\hat{\mathbf{z}}$$

Uno podría tomar el ∇_x de Navier-Stokes para deshacernos de la presión en la descripción.

La ecuación de la vorticidad en 2D, si utilizamos el corchete de Poisson (sí, aquel de la mecánica clásica que en la mecánica cuántica pasa a ser el conmutador y tiene propiedades fascinantes) nos lleva a que resulta

$$\frac{\partial\omega_z}{\partial t} = [\psi,\omega_z] + \nu\nabla^2\omega_z$$

Este corchete es claramente el término no lineal pues

$$[\psi,\omega_z] = \frac{\partial\psi}{\partial x}\frac{\partial\omega_z}{\partial y} - \frac{\partial\psi}{\partial y}\frac{\partial\omega_z}{\partial x} = \frac{\partial}{\partial y}\left(\omega_z\frac{\partial\psi}{\partial x}\right) - \frac{\partial}{\partial x}\left(\omega_z\frac{\partial\psi}{\partial y}\right)$$

Podemos expandir Fourier el

$$\omega_z^N(x,y,t) = \sum_{\vec{k}}\omega_k\,\mathrm{e}^{i\mathbf{k}\cdot\mathbf{r}}$$

donde

$$\mathbf{k} = (k_x,k_y) \qquad \mathbf{r} = (x,y)$$

de manera que

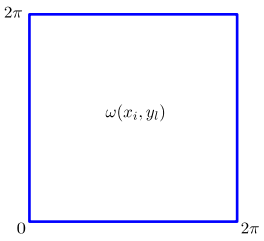
$$\omega_z^N(x,y,t) = \sum_{k_x,k_y=-N/2}^{N/2-1}\omega_k(t)\,\mathrm{e}^{ik_x x}\,\mathrm{e}^{ik_y y}$$

y podemos aplicar Galerkin, aunque en la práctica no se usa. Mejor aún, el método pseudoespectral. Discretizaremos en x,y y

$$x_j = \frac{2\pi j}{N} \qquad y_l = \frac{2\pi l}{N}$$

$$j = 0,1,...,N-1 \qquad l = 0,1,...,N-1$$

es decir que consideramos



y ω periódica. Dado el vínculo a través de la DFT resulta que

$$\omega_{\mathbf{k}}(t) = \frac{1}{N^2} \sum_{j,l=0}^{N-1} \omega^N(x_j, y_l, t)\, \mathrm{e}^{-ik_x x_j}\, \mathrm{e}^{-ik_y y_l}$$

existiendo el pasaje

$$\omega^N \longleftrightarrow \omega_k$$

a través de la DFT. Las derivadas serán

$$\frac{\partial \omega}{\partial x} \equiv ik_x \omega_{\mathbf{k}}$$

$$\frac{\partial \omega}{\partial y} \equiv ik_y \omega_{\mathbf{k}}$$

$$\nabla^2 \omega \to -k_x^2 \omega_{\mathbf{k}} - k_y^2 \omega_{\mathbf{k}} = -k^2 \omega_{\mathbf{k}}$$

donde $k^2 = k_x^2 + k_y^2$ y

$$\omega = -\nabla^2 \psi \quad \text{y} \quad \omega_k = k^2 \psi_k$$

y entonces la función corriente

$$\psi_k = \omega_k / k^2$$

Elijo condición modo inicial $\psi_{\mathbf{k}}(\mathbf{k} = 0) = 0$. En la resolución de Navier-Stokes se utiliza la inversión de ψ en función de la vorticidad.

En resumen podemos establecer el siguiente protocolo de resolución. Lo llamaremos Método de cálculo ecuación fluidos 2D

1. Dada
$$\omega(t = 0) \xrightarrow{\text{FFT}} \omega_{\mathbf{k}}(0)$$
2. Calculamos $\partial_x, \partial_y, \nabla^2$ en $\mathbf{k}$ entonces
$$ik_x \omega_k, ik_y \omega_k, -k^2 \omega_x, \psi_k = \frac{1}{k^2} \omega_k, ik_x \psi_x, ik_y \psi_y$$
3. Volvemos al espacio físico
$$\omega_k \xrightarrow{\text{FFT}^{-1}} \omega$$
$$ik_x \psi_k \xrightarrow{\text{FFT}^{-1}} \frac{\partial \psi}{\partial x}$$
$$ik_y \psi_k \xrightarrow{\text{FFT}^{-1}} \frac{\partial \psi}{\partial y}$$

$$\begin{array}{cc} \omega \dfrac{\partial \psi}{\partial x}(x_j, y_l) & \omega \dfrac{\partial \psi}{\partial y}(x_j, y_l) \\ \Big\downarrow FFT & \Big\downarrow FFT \\ \dfrac{\partial}{\partial y}\left(\left[\omega \dfrac{\partial \psi}{\partial x}\right]_{\mathbf{k}}\right) & \dfrac{\partial}{\partial x}\left(\left[\omega \dfrac{\partial \psi}{\partial y}\right]_{\mathbf{k}}\right) \\ ik_y \left(\left[\omega \dfrac{\partial \psi}{\partial x}\right]_{\mathbf{k}}\right) & ik_x \left(\left[\omega \dfrac{\partial \psi}{\partial y}\right]_{\mathbf{k}}\right) \end{array}$$

4. Integramos en el tiempo
$$\omega_{\mathbf{k}}(t + \Delta t)$$
5. Regresamos al punto 2.
6. Realizamos un *dealiasing*. En 2D es anular los modos tales que
$$|k_x| > N/3 \qquad |k_y| > N/3$$
y este filtro se aplica luego de cada paso de cuatro.

Capítulo 8

Introducción al Método de elementos finitos

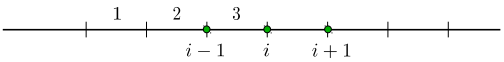
Estos métodos de elementos finitos (FEM, según sus siglas en inglés) tienen muchas aplicaciones ingenieriles. Para fluidos, no obstante, se suelen usar los métodos de volúmenes finitos, que son una especialización. Entre la bibliografía más reconocida de elementos finitos vale la pena citar a

- Hughes, *The Finite Element Method*, 1989.
- Zienkiewicz, *The Finite Element Method*, 1977.

y además, aunque más asociado a fluidos computacionales, a

- Hirsch, *Numerical Comparison of Internal and External Flows*, 1988.

Todo parte desde una PDE que quiero resolver en una región del espacio. Se subdivide ese espacio en elementos no superpuestos



y considero nodos entre las uniones de los elementos; serán los vértices de los elementos. Un nodo puede entonces pertenecer a más de un elemento. En la figura sobre estas líneas los nodos $i-1$ y i están asociados al elmento 3. Para dos dimensiones (ver Figura 26.1) suelen utilizarse triangulos y para 3D tetraedros. El mallado es una operación bastante complicada, porque implica dar una lista de nodos y elementos a los cuales pertenece cada nodo.

Luego se aproximará la solución de la PDE, supongamos

$$f(U(\mathbf{x}), t) = 0,$$

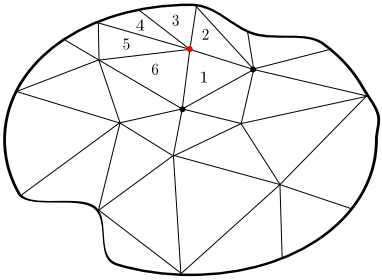


Figura 26.1 Esquema de mallado triangular en un dominio de forma general. Cada nodo tiene asociado cierta cantidad de elementos. El nodo en color rojo, pertenece a los elementos 1,2,3,4,5,6.

con funciones de interpolación o funciones de forma o base

$$U(\mathbf{x}) = U^N(\mathbf{x}) = \sum_i \alpha_i N_i(\mathbf{x}) \tag{26.1}$$

donde las N_i están asociadas a cada nodo i. Es una suma de soluciones locales. Los N_i satisfacen

$$N_i(\mathbf{x}_j) = \delta_{ij} \tag{26.2}$$

lo cual lleva a que

$$U(\mathbf{x_i}) = \alpha_i = U_i \quad \Rightarrow \quad U(\mathbf{x}) = \sum_i U_i N_i(\mathbf{x})$$

Además estas N_i suelen ser funciones a trozos donde cada uno de ellos se asocia con un elemento "e" y verifican $N_i^{(\mathrm{e})}(\mathbf{x}) = 0$ si $\mathbf{x} \notin$ elemento (e). Así, por ejemplo, en una dimensión se utilizan funciones lineales a trozos

$$N_i(x) = N_i^{(1)}(x) + N_i^{(2)}(x)$$

puesto que cada nodo abarca a lo sumo dos elementos.

Si se pide continuidad de las N_i son los llamados "elementos lagrangianos" y si se pide continuidad de las derivadas son "elementos hermitianos".

Luego reemplazamos la solución propuesta (26.1) en la ecuación diferencial y defino el residuo, por ejemplo para la ecuación de difusión, como

$$R = \frac{\partial U^N}{\partial t} - \nu \frac{\partial^2 U^N}{\partial x^2}$$

solicitando que

$$\int_\Omega R\, W \, d\Omega = 0$$

donde Ω es la región de interés. Es el método de los residuos pesados MWR (*Method of Weighted Residuals*). En FEM se suele elegir como funciones de peso a la base $N_i(x) = W$ con $i = 1, 2, ..., N$.

Es la misma idea que el método espectral pero con funciones de base muy sencillas. Es un Galerkin, si se quiere.

§ 27. FEM en una dimensión

Para una dimensión FEM no tiene ninguna ventaja sobre otros métodos. No obstante es instructivo para comprender cómo opera.

Para cada nodo i en una grilla equiespaciada (ver Figura 27.2) tendremos

$$N_i(x) = N_i^{(1)}(x) + N_i^{(2)}(x)$$

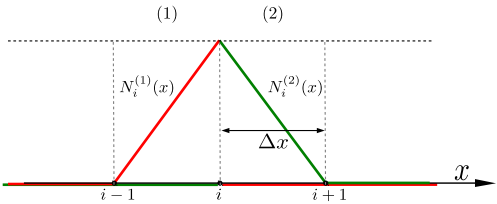


Figura 27.2 Función genérica de base $N_i(x)$

donde serán

$$N_i^{(1)} = \begin{cases} \frac{x - x_{i-1}}{\Delta x}, & x_{i-1} \le x \le x_i \\ 0, & \text{otro caso} \end{cases}$$

$$N_i^{(2)} = \begin{cases} \frac{x_{i+1} - x}{\Delta x}, & x_i \le x \le x_{i+1} \\ 0, & \text{otro caso} \end{cases}$$

y entonces $N_i(x)$ cumple la condición (26.2).

Se elige la suma porque se pide continuidad. Cada $N_i^{(e)}$ por separado cumple δ_{ij} pero no son continuas (véase en la Figura 27.2 que $N_i^{(1)}$ en color rojo y $N_i^{(2)}$ en verde son ambas discontinuas). En $x = x_i$ es $N_i = 1$, así que sólo sumo la contribución unitaria de una sola función. Es como excluir uno de los bordes.

Podemos hacer un *mapeo* para representar una función dada del siguiente modo

$$(x_{i-1}, x_i) \to (0, 1)$$

$$\xi = \frac{x - x_{i-1}}{\Delta x_i} \qquad \begin{cases} N_{i-1}^{(1)}(\xi) = 1 - \xi \\ N_i^{(1)}(\xi) = \xi \end{cases}$$

$$(x_i, x_{i+1}) \to (0,1)$$

$$\xi = \frac{x - x_i}{\Delta x_{i+1}} \qquad \begin{cases} N_i^{(2)}(\xi) = 1 - \xi \\ N_{i+1}^{(2)}(\xi) = \xi \end{cases}$$

Sobre el elemento (1)

$$U(x) = U_{i-1} N_{i-1}^{(1)} + U_i N_i^{(1)}, \quad x_{i-1} \le x \le x_i$$

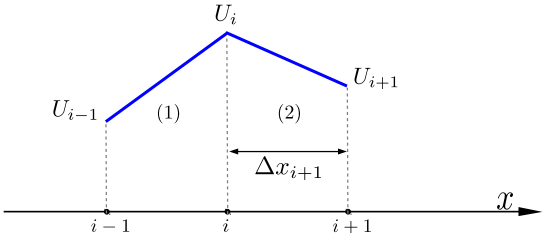
$$U(x) = U_{i-1} + \frac{x - x_{i-1}}{\Delta x_i}(U_i - U_{i-1})$$

y sobre el elemento (2)

$$U(x) = U_i N_i^{(2)} + U_{i+1} N_{i+1}^{(2)}, \quad x_i \le x \le x_{i+1}$$

$$U(x) = U_i + \frac{x - x_i}{\Delta x_{i+1}}(U_{i+1} - U_i)$$

de modo que ha resultado una función continua pero no derivable. Véase la Figura bajo estas líneas



Si se utiliza el elemento (1) la derivada es

$$\left.\frac{\partial U}{\partial x}\right|_i^{(1)} = \frac{U_i - U_{i-1}}{\Delta x_i}$$

y esto es una diferencia finita *backward*. Pero puede del mismo modo utilizarse el elemento (2) y entonces

$$\left.\frac{\partial U}{\partial x}\right|_i^{(2)} = \frac{U_{i+1} - U_i}{\Delta x_{i+1}}$$

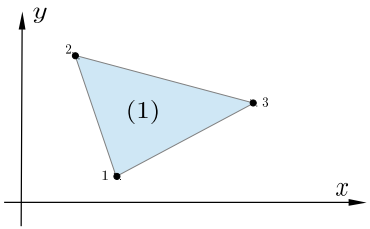
y es una diferencia finita *forward*. Para evitar la indeterminación se utiliza el promedio

$$\left(\frac{\partial U}{\partial x}\right)_i = \frac{1}{2}\left[\frac{U_i - U_{i-1}}{\Delta x_i} + \frac{U_{i+1} - U_i}{\Delta x_{i+1}}\right]$$

o bien se elige de manera consistente una u otra. Para el promedio, en el caso particular $\Delta x_{i+1} = \Delta x_i$ resulta el esquema de segundo orden. Para que sea de segundo orden en general

$$\left(\frac{\partial U}{\partial x}\right)_i = \frac{1}{\Delta x_{i+1} + \Delta x_i}\left[\Delta x_{i+1}\frac{\partial U}{\partial x}\bigg|_i^{(1)} + \Delta x_i \frac{\partial U}{\partial x}\bigg|_i^{(2)}\right]$$

En dos dimensiones suelen utilizarse triángulos



siendo la función que interpola

$$N_i^{(1)}(x,y) = \frac{(y_2 - y_3)(x - x_2)}{2A} + \frac{(y - y_2)(x_2 - x_2)}{2A}$$

donde

$$2A = (x_1 - x_2)(y_2 - y_3) + (y_1 - y_2)(x_3 - x_2).$$

Esta función es nula en los nodos 2 y 3 y vale la unidad en el nodo $i = 1$. Esto es una manera de hacer FEM. Pero hay otras.

§ 28. Formulación variacional de FEM

Veamos como ejemplo la ecuación de difusión

$$\frac{\partial}{\partial x}\left(k\frac{\partial u}{\partial x}\right) + \frac{\partial}{\partial y}\left(k\frac{\partial u}{\partial y}\right) = q.$$

donde $q(x,y)$ es la función fuente y queremos resolver en Ω siendo k un parámetro. Llamaremos $L(u)$ al operador con las derivadas. El resto será

$$R(u) \equiv L(u) - q$$

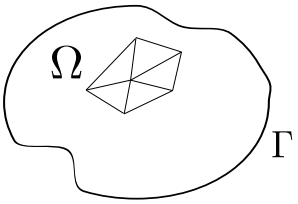
Sea

$$u^N = \sum_i u_i N_i, \tag{28.1}$$

entonces el método de los residuos pesados requerirá

$$\int_\Omega R(u^N) W d\Omega = 0$$

con $W = N_i(\mathbf{x})$ para $i = 1, 2, ..., N$ (nodos).



En el ejemplo introducido es

$$L(u) = \nabla \cdot (k \nabla u)$$

y

$$\int_\Omega (\nabla \cdot (k \nabla u) - q) W d\Omega = 0$$

$$\int_\Omega \nabla \cdot (k \nabla u) W d\Omega = \int_\Omega q W d\Omega.$$

Si aplicamos el teorema de Green aparece una integral sobre la superficie Γ que rodea al volumen Ω, es decir

$$\int_\Gamma k \frac{\partial u}{\partial n} W d\Gamma - \int_\Omega k \nabla W \cdot \nabla u d\Omega = \int_\Omega q W d\Omega. \tag{28.2}$$

Podemos utilizar $W = N_j(x)$ aprovechando la propiedad de que estas funciones N_j valen 0 si $\mathbf{x}$ está en un elemento que no contiene al nodo j.

Introduciendo estos pesos en la relación integral (28.2) resulta para el miembro derecho

$$\int_{\Omega_j} q \, N_j d\Omega,$$

donde Ω_j es una subregión. Para el miembro izquierdo la integral en el Ω tiene una sumatoria que sale fuera cuando escribimos u de acuerdo a (28.1),

$$\int_{\Gamma_j} k \frac{\partial u}{\partial n} N_j d\Gamma - \sum_i u_i \int_{\Omega_j} k \nabla N_j \cdot \nabla N_i d\Omega$$

Podemos elegir las funciones de interpolación para que $N_j(\mathbf{x}) = 0$ si se da que $x \in \Gamma$. Se suele llamar a

$$K_{ij} = -\int_{\Omega_j} k\nabla N_j \cdot \nabla N_i d\Omega$$

matriz de rigidez (*stiffness*), la cual es un factor geométrico y depende de los elementos. El vector de cargas es

$$f_j = \int_{\Omega_j} qN_j d\Omega,$$

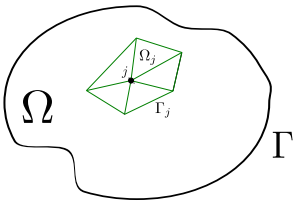
y entonces desembocamos en un sistema matricial

$$K\mathbf{u} = \mathbf{f}.$$

Escribiendo completa la relación,

$$\int_{\Gamma_j} k\frac{\partial u}{\partial n} N_j d\Gamma - \sum_i u_i \int_{\Omega_j} k\nabla N_j \cdot \nabla N_i d\Omega = \int_{\Omega_j} qN_j d\Omega$$

vemos que se integra en cada subregión j-ésima siendo la suma sobre todos los nodos i pertenecientes al elemento Ω_j. Consúltese la figura bajo estas líneas.



Veamos ahora un par de ejemplos. Para una dimensión la ecuación es

$$\frac{\partial}{\partial x}\left(k\frac{\partial u}{\partial x}\right) = q$$

y el resultado de la implementación es

$$-\sum_{i=j-1}^{j+1} u_i \int_{j-1}^{j+1} k\frac{\partial N_j}{\partial x}\frac{\partial N_i}{\partial x} dx = \int_{j-1}^{j+1} qN_j dx$$

que al especializarlo

$$-\left(u_{j-1}\left[-k\frac{\Delta x}{\Delta x^2} + 0\right] + u_j\left[k\frac{\Delta x}{\Delta x^2} + k\frac{\Delta x}{\Delta x^2}\right] + u_{j+1}\left[0 - k\frac{\Delta x}{\Delta x^2}\right]\right) = \int_{j-1}^{j} q\frac{x - x_{j-1}}{\Delta x} dx + \int_{j}^{j+1} q\frac{x_{j+1} - x}{\Delta x} dx$$

se reduce a

$$k\frac{u_{j-1}-2u_j+u_{j+1}}{\Delta x}=\frac{1}{6}(q_{j-1}+4q_j+q_{j+1})$$

donde k es constante y q es lineal y suponemos que

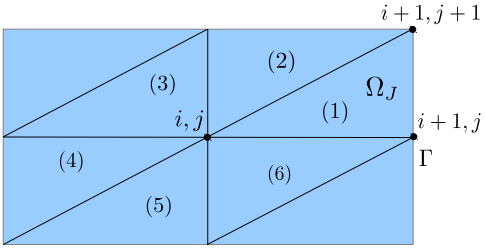
$$q=q_{j-1}+\frac{q_j-q_{j-1}}{\Delta x}(x-x_{j-1}) \qquad \text{en} \quad x_{j-1}\le x\le x_j$$

$$q=q_{j+1}-\frac{q_j-q_{j+1}}{\Delta x}(x-x_{j+1}) \qquad \text{en} \quad x_j\le x\le x_{j+1}$$

Ahora si vamos a dos dimensiones elegimos una malla triangular regular (triángulos del mismo tamaño). Queremos resolver

$$\nabla^2 u=q$$

en el Ω. Se cumple que $u=u_0$ en Γ (borde).



Suponemos $\Delta x=\Delta y=$ cte. El nodo $ij=J$ y el índice I recorre nodos

$$-\sum_I u_I\int_{\Omega_J}\left[\frac{\partial N_I}{\partial x}\frac{\partial N_J}{\partial x}+\frac{\partial N_I}{\partial y}\frac{\partial N_J}{\partial y}\right]dxdy=\int qN_I dxdy$$

donde he elegido $N_j=0$ en Γ (la frontera de Ω).

$$\Omega_J=\text{triángulos } (1)-(6)$$

$$N_J^{(1)}=N_{ij}^{(1)}=1-\frac{x-x_{ij}}{\Delta x}$$

es sólo nulo en $i+1,j$ y $i+1,j+1$ y vale 1 en i,j

$$N_{i+1,j}^{(1)}=1+\frac{x-x_{i+1,j}}{\Delta x}-\frac{y-y_{i+1,j}}{\Delta y}$$

$$N_{i+1,j+1}^{(1)}=1+\frac{y-y_{i+1,j+1}}{\Delta y}$$

Ahora, para el elemento (2)

$$N_{ij}^{(2)} = 1 - \frac{y - y_{ij}}{\Delta y}$$

vale 0 en $i, j+1$, $i+1, j+1$ y uno en el ij. Asimismo

$$N_{i,j}^{(3)} = 1 + \frac{x - x_{ij}}{\Delta x} - \frac{y - y_{ij}}{\Delta y}$$

que será uno en i, j, cero en $i, j-1$ y cero en $i, j+1$. Ahora hay que ver cuál es la contribución del elemento (1) a la matriz de stiff K_{IJ}, tal que los I refieren a los nodos del elemento 1.

$$K_{i+1,j\ ij}^{(1)} = \int_{(1)} \left[\frac{\partial N_{ij}^{(1)}}{\partial x} \frac{\partial N_{i+1,j}^{(1)}}{\partial x} + \frac{\partial N_{ij}^{(1)}}{\partial y} \frac{\partial N_{i+1,j}^{(1)}}{\partial y} \right] dxdy$$
$$= -\int_{(1)} \left(\frac{1}{\Delta x} \right) \left(\frac{1}{\Delta x} \right) dxdy = -\frac{1}{2}$$

$$K_{i+1,j+1\ ij}^{(1)} = 0$$

$$K_{ij\ ij}^{(1)} = \frac{1}{2}$$

Sumando las contribuciones de todos los elementos (1)-(6) de Ω_j

$$K_{ij\ ij} = 4$$

$$K_{i+1,j\ ij} = K_{i-1,j\ ij} = K_{i-1,j\ ij} = -1$$

$$K_{i+1,j+1\ } = 0$$

la ecuación total resulta

$$-4u_{ij} + u_{i+1,j} + u_{i-1,j} + u_{i,j+1} + u_{i,j-1}$$
$$= \frac{\Delta x^2}{12}(6q_{ij} + q_{i+1,j} + q_{i-1,j} + q_{i,j+1} + q_{i,j-1} + q_{i+1,j+1} + q_{i-1,j-1})$$

donde el miembro izquierdo es el laplaciano $\nabla^2 u$ y hemos supuesto además q lineal e interpolado linealmente q entre puntos de grilla.

§ 29. FEM para una ley de conservación

Sea la ecuación

$$\frac{\partial u}{\partial t} + \boldsymbol{\nabla} \cdot \mathbf{f} = q \tag{29.1}$$

la cual se puede aplicar a diferentes situaciones físicas como ser el modelado de fluidos (para lo cual $u \to \rho$ y $\mathbf{f} \to \rho\mathbf{U}$).

En un fluido con $q = 0$ la ecuación (29.1) representa la conservación de la masa. Si aplicamos el método de los residuos pesados en alguna región Ω estaremos pidiendo que se verifique

$$\int_\Omega RW d\Omega = 0$$

siendo

$$R(u) = \frac{\partial u}{\partial t} + \boldsymbol{\nabla} \cdot \mathbf{f} - q$$

lo que nos conduce a

$$\int_\Omega \frac{\partial u}{\partial t} W d\Omega + \int_\Omega \boldsymbol{\nabla} \cdot \mathbf{f}\, W d\Omega = \int_\Omega qW d\Omega. \tag{29.2}$$

Mediante identidades de Green transformamos la integral de la divergencia[1] en dos integrales

$$\int_\Omega \boldsymbol{\nabla} \cdot \mathbf{f} W d\Omega = \int_\Gamma \mathbf{f} \cdot d\mathbf{S} - \int_\Omega \boldsymbol{\nabla} W \cdot \mathbf{f}\, d\Omega$$

y si ahora escribimos u y f en términos de la base

$$u = \sum_i U_i N_i(\mathbf{x})$$

$$\mathbf{f} = \sum_i \mathbf{f_i} N_i(\mathbf{x}),$$

y utilizamos los pesos según Galerkin,

$$W = N_j(\mathbf{x})$$

podemos transformar (29.2) en

$$\sum_i \frac{\partial U_i}{\partial t} \int_{\Omega_j} N_i(\mathbf{x}) N_j(\mathbf{x}) d\Omega - \sum_i \mathbf{f_i} \cdot \int_{\Omega_j} \boldsymbol{\nabla} N_j(\mathbf{x}) N_i(\mathbf{x}) d\Omega = \int_{\Omega_j} qN_j d\Omega$$

[1] En efecto $W\partial_i f_i = \partial_i(Wf_i) - (\partial_i W)f$, que en forma vectorial tiene el aspecto que mostramos.

donde hemos supuesto además condiciones tipo Dirichlet en los bordes

$$N_j(\mathbf{x}) = 0 \qquad \text{en } \Gamma$$

de forma que la integral sobre Γ es nula y las integrales son en todos los elementos que pertenecen al nodo j. La sumatoria se realiza en los nodos pertenecientes a Ω_j.

Redefiniendo

$$M_{ij} \equiv \int_{\Omega_j} N_i(\mathbf{x}) N_j(\mathbf{x}) d\Omega \qquad \text{[Matriz de masa]}$$

$$K_{ij} \equiv \int_{\Omega_j} \boldsymbol{\nabla} N_j(\mathbf{x}) N_i(\mathbf{x}) d\Omega \qquad \text{[Matriz de rigidez]}$$

$$q_j \equiv \int_{\Omega_j} q N_j(\mathbf{x}) d\Omega$$

el sistema puede escribirse elegantemente

$$\sum_i \left(M_{ij} \frac{\partial u_i}{\partial t} - K_{ij} f_i \right) = q_j.$$

§ 30. Método de volúmenes finitos (FVM)

Es el método utilizado en fluidos (ingeniería) y la base del área de CFD (*Computational Fluid Dynamics*).

Se acaba de ver en la sección anterior que para una ley de conservación que emana de alguna relación como (29.1), donde q es un flujo, el planteamiento integral lleva a

$$\int_\Omega \frac{\partial u}{\partial t} W d\Omega + \int_\Omega \boldsymbol{\nabla} \cdot \mathbf{f} W d\Omega = \int_\Omega q W d\Omega$$

y para FEM expandíamos u en funciones base con W los pesos, que eran lineales.

Ahora podemos utilizar funciones más sencillas

$$W(x) = \begin{cases} 1 & \text{si } x \in \Omega_j \\ 0 & \text{si } x \notin \Omega_j \end{cases}$$

y entonces podemos escribir

$$\int_{\Omega_j} \frac{\partial u}{\partial t} d\Omega + \int_{\Omega_j} \boldsymbol{\nabla} \cdot \mathbf{f} d\Omega = \int_{\Omega_j} q d\Omega$$

de modo que empleando el teorema de Gauss para transformar la integral de la divergencia,

$$\int_{\Omega_j} \frac{\partial u}{\partial t} d\Omega + \int_{\Gamma_j} \mathbf{f} \cdot d\mathbf{S} = \int_{\Omega_j} q d\Omega$$

y definiendo

$$u_J \equiv \frac{1}{\Omega_j} \int_{\Omega_j} u d\Omega$$

$$q_J \equiv \frac{1}{\Omega_j} \int_{\Omega_j} q d\Omega$$

que son especies de promedio, arribar a

$$\frac{\partial}{\partial t}(u_J \Omega_j) + \sum_{\text{caras}} \mathbf{f} \cdot \Delta\mathbf{S} = q_J \Omega_j$$

Es una ley de conservación local, flujos evaluados en los bordes. Faltaría definir la regla para asignar valores a la función en las caras.

EJEMPLO 8.1. Difusión en 2D

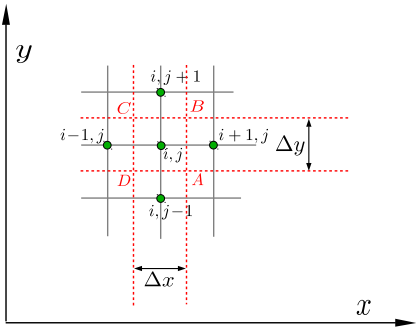
Veamos como ejemplo de aplicación la ecuación de difusión 2D,

$$\frac{\partial u}{\partial t} + \boldsymbol{\nabla} \cdot \mathbf{f} = 0 \qquad \text{con} \quad \mathbf{f} = k \boldsymbol{\nabla} u$$

que en coordenadas es una cosa del tipo

$$\frac{\partial u}{\partial t} + \frac{\partial}{\partial x}\left(k \frac{\partial u}{\partial x}\right) + \frac{\partial}{\partial y}\left(k \frac{\partial u}{\partial y}\right) = 0.$$

Utilizamos elementos cuadrados y los nodos de grilla en los centros, según se muestra a continuación



La idea es aplicar la conservación en cada celda. Considero normales externas

$$\frac{\partial}{\partial t}(u_j \Omega_j) + \sum_{\text{caras}} \mathbf{f} \cdot \Delta\mathbf{S} = q_j \Omega_j = 0$$

$$\frac{\partial}{\partial t}(u_{ij}\Delta x\Delta y) + (f_{AB} - f_{CD})\Delta y + (f_{BC} - f_{AD})\Delta x = 0$$

$$f_{AB} = k\left(k\frac{\partial u}{\partial x}\right)_{AB} = k\frac{u_{i+1,j} - u_{i,j}}{\Delta x},$$

$$f_{CD} = k\left(k\frac{\partial u}{\partial x}\right)_{CD} = k\frac{u_{i,j} - u_{i-1,j}}{\Delta x},$$

y de manera idéntica

$$f_{BC} = k\left(k\frac{\partial u}{\partial y}\right)_{BC} = k\frac{u_{i,j+1} - u_{i,j}}{\Delta x}$$

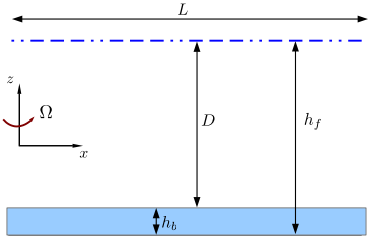
$$f_{AD} = k\left(k\frac{\partial u}{\partial y}\right)_{AD} = k\frac{u_{i,j} - u_{i,j-1}}{\Delta x}$$

suponiendo que $\Delta x = \Delta y$

$$\frac{\partial u_{i,j}}{\partial t} + \frac{k}{\Delta x^2}(u_{i+1,j} + u_{i-1,j} + u_{i,j+1} + u_{i,j-1} - 4u_{i,j}) = 0$$

y esto ha resultado igual a diferencias finitas de orden dos.

EJEMPLO 8.2. Aguas poco profundas (*shallow water*)
En aguas poco profundas suponiendo $\rho = cte$ y viscosidad nula (ideal).



Además suponemos, ver figura sobre esta línea, que $L \gg D$. Podemos plantear entonces el problema como

$$h(x,y,t) = h_f(x,y,t) - h_b(x,y,t)$$
$$\mathbf{u} = (u,v,w) = (u_x, u_y, u_z)$$
$$\omega \ll u, v \quad \text{(suposición)}$$
$$\left.\begin{aligned} u &= u(x,y,t) \\ v &= v(x,y,t) \\ h &= h(x,y,t) \end{aligned}\right\} \text{problema 2D}$$

De las ecuaciones de fluidos resulta

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} + v\frac{\partial u}{\partial y} - 2\Omega v = -g\frac{\partial h}{\partial x} \tag{30.1}$$

$$\frac{\partial v}{\partial t} + u\frac{\partial v}{\partial x} + v\frac{\partial v}{\partial y} + 2\Omega u = -g\frac{\partial h}{\partial y} \tag{30.2}$$

$$\frac{\partial h}{\partial t} + \frac{\partial hu}{\partial x} + \frac{\partial hv}{\partial y} = 0, \tag{30.3}$$

y los términos de la derecha salen de la aproximación hidrostática.

Estas ecuaciones no están escritas en forma de ley de conservación. Pero puede lograrse; multiplicamos (30.1) y (30.2) por h y utilizamos (30.3).

$$\frac{\partial}{\partial t}(hu) + \frac{\partial}{\partial x}\left(hu^2 + \frac{gh^2}{2}\right) + \frac{\partial}{\partial y}(huv) = 2\Omega hv \tag{30.4}$$

$$\frac{\partial}{\partial t}(hv) + \frac{\partial}{\partial y}\left(hv^2 + \frac{gh^2}{2}\right) + \frac{\partial}{\partial x}(huv) = -2\Omega hu \tag{30.5}$$

Puedo elegir en (30.4)

$$\bar{u} = hu, \qquad q = 2\,\Omega\, hv$$
$$\mathbf{f} = (hu^2 + \frac{gh^2}{2}, huv),$$

en (30.5)

$$\bar{u} = hv, q = -2\Omega hu$$
$$\mathbf{f} = (huv, hu^2 + \frac{gh^2}{2})$$

y en (30.3)

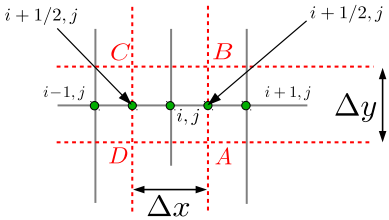
$$\bar{u} = h, q = 0$$
$$\mathbf{f} = (hu, hv)$$

y ahora finalmente el show,

$$\frac{\partial}{\partial t}(\bar{u}_j\Omega_j) + \sum_{\text{caras}} \mathbf{f}\Delta\mathbf{S} = q_j\Omega_j \tag{30.6}$$

$$\bar{u}_j \equiv \frac{1}{\Omega_j}\int_{\Omega_j} u d\Omega$$

Considerando nodos en las paredes del dominio según el esquema siguiente,



resultará, por ejemplo para (30.6),

$$\frac{\partial h_{ij}}{\partial x}\Delta x\Delta y + (hv)_{i+1/2,j}\Delta y - (hu)_{i-1/2,j}\Delta y + (hv)_{i,j+1/2}\Delta x - (hu)_{i,j-1/2}\Delta x = 0$$

$$(hu)_{i+1/2,j} = \frac{1}{4}(h_{ij} + h_{i+1,j})(u_{ij} + u_{i+1/2,j})$$

Queda un sistema para h_{ij}, u_{ij}, v_{ij}.

Índice alfabético