UNIVERSIDAD DE VALLADOLID

FACULTAD DE CIENCIAS

Métodos pseudoespectrales para la resolución numérica de ecuaciones en derivadas parciales

TRABAJO FIN DE GRADO
GRADO EN MATEMÁTICAS

Autor: Álvaro Cía Mina

Tutor: Luis María Abia Llera
Departamento de Matemática Aplicada

Índice general

Resumen

En este Trabajo Fin de Grado presentaremos los fundamentos matemáticos en que se basan los métodos espectrales y pseudoespectrales y su aplicación a la resolución numérica de ecuaciones en derivadas parciales. En la primera parte del trabajo describiremos los algoritmos de diferenciación espectral que emplean la transformada de Fourier discreta, así como las bases de la teoría de la aproximación que nos permitirán estudiar la convergencia de la diferenciación espectral. Posteriormente abordaremos la convergencia de un método Fourier-Galerkin y un método de colocación pseudoespectral para la ecuación Korteweg-de Vries (KdV). En la última parte del trabajo nos centraremos en los aspectos relacionados con la implementación de estos métodos, como es la transformada rápida de Fourier, para terminar con la programación de la resolución numérica de la ecuación KdV. El orden exponencial de convergencia de los métodos espectrales y pseudoespectrales nos permitirá hacer simulaciones precisas de la interacción entre solitones de la ecuación KdV.

Capítulo 1

Introducción

Las ecuaciones en derivadas parciales surgen en prácticamente todas las ramas de la ciencia, especialmente en la Física, en temas como la Termodinámica, la Mecánica de Fluidos, la Mecánica Cuántica, la Física de la Atmósfera o la Mecánica Ondulatoria. Rara vez se pueden dar expresiones cerradas para sus soluciones, o simplemente no es práctico trabajar con ellas. Surge, por tanto, la necesidad de la resolución numérica de estas ecuaciones.

Los tres métodos más usados para este fin son los de diferencias finitas, elementos finitos y métodos espectrales. Mientras que los métodos de elementos finitos son especialmente adecuados para resolver problemas en dominios con geometrías complicadas, los métodos espectrales presentan una mayor precisión y eficiencia principalmente en dominios sencillos. Los métodos de diferencias finitas responden bien ante dominios moderadamente complicados y ante un amplio rango de requerimientos de precisión.

Los métodos espectrales surgieron en la primera mitad del siglo XX como una herramienta para realizar cálculos en Mecánica de Fluidos y en Física de la Atmósfera. Sin embargo, el auge de estos métodos se produjo en los años 70, tras los trabajos de Gottlieb y Orszag [12]. A partir de los años 80 la investigación se centró en la extensión de estos métodos a geometrías más complicadas y ya en los años 90 se establecieron como métodos habituales en la computación. Especialmente los métodos espectrales son usados para problemas como propagación de ondas (acústicas, elásticas o electromagnéticas), astrofísica o análisis de estructuras.

Los métodos de diferencias finitas, para calcular aproximaciones a las derivadas de una función parten de un enfoque local. Por ejemplo, si queremos aproximar la derivada de $u : \mathbb{R} \to \mathbb{R}$, podemos utilizar la fórmula de diferencias centradas de segundo orden

$$u'(x) \approx \frac{u(x+h) - u(x-h)}{2h}, \quad x \in \mathbb{R},$$

que se obtiene calculando los desarrollos de Taylor de $u(x \pm h)$ cuando $h \to 0$. Este enfoque está justificado debido a la naturaleza local de la derivada. Sin embargo, los métodos espectrales parten de un enfoque global. Se aproxima la función u mediante una combinación lineal de funciones regulares ϕ_k,

$$u(x) \approx \sum_{k=0}^{N} a_k \phi_k(x),$$

por ejemplo, los polinomios de Chebyshev o los polinomios trigonométricos. Posteriormente se derivan (de forma analítica) estas funciones y se aproxima la derivada de u como

$$u'(x) \approx \sum_{k=0}^{N} a_k \phi_k'(x)\,.$$

Este planteamiento veremos que nos lleva a que para funciones analíticas el error de diferenciación decrece de forma exponencial al aumentar N, a diferencia de los métodos de diferencias finitas en los que lo hace de forma polinómica.

Para entender mejor cómo se aplican los métodos espectrales consideremos el ejemplo del problema

$$\begin{cases} u_t(x,t) - u_x(x,t) = 0\,, \quad x \in [0,2\pi]\,, 0 \le t \le T\,, \\ u(x,0) = f(x)\,, \quad x \in [0,2\pi]\,, 0 \le t \le T\,, \\ u(0,t) = u(2\pi,t) \quad 0 \le t \le T\,. \end{cases} \tag{1.1}$$

La elección del intervalo espacial, así como de las condiciones frontera, las discutiremos en capítulos posteriores del trabajo. En este punto nos limitaremos a dar un ejemplo del planteamiento de los métodos.

Utilizaremos el método espectral para la discretización espacial, donde consideraremos la aproximación por polinomios trigonométricos $\phi_k(x) = e^{ikx}$. Supondremos que N es par y para cada instante de tiempo t denotamos por $u_N(x,t)$ la solución aproximada del problema (1.1), que representaremos como

$$u_N(x,t) = \sum_{k=-N/2}^{N/2} a_k(t)\phi_k(x) = \sum_{k=-N/2}^{N/2} a_k(t) e^{ikx}\,.$$

Definimos el *residuo* asociado a u_N como

$$R_N(x,t) = \frac{\partial u_N}{\partial t}(x,t) - \frac{\partial u_N}{\partial x}(x,t) = \sum_{k=-N/2}^{N/2} \left(a_k'(t) - ika_k(t)\right) e^{ikx}\,. \tag{1.2}$$

El residuo asociado a la solución exacta es nulo. Por tanto, buscaremos minimizar de alguna forma el residuo para obtener la aproximación a la solución. Los métodos de *colocación* o *pseudoespectrales* que trataremos en posteriores capítulos, por ejemplo, exigen que el residuo (1.2) se anule sobre un conjunto de puntos $\{x_j : j = 0, 1, \ldots, N\}$, donde habitualmente $x_0 = 0$, $x_N = 2\pi$:

$$\frac{\partial u_N}{\partial t}(x_j,t) - \frac{\partial u_N}{\partial x}(x_j,t) = 0, \quad j = 0,\ldots,N\,,\ 0 \le t \le T\,. \tag{1.3}$$

Las condiciones frontera y la condición inicial necesarias para completar el sistema de ecuaciones son

$$\begin{aligned} u_N(x_0,t) &= u_N(x_N,t)\,, \quad 0 \le t \le T\,, \\ u_N(x_j,0) &= f(x_j)\,, \quad j = 0,\ldots,N\,. \end{aligned} \tag{1.4}$$

Los métodos espectrales de *Galerkin*, por el contrario, se obtienen imponiendo que

$$\int_0^{2\pi} R_N(x,t)\psi_k(x)\,dx = 0\,, \quad 0 \le t \le T\,, \quad k = -\frac{N}{2},\ldots,\frac{N}{2}, \tag{1.5}$$

donde ψ_k son las llamadas *funciones test*. Si tomamos $\psi_k(x) = \frac{1}{2\pi}e^{-ikx}$, como se satisface la condición de ortogonalidad

$$\int_0^{2\pi} \phi_k(x)\psi_{k'}(x)\,dx = \delta_{k,k'}\,,$$

las condiciones (1.5) se traducen en el sistema de ecuaciones diferenciales ordinarias

$$a_k'(t) - ika_k(t) = 0, \quad k = -\frac{N}{2}, \dots, \frac{N}{2}\,, \quad 0 \le t \le T\,. \tag{1.6}$$

Las condiciones iniciales son

$$a_k(0) = \int_0^{2\pi} f(x)\psi_k(x)\,dx\,, \quad k = -\frac{N}{2}, \dots, \frac{N}{2}\,, \tag{1.7}$$

las cuales surgen de imponer

$$\int_0^{2\pi} \left(f(x) - \sum_{k'=-N/2}^{N/2} a_{k'}(0)e^{ik'x} \right) \psi_k(x)\,dx = 0\,, \quad k = -\frac{N}{2}, \dots, \frac{N}{2}\,.$$

En los métodos de Galerkin estrictos, estas integrales (1.7) se calculan analíticamente, pero en los métodos de Galerkin-NI (Galerkin con integración numérica) se utilizan reglas de cuadratura para calcularlas numéricamente. Observamos que los métodos pseudoespectrales parten de la formulación fuerte del problema, ya que se requiere que la solución numérica satisfaga exactamente la ecuación diferencial en un conjunto de puntos. Por el contrario, los métodos de Galerkin utilizan la formulación débil, ya que requieren que el residuo satisfaga la condición (1.5).

Estos métodos se pueden aplicar también a problemas no lineales, como veremos a lo largo del trabajo con el estudio de la ecuación Korteweg-de Vries. Para esta ecuación, que presenta un término no lineal de la forma uu_x, habitualmente se prefieren los métodos pseudoespectrales a los de Galerkin puros, porque mediante las técnicas de la transformada de Fourier discreta se consigue una implementación muy eficiente del método. Por ello, inicialmente el objetivo de este trabajo incluía el análisis de la convergencia de un método pseudoespectral para la ecuación KdV mediante el método de la energía, siguiendo las técnicas abordadas en [1]. Sin embargo, durante el desarrollo de esta parte del trabajo surgió una dificultad técnica que no nos permitió avanzar por este camino y nos vimos obligados a abandonarlo. Revisando la bibliografía solo pudimos encontrar el análisis detallado de un método pseudoespectral para la ecuación KdV en el artículo de Maday y Quarteroni [16], pero al examinar la prueba que presentan, vimos que era excesivamente técnica y extensa para incluirla en el trabajo. Por ello hemos decidido hacer el estudio detallado del método espectral de Galerkin, ya que las técnicas utilizadas y el esquema de la demostración son exactamente los mismos que se utilizan para el método pseudoespectral. Además, el resultado de convergencia que se obtiene es el mismo para ambos métodos, y es que el error decrece más rápido que cualquier potencia de $1/N$ (lo que se conoce como convergencia exponencial, característica de los métodos espectrales). Para el método Galerkin la demostración es algo menos técnica, aunque no es en absoluto sencilla.

Por tanto, aunque en la última parte del trabajo implementaremos un método pseudoespectral para la resolución numérica de la ecuación KdV debido a su eficiente tratamiento del término no lineal

uu_x, en el Capítulo 4 incluiremos la demostración detallada del método espectral de Galerkin para la ecuación KdV y resumiremos los resultados análogos que se obtienen para el método espectral de colocación.

En la primera parte del trabajo introduciremos los conceptos de diferenciación espectral, que es el punto de partida de los métodos espectrales. Lo abordaremos desde dos puntos de vista, que son los métodos de diferencias finitas y la interpolación trigonométrica. En el tercer capítulo estudiaremos la convergencia de la diferenciación espectral, para lo que será necesario en primer lugar introducir los conceptos básicos de la teoría de aproximación. Como hemos mencionado antes, en el Capítulo 4 estudiaremos la convergencia de los métodos espectrales de Galerkin y espectrales de colocación (pseudoespectrales) para la ecuación KdV.

Para la implementación de los métodos espectrales nos centraremos en los métodos de colocación de Fourier, los cuales emplean los polinomios trigonométricos para la expansión de la función aproximada y toman como puntos de colocación los de la red equiespaciada en el intervalo $[0, 2\pi]$. Esta elección tiene importantes implicaciones desde el punto de vista de la implementación, ya que nos permite emplear la transformada rápida de Fourier para los cálculos. Por ello dedicamos un capítulo al desarrollo matemático de los fundamentos de la transformada rápida de Fourier, en el que describiremos el algoritmo de Cooley y Tukey y el algoritmo de Gentleman y Sande que permiten implementar de forma muy eficiente esta transformación. Al tratarse de una transformación compleja, en general, para su aplicación a la resolución de ecuaciones en derivadas parciales en las que la solución es una función real se puede establecer un algoritmo que reduce todavía más el costo computacional del cálculo de la transformada.

Para finalizar el trabajo implementaremos el método descrito comprobando numéricamente las propiedades de los métodos espectrales. Además estudiaremos el comportamiento de varios tipos de soluciones solitónicas que presenta la ecuación KdV y proporcionaremos una programación alternativa del método. Se tratará de la inclusión de un factor integrante que repercutirá de forma notable en la estabilidad del método.

Capítulo 2

Diferenciación espectral

Los métodos espectrales pueden introducirse desde varios puntos de vista. En primer lugar partiremos de los métodos de diferencias finitas, ya conocidos, y estableceremos los métodos espectrales como "límite" al incrementar la precisión. Para establecer esta relación que existe entre los dos tipos de métodos, comenzaremos recordando las ideas básicas de las diferencias finitas para aproximar las derivadas de una función conociendo sus valores sobre una red de nodos.

2.1. Matrices de derivación

Consideramos la red de N nodos equiespaciada $\{x_0, \ldots, x_{N-1}\}$, con $h = x_{j+1} - x_j$ para cada $j = 0, \ldots, N-2$. Si suponemos que se conocen los valores de una función $u : \mathbb{R} \to \mathbb{R}$ sobre la red, que los llamaremos $\{u_0, \ldots, u_{N-1}\}$, nos preguntamos cómo calcular con estos datos aproximaciones a los valores de la derivada de u sobre la red.

Una forma de obtener estas aproximaciones es el método de diferencias finitas. Por ejemplo, si llamamos w_j a la aproximación a $u'(x_j)$, la fórmula habitual de diferencias finitas en segundo orden es:

$$w_j = \frac{u_{j+1} - u_{j-1}}{2h}\,, \quad j = 1, \ldots, N-2\,, \tag{2.1}$$

la cual se puede obtener considerando los desarrollos de Taylor de $u(x_{j+1})$ y $u(x_{j-1})$ en $x = x_j$.

Esta fórmula en principio es válida para los nodos interiores, con $j = 1, \ldots, N-2$. Ahora bien, podemos suponer a partir de aquí que el problema es periódico y considerar la red formada por los puntos $\{x_j = jh : j \in \mathbb{Z}\}$, de forma que $u_j = u_{j+kN}$, para $k, j \in \mathbb{Z}$. Más adelante justificaremos esta suposición y proporcionaremos ejemplos donde es válida.

En ese caso podemos extender las fórmulas anteriores también para los extremos x_0 y x_{N-1} de la

red acotada y agruparlas en una expresión matricial de la forma siguiente:

$$\begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_{N-2} \\ w_{N-1} \end{bmatrix} = \frac{1}{h} \begin{bmatrix} 0 & \frac{1}{2} & & & -\frac{1}{2} \\ -\frac{1}{2} & 0 & \frac{1}{2} & & \\ & \ddots & \ddots & \ddots & \\ & & -\frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{2} & & & -\frac{1}{2} & 0 \end{bmatrix} \begin{bmatrix} u_0 \\ u_1 \\ \vdots \\ u_{N-2} \\ u_{N-1} \end{bmatrix}, \tag{2.2}$$

donde los elementos que no se muestran son nulos. Esta matriz cumple la propiedad llamada de *Toeplitz*, es decir, que los elementos a_{ij} solo dependen de $(i-j)$. También es *circulante*, ya que a_{ij} solo dependen de $(i-j) \pmod N$.

Obviamente si queremos calcular las aproximaciones a la derivada de una función sobre una red no necesitamos construir esta matriz, simplemente aplicamos la formula (2.1), pero introducimos la notación matricial porque nos ayudará a relacionar estos métodos con los métodos espectrales.

Otra forma de llegar a las expresiones (2.1) y (2.2) es construyendo para cada $j = 0, \dots, N-1$ el único polinomio p_j de grado menor o igual que 2 que interpola a la función u en los nodos $\{x_{j-1}, x_j, x_{j+1}\}$ y tomar $w_j = p_j'(x_j)$. Para los nodos x_0 y x_{N-1} utilizamos la periodicidad de la red.

Para orden 2 no hay gran diferencia entre ambos procedimientos. Sin embargo, el método de interpolación se puede generalizar fácilmente para órdenes mayores. Por ejemplo, si calculamos el polinomio interpolador de grado menor o igual que 4 que coincida con la función u sobre los nodos $\{x_{j-2}, x_{j-1}, x_j, x_{j+1}, x_{j+2}\}$, para $j = 0, 1, \dots, N-1$ (suponiendo también periodicidad) y hallamos su derivada, las ecuaciones que resultan las podemos expresar en forma matricial como:

$$\begin{bmatrix} w_0 \\ w_1 \\ w_2 \\ \vdots \\ \vdots \\ \vdots \\ w_{N-3} \\ w_{N-2} \\ w_{N-1} \end{bmatrix} = \frac{1}{h} \begin{bmatrix} 0 & \frac{2}{3} & -\frac{1}{12} & & & & & \frac{1}{12} & -\frac{2}{3} \\ -\frac{2}{3} & 0 & \frac{2}{3} & -\frac{1}{12} & & & & & \frac{1}{12} \\ \frac{1}{12} & -\frac{2}{3} & 0 & \frac{2}{3} & -\frac{1}{12} & & & & \\ & \ddots & \ddots & \ddots & \ddots & \ddots & & & \\ & & \frac{1}{12} & -\frac{2}{3} & 0 & \frac{2}{3} & -\frac{1}{12} & & \\ & & & \ddots & \ddots & \ddots & \ddots & \ddots & \\ & & & & \frac{1}{12} & -\frac{2}{3} & 0 & \frac{2}{3} & -\frac{1}{12} \\ -\frac{1}{12} & & & & & \frac{1}{12} & -\frac{2}{3} & 0 & \frac{2}{3} \\ \frac{2}{3} & -\frac{1}{12} & & & & & \frac{1}{12} & -\frac{2}{3} & 0 \end{bmatrix} \begin{bmatrix} u_0 \\ u_1 \\ u_2 \\ \vdots \\ \vdots \\ \vdots \\ u_{N-3} \\ u_{N-2} \\ u_{N-1} \end{bmatrix}. \tag{2.3}$$

En este caso obtenemos una matriz circulante pentadiagonal.

Podemos seguir aumentando el orden de la misma forma y obtener matrices de derivación para orden 6, 8, 10, etc. las cuales serán circulantes con un ancho de banda mayor. Este es precisamente el punto de partida de los métodos espectrales, los cuales se pueden entender como un paso al límite de este proceso para obtener una fórmula de derivación de "orden infinito".

Para precisar un poco más este aspecto, vamos a calcular explícitamente los coeficientes de la fórmula de derivación de orden par $2N$ para aproximar la derivada de la función u en un nodo.

Para que las fórmulas sean más sencillas, supondremos la red equiespaciada formada por los nodos $\{x_j = jh : j \in \mathbb{Z}\}$ sobre la que conocemos los valores de una función u y hallaremos la aproximación a $u'(0)$. Las fórmulas halladas serán válidas para cualquier nodo, sin más que cambiar los índices.

Como hemos mencionado antes, debemos calcular la derivada del polinomio interpolador de grado menor o igual que $2N$ que interpola a la función u en los nodos $\{x_{-N}, \dots, 0, \dots, x_N\}$. Escribimos este polinomio interpolador en la base de Lagrange, de forma que

$$p_N(x) = \sum_{j=-N}^{N} L_j(x) u_j \,,$$

donde los polinomios de la base de Lagrange son

$$L_j(x) = \frac{(x - x_{-N}) \cdots (x - x_{j-1})(x - x_{j+1}) \cdots (x - x_N)}{(x_j - x_{-N}) \cdots (x_j - x_{j-1})(x_j - x_{j+1}) \cdots (x_j - x_N)} = \frac{W_j(x)}{W_j(x_j)} \,.$$

Para simplificar la notación llamamos $W_j(x)$ al producto $(x - x_{-N}) \cdots (x - x_N)$ en el que se ha eliminado el término $(x - x_j)$.

De esta forma, nuestro objetivo es calcular la derivada del polinomio en $x = 0$, para lo que necesitamos calcular $W_j'(0)$. Para ello, como W_j es producto de términos lineales $x - x_j$, su derivada será la suma de todos los productos posibles resultantes de eliminar cada uno de los términos lineales. Si denotamos por $W_{j,i}(x)$ al producto $(x - x_{-N}) \cdots (x - x_N)$ en el que se ha eliminado el término $(x - x_j)$ y también el término $(x - x_i)$, para $i \neq j$, entonces

$$W_j'(x) = \sum_{i=-N,\, i\neq j}^{N} W_{j,i}(x) \,.$$

Si $j \neq 0$, entonces si evaluamos la suma anterior en $x = 0$ solo uno de los términos es no nulo: $W_{j,0}(0)$. Y teniendo en cuenta que $x_i = ih$ obtenemos el valor:

$$W_j'(0) = W_{j,0}(0) = \frac{(-1)^{N+1}(N!)^2 h^{2N-1}}{j}, \qquad j \neq 0 \,.$$

Por el contrario, si $j = 0$ obtenemos

$$W_0'(0) = \sum_{i=-N,\, i\neq 0}^{N} W_{0,i}(0) = \sum_{i=-N,\, i\neq 0}^{N} \frac{(-1)^{N+1}(N!)^2 h^{2N-1}}{i} = 0 \,.$$

Ahora, para obtener la expresión de la derivada del polinomio de Lagrange nos falta calcular $W_j(x_j)$ para $j \neq 0$. Esta expresión toma la forma del producto $h^{2N}(j + N)(j + N - 1) \cdots (j - N)$ donde se ha eliminado el término $(j - j)$. Este producto se puede expresar como

$$W_j(x_j) = h^{2N}(N + j)!(N - j)!(-1)^{N-j} \,.$$

Por tanto, la derivada del polinomio de Lagrange es

$$L_j'(0) = \begin{cases} 0, & \text{si } j = 0 \,, \\ \dfrac{(-1)^{j+1}(N!)^2}{jh(N + j)!(N - j)!}, & \text{si } j \neq 0 \,. \end{cases} \tag{2.4}$$

Y con estos coeficientes, la derivada del polinomio interpolador, que constituye la aproximación de orden $2N$ a la derivada, será

$$w_0 = p'_N(0) = \sum_{j=-N}^{N} L'_j(0) u_j \,.$$

Ahora que tenemos las expresiones explícitas de los coeficientes en la ecuación (2.4), podemos tomar el límite $N \to \infty$ para cada j fijo. Podemos hacerlo, por ejemplo, usando la fórmula de Stirling. Con este límite obtenemos los coeficientes

$$c_j = \begin{cases} 0, & \text{si } j = 0\,, \\ \dfrac{(-1)^{j+1}}{jh}, & \text{si } j \neq 0\,, \end{cases} \tag{2.5}$$

que representan el operador de derivación formal $\mathcal{D}$ de "orden infinito" siguiente:

$$\mathcal{D} = \frac{1}{h} \begin{bmatrix} & \vdots & \\ \ddots & \frac{1}{3} & \\ \ddots & -\frac{1}{2} & \\ \ddots & 1 & \\ & 0 & \\ & -1 & \ddots \\ & \frac{1}{2} & \ddots \\ & -\frac{1}{3} & \ddots \\ & \vdots & \end{bmatrix}.$$

Esta va a ser la conexión entre los métodos de diferencias finitas y los métodos espectrales, que quedará más clara cuando tratemos los métodos espectrales desde el punto de vista de la interpolación trigonométrica y las transformadas de Fourier discretas. A continuación nos disponemos a abordar esta aproximación.

2.2. Transformada de Fourier semidiscreta

Recordamos que para funciones $u \in L^2(\mathbb{R})$, la *transformada de Fourier* de u es la función $\hat{u}$ dada por

$$\hat{u}(k) = \int_{-\infty}^{\infty} e^{-ikx} u(x) dx, \qquad k \in \mathbb{R}\,. \tag{2.6}$$

La función u se puede recuperar a partir de su transformada a través de la transformada inversa de Fourier, que para $\hat{u} \in L^2(\mathbb{R})$ es:

$$u(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{ikx} \hat{u}(k) dk, \qquad x \in \mathbb{R}\,. \tag{2.7}$$

Los detalles se pueden consultar en el libro de Rudin [21, Cap. 9].

Para el caso de la malla infinita $h\mathbb{Z}$ formada por los puntos $x_j = jh$, para $j \in \mathbb{Z}$, definiremos análogamente una transformada semidiscreta. Llamaremos $\mathbb{Z}_h$ al conjunto de las funciones definidas sobre la malla $h\mathbb{Z}$ y escribiremos dichas funciones con negrita. Por ejemplo, $\boldsymbol{v}$ hace referencia a la función definida sobre la malla $h\mathbb{Z}$ cuyas componentes son v_j, $j \in \mathbb{Z}$. Si estos valores provienen de la restricción de una función (definida en $\mathbb{R}$), a esta la denotaremos sin negrita, simplemente como v, y a su restricción a la red la llamaremos $\boldsymbol{r}_h v$. Aunque cuando no haya riesgo de confusión utilizaremos la misma notación v para referirnos tanto a la función v como a su restricción a la red.

Por analogía con el caso continuo definimos (ver [25, Cap. 2]):

Definición 2.1. Para una función $\boldsymbol{v} \in \mathbb{Z}_h$ que toma el valor v_j en el nodo x_j, su *transformada de Fourier semidiscreta* se define como

$$\hat{v}(k) = h \sum_{j=-\infty}^{\infty} e^{-ikx_j} v_j, \qquad k \in \left[-\frac{\pi}{h}, \frac{\pi}{h}\right], \tag{2.8}$$

mientras que la *transformada inversa de Fourier semidiscreta* que nos permite recuperar la función $\boldsymbol{v}$ es:

$$v_j = \frac{1}{2\pi} \int_{-\pi/h}^{\pi/h} e^{ikx_j} \hat{v}(k) dk, \qquad j \in \mathbb{Z}. \tag{2.9}$$

Cabe destacar que el intervalo de definición de k se fija como $(-\pi/h, \pi/h)$ por el efecto de *aliasing*: $e^{ik_1 x_j} = e^{ik_2 x_j}$ si $k_1 - k_2$ es múltiplo de $\frac{2\pi}{h}$.

Estas fórmulas son válidas para $\boldsymbol{v} \in l^2(\mathbb{Z})$ y $\hat{v} \in L^2[-\pi/h, \pi/h]$, aunque en esta introducción de los métodos espectrales evitaremos las discusiones técnicas. Simplemente partiremos de estos resultados conocidos para establecer más adelante la transformada discreta, que será la que usaremos en la implementación de los métodos.

Para abordar la derivación espectral necesitamos un interpolante de la función $\boldsymbol{v}$, como hemos calculado en la sección anterior. Sin embargo, observando la expresión de la transformada inversa (2.9), tenemos una expresión que nos da los valores v_j de la función $\boldsymbol{v}$. Si extendemos esta fórmula para todo $x \in \mathbb{R}$ en lugar de solo para $x_j \in h\mathbb{Z}$ conseguimos un interpolante de la función $\boldsymbol{v}$:

$$p(x) = \frac{1}{2\pi} \int_{-\pi/h}^{\pi/h} e^{ikx} \hat{v}(k) dk, \qquad x \in \mathbb{R}. \tag{2.10}$$

Se trata de una función que cumple $p(x_j) = v_j$ para cada $j \in \mathbb{R}$. Además, su transformada de Fourier (2.6) es:

$$\hat{p}(k) = \begin{cases} \hat{v}(k), & k \in \left[-\frac{\pi}{h}, \frac{\pi}{h}\right], \\ 0, & \text{en otro caso.} \end{cases}$$

Se dice que p es el *interpolante de banda limitada* de v, lo que significa que $\hat{p}$ tiene soporte compacto contenido en el intervalo $\left[-\frac{\pi}{h}, \frac{\pi}{h}\right]$.

A partir de estos resultados podemos hacer una descripción de la derivación espectral de una función $\boldsymbol{v} \in \mathbb{Z}_h$ mediante los dos pasos siguientes:

- Dada la función $\boldsymbol{v}$, determinar su interpolante de banda limitada p a través de (2.10).
- Calcular $w_j = p'(x_j)$.

Equivalentemente, en función de su transformada de Fourier, dado que $\left(e^{ikx}\right)' = ike^{ikx}$:

- Dada la función $\boldsymbol{v}$, determinar su transformada de Fourier semidiscreta $\hat{v}$ a través de (2.8).
- $\hat{w}(k) = ik\hat{v}(k)$.
- Calcular w_j a partir de $\hat{w}$ haciendo la transformada inversa (2.9).

Para completar el estudio de esta forma de calcular la aproximación a la derivada de la función, vamos a calcular explícitamente los coeficientes del operador correspondiente a este método. Para ello, en primer lugar, expresaremos la función $\boldsymbol{v}$ en una base adecuada de $\mathbb{Z}_h$, formada por las deltas centradas en cada uno de los nodos.

Sea $\boldsymbol{\delta} \in \mathbb{Z}_h$ la *delta del Kronecker*:

$$\delta_j = \begin{cases} 1, & j = 0, \\ 0, & j \neq 0. \end{cases}$$

Por la expresión (2.8), su transformada semidiscreta es

$$\hat{\delta}(k) = \begin{cases} h, & k \in \left[-\frac{\pi}{h}, \frac{\pi}{h}\right], \\ 0, & \text{en otro caso,} \end{cases}$$

y su interpolante de banda limitada es

$$p(x) = \frac{h}{2\pi} \int_{-\pi/h}^{\pi/h} e^{ikx} dk = \frac{\sin(\pi x/h)}{\pi x/h} = \text{sinc}\left(\frac{\pi x}{h}\right).$$

Ahora podemos escribir la función v como

$$v_j = \sum_{m=-\infty}^{\infty} v_m \delta_{j-m}.$$

Por la linealidad de la transformada de Fourier semidiscreta tenemos que el interpolante de banda limitada de la función v será:

$$p(x) = \sum_{m=-\infty}^{\infty} v_m \text{sinc}\left(\frac{\pi(x - x_m)}{h}\right).$$

Teniendo en cuenta que la derivada de la función sinc es:

$$\frac{d}{dx}\text{sinc}(x) = \frac{\cos x}{x} - \frac{\sin x}{x^2},$$

llegamos a que la aproximación para la derivada de $\boldsymbol{v}$ es

$$w_j = \sum_{m=-\infty}^{\infty} v_m d_{j-m}, \tag{2.11}$$

donde los coeficientes d_j toman la forma:

$$d_j = \begin{cases} 0, & j = 0\,, \\ \frac{(-1)^j}{jh}, & j \neq 0\,. \end{cases}$$

La expresión (2.11) se puede interpretar como un producto matriz por vector de dimensión infinita, donde el operador $\mathcal{D}$ es el mismo que habíamos obtenido en el apartado anterior como límite de métodos de diferencias finitas:

$$\mathcal{D} = \frac{1}{h} \begin{bmatrix} & \vdots & \\ \ddots & \frac{1}{3} & \\ \ddots & -\frac{1}{2} & \\ \ddots & 1 & \\ & 0 & \\ & -1 & \ddots \\ & \frac{1}{2} & \ddots \\ & -\frac{1}{3} & \ddots \\ & \vdots & \end{bmatrix}.$$

2.3. Transformada de Fourier discreta

Hasta ahora hemos trabajado con la red infinita $h\mathbb{Z}$. Pero en la práctica, para resolver numéricamente problemas en un ordenador, debemos trabajar con una red finita de puntos. Por esa razón estudiaremos cómo se pueden extender los resultados que acabamos de ver al caso de una red finita cuando la función es periódica.

A partir de ahora consideraremos por simplicidad en la notación la red en el intervalo $[0, 2\pi]$. Los demás intervalos se pueden conseguir a partir de factores de escala y traslaciones. Cuando hablamos de una red periódica nos referimos a que los valores en los puntos de la red vienen de la evaluación de una función periódica. Es decir, los datos satisfacen:

$$v_{j+mN} = v_j, \quad \forall j, m \in \mathbb{Z}\,. \tag{2.12}$$

Esta suposición de red periódica no es muy restrictiva. Por ejemplo, en apartados posteriores de este trabajo aplicaremos estos métodos para resolver la ecuación KdV en una dimensión espacial. Supondremos que el dato inicial es un solitón (una onda localizada en el espacio que decae de forma rápida al alejarse de su pico). La función no es periódica, pero como decae de forma rápida, podemos suponer que en los extremos del intervalo espacial la función es prácticamente nula y el problema es periódico.

También supondremos que N es par para simplificar la notación. Los resultados son análogos para N impar, pero las fórmulas son diferentes. Además, al aplicar el algoritmo FFT nos centraremos en el caso en el que N sea potencia de 2.

El espaciado de la red será $h = \frac{2\pi}{N}$. Por analogía con los casos anteriores, la forma natural de definir la transformada de Fourier y la transformada inversa será:

Definición 2.2. Para una función $\boldsymbol{v}$ definida en la red, la *transformada de Fourier discreta* es:

$$\hat{v}_k = h \sum_{j=1}^{N} e^{-ikx_j} v_j, \qquad k = \frac{-N}{2} + 1, \ldots, \frac{N}{2}\,, \tag{2.13}$$

y la *transformada inversa de Fourier discreta* es:

$$v_j = \frac{1}{2\pi} \sum_{k=-N/2+1}^{N/2} e^{ikx_j} \hat{v}_k, \qquad j = 1, \ldots, N\,. \tag{2.14}$$

La comprobación de que al efectuar la transformada inversa se recupera la función original es directa:

$$\begin{aligned} \frac{1}{2\pi} \sum_{k=-N/2+1}^{N/2} e^{ikx_j} \hat{v}_k &= \frac{1}{2\pi} \sum_{k=-N/2+1}^{N/2} e^{ikx_j} h \sum_{j'=1}^{N} e^{-ikx'_j} v_{j'} \\ &= \frac{1}{N} \sum_{j'=1}^{N} v_{j'} \sum_{k=-N/2+1}^{N/2} e^{ik(x_{j'} - x_j)}\,. \end{aligned}$$

Como $x_j - x_{j'} = (j - j')2\pi/N$, al efectuar la suma de la progresión geométrica solo proporciona un resultado distinto de cero el término $j = j'$, recuperando el valor v_j.

Para la derivación espectral seguiremos el mismo procedimiento que en la sección anterior. Necesitamos, en primer lugar, un interpolante de banda limitada, que lo podemos obtener evaluando la transformada inversa para todo x, en lugar de solo en los nodos de la red. Pero al derivar este interpolante se nos presenta una dificultad.

Pongamos como ejemplo los datos en los nodos que se corresponden con la llamada función diente de sierra, que toma alternativamente los valores ± 1:

$$v_j = (-1)^j, \qquad j = 1, \ldots, N\,.$$

Su transformada discreta según (2.13) será:

$$\hat{v}_k = h \sum_{j=1}^{N} e^{-ikx_j} v_j = h \sum_{j=1}^{N} \left(-e^{-ikh}\right)^j = h \frac{e^{-ikNh} - 1}{e^{ikh} + 1}$$

si $kh \neq \pi$. Mientras que para $kh = \pi$ vale 2π. Además, teniendo en cuenta que $Nh = 2\pi$ tenemos:

$$\hat{v}_k = \begin{cases} 2\pi & \text{si } kh = \pi\,, \\ 0 & \text{si } kh \neq \pi\,. \end{cases}$$

Al calcular el interpolador de banda limitada solo es no nulo un único término:

$$p(x) = \frac{1}{2\pi} \sum_{k=-N/2+1}^{N/2} e^{ikx} \hat{v}_k = e^{iNx/2}\,,$$

y su derivada será:

$$p'(x) = \frac{iN}{2} e^{iNx/2}\,.$$

Pero si pensamos que la función diente de sierra proviene de la evaluación de la función coseno sobre los nodos de la red, cabría esperar que la derivada fuese nula en todos ellos, y no una exponencial compleja.

Esto es debido a la asimetría existente entre los nodos de los extremos del intervalo. Para solventar este problema se definen los coeficientes de Fourier de la siguiente forma:

$$\hat{v}_k = h \sum_{j=0}^{N}{}' e^{-ikx_j} v_j, \qquad k = \frac{-N}{2}, \ldots, \frac{N}{2}\,, \tag{2.15}$$

donde la prima indica que el primer y último sumando están multiplicados por $1/2$. También hemos definido $\hat{v}_{-N/2} = \hat{v}_{N/2}$. Y se sustituye la fórmula (2.14) por la siguiente:

$$v_j = \frac{1}{2\pi} \sum_{k=-N/2}^{N/2}{}' e^{ikx_j} \hat{v}_k, \qquad j = 1, \ldots, N\,. \tag{2.16}$$

Ahora estas fórmulas ya son simétricas respecto a los extremos del intervalo. Nótese que no estamos cambiando la definición, las fórmulas (2.13) y (2.14) siguen siendo válidas para calcular la transformada y la transformada inversa. Pero ahora el interpolante lo calculamos a partir de la fórmula simétrica para la transformada inversa:

$$p(x) = \frac{1}{2\pi} \sum_{k=-N/2}^{N/2}{}' e^{ikx} \hat{v}_k, \qquad x \in [0, 2\pi]\,, \tag{2.17}$$

que es un polinomio trigonométrico de grado menor o igual que $N/2$.

Tras estas consideraciones, si retomamos el ejemplo anterior de la función diente de sierra, con esta elección del interpolador de banda limitada tenemos:

$$p(x) = \frac{1}{4\pi}\left(e^{-i\frac{N}{2}x} 2\pi + e^{i\frac{N}{2}x} 2\pi\right) = \cos\left(\frac{N}{2}x\right)$$

Luego su derivada será un seno, que se anula en todos los puntos de la red, como cabría esperar.

Una vez que tenemos expresada la forma del interpolante, al igual que en las secciones anteriores podemos calcular los coeficientes de la matriz de derivación. En primer lugar, como hicimos para el caso de la red infinita, interpolamos la función delta (que ahora es periódica):

$$\delta_j = \begin{cases} 1, & j \equiv 0 \pmod{N}, \\ 0, & j \not\equiv 0 \pmod{N}, \end{cases}$$

luego tenemos que $\hat{\delta}_k = h$ para todo k.

Entonces, el interpolante de banda limitada de la función δ es:

$$p(x) = \frac{1}{2\pi} \sum_{k=-N/2}^{N/2\prime} e^{ikx} h = \frac{h}{2\pi} \left(\frac{1}{2} \sum_{k=-N/2}^{N/2-1} e^{ikx} + \frac{1}{2} \sum_{k=-N/2+1}^{N/2} e^{ikx} \right) .$$

Para agrupar los sumatorios convenientemente sacamos factor común $e^{-ix/2}$ del primer término y $e^{ix/2}$ del segundo:

$$\begin{aligned}
p(x) &= \frac{h}{2\pi} \left(\frac{1}{2} e^{-ix/2} \sum_{k=-N/2}^{N/2-1} e^{i\left(k+\frac{1}{2}\right)x} + \frac{1}{2} e^{ix/2} \sum_{k=-N/2+1}^{N/2} e^{i\left(k-\frac{1}{2}\right)x} \right) \\
&= \frac{h}{2\pi} \left(\frac{e^{-ix/2} + e^{ix/2}}{2} \right) \sum_{l=-N/2+1/2}^{N/2-1/2} e^{ilx} \\
&= \frac{h}{2\pi} \cos\left(\frac{x}{2}\right) \sum_{l=-N/2+1/2}^{N/2-1/2} e^{ilx} \\
&= \frac{h}{2\pi} \cos\left(\frac{x}{2}\right) \frac{e^{i(-N/2+1/2)x} - e^{i(N/2+1/2)x}}{1 - e^{ix}} \\
&= \frac{h}{2\pi} \cos\left(\frac{x}{2}\right) \frac{e^{-i(N/2)x} - e^{i(N/2)x}}{e^{-ix/2} - e^{ix/2}} \\
&= \frac{h}{2\pi} \cos\left(\frac{x}{2}\right) \frac{\sin\left(\frac{Nx}{2}\right)}{\sin\left(\frac{x}{2}\right)} \\
&= \frac{h}{2\pi} \frac{\sin\left(\frac{\pi x}{h}\right)}{\tan\left(\frac{x}{2}\right)} ,
\end{aligned}$$

donde se entiende que $p(2j\pi) = 1$ y $p\left((2j-1)\pi\right) = 0$, para $j \in \mathbb{Z}$.

Para calcular los coeficientes de la matriz de derivación necesitamos derivar esta expresión:

$$p'(x) = \frac{h}{2\pi} \frac{\frac{\pi}{h} \cos\left(\frac{\pi x}{h}\right) \tan\left(\frac{x}{2}\right) - \frac{1}{2} \sin\left(\frac{\pi x}{h}\right) \sec^2\left(\frac{x}{2}\right)}{\tan^2\left(\frac{x}{2}\right)} ,$$

y evaluando en los puntos $x_j = jh, j \not\equiv 0 \pmod N$ tenemos

$$p'(x_j) = \frac{h}{2\pi} \frac{\frac{\pi}{h} \cos(j\pi) \tan(jh/2) - \frac{1}{2} \sin(j\pi) \sec^2(jh/2)}{\tan^2(jh/2)} = \frac{1}{2} (-1)^j \cot\left(\frac{jh}{2}\right) , \qquad (2.18)$$

mientras que vale 0 para $j \equiv 0 \pmod N$. Ahora, teniendo en cuenta el desarrollo de la función $\boldsymbol{v}$

$$v_j = \sum_{m=1}^{N} v_m \delta_{j-m} ,$$

el interpolante de banda limitada de la función $\boldsymbol{v}$ se puede escribir como:

$$q(x) = \sum_{m=1}^{N} v_m p(x - x_m) .$$

Utilizando la expresión de la derivada (2.18) podemos escribir

$$w_j = \sum_{m=1}^{N} v_m d_{j-m}\,,$$

donde los coeficientes d_j son:

$$d_j = \begin{cases} 0, & j \equiv 0 \pmod N, \\ \frac{1}{2}(-1)^j \cot\left(\frac{jh}{2}\right), & j \not\equiv 0 \pmod N. \end{cases}$$

En este caso correspondiente a la red acotada, podemos definir la matriz de derivación espectral $\mathcal{D}_N$, de tamaño $N \times N$, a partir de los coeficientes d_j:

$$\mathcal{D}_N = \begin{bmatrix} 0 & & & & -\frac{1}{2}\cot\left(\frac{1h}{2}\right) \\ -\frac{1}{2}\cot\left(\frac{1h}{2}\right) & \ddots & & \ddots & \frac{1}{2}\cot\left(\frac{2h}{2}\right) \\ \frac{1}{2}\cot\left(\frac{2h}{2}\right) & & \ddots & & -\frac{1}{2}\cot\left(\frac{3h}{2}\right) \\ -\frac{1}{2}\cot\left(\frac{3h}{2}\right) & & & \ddots & \vdots \\ \vdots & & \ddots & & \frac{1}{2}\cot\left(\frac{1h}{2}\right) \\ \frac{1}{2}\cot\left(\frac{1h}{2}\right) & & & & 0 \end{bmatrix}$$

Análogamente podemos construir las matrices de derivación para derivadas de orden mayor si efectuamos sucesivas derivadas del interpolante de banda limitada.

Pero también podemos aplicar el mismo esquema que en la sección anterior, trabajando directamente con la transformada. Para calcular la derivada espectral m-ésima de la función $\boldsymbol{v}$, los pasos a seguir son los siguientes, si utilizamos para las transformadas las ecuaciones (2.13) y (2.14):

- Se calcula $\hat{v}$.
- Se define $\hat{w}_k = (ik)^m \hat{v}_k$ excepto $\hat{w}_{N/2} = 0$ si m es impar.
- Se calcula w_j a partir de $\hat{w}$.

Más adelante, cuando introduzcamos la transformada rápida de Fourier, apreciaremos el potencial de este método, ya que no será necesario calcular directamente el producto matriz por vector como lo hemos expresado aquí.

La justificación de definir $\hat{w}_{N/2} = 0$ si m es impar es la siguiente. La fórmula del interpolante trigonométrico (2.17) la podemos escribir, teniendo en cuenta que $\hat{v}_{N/2} = \hat{v}_{-N/2}$:

$$\begin{aligned} p(x) &= \frac{1}{2\pi} \sum_{k=-N/2+1}^{N/2-1} e^{ikx}\hat{v}_k + \frac{1}{2\pi}\hat{v}_{N/2}\left(\frac{e^{iNx/2} + e^{-iNx/2}}{2}\right) \\ &= \frac{1}{2\pi} \sum_{k=-N/2+1}^{N/2-1} e^{ikx}\hat{v}_k + \frac{1}{2\pi}\hat{v}_{N/2}\cos\left(\frac{Nx}{2}\right), \qquad x \in [0, 2\pi]\,. \end{aligned}$$

Por tanto, al calcular la derivada tenemos

$$p'(x) = \frac{1}{2\pi} \sum_{k=-N/2+1}^{N/2-1} ike^{ikx}\hat{v}_k - \frac{N}{4\pi}\hat{v}_{N/2}\sin\left(\frac{Nx}{2}\right), \qquad x \in [0, 2\pi]\,.$$

Entonces, evaluando esta expresión en los puntos de la red $x_j = jh$, el segundo sumando se anula, de forma que

$$w_j = p'(x_j) = \frac{1}{2\pi} \sum_{k=-N/2+1}^{N/2-1} ike^{ikx_j}\hat{v}_k.$$

Es decir, w_j se obtiene calculando la transformada inversa según la fórmula (2.14) en la que los coeficientes son $\hat{w}_k = ik\hat{v}_k$ excepto el término $\hat{w}_{N/2} = 0$. Y ocurre lo mismo para todas las derivadas de orden impar, ya que el seno resultante se anula. Por el contrario, para las derivadas de orden par aparece un coseno que no se anula y no surgen problemas. Aquí es importante notar que esta excepción para $\hat{w}_{N/2}$ solo es necesaria si calculamos la transformada inversa mediante la fórmula (2.14). Por el contrario, si utilizamos la fórmula (2.16) no hace falta hacer esa excepción, ya que precisamente el interpolante trigonométrico lo hemos definido a partir de esa misma fórmula, y al derivar simplemente aparecen los factores ik. Pero habitualmente los programas de cálculo de transformadas discretas utilizan la formulación (2.14), luego es importante tener en cuenta este detalle para calcular las derivadas espectrales.

Este procedimiento que acabamos de describir constituye el fundamento de los métodos de diferenciación espectral. Posteriormente veremos cómo se aplican estos métodos a la resolución numérica de ecuaciones en derivadas parciales. En particular, implementaremos los métodos pseudoespectrales que, como hemos mencionado en la introducción, son los métodos espectrales de colocación. Es decir, habitualmente se discretiza espectralmente en espacio y se impone que la solución aproximada cumpla la ecuación diferencial en un conjunto de puntos. Este procedimiento quedará más claro cuando describamos el método pseudoespectral para la ecuación KdV.

2.4. Operador de diferenciación espectral

Tras las consideraciones anteriores, podemos definir el operador de derivación espectral (a partir de este punto supondremos que las funciones implicadas son periódicas, como hemos mencionado anteriormente), de forma que, actuando sobre una función de $\mathbb{Z}_h$, nos proporciona su derivada espectral. Por tanto lo definiremos derivando la ecuación (2.17):

Definición 2.3. El *operador de diferenciación espectral* $D : \mathbb{Z}_h \mapsto \mathbb{Z}_h$ se define por sus componentes como:

$$(D\boldsymbol{v})_j = \frac{1}{2\pi} \sum_{k=-N/2}^{N/2\,\prime} (ik)e^{ikx_j}\hat{v}_k, \qquad \boldsymbol{v} \in \mathbb{Z}_h,\ j = 1, \ldots, N\,.$$

También podemos expresar la relación en términos de los coeficientes de Fourier:

$$\widehat{(D\boldsymbol{v})}_k = ik\hat{v}_k, \qquad k = \frac{-N}{2}, \ldots, \frac{N}{2}\,. \tag{2.19}$$

Demostraremos a continuación una de las propiedades del operador de diferenciación que se usan habitualmente para analizar los métodos espectrales.

Proposición 2.4. El operador de diferenciación espectral es antisimétrico, es decir, para todo $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{Z}_h$ se cumple que

$$[D\boldsymbol{A}, \boldsymbol{B}] = -[\boldsymbol{A}, D\boldsymbol{B}] \ , \tag{2.20}$$

donde $[\cdot, \cdot]$ representa el producto interno definido por $[\boldsymbol{u}, \boldsymbol{v}] = h \sum\nolimits_{j=0}^{\prime N} u_j \bar{v}_j$.

Demostración.

$$\begin{aligned}
[D\boldsymbol{A}, \boldsymbol{B}] &= h \sum_{j=0}^{N}{}' (D\boldsymbol{A})_j \bar{B}_j \\
&= h \sum_{j=0}^{N}{}' \left[\frac{1}{2\pi} \sum_{k=-N/2}^{N/2}{}' (ik) e^{ikx_j} \hat{A}_k \right] \bar{B}_j \\
&= h \sum_{j=0}^{N}{}' \left[\frac{1}{2\pi} \sum_{k=-N/2}^{N/2}{}' (ik) e^{ikx_j} \left(h \sum_{l=0}^{N}{}' e^{-ikx_l} A_l \right) \right] \bar{B}_j \, .
\end{aligned}$$

Si ahora reordenamos los sumatorios:

$$\begin{aligned}
[D\boldsymbol{A}, \boldsymbol{B}] &= h \sum_{l=0}^{N}{}' \left[\frac{1}{2\pi} \sum_{k=-N/2}^{N/2}{}' (ik) e^{-ikx_l} \left(h \sum_{j=0}^{N}{}' e^{ikx_j} \bar{B}_j \right) \right] A_l \\
&= h \sum_{l=0}^{N}{}' \left[\frac{1}{2\pi} \sum_{k=-N/2}^{N/2}{}' (ik) e^{-ikx_l} \bar{\hat{B}}_k \right] A_l \\
&= -h \sum_{l=0}^{N}{}' (\overline{D\boldsymbol{B}})_l A_l = -[\boldsymbol{A}, D\boldsymbol{B}] \ .
\end{aligned}$$

□

Capítulo 3

Aproximación espectral

En este capítulo introduciremos los resultados fundamentales sobre interpolación y aproximación por polinomios trigonométricos, los cuales nos permitirán establecer acotaciones para el error de la diferenciación espectral que hemos tratado en el capítulo anterior.

En primer lugar vamos a introducir unos resultados previos de Análisis Matemático que serán necesarios tanto para el estudio del error de la diferenciación espectral como para el posterior análisis de los métodos espectrales que incluiremos en el capítulo siguiente. Fundamentalmente seguiremos el libro de Canuto, Hussaini, Quarteroni y Zang [4].

Recordamos que la norma en el espacio $L^p(0,2\pi)$ para p entero se define para funciones $u : (0,2\pi) \to \mathbb{C}$ como

$$\|u\|_{L^p(0,2\pi)} = \left(\int_0^{2\pi} |u(x)|^p \, dx \right)^{\frac{1}{p}} . \tag{3.1}$$

Habitualmente trabajaremos en el espacio $L^2(0,2\pi)$ cuya norma denotaremos simplemente como $\|\cdot\|$ para simplificar la notación. En este espacio sabemos que la familia

$$\left\{ \phi_k(x) = \frac{1}{\sqrt{2\pi}} e^{ikx}, k \in \mathbb{Z} \right\} \tag{3.2}$$

es un sistema ortonormal y completo. Definiremos el *espacio de polinomios trigonométricos* de grado menor o igual que $N/2$, para $N \in 2\mathbb{Z}$ como

$$\mathcal{S}_N = \text{span} \left\{ \phi_k \, , \frac{N}{2} \leq k \leq \frac{N}{2} \right\} \tag{3.3}$$

y denotaremos por $P_N : L^2(0,2\pi) \to \mathcal{S}_N$ el operador de proyección sobre $\mathcal{S}_N$ con respecto al producto interno de $L^2(0,2\pi)$, $\langle f, g \rangle = \int_0^{2\pi} f(x)\overline{g(x)}dx$, el cual actúa de la siguiente forma:

$$\forall u \in L^2(0,2\pi) \, , \quad P_N u = \sum_{k=-\frac{N}{2}}^{\frac{N}{2}} \hat{u}(k)\phi_k \, ,$$

donde los coeficientes de la expansión son los *coeficientes de Fourier*

$$\hat{u}(k) = \langle u, \phi_k \rangle = \int_0^{2\pi} u(x)\overline{\phi_k(x)} \, dx \, , \quad k \in \mathbb{Z} \, . \tag{3.4}$$

Esta definición de los coeficientes de Fourier varía según el texto de referencia que se tome. A veces se añade un factor $1/(2\pi)$ y las funciones generadoras de $\mathcal{S}_N$ se toman sin el factor $1/\sqrt{2\pi}$.

Por ser P_N el operador de proyección sobre $\mathcal{S}_N$ tenemos equivalentemente que

$$\int_0^{2\pi} (P_N u - u)\,\Psi = 0\,, \quad \forall \Psi \in \mathcal{S}_N\,. \tag{3.5}$$

A partir de los coeficientes de Fourier se define la *serie de Fourier* como la expansión formal de u en términos del sistema ortonormal (3.2)

$$Su = \sum_{k=-\infty}^{\infty} \hat{u}(k)\phi_k\,. \tag{3.6}$$

Por tanto, $P_N u$ es la serie de Fourier de u truncada. La convergencia de esta serie se puede garantizar en virtud del siguiente resultado conocido (no incluimos su demostración por salirse de los objetivos de este trabajo; se puede encontrar en [8]):

Proposición 3.1. Si $u \in L^2(0, 2\pi)$, entonces su serie de Fourier converge a u en $L^2(0, 2\pi)$ y

$$\|u\|^2 = \sum_{k=-\infty}^{\infty} |\hat{u}(k)|^2\,, \quad \text{(Identidad de Parseval.)} \tag{3.7}$$

A partir de este resultado podemos deducir fácilmente una desigualdad para polinomios trigonométricos que será útil más adelante.

Proposición 3.2. La norma en $L^2(0, 2\pi)$ de la derivada de orden r (natural) de los polinomios trigonométricos se puede acotar

$$\|\phi^{(r)}\| \leq N^r \|\phi\|\,, \quad \forall \phi \in \mathcal{S}_N\,. \tag{3.8}$$

Demostración. Sea $\phi = \sum_{k=-\frac{N}{2}}^{\frac{N}{2}} c_k \phi_k$. Por la definición de los coeficientes de Fourier (3.4) y la ortogonalidad de la familia (3.2) deducimos que los únicos coeficientes de Fourier no nulos de ϕ son los correspondientes a $-N/2 \leq k \leq N/2$ y precisamente coinciden con c_k. Entonces, aplicando la identidad de Parseval llegamos a que

$$\|\phi^{(r)}\|^2 = \sum_{k=-\frac{N}{2}}^{\frac{N}{2}} k^{2r} |c_k|^2 \leq N^{2r} \sum_{k=-\frac{N}{2}}^{\frac{N}{2}} |c_k|^2 = N^{2r} \|\phi\|^2\,. \tag{3.9}$$

□

Para el análisis numérico moderno de las ecuaciones diferenciales la familia de normas naturales con las que se trabaja son las normas de Sobolev. Por ello, los resultados de aproximación que presentaremos en este capítulo, así como el análisis de los métodos espectrales que incluiremos en el capítulo siguiente harán uso de estas normas. Como referencia general para este apartado puede tomarse el libro de Adams y Fournier [3].

Definición 3.3. Los espacios de Sobolev $W^{m,r}(0,2\pi)$ se definen para r, m enteros como

$$W^{m,r}(0,2\pi) = \left\{ u \in L^r(0,2\pi) : \|u\|_{W^{m,r}(0,2\pi)} = \left(\sum_{k=0}^{m} \left\| \frac{d^k u}{dx^k} \right\|_{L^r(0,2\pi)}^r \right)^{\frac{1}{r}} < \infty \right\} \tag{3.10}$$

Para el caso en el que $r = 2$, que será el que usemos en los siguientes apartados, denotaremos el espacio como $H^m(0,2\pi) = W^{m,2}(0,2\pi)$ y la norma como $\|\cdot\|_m$:

$$\|u\|_m = \left(\sum_{k=0}^{m} \left\| \frac{d^k u}{dx^k} \right\|^2 \right)^{\frac{1}{2}} \tag{3.11}$$

Asimismo, denotaremos por $H_p^m(0,2\pi)$ el subespacio de $H^m(0,2\pi)$ formado por las funciones cuyas $m-1$ primeras derivadas son 2π-periódicas.

Como se incluye en [4, Cap. 5], las funciones del espacio $H_p^m(0,2\pi)$ son aquellas para las que se puede diferenciar término a término la serie de Fourier m veces. Por ejemplo, $H_p^1(0,2\pi)$ es el espacio de todas las funciones u para las que

$$u' = \sum_{k=-\infty}^{\infty} ik\hat{u}(k)\phi_k \quad \text{en } L^2(0,2\pi)\,.$$

Este resultado es una consecuencia directa de la conmutabilidad de los operadores P_N y d/dx en $H_p^1(0,2\pi)$, que será un resultado que usaremos más adelante:

Proposición 3.4.

$$(P_N u)' = P_N u'\,, \quad \forall u \in H_p^1(0,2\pi)\,. \tag{3.12}$$

Demostración. De la definición de los coeficientes de Fourier y efectuando una integración por partes se obtiene

$$\widehat{(u')}(k) = \langle u', \phi_k \rangle = -\langle u, \phi_k' \rangle = ik\langle u, \phi_k \rangle = ik\hat{u}(k)\,. \tag{3.13}$$

Por tanto, cada término de la suma $P_N u'$ coincide con la derivada de cada término de la suma $P_N u$. □

Otra propiedad que usaremos para el análisis del método será la *desigualdad inversa* en las normas de Sobolev para los polinomios trigonométricos, la cual es una consecuencia inmediata de la desigualdad (3.8):

$$\|\phi\|_s \leq \gamma(s) N^{s-r} \|\phi\|_r\,, \quad 0 \leq r \leq s\,,\ \phi \in \mathcal{S}_N\,, \tag{3.14}$$

donde $\gamma(s)$ es una constante que únicamente depende de s.

Por último, una desigualdad también inmediata a partir de la identidad de Parseval que usaremos en el Capítulo 4 es:

$$\|u\|_1^2 = \sum_{k=-\frac{N}{2}}^{\frac{N}{2}} (1+k^2)\,|\hat{u}(k)|^2 \leq |\hat{u}(0)|^2 + 2\left\| \frac{du}{dx} \right\|^2, \quad \forall u \in \mathcal{S}_N\,. \tag{3.15}$$

Una vez introducidos estos conceptos, incluiremos los estimativos para el error de truncación y de interpolación. No efectuaremos las demostraciones de estos estimativos (el estudio detallado se puede encontrar en el libro de Canuto, Hussaini, Quarteroni y Zang [4]) porque sería necesario un estudio en profundidad de la teoría de la aproximación que se saldría de los objetivos de este trabajo. Por el contrario, consideramos más apropiado dar más peso en la exposición al análisis del orden de convergencia de un método espectral, el cual nos supondrá un gran esfuerzo, así como a las cuestiones relativas a la implementación de los métodos espectrales como la transformada rápida, que son fundamentales para su aplicación práctica.

Proposición 3.5. Estimativo para el error de truncación. Para l, m enteros tales que $m \geq 0$ y $0 \leq l \leq m$, se cumple

$$\|u - P_N u\|_l \leq c\, N^{l-m} \|u^{(m)}\|\,, \quad \forall u \in H_p^m(0, 2\pi)\,, \tag{3.16}$$

donde, a partir de este punto del trabajo, c denotará una constante positiva independiente de N, no necesariamente la misma en distintas ecuaciones.

Para estudiar el orden de convergencia de la diferenciación espectral será necesario introducir un resultado para el orden de convergencia de la interpolación. Denotaremos por $I_N u$ el interpolante trigonométrico de una función u en los nodos $x_j = 2\pi j/N,\ j = 0, ..., N-1$.

Proposición 3.6. Estimativo para el error de interpolación. Para l, m enteros tales que $m \geq 1$ y $0 \leq l \leq m$, se cumple

$$\|u - I_N u\|_l \leq c\, N^{l-m} \|u^{(m)}\|\,, \quad \forall u \in H_p^m(0, 2\pi)\,. \tag{3.17}$$

Estos dos resultados nos muestran que asintóticamente cuando $N \to \infty$, tanto el error de truncación como el de interpolación se pueden acotar de la misma forma. Además, observamos que el orden de convergencia únicamente depende de la regularidad de la función u. Podemos afirmar que si u es suficientemente regular, entonces tanto el error de truncación como el de interpolación decrecen más rápido que cualquier potencia de h. Estos resultados nos van a permitir demostrar que la diferenciación espectral también cumple esta propiedad, como estudiamos a continuación.

3.1. Convergencia de la diferenciación espectral

Recordamos que, como tratamos en el Capítulo 2, la diferenciación espectral se basa en calcular el interpolante trigonométrico de la función sobre los nodos de la red equiespaciada para posteriormente calcular la derivada del polinomio interpolador. Esto nos va a permitir aplicar el resultado que acabamos de obtener para el error de interpolación para obtener la convergencia de la diferenciación espectral.

En este apartado denotaremos el operador de diferenciación espectral como D_N para indicar explícitamente el número de nodos de la red como subíndice, ya que nuestro objetivo es estudiar cómo se comporta el error de la diferenciación cuando $N \to \infty$.

Proposición 3.7. Se cumple la siguiente acotación para el error de la diferenciación espectral:

$$\left\|u' - D_N u\right\| \leq c\, N^{1-m} \|u^{(m)}\|\,, \quad \forall u \in H_p^m(0, 2\pi),\ m \geq 1.$$

Demostración. Por definición, $D_N u = (I_N u)'$. Además, por la definición de la norma en $H^1(0, 2\pi)$ tenemos

$$\left\|u' - D_N u\right\| = \left\|(u - I_N u)'\right\| \leq \|u - I_N u\|_1\ ,$$

y aplicando el resultado (3.17) para $l = 1$ llegamos a la desigualdad buscada. □

Podemos generalizar este resultado para derivadas de orden superior. Tengamos en cuenta que, por ejemplo, para la derivada de segundo orden:

$$u'' - D_N^2 u = u'' - D_N(D_N u) = u'' - D_N(I_N u)' = u'' - [I_N(I_N u)']'.$$

Ahora bien, como $I_N u$ es un polinomio trigonométrico, entonces $(I_N u)'$ también lo es, luego su interpolante trigonométrico coincide con él:

$$u'' - [I_N(I_N u)']' = u'' - (I_N u)''.$$

En general se cumple que $D_N^j u = (I_N u)^{(j)}$. Y de aquí podemos deducir la siguiente acotación para el error de la diferenciación, que se demuestra de manera análoga a la proposición (3.7).

Proposición 3.8. Se cumple la siguiente acotación para el error de la diferenciación espectral de orden $j \geq 1$:

$$\|u^{(j)} - D_N^j u\| \leq c\, N^{j-m} \|u^{(m)}\|\,, \quad \forall u \in H_p^m(0, 2\pi)\,,\ m \geq j.$$

Por tanto, hemos demostrado que el error de la diferenciación espectral decrece más rápido que cualquier potencia de h en la norma $L^2(0, 2\pi)$ si u es suficientemente regular. Esta va a ser la principal ventaja de los métodos espectrales y pseudoespectrales frente a los de diferencias finitas o elementos finitos cuando los apliquemos a la resolución numérica de ecuaciones en derivadas parciales, ya que para estos dos últimos el orden de convergencia es polinómico.

Capítulo 4

Análisis de los métodos espectrales para la ecuación KdV

A partir de este punto nos centraremos en la aplicación de los métodos de diferenciación espectral a la resolución numérica de ecuaciones en derivadas parciales. En este capítulo propondremos y analizaremos un método Fourier-Galerkin y un método pseudoespectral que utilizan la discretización espectral en espacio para resolver numéricamente la ecuación Korteweg-de Vries. En ambos casos se trata de métodos continuos en tiempo.

La ecuación Korteweg-de Vries (KdV) es una ecuación en derivadas parciales no lineal de tercer orden que se puede expresar como

$$u_t + uu_x + \alpha u_{xxx} = 0\,,$$

donde α es un parámetro real distinto de cero. En este caso será importante expresar el segundo término de la forma $\frac{1}{2}(u^2)_x$, ya que facilita mucho la implementación de los métodos tanto de Fourier-Galerkin como los pseudoespectrales (de colocación). Por ello es habitual el uso de este tipo de métodos en muchas aplicaciones de la ecuación KdV. El objetivo de este capítulo será demostrar que para estos métodos se mantiene el orden de convergencia llamado *exponencial* característico de los métodos espectrales cuando se discretizan ecuaciones lineales.

Consideramos el problema de valores iniciales 2π-periódico para la ecuación KdV

$$\begin{cases} u_t + uu_x + \alpha u_{xxx} = 0, \quad x \in \mathbb{R}, \quad t > 0\,, & (4.1)\\ u(x+2\pi,t) = u(x,t), \quad x \in \mathbb{R}, \quad t > 0\,, & (4.2)\\ u(x,0) = u^0(x), \quad x \in \mathbb{R}\,, & (4.3) \end{cases}$$

donde el dato inicial $u^0(x)$ es una función real 2π-periódica.

Para el análisis de los métodos espectrales utilizaremos la misma notación que en el capítulo anterior: $\|\cdot\|$ denotará la norma en $L^2(0,2\pi)$ mientras que $\|\cdot\|_m$ denotará la norma en $H^m(0,2\pi)$. Además, si Λ es un intervalo de la recta $\mathbb{R}$ y X es un espacio de Banach, para las funciones $f : \mathbb{R} \to X$ denotaremos

$$\|f\|_{L^\infty(\Lambda,X)} = \sup_{t\in\Lambda} \|f(t)\|_X\ . \tag{4.4}$$

Con esta notación podemos establecer (ver [24]) una acotación para la solución del problema de valores iniciales de la ecuación KdV dado por el sistema (4.1), (4.2), (4.3). Si u^0 pertenece a $H_p^m(0,2\pi)$, con m natural, entonces la solución del sistema verifica

$$\|u\|_{L^\infty(0,T;H_p^m(0,2\pi))} \leq \eta(m,T;\left\|u^0\right\|_m)\,, \quad m \geq 1\,, \tag{4.5}$$

para todo $T > 0$, donde la constante η solo depende de los términos indicados entre paréntesis.

También introduciremos en este punto el *lema de Gronwall* junto con su demostración, que será útil para el análisis del método.

Lema 4.1. de Gronwall. Sea J un intervalo en $\mathbb{R}$, $t_0 \in J$ y sean $a, \beta, u \in C(J, \mathbb{R}_+)$. Supongamos además que

$$u(t) \leq a(t) + \left|\int_{t_0}^t \beta(s)u(s)ds\right|, \qquad \forall t \in J\,. \tag{4.6}$$

Entonces se cumple que

$$u(t) \leq a(t) + \left|\int_{t_0}^t a(s)\beta(s)e^{|\int_s^t \beta(\sigma)d\sigma|}ds\right|, \qquad \forall t \in J\,. \tag{4.7}$$

Demostración. Sea $v(t) := \int_{t_0}^t \beta(s)u(s)ds$. Entonces, por (4.6) se verifica

$$v'(t) = \beta(t)u(t) \leq a(t)\beta(t) + \text{sgn}(t-t_0)\beta(t)v(t), \qquad \forall t \in J\,. \tag{4.8}$$

Sea ahora

$$\begin{aligned}\gamma(t) := \exp\left(-\left|\int_{t_0}^t \beta(s)ds\right|\right) &= \exp\left(-\int_{t_0}^t \text{sgn}(t-t_0)\beta(s)ds\right)\\ &= \exp\left(-\int_{t_0}^t \text{sgn}(s-t_0)\beta(s)ds\right)\,.\end{aligned}$$

Como $\gamma'(t) = \gamma(t)\text{sgn}(t_0 - t)\beta(t)$, multiplicando la ecuación (4.8) por $\gamma(t) > 0$ tenemos

$$\gamma v' \leq a\beta\gamma - \gamma' v\,,$$

luego $(\gamma v)' \leq a\beta\gamma$. Integramos teniendo en cuenta que $v(t_0) = 0$ y que en general t puede ser mayor o menor que t_0:

$$\text{sgn}(t-t_0)\gamma(t)v(t) \leq \text{sgn}(t-t_0)\int_{t_0}^t a(s)\beta(s)\gamma(s)ds, \qquad \forall t \in J\,.$$

Como $\gamma(t) > 0$,

$$\text{sgn}(t-t_0)v(t) \leq \text{sgn}(t-t_0)\int_{t_0}^t \frac{a(s)\beta(s)\gamma(s)}{\gamma(t)}ds = \left|\int_{t_0}^t \frac{a(s)\beta(s)\gamma(s)}{\gamma(t)}ds\right|, \qquad \forall t \in J\,.$$

Si s está en el intervalo de extremos t y t_0 se cumple la igualdad:

$$\text{sgn}(s-t_0) = \text{sgn}(t-t_0) = \text{sgn}(t-s)\,.$$

Por la definición de γ tenemos, en ese caso:

$$\begin{aligned}\frac{\gamma(s)}{\gamma(t)} &= \exp\left(-\int_{t_0}^{t} \operatorname{sgn}(t-t_0)\beta(\sigma)d\sigma + \int_{t_0}^{s} \operatorname{sgn}(s-t_0)\beta(\sigma)d\sigma\right) \\ &= \exp\left(\operatorname{sgn}(t-s)\int_{s}^{t}\beta(\sigma)d\sigma\right) \\ &= \exp\left(\left|\int_{s}^{t}\beta(\sigma)d\sigma\right|\right)\,.\end{aligned}$$

A partir de (4.6) llegamos a la desigualdad buscada:

$$u(t) \le a(t) + \operatorname{sgn}(t-t_0)v(t) \le a(t) + \left|\int_{t_0}^{t} a(s)\beta(s)e^{\left|\int_s^t \beta(\sigma)d\sigma\right|}ds\right|, \qquad \forall t \in J\,.$$

□

Tras introducir estos resultados podemos pasar al análisis de los métodos espectrales. Como hemos comentado en la introducción del trabajo, inicialmente nuestro objetivo era el análisis de un método pseudoespectral para la ecuación KdV mediante el método de la energía, pero durante el estudio del mismo hallamos una dificultad técnica que nos impidió dar una prueba de la convergencia del método. Por ello hemos optado por hacer el estudio de un método espectral Fourier-Galerkin, dado que las técnicas empleadas son exactamente las mismas que se emplean para los métodos pseudo-espectrales y en ambos métodos el resultado final que se obtiene tras el análisis es que se mantiene la convergencia espectral (el error decrece más rápido que cualquier potencia de h si la condición inicial es suficientemente regular). Tras el estudio del método Fourier-Galerkin indicaremos los resultados análogos que se obtienen para un método pseudoespectral, cuya demostración no incluiremos por ser excesivamente técnica y extensa. Para el estudio de ambos métodos tomaremos como referencia el artículo de Maday y Quarteroni [16], que es el único que hemos encontrado en el que se presenta una prueba para un método pseudoespectral de la ecuación KdV, la cual es bastante más técnica y extensa que la correspondiente al método espectral Fourier-Galerkin.

4.1. Método Fourier-Galerkin

Vamos a analizar en primer lugar el método de Fourier-Galerkin, el cual consiste en encontrar una función $u_N : [0,T] \to \mathcal{S}_N$ tal que

$$\begin{cases} \langle \dfrac{\partial u_N}{\partial t} + u_N \dfrac{\partial u_N}{\partial x} + \alpha \dfrac{\partial^3 u_N}{\partial x^3}, \varphi \rangle = 0, \quad \forall \varphi \in \mathcal{S}_N, \quad \forall t \in [0,T], & (4.9) \\ u_N(0) = P_N u^0. & (4.10) \end{cases}$$

Si escribimos $u_N(x,t)$ en la forma

$$u_N(x,t) = \sum_{k=-N/2}^{N/2} c_k(t)\phi_k(x),$$

el sistema (4.9),(4.10) representa un sistema no lineal de ecuaciones diferenciales ordinarias en los coeficientes $c_k(t), k = -N/2, \ldots, N/2$, de la aproximación Fourier-Galerkin $u_N(x,t)$, imponiendo la condición (4.9) para las funciones $\phi_k \in \mathcal{S}_N$.

Vamos a establecer un lema que afirma que la solución numérica para la discretización Fourier-Galerkin verifica las tres leyes fundamentales de conservación que verifican las soluciones de la ecuación KdV, las cuales se pueden consultar en el capítulo 5 del libro de Drazin [7].

Lema 4.2. Existe una única solución $u_N(t)$ de (4.9),(4.10). Además $u_N(t)$ deja invariante las tres primeras integrales de la energía de la ecuación KdV:

$$\frac{d}{dt}\left[\int_0^{2\pi} u_N(x,t)dx\right] = 0, \tag{4.11}$$

$$\frac{d}{dt}\left[\int_0^{2\pi} |u_n(x,t)|^2 dx\right] = 0, \tag{4.12}$$

$$\frac{d}{dt}\left[\int_0^{2\pi} \left(\alpha\left(\frac{\partial u_N}{\partial x}(x,t)\right)^2 - \frac{u_N^3(x,t)}{3}\right)dx\right] = 0. \tag{4.13}$$

Demostración. La existencia de un intervalo maximal de existencia $(0, t_0)$, con $0 < t_0 \leq T$, tal que para $t < t_0$ existe una única solución $u_N(x,t)$ del sistema (4.9), (4.10) es un resultado clásico de la teoría de ecuaciones diferenciales ordinarias.

Para probar (4.11) utilizamos la función test $\varphi \equiv 1$ en (4.9) y la derivación bajo el signo integral. Entonces, para $0 \leq t < t_0$,

$$\frac{d}{dt}\int_0^{2\pi} u_N(x,t)dx + \frac{1}{2}\int_0^{2\pi} \frac{\partial u_N^2}{\partial x}(x,t)dx + \alpha\int_0^{2\pi} \frac{\partial^3 u_N}{\partial x^3}(x,t)dx = 0.$$

Usando la periodicidad de $u_N(\cdot,t)$, el segundo y tercer sumandos son nulos y deducimos (4.11).

Por otra parte, si utilizamos la función test $\varphi = u_N$ en (4.9) análogamente obtenemos

$$\frac{d}{dt}\frac{1}{2}\int_0^{2\pi} u_N^2(x,t)dx + \frac{1}{3}\int_0^{2\pi} \frac{\partial u_N^3(x,t)}{\partial x}dx + \alpha\int_0^{2\pi} u_N(x,t)\frac{\partial^3 u_N(x,t)}{\partial x^3}dx = 0.$$

El segundo sumando es nulo por la periodicidad de $u_N(\cdot,t)$. Integrando por partes en el tercer sumando y usando de nuevo la periodicidad de $u_N(\cdot,t)$ se concluye también que es idénticametne nulo. En definitiva, el primer sumando debe ser nulo, que es la norma L^2 de $u_N(\cdot,t)$, y (4.12) queda establecida.

Si ahora integramos (4.12) entre 0 y t, con $0 \leq t < t_0$ tenemos que

$$\|u_N(\cdot,t)\| = \|u_N(\cdot,0)\| \leq \|u^0\|$$

y la solución en $0 \leq t < t_0$ puede prolongarse hasta $t_0 = T$. Es decir, tenemos la existencia y unicidad de la solución u_N para todo t en $[0,T]$.

Para probar (4.13), tomamos como función test $\varphi = P_N[u_N^2 + 2\,\alpha\,\partial^2 u_N/\partial x^2](\cdot,t)$. Entonces,

$$\int_0^{2\pi} \frac{\partial u_N}{\partial t} P_N\left[u_N^2 + 2\alpha\frac{\partial^2 u_N}{\partial x^2}\right]dx$$
$$+ \frac{1}{2}\int_0^{2\pi} \frac{\partial}{\partial x}\left[u_N^2 + 2\alpha\frac{\partial^2 u_N}{\partial x^2}\right] P_N\left[u_N^2 + 2\alpha\frac{\partial^2 u_N}{\partial x^2}\right]dx = 0.$$

Puesto que $\partial u_N/\partial t$ es un elemento de $\mathcal{S}_N$ y usando (3.5) tenemos

$$\begin{aligned}\int_0^{2\pi} \frac{\partial u_N}{\partial t} P_N \left[u_N^2 + 2\alpha \frac{\partial^2 u_N}{\partial x^2}\right] dx &= \int_0^{2\pi} \frac{\partial u_N}{\partial t} \left[u_N^2 + 2\alpha \frac{\partial^2 u_N}{\partial x^2}\right] dx \\ &= \frac{d}{dt}\left[\int_0^{2\pi} \left(\frac{u_N^3}{3} - \alpha \left(\frac{\partial u_N}{\partial x}\right)^2\right) dx\right],\end{aligned}$$

donde para establecer la última igualdad hemos efectuado una integración por partes.

Por otro lado, usando también (3.5) tenemos que

$$\begin{aligned}&\int_0^{2\pi} \frac{\partial}{\partial x}\left[u_N^2 + 2\alpha \frac{\partial^2 u_N}{\partial x^2}\right] P_N \left[u_N^2 + 2\alpha \frac{\partial^2 u_N}{\partial x^2}\right] dx \\ &= \frac{1}{2}\int_0^{2\pi} \frac{\partial}{\partial x}\left(P_N \left[u_N^2 + 2\alpha \frac{\partial^2 u_N}{\partial x^2}\right]\right)^2 dx = 0,\end{aligned}$$

luego (4.13) queda probado. □

Los dos lemas siguientes, bastante técnicos, establecen cotas *a priori* de la solución Fourier-Galerkin $u_N(\cdot, t)$ tanto en la norma de $H^1(0, 2\pi)$ como en la norma de $H^2(0, 2\pi)$.

A partir de este punto, c denotará una constante positiva independiente de N, no necesariamente la misma en distintas ecuaciones.

Lema 4.3. Supongamos que u^0 pertence al espacio $H_p^1(0, 2\pi)$. Entonce existe una constante $c > 0$ independiente de N tal que para todo t, $0 \leq t \leq T$,

$$\|u_N(\cdot, t)\|_1 \leq c. \tag{4.14}$$

Demostración. Integrando (4.13), se tiene para $0 \leq t \leq T$

$$\int_0^{2\pi} \left(\alpha \left(\frac{\partial u_N}{\partial x}\right)^2 - \frac{u_N^3}{3}\right)(x, t)dx = \int_0^{2\pi} \left(\alpha \left(\frac{\partial u_N}{\partial x}\right)^2 - \frac{u_N^3}{3}\right)(x, 0)dx. \tag{4.15}$$

Utilizando la inyección continua de $H_p^1(0, 2\pi) \hookrightarrow L^\infty(0, 2\pi)$ (consultar [3, Cap. 4]) tenemos

$$\int_0^{2\pi} \frac{u_N^3}{3}(x, 0)dx \leq \frac{1}{3}\|u_N(\cdot, 0)\|_{L^\infty}\|u_N(\cdot, 0)\|^2 \leq c\|u_N(\cdot, 0)\|_1\|u_N(\cdot, 0)\|^2.$$

Por la definición de $u_N(\cdot, 0)$ y usando (3.16) llegamos a que

$$\int_0^{2\pi} \frac{u_N^3}{3}(x, 0)dx \leq c\|u^0\|_1\|u^0\|^2 \leq c(\|u^0\|_1^2 + \|u^0\|^4).$$

Con el mismo argumento

$$\begin{aligned}\int_0^{2\pi} \frac{u_N^3}{3}(x, t)dx &\leq c\|u_N(\cdot, t)\|_1\|u_N(\cdot, t)\|^2 \leq c\|u_N(\cdot, t)\|_1\|u^0\|^2 \\ &\leq \frac{|\alpha|}{2}\|u_N(\cdot, t)\|_1^2 + c\|u^0\|^4,\end{aligned}$$

donde hemos utilizado en la última desigualdad que $ab \leq \frac{1}{2\varepsilon}a^2 + \frac{\varepsilon}{2}b^2$, para $a, b \geq 0$ y $\varepsilon > 0$.

Para obtener la cota deseada de $\|u_N(\cdot,t)\|_1$, notemos primero que

$$\begin{aligned}\|u_N(\cdot,t)\|_1^2 &= \left(\int_0^{2\pi} u_N^2(\cdot,t)dx\right)^2 + \left\|\frac{d}{dx}u_N(\cdot,t)\right\|^2 \\ &= \left(\int_0^{2\pi} u_N^2(\cdot,0)dx\right)^2 + \left\|\frac{d}{dx}u_N(\cdot,t)\right\|^2\end{aligned}$$

y que, por (4.15), el segundo sumando de la ecuación anterior lo podemos acotar por

$$\begin{aligned}\int_0^{2\pi}\left(\frac{\partial u_N}{\partial x}(x,t)\right)^2 dx &= \int_0^{2\pi}\left(\frac{\partial u_N}{\partial x}(x,0)\right)^2 dx \\ &\quad + \frac{1}{|\alpha|}\int_0^{2\pi}\frac{u_N^3(x,t)}{3}dx - \frac{1}{|\alpha|}\int_0^{2\pi}\frac{u_N^3(x,0)}{3}dx \\ &\leq \int_0^{2\pi}\left(\frac{\partial u_N}{\partial x}(x,0)\right)^2 dx \\ &\quad + \frac{1}{2}\|u_N(\cdot,t)\|_1^2 + \frac{c}{|\alpha|}(\|u^0\|_1^2 + 2\|u^0\|^4).\end{aligned}$$

Por tanto,

$$\begin{aligned}\|u_N(\cdot,t)\|_1^2 \leq &\left(\int_0^{2\pi} u_N^2(\cdot,0)dx\right)^2 + 2\int_0^{2\pi}\left(\frac{\partial u_N}{\partial x}(x,0)\right)^2 dx \\ &+ \frac{2c}{|\alpha|}\left(\|u^0\|_1^2 + 2\|u^0\|^4\right),\end{aligned}$$

que prueba la cota a priori de la norma en $H_p^1(0,2\pi)$ de $u_N(\cdot,t)$, □

Lema 4.4. Supongamos que u^0 pertence al espacio $H_p^2(0,2\pi)$. Entonce existe una constante $c > 0$ independiente de N tal que para todo t, $0 \leq t \leq T$,

$$\|u_N(\cdot,t)\|_2 \leq c. \tag{4.16}$$

Demostración. Si tomamos como función test en (4.9) a

$$\varphi = P_N\left[u_N^3 + 3\alpha\left(\frac{\partial u_N}{\partial x}\right)^2 + 6\alpha u_N\frac{\partial^2 u_N}{\partial x^2} + \frac{18}{5}\alpha^2\frac{\partial^4 u_N}{\partial x^4}\right](\cdot,t),$$

resulta

$$\int_0^{2\pi}\left[\frac{\partial u_N}{\partial t}\varphi + u_N\frac{\partial u_N}{\partial x}\varphi + \alpha\frac{\partial^3 u_N}{\partial x^3}\varphi\right]dx = 0. \tag{4.17}$$

Para el primer término en (4.17), como $\partial u_N/\partial t \in \mathcal{S}_N$, por la relación (3.5) tenemos

$$\int_0^{2\pi}\frac{\partial u_N}{\partial t}\varphi dx = \int_0^{2\pi}\frac{\partial u_N}{\partial t}\left(u_N^3 + 3\alpha\left(\frac{\partial u_N}{\partial x}\right)^2 + 6\alpha u_N\frac{\partial^2 u_N}{\partial x^2} + \frac{18}{5}\alpha^2\frac{\partial^4 u_N}{\partial x^4}\right)dx,$$

e, integrando por partes,

$$\int_0^{2\pi} \frac{\partial u_N}{\partial t}\varphi dx = \frac{d}{dt}\int_0^{2\pi}\left[\frac{u_N^4}{4} - 3\alpha u_N\left(\frac{\partial u_N}{\partial x}\right)^2 + \frac{9}{5}\alpha^2\left(\frac{\partial^2 u_N}{\partial x^2}\right)^2\right]dx. \tag{4.18}$$

Para continuar, notemos que

$$\varphi = P_N\left[u_N^3 + 3\alpha\left(\frac{\partial u_N}{\partial x}\right)^2 + 6\alpha u_N\frac{\partial^2 u_N}{\partial x^2}\right] + \frac{18}{5}\alpha^2\frac{\partial^4 u_N}{\partial x^4},$$

y, por tanto

$$\int_0^{2\pi}\left[u_N\frac{\partial u_N}{\partial x} + \alpha\frac{\partial^3 u_N}{\partial x^3}\right]\varphi dx = A + B + C + D + E, \tag{4.19}$$

donde hemos agrupado los términos de la integral de la siguiente forma:

$$\begin{aligned}
A &= \frac{18}{5}\alpha^3\int_0^{2\pi}\frac{\partial^3 u_N}{\partial x^3}\frac{\partial^4 u_N}{\partial x^4}dx = 0, \quad \text{(por periodicidad)}\\
B &= \int_0^{2\pi} u_N\frac{\partial u_N}{\partial x}P_N(u_N^3)dx,\\
C &= \alpha\int_0^{2\pi}\left[\frac{\partial^3 u_N}{\partial x^3}u_N^3 + 3u_N\frac{\partial u_N}{\partial x}P_N\left[u_N\frac{\partial^2 u_N}{\partial x^2}\right]\right]dx,\\
D &= 3\alpha\int_0^{2\pi} u_N\frac{\partial u_N}{\partial x}P_N\left[\left(\frac{\partial u_N}{\partial x}\right)^2 + u_N\frac{\partial^2 u_N}{\partial x^2}\right]dx,\\
E &= \alpha^2\int_0^{2\pi}\left[\frac{18}{5}u_N\frac{\partial u_N}{\partial x}\frac{\partial^4 u_N}{\partial x^4} + 3\frac{\partial^3 u_N}{\partial x^3}\left(\left(\frac{\partial u_N}{\partial x}\right)^2 + 2u_N\frac{\partial^2 u_N}{\partial x^2}\right)\right]dx.
\end{aligned}$$

Acotaremos a continuación cada uno de los términos en el segundo miembro de (4.19). Para acotar B utilizamos de nuevos la inyección continua $H_p^1(0,2\pi)\hookrightarrow L^\infty(0,2\pi)$, para escribir

$$B \le c\|u_N\|_1\left\|\frac{\partial u_N}{\partial x}\right\|_0\|P_N(u_N^3)\| \le c\|u_N\|_1^2\|u_N^3\| \le c\|u_N\|_1^2\|u_N^3\|_1.$$

Ahora, como $H_p^1(0,2\pi)$ es un álgebra (ver [3, Cap. 4]), $\|u_N^3\|_1 \le c\|u_N\|_1^3$, y por tanto, puesto que $\|u_N(\cdot,t)\|_1 \le c$, por el lema anterior concluimos que

$$|B| \le c. \tag{4.20}$$

Consideremos ahora el término C. Después de integrar por partes el primer sumando tenemos

$$\begin{aligned}
C &= -\alpha\int_0^{2\pi}\left[\frac{\partial^2 u_N}{\partial x^2}\left(3u_N^2\frac{\partial u_N}{\partial x}\right) - 3u_N\frac{\partial u_N}{\partial x}P_N\left(u_N\frac{\partial^2 u_N}{\partial x^2}\right)\right]dx\\
&= 3\alpha\int_0^{2\pi} u_N\frac{\partial u_N}{\partial x}\left[P_N\left(u_N\frac{\partial^2 u_N}{\partial x^2}\right) - \left(u_N\frac{\partial^2 u_N}{\partial x^2}\right)\right]dx,
\end{aligned}$$

y podemos acotar

$$|C| \le 3\alpha\|u_N\|_{L^\infty}\|u_N\|_1\left\|P_N\left(u_N\frac{\partial^2 u_N}{\partial x^2}\right) - \left(u_N\frac{\partial^2 u_N}{\partial x^2}\right)\right\|.$$

Finalmente, utilizando la propiedad (3.16) y que $H_p^1(0,2\pi) \hookrightarrow L^\infty(0,2\pi)$,

$$\begin{aligned}|C| &\leq c\|u_N\|_{L^\infty}\,\|u_N\|_1\,\left\|u_N\frac{\partial^2 u_N}{\partial x^2}\right\| \leq c\,\|u_N\|_{L^\infty}\,\|u_N\|_1\,\|u_N\|_{L^\infty}\,\left\|\frac{\partial^2 u_N}{\partial x^2}\right\| \\ &\leq c\,\|u_N\|_1^3\,\|u_N\|_2\,,\end{aligned}$$

y, por la acotación de $\|u_N\|_1$ (4.14), concluimos que

$$|C| \leq c\|u_N\|_2. \tag{4.21}$$

Para el término D tenemos

$$\begin{aligned}D &= 3\alpha\int_0^{2\pi} u_N\frac{\partial u_N}{\partial x}P_N\left(\frac{\partial}{\partial x}\left(u_N\frac{\partial u_N}{\partial x}\right)\right)dx \\ &= 3\alpha\int_0^{2\pi} P_N\left(u_N\frac{\partial u_N}{\partial x}\right)P_N\left(\frac{\partial}{\partial x}\left(u_N\frac{\partial u_N}{\partial x}\right)\right)dx \\ &= 3\alpha\int_0^{2\pi} P_N\left(u_N\frac{\partial u_N}{\partial x}\right)\frac{\partial}{\partial x}P_N\left(u_N\frac{\partial u_N}{\partial x}\right)dx = 0.\end{aligned}$$

Por último, de forma similar, si integramos por partes el primer sumando de E llegamos a

$$\begin{aligned}E &= \alpha^2\int_0^{2\pi}\left[-\frac{18}{5}\left(\left(\frac{\partial u_N}{\partial x}\right)^2 + u_N\frac{\partial^2 u_N}{\partial x^2}\right)\left(\frac{\partial^3 u_N}{\partial x^3}\right)\right. \\ &\quad \left. + 3\frac{\partial^3 u_N}{\partial x^3}\left(\left(\frac{\partial u_N}{\partial x}\right)^2 + 2u_N\frac{\partial^2 u_N}{\partial x^2}\right)\right]dx \\ &= -\alpha^2\int_0^{2\pi}\left[\frac{3}{5}\left(\frac{\partial u_N}{\partial x}\right)^2\frac{\partial^3 u_N}{\partial x^3} - \frac{6}{5}u_N\frac{\partial}{\partial x}\left(\frac{\partial^2 u_N}{\partial x^2}\right)^2\right]dx \\ &= \alpha^2\int_0^{2\pi}\left[\frac{3}{5}\frac{\partial}{\partial x}\left(\left(\frac{\partial u_N}{\partial x}\right)^2\right)\frac{\partial^2 u_N}{\partial x^2} + \frac{6}{5}u_N\frac{\partial}{\partial x}\left(\frac{\partial^2 u_N}{\partial x^2}\right)^2\right]dx\end{aligned}$$

y, tras integrar por partes, obtenemos que todos los términos se cancelan y $E = 0$.

Podemos acotar entonces el primer miembro de (4.19)

$$\left|\int_0^{2\pi}\left[\frac{1}{2}\frac{\partial u_N^2}{\partial x} + \alpha\frac{\partial^3 u_N}{\partial x^3}\right]\varphi\,dx\right| \leq c\,(1+\|u_N\|_2)$$

y, obtener de (4.17) y (4.18)

$$\frac{d}{dt}\int_0^{2\pi}\left[\frac{u_N^4}{4} - 3\alpha u_N\left(\frac{\partial u_N}{\partial x}\right)^2 + \frac{9}{5}\alpha^2\left(\frac{\partial^2 u_N}{\partial x^2}\right)^2\right]dx \leq c\,(1+\|u_N\|_2).$$

Integrando esta expresión entre 0 y t, con $0 < t \leq T$, y usando que $\|u_N\|_1 \leq c$, finalmente tenemos

$$\int_0^{2\pi}\frac{9}{5}\alpha^2\left(\frac{\partial^2 u_N}{\partial x^2}\right)^2 dx \leq c\left(1 + \int_0^t \|u_N(\cdot,s)\|_2^2 ds\right), \tag{4.22}$$

donde c depende de la norma en $H_p^2(0,2\pi)$ de u^0 y de T. Por último, observamos que

$$\|u_N(\cdot,s)\|_2^2 \leq c + \left\| \frac{\partial^2 u_N}{\partial x^2}(\cdot,s) \right\|^2,$$

y utilizando esta expresión junto con el lema de Gronwall en (4.22) para acotar $\left\| \frac{\partial^2 u_N}{\partial x^2} \right\|^2$ obtenemos (4.16). □

Tras introducir estos lemas abordamos ahora la convergencia de la aproximación Fourier-Galerkin a la solución de la ecuación KdV, que va a ser el resultado fundamental de este apartado:

Teorema 4.5. Supongamos que u^0 pertenece a $H_p^m(0,2\pi)$, para algún entero $m \geq 2$. Entonces existe una constante $c > 0$ independiente de N tal que para todo t, con $0 \leq t \leq T$, se tiene

$$\|u(\cdot,t) - u_N(\cdot,t)\| \leq c\, N^{1-m}. \tag{4.23}$$

Demostración. Para todo t, $0 \leq t \leq T$, definimos

$$e(t) = P_N u(t) - u_N(t)\,.$$

En este caso, aplicando la propiedad (3.12) se cumple que

$$\frac{\partial e}{\partial t} + \alpha \frac{\partial^3 e}{\partial x^3} = P_N \left(\frac{\partial u}{\partial t} + \alpha \frac{\partial^3 u}{\partial x^3} \right) - \left(\frac{\partial u_N}{\partial t} + \alpha \frac{\partial^3 u_N}{\partial x^3} \right).$$

Introduciremos también la notación $E[f,g] = f f_x - g g_x$. Entonces, de (4.1) y (4.9) deducimos que $\forall \varphi \in \mathcal{S}_N$ se cumple

$$\begin{aligned} \langle \frac{\partial e}{\partial t} + \alpha \frac{\partial^3 e}{\partial x^3}, \varphi \rangle &= \langle -P_N \left(u \frac{\partial u}{\partial x} \right) + u_N \frac{\partial u_N}{\partial x}, \varphi \rangle \\ &= \langle -u \frac{\partial u}{\partial x} + u_N \frac{\partial u_N}{\partial x}, \varphi \rangle \\ &= \langle E[P_N u, u] - E[P_N u, u_N], \varphi \rangle\,, \end{aligned} \tag{4.24}$$

donde hemos usado también la propiedad (3.5). Ahora elegimos como función test en (4.24) $\varphi = e$, es decir

$$\frac{d}{dt}\|e\|^2 + \langle \alpha \frac{\partial^3 e}{\partial x^3}, e \rangle = \langle E[P_N u, u], e \rangle - \langle E[P_N u, u_N], e \rangle\,. \tag{4.25}$$

Además se tiene que $\langle \alpha \frac{\partial^3 e}{\partial x^3}, e \rangle = 0$ integrando por partes y aplicando la periodicidad. Para el primer sumando del segundo miembro, como $E[P_N u, u] = \frac{1}{2}\frac{\partial}{\partial x}((P_N u)^2 - u^2)$, usando la desigualdad (3.16) y (4.5) para $\|P_N u\|_1$ tenemos

$$\begin{aligned} |\langle E[P_N u, u], e \rangle| &\leq c\,(\|u\|_1 + \|P_N u\|_1)\,\|u - P_N u\|_1\,\|e\| \\ &\leq c\,\|u - P_N u\|_1\,\|e\| \\ &\leq c\,(\|u - P_N u\|_1^2 + \|e\|^2)\,. \end{aligned} \tag{4.26}$$

Para el segundo sumando integramos por partes y usamos la definición de e

$$\begin{aligned}|\langle E[P_N u, u_N], e\rangle| &= \left|\frac{1}{2}\int_0^{2\pi} \frac{\partial}{\partial x}\left[(P_N u)^2 - u_N^2\right] e\, dx\right| \\ &= \left|\frac{1}{2}\int_0^{2\pi} \left[P_N u + u_N\right] e \frac{\partial e}{\partial x}\, dx\right| \\ &= \left|\frac{1}{4}\int_0^{2\pi} \left[P_N u + u_N\right] \frac{\partial e^2}{\partial x}\, dx\right| \\ &= \left|\frac{1}{4}\int_0^{2\pi} \frac{\partial}{\partial x}\left[P_N u + u_N\right] e^2\, dx\right| .\end{aligned}$$

Ahora, usando (4.5), (4.16) y $H_p^1(0,2\pi) \hookrightarrow L^\infty(0,2\pi)$ llegamos a que

$$|\langle E[P_N u, u_N], e\rangle| \leq c \left\| \frac{\partial}{\partial x}[P_N u + u_N]\right\|_{L^\infty} \|e\|^2 \leq c\, \|e\|^2 . \tag{4.27}$$

Por tanto, de las ecuaciones (4.25), (4.26) y (4.27) tenemos que

$$\frac{d}{dt}\|e\|^2 \leq c\left(\|u - P_N u\|_1^2 + \|e\|^2\right) . \tag{4.28}$$

Además, por la definición de e tenemos que $e(0) = 0$. Aplicando el lema de Gronwall obtenemos la desigualdad

$$\|e(t)\| \leq c\, e^{ct} \left(\int_0^t \|u(s) - P_N u(s)\|_1^2\, ds\right)^{\frac{1}{2}} .$$

Para finalizar con la demostración observamos que, por (3.16), el integrando se puede acotar

$$\|u - P_N u\|_1 \leq c N^{1-m} \|u\|_m .$$

Además, el término de la derecha lo podemos acotar mediante (4.5) si $u^0 \in H_p^m(0,2\pi)$ para algún m. Por último, como

$$\|u - u_N\| \leq \|u - P_N u\| + \|e\| ,$$

solo nos falta acotar $\|u - P_N u\|$ de la misma forma usando (4.5) y (3.16) y obtenemos el resultado buscado:

$$\|u(\cdot,t) - u_N(\cdot,t)\| \leq c\, N^{1-m}.$$

□

En resumen, el resultado que hemos obtenido en el teorema (4.5) nos dice que el método Fourier-Galerkin es convergente y además el orden de convergencia únicamente depende de la regularidad de la condición inicial u^0. Esto quiere decir que si u^0 es suficientemente regular, entonces el error en norma $L^2(0,2\pi)$ tiende a cero más rápido que cualquier potencia de $1/N$. Este comportamiento, al que habitualmente nos referiremos como *convergencia exponencial* es la principal ventaja que presentan los métodos espectrales frente a los métodos también muy usados de diferencias finitas o elementos finitos.

Podemos obtener como corolario de este teorema el orden de convergencia en la norma de $H_p^1(0,2\pi)$:

Corolario 4.6. Si u^0 pertenece a $H_p^m(0, 2\pi)$ para algún entero $m \geq 2$, entonces existe una constante $c > 0$ independiente de N tal que para todo t, con $0 \leq t \leq T$, se tiene

$$\|u(\cdot, t) - u_N(\cdot, t)\|_1 \leq c\, N^{2-m}. \tag{4.29}$$

Demostración. Aplicando la desigualdad triangular tenemos que

$$\|u(\cdot, t) - u_N(\cdot, t)\|_1 \leq \|u(\cdot, t) - P_N u(\cdot, t)\|_1 + \|u_N(\cdot, t) - P_N u(\cdot, t)\|_1 \ . \tag{4.30}$$

El primer sumando lo podemos acotar directamente mediante (3.16), mientras que para el segundo sumando podemos aplicar la desigualdad inversa (3.14), de forma que

$$\begin{aligned} \|u_N(\cdot, t) - P_N u(\cdot, t)\|_1 &\leq c\, N \|u_N(\cdot, t) - P_N u(\cdot, t)\| \\ &\leq c\, N \left(\|u(\cdot, t) - u_N(\cdot, t)\| + \|u(\cdot, t) - P_N u(\cdot, t)\|\right) \ . \end{aligned}$$

Entonces, de (3.16) y (4.23) deducimos

$$\|u_N(\cdot, t) - P_N u(\cdot, t)\|_1 \leq c\, N^{2-m} \, .$$

Por tanto, acotando los dos sumandos de (4.30) demostramos la propiedad de convergencia. □

Como ya hemos comentado en los capítulos previos, posteriormente implementaremos un método espectral para la resolución numérica de la ecuación KdV y comprobaremos la convergencia exponencial. A continuación incluimos los resultados del único análisis de un método pseudoespectral que hemos encontrado para la ecuación KdV. No es exactamente el que vamos a implementar, pero consideramos ilustrativo incluir los resultados que prueban esa convergencia exponencial para un método de colocación pseudoespectral. La demostración de los resultados siguientes es bastante más técnica y extensa que para el método de Galerkin que acabamos de analizar, de forma que nos limitaremos a enunciar los resultados sin demostración.

4.2. Método de colocación pseudoespectral

De la misma forma que en los capítulos anteriores denotaremos como $h = 2\pi/N$ el espaciado de la red de puntos formada por $x_j = jh$ y el operador de diferenciación espectral será D_N.

El método de colocación pseudoespectral continuo en tiempo que se analiza en [16] consiste en encontrar una función $u_N : [0, T] \to \mathcal{S}_N$ tal que:

$$\begin{cases} \left(\dfrac{\partial u_N}{\partial t} + \dfrac{1}{2} D_N(u_N^2) + \alpha \dfrac{\partial^3 u_N}{\partial x^3}\right)(x_j) = 0, \quad \forall t,\, 0 \leq t \leq T, \quad \forall j,\, 0 \leq j \leq N-1, & (4.31) \\ u_N(0, x_j) = u^0(x_j)\,, \quad \forall j,\, 0 \leq j \leq N-1\,. & (4.32) \end{cases}$$

Observamos que este método solo incluye la diferenciación espectral para el término no lineal de la ecuación KdV, mientras que a la hora de implementar el método en el Capítulo 6 discretizaremos espectralmente también el término de tercer orden en espacio.

Para el análisis de este método será útil definir dos problemas, el primero de ellos será un problema continuo para la ecuación KdV, mientras que el segundo será un problema de colocación pseudoespectral con una condición inicial distinta.

Definimos el problema para la ecuación KdV con condición inical v^0:

$$\begin{cases} \dfrac{\partial v}{\partial t} + v\dfrac{\partial v}{\partial x} + \alpha\dfrac{\partial^3 v}{\partial x^3} = 0, \quad x \in \mathbb{R}\,,\ t > 0\,, & (4.33) \\ v(x+2\pi,t) = v(x,t)\,, \quad x \in \mathbb{R}\,,\ t > 0\,, & (4.34) \\ v(x,0) = v^0(x)\,, \quad x \in \mathbb{R}\,. & (4.35) \end{cases}$$

Por otra parte, definimos el problema de colocación pseudoespectral cuya condición inicial es $v_N^0 \in \mathcal{S}_N$, una aproximación a v^0 que no tiene por qué coincidir con su interpolante $I_N v^0$ (este aspecto quedará más claro en los resultados posteriores):

Encontrar una función $v_N : [0,T] \to \mathcal{S}_N$ tal que

$$\begin{cases} \left(\dfrac{\partial v_N}{\partial t} + \dfrac{1}{2}D_N(v_N^2) + \alpha\dfrac{\partial^3 v_N}{\partial x^3}\right)(x_j) = 0, \quad \forall t,\, 0 \le t \le T, \quad \forall j,\, 0 \le j \le N-1, & (4.36) \\ v_N(0,x_j) = v_N^0(x_j)\,, \quad \forall j,\, 0 \le j \le N-1\,. & (4.37) \end{cases}$$

El esquema de demostración será análogo al del método Fourier-Galerkin. En primer lugar se demostrarán unos lemas de estabilidad para la solución numérica del problema (4.36), (4.37).

Lema 4.7. Para todo número real R, existen tres constantes positivas $\tilde{t}_1$, β_1, γ_1 que únicamente dependen de R, tales que para cada dato inicial v_N^0 verificando

$$\left\|v_N^0\right\|_1 \le R \tag{4.38}$$

y para cada t, $0 \le t \le \tilde{t}_1$ se cumple

$$\|v_N(\cdot,t)\| \le \beta_1\,, \tag{4.39}$$
$$\|v_N(\cdot,t)\|_1 \le \gamma_1\,. \tag{4.40}$$

Lema 4.8. Para todo número real R, existen tres constantes positivas $t_1^* \le \tilde{t}_1$, β_1^*, γ_1^* que únicamente dependen de R, tales que para cada dato inicial v_N^0 verificando

$$\left\|v_N^0\right\|_4 \le R \tag{4.41}$$

y para cada t, $0 \le t < t_1^*$ se cumple

$$\left\|\frac{\partial v_N}{\partial t}(\cdot,t)\right\| \le \beta_1^*\,, \tag{4.42}$$
$$\left\|\frac{\partial v_N}{\partial t}(\cdot,t)\right\|_1 \le \gamma_1^*\,. \tag{4.43}$$

Para poder estudiar la convergencia se necesita un último resultado de estabilidad en la norma de $H^3(0,2\pi)$.

Lema 4.9. Para todo número real R, existe una constante γ_3 que únicamente depende de R, tal que para cada dato inicial v_N^0 verificando

$$\left\|v_N^0\right\|_4 \leq R \tag{4.44}$$

y para cada t, $0 \leq t \leq t_1^*$ se cumple

$$\|v_N(\cdot,t)\|_3 \leq \gamma_3\,. \tag{4.45}$$

Usando estos lemas se deduce en primer lugar un resultado para la convergencia local de la solución del problema de colocación (4.36), (4.37) a la solución del problema de la ecuación KdV (4.33), (4.34), (4.35) en el intervalo temporal $[0, t_1^*]$:

Proposición 4.10. Supongamos que v^0 pertenece a $H_p^m(0,2\pi)$ para algún $m \geq 4$ y que v_N^0 está acotada en $H_p^4(0,2\pi)$ independientemente de N. Entonces existe una constante $\Lambda_m > 0$, que depende continuamente de $\left\|v_N^0\right\|$ pero independiente de N, tal que para cada t, $0 \leq t \leq t_1^*$:

$$\|v_N(\cdot,t) - v(\cdot,t)\|_1 \leq \Lambda_m N^{2-m} + \left\|v_N^0 - v^0\right\|_1\,. \tag{4.46}$$

Por último, los autores consiguen extender este resultado de convergencia local y demostrar el principal resultado de convergencia global para el problema (4.31), (4.32):

Teorema 4.11. Supongamos que u^0 pertenece a $H_p^m(0,2\pi)$ para algún $m > 4$. Entonces, para cada t, $0 \leq t \leq T$ y cada N suficientemente grande se verifica

$$\|u(\cdot,t) - u_N(\cdot,t)\|_1 \leq c\,N^{2-m}. \tag{4.47}$$

Por tanto, el método pseudoespectral mantiene la convergencia exponencial que demostramos para el método Galerkin.

Como hemos mencionado, este método es ligeramente distinto al que implementaremos en el último capítulo del trabajo. Pero verificaremos numéricamente que se mantiene la convergencia exponencial característica de los métodos espectrales.

Capítulo 5

Implementación de los métodos espectrales

Tras introducir los métodos espectrales y realizar el análisis de convergencia de los mismos, en los siguientes apartados centraremos nuestro estudio en la aplicación eficiente de estos métodos espectrales para la resolución numérica de ecuaciones en derivadas parciales.

Hemos visto que una opción para calcular las derivadas espectrales de una función definida en una red de nodos equiespaciada es aplicar el correspondiente operador de diferenciación espectral D_N. Esto se traduce en la práctica en el cómputo de un producto matriz por vector de dimensión N, lo cual supone N^2 productos y $N(N-1)$ sumas. A esto se le añade que si necesitamos calcular derivadas espectrales de orden superior, debemos en primer lugar calcular cuál es la expresión concreta del operador para luego aplicarlo.

También hemos comentado la posibilidad de trabajar con los coeficientes de Fourier, ya que la relación entre una función y su derivada espectral de cualquier orden es mucho más sencilla. Simplemente es necesario multiplicar los coeficientes de Fourier por el factor $(ik)^m$. Para implementarlo es necesario en primer lugar calcular la transformada de Fourier de la función inicial y, tras efectuar el producto por $(ik)^m$, hacer la transformada inversa para obtener los valores buscados.

Si nos quedamos únicamente con la definición que dimos de la transformada discreta, observamos que para su cálculo necesitamos N^2 productos de números complejos, dado que se puede expresar como un producto matriz por vector.

En los dos casos estaríamos desaprovechando información, porque las matrices implicadas tienen una estructura definida que hace que muchos de sus elementos guarden relación entre ellos. El algoritmo de la transformada rápida de Fourier (FFT) que describiremos a continuación utiliza esta propiedad para permitir el cálculo de la transformada discreta con un costo operativo de tan solo $\frac{N}{2}\log_2\frac{N}{2}$.

Ciertamente este es un ahorro sustancial. Pensemos, por ejemplo en $N = 2^{10}$. En este caso $N^2 = 1048576$, mientras que $\frac{N}{2}\log_2\frac{N}{2} = 4608$. La diferencia es nada menos que un factor de más de 200.

5.1. La transformada rápida de Fourier

En esta sección y en las subsiguientes, como es habitual en la literatura, adoptaremos una notación ligeramente diferente a la que empleamos cuando introdujimos la transformada discreta.

Supondremos que el vector que queremos transformar tiene tamaño N (par), con índices desde 0 hasta $N-1$: $\boldsymbol{x} = (x_0, \dots, x_{N-1})$ El vector transformado lo denotaremos por $\boldsymbol{X}$ (también de tamaño N) y sus componentes serán también $\boldsymbol{X} = (X_0, \dots, X_{N-1})$.

La diferencia estriba en la forma de ordenar las componentes del vector transformado, ya que en las secciones anteriores los números de onda los considerábamos tanto positivos como negativos

$$k = -\frac{N}{2}+1, -\frac{N}{2}+2, \dots, -1, 0, 1, \dots, \frac{N}{2}-1, \frac{N}{2},$$

mientras que ahora solo los consideramos positivos, desde 0 hasta $N-1$. Este cambio solo implica una reorganización de los términos, porque ya sabemos que los números de onda para una red discreta solo están definidos módulo N. De esta forma el término correspondiente a $k = -\frac{N}{2}+1$ es el mismo que el correspondiente a $k = \frac{N}{2}+1$ y así sucesivamente.

Entonces en esta nueva notación estamos organizando los números de onda que definimos antes como

$$k = 0, 1, \dots, \frac{N}{2}, -\frac{N}{2}+1, -\frac{N}{2}+2 \dots, -1\,.$$

Con esta nueva notación y la interpretación matricial de las fórmulas de la transformada, podemos ver la transformada de Fourier discerta (2.13) como una aplicación lineal de $\mathbb{C}^N$ en sí mismo dada por $\boldsymbol{x} \to \boldsymbol{X} = F_N \boldsymbol{x}$ (como es habitual al tratar la transformada rápida, omitimos el factor h que aparece en la fórmula (2.13) por simplicidad). Aquí F_N es la matriz compleja $N \times N$ constituida por los elementos $(F_N)_{jk} = \omega_N^{jk}$, siendo

$$\omega_N = e^{-\frac{2\pi}{N}i}$$

la raíz N-ésima principal de la unidad de argumento negativo más pequeño.

Es decir, la forma de la matriz es la siguiente:

$$F_N = \begin{pmatrix} \omega_N^0 & \omega_N^0 & \omega_N^0 & \cdots & \omega_N^0 \\ \omega_N^0 & \omega_N^1 & \omega_N^2 & \cdots & \omega_N^{N-1} \\ \omega_N^0 & \omega_N^2 & \omega_N^4 & \cdots & \omega_N^{2(N-1)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \omega_N^0 & \omega_N^{N-1} & \omega_N^{2(N-1)} & \cdots & \omega_N^{(N-1)(N-1)} \end{pmatrix}.$$

Obviamente, como w_N es una raíz $N-$ésima de la unidad, se cumple que $w_N^N = 1$, lo que nos permite expresar la matriz únicamente en función de las potencias de w_N de grado menor que N: $\omega_N^{jk} = \omega_N^{jk \pmod N}$

Además, debido a que la matriz F_N es simétrica, podemos dar una expresión sencilla para la matriz inversa, que va a ser la que nos proporcione la transformada inversa de Fourier:

Proposición 5.1. La matriz F_N es invertible y $F_N^{-1} = \frac{1}{N} F_N^H$, donde F_N^H es la matriz que resulta al conjugar cada elemento de F_N^T.

Demostración. Teniendo en cuenta que F_N es simétrica, basta ver que $F_N\overline{F_N} = NI$:

$$(F_N\overline{F_N})_{jk} = \sum_{l=0}^{N-1} \omega_N^{jl}\bar{\omega}_N^{lk} = \sum_{l=0}^{N-1} \omega_N^{jl}\omega_N^{-lk} = \sum_{l=0}^{N-1} (\omega_N^{j-k})^l \,.$$

Si $j = k$ se tiene que $(F_N\overline{F_N})_{jk} = N$. Si $j \neq k$, la suma de la progresión geométrica de razón $\omega_N^{j-k} \neq 1$ resulta:

$$(F_N\overline{F_N})_{jk} = \frac{(\omega_N^{j-k})^N - 1}{\omega_N^{j-k} - 1} = 0\,,$$

dado que $\omega_N^N = 1$. □

El algoritmo de Cooley y Tukey de 1965 [6] que vamos a estudiar a continuación va a ser el que nos permita reducir el costo operativo del cálculo de la transformada hasta $\frac{N}{2}\log_2\frac{N}{2}$ productos de números complejos.

5.2. Algoritmo de Cooley y Tukey

En la versión más habitual del algoritmo se trabaja con un vector $\boldsymbol{x}$ de longitud $N = 2^m$ y se aplica la técnica de diezmación en tiempos. Esta nos permite dividir el vector en dos mitades, relacionando la transformada original con las transformadas de tamaño mitad de los dos nuevos vectores.

Dividiremos el vector original de tamaño 2^m en los vectores

$$\boldsymbol{x}^P = (x_0, x_2, ..., x_{N-2})^T, \qquad \boldsymbol{x}^I = (x_1, x_3, ..., x_{N-1})^T, \tag{5.1}$$

formados por las componentes pares e impares del original, respectivamente.

El siguiente teorema es el que nos permite calcular la transformada del vector $\boldsymbol{x}$ a partir de las transformadas de los vectores $\boldsymbol{x}^P$ y $\boldsymbol{x}^I$, ambas de tamaño $N/2$.

Proposición 5.2. Si llamamos $\boldsymbol{X}^P$ y $\boldsymbol{X}^I$ a las transformadas de los vectores $\boldsymbol{x}^P$ y $\boldsymbol{x}^I$:

$$\boldsymbol{X}^P = F_{\frac{N}{2}}\boldsymbol{x}^P, \qquad \boldsymbol{X}^I = F_{\frac{N}{2}}\boldsymbol{x}^I,$$

entonces las componentes de $\boldsymbol{X}$ están dadas por

$$\begin{aligned} X_k &= X_k^P + \omega_N^k X_k^I, \quad k = 0, 1, ..., \frac{N}{2} - 1\,, \\ X_{\frac{N}{2}+k} &= X_k^P - \omega_N^k X_k^I, \quad k = 0, 1, ..., \frac{N}{2} - 1\,, \end{aligned} \tag{5.2}$$

Demostración. Separando en el sumatorio que define la transformada los términos con índice j par e impar

$$X_k = \sum_{j=0}^{N-1} \omega_N^{jk} x_j = \sum_{j=0}^{N/2-1} \omega_N^{2jk} x_{2j} + \sum_{j=0}^{N/2-1} \omega_N^{(2j+1)k} x_{2j+1}\,.$$

Teniendo en cuenta que $\omega_{\frac{N}{2}} = \omega_N^2$

$$\begin{aligned} X_k &= \sum_{j=0}^{N/2-1} \omega_{\frac{N}{2}}^{jk} x_j^P + \omega_N^k \sum_{j=0}^{N/2-1} \omega_{\frac{N}{2}}^{jk} x_j^I \\ &= X_k^P + \omega_N^k X_k^I, \quad k = 0, 1, ..., \frac{N}{2} - 1\,. \end{aligned}$$

Análogamente, para los índices de la forma $\frac{N}{2} + k$, con $k = 0, 1, ..., \frac{N}{2} - 1$

$$\begin{aligned} X_{\frac{N}{2}+k} &= \sum_{j=0}^{N/2-1} \omega_{\frac{N}{2}}^{j\frac{N}{2}} \omega_{\frac{N}{2}}^{jk} x_j^P + \omega_N^{\frac{N}{2}} \omega_N^k \sum_{j=0}^{N/2-1} \omega_{\frac{N}{2}}^{j\frac{N}{2}} \omega_{\frac{N}{2}}^{jk} x_j^I \\ &= X_k^P - \omega_N^k X_k^I, \quad k = 0, 1, ..., \frac{N}{2} - 1\,, \end{aligned}$$

ya que $\omega_{\frac{N}{2}}^{\frac{N}{2}} = 1$ y $\omega_N^{\frac{N}{2}} = -1$. □

Este resultado nos permite calcular una transformada N-dimensional de forma recursiva, dividiendo sucesivamente entre 2 la dimensión. Por tanto, el problema se reduce a calcular transformadas de dos elementos, las cuales son inmediatas: Si $\boldsymbol{x} = (x_0, x_1)^T$, entonces $F_2(\boldsymbol{x}) = (x_0 + x_1, x_0 - x_1)^T$.

Para calcular el costo operativo de este algoritmo denotamos por $M(m)$ el número de multiplicaciones de números complejos necesarias para realizar la transformada de Fourier discreta de $N = 2^m$ elementos mediante el algoritmo descrito. Por (5.2), las multiplicaciones que necesitaremos serán las necesarias para calcular la transformada de $\boldsymbol{x}^P$ y de $\boldsymbol{x}^I$, de tamaño 2^{m-1} y los 2^{m-1} productos $\omega_N^k X_k^I$ (suponiendo que tenemos almacenadas las potencias de ω_N). Luego $M(m)$ cumple la recurrencia:

$$M(m) = \begin{cases} 2M(m-1) + 2^{m-1}, & m > 1\,, \\ 0, & m = 1\,. \end{cases}$$

Iterando la expresión anterior podemos escribir:

$$\begin{aligned} M(m) &= 2\left(2M(m-2) + 2^{m-2}\right) + 2^{m-1} \\ &= 2^2 M(m-2) + 2 \cdot 2^{m-1} \\ &= 2^i M(m-i) + i2^{m-1}, \quad i = 3, ..., m-2 \\ &= 2^{m-1} M(1) + (m-1)2^{m-1} \\ &= (m-1)2^{m-1}\,. \end{aligned}$$

Si lo expresamos en función de N tenemos que el costo operativo es de $(\log_2 N - 1)\frac{N}{2} = \frac{N}{2}\log_2 \frac{N}{2}$, lo que supone un ahorro considerable frente a los N^2 productos de números complejos necesarios

para hacer el cálculo directo.

Si queremos calcular el costo en operaciones de punto flotante entre números reales, hay que tener en cuenta que cada suma compleja requiere 2 flops, y cada producto complejo requiere 6 flops. Llamando $C(m)$ a este costo operativo, la recurrencia que satisface es:

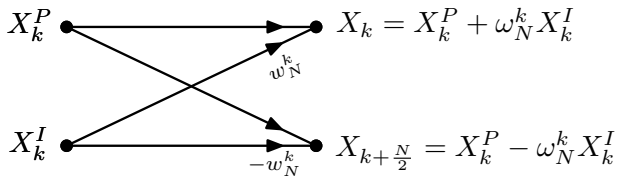
$$C(m) = \begin{cases} 2C(m-1) + 5 \cdot 2^m, & m > 1\,, \\ 4, & m = 1\,. \end{cases}$$

Para hallar el costo computacional en flops operamos de forma análoga:

$$\begin{aligned} C(m) &= 2(2C(m-2) + 5 \cdot 2^{m-1}) + 5 \cdot 2^m \\ &= 2^2 C(m-2) + 5 \cdot 2 \cdot 2^m \\ &= 2^i C(m-i) + 5i2^m, \quad i = 3, ..., m-2 \\ &= 2^{m-1} C(1) + 5(m-1)2^m \\ &= 4 \cdot 2^{m-1} + 5(m-1)2^m \\ &= (5m-3)2^m, \end{aligned}$$

lo que lleva a un costo operativo, expresado en función de N, de $5N \log_2 N - 3N$ flops.

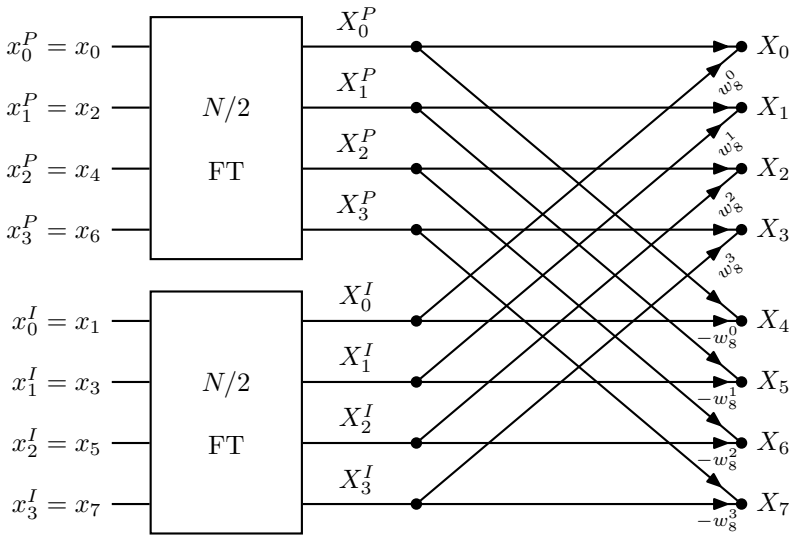
Habitualmente el algoritmo de Cooley y Tukey se esquematiza con una "mariposa" que indica qué componentes deben combinarse y cómo lo hacen:



GRÁFICA 5.1: Mariposa del algoritmo de Cooley y Tukey.

Las flechas que convergen a los nodos de la derecha indican qué componentes deben sumarse, y la presencia de ω_N^k en dos de las flechas indica que esos términos deben ser multiplicados por el factor correspondiente.

Para ver cómo se combinan los distintos términos en una iteración del algoritmo de Cooley y Tukey presentamos el siguiente esquema con las mariposas correspondientes para un tamaño de $N = 8$. En él se muestran las combinaciones necesarias para, tras calcular las dos transformadas de tamaño 4 correspondientes a los vectores de índice par e impar, obtener las componentes de la transformada completa.



GRÁFICA 5.2: Esquema de aplicación de una iteración del algoritmo de Cooley y Tukey para $N = 8$.

Como hemos mencionado antes, este algoritmo se utiliza de forma recursiva, de forma que las cajas con las que hemos representado las transformadas, a su vez se pueden descomponer sucesivamente en dos transformadas de tamaño mitad, además de las combinaciones correspondientes.

5.3. Algoritmo de Gentleman y Sande

A diferencia del algoritmo de Cooley y Tukey que utiliza diezmación en tiempo, el algoritmo de Gentleman y Sande (o también llamado de Sande y Tukey) utiliza lo que se conoce como diezmación en frecuencias. Esto se traduce en el cálculo de los elementos pares e impares de la transformada por separado.

Para los términos pares:

$$X_{2k} = \sum_{j=0}^{N-1} \omega_N^{2kj} x_j, \qquad k = 0, 1, \dots, \frac{N}{2} - 1\,,$$

Si agrupamos la primera y la segunda mitad de los términos por separado:

$$\begin{aligned} X_{2k} &= \sum_{j=0}^{\frac{N}{2}-1} \omega_N^{2kj} x_j + \sum_{j=0}^{\frac{N}{2}-1} \omega_N^{2k\left(j+\frac{N}{2}\right)} x_{j+\frac{N}{2}} \\ &= \sum_{j=0}^{\frac{N}{2}-1} \left(x_j + x_{j+\frac{N}{2}}\right) \omega_{\frac{N}{2}}^{kj} \\ &= \sum_{j=0}^{\frac{N}{2}-1} y_j \omega_{\frac{N}{2}}^{kj}, \qquad k = 0, 1, \dots, \frac{N}{2} - 1\,. \end{aligned}$$

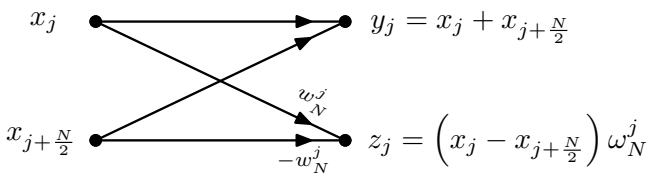
Análogamente, para los términos impares:

$$\begin{aligned} X_{2k+1} &= \sum_{j=0}^{N-1} \omega_N^{(2k+1)j} x_j \\ &= \sum_{j=0}^{\frac{N}{2}-1} \omega_N^{(2k+1)j} x_j + \sum_{j=0}^{\frac{N}{2}-1} \omega_N^{(2k+1)(j+\frac{N}{2})} x_{j+\frac{N}{2}} \\ &= \sum_{j=0}^{\frac{N}{2}-1} \left(x_j + x_{j+\frac{N}{2}} \omega_N^{(2k+1)\frac{N}{2}} \right) \omega_N^{(2k+1)j} \\ &= \sum_{j=0}^{\frac{N}{2}-1} \left(\left(x_j - x_{j+\frac{N}{2}} \right) \omega_N^j \right) \omega_{\frac{N}{2}}^{kj} \\ &= \sum_{j=0}^{\frac{N}{2}-1} z_j \omega_{\frac{N}{2}}^{kj}, \qquad k = 0, 1, \dots, \frac{N}{2} - 1 \,. \end{aligned}$$

Estas dos ecuaciones nos permiten calcular la transformada del vector $\boldsymbol{x}$ a partir de dos transformadas de tamaño $\frac{N}{2}$ efectuadas a los vectores $\boldsymbol{y}$ y $\boldsymbol{z}$ definidos para $j = 0, 1, \dots, \frac{N}{2} - 1$ como

$$\begin{aligned} y_j &= x_j + x_{j+\frac{N}{2}} \,, \\ z_j &= \left(x_j - x_{j+\frac{N}{2}} \right) \omega_N^j \,. \end{aligned}$$

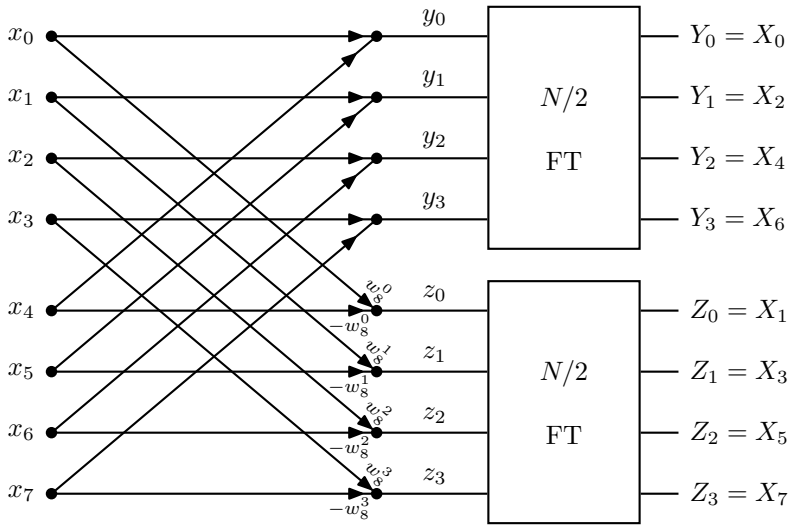
La mariposa correspondiente que sintetiza las combinaciones necesarias para aplicar este método es la siguiente, donde los diferentes elementos que aparecen tienen el mismo significado que para la del algoritmo de Cooley y Tukey.



GRÁFICA 5.3: Mariposa del algoritmo de Gentleman y Sande.

Para calcular los vectores $\boldsymbol{y}$ y $\boldsymbol{z}$ se necesitan N sumas y $\frac{N}{2}$ productos de números complejos, que es el mismo que necesitábamos para el algoritmo de Cooley y Tukey. Por tanto el costo computacional de este algoritmo será el mismo.

Incluimos también un esquema de la primera iteración del método para $N = 8$, que es en cierta forma dual con respecto al de Cooley y Tukey.



GRÁFICA 5.4: Esquema de aplicación de una iteración del algoritmo de Gentleman y Sande para $N = 8$.

Estos dos métodos que hemos descrito se pueden efectuar sobreescribiendo el vector inicial, sin necesidad de más posiciones de memoria (además de los factores correspondientes ω_N^k que tendremos almacenados). Sin embargo, a la vista del ejemplo de aplicación del algoritmo de Gentleman y Sande para $N = 8$ observamos que el vector de salida está desordenado. De la misma forma que para aplicar el algoritmo de Cooley y Tukey necesitamos ordenar de una determinada manera el vector inicial para conseguir que la salida esté en el orden correcto.

Por ejemplo, si efectuamos todas las iteraciones del algoritmo de Gentleman y Sande para el ejemplo anterior con

$$\boldsymbol{x} = (x_0, x_1, x_2, x_3, x_4, x_5, x_6, x_7)\,,$$

el vector de salida estará ordenado de la siguiente forma:

$$\boldsymbol{X} = (X_0, X_4, X_2, X_6, X_1, X_5, X_3, X_7)\,.$$

Ocurre que, si escribimos en código binario estos índices (teniendo en cuenta también los ceros a la izquierda), coinciden con los índices iniciales pero invertidos, es decir, leídos de derecha a izquierda. Este hecho tiene implicaciones importantes a la hora de programar los algoritmos FFT y actualmente existen muchas variantes de estos algoritmos que también tienen en cuenta cómo efectuar de forma eficiente esta reordenación de los vectores (llamada bit-inversión).

Tras introducir los conceptos fundamentales de la transformada de Fourier y algunos métodos que permiten su cálculo de manera eficiente, ahora nos planteamos el caso particular en el que el vector $\boldsymbol{x}$ es real. Si aplicásemos el algoritmo de la FFT directamente estaríamos haciendo muchas operaciones redundantes, ya que las partes imaginarias de los elementos x_j son nulas. Por tanto, es necesario tratar este caso de una forma más eficiente.

Daremos una serie de resultados previos que serán útiles para nuestro propósito. El primer paso será conseguir hacer dos transformadas reales por medio de una única transformada compleja.

5.4. Cálculo de dos FFT reales simultáneamente

Supongamos que queremos hacer dos FFT reales de tamaño N, cuyos datos de entrada llamamos, respectivamente f_l y g_l, para $0 \leq l \leq N-1$. A partir de estos datos creamos un nuevo vector $\boldsymbol{x}$ complejo como sigue:

$$x_l = f_l + ig_l, \qquad 0 \leq l \leq N-1\,.$$

Por definición, las componentes de la transformada de $\boldsymbol{x}$ son:

$$\begin{aligned} X_k &= \sum_{l=0}^{N-1} x_l \omega_N^{kl} \\ &= \sum_{l=0}^{N-1} f_l \omega_N^{kl} + i \sum_{l=0}^{N-1} g_l \omega_N^{kl} \\ &= F_k + iG_k \qquad 0 \leq k \leq N-1\,, \end{aligned} \tag{5.3}$$

donde F_k y G_k son las componentes de las transformadas de $\boldsymbol{f}$ y $\boldsymbol{g}$, respectivamente, que son las que queremos calcular. Recordemos que en general serán valores complejos aunque $\boldsymbol{f}$ y $\boldsymbol{g}$ sean reales.

Nuestro objetivo es recuperar los valores de F_k y G_k a partir de $\boldsymbol{X}$. Para ello utilizaremos la siguiente propiedad de simetría. Aunque en principio los índices de la transformada varían entre 0 y $N-1$, definiremos $F_N = F_0$, $G_N = G_0$:

Proposición 5.3. Si el vector $\boldsymbol{f}$ es real y llamamos $\boldsymbol{F}$ a su transformada, entonces se cumple que

$$\bar{F}_{N-k} = F_k\,, \qquad 0 \leq k \leq N-1\,. \tag{5.4}$$

Demostración. Por la definición de transformada, teniendo en cuenta que los valores f_l son reales:

$$\bar{F}_{N-k} = \sum_{l=0}^{N-1} \bar{f}_l \bar{\omega}_N^{(N-k)l} = \sum_{l=0}^{N-1} \bar{f}_l \bar{\omega}_N^{Nl} \bar{\omega}_N^{-kl} = \sum_{l=0}^{N-1} f_l \omega_N^{kl} = F_k\,, \qquad 0 \leq k \leq N-1\,.$$

□

Análogamente, dado que los g_l son reales, se cumple que $\bar{G}_{N-k} = G_k$ para $0 \leq k \leq N-1$.

Haciendo uso de estas propiedades de simetría se tiene que:

$$\bar{X}_{N-k} = \bar{F}_{N-k} - i\bar{G}_{N-k} = F_k - iG_k\,, \qquad 0 \leq k \leq N-1\,. \tag{5.5}$$

Finalmente, combinando (5.3) y (5.5) llegamos a la importante relación

$$F_k = \frac{1}{2}(X_k + \bar{X}_{N-k}), \qquad G_k = \frac{i}{2}(\bar{X}_{N-k} - X_k)\,, \qquad 0 \leq k \leq N-1\,. \tag{5.6}$$

Por tanto, con estas sumas adicionales de números complejos podemos recuperar las dos FFTs reales después de hacer la FFT compleja. Esto significa que mediante este procedimiento nos podemos ahorrar casi la mitad de las operaciones que habría supuesto hacer directamente las dos FFT.

Como aplicación de este resultado se deduce fácilmente cómo se puede calcular una FFT real.

5.5. Cálculo de una FFT real

Para calcular una FFT cuando los datos x_l son reales, lo primero que hacemos es reducir el problema a calcular dos FFT de tamaño mitad y aplicar el método de la sección anterior.

Como vimos en la proposición (5.2), a partir de las transformadas de los vectores $\boldsymbol{x}^P$ y $\boldsymbol{x}^I$ podemos calcular fácilmente la transformada de $\boldsymbol{x}$. Ahora bien, como los vectores $\boldsymbol{x}^P$ y $\boldsymbol{x}^I$ son reales, podemos calcular su FFT por el método de la sección anterior.

Pero todavía podemos ahorrarnos más cálculos. Dado que $\boldsymbol{x}$ es real, como vimos en (5.4) se cumple que $X_{N-k} = \bar{X}_k$. Esto quiere decir que, a partir de X^P y X^I solo necesitamos calcular a través de la fórmula (5.2) los X_k con $k = 0, 1, ..., \frac{N}{2} - 1$, además de $X_{\frac{N}{2}}$, ya que los demás vienen dados por la propiedad de simetría.

En resumen, vamos a mostrar los pasos a seguir para calcular la FFT de un vector $\boldsymbol{x} = (x_0, ..., x_{N-1})$ real. Sean $\boldsymbol{x}^P$ y $\boldsymbol{x}^I$ los vectores (reales) de tamaño $N/2$ formados por las componentes de $\boldsymbol{x}$ con índice par e impar, respectivamente. Sea $\boldsymbol{y} = \boldsymbol{x}^P + i\boldsymbol{x}^I$.
Los pasos a seguir son:

- Calcular $\boldsymbol{Y}$, la FFT de $\boldsymbol{y}$ de tamaño $N/2$.
- Calcular las componentes de la transformada de los vectores $\boldsymbol{x}^P$ y $\boldsymbol{x}^I$ según las fórmulas

$$X_k^P = \frac{1}{2}(Y_k + \bar{Y}_{\frac{N}{2}-k}), \qquad X_k^I = \frac{i}{2}(\bar{Y}_{\frac{N}{2}-k} - Y_k), \qquad k = 0, 1, ..., \frac{N}{2} - 1\,.$$

  Para ello necesitamos N sumas de números complejos. Además, podemos multiplicar primero $\frac{1}{2}\boldsymbol{Y}$, lo que requiere N productos reales, y después hacer las sumas. El producto por i consiste en permutar las partes real e imaginaria y posteriormente cambiar el signo a la parte real.

- Calcular las componentes $X_0, ..., X_{\frac{N}{2}-1}$ como

$$X_k = X_k^P + \omega_N^k X_k^I, \quad k = 0, 1, ..., \frac{N}{2} - 1\,.$$

  Esto requiere $\frac{N}{2}$ productos y $\frac{N}{2}$ sumas de números complejos.

- Calcular $X_{\frac{N}{2}} = X_0^P - X_0^I$, que requiere una suma de números complejos.
- Calcular las componentes $X_{\frac{N}{2}-1}, ..., X_{N-1}$ a través de la propiedad de simetría:

$$X_{N-k} = \bar{X}_k, \qquad k = 1, 2, ..., \frac{N}{2} - 1\,.$$

Entonces, si N es grande, se pueden ahorrar casi la mitad de las operaciones aritméticas haciendo la FFT de $N/2$ números complejos en lugar de hacer la FFT de N números reales directamente considerándolos complejos sin ninguna simplificación.

Capítulo 6

Ejemplo de aplicación práctica

A continuación vamos a utilizar los métodos pseudoespectrales para programar la resolución numérica de la ecuación KdV. Estudiaremos el comportamiento de las soluciones llamadas "solitones" y la interacción entre ellos. Para conocer qué son los solitones y dónde aparecen este tipo de soluciones introduciremos una breve e interesante reseña histórica sobre el tema.

6.1. La ecuación KdV y los solitones

Las soluciones solitónicas de la ecuación KdV son de las más estudiadas desde el punto de vista de las Matemáticas y la Física. Su historia comienza con las observaciones de John Scott Russell en la primera mitad del siglo XIX, que observó en el canal de Edimburgo un comportamiento curioso de una onda en la superficie del canal. En un artículo publicado en 1844 describe cómo siguió a caballo el recorrido de esta onda durante más de un kilómetro por el canal, sin aparente variación de su forma o su velocidad:

> "I was observing the motion of a boat which was rapidly drawn along a narrow channel by a pair of horses, when the boat suddenly stopped - not so the mass of water in the channel which it had put in motion; it accumulated round the prow of the vessel in a state of violent agitation, then suddenly leaving it behind, rolled forward with great velocity, assuming the form of a large solitary elevation, a rounded, smooth and well-defined heap of water, which continued its course along the channel apparently without change of form or diminution of speed. I followed it on horseback, and overtook it still rolling on at a rate of some eight or nine miles an hour, preserving its original figure some thirty feet long and a foot to a foot and a half in height. Its height gradually diminished, and after a chase of one or two miles I lost it in the windings of the channel.
>
> SCOTT RUSSELL - *Report on Waves* [22]

Russell llevó a cabo experimentos de generación de ondas solitarias en canales y dedujo empíricamente que la velocidad de la onda crecía con la amplitud de la misma y con la profundidad del

canal. Posteriormente, Boussinesq y Rayleigh en los años 70 del siglo XIX estudiaron este comportamiento teóricamente partiendo de las ecuaciones de los fluidos incompresibles y corroboraron los resultados de los experimentos de Russell.

Finalmente, Korteweg & de Vries en 1895 dedujeron la ecuación que debían satisfacer estas ondas, y que permitió explicar el comportamiento de estas ondas solitarias. Hallaron que el prefil de estas ondas se podía expresar matemáticamente como una secante hiperbólica al cuadrado.

En sus experimentos, Russell también observó cómo interaccionaban dos de estas ondas solitarias. Comprobó que si una onda alcanzaba a otra, tras interaccionar, cada una recuperaba su forma original y se desplazaba con la misma velocidad que al principio. Ya en el siglo XX el desarrollo de la informática permitió la resolución numérica de esta ecuación y se comprobó este comportamiento no solo en la interacción entre dos ondas solitarias, sino también la interacción de una onda solitaria y otro perfil. Debido a este comportamiento, por el que parece que las ondas solitarias mantienen su identidad tras interaccionar, se llamó a estas ondas "solitones" (por analogía con fotón, protón, etc.).

Hay varias formas de expresar la ecuación KdV, todas ellas equivalentes, correspondientes a cambios de escala temporal, espacial o de la función de onda. La forma que presentaremos aquí es la tratada en el apartado correspondiente al estudio de los métodos espectrales con $\alpha = 1$, es decir:

$$u_t + uu_x + u_{xxx} = 0\,.$$

6.2. Soluciones de ondas solitarias

Nos disponemos a obtener las soluciones de ondas solitarias para la ecuación KdV. Para ello, en primer lugar buscaremos soluciones de la forma $u(x,t) = f(\xi)$, donde $\xi = x - ct$, es decir, ondas viajeras que se desplazan con velocidad constante c.

De esta forma, la ecuación KdV se transforma en

$$-cf' + ff' + f''' = 0\,.$$

Integrando una vez esta ecuación obtenemos

$$-cf + \frac{f^2}{2} + f'' = A\,,$$

donde A es una constante arbitraria. Si ahora multiplicamos por f', la ecuación resultante se puede integrar otra vez, resultando

$$-\frac{c}{2}f^2 + \frac{1}{6}f^3 + \frac{1}{2}(f')^2 = Af + B\,,$$

donde B es otra constante arbitraria. En este punto, en lugar de estudiar el problema de buscar ondas viajeras en general, para buscar ondas solitarias suponemos que tanto f como sus derivadas tienden a 0 cuando $\xi \to \pm\infty$. En este caso, de las dos ecuaciones anteriores deducimos que $A = B = 0$, por lo que nos queda la ecuación diferencial

$$(f')^2 = f^2\left(c - \frac{1}{3}f\right)\,.$$

Por tanto, para que exista una solución de onda solitaria se debe cumplir que $c - \frac{1}{3}f \geq 0$.

Al tratarse de una ecuación diferencial de primer orden de variables separadas, podemos integrarla fácilmente:

$$\int \frac{df}{f\sqrt{c - \frac{1}{3}f}} = \pm \int d\xi \, .$$

Si probamos con el cambio de variable $f = 3c\,\text{sech}^2\theta$ obtenemos

$$\int \frac{-6c\tanh\theta\,\text{sech}^2\theta\,d\theta}{3c\,\text{sech}^2\theta\sqrt{c - c\,\text{sech}^2\theta}} = \pm \int d\xi \, .$$

Como $1 - \text{sech}^2\theta = \tanh^2\theta$, la integral se simplifica y solo queda la expresión

$$\int \frac{-2c\tanh\theta\,d\theta}{\sqrt{c}\,|\tanh\theta\,|} = \pm \int d\xi \, .$$

Observamos que, como la secante hiperbólica es una función par, el signo de ξ no influye en f. Por tanto, podemos simplificar directamente el cociente de la primera integral y obtener que

$$\theta = \frac{\sqrt{c}}{2}(x - x_0 - ct) \, ,$$

donde x_0 es la constante de integración, que tiene el significado del punto en el que se encuentra centrada la onda en el instante inicial.

Por tanto las soluciones de ondas solitarias para la ecuación KdV se expresan como

$$u(x,t) = 3c\,\text{sech}^2\left(\frac{\sqrt{c}}{2}(x - x_0) - \frac{c^{3/2}}{2}t\right) .$$

Otra forma habitual de expresar esta solución, teniendo en cuenta que $c \geq 0$ es hacer el cambio de parámetro $c = A^2$ con $A > 0$, de forma que la solución se escribe

$$u(x,t) = 3A^2\text{sech}^2\left(\frac{A}{2}(x - x_0) - \frac{A^3}{2}t\right) . \tag{6.1}$$

La amplitud de esta onda es $3A^2$, mientras que su velocidad es A^2, es decir, la velocidad es proporcional a la amplitud (comportamiento que ya hemos comentado que fue estudiado por Russell). Este comportamiento difiere respecto al caso de las ondas para las ecuaciones lineales, en las cuales la velocidad y la amplitud son independientes. Además se trata de ondas localizadas en el espacio, ya que el valor de u decae rápidamente en el espacio al alejarse del punto central de la onda, $x = x_0 + A^2t$. Por ello podremos suponer la periodicidad de la red, como hemos comentado anteriormente.

El comportamiento característico de estas ondas (llamadas solitones) de la ecuación KdV proviene de un delicado equilibrio entre el término dispersivo u_{xxx} de la ecuación y el término no lineal uu_x que da lugar a las conocidas leyes de conservación (ver [7]). Es por ello que necesitamos discretizar la ecuación con una precisión suficientemente alta para para estudiar problemas como la interacción

de solitones. De lo contrario, con una precisión baja se apreciarían efectos de pérdida de amplitud o de oscilaciones en los solitones de la solución numérica. Esto justifica la elección de los métodos espectrales o pseudoespectrales, ya que el orden de convergencia exponencial nos permite alcanzar precisión suficiente con un espaciado de red no muy pequeño. Elegiremos los métodos pseudoespectrales por la eficiencia en el tratamiento del término no lineal uu_x mediante la transformada rápida de Fourier, como veremos a continuación.

6.3. Programación del método

En este apartado programaremos en Matlab la resolución numérica de la ecuación KdV. Las aproximaciones para las derivadas espaciales las efectuaremos con métodos pseudoespectrales y para avanzar en tiempo utilizaremos el método de Runge-Kutta clásico de cuatro etapas y orden 4.

Para el planteamiento del método podemos utilizar la formulación que empleamos en el Capítulo 4, consistente en encontrar una función $u_N : [0,T] \to \mathcal{S}_N$. Pero esta formulación es más adecuada para el análisis del método que para la implementación, porque lo que vamos a calcular en un ordenador van a ser funciones definidas en una red de nodos, no funciones definidas en $\mathbb{R}$. Por tanto, la formulación habitual en estos casos es la siguiente:

El método pseudoespectral continuo en tiempo para el problema

$$\begin{cases} u_t + u_{xxx} + \dfrac{1}{2}(u^2)_x = 0\,, \\ u(x,t) = u(x+2\pi,t), \qquad x \in \mathbb{R},\ t \in [0,T]\,, \\ u(x,0) = q(x), \qquad x \in \mathbb{R}\,, \end{cases} \tag{6.2}$$

consiste en encontrar una aplicación $\boldsymbol{U} : [0,T] \mapsto \mathbb{Z}_h$ tal que $\boldsymbol{U}(0)$ es una aproximación a la restricción a la red de q y

$$\frac{d}{dt}\boldsymbol{U}(t) + D^3\boldsymbol{U}(t) + \frac{1}{2}D\boldsymbol{U}^2(t) = \boldsymbol{0}, \qquad t \in [0,T]\,. \tag{6.3}$$

Es decir, $\boldsymbol{U}(t)$ aproxima la restricción a la red de la solución $u(\cdot,t)$. Podemos expresar la ecuación de manera equivalente, separando el término de la derivada temporal:

$$\frac{d}{dt}\boldsymbol{U}(t) = -D^3\boldsymbol{U}(t) - \frac{1}{2}D\boldsymbol{U}^2(t), \qquad t \in [0,T]\,,$$

donde las aplicaciones del operador de diferenciación espectral las calculamos mediante la transformada rápida de Fourier. Para ello empleamos la función de FFT que incluye Matlab, la cual en las últimas versiones del programa tiene en cuenta si los datos de entrada son reales, como hemos descrito en el capítulo anterior. A esta ecuación le aplicamos el método Runge Kutta de cuarto orden como sigue, con un espaciado temporal Δt. Si llamamos $\boldsymbol{U}_n$ a la aproximación de la solución en el instante de tiempo $t_n = n\Delta t$, entonces la aplicación del método para avanzar un paso resulta en los

cálculos de las pendientes $\boldsymbol{k}_i$, para $i = 1, \dots, 4$

$$\begin{aligned}
\boldsymbol{k}_1 &= f(\boldsymbol{U}_n)\,, \\
\boldsymbol{k}_2 &= f\left(\boldsymbol{U}_n + \frac{1}{2}\boldsymbol{k}_1\Delta t\right)\,, \\
\boldsymbol{k}_3 &= f\left(\boldsymbol{U}_n + \frac{1}{2}\boldsymbol{k}_2\Delta t\right)\,, \\
\boldsymbol{k}_4 &= f\left(\boldsymbol{U}_n + \boldsymbol{k}_3\Delta t\right)\,,
\end{aligned}$$

mediante la evaluación de la función

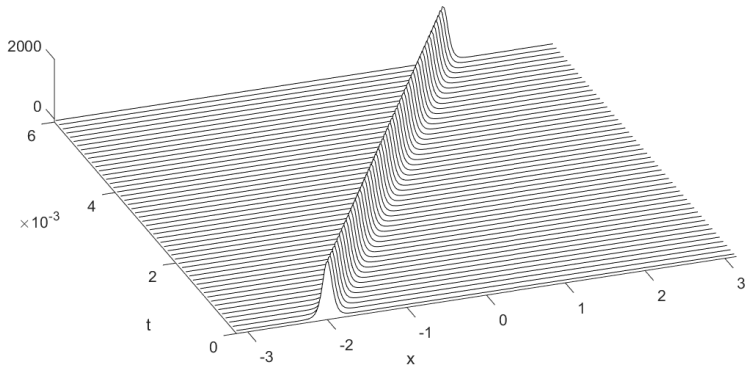
$$\begin{aligned}
f(\boldsymbol{U}) &= -\mathcal{F}^{-1}\left((ik)^3\mathcal{F}(\boldsymbol{U})\right) - \mathcal{F}^{-1}\left((ik)\mathcal{F}\left(\frac{\boldsymbol{U}^2}{2}\right)\right) \\
&= -\mathcal{F}^{-1}\left((ik)^3\mathcal{F}(\boldsymbol{U}) + (ik)\mathcal{F}\left(\frac{\boldsymbol{U}^2}{2}\right)\right)\,.
\end{aligned}$$

Aquí hemos denotado como producto por ik las multiplicaciones que tenemos que hacer componente a componente para cada coeficiente de la transformada como $ik\hat{u}_k$, salvo la excepción ya comentada de $k = N/2$ al tratarse de derivadas de orden impar. Posteriormente se combinan las pendientes de la forma

$$\boldsymbol{U}_{n+1} = \boldsymbol{U}_n + \frac{1}{6}(\boldsymbol{k}_1 + 2\boldsymbol{k}_2 + 2\boldsymbol{k}_3 + \boldsymbol{k}_4)\,.$$

El código en Matlab del programa se incluye en el Apéndice (Programa 1). Imponemos la condición inicial de un solitón, de acuerdo a la ecuación (6.1) con $A = 25$:

$$u(x,0) = 3 \cdot 25^2\,\mathrm{sech}^2\left(\frac{25}{2}(x+2)\right)\,.$$



GRÁFICA 6.1: Solitón de la ecuación KdV.

6.4. Programación con factor integrante

Otra opción más inteligente para programar el método es hacer uso de un factor integrante. Veamos cómo podemos aplicar esta técnica.

Si partimos de la ecuación KdV transformada

$$\hat{u}_t + \frac{i}{2}k\hat{u^2} - ik^3\hat{u} = 0\,,$$

y multiplicamos por e^{-ik^3t} tenemos

$$e^{-ik^3t}\hat{u}_t + \frac{i}{2}e^{-ik^3t}k\hat{u^2} - ie^{-ik^3t}k^3\hat{u} = 0\,.$$

Ahora definimos la nueva variable $\hat{v} = e^{-ik^3t}\hat{u}$. Entonces se cumple que $\hat{v}_t = -ik^3\hat{v} + e^{-ik^3t}\hat{u}_t$, y podemos simplificar la ecuación hasta obtener

$$\hat{v}_t + \frac{i}{2}e^{-ik^3}k\hat{u^2} = 0\,.$$

Hemos conseguido eliminar el término lineal. Esto tiene importantes consecuencias prácticas a la hora de implementar la resolución, ya que hemos eliminado el carácter rígido del problema y lo hemos transformado en un problema con un coeficiente de variación rápida. En este caso, para discretizar el problema hay que tener en cuenta que para calcular $\hat{u^2}$ a partir de $\hat{v}$, primero tenemos que multiplicar $e^{ik^3t}\hat{v}$, hacer la transformada inversa, elevar al cuadrado y por último hacer la transformada de Fourier discreta.

Tras estas consideraciones, la discretización del problema es:

$$\hat{v}_t + \frac{i}{2}e^{-ik^3t}k\mathcal{F}\left(\left(\mathcal{F}^{-1}\left(e^{ik^3t}\hat{v}\right)\right)^2\right) = 0\,.$$

Para programar este método utilizamos el mismo procedimiento que para el anterior. Usaremos también el método de Runge Kutta de cuarto orden para la discretización temporal. A diferencia del caso anterior, ahora nos aparece una dependencia temporal explícita. Esto hace que tengamos que tenerlo en cuenta al calcular las pendientes intermedias.

$$\begin{aligned}
\boldsymbol{k}_1 &= f(t_n, \hat{\boldsymbol{v}}_n)\,,\\
\boldsymbol{k}_2 &= f\left(t_n + \frac{1}{2}\Delta t, \hat{\boldsymbol{v}}_n + \frac{1}{2}\boldsymbol{k}_1\Delta t\right)\,,\\
\boldsymbol{k}_3 &= f\left(t_n + \frac{1}{2}\Delta t, \hat{\boldsymbol{v}}_n + \frac{1}{2}\boldsymbol{k}_2\Delta t\right)\,,\\
\boldsymbol{k}_4 &= f\left(t_n + \Delta t, \hat{\boldsymbol{v}}_n + \boldsymbol{k}_3\Delta t\right)\,.
\end{aligned}$$

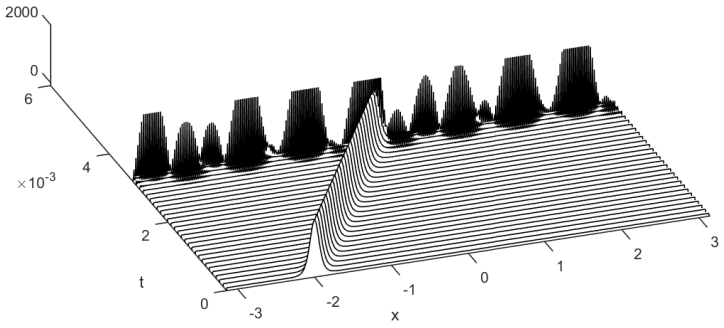
Para programarlo, en lugar de introducir directamente la expresión de la función f, podemos ahorrarnos operaciones si algunas de las exponenciales las multiplicamos en el lugar adecuado. Por ejemplo, cuando calculamos $\boldsymbol{k}_1$ deberíamos multiplicar todo por e^{-ik^3t}, pero para calcular $\boldsymbol{k}_2$ tenemos que multiplicar de nuevo $\boldsymbol{k}_1$ por $e^{ik^3\left(t+\frac{1}{2}\Delta t\right)}$. Y lo mismo ocurre para el cálculo de $\boldsymbol{k}_3$ y $\boldsymbol{k}_4$. Por ello factorizamos $e^{ik^3\frac{1}{2}\Delta t}$ y el otro factor exponencial lo multiplicamos al final.

Además en este caso estamos resolviendo la ecuación para $\hat{v}$, no para u como en el caso anterior. Pero la solución que buscamos es la onda para u. Para simplificar algún cálculo también empleamos que $\hat{u} = e^{ik^3t}\hat{v}$, por lo que no es necesario calcular $\hat{v}$ a partir de u, basta con calcular $\hat{u}$ porque al

evaluar la función se simplifica el factor exponencial (estas simplificaciones se aprecian de forma más clara al examinar el programa que incluimos).

Tras estas consideraciones, modificamos el programa anterior para tener en cuenta este aspecto. El programa se incluye en el Apéndice (Programa 2). Como hemos comentado, este programa permite elegir un paso de integración temporal mayor. Vamos a verificarlo numéricamente a la vista de los resultados obtenidos al ejecutar los programas.

Al igual que ocurre para las ecuaciones lineales, los problemas de estabilidad del método numérico se manifiestan habitualmente con un comportamiento de la solución numérica característico, que explota conforme avanza el tiempo. Si elegimos un intervalo de tiempo demasiado grande el resultado es el mostrado en la Gráfica 6.2. Para realizar dicha representación hemos usado la programación con factor integrante y con los datos $N = 512$ y $\Delta t = 1{,}58 \cdot 10^{-6}$.



GRÁFICA 6.2: Explosión de la solución numérica.

Para estudiar numéricamente la dependencia aproximada entre el espaciado de la red temporal y espacial para conservar la estabilidad ejecutamos el programa con un valor de N fijo y variamos el espaciado temporal hasta hallar el valor máximo que mantiene la estabilidad de la solución en el intervalo temporal que se representa. No se trata de un análisis riguroso de la estabilidad de la implementación, simplemente observamos de forma gráfica el comportamiento de la solución para deducir de forma aproximada la dependencia entre los espaciados de la red temporal y espacial. Por ejemplo, para el programa sin factor integrante, los pasos de integración temporal máximos que podemos tomar de forma que no explote la solución numérica dentro del intervalo temporal representado son, para distintos valores de N:

$\boldsymbol{N}$	$\boldsymbol{\Delta t_{max}}$
128	$1{,}141 \cdot 10^{-5}$
256	$1{,}384 \cdot 10^{-6}$
512	$1{,}706 \cdot 10^{-7}$

TABLA 6.1: Paso de integración temporal máximo para la programación sin factor integrante en función de N.

Observamos que al aumentar un factor 2 el valor de N, debemos dividir aproximadamente por un factor 2^3 el espaciado temporal. Esta es la razón por la que en el Programa 1 hemos calculado Δt de forma proporcional a N^{-3}. Haciendo lo mismo para el programa con factor integrante obtenemos:

$\boldsymbol{N}$	$\boldsymbol{\Delta t_{max}}$
256	$6{,}90 \cdot 10^{-6}$
512	$1{,}57 \cdot 10^{-6}$
1024	$3{,}52 \cdot 10^{-7}$

TABLA 6.2: Paso de integración temporal máximo para la programación con factor integrante en función de N.

Aquí ocurre algo distinto. En lugar de dividirse por 2^3 el espaciado temporal, aproximadamente se divide por 2^2. Es decir, la dependencia del espaciado temporal con el número de nodos espaciales es N^{-2}. Por tanto, aquí observamos numéricamente la importancia del factor integrante en términos de la estabilidad del método.

Para continuar con el análisis numérico del método, estudiaremos la dependencia del error global con el espaciado temporal y espacial. Para este análisis utilizaremos la programación con factor integrante ya que reduce notablemente el tiempo de computación y nos permite abarcar un rango más amplio de valores de los parámetros espacial y temporal.

En primer lugar, fijaremos un intervalo temporal determinado y variaremos N para estudiar cómo depende el error con este parámetro. El error global lo calculamos como la norma infinito de la diferencia entre la solución numérica y la solución exacta (6.1) para el instante de tiempo en el que se calcula la última aproximación.

$\boldsymbol{N}$	error global
1024	$1{,}6657 \cdot 10^{-4}$
512	$1{,}6657 \cdot 10^{-4}$
256	0,0011
128	335,5980

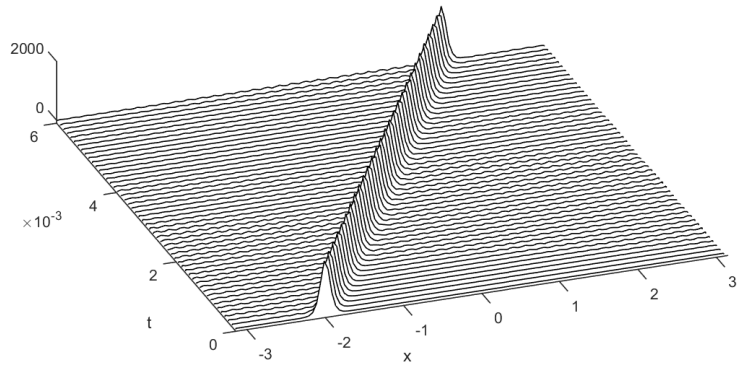
TABLA 6.3: Dependencia del error global con N para $\Delta t = 3{,}52 \cdot 10^{-7}$.

Observamos que para $N = 512$ y 1024 el error no varía. Esto es debido a que para estos valores de N, el error se debe principalmente a la discretización temporal y no a la espacial. Sin embargo, para valores de N más pequeños comienza a reflejarse el efecto de la discretización espacial.

Llama la atención el fuerte descenso del error que se produce entre los valores de $N = 128$ y $N = 256$. Con tan solo duplicar el número de nodos espaciales, el error global disminuye en nada menos que 6 órdenes de magnitud. Este comportamiento es característico de los métodos espectrales y pseudoespectrales y es una de las principales ventajas que presentan frente a otros métodos.

Aunque el valor del error obtenido para $N = 128$ parezca muy grande, en realidad no lo es tanto comparado con la altura del pico de la onda, que está en torno a 2000. De hecho, si dibujamos la

gráfica correspondiente a este caso, la solución de onda viajera se sigue apreciando con nitidez, pero ahora aparecen unas oscilaciones que hacen aumentar el error:



GRÁFICA 6.3: Aparición de oscilaciones en la solución numérica.

Por último, estudiamos la dependencia del error para N fijo y variando el intervalo temporal de integración:

Δt	error global
$3{,}52 \cdot 10^{-7}$	$1{,}6657 \cdot 10^{-4}$
$1{,}76 \cdot 10^{-7}$	$7{,}0631 \cdot 10^{-6}$
$8{,}8 \cdot 10^{-8}$	$3{,}339 \cdot 10^{-7}$

TABLA 6.4: Dependencia del error global con Δt para $N = 1024$.

En este caso hemos dividido sucesivamente el intervalo temporal por 2, obteniendo que el error global aproximadamente se divide por 20. Este comportamiento es el que cabría esperar ya que para la integración temporal estamos usando un método Runge Kutta de orden 4, en el que el error global decrece aproximadamente como $(\Delta t)^4$.

6.5. Interacción entre solitones

La ecuación KdV, además de los solitones, posee un tipo de soluciones que, asintóticamente cuando $t \to \pm\infty$ se comportan como la suma de varios solitones del tipo (6.1). Como ya vimos que la velocidad es proporcional a la amplitud del solitón, los solitones se desplazarán a diferente velocidad y se alcanzarán los unos a los otros. Pero la peculiaridad es que después de interaccionar mantienen su forma original y el único efecto que pone de manifiesto la interacción es un cambio de fase de los solitones. Este comportamiento de las soluciones se puede estudiar matemáticamente con el llamado método de scattering inverso, introducido en el año 1967 por Gardner, Greene, Kruskal y Miura [11].

Realizando una transformación a la ecuación KdV, llamada transformación de Miura, nos permite asociar a esta un problema físico de scattering, cuya representación matemática se realiza con una

ecuación de Sturm-Liouville del tipo

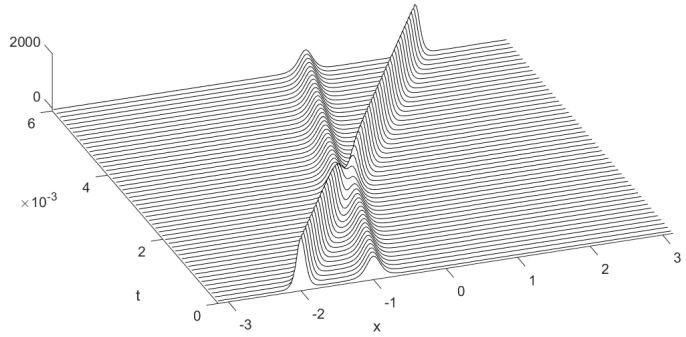
$$-\Psi_{xx} + u\Psi = \lambda\Psi\,.$$

Y estudiando las propiedades de scattering de este problema, como son los coeficientes de reflexión y transmisión, se puede estudiar el comportamiento de las soluciones de la ecuación KdV. De esta forma, el método de scattering inverso permite mostrar que existen soluciones de la ecuación KdV que asintóticamente se comportan como suma de solitones cuando $x \to \pm\infty$.

Para estudiar la interacción entre dos solitones propondremos la condición inicial:

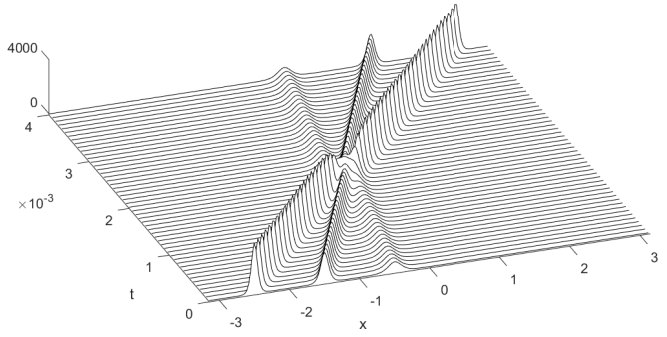
$$u(x,0) = 3A^2\text{sech}^2\left(\frac{A}{2}(x-x_0)\right) + 3B^2\text{sech}^2\left(\frac{B}{2}(x-x_0')\right)\,. \tag{6.4}$$

El resultado obtenido al ejecutar el programa con dicha condición inicial se presenta en la siguiente gráfica:



GRÁFICA 6.4: Interacción entre dos solitones de la ecuación KdV.

Observamos que, tras interaccionar, los solitones mantienen su amplitud original (y, por tanto, su velocidad). Si hacemos lo mismo para 3 solitones que interaccionan en el mismo punto, el efecto de la interacción es análogo: recuperan su forma y sufren un cambio de fase.



GRÁFICA 6.5: Interacción entre tres solitones de la ecuación KdV.

6.6. Conclusiones

A lo largo de este trabajo hemos expuesto los fundamentos teóricos y prácticos de los métodos espectrales y pseudoespectrales, aplicándolos a la ecuación KdV. Como hemos visto en este último capítulo, la programación de estos métodos en Matlab es sencilla y muy eficiente al contar con un algoritmo de FFT ya programado. Hemos conseguido hacer simulaciones muy precisas de la interacción entre solitones, que son problemas que requieren una precisión elevada.

Un aspecto que no hemos desarrollado en este trabajo y que sería la continuación natural del estudio de los métodos espectrales y pseudoespectrales es la elección del integrador temporal. En la programación que hemos realizado hemos optado por un método Runge Kutta, pero hemos constatado las restricciones de estabilidad que presentan, siendo necesario un paso de integración temporal muy pequeño. Aunque con la introducción del factor integrante hemos conseguido aumentar el paso de integración temporal, la restricción todavía sigue siendo importante.

Una futura línea de trabajo podrían ser los métodos llamados *exponenciales*, que son una interesante clase de métodos numéricos para la integración de sistemas de ecuaciones diferenciales rígidos de la forma

$$u'(t) = F(t, u(t)), \quad u(0) = u_0\,. \tag{6.5}$$

La idea básica de estos métodos consiste en linealizar (6.5) en torno a un cierto w. Si el sistema es autónomo obtenemos

$$v'(t) + Av(t) = g(v(t))\,, \quad v(0) = u_0 - w\,, \tag{6.6}$$

con $A = -DF(w)$ y $v(t) = u(t) - w$. Esta linealización proporciona una ecuación semilineal, cuya solución habitualmente se expresa mediante la ecuación integral de Volterra que se obtiene mediante variación de los parámetros:

$$u(t) = e^{-tA}u_0 + \int_0^t e^{-(t-\tau)A} g(u(\tau))\, d\tau\,. \tag{6.7}$$

Los integradores exponenciales parten de una discretización de esta ecuación y pueden estar basados en discretizaciones Runge-Kutta, métodos lineales multipaso u otros métodos. Por tanto en los métodos exponenciales el término lineal se integra exactamente, lo que contribuye a reducir los problemas de estabilidad de los esquemas utilizados (en nuestro caso el término lineal de la ecuación KdV sería el término u_{xxx}, discretizado como $D^3\boldsymbol{U}$ en la ecuación (6.3)). Se puede consultar [13] para una descripción detallada de estos métodos.

Por último, mencionar que también se podría continuar este trabajo estudiando más en profundidad los métodos espectrales. En particular, cómo aplicar los métodos de Galerkin para tratar de forma eficiente los términos no lineales, que en el método Galerkin original incrementan el costo computacional de forma notable. También se podrían estudiar otro tipo de métodos espectrales, como los métodos *tau*, que se pueden entender como una modificación de los métodos de Galerkin y son aplicables a problemas con condiciones frontera no periódicas.

Apéndices

Apéndice A

Programas en Matlab

A continuación incluimos los programas desarrollados en Matlab para las representaciones gráficas que se han ido adjuntando a lo largo del presente Trabajo Fin de Grado en Matemáticas.

Programa 1

Calcula y representa la solución numérica de la ecuación KdV mediante el método pseudoespectral con la condición inicial

$$u(x,0) = 3 \cdot 25^2 \operatorname{sech}^2\left(\frac{25}{2}(x+2)\right).$$

Véase la Gráfica 6.1.

```
%Malla y condicion inicial
N = 512; dt = 20/N^3; x = (2*pi/N)*(-N/2:N/2-1)';
A = 25; u = 3*A^2*sech(.5*(A*(x+2))).^2;
k = [0:N/2-1 0 -N/2+1:-1]'; ik3 = 1i*k.^3; ik=1i*k;
% Resolucion numerica:
tmax = 0.006; nplt = floor((tmax/50)/dt); nmax = round(tmax/dt);
udata = u; tdata = 0; h = waitbar(0,'please wait...');
for n = 1:nmax
    t = n*dt;
    % Metodo RK4:
    k1 = -real(ifft(ik.*fft(       0.5*u.^2      )...
          -ik3.*fft(      u      )));
    k2 = -real(ifft(ik.*fft(0.5*(u+dt*k1/2).^2)...
          -ik3.*fft(u+dt*k1/2)));
    k3 = -real(ifft(ik.*fft(0.5*(u+dt*k2/2).^2)...
          -ik3.*fft(u+dt*k2/2)));
    k4 = -real(ifft(ik.*fft( 0.5*(u+dt*k3).^2  )...
          -ik3.*fft( u+dt*k3 )));
    u  =  u + dt*(k1 + 2*(k2+k3) + k4)/6;
    if mod(n,nplt) == 0
      waitbar(n/nmax)
      udata = [udata u]; tdata = [tdata t];
```

```
    end
end
waterfall(x,tdata,udata'), colormap([0 0 0]); view(-20,25)
xlabel x, ylabel t, axis([-pi pi 0 tmax -100 2000]), grid off
set(gca,'ztick',[0 2000]), close(h), pbaspect([1 1 .13])
```

Programa 2

Programación del método pseudoespectral para la ecuación KdV con factor integrante. Véase la sección 6.4.

```
%Malla y condicion inicial
N = 256; dt = 0.4/N^2; x = (2*pi/N)*(-N/2:N/2-1)';
A = 25; u = 3*A^2*sech(.5*(A*(x+2))).^2;
w = fft(u); k = [0:N/2-1 0 -N/2+1:-1]'; ik3 = 1i*k.^3;
% Resolucion numerica:
tmax = 0.006; nplt = floor((tmax/50)/dt); nmax = round(tmax/dt);
udata = u; tdata = 0; h = waitbar(0,'please wait...');
for n = 1:nmax
  t = n*dt;
  % Metodo RK4:
  g = -.5i*dt*k;
  E = exp(dt*ik3/2); E2 = E.^2;
  k1 = g.*fft(real( ifft(     w     ).^2));
  k2 = g.*fft(real( ifft(E.*(w+k1/2)).^2));
  k3 = g.*fft(real( ifft(E.*w + k2/2) .^2));
  k4 = g.*fft(real( ifft(E2.*w+E.*k3).^2));
  w = E2.*w + (E2.*k1 + 2*E.*(k2+k3) + k4)/6;
  if mod(n,nplt) == 0
    u = real(ifft(w)); waitbar(n/nmax)
    udata = [udata u]; tdata = [tdata t];
  end
end
waterfall(x,tdata,udata'), colormap([0 0 0]); view(-20,25)
xlabel x, ylabel t, axis([-pi pi 0 tmax -100 2000]), grid off
set(gca,'ztick',[0 2000]), close(h), pbaspect([1 1 .13])
```

Bibliografía

[1] L. Abia, J. M. Sanz-Serna, *A spectral method for a nonlinear equation arising in fluidized bed modelling.* Numerical Treatment of Differential Equations, Proceedings of the Fifth Seminar "NUMDIFF-5" held in Halle, 1989, Karl Strehmel (editor). Leipzig: Teubner.

[2] L. Abia, J. M. Sanz-Serna, *The spectral accuracy of a fully-discrete scheme for a nonlinear third order equation.* Computing **44**, 187-196 (1990).

[3] R. A. Adams, J. J. F. Fournier, *Sobolev spaces*. (Elsevier, Oxford, 2003).

[4] C. Canuto, M. Y. Hussaini, A. Quarteroni, T. A. Zang, *Spectral Mehods. Fundamentals in Single Domains.* (Springer, Berlin, 2006).

[5] E. Chu, A. George, *Inside the FFT black box. Serial and Parallel Fast Fourier Transform Algorithms*. (CRC Press, New York, 2000).

[6] J. W. Cooley, J. W. Tuckey, *An algorithm for the machine calculation of complex Fourier series.* Math. Comput. **19**, 297-301 (1965).

[7] P. G. Drazin, R. S. Johnson, *Solitons: an introduction.* (Cambridge University Press, Cambridge, 1989).

[8] G. B. Folland, *Fourier analysis and its applications.* (Brooks/Cole Publishing Company, California, 1992).

[9] B. Fornberg, *On a Fourier method for the integration of hyperbolic equations*. SIAM J. Numer. Anal. **12**, 509-528 (1975).

[10] B. Fornberg, *A practical guide to pseudospectral methods*. (Cambridge University Press, Cambridge, 1998).

[11] C. S. Gardner, J. M Greene, M. D. Kruskal, R. M. Miura, *Method for solving the Korteweg de Vries equation.* Phys. Rev. Lett. **19**, 1095-1097 (1967).

[12] D. Gottlieb, S. A. Orszag, *Numerical Analysis of Spectral Methods: Theory and Applications.* (SIAM, Philadelphia, 1977).

[13] M. Hochbruck, A. Ostermann, *Exponential integrators.* Acta Numerica **19**, 209-286 (2010).

[14] D. J. Korteweg, G. de Vries, *On the change of form of long waves advancing in a rectangular canal, and on a new type of long stationary waves.* Phil. Mag. (5) **39**, 422–443 (1895).

[15] J. L. Lions, E. Magenes, *Non-homogeneous boundary value problems and applications,* Vol. 1. (Springer Verlag, Berlin, 1972).

[16] Y. Maday, A. Quarteroni, *Error analysis for spectral approximation of the Korteweg-de Vries equation*. Modélisation mathématique et analyse numérique **3**, 499-529 (1988).

[17] W. Malfliet, *Solitary wave solutions of nonlinear wave equations*. Am. J. Phys. **60**, 650-654 (1992).

[18] V. B. Matveev, M. A. Salle, *Darboux Transformations and Solitons.* (Springer, Berlin, 1991).

[19] R. M. Miura, *The Korteweg-de Vries Equation: A survey or Results.* SIAM Rev. **18**, 412–459 (1976).

[20] A. V. Oppenheimer, R. W. Schafer, *Discrete-Time Signal Processing*. (Prentice-Hall, New Jersey, 1999).

[21] W. Rudin, *Real and Complex Analysis.* (Mc. Graw-Hill, New York, 1987).

[22] J. S. Russell, *Report on Waves*. Rep. 14th Meet. Brit. Assoc. Adv. Sci. (J. Murray, London, 1844) 311–390.

[23] E. Tadmor, *The exponential accuracy of Fourier and Chebyshev differencing methods*. SIAM J. Numer. Anal. **23**, 1-10 (1986).

[24] R. Témam, *Sur un problème non linéaire.* J. Math. Pures Appl. **48**, 159-172 (1969).

[25] L. N. Trefethen, *Spectral methods in MATLAB.* (SIAM, Philadelphia, 2000).