Interpolación polinómica de alto orden Métodos espectrales

Aplicación a problemas de contorno y de condiciones iniciales

Juan A. Hernández Ramos
Mario A. Zamecnik Barros

Departamento de Matemática Aplicada a la Ingeniería Aeroespacial
Escuela Técnica Superior de Ingeniería Aeronáutica y del Espacio
Universidad Politécnica de Madrid

Portada:

Conquistar la cima de tu propósito, requiere confianza y perseverancia.

“Ascensión al Pico del Lobo”(febrero de 2019).

Fotografía y diseño de cubierta por Belén Moreno Santamaría.

ISBN 978-1076625595

Índice general

Prefacio

El uso de esquemas numéricos de alto orden para la simulación o integración de problemas físicos es hoy en día un tema de mucha importancia. Las dos técnicas mayormente empleadas son: métodos espectrales e interpolación polinómica de alto orden. Sin embargo, existen problemas asociados al mal condicionamiento numérico de la semidiscretización espacial cuando se utilizan interpolantes de alto orden.

El objetivo de este libro es poner de manifiesto la importancia de la interpolación en el resultado numérico obtenido por los diferentes esquemas numéricos o métodos. En concreto, se muestra que la concentración de la nube de puntos en los extremos del intervalo o dominio de integración es vital para el éxito de los métodos de interpolación de alto orden. Aunque todo el material que aparece en este libro es bien conocido, los autores han querido recoger de forma didáctica y ordenada algunos de los aspectos más relevantes para entender el origen de los problemas en la integración con los esquemas de alto orden.

En el capítulo 1 se hace una clasificación de los problemas objeto de estudio. En los problemas de evolución, la variable independiente temporal permite clasificar matemáticamente los problemas en: hiperbólicos o parabólicos. La evolución temporal de estos sistemas permite obtener la solución numérica aproximada en función de instantes anteriores. Es decir, a partir de la condición inicial, se puede determinar la solución aproximada en el siguiente paso de integración y así sucesivamente. Cuando los problemas no son de evolución, generalmente, son problemas elípticos en los que la información se propaga de manera instantánea y los algoritmos de cálculo son completamente diferentes. En concreto, una vez discretizado el problema espacialmente, la solución se debe obtener mediante la resolución de un sistema lineal o no lineal de ecuaciones. Aunque gran parte del contenido de este libro está dedicado a la interpolación, el principal objetivo de este capítulo es poner el foco en la simulación de los problemas gobernados por sistemas de ecuaciones con condiciones iniciales y de contorno.

En el capítulo 2 se hace una revisión somera de las técnicas de aproximación en espacios de dimensión finita e infinita. Se define el conjunto de funciones base del espacio junto con su producto interno. El uso del ordenador para la aproximación de funciones exige siempre el tratamiento en espacios de dimensión finita. Es por

esta razón que la idea principal de este capítulo es relacionar el error de aproximación de una serie discreta vinculada a un espacio de dimensión finita y una serie truncada vinculada a un espacio de dimensión infinita. La serie discreta se obtiene haciendo pasar el interpolante por una serie de puntos discretos o de colocación. El mayor atractivo de los métodos espectrales o desarrollos en serie en espacios de dimensión infinita reside en la precisión espectral. La precisión espectral está asociada a desarrollos de funciones infinitamente derivables mediante ciertas bases ortogonales. La idea principal de la precisión espectral es que mediante unos cuantos grados de libertad o términos del desarrollo en serie, el error de aproximación se hace extremadamente pequeño. Por esta razón, se busca que la aproximación mediante una serie discreta en un espacio de dimensión finita goce de propiedades similares a la precisión espectral. Para ello, se revisan las fórmulas de cuadratura de Gauss que permiten realizar una integral de manera exacta de ciertos polinomios mediante un conjunto de puntos de colocación. Si se eligen estos puntos de colocación de las cuadraturas de Gauss como puntos de interpolación en el espacio de dimensión finita, se demuestra que el producto interno coincide con el producto interno en dimensión infinita y la propiedad de precisión espectral se puede llevar a los espacios de dimensión finita. Es decir, siempre que la nube de puntos de interpolación o discretización se corresponda con los puntos gaussianos, la aproximación con una serie discreta y con una serie truncada se comportan asintóticamente con el mismo error.

En el capítulo 3 se hace un estudio en detalle de las diferentes técnicas de interpolación polinómica: método de Vandermonde, interpolación de Lagrange y fórmula de Newton. A continuación, se define el error de interpolación como la suma del error de truncamiento más el error de redondeo y se entra en el origen de ambos errores. Se demuestra que el error de truncamiento está gobernado por la función de error $\pi(x)$ y se analiza su forma con la distribución de la nube de puntos. Para distribuciones con puntos equiespaciados, se observa la presencia de un máximo de la función $\pi(x)$ próximo a los extremos del intervalo de aproximación. De igual forma, el error de redondeo se asocia a la función de Lebesgue que presenta un máximo cerca de los extremos del intervalo. Sin embargo, mientras que la función $\pi(x)$ tiende a cero con el número de puntos N tendiendo a infinito, la función de Lebesgue no tiende a cero con N. Éste es el principal problema de la interpolación de alto orden en distribuciones de puntos arbitrarias. Para paliar este problema, se proponen distribuciones de puntos no equiespaciados que arrojan valores moderados para la el máximo de la función de Lebesgue o constante de Lebesgue.

El capítulo 4 está dedicado a los interpolantes de Chebyshev o la serie discreta de Chebyshev. La motivación de los puntos de Chebyshev pasa por buscar distribuciones de puntos que hagan uniforme los máximos o los extremos de la función $\|\pi(x)\|$. Se definen los polinomios de Chebyshev como $T_k(x) = \cos(k\theta)$ con $x = \cos(\theta)$, y se demuestra que los ceros de Chebyshev o los puntos que hacen que $T_{N+1}(x) = 0$ verifican que los máximos $\|\pi(x)\|$ son todos iguales. Por otra parte,

se sabe que estos puntos coinciden con los puntos de la cuadratura de Gauss que hace que la serie discreta de Chebyshev y la serie truncada se comporten asintóticamente con el mismo error. Sin embargo, los ceros de Chebyshev no incluyen los extremos del intervalo de integración $[-1, +1]$ y surge la definición de los extremos de Chebyshev. Aunque estos puntos no verifican que los máximos de la función $\pi(x)$ sean uniformes, su comportamiento es mucho mejor que el comportamiento de los puntos equiespaciados. Además, la principal ventaja de los extremos de Chebyshev es que coinciden con los puntos Gauss–Lobatto que permiten realizar de forma exacta la integral para ciertos polinomios. Esta propiedad hace que el producto interno en el espacio de dimensión finita y dimensión infinita coincidan y que el error de la serie truncada y discreta tenga el mismo comportamiento. Por otra parte, en este capítulo se expone la posibilidad de representar la función mediante los coeficientes de la serie discreta o representación en el plano espectral o mediante los valores de la función en los nodos o puntos de Chebyshev o representación en el plano físico. El interés de ambas representaciones reside en la existencia de una transformada rápida entre el plano espectral y el plano físico que involucra $O(Nlog_2N)$ operaciones. Mientras que el cálculo de una derivada espacial en el plano físico cuesta $O(N^2)$ operaciones, el cálculo de la misma derivada en el plano espectral cuesta $O(N)$ operaciones. Por lo tanto, aunque tengamos que hacer un par de transformadas del plano físico al espectral y del espectral al físico, el cálculo de la derivada espectral en el plano espectral es siempre mas atractivo desde el punto de vista computacional. Por el contrario, el cálculo de un término no lineal en el plano espectral cuesta $O(N^2)$ operaciones mientras que en el plano físico cuesta tan solo $O(N)$ operaciones. Por estas razones, los esquemas numéricos utilizan ambas representaciones física y espectral para realizar los cálculos con la mínima carga computacional.

En el capítulo 5 se define la interpolación continua a trozos. Es decir, se fija el grado del interpolante q y, en consecuencia, el número de puntos $q + 1$ por los que pasa el interpolante. De esta forma, el interpolante del dominio de integración está definido por un conjunto de funciones polinómicas de grado q. Desde el punto de vista intuitivo, el comportamiento de la función en un punto depende de unos cuantos puntos vecinos a la derecha y a la izquierda de este punto. Se usa la teoría de interpolación desarrollada en el capítulo 3 para determinar el error de truncamiento y el error de redondeo. Cuando se interpola una función dada en una nube de $N + 1$ puntos mediante interpolantes continuos a trozos de grado q, el error de truncamiento es $O(\Delta x^{q+1})$ siendo Δx la distancia entre los puntos nodales. En este caso, se dice que el método de aproximación tiene una convergencia algebraica que es muy inferior a la convergencia espectral. En este capítulo se obtienen las fórmulas clásicas en diferencias finitas con tres puntos ($q = 2$) para las derivadas primera y segunda y se analiza el error de truncamiento y redondeo. Finalmente, se estudia el paso Δx óptimo que minimiza el error total. Como el error de truncamiento tiende a cero cuando Δx tiende a cero y el error de redondeo tiende a infinito cuando Δx tiende a cero, existe un paso óptimo que minimiza el error. Si nos empeñáramos en disminuir el Δx por debajo de este valor óptimo, el

error en la derivada primera o en la derivada segunda aumentaría al aumentar el error de redondeo.

Una vez entendido en detalle el error de truncamiento y el error de redondeo en el cálculo de derivadas de interpolantes, se procede al uso de estos interpolantes globales o continuos a trozos para resolver problemas de contorno, problemas de condiciones iniciales y problemas de condiciones iniciales y de contorno.

En el capítulo 6 se aborda el problema de contorno en ecuaciones diferenciales ordinarias y ecuaciones en derivadas parciales. Se plantea el algoritmo numérico para resolver estos problemas. El resultado de la discretización espacial conduce a un conjunto de ecuaciones lineales o no lineales para la variable dependiente en el conjunto de puntos nodales o puntos de interpolación. La extensión a problemas bidimensionales se hace mediante el producto tensorial de dos interpolantes unidimensionales continuos a trozos o globales.

En el capítulo 7 se trata el problema de condiciones iniciales o de Cauchy en ecuaciones diferenciales ordinarias. Mediante interpolantes para la función F del problema de Cauchy, se obtienen los métodos de Adams cuyos coeficientes son integrales de la funciones de Lagrange. Se define el error global y el error local de truncamiento de un esquema temporal y se obtiene la ecuación en diferencias que gobierna el error. Cuando el problema de Cauchy es lineal y de coeficientes constantes, se puede resolver analíticamente la ecuación en diferencias del error. La acotación del error para un número de pasos de integración fijado permite definir el radio espectral de la matriz del sistema que determina la estabilidad de la solución del error. Se demuestra que si el radio espectral es menor que uno, el error global de la solución está acotado por el error local de truncamiento. Si el radio espectral es mayor que uno, el error tiende a infinito con el número de pasos tendiendo a infinito. Sin embargo, la determinación del radio espectral es complicada y mediante el teorema de transformación espectral y la definición de la región de estabilidad absoluta se puede determinar de una manera muy rápida y efectiva si el radio espectral es mayor o menor que la unidad. Si los autovalores de la matriz del sistema multiplicados por el paso temporal Δt están dentro de la región de estabilidad absoluta, entonces el radio espectral es menor que uno y podemos esperar que el error esté acotado por el error local de truncamiento. Si embargo, si los autovalores multiplicados por el Δt están fuera de la región de estabilidad absoluta, no tenemos ninguna esperanza de obtener una solución numérica.

En el capítulo 8 se aborda el problema de condiciones iniciales y de contorno. En este caso, la discretización de las variables espaciales se trata de forma independiente a la discretización de la variable temporal. Primero se realiza la discretización espacial y se reduce el problema a un conjunto de ecuaciones diferenciales con condiciones iniciales mediante las técnicas aprendidas en el capítulo 6. A este procedimiento se le conoce con el nombre del método de las líneas. Se define el error de la semidiscretización espacial y se plantea la ecuación diferencial ordinaria para

la evolución temporal del error espacial. Cuando la ecuación en derivadas parciales es lineal de coeficientes constantes, esta ecuación se puede resolver analíticamente y la acotación de la solución para un tiempo de integración dado permite definir la abscisa espectral de la matriz del sistema que determina la estabilidad de la solución. Si la abscisa espectral es mayor que cero, el problema está mal condicionado y el error espacial crece exponencialmente con el tiempo. Sin embargo, si la abscisa espectral es menor que cero, el error espacial está acotado por el error de truncamiento de la semidiscretización espacial. Generalmente, cuando se discretiza la ecuación de ondas mediante esquemas centrados de diferencias finitas de alto orden en mallas equiespaciadas, se obtienen autovalores que tienen parte real mayor que cero que dan origen a una abscisa espectral mayor que cero. Esta problemática está asociada al mal comportamiento de los interpolantes de alto orden en los contornos cuando la malla es equiespaciada. Este problema se suele resolver concentrando puntos en los contornos como lo hacen los puntos de Chebyshev.

En resumen, este libro trata de dar pistas y de entender el origen de los problemas mas comunes que nos podemos encontrar en el tratamiento con el ordenador de los problemas gobernados por ecuaciones diferenciales. Por otra parte, este libro trata de fusionar el mundo de la interpolación en mallas no equiespaciadas con los métodos espectrales o desarrollos en serie discretos. Desde el punto de vista de los autores, el uso del ordenador es imprescindible para entender en profundidad los retos aquí planteados. En esa línea de trabajo, el libro del mismo autor *"How to learn Applied Mathematics through modern FORTRAN"* trata de fusionar el mundo de la programación con el mundo del cálculo numérico a través de abstracciones de alto nivel utilizando un paradigma funcional. La experiencia en las aulas nos ha enseñado que la mejor manera de entender un operador diferencial o la influencia de una condición de contorno es en el ordenador. Esperemos que este libro permita al lector crear sus propios algoritmos basados en los pilares básicos de la interpolación polinómica y su estrecha relación con los métodos espectrales.

Juan A. Hernández
Madrid
Julio 2019

Capítulo 1

Problemas objeto de estudio

1.1. Introducción

De forma general, los problemas que provienen de la física y de la ingeniería se pueden clasificar en tres grandes enunciados: los *problemas de condiciones iniciales y de contorno*, los *problemas de contorno* y los *problemas mixtos*. Desde un punto de vista matemático, los problemas de condiciones iniciales pueden dar lugar a ecuaciones hiperbólicas o parabólicas, los problemas de contorno derivan en ecuaciones de tipo elípticas y los problemas mixtos dan lugar a ecuaciones de tipo parabólico y elíptico. La obtención de soluciones numéricas para estos tres tipos de problemas implica realizar discretizaciones parciales, que se indican en el cuadro (1.1) y que determinan las necesidades o especificaciones siguientes:

En el problema de condiciones iniciales, el modelo matemático contiene derivadas parciales, de la función incógnita $u : \mathbb{R}^3 \times \mathbb{R} \to \mathbb{R}^d$, respecto de variables independientes espaciales y temporales. Para evaluar estas derivadas, se da un conjunto de puntos discretos en el dominio espacial y se aplica la *teoría de interpolación polinómica* y la *teoría de derivación numérica*, dando lugar a los métodos de discretización espacial, y que dan origen a los métodos de *diferencias finitas* (*FDM*), *volúmenes finitos* (*FVM*) y *elementos finitos* (*FEM*). Por otro lado, la *teoría de interpolación trigonométrica* permite obtener los denominados *métodos espectrales*. Este procedimiento se conoce como semi–discretización espacial y el resultado es un problema de valores iniciales para un sistema de ecuaciones diferenciales ordinarias en la función incógnita $U : \mathbb{R} \to \mathbb{R}^N$ siendo $N = dM$ con M el número de puntos de la semi–discretización espacial. Para evaluar las derivadas en la variable temporal, se define un conjunto de puntos discretos en el dominio temporal y, nuevamente,

se aplica la *teoría de interpolación polinómica* y la *teoría de derivación y cuadratura numérica*. En este caso, se obtienen los métodos numéricos *lineales multipaso* y *predictores–correctores*, que pueden ser tanto explícitos como implícitos. Otras técnicas de aproximación permiten obtener otro conjunto importante de métodos que son los denominados *Runge-Kutta* y que, de nuevo, pueden ser explícitos o implícitos. Los métodos implícitos, cuando se aplican a problemas diferenciales no lineales, dan lugar a sistemas de ecuaciones no lineales, que se resuelven aplicando los métodos numéricos correspondientes, como son el método de Newton-Raphson o el método de iteración de punto fijo. Esta aproximación se conoce como discretización temporal y el resultado es un problema de valores iniciales para un sistema de ecuaciones en diferencias. La primera columna del cuadro (1.1) resume el método de cálculo en problemas de valores iniciales explicado anteriormente.

En el problema de contorno, al igual que en el problema de valores iniciales, el modelo matemático contiene derivadas parciales, de la función incógnita $u : \mathbb{R}^3 \rightarrow \mathbb{R}^d$, respecto de las variables espaciales. Para evaluar estas derivadas, se emplean las técnicas de discretización espacial mencionadas en el problema de valores iniciales y el resultado es un sistema de ecuaciones algebraicas. Si el sistema es lineal, la solución se obtiene aplicando métodos numéricos para sistemas lineales como son el método de Gauss o factorización LU. Si el sistema es no lineal, la solución se obtiene aplicando métodos numéricos para sistemas no lineales como son el método de Newton o el método de iteración de punto fijo. La segunda columna del cuadro (1.1) resume el método de cálculo en problemas de contorno explicado anteriormente.

En un problema mixto, aparecen ecuaciones en derivadas parciales que evolucionan en el tiempo, por lo general de tipo parabólico, junto con ecuaciones de tipo elíptico. Para resolver los problemas mixtos se emplean las técnicas de discretización espacial y temporal junto con los métodos en ecuaciones algebraicas mencionadas en el tratamiento de los problemas de condiciones iniciales y de contorno explicados anteriormente.

Enunciado	Problema hiperbólico o parabólico.	Problema elíptico
Modelo matemático	Conjunto de ecuaciones en derivadas parciales $\frac{\partial u}{\partial t} = \mathcal{L}(u, t, x)$ + condiciones iniciales + condiciones de contorno	Conjunto de ecuaciones en derivadas parciales $\mathcal{L}(u, x) = 0$ + condiciones de contorno
Discretización espacial	Conjunto de ecuaciones diferenciales ordinarias $\frac{dU}{dt} = F(U, t)$ + condiciones iniciales	Conjunto de ecuaciones algebraicas $F(U) = 0$
Discretización temporal	Conjunto de ecuaciones en diferencias (DEs) $G\left(U^{n+1}, U^n, \ldots\right) = 0$ + condiciones iniciales	
Solución discreta	U^n	U

Cuadro 1.1: Discretización espacial y temporal para problemas de condiciones iniciales y de contorno en ecuaciones en derivadas parciales.

Capítulo 2

Aproximación de funciones

2.1. Introducción

Un problema de gran interés en matemática aplicada es la evaluación de una función en un punto, cuando no se conoce la expresión analítica de la misma, o cuando, sí se dispone de tal expresión, pero es difícil de evaluar en el punto de estudio. Otro problema habitual, por ejemplo en cálculo numérico, es la obtención de derivadas de distinto orden e integrales de funciones de las que sólo se disponen de algunos valores puntuales. Para abordar el estudio de estos problemas, se dispone de un tema clásico de la matemática que es la *teoría de aproximación*.

Dada una función f, la teoría de aproximación estudia la obtención de otra función g más sencilla, que es *próxima* a f en un cierto sentido, y que debe precisarse. Dicha función sencilla g puede ser un polinomio, una función trigonométrica, una función racional, y en general, cualquier función elemental de las estudiadas en el cálculo infinitesimal. Particularmente, las funciones polinómicas y trigonométricas se han estudiado con detalle dando lugar al desarrollo de la teoría de aproximación polinómica y a la teoría de aproximación trigonométrica. La evolución en la teoría de aproximación se puede resumir a partir de los aportes fundamentales de Newton (1643), Euler (1707), Lagrange (1736), Weierstrass (1815), Chebyshev (1821) y Lebesgue (1875).

El teorema siguiente asegura la existencia de polinomios que convergen uniformemente a f.

Teorema de aproximación de Weierstrass.
Sea f una función continua definida en un intervalo $[a, b]$. Para todo $\epsilon > 0$, le corresponde un polinomio P tal que:

$$||f - P|| < \epsilon.$$

Es decir que:

$$|f(x) - P(x)| < \epsilon, \qquad \textit{para todo } x \in [a, b].$$

Es bien conocida la utilidad de expresar una función como un desarrollo en serie de funciones elementales de la forma:

$$f(x) \;=\; \hat{c}_0\ \phi_0(x) \;+\; \hat{c}_1\ \phi_1(x) \;+\; \hat{c}_2\ \phi_2(x) \;+\; \cdots$$

En la expresión anterior, $\hat{c}_k$ son coeficientes a determinar y ϕ_k son funciones elementales o funciones base como polinomios, funciones trigonométricas o exponenciales.

Si $f(x)$ es un elemento de un espacio vectorial de dimensión infinita y existe un conjunto de funciones base $\phi_k(x)$, entonces las función $f(x)$ se puede poner como:

$$f(x) \;= \sum_{k=0}^{\infty} \hat{c}_k\ \phi_k(x), \tag{2.1}$$

donde $\hat{c}_k$ constituyen los coeficientes del desarrollo en serie. Siguiendo la analogía de un espacio vectorial, los coeficientes se obtienen proyectando la función $f(x)$ en las funciones base $\phi_k(x)$. De manera más precisa, si definimos el producto interno siguiente con la función de peso $\omega(x)$:

$$< \phi_m, \phi_k >_w = \int_a^b \phi_m(x)\ \phi_k(x)\ \omega(x)dx, \tag{2.2}$$

entonces la proyección de $f(x)$ en $\phi_m(x)$ queda:

$$\int_a^b f(x)\ \phi_m(x)\ \omega(x)dx = \sum_{k=0}^{\infty} \hat{c}_k \int_a^b \phi_k(x)\ \phi_m(x)\ \omega(x)dx \tag{2.3}$$

que permiten obtener los coeficientes $\hat{c}_k$. Generalmente, se eligen funciones ortogonales, es decir:

$$< \phi_m, \phi_k >_w = \begin{cases} 0 & k \neq m \\ \gamma_m & k = m \end{cases}$$

En este caso, los coeficientes $\hat{c}_k$ se calculan a partir de la siguiente expresión:

$$\hat{c}_m \;=\; \frac{1}{\gamma_m} \int_a^b f(x)\ \phi_m(x)\ \omega(x)\ dx. \tag{2.4}$$

2.2. Serie truncada y precisión espectral

Dado un desarrollo en serie de una función $f(x)$, se define la serie truncada como los N primeros términos del desarrollo

$$P_N(x) \;=\; \sum_{k=0}^{N} \hat{c}_k \;\phi_k(x). \tag{2.5}$$

El error de la serie truncada se define como el error de truncamiento y vale:

$$E_N(x) \;=\; \sum_{k=N+1}^{\infty} \hat{c}_k \;\phi_k(x).$$

Desde el punto de vista computacional interesa conocer cuantos términos debemos retener del desarrollo en serie para que el error de truncamiento sea pequeño. La regularidad de la función $f(x)$ a desarrollar y las propiedades de las funciones base permiten conocer el número de términos que se deben retener para un error prefijado. La elección de las funciones base permite acelerar la convergencia del desarrollo en serie o minimizar el número de grados de libertad o coeficientes $\hat{c}_k$ necesarios para representar de manera precisa la función $f(x)$.

Desde el punto de vista numérico, solo podemos manejar un número finito de modos k por lo que nos interesa que la convergencia de estos desarrollos en serie sea lo más rápida posible. De esta forma, la serie truncada aproxima bien a la función objeto de estudio. La aproximación más familiar es el desarrollo en serie de Fourier en donde las funciones ortogonales son senos y cosenos $\phi_k(x) = e^{ikx}$. Si la función es infinitamente suave, con todas sus derivadas periódicas en el contorno, el coeficiente k–ésimo del desarrollo en serie decae más rápidamente que cualquier potencia inversa de k. En la práctica se observa este comportamiento cuando tenemos suficientes términos del desarrollo en serie como para representar bien la estructura de la función $f(x)$. El rápido decaimiento de los coeficientes nos permite truncar las series con unos pocos términos y tener una excelente aproximación de la función. Esta característica básica de los métodos espectrales se denominada precisión espectral.

La precisión espectral también se puede conseguir para funciones suaves pero que no sean periódicas utilizando funciones $\phi_k(x)$ especiales. Para cualquier base ortogonal de funciones no se verifica, en principio, que los coeficientes del desarrollo en serie de una función suave en esta base tiendan a cero más rápidamente que cualquier potencia inversa de k. Usualmente, la precisión espectral se consigue solamente cuando la función tiene un comportamiento en el contorno muy especial. Las autofunciones del problema singular de Sturm–Liouville permiten precisión espectral en el desarrollo de cualquier función suave sin que sea necesario ninguna restricción adicional en el comportamiento en el contorno.

Ejemplo: Desarrollar en serie de senos y cosenos una función $f(x)$.

Para analizar el comportamiento de los coeficientes del desarrollo en serie de Fourier con el número de onda k, integramos por partes en (2.4), para lo cual tenemos que considerar que $f(x)$ es diferenciable en $[0, 2\pi]$, entonces para $k \neq 0$

$$\hat{c}_k = \frac{-1}{ik}(f(2\pi) - f(0)) + \frac{1}{ik}\int_0^\infty f'(x)e^{-ikx}dx. \tag{2.6}$$

Si no imponemos ninguna condición adicional, entonces se verifica que

$$|\hat{c}_k| = O\left(\frac{1}{k}\right).$$

Si además, $f'(x)$ es diferenciable en $[0, 2\pi]$ y $f(2\pi) = f(0)$, integrando nuevamente por partes en (2.6) queda

$$\hat{c}_k = \frac{1}{(ik)^2}(f'(2\pi) - f'(0)) + \frac{1}{(ik)^2}\int_0^\infty f''(x)e^{-ikx}dx, \tag{2.7}$$

y los coeficientes

$$|\hat{c}_k| = O\left(\frac{1}{k^2}\right).$$

De esta forma, si $f(x)$ es diferenciable m veces en $[0, 2\pi]$ y todas sus derivadas hasta orden $m - 1$ son periódicas, entonces

$$|\hat{c}_k| = O\left(\frac{1}{k^m}\right). \tag{2.8}$$

Como corolario, se puede concluir que el coeficiente $\hat{c}_k$ del desarrollo en serie de Fourier de una función que es infinitamente diferenciable y todas sus derivadas periódicas en $[0, 2\pi]$ decae más rápidamente que cualquier potencia negativa de k.

2.3. Serie discreta o interpolante

Se define la serie discreta o interpolante $I_N(x)$ como el siguiente desarrollo en serie de $f(x)$ para el mismo conjunto de funciones base $\phi_k(x)$

$$I_N(x) \;=\; \sum_{k=0}^{N} \tilde{c}_k \;\phi_k(x), \tag{2.9}$$

donde los coeficientes $\tilde{c}_k$ se obtienen forzando a que $I_N(x)$ coincida con el valor de la función $f(x)$ en ciertos puntos nodales x_j.

La diferencia fundamental de la serie discreta frente a la serie truncada está en el método de obtención de los coeficientes. Si forzamos a que el interpolante $I_N(x)$ pase por un conjunto de puntos $\{x_j, j = 0, \dots N\}$, se obtiene

$$I_N(x_j) \;=\; f(x_j) \;=\; \sum_{k=0}^{N} \tilde{c}_k \; \phi_k(x_j). \tag{2.10}$$

El sistema anterior constituye un sistema lineal de ecuaciones para la determinación de los coeficientes $\tilde{c}_k$. En general, salvo que la nube de puntos x_j sea óptima, la convergencia de la serie discreta (2.10) es mala. Existen nubes de puntos óptimas desde el punto de vista de interpolación. Es objeto de este libro entender el origen de esta nube de puntos y en donde reside su bondad.

Desde el punto de vista intuitivo se busca una nube de puntos que con un producto interno en un espacio de dimensión finita goce de las propiedades de ortogonalidad de los espacios de dimensión finita. Dada una función $f(x)$ definida en un intervalo $[a,\ b]$, se considera una partición de $N+1$ puntos discretos x_j con $x_0 = a$ y $x_N = b$. Dada una función base $\phi_m(x)$, se pueden definir el vector base Φ_m cuyas componentes son la evaluación de la función base en los puntos nodales:

$$\Phi_m \;=\; \{\ \phi_m(x_0),\ \phi_m(x_1),\ \dots\ ,\ \phi_m(x_j),\ \dots\ ,\ \phi_m(x_N)\ \}.$$

A partir de estos vectores base y mediante los coeficientes de peso α_j se define el producto interno en un espacio de dimensión finita mediante:

$$< \phi_m, \phi_k >_N = \sum_{j=0}^{N} \phi_m(x_j)\ \phi_k(x_j)\ \alpha_j. \tag{2.11}$$

Consideremos una nube de puntos que verifique las siguientes ecuaciones:

$$\sum_{j=0}^{N} \phi_k(x_j)\ \phi_m(x_j)\ \alpha_j = \int_a^b \phi_m(x)\ \phi_k(x)\ \omega(x)\ dx. \tag{2.12}$$

Como disponemos de $N+1$ puntos nodales a determinar y $N+1$ coeficientes α_j desconocidos, podemos forzar al cumplimiento de $2N+2$ ecuaciones de (2.12). Es decir, podemos forzar el cumplimiento de (2.12) con $k = 0, \dots N$ y $m = 0, \dots N$. O lo que es lo mismo, si se elige adecuadamente la nube de puntos, entonces los vectores base del espacio de dimensión finita son ortogonales:

$$\sum_{j=0}^{N} \phi_m(x_j)\ \phi_k(x_j)\ \alpha_j = \begin{cases} 0 & k \neq m, \\ \gamma_m & k = m. \end{cases} \tag{2.13}$$

Es importante hacer notar que la ortogonalidad se verifica exclusivamente para valores de $k \leq N$ y $m \leq N$. Para estos puntos nodales, los coeficientes $\tilde{c}_m$ del

interpolante se obtienen proyectando el vector

$$\mathbf{f} \;=\; \{\; f(x_0),\; f(x_1),\; \ldots\;,\; f(x_j),\; \ldots\;,\; f(x_N)\;\}$$

en los vectores de la base. Es decir,

$$\tilde{c}_m \;=\; \frac{1}{\gamma_m}\;\sum_{j=0}^{N}\; f(x_j)\;\phi_m(x_j)\;\alpha_j, \tag{2.14}$$

o mediante la definición del producto interno (2.11) como la proyección de $\mathbf{f}$ sobre el vector de la base Φ_m

$$\tilde{c}_m \;=\; \frac{1}{\gamma_m}\;<f,\;\phi_m>_N\;. \tag{2.15}$$

2.4. Serie discreta y serie truncada

En esta sección se estudia la relación que existe entre la aproximación por series y la aproximación por interpolación. En particular, se analiza la relación entre los coeficientes $\hat{c}_k$ de la serie truncada y los coeficientes $\tilde{c}_k$ de la serie discreta.

Para relacionar estos coeficientes, se particulariza el interpolante (2.10) en x_j y el desarrollo en serie dado por (2.1) y se igualan sus valores

$$\sum_{k=0}^{N}\;\tilde{c}_k\;\phi_k(x_j) = \sum_{k=0}^{\infty}\hat{c}_k\;\phi_k(x_j). \tag{2.16}$$

En esta expresión multiplicamos por $\phi_m(x_j)\alpha_j$ y sumamos desde $j=0$ hasta $j=N$

$$\sum_{k=0}^{N}\;\tilde{c}_k\;\sum_{j=0}^{N}\phi_m(x_j)\;\phi_k(x_j)\;\alpha_j = \sum_{k=0}^{\infty}\;\hat{c}_k\;\sum_{j=0}^{N}\phi_m(x_j)\;\phi_k(x_j)\;\alpha_j. \tag{2.17}$$

Para la nube de puntos x_j que verifica las ecuaciones (2.12) y utilizando la ortogonalidad dada por (2.13), los coeficientes de la serie discreta se expresan como:

$$\tilde{c}_m = \;\hat{c}_m + \frac{1}{\gamma_m}\;\sum_{k=N+1}^{\infty}\;\hat{c}_k\;\sum_{j=0}^{N}\phi_m(x_j)\;\phi_k(x_j)\;\alpha_j. \tag{2.18}$$

Esa misma expresión se puede poner mediante la definición del producto interno discreto (2.11) como:

$$\tilde{c}_m = \;\hat{c}_m + \frac{1}{\gamma_m}\;\sum_{k=N+1}^{\infty}\;<\phi_m,\phi_k>_N\;\;\hat{c}_k. \tag{2.19}$$

La diferencia entre los coeficientes de la serie discreta y los coeficientes de la serie truncada se denomina error de aliasing. Es importante hacer notar que si el desarrollo en serie converge con precisión espectral, las colas de desarrollo en serie a partir del valor $k = N + 1$ pueden ser muy pequeñas y la diferencia entre los coeficientes de la serie discreta y los coeficientes de la serie truncada ser despreciable.

Desde el punto de vista computacional, las conclusiones mas importantes relacionadas con los desarrollos en serie y los interpolantes o series discretas son las siguientes:

1. El tratamiento con ordenador de una determinada función obliga a aproximar la misma con una serie de grados de libertad. Estos grados de libertad pueden ser los coeficientes de un desarrollo en serie o los valores de la función en un conjunto de puntos nodales.

2. Si la función es infinitamente derivable, el desarrollo en serie de funciones base puede tener convergencia espectral. Es decir, que el módulo de los coeficientes de la serie tiendan a cero mucho más rápidamente que cualquier potencia negativa del índice del coeficiente. Esta propiedad es especialmente importante para reducir de forma drástica el número de grados de libertad cuando se aproxima una función en el ordenador y conduce a lo que se denomina esquemas con precisión espectral.

3. Existen puntos nodales óptimos que permiten interpolar la función con un error similar a los de los desarrollos en serie espectrales salvo un pequeño error de aliasing. Generalmente, la simulación o tratamiento con ordenador de los problemas de matemática aplicada se implementan de manera mucho mas sencilla mediante mallas o nubes de puntos que mediante desarrollos en serie. La confianza en los métodos de interpolación como métodos de precisión espectral hace muy interesante su uso en la aplicación práctica.

Los siguientes capítulos del libro están enfocados a entender con precisión el origen y las implicaciones de estas conclusiones para poder abordar con confianza los problemas no lineales de simulación en ecuaciones en derivadas parciales.

2.5. Fórmulas de cuadratura de Gauss

En esta sección trataremos de obtener las nubes de puntos óptimas de las que hablábamos en la sección anterior. Para entender el origen de esta nube de puntos, tenemos que profundizar en las fórmulas de cuadratura de Gauss. Una cuadratura de Gauss es una fórmula que selecciona $N+1$ puntos nodales x_j y $N+1$ pesos α_j de manera óptima para dar un resultado exacto para polinomios $g(x)$ de grado $2N+1$ con una función de ponderación $\omega(x)$ no negativa en intervalo $[a,b]$. Es decir,

$$\int_a^b g(x)\ \omega(x)\ dx = \sum_{j=0}^{N}\ \alpha_j\ g(x_j). \tag{2.20}$$

Para llevar a cabo la demostración, expresamos $g(x)$ mediante

$$g(x) = c(x)\ \phi_{N+1}(x) + r(x) \tag{2.21}$$

donde $c(x)$ es el polinomio cociente y $r(x)$ es el polinomio resto. Como $g(x)$ es un polinomio de grado $2N+1$, el polinomio cociente en serie de las funciones base es

$$c(x) = \beta_0\ \phi_0(x) + \ldots \beta_N\ \phi_N(x). \tag{2.22}$$

Si se lleva este desarrollo a (2.20) y debido a la ortogonalidad de ϕ_k y ϕ_{N+1} para todo $k \leq N$, se comprueba

$$\int_a^b g(x)\ \omega(x)\ dx = \int_a^b r(x)\ \omega(x)\ dx. \tag{2.23}$$

Como las integrales de (2.23) se pueden calcular mediante (2.20), $r(x_j)$ debe coincidir con $g(x_j)$. Teniendo en cuenta (2.21) y para que $g(x_j) = r(x_j)$, x_j deben ser los ceros de $\phi_{N+1}(x)$.

$$\int_a^b r(x)\ \omega(x)\ dx = \sum_{j=0}^{N}\ \alpha_j\ g(x_j). \tag{2.24}$$

Por último, falta por determinar los pesos α_j. Como $r(x)$ es un polinomio de grado N, $r(x)$ se puede expresar

$$r(x) = a_0 + a_1 x + \ldots + a_N x^N. \tag{2.25}$$

donde a_k son coeficientes arbitrarios. Si llevamos este desarrollo a (2.24), se obtiene

$$a_0 \int_a^b\ \omega(x)dx + a_1 \int_a^b x\ \omega(x)dx + \ldots + a_N \int_a^b x^N \omega(x)\ dx =$$
$$a_0 \sum_{j=0}^{N}\ \alpha_j + a_1 \sum_{j=0}^{N}\ x_j\ \alpha_j + \ldots + a_N \sum_{j=0}^{N}\ x_j^N\ \alpha_j.$$

para cualquier valor de a_k. De esta forma y dada una función de peso $\omega(x)$, el sistema lineal de ecuaciones

$$\int_a^b x^k \ \omega(x)dx = \sum_{j=0}^{N} \ x_j^k \ \alpha_j, \qquad k = 0, \dots N \tag{2.26}$$

permite obtener los pesos α_j.

En conclusión, si los puntos nodales $\{x_j, j = 0, \dots N\}$ son los ceros de $\phi_{N+1}(x)$ y los pesos α_j están determinados por el sistema lineal (2.26), la fórmula de cuadratura de Gauss (2.20) es exacta para polinomios de grado $2N + 1$.

Con estos pesos α_j y estos puntos nodales x_j, la fórmula de cuadratura de Gauss permite calcular de forma exacta el producto interno de un espacio de dimensión infinita mediante el producto interno de un espacio de dimensión finita

$$\int_a^b \phi_m(x) \ \phi_k(x) \ \omega(x) \ dx = \sum_{j=0}^{N} \phi_k(x_j) \ \phi_m(x_j) \ \alpha_j \tag{2.27}$$

para todo $k \leq N$ y todo $m \leq N$.

En esta fórmula de integración gaussiana los puntos nodales x_j no incluyen los contornos $x = +1$ y $x = -1$. Cuando el problema que queramos resolver incluya condiciones de contorno en los extremos del intervalo, necesitamos generalizar las fórmulas de integración gaussiana que incluyan estos puntos. Éste es el objeto de la siguiente sección.

2.6. Fórmulas de cuadratura de Gauss–Lobatto

Para obtener las fórmulas de integración gaussiana que incluyan los extremos del intervalo, se considera el siguiente polinomio:

$$q_{N+1}(x) = \phi_{N+1}(x) + c_1 \ \phi_N(x) + c_2 \ \phi_{N-1}(x), \tag{2.28}$$

donde c_1 y c_2 se eligen para que $q_{N+1}(-1) = q_{N+1}(+1) = 0$. De esta forma, las $N + 1$ raíces de $q_{N+1}(x)$ incluyen $x_0 = -1$ y $x_N = +1$.

Una cuadratura de Gauss–Lobatto es una fórmula que selecciona $N - 1$ puntos nodales $x_1, \dots x_{N-1}$ y $N + 1$ pesos α_j de manera óptima para dar un resultado exacto para polinomios $g(x)$ de grado $2N - 1$. Es decir,

$$\int_a^b g(x) \ \omega(x) \ dx = \sum_{j=0}^{N} \ \alpha_j \ g(x_j). \tag{2.29}$$

La demostración se hace de manera similar a la cuadratura de Gauss. Consideremos que $g(x)$ es un polinomio de grado $2N-1$ que lo expresamos de la siguiente forma:

$$g(x) = c(x)\ q_{N+1}(x) + r(x), \tag{2.30}$$

donde $r(x)$ es el polinomio resto de grado N y $c(x)$ es el polinomio cociente de grado $N-2$ siguiente:

$$c(x) = \beta_0\ \phi_0(x) + \ldots \beta_{N-2}\ \phi_{N-2}(x). \tag{2.31}$$

Es importante hacer notar que como (2.28) incluye $\phi_{N-1}(x)$, $c(x)$ puede llegar a grado $N-2$ para que

$$\int_a^b c(x)\ q_{N+1}(x)\ \omega(x)\ dx = 0, \tag{2.32}$$

y podamos calcular la integral mediante (2.29) con los pesos α_j dados por (2.26).

En resumen, si los puntos nodales $\{x_j, j = 0, \ldots N\}$ son los ceros de $q_{N+1}(x)$ que incluyen los extremos del intervalo y los pesos α_j están determinados por el sistema lineal (2.26), la formula de cuadratura de Gauss (2.29) es exacta para polinomios de grado $2N-1$.

Con estos pesos α_j y estos puntos nodales x_j, la fórmula de cuadratura de Gauss permite calcular de forma exacta el producto interno de un espacio de dimensión infinita mediante el producto interno de un espacio de dimensión finita

$$\int_a^b \phi_m(x)\ \phi_k(x)\ \omega(x)\ dx = \sum_{j=0}^{N} \phi_k(x_j)\ \phi_m(x_j)\ \alpha_j \tag{2.33}$$

para todo $k \leq N$ y todo $m \leq N-1$.

Capítulo 3

Interpolación

3.1. Interpolación polinómica

Definición: Problema general de interpolación.
Dada una función f y un conjunto de $N+1$ puntos discretos $x_0, x_1, \ldots, x_N$ del dominio de definición, el problema general de interpolación consiste en obtener otra función I, denominada función interpolante, de la forma:

$$I_N(x) \;=\; \tilde{c}_0\ \phi_0(x) \;+\; \tilde{c}_1\ \phi_1(x) \;+\; \cdots \;+\; \tilde{c}_N \phi_N(x),$$

que es combinación lineal de funciones elementales $\phi_0,\ \phi_1,\ \ldots,\ \phi_N$, y que debe verificar:

$$I(x_j) \;=\; f(x_j), \qquad j = 0, 1, \ldots, N.$$

Los puntos x_j se denominan *abscisas* o *nodos* de interpolación y los valores $\tilde{c}_j$ son los *coeficientes* de interpolación. Resolver el problema de interpolación implica obtener los coeficientes a_j. Esta definición también puede enunciarse diciendo que dado un espacio vectorial V, de dimensión finita N, y dada una base $\phi_0,\ \phi_1,\ \ldots,\ \phi_N$, cualquier elemento de V se puede expresar como una combinación lineal de los elementos de la base.

Funciones base polinómicas dan lugar a problemas de *interpolación polinómica* y funciones base trigonométricas dan lugar a problemas de *interpolación trigonométrica*.

A continuación, se presentan tres formas distintas de expresar un polinomio interpolante que son: la interpolación en monomios, la interpolación de Lagrange

y la forma de Newton. En estas tres formas, los polinomios $\phi_0,\ \phi_1,\ \ldots,\ \phi_N$ tienen expresiones diferentes y cada una de ellas permite justificar distintos aspectos de la teoría de la interpolación polinómica.

3.2. Matriz de Vandermonde

Se consideran $N+1$ puntos $x_0, x_1, \ldots, x_N$ del dominio de definición de una función f, en los que se conocen los valores $f(x_j) = y_j, \quad j = 0, 1, \ldots, N$. El problema de interpolación consiste en obtener una función interpolante I_N de grado $\leq N$, como una combinación lineal de monomios de la forma siguiente:

$$I_N(x) \ = \ a_0 \ + \ a_1 \ x \ + \ \cdots \ + \ a_N \ x^N.$$

Por definición de interpolación, el interpolante I_N debe verificar que:

$$I_N(x_j) \ = \ f(x_j), \quad j = 0, 1, \ldots, N.$$

En la expresión anterior, $a_0, a_1, \ldots, a_N$ son los coeficientes del polinomio y la obtención de los mismos, resuelve el problema de interpolación. La condición anterior conduce al sistema lineal de ecuaciones siguiente:

$$\begin{aligned}
a_0 \ + \ a_1 \ x_0 \ + \ a_2 \ x_0^2 \ + \ \cdots \ + \ a_N \ x_0^N \ &= \ f(x_0), \\
a_0 \ + \ a_1 \ x_1 \ + \ a_2 \ x_1^2 \ + \ \cdots \ + \ a_N \ x_1^N \ &= \ f(x_1), \\
&\vdots \\
a_0 \ + \ a_1 \ x_N \ + \ a_2 \ x_N^2 \ + \ \cdots \ + \ a_N \ x_N^N \ &= \ f(x_N),
\end{aligned}$$

que, en forma matricial, puede expresarse como sigue

$$\begin{pmatrix} 1 & x_0 & x_0^2 & \cdots & x_0^N \\ 1 & x_1 & x_1^2 & \cdots & x_1^N \\ \vdots & & & & \vdots \\ 1 & x_N & x_N^2 & \cdots & x_N^N \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_N \end{pmatrix} = \begin{pmatrix} f(x_0) \\ f(x_1) \\ \vdots \\ f(x_N) \end{pmatrix}. \tag{3.1}$$

La matriz de la expresión anterior, se conoce como matriz de Vandermonde. Si los puntos de interpolación son distintos, la matriz de Vandermonde no es singular y el sistema (3.1) tiene solución única. La solución del sistema lineal (3.1) se puede obtener mediante cualquier método directo de resolución de sistemas lineales como, por ejemplo, el método de eliminación Gaussiana o de factorización LU. El

problema de interpolación queda finalmente resuelto con el polinomio siguiente:

$$I_N(x) \;=\; a_0 \;+\; a_1\, x \;+\; \cdots \;+\; a_N\, x^N,$$

expresado como una combinación lineal de monomios. Una dificultad que presenta esta forma de resolver el problema de interpolación es que la matriz de Vandermonde puede presentar un número de condición grande, dando lugar a sistemas mal condicionados. A continuación se presenta un ejemplo de interpolación en monomios con tres puntos de interpolación.

Ejemplo:
Dada una función f y 3 puntos discretos x_0, x_1, x_2 del dominio de definición, se propone obtener el polinomio interpolante I_2 de grado ≤ 2, expresado como combinación lineal de monomios. Por definición del problema de interpolación, I_2 debe verificar que:

$$I_2(x_0) \;=\; f(x_0) \;=\; y_0, \quad I_2(x_1) \;=\; f(x_1) \;=\; y_1, \quad I_2(x_2) \;=\; f(x_2) \;=\; y_2.$$

Tal como se observa en las expresiones anteriores, se reemplaza $f(x_j)$ por y_j para simplificar las expresiones finales.

Para resolver este problema se plantea el polinomio de grado ≤ 2 siguiente:

$$I_2(x) \;=\; a_0 \;+\; a_1\, x \;+\; a_2\, x^2,$$

donde a_0, a_1 y a_2 son coeficientes que deben determinarse. Para ello, se fuerza a que el polinomio I_2 coincida con la función f en los 3 puntos discretos:

$$I_2(x_0) \;=\; y_0, \qquad I_2(x_1) \;=\; y_1, \qquad I_2(x_2) \;=\; y_2,$$

que conduce a un sistema de tres ecuaciones algebraicas lineales, cuya matriz de Vandermonde (matriz ampliada) es:

$$\left(\begin{array}{ccc|c} 1 & x_0 & x_0^2 & y_0 \\ 1 & x_1 & x_1^2 & y_1 \\ 1 & x_2 & x_2^2 & y_2 \end{array}\right).$$

El sistema se resuelve aplicando el método de Gauss. Mediante operaciones elementales de filas se obtienen los sistemas equivalentes siguientes:

$$\left(\begin{array}{ccc|c} 1 & x_0 & x_0^2 & y_0 \\ 0 & x_1 - x_0 & x_1^2 - x_0^2 & y_1 - y_0 \\ 0 & x_2 - x_1 & x_2^2 - x_1^2 & y_2 - y_1 \end{array}\right),$$

$$\left(\begin{array}{ccc|c} 1 & x_0 & x_0^2 & y_0 \\ 0 & 1 & x_1+x_0 & \dfrac{y_1-y_0}{x_1-x_0} \\ 0 & 1 & x_2+x_1 & \dfrac{y_2-y_1}{x_2-x_1} \end{array}\right),$$

$$\left(\begin{array}{ccc|c} 1 & x_0 & x_0^2 & y_0 \\ 0 & 1 & x_1+x_0 & \dfrac{y_1-y_0}{x_1-x_0} \\ 0 & 0 & x_2-x_0 & \left(\dfrac{y_2-y_1}{x_2-x_1} - \dfrac{y_1-y_0}{x_1-x_0}\right) \end{array}\right).$$

Aplicando sustitución inversa al sistema anterior, se calculan los coeficientes del polinomio:

$$\begin{aligned} a_2 &= \frac{1}{x_2-x_0}\left(\frac{y_2-y_1}{x_2-x_1} - \frac{y_1-y_0}{x_1-x_0}\right), \\ a_1 &= \left(\frac{y_1-y_0}{x_1-x_0}\right) - \frac{x_1+x_0}{x_2-x_0}\left(\frac{y_2-y_1}{x_2-x_1} - \frac{y_1-y_0}{x_1-x_0}\right), \\ a_0 &= y_0 - x_0 a_1 - x_0^2 a_2. \end{aligned}$$

La forma de solución del problema de interpolación, expresada como una combinación lineal de monomios, no es la habitualmente empleada, dado que requiere resolver el sistema lineal (3.1) cada vez que se modifica el conjunto de nodos de interpolación. Esto hace que este método resulte ser muy laborioso. Por otro lado, desde un punto de vista computacional, el mal condicionamiento de la matriz de Vandermonde hace que el método no sea el adecuado. A continuación, se presenta otra forma de resolver el problema de interpolación que resulta ser mucho más práctica al momento de generar esquemas numéricos para la solución de ecuaciones diferenciales.

3.3. Interpolación de Lagrange

En este apartado, se presenta otra forma de resolver el problema de interpolación polinómica, que resulta ser más eficiente desde el punto de vista de la

obtención de esquemas numéricos para la solución de ecuaciones diferenciales. Se consideran $N+1$ puntos $x_0, x_1, \ldots, x_N$ del dominio de definición de una función f, en los que se conocen los valores $f(x_j), \quad j = 0, 1, \ldots, N$. A continuación, se expresa el polinomio I_N de grado N en la forma siguiente:

$$I_N(x) = b_0\ \ell_0(x)\ +\ b_1\ \ell_1(x)\ +\ \cdots\ +\ b_N\ \ell_N(x), \tag{3.2}$$

donde $b_0, b_1, \ldots, b_N$ son coeficientes a calcular y $\ell_j(x)$ son polinomios de grado N que tienen la expresión siguiente:

$$\ell_j(x)\ =\ \frac{(x-x_0)(x-x_1)\ldots(x-x_{j-1})(x-x_{j+1})\ldots(x-x_N)}{(x_j-x_0)(x_j-x_1)\ldots(x_j-x_{j-1})(x_j-x_{j+1})\ldots(x_j-x_N)}, \tag{3.3}$$

o en su forma mas compacta

$$\ell_j(x)\ =\prod_{\substack{i=0\\ i\neq j}}^{N}\frac{(x-x_i)}{(x_j-x_i)}. \tag{3.4}$$

La forma (3.2) se conoce como interpolación de Lagrange y los polinomios (3.3) son los polinomios de Lagrange. Como los polinomios (3.3) toman los valores siguientes:

$$\begin{cases} \ell_j(x) = 1, & x = x_j, \\ \ell_j(x) = 0, & x = x_i, \qquad i = 0, 1, \ldots, j-1, j+1, \ldots, N, \end{cases} \tag{3.5}$$

entonces la interpolación de Lagrange se puede expresar como se indica a continuación:

$$I_N(x)\ =\ \sum_{j=0}^{N}\ f(x_j)\ \ell_j(x). \tag{3.6}$$

A continuación se presenta un ejemplo de interpolación en la forma de Lagrange, con tres puntos de interpolación.

Ejemplo:
Dada una función f y 3 puntos de interpolación x_0, x_1, x_2 del dominio de definición, se propone obtener el polinomio interpolante I_2, de grado ≤ 2, mediante la interpolación de Lagrange, que verifique que

$$I_2(x_0)\ =\ f(x_0)\ =\ y_0, \qquad I_2(x_1)\ =\ f(x_1)\ =\ y_1, \qquad I_2(x_2)\ =\ f(x_2)\ =\ y_2.$$

Aplicando la expresión (3.2), el interpolante de grado ≤ 2 es:

$$I_2(x)\ =\ b_0\ \ell_0(x)\ +\ b_1\ \ell_1(x)\ +\ b_2\ \ell_2(x),$$

donde $\ell_0,\ \ell_1,\ \ell_2$ son polinomios de grado ≤ 2 que, según (3.3), tienen las expresiones siguientes:

$$\ell_0(x) \ = \ \frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)},$$

$$\ell_1(x) \ = \ \frac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)},$$

$$\ell_2(x) \ = \ \frac{(x-x_0)(x-x_1)}{(x_2-x_0)(x_2-x_1)}.$$

Obsérvese que los polinomios anteriores verifican los valores siguientes (ver 3.5):

$$\ell_0(x_0) \ = \ 1, \qquad \ell_0(x_1) \ = \ 0, \qquad \ell_0(x_2) \ = \ 0,$$

$$\ell_1(x_0) \ = \ 0, \qquad \ell_1(x_1) \ = \ 1, \qquad \ell_1(x_2) \ = \ 0,$$

$$\ell_2(x_0) \ = \ 0, \qquad \ell_2(x_1) \ = \ 0, \qquad \ell_2(x_2) \ = \ 1.$$

Finalmente, los coeficientes del interpolante son:

$$b_0 \ = \ f(x_0),\ b_1 \ = \ f(x_1),\ b_2 \ = \ f(x_2),$$

que son datos del problema. El polinomio buscado es:

$$I_2(x) \ = \ f(x_0)\ \ell_0(x) \ + \ f(x_1)\ \ell_1(x) \ + \ f(x_2)\ \ell_2(x).$$

3.4. Método recursivo

En la resolución del problema de interpolación polinómica, resulta de gran utilidad contar con un procedimiento recursivo que permita obtener el polinomio interpolante de grado N a partir del polinomio interpolante de grado $N-1$. Dicho de otra forma, al incorporar un nuevo punto de interpolación y a partir del interpolante construido con N puntos, interesa obtener el nuevo interpolante construido con $N+1$ puntos de interpolación. En este apartado se describe un método recursivo para obtener los polinomios de Lagrange cuando, a partir de un conjunto de puntos dados, se considera un nuevo punto de interpolación. Con el fin de desarrollar este tema, a modo de notación, se introduce un subíndice adicional a los polinomios de Lagrange. En lo que sigue, se indicará con ℓ_{ji} al polinomio de Lagrange que verifica el punto de interpolación j construido a partir de i puntos de interpolación, de grado $i-1$. Por tanto, a partir de la expresión (3.3), el polinomio de Lagrange obtenido a partir de i puntos, puede expresarse de la forma siguiente:

$$\ell_{ji}(x) \ = \ \frac{(x-x_0)(x-x_1)\ldots(x-x_{j-1})(x-x_{j+1})\ldots(x-x_{i-1})}{(x_j-x_0)(x_j-x_1)\ldots(x_j-x_{j-1})(x_j-x_{j+1})\ldots(x_j-x_{i-1})},$$

$$j=0,\ldots,N, \qquad i=1,\ldots,N+1.$$

La expresión anterior puede escribirse como:

$$\ell_{ji}(x) = \frac{(x-x_0)\dots(x-x_{i-2})}{(x_j-x_0)\dots(x_j-x_{i-2})}\left(\frac{x-x_{i-1}}{x_j-x_{i-1}}\right), \quad j=0,\dots,N, i=1,\dots,N+1.$$

De esta forma, $\ell_{ji}(x)$ se puede expresar mediante el polinomio de Lagrange ℓ_{ji-1} obtenido a partir de $i-1$ puntos. Finalmente, se obtiene la recursión

$$\ell_{ji}(x) \;=\; \ell_{ji-1}(x)\;\left(\frac{x-x_{i-1}}{x_j-x_{i-1}}\right), \qquad j=0,\dots,N, \qquad i=1,\dots,N+1,$$

que permite obtener el polinomio de Lagrange de grado $i-1$ a partir del polinomio de Lagrange de grado $i-2$, cuando se incorpora el punto de interpolación x_i. La condición inicial del método recursivo propuesto es $\ell_{j1} \;=\; 1$ que corresponde al polinomio de Lagrange, de grado 0, construido con un único punto de interpolación x_j. A continuación, se presenta un ejemplo donde se aplica el procedimiento recursivo explicado.

Ejemplo:
Dada una función f se propone obtener el polinomio interpolante I_2, de grado ≤ 2, partiendo del polinomio de Lagrange de grado 0, que verifique el punto x_0, posteriormente incorporar el punto x_1 para obtener el polinomio de grado 1, y por último, agregar el punto de interpolación x_2 y obtener el polinomio de grado 2. En la figura (3.1) se representa gráficamente el algoritmo recursivo que se explica a continuación.

Se comienza con el punto de interpolación x_0 y el polinomio de Lagrange de grado 0 es:

$$\ell_{01} \;=\; 1,$$

siendo esta la condición inicial del método recursivo. A continuación, se considera el punto x_1 y se calcula el polinomio de Lagrange de grado 1, empleando la expresión (3.4)

$$\ell_{02}(x) \;=\; \ell_{01}\;\frac{x-x_1}{x_0-x_1} \;=\; \frac{x-x_1}{x_0-x_1}.$$

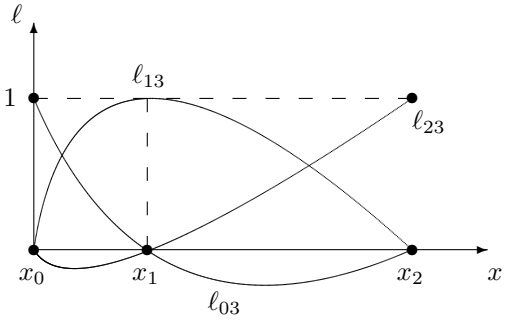
Por último, se considera el punto x_2 y se obtiene el polinomio de Lagrange de grado 2, aplicando nuevamente la expresión (3.4)

$$\ell_{03}(x) \;=\; \ell_{02}(x)\;\frac{x-x_2}{x_0-x_2} \;=\; \frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)}.$$

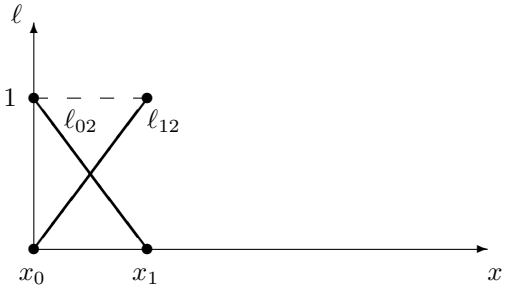
Los polinomios de Lagrange $\ell_{13}(x)$ y $\ell_{23}(x)$ se obtienen de igual forma que $\ell_{03}(x)$. Finalmente, el interpolante de grado ≤ 2 de la función f es:

$$I_2(x) \;=\; f(x_0)\;\ell_{03}(x) \;+\; f(x_1)\;\ell_{13}(x) \;+\; f(x_2)\;\ell_{23}(x),$$

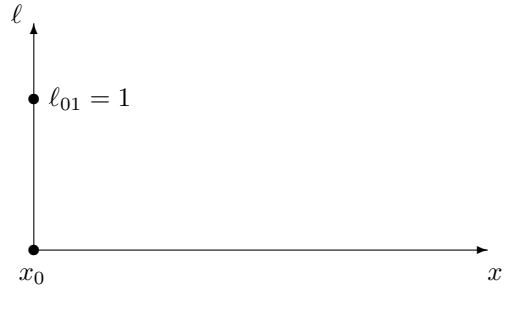
que se representa en la figura (3.2).



grado 2



grado 1



grado 0

Figura 3.1: Algoritmo recursivo para la obtención de los polinomios de Lagrange

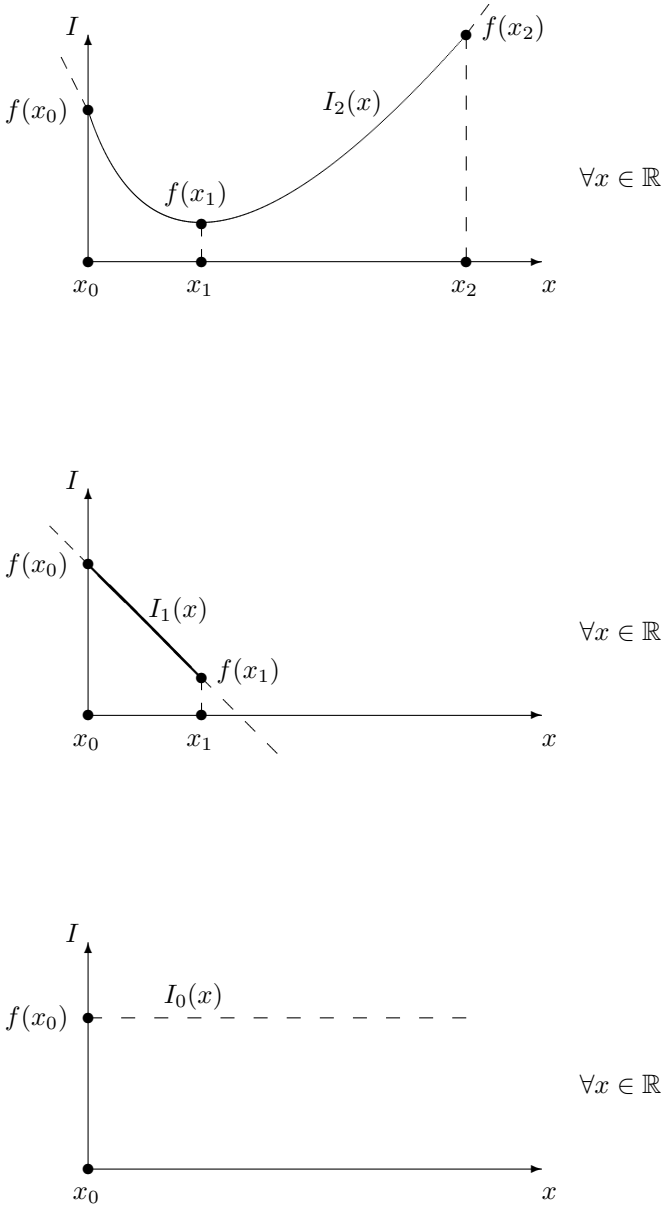


Figura 3.2: Interpolante de Lagrange de grado ≤ 2 de una función f

3.5. Forma de Newton

En esta sección, se presenta otra forma de resolver el problema de interpolación polinómica que se conoce como forma de Newton. Al igual que en la interpolación de Lagrange, este método aporta un procedimiento recursivo para construir polinomios de grado superior, a medida que se añaden puntos de interpolación. Además, este método permite introducir el concepto de cociente incremental, también conocido como de diferencia dividida, que permite definir otro método recursivo para determinar los coeficientes de la forma de Newton.

Se consideran $N+1$ puntos distintos $x_0, x_1, \ldots, x_N$ del dominio de definición de una función f, en los que se conocen los valores $f(x_j), \quad j = 0, 1, \ldots, N$. A continuación, se expresa el polinomio I_N de grado $\leq N$, en la forma siguiente:

$$\begin{aligned} I_N(x) = \; & c_0 + c_1(x - x_0) + c_2(x - x_0)(x - x_1) + \cdots + \\ & + c_N(x - x_0)(x - x_1)(x - x_2)\ldots(x - x_{N-2})(x - x_{N-1}), \end{aligned} \tag{3.7}$$

donde $c_0, c_1, \ldots, c_N$ son los coeficientes de interpolación a calcular. La expresión (3.7) se conoce como forma de Newton del problema de interpolación polinómica y, de forma compacta, se puede expresar como:

$$I_N(x) = \sum_{j=0}^{N} c_j \prod_{i=0}^{j-1}(x - x_i). \tag{3.8}$$

Para calcular los coeficientes de interpolación, se recurre al procedimiento recursivo que se describe a continuación. Como condición inicial, se consideran los datos siguientes: el punto de interpolación x_0 y la imagen de la función $f(x_0)$. A partir del punto x_0, se obtiene el interpolante $I_0(x)$ de grado 0 que verifica la función f en el nodo de interpolación x_0. De esta forma, se obtiene:

$$I_0(x) = f(x_0) = c_0.$$

A partir de $I_0(x)$, se añade un segundo punto x_1 y se calcula el interpolante $I_1(x)$ de grado ≤ 1 de la forma siguiente:

$$I_1(x) = I_0(x) + \varphi_1(x).$$

Para obtener la función $\varphi_1(x)$ se imponen las dos condiciones siguientes: debe tratarse de un polinomio de grado ≤ 1 dado que I_0 es de grado 0, y debe anularse en el punto x_0 para que se cumpla que $I_1(x_0) = I_0(x_0)$. Por tanto, la función $\varphi_1(x)$ debe tener la expresión siguiente:

$$\varphi_1(x) = c_1(x - x_0),$$

donde c_1 es un coeficiente que se determina a partir de la condición $I_1(x_1) = f(x_1)$. Una vez calculado el polinomio I_1, y a partir del mismo, se incorpora otro punto

de interpolación x_2 y se busca el interpolante $I_2(x)$ de grado ≤ 2, que tenga la expresión general

$$I_2(x) = I_1(x) + \varphi_2(x).$$

La función $\varphi_2(x)$ debe tratarse de un polinomio de grado ≤ 2 dado que I_1 es de grado ≤ 1, y debe anularse en los puntos x_1 para que se cumpla que $I_2(x_1) = I_1(x_1)$, y x_0 para que $I_2(x_0) = I_0(x_0)$. A partir de estas condiciones, $\varphi_2(x)$ tendrá la forma:

$$\varphi_2(x) = c_2(x - x_0)(x - x_1),$$

donde el coeficiente c_2 se determina imponiendo la condición $I_2(x_2) = f(x_2)$. Finalmente, el interpolante $I_N(x)$ de grado $\leq N$ se obtiene a partir del polinomio I_{N-1}, añadiendo el punto x_N. La expresión general para este paso es:

$$I_N(x) = I_{N-1}(x) + \varphi_N(x),$$

donde $\varphi_N(x)$ debe ser un polinomio de grado $\leq N$ y debe anularse en los puntos $x_{N-1}, x_{N-2}, \ldots, x_1, x_0$. La función $\varphi_N(x)$ que cumple las condiciones citadas anteriormente es:

$$\varphi_N(x) = c_N(x - x_0)(x - x_1)\ldots(x - x_{N-2})(x - x_{N-1}).$$

El coeficiente c_N se determina imponiendo la condición $I_N(x_N) = f(x_N)$. Finalmente, se puede expresar la recursión siguiente:

$$I_i(x) = I_{i-1}(x) + \varphi_i(x), \tag{3.9}$$

siendo

$$\varphi_i(x) = c_i(x - x_0)(x - x_1)\ldots,(x - x_{i-1}) = c_i \prod_{j=0}^{i-1}(x - x_j),$$

donde los coeficientes c_i se obtienen imponiendo

$$I_i(x_i) = f(x_i), \quad i = 0, \ldots, N.$$

Esta recursión permite construir el interpolante de grado i, en la forma de Newton, a partir del polinomio de Newton de grado $i - 1$, cuando se incorpora el punto de interpolación x_i.

Ejemplo:
Dada una función f y 3 puntos de interpolación x_0, x_1, x_2 del dominio de definición, se propone obtener el polinomio interpolante I_2, de grado ≤ 2, en la forma de Newton, que verifique que

$$I_2(x_0) = f(x_0) = y_0, \quad I_2(x_1) = f(x_1) = y_1, \quad I_2(x_2) = f(x_2) = y_2.$$

Se parte de los datos iniciales x_0 y $f(x_0)$ y se obtiene el polinomio de grado 0 siguiente:

$$I_0(x) = f(x_0) = c_0.$$

A partir del interpolante anterior, se añade el punto de interpolación x_1 y se calcula el interpolante de grado ≤ 1, aplicando la recursión (3.9). La expresión resultante es:

$$I_1(x) = I_0(x) + c_1(x - x_0),$$

donde c_1 se calcula forzando a que $I_1(x_1) = f(x_1)$, de la forma siguiente:

$$I_1(x_1) = f(x_1),$$

$$I_0(x_1) + c_1(x_1 - x_0) = f(x_1),$$

$$c_1 = \frac{f(x_1) - I_0(x_1)}{x_1 - x_0},$$

$$c_1 = \frac{f(x_1) - f(x_0)}{x_1 - x_0}.$$

El interpolante de grado ≤ 1 resulta:

$$I_1(x) = f(x_0) + \frac{f(x_1) - f(x_0)}{x_1 - x_0}\ (x - x_0).$$

A partir de I_1, se incorpora el punto x_2 y, aplicando la recursión (3.9), se construye $I_2(x)$ de grado ≤ 2:

$$I_2(x) = I_1(x) + c_2(x - x_0)(x - x_1),$$

donde c_2 se calcula forzando a que $I_2(x_2) = f(x_2)$, de la forma siguiente:

$$I_2(x_2) = f(x_2),$$

$$I_1(x_2) + c_2(x_2 - x_0)(x_2 - x_1) = f(x_2),$$

$$c_2 = \frac{f(x_2) - I_1(x_2)}{(x_2 - x_0)(x_2 - x_1)},$$

$$c_2 = \frac{f(x_2) - f(x_0) - \dfrac{f(x_1) - f(x_0)}{x_1 - x_0}\ (x_2 - x_0)}{(x_2 - x_0)(x_2 - x_1)},$$

$$c_2 = \frac{\dfrac{f(x_2) - f(x_1)}{x_2 - x_1} - \dfrac{f(x_1) - f(x_0)}{x_1 - x_0}}{(x_2 - x_0)}.$$

El interpolante de grado ≤ 2 resulta:

$$\begin{aligned} I_2(x) =& f(x_0)+ \\ & \left[\frac{f(x_1)-f(x_0)}{x_1-x_0}\right](x-x_0)+ \\ & \left[\frac{\dfrac{f(x_2)-f(x_1)}{x_2-x_1} - \dfrac{f(x_1)-f(x_0)}{x_1-x_0}}{(x_2-x_0)}\right](x-x_0)(x-x_1). \end{aligned}$$

La forma de Newton permite introducir un nuevo concepto que se conoce como *cociente de incrementos* o *diferencias divididas*, que aporta un algoritmo recursivo para la determinación de los coeficientes de interpolación.

Definición: Diferencia dividida.

Dada una función f y un conjunto de $N+1$ puntos distintos $x_0, x_1, \ldots, x_N$ del dominio de definición de f, se define como diferencia dividida de la función f en los $N+1$ puntos a los coeficientes del interpolante polinómico en la forma de Newton.

La diferencia dividida de orden 0 de la función f con respecto al punto x_0 es el valor de la función en dicho punto y se indica de la forma siguiente:

$$f[x_0] = f(x_0).$$

La diferencia dividida de orden 1 de la función f con respecto a los puntos x_0 y x_1 tiene la expresión siguiente:

$$f[x_0, x_1] = \frac{f(x_1)-f(x_0)}{x_1-x_0},$$

que es, por definición, el coeficiente del monomio x del interpolante de Newton (ver ejemplo anterior).

Si se añade el punto siguiente x_2, la diferencia dividida de orden 2 de la función f con respecto a los puntos x_0, x_1 y x_2, resulta:

$$f[x_0, x_1, x_2] = \frac{\dfrac{f(x_2)-f(x_1)}{x_2-x_1} - \dfrac{f(x_1)-f(x_0)}{x_1-x_0}}{(x_2-x_0)},$$

que es, por definición, el coeficiente del monomio x^2 del interpolante de Newton (ver ejemplo anterior). Se puede observar que la diferencia dividida de segundo orden, se puede expresar en función de las diferencias divididas $f[x_0, x_1]$ y $f[x_1, x_2]$, de

primer orden, tal como se indica a continuación:

$$f[x_0, x_1, x_2] = \frac{f[x_1, x_2] \; - \; f[x_0, x_1]}{(x_2 - x_0)}.$$

De forma recursiva, se pueden ir añadiendo puntos e ir obteniendo diferencias divididas de orden superior, o lo que es equivalente, se pueden ir obteniendo los coeficientes de los mononios de mayor potencia del interpolante correspondiente. Por tanto, a partir de las diferencias divididas de orden $i-1$, se puede construir la diferencia dividida de orden i mediante la expresión siguiente:

$$f[x_0, x_1, \ldots, x_i] = \frac{f[x_1, x_2, \ldots, x_i] - f[x_0, x_1, \ldots, x_{i-1}]}{x_i - x_0}. \tag{3.10}$$

Obsérvese que si se permutan los argumentos x_0 y x_1 en la diferencia dividida de orden 1, el resultado no cambia. Es decir que:

$$f[x_0, x_1] \; = \; f[x_1, x_0].$$

Por otro lado, si se permutan los argumentos x_0, x_1 y x_2 en la diferencia dividida de orden 2, el resultado de las $3! = 6$ permutaciones posibles dan el mismo resultado. Es decir que:

$$\begin{aligned} f[x_0, x_1, x_2] \; &= \; f[x_0, x_2, x_1] \; = \; f[x_1, x_0, x_2] \; = \\ f[x_1, x_2, x_0] \; &= \; f[x_2, x_0, x_1] \; = \; f[x_2, x_1, x_0]. \end{aligned}$$

Se puede demostrar que cualquier permutación de los argumentos no modifica el valor de la diferencia dividida. Es decir que para la diferencia dividida de orden N, se puede expresar que:

$$f[x_0, x_1, \ldots, x_{N-1}, x_N] \; = \; f[x_{\alpha_0}, x_{\alpha_1}, \ldots, x_{\alpha_N}],$$

donde la serie finita de enteros $\alpha_0, \alpha_1, \ldots, \alpha_N$ es cualquier permutación de $0, 1, \ldots, N$.

Para demonstrar las anteriores propiedades se considera la fórmula $P_1(x)$ de Newton del polinomio interpolante de grado N que pasa los puntos: $x_0, x_1, \ldots x_N$

$$P_1(x) = f[x_0] + f[x_0, x_1](x - x_0) \ldots + f[x_0, \ldots x_N](x - x_0) \ldots (x - x_{N-1}).$$

Ahora consideramos la siguiente permutación de los valores nodales: $x_N, x_{N-1}, \ldots x_0$ y construimos el polinomio interpolante $P_2(x)$

$$P_2(x) = f[x_N] + f[x_N, x_{N-1}](x - x_N) \ldots + f[x_N, \ldots x_0](x - x_N) \ldots (x - x_1).$$

Por la unicidad del polinomio interpolante $P_1(x)$ y $P_2(x)$ deben ser el mismo. Identificando los coeficientes de orden máximo x^N se obtiene:

$$f[x_0, \ldots x_N] = f[x_N, \ldots x_0]. \tag{3.11}$$

Es decir, la diferencia dividida de orden N es la misma para la secuencia de puntos $x_0, \dots x_N$ y para la secuencia $x_N, \dots x_0$ y, en general, para cualquier otra permutación.

Si identificamos ahora el coeficiente de orden x^{N-1}, se obtiene:

$$f[x_0, x_1, \dots x_{N-1}] - f[x_0, x_1, \dots x_N](x_0 + x_1 + \dots x_{N-1}) =$$
$$f[x_N, x_{N-1}, \dots x_1] - f[x_N, x_{N-1}, \dots x_0](x_N + x_{N-1} + \dots x_1).$$

Si utilizamos la propiedad (3.11) y despejamos la diferencia dividida de orden N en función de la diferencia dividida de orden $N-1$, se obtiene

$$f[x_0, \dots x_N] = \frac{f[x_1, \dots x_N] - f[x_0, \dots x_{N-1}]}{x_N - x_0}. \tag{3.12}$$

Por último, dados $N+1$ puntos distintos $x_0, x_1, \dots, x_N$ del dominio de definición de una función $f(x)$ y si se definen las diferencias divididas de orden cero como $f[x_k] = f(x_k), \quad k = 0, 1, \dots, N$, empleando los conceptos explicados anteriormente, se puede construir la tabla de diferencias divididas siguiente:

$$\begin{array}{llllll}
x_0 & f[x_0] & & & & \\
 & & f[x_0, x_1] & & & \\
x_1 & f[x_1] & & f[x_0, x_1, x_2] & & \\
 & & f[x_1, x_2] & & \ddots & \\
x_2 & f[x_2] & & f[x_1, x_2, x_3] & & \\
 & & f[x_2, x_3] & & \ddots & \\
 & & & & & \\
\vdots & \vdots & \vdots & \vdots & \dots & f[x_0, x_1, \dots, x_N] \\
 & & & & & \\
 & & f[x_{N-3}, x_{N-2}] & & & \\
x_{N-2} & f[x_{N-2}] & & f[x_{N-3}, x_{N-2}, x_{N-1}] & & \\
 & & f[x_{N-2}, x_{N-1}] & & & \ddots \\
x_{N-1} & f[x_{N-1}] & & f[x_{N-2}, x_{N-1}, x_N] & & \\
 & & f[x_{N-1}, x_N] & & & \ddots \\
x_N & f[x_N] & & & &
\end{array} \tag{3.13}$$

La columna primera indica el conjunto de $N+1$ puntos distintos del dominio de definición de f, y cada columna, a partir de la segunda, representa las diferencias divididas de orden 0, 1, 2 hasta orden N. Moviéndose a través de la diagonal de la tabla de diferencias divididas se puede obtener la diferencia dividida de orden i a partir de las diferencias divididas de orden $i-1$.

Finalmente, el interpolante de grado $\leq N$, en la forma de Newton, se puede expresar en la forma siguiente:

$$\begin{aligned} I_N(x) = \; & f[x_0] + f[x_0, x_1]\ (x - x_0) + f[x_0, x_1, x_2]\ (x - x_0)(x - x_1) + \cdots + \\ & + f[x_0, x_1, \ldots, x_N]\ (x - x_0)(x - x_1)(x - x_2)\ldots(x - x_{N-2})(x - x_{N-1}), \end{aligned}$$

donde los coeficientes han sido reemplazados por las diferencias divididas del orden correspondiente que obtienen de forma recursiva mediante la expresión (3.12).

Es importante hacer notar que con la fórmula de Newton, el polinomio de grado N se construye añadiendo un punto $(x_N, f(x_N))$ al polinomio de grado $N-1$. Es decir,

$$I_N(x) = I_{N-1}(x) + f[x_0, x_1, \ldots, x_N]\ (x - x_0)\ldots(x - x_{N-1}). \tag{3.14}$$

3.6. Error de interpolación

Una vez analizadas las tres posibles formas de un polinomio interpolante, resulta de gran importancia conocer la diferencia que existe entre la función $f(x)$ a interpolar y sus derivadas y el interpolante $I_N(x)$ en el dominio de definición de la función.

De forma general, el error total al aproximar una función proviene de dos fuentes o componentes de error que se explican a continuación. Por un lado, existe un error asociado al método de aproximación empleado que se conoce como *error de truncamiento*. Este error se puede analizar cuestionando la igualdad siguiente:

$$\lim_{N \to \infty} I_N(x) \;=\; f(x). \tag{3.15}$$

Por otro lado, existe un error asociado a la sensibilidad del método de aproximación a perturbaciones que se conoce como condicionamiento del método de aproximación. Al implementar un método de aproximación en el ordenador, aparecen perturbaciones asociadas a la aritmética de precisión finita por lo que este error se lo conoce como *error de redondeo*. Por lo tanto, el error total de aproximación se expresa de la forma siguiente:

$$E_{\text{total}} \;=\; R_N(x) \;+\; R_L(x), \tag{3.16}$$

donde $R_N(x)$ es el error de truncamiento y $R_L(x)$ es el error de redondeo asociado a las perturbaciones de la función $f(x)$ a interpolar.

En las secciones siguientes se estudian con detalle el error de truncamiento a partir del teorema del error de interpolación y el error de redondeo a partir de la función de Lebesgue.

3.6.1. Error de truncamiento

Se pretende obtener una expresión analítica para el error de truncamiento que permita analizar las diferentes fuentes de error. El siguiente teorema nos relaciona la nube de puntos y la regularidad de la función con el error de truncamiento.

Teorema: Error de truncamiento.
Sea f una función de clase C^{N+1}, definida en un intervalo $[a,b]$ y sean $x_0, x_1, \ldots, x_N$, $N+1$ puntos de interpolación distintos, pertenecientes al intervalo $[a,b]$, de los que se obtiene el polinomio interpolante I_N de grado $\leq N$. Entonces, a cada punto x perteneciente al intervalo $[a,b]$, le corresponde un punto ξ, que corresponde al intervalo:

$$\text{mín}(x_0, x_1, \ldots, x_N, x) < \xi < \text{máx}(x_0, x_1, \ldots, x_N, x), \tag{3.17}$$

para el que se cumple la expresión siguiente que corresponde al error de interpolación y que se indica con $R_N(x)$:

$$R_N(x) \;=\; f(x) \;-\; I_N(x) \;=\; \pi_{N+1}(x)\;\frac{f^{(N+1)}(\xi)}{(N+1)!}, \tag{3.18}$$

donde π_{N+1} es la función de error dada por :

$$\pi_{N+1}(x) = (x - x_0)(x - x_1)\ldots(x - x_N). \tag{3.19}$$

Es importante destacar que, a priori, el punto ξ se desconoce y que es función de x, es decir que $\xi = \xi(x)$.

Demostración:
Para demostrar el teorema del error de interpolación, se plantean dos pasos previos, que se indican a continuación. En un primer paso se demostrará que el error de interpolación de grado N se puede expresar a partir de las diferencias divididas mediante la expresión siguiente:

$$R_N(x) \;=\; \pi_{N+1}(x)\; f[x, x_0, x_1, \ldots, x_N], \tag{3.20}$$

y en un segundo paso, se demostrará que:

$$f[x, x_0, x_1, \ldots, x_N] \;=\; \frac{f^{N+1}(\xi)}{(N+1)!}. \tag{3.21}$$

Para comprobar (3.20) se demuestra mediante la relación de recurrencia (3.14) que el error de interpolación $R_N(x)$ es:

$$R_N(x) \;=\; \pi_{N+1}(x)\; f[x, x_0, x_1, \ldots, x_N]. \tag{3.22}$$

Se parte de $N+1$ puntos de interpolación $x_0, \dots x_N$ y se construye el interpolante de grado N. A continuación, se añade el punto genérico $(t, f(t))$ y se construye el interpolante mediante la relación de recurrencia (3.14),

$$I(x) \ = I_N(x) + f[x_0, x_1, \dots x_N, t] \ (x - x_0) \dots (x - x_{N-1})(x - x_N). \tag{3.23}$$

Como el punto añadido es genérico el interpolante con $t = x$ coincide con la función a interpolar ($I(x) = f(x)$). Así, la diferencia entre la función $f(x)$ y el interpolante $I_N(x)$ es el error de truncamiento

$$R_N(x) \ = f[x_0, x_1, \dots x_N, x] \ (x - x_0) \dots (x - x_{N-1})(x - x_N),$$

que es la expresión que queríamos demostrar.

En un segundo paso, se pretende demostrar (3.21), es decir que la diferencia dividida del paso anterior es:

$$f[x, x_0, x_1, \dots, x_N] \ = \ \frac{f^{N+1}(\xi)}{(N+1)!}. \tag{3.24}$$

En principio se sabe que:

$$R_N(x) \ = \ \pi_{N+1}(x) \ f[x, x_0, x_1, \dots, x_N] \ = \ f(x) \ - \ I_N(x). \tag{3.25}$$

Para esta demostración, se construye una función $Q(x)$ a partir de la expresión siguiente:

$$Q(x) \ = \ f(x) \ - \ I_N(x) \ - \ \pi_{N+1}(x) \ G(t), \tag{3.26}$$

donde $G(t)$ es:

$$G(t) \ = \ f[t, x_0, \dots, x_N]. \tag{3.27}$$

A continuación se analizan los ceros de la función Q y de sus derivadas sucesivas.

La función Q tiene $N+2$ ceros que se describen a continuación:

$$\begin{cases} \text{N + 1 ceros en:} & x_j, \quad j = 0, \dots, N, \\ \\ \text{1 cero en :} & x = t. \end{cases} \tag{3.28}$$

Los $N+1$ primeros ceros provienen de los $N+1$ puntos de interpolación

$$x_j, \quad j = 0, \dots, N.$$

En ellos, por definición de interpolación, se cumple que:

$$f(x_j) \ - \ I_N(x_j) \ = \ 0. \tag{3.29}$$

Por otro lado, la función $\pi_{N+1}(x_j) = 0$. Finalmente, la función Q se anula en los $N+1$ puntos de interpolación:

$$Q(x_j) \ = \ f(x_j) \ - \ I_N(x_j) \ - \ \pi_{N+1}(x_j) \ G(t) \ = \ 0, \qquad j = 0, \dots, N. \tag{3.30}$$

Además, la función Q tiene un cero en $x = t$, por definición de interpolación:

$$Q(t) \;=\; f(t) \;-\; I_N(t) \;-\; \pi_{N+1}(t)\, G(t) \;=\; 0. \tag{3.31}$$

Si se deriva Q una vez respecto de x, se observa que Q' tiene $N+1$ ceros, que Q'' tiene N ceros, Q''' tiene $N-1$ ceros y así hasta la derivada de orden $N+1$ de Q que tiene la expresión siguiente:

$$Q^{(N+1)}(x) \;=\; f^{(N+1)}(x) \;-\; I_N^{(N+1)}(x) \;-\; (N+1)!\, G(t), \tag{3.32}$$

en la que $I_N^{(N+1)} = 0$. Finalmente, la función $Q^{(N+1)}$ tiene la expresión siguiente:

$$Q^{(N+1)}(x) \;=\; f^{(N+1)}(x) \;-\; (N+1)!\, f[t, x_0, x_1, \ldots, x_N], \tag{3.33}$$

donde por el teorema de Rolle, tiene un cero en algún valor $x = \xi$, que permite expresar que:

$$f^{(N+1)}(\xi) \;=\; (N+1)!\, G(t) \;=\; (N+1)!\, f[t, x_0, x_1, \ldots, x_N], \tag{3.34}$$

y donde ξ es función de $t, x_0, x_1, \ldots, x_N$. De esta forma queda demostrado (3.21)

$$f[x, x_0, x_1, \ldots, x_N] \;=\; \frac{f^{N+1}(\xi)}{(N+1)!}. \tag{3.35}$$

A partir de los resultados (3.20) y (3.21) se demuestra el teorema del error de interpolación:

$$R_N(x) \;=\; \pi_{N+1}(x)\, \frac{f^{(N+1)}(\xi)}{(N+1)!}, \tag{3.36}$$

donde ξ pertenece al intervalo:

$$\text{mín}(x_0, x_1, \ldots, x_N, x) \;<\; \xi \;<\; \text{máx}(x_0, x_1, \ldots, x_N, x).$$

De la expresión del error de truncamiento (3.36) se observa que el error está relacionado con dos cuestiones, una de ellas es la regularidad de la función a interpolar y la otra son los puntos de interpolación. Con respecto a la regularidad de la función, se observa que la misma debe ser derivable hasta el orden $N+1$. Independientemente de su regularidad, si f^{N+1} es grande o pequeña, el error de interpolación será grande o pequeño. En particular, si f^{N+1} no está acotada con N en el dominio de definición, el error de interpolación resultará no acotado. Con respecto a los puntos de interpolación, se observa que afectan al error de interpolación en el número de puntos a través de $(N+1)!$ y en la distribución de puntos de interpolación a través de la función $\pi_{N+1}(x)$.

3.6.2. Error de redondeo

Siempre que interpolamos una función aparecen errores de redondeo asociados a la precisión del ordenador. Si trabajamos con precisión simple estos errores son del $O(10^{-7})$ y con precisión doble de $O(10^{-15})$. Este error que introduce el ordenador por la precisión finita del mismo se le conoce como error de redondeo o *round–off* y se le denomina por la letra ϵ. De esta manera, al introducir en el ordenador para el cálculo del interpolante la evaluación de la función $f(x)$ en el punto nodal x_j aparece un error de redondeo que denominamos por la variable ϵ_j. Es decir, ϵ_j es la indeterminación que tenemos en el ordenador al evaluar la función en el nodo x_j. Sin embargo, estos valores se pueden acotar por el valor ϵ asociado a la precisión con la que trabajemos.

De esta manera, en el interpolante sustituimos $f(x_j)$ por $f(x_j)+\epsilon_j$ para tener en cuenta precisión finita del ordenador. Esta indeterminación la denominamos por error de redondeo y tiene la expresión:

$$R_L(x) \;=\; \epsilon_0\;\ell_0(x)+\epsilon_1\;\ell_1(x)+\ldots+\epsilon_N\;\ell_N(x). \tag{3.37}$$

Aunque no podamos conocer el error de interpolación, nos interesa acotar este error para saber si el problema de interpolación está bien condicionado. Si tomamos el valor absoluto en (3.37), se obtiene la siguiente expresión mediante la desigualdad triangular:

$$|R_L(x)| \;\leq\; |\epsilon_0|\;|\ell_0(x)|+\ldots+|\epsilon_N|\;|\ell_N(x)|. \tag{3.38}$$

Si consideramos que todos los errores de redondeo ϵ_j de cada nodo x_j están acotados por ϵ, entonces

$$|R_L(x)| \;\leq\; \epsilon\;\lambda_N(x), \tag{3.39}$$

donde $\lambda_N(x)$ es la función de Lebesgue definida como:

$$\lambda_N(x) \;=\; \sum_{j=0}^{N}\;|\ell_j(x)|\,. \tag{3.40}$$

Si ahora calculamos el máximo del error de interpolación en todo el dominio de interpolación, la cota superior viene determinada por la constante de Lebesgue.

$$\max_{x\in[a,b]}|R_L(x)| \;\leq\; \epsilon\;\Lambda_N, \tag{3.41}$$

donde Λ_N es la constante de Lebesgue dada por:

$$\Lambda_N \;=\; \max_{x\in[a,b]}\lambda_N(x). \tag{3.42}$$

La constante de Lebesgue también está relacionada con la mejor aproximación que podemos hacer mediante un interpolante.

Teorema: Error de interpolación con respecto al mejor interpolante.
Sea Λ_N la constante de Lebesgue dada por (3.42), f una función definida en $[a,b]$, I_N el polinomio interpolante correspondiente y I_N^* el mejor polinomio interpolante. Entonces,

$$\|f - I_N\| \leq (\Lambda_N + 1)\ \|f - I_N^*\|. \tag{3.43}$$

Si la constante de Lebesgue de nuestro conjunto de puntos $x_j, j = 0, \ldots N$ es grande, el error interpolación con respecto al error del interpolante óptimo es grande.

Demostración:

En primer lugar tenemos que relacionar la norma del interpolante con la constante de Lebesgue mediante su definición. Si I_N es un interpolante de f de grado N, mediante la definición (3.42) entonces $\|I_N\| \leq \|f\|\ \Lambda_N$. Si expresamos $f - I_N$ como $f - I_N^* - I_N + I_N^*$ y utilizamos la desigualdad triangular, entonces

$$\|f - I_N\| \leq\ \|f - I_N^*\| + \|I_N - I_N^*\|.$$

El polinomio $Q_N = I_N - I_N^*$ es un interpolante $f - I_N^*$ porque

$$Q_N(x_j) = I_N(x_j) - I_N^*(x_j) = f(x_j) - I_N^*(x_j).$$

De esta manera y mediante la definición de la constante de Lebesgue

$$\|I_N - I_N^*\| \leq \Lambda_N \|f - I_N^*\|,$$

y

$$\|f - I_N\| \leq (\Lambda_N + 1)\ \|f - I_N^*\|, \tag{3.44}$$

con lo que el teorema queda demostrado.

3.6.3. Acotación del error de truncamiento y redondeo

Una vez analizado el origen y las expresiones analíticas para el error de interpolación, pasamos a estudiar su comportamiento con las distribuciones nodales de puntos y con el número de puntos N. La expresión para el error total de interpolación es:

$$E_N(x)\ =\ \pi_{N+1}(x)\ \frac{f^{(N+1)}(\xi)}{(N+1)!}\ +\ \sum_{j=0}^{N}\ \epsilon_j\ \ell_j(x), \tag{3.45}$$

donde el primer término es el error de truncamiento $R_N(x)$ y el segundo el error de redondeo $R_L(x)$. Tomando la norma del supremo en esta expresión y utilizando la desigualdad triangular, se obtiene:

$$\|E_N\|\ \leq\ \|\pi_{N+1}\|\ \frac{\|f^{(N+1)}(\xi)\|}{(N+1)!}\ +\ \|\epsilon\|\ \Lambda_N. \tag{3.46}$$

Antes de estudiar el comportamiento de los términos anteriores hacemos un cambio de variable x independiente para transformar el dominio $[a, b]$ en el dominio $[-1, 1]$ mediante la siguiente expresión:

$$x = \frac{a+b}{2} + \frac{b-a}{2}t, \tag{3.47}$$

donde $t \in [-1, 1]$. Tanto en los polinomio de Lagrange con en la función de error π_{N+1} aparecen los factores $x - x_j$ que al hacer el cambio de variable quedan:

$$x - x_j = \frac{b-a}{2}(t - t_j), \tag{3.48}$$

siendo t_j una partición del intervalo $[-1, 1]$. Así, la función de error π_{N+1} queda

$$\pi_{N+1}(t) = (t - t_0)(t - t_1)\dots(t - t_N)\left(\frac{b-a}{2}\right)^{N+1}, \tag{3.49}$$

y los polinomios de Lagrange

$$\ell_j(t) = \frac{(t - t_0)\dots(t - t_{j-1})(t - t_{j+1})\dots(t - t_N)}{(t_j - t_0)\dots(t_j - t_{j-1})(t_j - t_{j+1})\dots(t_j - t_N)} \tag{3.50}$$

Por otra parte, la derivada $N + 1$ veces de f con respecto a t queda:

$$\frac{d^{N+1}f}{dx^{N+1}} = \frac{d^{N+1}f}{dt^{N+1}}\left(\frac{2}{b-a}\right)^{N+1}. \tag{3.51}$$

Al llevar estas expresiones al error de truncamiento, el segundo factor de (3.51) se cancela con el segundo factor de (3.49) quedando las mismas expresiones que en la variable independiente x. Por lo tanto, sin ninguna pérdida de generalidad se puede suponer que en todo el análisis que sigue el dominio de interpolación es el $[-1, 1]$.

A la vista de la expresión (3.46) se plantean las siguientes cuestiones:

1. Es necesario conocer el comportamiento de $\|\pi_{N+1}\|$ con $N \to \infty$ y la distribución de la nube de puntos.
2. Cuando se interpola con alto orden N, se debe exigir la existencia al menos de la derivada $N + 1$ de la función a interpolar.
3. Es necesario conocer el comportamiento de la constante de Lebesgue Λ_N con N y con la distribución de puntos.

Si los puntos de interpolación están equiespaciados a una distancia Δx en el dominio $[-1, 1]$, veremos en las siguientes secciones que la función de error π_{N+1} está acotada por

$$\|\pi_{N+1}\| \;\le\; \frac{\Delta x^{N+1}}{4}\; N! \;=\; \left(\frac{2}{e}\right)^N \sqrt{\frac{2\pi}{N}},$$

que tiende a cero con $N \to \infty$. Sin embargo, la constante de Lebesgue se puede acotar inferior y superiormente por

$$\frac{2^{N-2}}{N^2} \ < \ \Lambda_N \ < \ \frac{2^{N+3}}{N},$$

y se observa que no tiende a cero con $N \to \infty$. Esta característica asociada a las distribuciones de puntos equiespaciadas invalida su uso en las interpolaciones de alto orden.

En la sección siguiente se analizará el comportamiento del error con diferentes distribuciones de puntos y se llegará a la conclusión de que para paliar el efecto anteriormente descrito se deben concentrar puntos en los extremos del intervalo de interpolación.

3.7. Distribuciones equiespaciadas de puntos

En este caso, se considera una función f y $N+1$ puntos del dominio de definición de f a partir de los cuales se construye el interpolante I_N. Se propone estudiar el comportamiento del error de interpolación a través de $\pi_{N+1}(x)$ y de Λ_N con $N \to \infty$ y una distribución equiespaciada de puntos.

3.7.1. Error de truncamiento

En este apartado se estudia el comportamiento del error de interpolación cuando el número de puntos tiende a infinito, con una distribución equiespaciada de los mismos. En particular, se explica por qué el error de interpolación es mayor en los extremos del intervalo de interpolación y se determina cuál es el valor máximo de este error. Se plantea una distribución equiespaciada de $N+1$ puntos de interpolación donde la distancia entre puntos consecutivos es Δx. Para dos puntos consecutivos genéricos (x_j, x_{j+1}) la distancia tiene la expresión siguiente:

$$\Delta x = x_{j+1} - x_j.$$

Esta distribución se muestra en la figura (3.3).

Del teorema del error de interpolación se tiene que $\pi_{N+1}(x)$ es:

$$\pi_{N+1}(x) = (x - x_0)(x - x_1) \cdots (x - x_N). \tag{3.52}$$

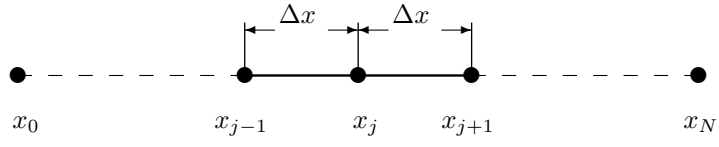


Figura 3.3: Distribución equiespaciada de puntos de interpolación

Se consideran dos puntos de interpolación consecutivos genéricos x_j y x_{j+1}. Suponiendo que $\pi_{N+1}(x)$ alcanza un extremo en un valor de x que pertenece al intervalo (x_j, x_{j+1}), se propone demostrar que:

$$|(x - x_j)(x - x_{j+1})| \leq \frac{\Delta x^2}{4}. \tag{3.53}$$

Para ello, en los pasos siguientes, se busca el extremo de la función $y = (x - x_j)(x - x_{j+1})$:

$$y = (x - x_j)(x - x_{j+1}),$$

$$y' = x - x_{j+1} + x - x_j = 0,$$

$$x_{\text{extremo}} = \frac{1}{2}(x_j + x_{j+1}),$$

$$y_{\text{extremo}} = \left[\frac{1}{2}(x_j + x_{j+1}) - x_j\right]\left[\frac{1}{2}(x_j + x_{j+1}) - x_{j+1}\right],$$

$$y_{\text{extremo}} = \left[\tfrac{1}{2}(x_{j+1} - x_j)\right]\left[\tfrac{1}{2}(x_j - x_{j+1})\right],$$

$$y_{\text{extremo}} = \left(\frac{\Delta x}{2}\right)\left(-\frac{\Delta x}{2}\right) = -\frac{\Delta x^2}{4}.$$

Se observa que el extremo de $\pi_{N+1}(x)$, en caso de obtenerse en el intervalo (x_j, x_{j+1}), se alcanza en el punto medio de intervalo y su valor es $\dfrac{\Delta x^2}{4}$ (ver figura 3.4). Aplicando módulos a la expresión anterior, se demuestra (3.53).

A continuación, se plantea el caso general de una interpolación global tal como se representa en la figura (3.5). Observando la misma se puede comprobar que:

$$\begin{aligned} &|x - x_i| \leq (j - i + 1)\Delta x, && \text{para} \quad i < j, \\ &|x - x_i| \leq (i - j)\Delta x, && \text{para} \quad i > j + 1. \end{aligned} \tag{3.54}$$

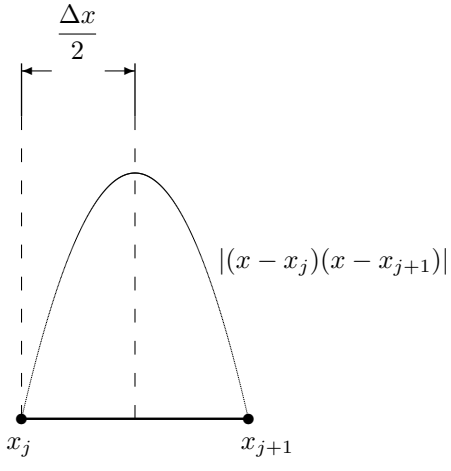


Figura 3.4: Representación de $|\pi_{N+1}(x)|$ entre los puntos de interpolación x_j y x_{j+1}

En la figura (3.5) el índice i indica puntos de interpolación genéricos por delante y por detrás del intervalo (x_j, x_{j+1}) donde se supone que existe el máximo de $|\pi_{N+1}(x)|$.

Aplicando módulos a la expresión (3.52), se tiene que:

$$|\pi_{N+1}(x)| \leq |(x-x_0)|\cdots|(x-x_{j-1})|\ \frac{\Delta x^2}{4}\ |(x-x_{j+2})|\cdots|(x-x_{N-1})||(x-x_N)|,$$

y haciendo uso de las desigualdades (3.54), se llega a que:

$$|\pi_{N+1}(x)| \leq (j+1)\ j\ (j-1)\cdots 2\ \Delta x^j\ \frac{\Delta x^2}{4}\ \ 2\ \ 3\ \cdots\ (N-j)\Delta x^{N-j-1},$$

$$|\pi_{N+1}(x)| \leq (j+1)!\ (N-j)!\ \frac{\Delta x^{N+1}}{4}.$$

La expresión anterior da una cota de $|\pi_{N+1}(x)|$, si se considera que el máximo está en el intervalo (x_j, x_{j+1}). Si ahora se mueve el índice j desde $j = 0, 1, \ldots, N-1$, se puede comprobar que los valores máximos se presentan en los extremos del dominio de interpolación, correspondiendo a los índices $j = 0$ y $j = N-1$. Es decir, los máximos de $|\pi_{N+1}(x)|$ se alcanzan en los intervalos (x_0, x_1) y (x_{N-1}, x_N) y están acotados por:

$$\max_{x\in[a,b]} |\pi_{N+1}(x)| \leq\ N!\ \frac{\Delta x^{N+1}}{4}. \tag{3.55}$$

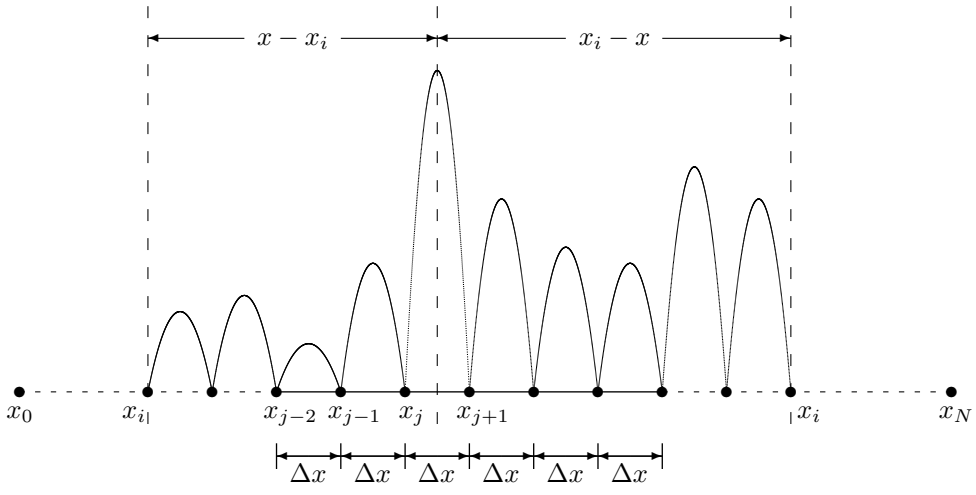


Figura 3.5: Extremos de $|\pi_{N+1}(x)|$ en un caso general de una interpolación global equiespaciada

Utilizando la fórmula de Stirling para el factorial de N y sustituyendo Δx por $2/N$, se obtiene:

$$|\pi_{N+1}(x)| \leq \left(\frac{2}{e}\right)^N \sqrt{\frac{2\pi}{N}}. \tag{3.56}$$

El resultado (3.56) tiene una gran importancia ya que permite conocer una cota del error de interpolación en función del número de puntos elegido. La cota superior de la función de error π_{N+1} tiende a cero con $N \to \infty$. De forma esquemática, en la figura (3.6) se representa $|\pi_{N+1}(x)|$ para una interpolación equiespaciada genérica donde puede observarse que los errores son importantes en los extremos del dominio de interpolación y son pequeños en las proximidades del punto medio de dicho dominio.

Tomando logaritmos en la expresión (3.56) se obtiene:

$$\log \|\pi_{N+1}(x)\| \leq -\frac{1}{2}\log N - N \log\left(\frac{e}{2}\right) + \frac{\log(2\pi)}{2}.$$

De forma esquemática, la expresión anterior tiene la representación gráfica que se indica en la figura (3.7) y que representa la convergencia de $\|\pi_{N+1}(x)\|$ con el número de puntos N en una interpolación equiespaciada.

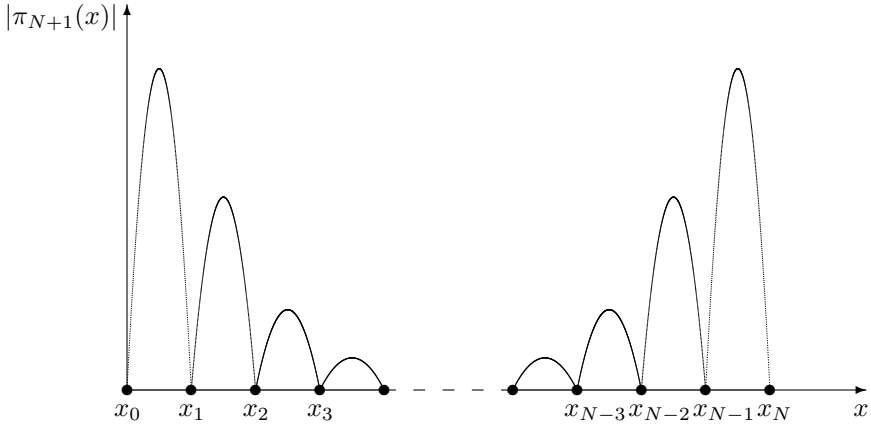


Figura 3.6: $|\pi_{N+1}(x)|$ en una interpolación global equiespaciada

Una vez estudiado el comportamiento de $\pi_{N+1}(x)$, se estudia una cota del error de interpolación $R_N(x)$. Mediante la expresión (3.46) y la acotación (3.55), se obtiene la desigualdad siguiente:

$$\|R_N(x)\| \leq \frac{\Delta x^{N+1}}{4} \ \frac{|f^{(N+1)}(\xi)|}{N+1}.$$

Si la función f a interpolar tiene derivadas de cualquier orden acotadas, como por ejemplo es la función $f(x) = \sin(x)$, y se estudia el límite del módulo del error con $N \to \infty$ se tiene que:

$$\lim_{N\to\infty} |R_N(x)| \leq \lim_{N\to\infty} \left[\frac{\Delta x^{N+1}}{4} \ \frac{1}{(N+1)} \right] = 0.$$

Es decir, el error de truncamiento tiende a cero cuando $N \to \infty$. Si por el contrario, se interpola una función que tiene derivadas no acotadas como por ejemplo es la función de Runge, que tiene la expresión siguiente:

$$f(x) = \frac{1}{1+25x^2}, \qquad x \in [-1, 1],$$

se tiene

$$\lim_{N\to\infty} |R_N(x)| = \infty,$$

porque las derivadas $f^{(N+1)}$ para la función de Runge no están acotadas cuando $N \to \infty$.

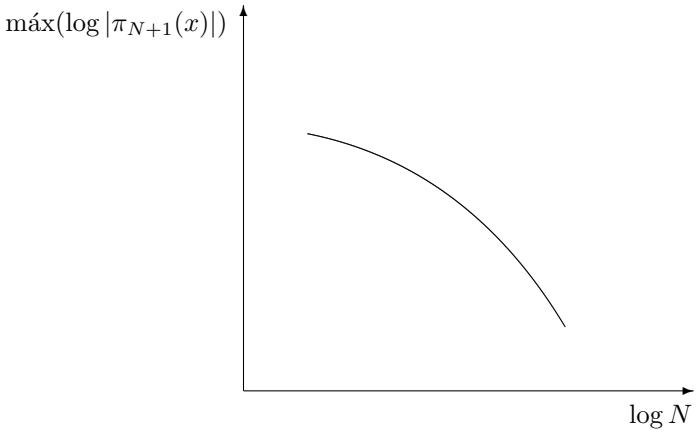


Figura 3.7: Convergencia de $|\pi_{N+1}(x)|$ en función de N para una interpolación global equiespaciada

3.7.2. Error de redondeo

Como hemos comentado con anterioridad, el error de redondeo $R_L(x)$ está asociado a la precisión finita del ordenador y se puede medir a través de la constante de Lebesgue

$$\Lambda_N = \max_{x\in[a,b]} \sum_{j=0}^{N} |\ell_j(x)|.$$

Cuando seleccionamos la nube de puntos $x_0, \dots x_N$ nos interesa: *(i)* que la constante de Lebesgue sea lo menor posible para que la aproximación esté próxima a la óptima (3.44) y *(ii)* que el problema esté bien condicionado. Es decir, que una perturbación pequeña ϵ de los valores de la función a interpolar no introduzcan grandes variaciones en el interpolante.

Procedemos a buscar una cota inferior para la constante de Lebesgue. Como la función de Lebesgue $\lambda_N(x)$ es una suma de $N+1$ términos en valor absoluto, podemos acotarla inferiormente mediante el término central del sumatorio. Es decir,

$$\lambda_N(x) = |\ell_0(x)| + \dots |\ell_{N/2}(x)| + \dots + |\ell_N(x)| \geq |\ \ell_{N/2}(x)|. \tag{3.57}$$

Por otra parte, si la constante de Lebesgue Λ_N es el máximo de la función de Lebesgue $\lambda_N(x)$ en $x \in [a,b]$, podemos calcular el máximo en la expresión anterior y acotar particularizando por el valor medio $x = a + \Delta x/2$

$$\max_{x\in[a,b]} \lambda_N(x) \geq |\max_{x\in[a,b]} |\ell_{N/2}(x)| \geq |\ell_{N/2}\left(a+\frac{\Delta x}{2}\right)|. \tag{3.58}$$

Consideramos la expresión del polinomio de Lagrange

$$\ell_j(x) = \frac{(x-x_0)\dots(x-x_{j-1})(x-x_{j+1})\dots(x-x_N)}{(x_j-x_0)\dots(x_j-x_{j-1})(x_j-x_{j+1})\dots(x_j-x_N)}, \tag{3.59}$$

con una distribución equiespaciada de puntos

$$x_j = a + \Delta x\ j, \qquad j = 0,\dots N,$$

con $\Delta x = (b-a)/N$. Si definimos $y = (x-x_0)/\Delta x$, la expresión (3.59) para una distribución equiespaciada queda:

$$\ell_j(y) = \frac{y(y-1)\dots(y-j+1)(y-j-1)\dots(y-N)}{j(j-1)\dots 1\ \ 1\dots(j-N)}, \tag{3.60}$$

con $y \in [0,N]$. Si en esta expresión hacemos $y = 1/2$ y $j = N/2$, se obtiene:

$$\ell_j(N/2) = \frac{\frac{1}{2}(-\frac{1}{2})(-\frac{3}{2})(-\frac{5}{2})\dots(-\frac{2N-1}{2})}{(\frac{N}{2})!\ (\frac{N}{2})!}. \tag{3.61}$$

Es importante hacer notar que en el numerador de la expresión anterior no está el factor $(N-1)/2$. De esta forma, el valor absoluto de la expresión (3.62) queda como:

$$|\ell_j(N/2)| = \frac{2\ (2N-1)(2N-3)\dots 1}{2^N\ (N-1)\ \ (\frac{N}{2})!\ \ (\frac{N}{2})!}. \tag{3.62}$$

El numerador está formado por el producto de los $N-1$ factores impares que se puede poner como

$$(2N-1)(2N-3)(2N-4)\dots 1 = \frac{(2N-1)(2N-2)(2N-3)(2N-4)\dots 1}{(2N-2)(2N-4)\dots 2}. \tag{3.63}$$

O lo que es lo mismo

$$(2N-1)(2N-3)(2N-4)\dots 1 = \frac{(2N-1)!}{2^{N-1}(N-1)!} = \frac{(2N)!}{2^N N!}. \tag{3.64}$$

Utilizando la expresión (3.64) y la fórmula de Stirling $N!\ \sim \sqrt{2\pi N}(N/e)^N$, la cota (3.62) queda:

$$|\ell_j(N/2)| = \frac{2^N\ \sqrt{2}}{\pi N(N-1)}. \tag{3.65}$$

De esta manera la constante de Lebesgue es

$$\Lambda_N \geq \frac{2^N\ \sqrt{2}}{\pi N(N-1)}. \tag{3.66}$$

Incluso para valores moderados de N, la constante de Lebesgue puede ser muy grande haciendo que la interpolación en una malla equiespaciada presente errores muy grandes. En conclusión, este valor de la constante de Lebesgue para mallas equiespaciadas invalida el uso de interpolantes de alto orden en estas distribuciones de puntos.

3.8. Distribuciones de puntos no equiespaciados

Como se ha visto en el apartado anterior, una distribución equiespaciada de puntos de interpolación da lugar a una constante de Lebesgue grande que hace que el problema esté mal condicionado que es donde se encuentra el máximo de la función de Lebesgue $\lambda_N(x)$. Aunque los máximos de la función que mide el error de truncamiento $\pi_{N+1}(x)$ también se encuentran en las proximidades del dominio de integración, la cota superior de π_{N+1} tiende a cero con N tendiendo a infinito. La pregunta que surge ahora es si es posible encontrar distribuciones de puntos que tengan valores de la función de Lebesgue próximos a la unidad para todo $x \in [a, b]$. La respuesta a esta pregunta es afirmativa y pasa por concentrar puntos en los extremos del intervalo $[a, b]$. Aunque la motivación de Chebyshev fue encontrar distribuciones de puntos que hacen que los extremos π_{N+1} sean uniformes, además con estas distribuciones arrojan valores moderados de la constante de Lebesgue.

Los puntos nodales de Chebyshev se basan en funciones con extremos iguales tales como senos y cosenos trigonométricos. Para estas funciones, una distribución equiespaciada de la variable $\theta \in [0, \pi]$, produce una distribución que concentra puntos de la variable $x \in [a, b]$. Si se realiza un cambio de variable de la forma:

$$x = \cos\theta, \tag{3.67}$$

se observa que los puntos x se concentran en los extremos del intervalo.

Capítulo 4

Interpolantes de Chebyshev

4.1. Distribuciones de puntos de Chebyshev

La motivación de los puntos de Chebyshev pasa por buscar distribuciones de puntos que tengan extremos iguales para la función de error π_{N+1}. Un ejemplo de funciones con extremos iguales son los senos y los cosenos. Definimos la función de error π_{N+1} mediante un coseno que tenga $N+1$ ceros en nuestro intervalo de integración de la forma siguiente:

$$\pi_{N+1}(x) \;=\; K\ \cos\,(N+1)\theta, \tag{4.1}$$

con $x = \cos\theta$. Tenemos que demostrar que la función π_{N+1} así definida es un polinomio de grado $N+1$. Si utilizamos la fórmula de Moivre

$$\cos\,(N+1)\theta + i\ \operatorname{sen}\,(N+1)\theta = (\cos\theta + i\ \operatorname{sen}\theta)^{N+1}\,,$$

y sustituimos $x = \cos\theta$ en la expresión anterior se obtiene:

$$\cos\,(N+1)\theta + i\ \operatorname{sen}\,(N+1)\theta = \left(x + i\ \sqrt{1-x^2}\right)^{N+1}. \tag{4.2}$$

Mediante la expresión del binomio de Newton

$$(x+y)^N = \sum_{k=0}^{N} x^{N-k}\ y^k\ \binom{N}{k}$$

donde el coeficiente binomial vale

$$\binom{N}{k} = \frac{N!}{(N-k)!\ k!}$$

y desarrollando el binomio (4.2), la expresión de $\cos(N+1)\theta$ proviene de los términos con parte real. El término $i\sqrt{1-x^2}$ debe estar elevado a una potencia par para que contribuya con parte real. Así, se demuestra que $\cos(N+1)\theta$ es una expresión polinómica de grado $N+1$.

Los ceros de $\pi_{N+1}(x)$ son:

$$(N+1)\theta_j \;=\; \frac{\pi}{2} \;+\; \pi j, \qquad j = 0, \ldots, N.$$

que se corresponden con los nodos de interpolación

$$x_j \;=\; \cos\left(\frac{\dfrac{\pi}{2} \;+\; \pi j}{N+1}\right), \qquad j = 0, \ldots, N. \tag{4.3}$$

Entre estos nodos x_j, $\pi_{N+1}(x)$ alcanza N extremos dados por:

$$y_j \;=\; \cos\left(\frac{\pi j}{N+1}\right), \qquad j = 1, \ldots, N. \tag{4.4}$$

A modo de ejemplo, la figura (4.1) presenta los ceros y extremos de $|\pi_{N+1}(x)|$ para 9 puntos distribuidos a partir de la expresión (4.3). En esta representación gráfica, θ_j se recorre desde $j = N$ hasta $j = 0$ para obtener el conjunto de puntos correspondiente $x_0, x_1, \ldots, x_8$.

Tal como puede observarse en la figura y como puede comprobarse en las expresiones correspondientes, un inconveniente que presenta la distribución de puntos (4.3) es que no contiene los puntos extremos del intervalo de interpolación. Los puntos nodales (4.3) se los conoce como puntos de Chebyshev de primera clase o ceros de Chebyshev y los puntos (4.5) se los conoce como puntos de Chebyshev de segunda clase o extremos de Chebyshev. Generalmente, los puntos dados por (4.5) junto con $x = -1$ y $x = 1$ son los puntos de Chebyshev de mayor uso en las aplicaciones numéricas. Estos puntos así completados quedan:

$$x_j \;=\; \cos\left(\frac{\pi j}{N}\right), \qquad j = 0, \ldots, N. \tag{4.5}$$

Es importante hacer notar que para esta distribución de puntos nodales, $\pi_{N+1}(x)$ y $\lambda_N(x)$ son no uniformes pero se comportan mucho mejor que para distribuciones equiespaciadas de puntos.

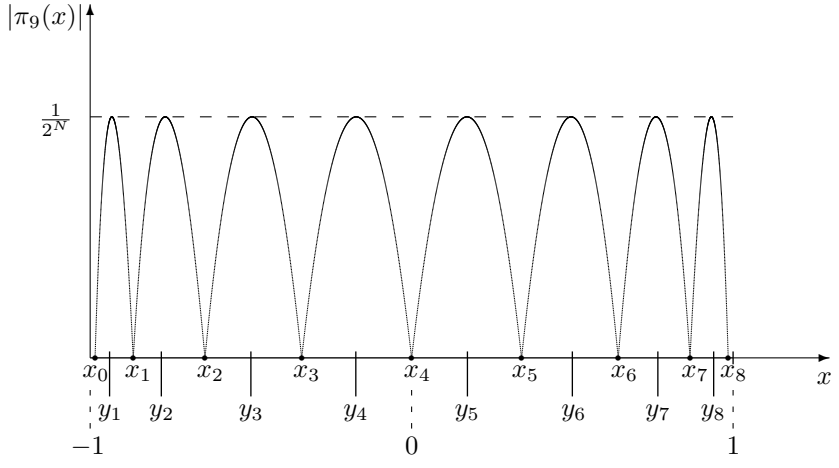


Figura 4.1: Ceros y extremos de $|\pi_{N+1}(x)|$ para un conjunto de 9 puntos ($N = 8$) no equiespaciados obtenidos a partir de los ceros de $\pi_{N+1}(x)$

4.2. Polinomios de Chebyshev

Los polinomios de Chebyshev son una secuencia de polinomios ortogonales que se definen de la siguiente forma:

$$T_k(x) \;=\; \cos(k\theta), \tag{4.6}$$

donde $x = \cos\theta$. Mediante la fórmula de Moivre se demuestra que la expresión anterior es un polinomio de grado k. La seria truncada asociada a esta base se define como:

$$P_N(x) = \sum_{k=0}^{N} \hat{c}_k T_k(x). \tag{4.7}$$

Asociado a los ceros de Chebyshev (4.3) o a los extremos de Chebyshev (4.5) se construye el interpolante o la serie discreta

$$I_N(x) = \sum_{k=0}^{N} \tilde{c}_k T_k(x). \tag{4.8}$$

La pregunta que surge ahora es determinar la bondad del ajuste del interpolante de Lagrange basado en las distribuciones anteriores y su relación con la serie truncada. Como vimos en las secciones 2.5 y 2.6, existen distribuciones de puntos que nos permiten igualar el producto interno $< u, v >_N$ del espacio vectorial de dimensión finita determinado por los puntos nodales (2.13) con el producto interno $< u, v >_w$

del espacio de dimensión infinita (2.3). Es decir,

$$< u, v >_N = < u, v >_w \tag{4.9}$$

para u v polinomios de grado máximo $2N+1$ para las fórmulas de cuadratura de Gauss o polinomios de grado máximo $2N-1$ para las fórmulas de cuadratura de Gauss–Lobatto.

4.2.1. Ceros de Chebyshev

Asociado a los ceros de Chebyshev (4.3) se construye el interpolante o la serie discreta

$$I_N(x) = \sum_{k=0}^{N} \tilde{c}_k T_k(x). \tag{4.10}$$

Imponiendo que el interpolante pase por los ceros de Chebyshev se determinan los coeficientes de esta serie discreta.

Las fórmulas de cuadratura de Gauss se obtienen cuando los puntos nodales son los $N+1$ ceros del polinomio de la base

$$T_{N+1}(x) = \cos(N+1)\theta.$$

Estos puntos coinciden con los ceros de Chebyshev (4.3) y permiten calcular los coeficientes de la serie discreta en función de los puntos nodales. En este caso, los coeficientes de peso son $\alpha_j = \pi/(N+1)$ y el producto interno queda:

$$< u, v >_N = \frac{\pi}{N+1} \sum_{j=0}^{N} u(x_j)v(x_j) \tag{4.11}$$

y los coeficientes de normalización γ_m

$$\gamma_m \ = \frac{\pi}{N+1} \sum_{j=0}^{N} \cos^2(m\theta_j) = \frac{\pi}{2}. \tag{4.12}$$

Así, los coeficientes de la serie discreta se obtienen mediante

$$\tilde{c}_m \ = \ \frac{2}{(N+1)} \sum_{j=0}^{N} f(x_j) \ \cos m \ \theta_j, \qquad m = 1, \ldots, N. \tag{4.13}$$

Cuando el interpolante se usa para integrar un problema en donde es necesario imponer condiciones de contorno, los ceros de Chebyshev no tienen mucha utilidad porque no incluyen los contornos $x = -1$ y $x = +1$. Sin embargo, los extremos de Chebyshev dados por (4.5) si incluyen los contornos y son los que tienen una mayor utilidad.

4.2.2. Extremos de Chebyshev

Asociado a los extremos de Chebyshev (4.3) se construye el interpolante o la serie discreta

$$I_N(x) = \sum_{k=0}^{N} \tilde{c}_k T_k(x). \tag{4.14}$$

Imponiendo que el interpolante pase por los extremos de Chebyshev se determinan los coeficientes de esta serie discreta.

Las fórmulas de cuadratura de Gauss–Lobatto se obtienen cuando los puntos nodales son los $N+1$ ceros de la combinación de polinomios de la base (2.28)

$$q_{N+1}(x) = T_{N+1}(x) + c_1\ T_N(x) + c_2\ T_{N-1}(x),$$

donde c_1 y c_2 se eligen para que $q_{N+1}(-1) = q_{N+1}(+1) = 0$. Resolviendo el sistema anterior se obtiene $c_1 = 0$ y $c_2 = -1$. Es decir,

$$q_{N+1}(x) = T_{N+1}(x) - T_{N-1}(x) = \cos\theta\ \operatorname{sen} N\theta.$$

Las raíces de este polinomio coinciden con los extremos de Chebyshev (4.5) y permiten calcular los coeficientes de la serie discreta en función de los puntos nodales. En este caso, la expresión de los coeficientes de peso (2.26) permite obtener:

$$\{\alpha_0,\ \alpha_1, \ldots, \alpha_{N-1},\ \alpha_N\} = \left\{\frac{\pi}{2N},\ \frac{\pi}{N}, \ldots, \frac{\pi}{N},\ \frac{\pi}{2N}\right\},$$

y el producto interno queda:

$$< u, v >_N = \frac{\pi}{2N}\,(u_0 v_0 + u_N v_N) + \frac{\pi}{N}\ \sum_{j=1}^{N-1}\ u_j v_j. \tag{4.15}$$

Los coeficientes de normalización γ_m son:

$$\gamma_m\ = \frac{\pi}{2N}\left(\cos^2(m\theta_0) + \cos^2(m\theta_N\right) + \frac{\pi}{N}\ \sum_{j=1}^{N-1} \cos^2(m\theta_j). \tag{4.16}$$

Como los extremos de Chebyshev incluyen $\theta_N = \pi$ y $\theta_0 = 0$, la expresión anterior se simplifica,

$$\gamma_m = \begin{cases} \dfrac{\pi}{2} & m < N, \\ \pi & m = N. \end{cases} \tag{4.17}$$

Así, los coeficientes de la serie discreta se obtienen mediante

$$\tilde{c}_m\ =\ \frac{1}{2N}\left(f(x_0) + (-1)^k f(x_N)\right) + \frac{1}{N}\ \sum_{j=1}^{N-1}\ f(x_j)\ \cos\left(\frac{m\ \pi j}{N}\right). \tag{4.18}$$

En el caso especial de los extremos de Chebyshev el error de aliasing (2.18) entre la serie discreta y la serie truncada tiene una expresión muy simple

$$\tilde{c}_m \;=\; \hat{c}_m \;+\; \sum_{|p|\geq 1} \hat{c}_{m+pN}. \tag{4.19}$$

Es importante hacer notar que si la serie truncada $\hat{c}_k$ tiene precisión espectral, la serie discreta también tiene comportamiento espectral y por lo tanto el interpolante de Lagrange con los puntos (4.5) tiene un error tan pequeño como el error de la serie truncada. Por otra parte, el paso entre el plano físico $f(x_j)$ y el plano espectral $\tilde{c}_m$ se puede realizar con una transformada rápida de Fourier o transformada coseno dada por (4.18).

4.3. Transformada rápida de Fourier

La transformada rápida de Fourier es un algoritmo recursivo para evaluar la transformada discreta de Fourier y su inversa. La transformada rápida de Fourier más simple requiere que N sea una potencia de 2. Si todos los datos son complejos, entonces la evaluación de una FFT necesita $5Nlog_2N - 6N$ operaciones reales. En muchas aplicaciones, u es real y el número de operaciones se divide por la mitad. Las transformadas de Fourier rápidas que permiten factores 2,3,4,5 y 6 ofrecen una reducción $10 - 20\,\%$ sobre las FFT de potencias de 2. La transformada de Fourier se usa para la evaluación de,

$$\tilde{f}_k = \frac{1}{N} \sum_{j=0}^{N-1} f_j \; e^{-ik2\pi j/N}, \quad k = -\frac{N}{2}, \ldots \frac{N}{2} - 1 \tag{4.20}$$

$$f_j = \sum_{-N/2}^{N/2-1} \tilde{f}_k \; e^{ik2\pi j/N}, \quad j = 0, \ldots N-1 \tag{4.21}$$

El algoritmo de Cooley–Turkey (1965) permite evaluar las sumas de (4.20) en $5Nlog_2N$ operaciones reales (cuando N es una potencia de 2) en lugar de $8N^2$ operaciones reales requeridas por la suma directa. Además, el cálculo de (4.20) por la transformada rápida de Fourier produce menos error debido a la falta de precisión de la suma directa. En la actualidad existen versiones de la FFT (Fast Fourier Transform) que permiten que N sea de la forma,

$$N = 2^p 3^q 4^r 5^s 6^t$$

y el número de operaciones es

$$N(5p + 3q + 4r + \frac{39}{5}s + \frac{13}{3}t - 6)$$

Si $N = 2^m$, entonces la transformada de Fourier se puede calcular de forma rápida. La transformada directa de Fourier se puede poner como la transformada de Fourier para los términos pares mas la transformada de Fourier para los términos impares.

$$\tilde{f}_k = \frac{1}{N}\sum_{j=0}^{N/2} f_{2j}\ e^{-ikx_{2j}} + \frac{1}{N}\sum_{j=0}^{N/2-1} f_{2j+1}\ e^{-ikx_{2j+1}}. \tag{4.22}$$

Como los puntos de colocación están equiespaciados, entonces

$$x_{2j+1} = x_{2j} + \frac{2\pi}{N}$$

y la expresión (4.22) se puede poner como:

$$\tilde{f}_k^{r+1} = \tilde{f}_{kp}^{r} + w^k \tilde{f}_{ki}, \quad w = e^{-i2\pi/N}.$$

La transformada de Fourier $\tilde{f}_k^{r+1}$ se puede calcular a partir de la transformada de Fourier para los términos pares mas la transformada de Fourier para los términos impares. Lo cual nos permite crear un árbol binario subdividiendo consecutivamente los puntos de colocación en pares e impares.

Ejemplo. Consideraremos la transformada de Fourier de un conjunto de 8 elementos como ejemplo de cálculo de la transformada rápida de Fourier. Las expresiones para calcular los coeficientes de la transformada de Fourier de 8 elementos son:

$$\begin{aligned}
\tilde{f}_0 &= f_0 + w_0 f_1 + w_0(f_2 + w_0 f_3) + w_0[f_4 + w_0 f_5 + w_0(f_6 + w_0 f_7)],\\
\tilde{f}_1 &= f_0 + w_4 f_1 + w_6(f_2 + w_4 f_3) + w_7[f_4 + w_4 f_5 + w_6(f_6 + w_4 f_7)],\\
\tilde{f}_2 &= f_0 + w_0 f_1 + w_4(f_2 + w_0 f_3) + w_6[f_4 + w_0 f_5 + w_4(f_6 + w_0 f_7)],\\
\tilde{f}_3 &= f_0 + w_4 f_1 + w_2(f_2 + w_4 f_3) + w_5[f_4 + w_4 f_5 + w_2(f_6 + w_4 f_7)],\\
\tilde{f}_{-4} &= f_0 + w_0 f_1 + w_0(f_2 + w_0 f_3) + w_4[f_4 + w_0 f_5 + w_0(f_6 + w_0 f_7)],\\
\tilde{f}_{-3} &= f_0 + w_4 f_1 + w_6(f_2 + w_4 f_3) + w_3[f_4 + w_4 f_5 + w_6(f_6 + w_4 f_7)],\\
\tilde{f}_{-2} &= f_0 + w_0 f_1 + w_4(f_2 + w_0 f_3) + w_2[f_4 + w_0 f_5 + w_4(f_6 + w_0 f_7)],\\
\tilde{f}_{-1} &= f_0 + w_4 f_1 + w_2(f_2 + w_4 f_3) + w_1[f_4 + w_4 f_5 + w_2(f_6 + w_4 f_7)]
\end{aligned}$$

con $w_k = e^{-ik2\pi/N}$. Como puede observase, para obtener los coeficientes de la serie de Fourier mediante la suma directa necesitaríamos realizar 14 operaciones ($2(N-1)$) para obtener cada coeficiente, y como tenemos que determinar 8 coeficientes necesitaríamos realizar un total de 112 operaciones ($2(N-1)N$). Sin embargo, vemos que muchas operaciones en el cálculo anterior están repetidas y ese hecho es el que nos permite reducir el número de operaciones estrictamente necesario. Las operaciones diferentes en la determinación de los coeficientes de Fourier son 8×3 ($Nlog_2(N)$).

En la gráfica (4.2) se observa como se obtiene la transformada rápida de Fourier de un conjunto de 8 elementos. Conocida la transformada de Fourier de las parejas de elementos $(f_0, f_4), (f_2, f_6)$ se puede obtener la transformada de Fourier

de f_0, f_2, f_4, f_6 . De igual forma, conocida la transformada de Fourier de las parejas de elementos $(f_1, f_5), (f_3, f_7)$ se obtiene la transformada de Fourier de los elementos f_1, f_3, f_5, f_7. Por último, se determina la transformada de Fourier de $f_0, f_1, f_2, f_3, f_4, f_5, f_6, f_7$ a partir de la transformada de Fourier de f_0, f_2, f_4, f_6 y de la transformada de f_1, f_3, f_5, f_7 .

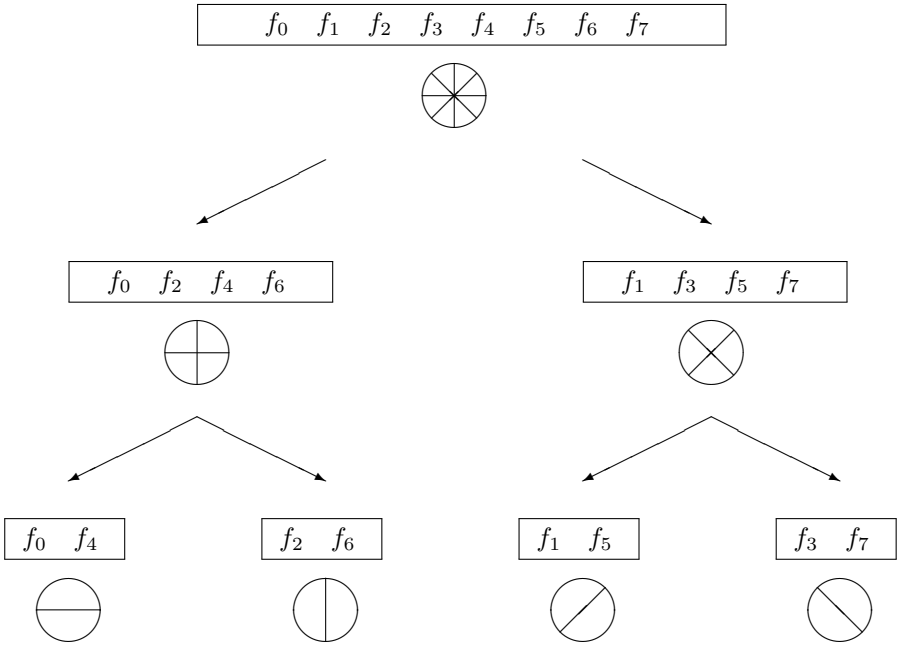


Figura 4.2: Transformada rápida de Fourier de ocho puntos mediante la regla de recursión de puntos pares e impares.

Generalmente, necesitaremos la transformada de Fourier de una función real. En este caso, y considerando que en la expresión (4.20) f_j es un número complejo, podríamos calcular dos transformadas de Fourier reales almacenando un conjunto de valores en la parte real de f_j y otro conjunto de valores en la parte imaginaria de f_j.

En el caso de necesitar transformadas de Fourier multidimensionales, las transformadas se pueden obtener por parejas. Supongamos que f_j^1 y $f_j^2, j = 0, \ldots N-1$ son dos conjuntos de datos reales. Entonces se puede definir,

$$f_j = f_j^1 + i f_j^2 \tag{4.23}$$

y calcular los $\tilde{f}_k$ mediante (4.20). Entonces, los coeficientes de Fourier $\tilde{f}_k^1, \tilde{f}_k^2$ se

pueden obtener con las siguientes expresiones,

$$\tilde{f}_k^1 = \frac{1}{2}(\tilde{f}_k + \overline{\tilde{f}_k}), \tag{4.24}$$

$$\tilde{f}_k^2 = -\frac{i}{2}(\tilde{f}_k - \overline{\tilde{f}_{-k}}). \tag{4.25}$$

4.3.1. Transformada rápida coseno mediante la FFT

La transformada de Fourier (4.20) con M puntos se expresa:

$$\tilde{f}_k = \frac{1}{M}\sum_{j=0}^{M-1} f_j\ e^{-ik\ \theta_j}, \quad k = -\frac{M}{2}, \dots \frac{M}{2} - 1 \tag{4.26}$$

con $\theta_j = 2\pi j/M$. Por otra parte, la transformada coseno (4.18)

$$\tilde{c}_k\ =\ \frac{1}{2N}\left(f_0 + (-1)^k f_N\right) + \frac{1}{N}\sum_{j=1}^{N-1} f_j\ \cos\left(\frac{k\ \pi j}{N}\right). \tag{4.27}$$

Prolongamos periódicamente f_j con $N-1$ puntos extra haciendo

$$\begin{aligned} f_{N+1} &= f_{N-1}, \\ f_{N+2} &= f_{N-2}, \\ &\vdots \\ f_{2N-1} &= f_1. \end{aligned}$$

Con $M = 2N$ usamos estos puntos prolongados periódicamente en la expresión (4.26) y dividimos la suma total en dos términos para los $N+1$ primeros puntos y para los siguiente $N-1$ puntos siguientes

$$M\tilde{f}_k = \sum_{j=0}^{N} f_j\ e^{-ik\ \theta_j} + \sum_{j=N+1}^{2N-1} f_j\ e^{-ik\ \theta_j}. \tag{4.28}$$

Del primer sumatorio de la expresión anterior separamos los términos correspondientes a $j = 0$ y $j = N$

$$M\tilde{f}_k = f_0 + f_N(-1)^N + \sum_{j=1}^{N-1} f_j\ e^{-ik\ \theta_j} + \sum_{j=N+1}^{2N-1} f_j\ e^{-ik\ \theta_j}. \tag{4.29}$$

Finalmente, los dos sumatorios de esta expresión los agrupamos por parejas f_j y f_{2N-j}

$$M\tilde{f}_k = f_0 + f_N(-1)^N + \sum_{j=1}^{N-1} f_j\ e^{-ik\ \theta_j} + f_{2N-j}\ e^{-ik\ \theta_{2N-j}}. \tag{4.30}$$

Como $\theta_{2N-j} = 2\pi - \theta_j$ y $f_{2N-j} = f_j$ la expresión anterior queda

$$M\tilde{f}_k = f_0 + f_N(-1)^N + 2\sum_{j=1}^{N-1} f_j \cos k\,\theta_j. \tag{4.31}$$

Es decir, la transformada coseno (4.27) se calcula mediante la transformada rápida de Fourier (4.26) haciendo

$$\tilde{c}_k = \frac{\tilde{f}_k}{2}. \tag{4.32}$$

4.3.2. Derivada del interpolante de Chebyshev

Una de la principales ventajas del interpolante de Chebyshev es la existencia de transformada rápida. Esto permite transformar entre el plano físico y el plano espectral en $O(NlogN)$ operaciones. Cualquier problema de simulación involucra el interpolante de Chebyshev y sus derivadas. Mientras que la derivación del interpolante en los $N+1$ nodales involucra $O(N^2)$ operaciones, la derivación en el plano espectral y su posterior transformación al plano físico involucra $O(NlogN)$ operaciones. Es por esta razón por la que nos interesa conocer una expresión de la derivada del polinomio de Chebyshev en el plano espectral.

La derivada del interpolante de Chebyshev dado por (4.8) vale

$$\frac{dI_N}{dx} = \sum_{k=0}^{N} \tilde{c}_k \frac{dT_k}{dx}, \tag{4.33}$$

donde

$$\frac{dT_k}{dx} = k\,\frac{\sin k\theta}{\sin\theta}. \tag{4.34}$$

Mediante la siguiente igualdad trigonométrica

$$\frac{\sin(k+1)\theta}{\sin\theta} - \frac{\sin(k+1)\theta}{\sin\theta} = 2\cos\theta, \tag{4.35}$$

expresamos las derivadas $T_k'(x)$ en función de los polinomios de Chebyshev,

$$\frac{T_{k+1}'}{k+1} - \frac{T_{k-1}'}{k-1} = 2T_k, \qquad k = 0, \ldots N-1. \tag{4.36}$$

El sistema anterior en forma matricial se puede expresar

$$\begin{pmatrix} 2 & 0 & 0 & 0 & \dots & \dots & 0 \\ 0 & 0 & 1/2 & 0 & \dots & \dots & 0 \\ 0 & -1 & 0 & 1/3 & \dots & \dots & 0 \\ 0 & 0 & -1/2 & 0 & 1/4 & \dots & 0 \\ \vdots & & \ddots & \ddots & \ddots & \dots & 0 \\ 0 & \dots & \dots & 0 & -1/(N-2) & 0 & 1/N \end{pmatrix} \begin{pmatrix} T_1' \\ T_2' \\ T_3' \\ T_4' \\ \vdots \\ T_N' \end{pmatrix} = 2 \begin{pmatrix} T_0 \\ T_1 \\ T_2 \\ T_3 \\ \vdots \\ T_{N-1} \end{pmatrix}. \quad (4.37)$$

Si definimos mediante $\tilde{c}_k^{(1)}$ los coeficientes del desarrollo en serie de Chebyshev de la derivada del interpolante

$$\frac{dI_N}{dx} = \sum_{k=0}^{N} \tilde{c}_k^{(1)} T_k', \quad (4.38)$$

entonces los coeficientes $\tilde{c}_k^{(1)}$ se relacionan con los coeficientes $\tilde{c}_k$ mediante,

$$2(\tilde{c}_0, \dots, \tilde{c}_{N-1}) = (\tilde{c}_1^{(1)}, \dots, \tilde{c}_N^{(1)})\ D, \quad (4.39)$$

donde D es la matriz del sistema (4.37). La componente genérica k del sistema (4.39) queda:

$$\tilde{c}_{k-1}^{(1)} = \tilde{c}_{k+1}^{(1)} + 2k\tilde{c}_k, \qquad k = N-1, \dots 1. \quad (4.40)$$

Esta es una ecuación en diferencias de segundo orden para los valores $\tilde{c}_k^{(1)}$ que se integra con $\tilde{c}_N^{(1)} = 0$ y $\tilde{c}_{N-1}^{(1)} = 2N\ \tilde{c}_N$. El número de operaciones necesario para obtener los coeficientes $\tilde{c}_k^{(1)}$ es $O(N)$. Posteriormente, para obtener la derivada del interpolante en los puntos nodales se hace una transformada coseno que involucra $O(N \log N)$ operaciones.

Capítulo 5

Interpolación continua a trozos

Los métodos para aproximar funciones mediante interpolantes son la *interpolación global* y la *interpolación continua a trozos*. La interpolación global consiste en utilizar todos los puntos de interpolación dentro del dominio de definición, para construir un interpolante que será continuo en dicho dominio. La interpolación continua a trozos consiste en utilizar una cantidad, por lo general pequeña, de puntos de interpolación para construir un interpolante continuo a trozos. En este caso, se habla de $q+1$ puntos nodales de interpolación y el interpolante resultante I_q es un polinomio de grado q. El método de interpolación continua a trozos es el más empleado cuando, a partir de la teoría de interpolación, se pretende obtener métodos numéricos para la solución de ecuaciones diferenciales. La estrategia de construcción de un interpolante continuo a trozos se explicará en secciones siguientes de este capítulo.

En el capítulo anterior se analizó el comportamiento del error cuando se emplean interpolaciones globales equiespaciadas y no equiespaciadas. En este capítulo se estudia la *interpolación continua a trozos* con distribuciones equiespaciadas y no equiespaciadas de puntos. Ésta es la forma más empleada al momento de generar esquemas numéricos para la solución de ecuaciones diferenciales. Para este tipo de interpolación se estudia el comportamiento del error de interpolación a través de la función de error π y de la función de Lebesgue $\lambda_N(x)$.

Para describir la interpolación equiespaciada continua a trozos se consideran $N+1$ puntos $x_0, x_1, \ldots, x_N$ del dominio de definición de una función f. A continuación, se consideran $q/2$ puntos de interpolación a la derecha y a la izquierda de un punto genérico x_j, donde q en un número natural par. De esta forma se tienen $q+1$ puntos de interpolación donde x_j es el punto medio. A esta configuración de

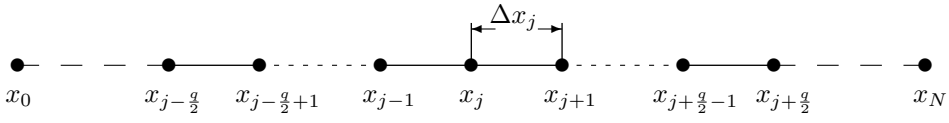


Figura 5.1: Molécula computacional centrada y equiespaciada

puntos se la suele denominar molécula computacional centrada y se representa en la figura (5.1). Para el conjunto de $q+1$ puntos de interpolación se puede construir un interpolante I_q de grado $\leq q$ y si se repite esta construcción a todo el dominio de interpolación, se tiene una interpolación continua a trozos. Es importante hacer notar que para interpolar en las proximidades de los extremos x_0 y x_N la molécula computacional se debe descentrar manteniendo el mismo grado del interpolante.

El error correspondiente a la interpolación polinómica a partir de los $q+1$ puntos equiespaciados tiene la expresión siguiente:

$$R_q(x) = \pi_{q+1}(x)\frac{f^{(q+1)}(\xi)}{(q+1)!}, \tag{5.1}$$

donde $\pi_{q+1}(x)$ es:

$$\begin{aligned}\pi_{q+1}(x) = (x - x_{j-\frac{q}{2}})(x - x_{j-\frac{q}{2}+1})\cdots(x - x_{j-1})(x - x_j)(x - x_{j+1})\cdots \\ \cdots(x - x_{j+\frac{q}{2}-1})(x - x_{j+\frac{q}{2}}).\end{aligned} \tag{5.2}$$

Al igual que se hizo en la interpolación global, a continuación se busca una cota del error de interpolación continua a trozos a través de la función (5.2). Si la distribución de puntos nodales es equiespaciada, la expresión (3.56) para la cota del error de truncamiento resulta:

$$\|\pi_{q+1}(x)\| \leq \; q!\; \frac{\Delta x^{q+1}}{4}. \tag{5.3}$$

En el caso de una nube de puntos equiespaciada, $\Delta x = 2/N$ y la expresión (5.3) queda:

$$\|\pi_{q+1}(x)\| \leq \; \frac{q!}{4}\; \left(\frac{2}{N}\right)^{q+1}. \tag{5.4}$$

Tomando logaritmos en la expresión (5.3), se tiene que:

$$\log\|\pi_{q+1}(x)\| \leq \log\left(\frac{q!}{4}\right) + (q+1)\big[\log 2 - \log N\big].$$

De forma esquemática, la figura (5.2) representa la expresión anterior e indica la convergencia de $|\pi_{q+1}(x)|$ con el número de puntos N en una interpolación continua a trozos equiespaciada.

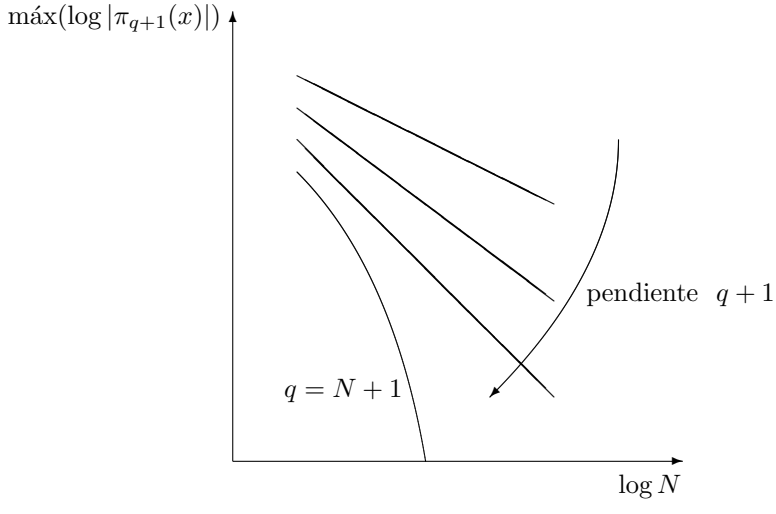


Figura 5.2: Convergencia de $|\pi_{q+1}(x)|$ en función de N para una interpolación continua a trozos equiespaciada con q+1 puntos

Se observa en la figura (5.2) que mientras que en la interpolación continua a trozos las expresiones son rectas con pendiente $q+1$, el error de truncamiento en la interpolación global disminuye con N más rápidamente que una recta de pendiente $N+1$.

5.1. Fórmulas para la derivada primera y segunda

En este apartado, se obtienen las fórmulas de las derivadas primera y segunda para interpolantes a trozos construidos a partir de $q+1$ puntos. En la figura (5.3) se representan los interpolantes de grado q y su dominio de validez representados con diferentes tramos. Es importante hacer notar que el dominio de validez del

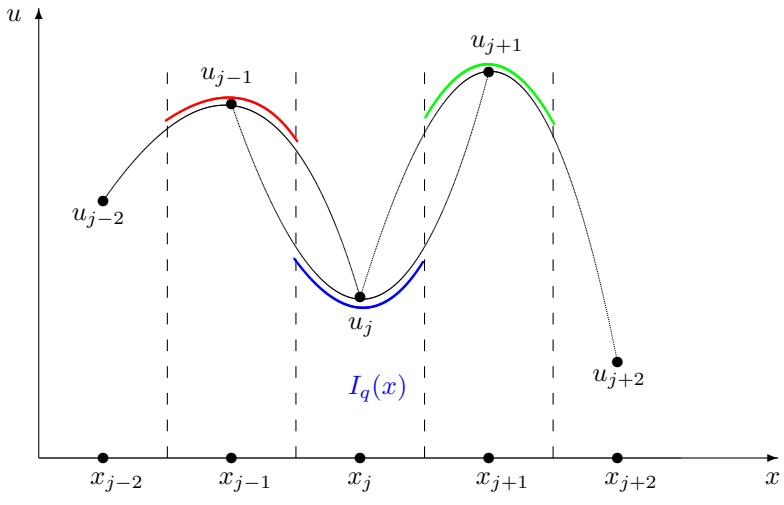


Figura 5.3: Interpolación continua a trozos y dominio de definición del inteprolante $I_q(x)$

interpolante I_q es:

$$\forall x \in \left(x_j - \frac{\Delta x_{j-1}}{2},\ x_j + \frac{\Delta x_{j+1}}{2} \right). \tag{5.5}$$

De esta forma, las expresiones de la derivada primera y segunda de I_q particularizadas en el punto nodal x_k perteneciente a su dominio de definición quedan:

$$\left(\frac{dI_q}{dx} \right)_{x_k} = \sum_{j=0}^{q} f_j \ \ell'_j(x_k), \tag{5.6}$$

$$\left(\frac{d^2 I_q}{dx^2} \right)_{x_k} = \sum_{j=0}^{q} f_j \ \ell''_j(x_k). \tag{5.7}$$

Obsérvese que tanto $\ell'_j(x_k)$ como $\ell''_j(x_k)$ son números y las expresiones (5.6) y (5.7) se conocen como las fórmulas de diferencias finitas para las derivadas primera y segunda.

5.2. Error de las derivadas primera y segunda

El error de truncamiento y redondeo del interpolante I_q es:

$$E_q(x) = \pi_{q+1}(x) \ \frac{f^{(q+1)}(\xi)}{(q+1)!} + \sum_{j=0}^{q} \epsilon_j \ \ell_j(x). \tag{5.8}$$

La expresión (5.8) es válida para su domino de definición correspondiente. Si la molécula computacional está centrada en x_k, el dominio de validez de este interpolante es un entorno de x_k dado por la expresión (5.5). Obsérvese que la interpolación no es continua en $x = x_k - \Delta x_{k-1}/2$ y en $x_k + \Delta x_k/2$. En el caso de los contornos, el dominio de validez es mayor.

Derivando la expresión (5.8) con respecto a x, se obtiene el error de la derivada primera:

$$E'_q(x) \;=\; \pi'_{q+1}(x)\;\frac{f^{(q+1)}(\xi)}{(q+1)!} \;+\; \pi_{q+1}(x)\;\frac{f^{(q+2)}(\xi)}{(q+1)!}\;\left(\frac{d\xi}{dx}\right) \;+\; \sum_{j=0}^{q}\;\epsilon_j\;\ell'_j(x). \quad (5.9)$$

y derivando nuevamente la expresión anterior con respecto a x, se obtiene el error de la derivada segunda:

$$\begin{aligned} E''_q(x) \;=\; & \pi''_{q+1}(x)\;\frac{f^{(q+1)}(\xi)}{(q+1)!} \;+\; 2\;\pi'_{q+1}(x)\;\frac{f^{(q+2)}(\xi)}{(q+1)!}\;\left(\frac{d\xi}{dx}\right) \;+ \\ & \pi_{q+1}(x)\;\left(\frac{f^{(q+2)}(\xi)}{(q+1)!}\;\left(\frac{d\xi}{dx}\right)\right)' \;+\; \sum_{j=0}^{q}\;\epsilon_j\;\ell''_j(x). \end{aligned} \quad (5.10)$$

En particular, las expresiones (5.8), (5.9) y (5.10) para los puntos *nodales* x_k, respectivamente quedan:

$$E_q(x_k) = \epsilon_k, \quad (5.11)$$

$$E'_q(x_k) = \pi'_{q+1}(x_k)\frac{f^{(q+1)}(\xi)}{(q+1)!} + \sum_{j=0}^{q}\epsilon_j\ell'_j(x_k), \quad (5.12)$$

$$E''_q(x_k) = \pi''_{q+1}(x_k)\frac{f^{(q+1)}(\xi)}{(q+1)!} + 2\pi'_{q+1}(x_k)\frac{f^{(q+2)}(\xi)}{(q+1)!}\left(\frac{d\xi}{dx}\right) + \sum_{j=0}^{q}\epsilon_j\ell''_j(x_k). \quad (5.13)$$

Si q es par, el número de puntos de interpolación es $q+1$ impar y las moléculas computacionales son centradas. Si q es impar, el número de puntos de interpolación es $q+1$ par y la molécula computacional no es centrada.

Es importante hacer notar que si la malla es equiespaciada y q es par, $\pi''_{q+1}(x)$ se anula en el punto central y el factor $\pi''_{q+1}(x)$ en (5.13) se hace cero, quedando la derivada primera y la derivada segunda con el mismo orden de error. Por esta razón, siempre se elige q par cuando se interpola con polinomios continuos a trozos.

En los siguientes apartados se discutirá el error de la derivada primera y segunda para una interpolación continua a trozos basada en tres puntos ($q = 2$).

5.3. Fórmulas para derivadas con tres puntos

Se consideran tres puntos de interpolación no equiespaciados tal como se representa en la figura (5.4).

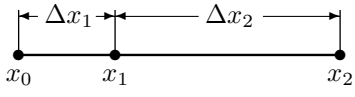


Figura 5.4: Tres puntos de interpolación no equiespaciados

Con estos tres puntos se define un interpolante de segundo orden $I_2(x)$

$$I_2(x) = f_0\ell_0(x) + f_1\ell_1(x) + f_2\ell_2(x) \tag{5.14}$$

con los polinomios de Lagrange asociados:

$$\ell_0(x) = \frac{(x-x_1)\ (x-x_2)}{(x_0-x_1)\ (x_0-x_2)}, \tag{5.15}$$

$$\ell_1(x) = \frac{(x-x_0)\ (x-x_2)}{(x_1-x_0)\ (x_1-x_2)}, \tag{5.16}$$

$$\ell_2(x) = \frac{(x-x_0)\ (x-x_1)}{(x_2-x_0)\ (x_2-x_1)}. \tag{5.17}$$

Para calcular las fórmulas de las derivadas primera y segunda se deriva el interpolante y se particulariza en el punto deseado. Estas expresiones para la derivada primera y segunda particularizadas en $x = x_1$ quedan:

$$I_2'(x_1) = -\frac{\Delta x_2 f_0}{(\Delta x_2+\Delta x_1)\ \Delta x_1} + \frac{(\Delta x_2-\Delta x_1)f_1}{\Delta x_1\ \Delta x_2} + \frac{\Delta x_1 f_2}{(\Delta x_2+\Delta x_1)\ \Delta x_2}. \tag{5.18}$$

$$I_2''(x_1) = \frac{2\ f_0}{(\Delta x_2+\Delta x_1)\ \Delta x_1} - \frac{2\ f_1}{\Delta x_1\ \Delta x_2} + \frac{2\ f_2}{(\Delta x_2+\Delta x_1)\ \Delta x_2}. \tag{5.19}$$

Cuando la derivada primera se particulariza en el extremo $x = x_0$, se obtiene:

$$I_2'(x_0) = -\frac{(\Delta x_2+2\Delta x_1)f_0}{(\Delta x_2+\Delta x_1)\ \Delta x_1} + \frac{(\Delta x_2+\Delta x_1)f_1}{\Delta x_1\ \Delta x_2} - \frac{\Delta x_1 f_2}{(\Delta x_2+\Delta x_1)\ \Delta x_2}. \tag{5.20}$$

Las fórmulas anteriores permiten aproximar las derivadas primera y segunda en un punto central o en un punto del contorno. Si la malla es equiespaciada, las

expresiones anteriores se reducen a las conocidas fórmulas de diferencias finitas para las derivadas primera y segunda

$$I_2'(x_1) = \frac{1}{2\ \Delta x}(f_2\ -\ f_0), \tag{5.21}$$

$$I_2''(x_1) = \frac{1}{\Delta x^2}(f_2\ -2f_1\ +\ f_0), \tag{5.22}$$

$$I_2'(x_0) = \frac{1}{2\Delta x}\ (-3f_0\ +\ 4f_1\ -\ f_2)\,, \tag{5.23}$$

$$I_2'(x_2) = \frac{1}{2\Delta x}\ (+3f_2\ -\ 4f_1\ +\ f_0)\,. \tag{5.24}$$

5.4. Error de interpolación con tres puntos

Las expresiones para el error de truncamiento y el error de redondeo de la derivada primera (5.12) y de la derivada segunda (5.13) particularizadas para $q = 2$ quedan:

$$R_2'(x_k) = \pi_3'(x_k)\frac{f^{(3)}(\xi)}{3!} + \sum_{j=0}^{2}\epsilon_j \ell_j'(x_k), \tag{5.25}$$

$$R_2''(x_k) = \pi_3''(x_k)\frac{f^{(3)}(\xi)}{3!} + 2\pi_3'(x_k)\frac{f^{(4)}(\xi)}{3!}\left(\frac{d\xi}{dx}\right) + \sum_{j=0}^{2}\epsilon_j \ell_j''(x_k). \tag{5.26}$$

5.4.1. Error de truncamiento

La influencia de la distribución de puntos en el error de truncamiento está caracterizada por $\pi_3'(x_k)$ y por $\pi_3''(x_k)$. La función π_3 para este caso es:

$$\pi_3(x) = (x - x_0)(x - x_1)(x - x_2), \tag{5.27}$$

y sus derivadas primera y segunda son:

$$\begin{aligned} \pi_3'(x) &= (x - x_0)(x - x_1) + (x - x_0)(x - x_2) + (x - x_1)(x - x_2), \\ \pi_3''(x) &= 6x - 2(x_0 + x_1 + x_2). \end{aligned} \tag{5.28}$$

Particularizamos la derivada primera y segunda de $\pi_3(x)$ en el punto nodal

medio x_1 de la figura (5.4)

$$\pi_3'(x_1) = (x_1 - x_0)(x_1 - x_2),$$

$$\pi_3''(x_1) = 6x_1 - 2x_0 - 2x_1 - 2x_2 = 2(x_1 - x_0 - x_2) = 4\left[x_1 - \frac{1}{2}(x_0 + x_2)\right],$$

y sustituyendo las distancias entre los puntos nodales por su espaciado correspondiente $\Delta x_1 = x_1 - x_0$ y $\Delta x_2 = x_2 - x_1$, las derivadas primera y segunda de π en x_1 resultan:

$$\pi_3'(x_1) = -\Delta x_1 \Delta x_2,$$

$$\pi_3''(x_1) = 2(\Delta x_1 - \Delta x_2).$$

Con estas expresiones, los errores de truncamiento para la derivada primera y segunda en el punto x_1 quedan:

$$R_2'(x_1) \;=\; -\Delta x_1 \;\Delta x_2 \;\frac{f^{(3)}(\xi)}{3!}, \tag{5.29}$$

$$R_2''(x_1) \;=\; -2\Delta x_1 \;\Delta x_2 \;\frac{f^{(4)}(\xi)}{3!}\left(\frac{d\xi}{dx}\right) + 2(\Delta x_1 - \Delta x_2)\frac{f^{(3)}(\xi)}{3!}. \tag{5.30}$$

A la vista de las anteriores expresiones (5.29) y (5.30) podemos extraer las siguientes conclusiones:

1. Si consideramos $N+1$ puntos nodales equiespaciados en el compacto $[-1, 1]$, entonces, $\Delta x = 2/N$ y de la expresión (5.29) vemos que el error truncamiento es inversamente proporcional al número de puntos al cuadrado. O lo que es lo mismo, si integramos numéricamente con $2N$ puntos, entonces el error de truncamiento se divide por 2^2. En general, se dice que un determinado esquema numérico es de orden q si al duplicar el número de puntos manteniendo la distribución de los puntos nodales constante, el error se divide por 2^q.

2. Por otra parte, no podemos asegurar nada acerca del comportamiento del error cuando se cambia el número de puntos y además se cambia su distribución. En este caso, si el cambio de la distribución es favorable a la solución, el error disminuirá aunque no sabemos con que ley. Sin embargo, si la nueva distribución es desfavorable a la solución, el error podrá aumentar aún incluso habiendo aumentado el número de puntos.

3. Mientras que el primer término del desarrollo (5.30) es el error local de truncamiento propio del esquema centrado de tres puntos y aparece incluso en

mallas con espaciamiento uniforme, el segundo término es propio de las mallas no uniformes. Si la malla es uniforme ($\Delta x_1 = \Delta x_2$), este segundo término es cero. En distribuciones de puntos en las que la variación del paso es del mismo orden que el paso, este término llega a ser $O(\Delta x)$. Para estas distribuciones nodales, el error de truncamiento de la derivada segunda es $O(\Delta x)$ en lugar de $O(\Delta x^2)$ como lo es el error de truncamiento de la derivada primera.

Por lo tanto, si queremos que el error de truncamiento de la derivada primera y de la derivada segunda sean del mismo orden, deberemos exigir condiciones de regularidad a la distribución de puntos nodales. Si

$$\Delta x_i = \Delta x_{i-1} + O(\Delta x^2),$$

entonces todos los términos de (5.30) son del mismo orden y el error de truncamiento es $O(\Delta x^2)$.

4. En zonas con gradientes fuertes, la distancia entre los puntos nodales se puede reducir de forma que el error de truncamiento se mantenga constante en todo el dominio. De esta manera, la distribución de puntos nodales se puede adaptar a la forma de la solución para minimizar el error de truncamiento.

5.4.2. Error de redondeo

Las expresiones para el error de redondeo de la derivada primera y segunda para las fórmulas con tres puntos particularizadas para el punto x_k son:

$$R'_L(x_k) \;=\; \sum_{j=0}^{2} \epsilon_j \; \ell'_j(x_k), \tag{5.31}$$

$$R''_L(x_k) \;=\; \sum_{j=0}^{2} \epsilon_j \; \ell''_j(x_k). \tag{5.32}$$

Estas expresiones coinciden con las fórmulas (5.18) y (5.19) de la derivada primera y de la derivada segunda cuando se sustituye (f_0, f_1, f_2) por $(\epsilon_0, \epsilon_1, \epsilon_2)$.

En el caso de una malla equiespaciada, las expresiones (5.31) y (5.32) particularizadas para $x_k = x_1$ quedan:

$$R'_L(x_1) \;=\; \frac{\epsilon_2 \;-\; \epsilon_1}{2\;\Delta x}, \qquad R''_L(x_1) \;=\; \frac{\epsilon_2 \;-\; 2\epsilon_1 \;+\; \epsilon_0}{\Delta x^2}. \tag{5.33}$$

5.4.3. Error de interpolación: truncamiento y redondeo

Una vez hemos obtenido el error de truncamiento y el error de redondeo pasamos a analizar el tamaño o la importancia de estos dos términos en distribuciones equiespaciadas de puntos en función del número de puntos nodales N. Las expresiones para el error de truncamiento (5.29) (5.30) junto con las del error de redondeo (5.33) permiten escribir el error total de interpolación para la derivada primera y segunda

$$E_2'(x_1) = -\Delta x^2 \ \frac{f^{(3)}(\xi)}{3!} \ + \ \frac{\epsilon_2 \ - \ \epsilon_1}{2 \ \Delta x}, \tag{5.34}$$

$$E_2''(x_1) = -2\Delta x^2 \ \frac{f^{(4)}(\xi)}{3!} \left(\frac{d\xi}{dx}\right) \ + \ \frac{\epsilon_2 \ - \ 2\epsilon_1 \ + \ \epsilon_0}{\Delta x^2}. \tag{5.35}$$

Acotando las expresiones anteriores se tiene:

$$|E_2'(x_1)| \ \leq \ \frac{\Delta x^2}{6} \ K_1 + \ \frac{2 \ \epsilon}{2 \ \Delta x}, \tag{5.36}$$

$$|E_2''(x_1)| \ \leq \ \frac{\Delta x^3}{3} \ K_2 \ + \ \frac{4 \ \epsilon}{\Delta x^2}, \tag{5.37}$$

donde $|f^{(3)}(\xi)| \leq K_1$, $\left| f^{(4)}(\xi) \left(\frac{d\xi}{dx}\right) \right| \leq K_2$ y $\epsilon \ = \ \text{máx}\,(\epsilon_0, \epsilon_1, \epsilon_2)$. Las expresiones (5.36) y (5.37) se representan respectivamente en las figuras (5.5) y (5.6) en función de Δx.

Para los dos casos se observa la presencia de un mínimo de error para un Δx óptimo. La razón reside en la importancia del error de redondeo frente al error de truncamiento.

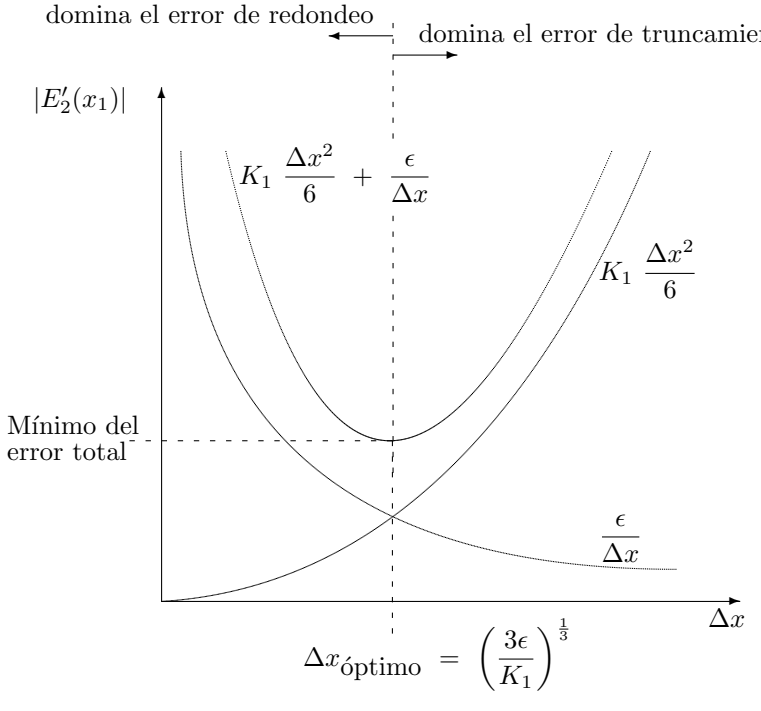


Figura 5.5: Error total de la fórmula para la derivada primera centrada con tres puntos en una malla equiespaciada

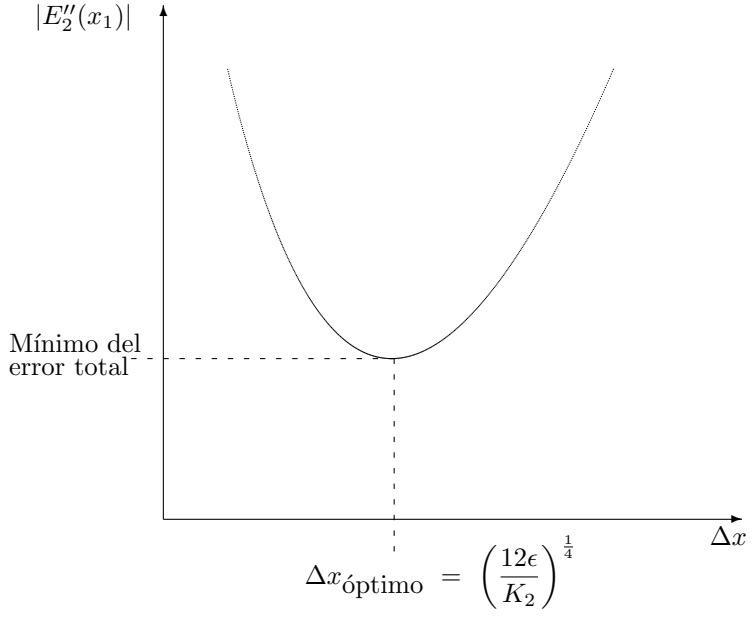


Figura 5.6: Error total de la fórmula para la derivada segunda centrada con tres puntos en una malla equiespaciada

Capítulo 6

Problema de contorno

En este capítulo se analizará por separado el problema de contorno en ecuaciones diferenciales ordinarias y el problema de contorno en ecuaciones en derivadas parciales.

Un problema de contorno se define por un operador diferencial $\mathcal{L}(u)$ y sus condiciones de contorno $BC(u) = 0$ siguientes:

$$\begin{aligned} \mathcal{L}(u) &= 0, \qquad \forall x \in \Omega, \qquad \Omega \subset \mathbb{R}^d \\ BC(u)|_{\partial\Omega} &= 0, \end{aligned} \tag{6.1}$$

donde $\mathcal{L}(u)$ constituye una relación entre los valores de la función y sus derivadas. El orden de un problema diferencial es el mayor orden de las derivadas que involucra el operador $\mathcal{L}(u)$. Esta relación es válida en un dominio Ω. El problema de contorno se caracteriza por dar las condiciones que debe cumplir la solución general de (6.1) en el contorno del dominio Ω.

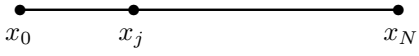
6.1. Ecuaciones diferenciales ordinarias

En particular, si el problema fuera unidimensional y la expresión $\mathcal{L}(u)$ fuera un operador escalar de segundo orden, se deben dar dos condiciones para determinar la solución, una en $x = x_0$ y la otra en $x = x_f$. Si el problema fuera de orden superior a dos, el problema se puede reducir a un problema vectorial de segundo orden mediante la inclusión de una variable auxiliar. Generalmente, las leyes de

conservación de la física en formulación primitiva constituyen sistemas de ecuaciones en derivadas parciales que involucran a lo sumo derivadas segundas de las magnitudes físicas.

A continuación, se expone el algoritmo para resolver un problema de contorno escalar en un dominio unidimensional:

1. Partición del dominio de integración mediante una malla equiespaciada o no equiespaciada.



2. Elegir el interpolante: global o continuo a trozos.
   Si es global, estaremos frente a métodos espectrales y si es continua a trozos frente a diferencias finitas.

3. Calcular las expresiones para las derivadas primera y segunda. En general, las condiciones de contorno se representan por el operador $BC(u) = 0$ que involucra a la función y a su derivada primera. Estas condiciones o relaciones diferenciales se deben cumplir en el contorno de Ω. En el caso unidimensional el contorno de Ω es $x = x_0$ y $x = x_f$. En el caso de diferencias finitas se utilizan las fórmulas centradas o descentradas obtenidas en los capítulos anteriores. En el caso de interpolación global, se pueden obtener las fórmulas de las derivadas en el plano espectral o en el plano físico. Si el problema se resuelve en el plano físico, las incógnitas son los valores de la función en los puntos nodales mientras que si el problema se resuelve en el plano espectral, las incógnitas del problema son las amplitudes de los armónicos de la serie discreta.

4. Imponemos el cumplimiento del operador diferencial $\mathcal{L}(u) = 0$ en cada uno de los puntos interiores $x_j, \quad j = 1, \ldots, N-1$.

5. Imponemos las condiciones de contorno $BC(u) = 0$ en $x = x_0$ y en $x = x_N$ mediante las fórmulas para las derivadas obtenidas en el paso 3.

6. Si el problema es escalar, el sistema anterior constituye un sistema de $N + 1$ ecuaciones con $N + 1$ incógnitas. Si la resolución se hace en el plano físico, las incógnitas son los valores de la función en los puntos nodales $u_j, \quad j = 0, \ldots, N$. Por el contrario, si la resolución se hace en el plano espectral, las incógnitas son las amplitudes de los armónicos del desarrollo en

serie:

$$I_N(x) \;=\; \sum_{k=0}^{N} \tilde{c}_k \;\phi_k(x), \qquad \{\tilde{c}_k, \qquad k = 0, \ldots, N\}.$$

Dependiendo de la linealidad del operador diferencial y de la linealidad de las condiciones de contorno, el sistema resultante de las ecuaciones algebraicas puede ser lineal o no lineal.

7. El último punto pasa por resolver el sistema algebraico del paso 6.

La figura (6.1) resume el algoritmo explicado anteriormente.

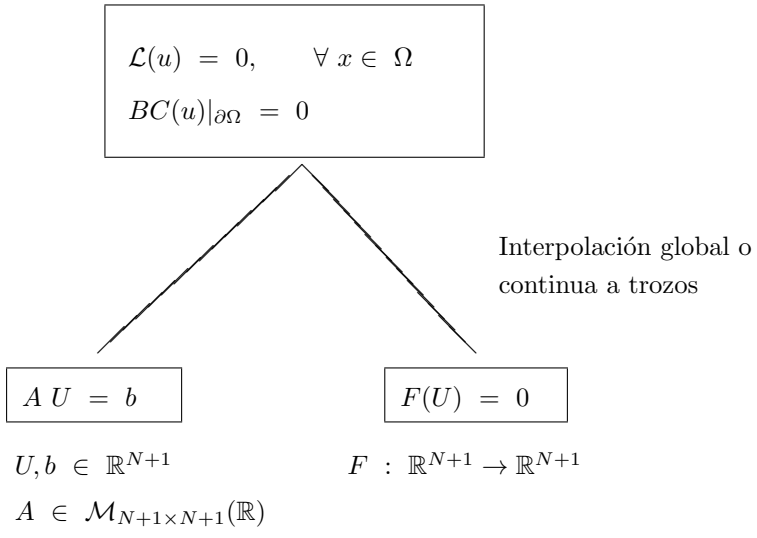


Figura 6.1: Esquema del algoritmo para la solución de un problema de contorno

A continuación, a modo de ejemplo, se presenta el siguiente problema de contorno:

$$\begin{aligned} u'' \;+\; \pi^2 u \;&=\; 0, \\ u(-1) \;&=\; 1, \\ u(+1) \;&=\; 1, \end{aligned} \tag{6.2}$$

donde la solución analítica es:

$$u(x) \;=\; \cos(\pi x). \tag{6.3}$$

Se plantea una partición espacial tal como se indica en la figura (6.2). Imponiendo que se verifique el operador diferencial en los puntos interiores y las condiciones de

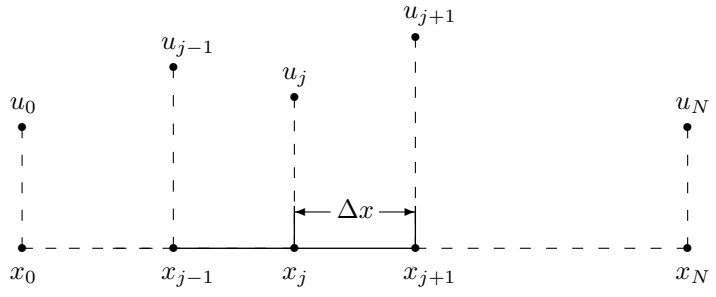


Figura 6.2: Partición de la malla espacial

contorno se obtienen el siguiente sistema de $N+1$ ecuaciones con $N+1$ incógnitas u_j o valores nodales.

$$\begin{aligned} &\frac{1}{\Delta x^2}\left(u_{j+1} \;-\; 2u_j \;+\; u_{j-1}\right) \;+\; \pi^2\; u_j \;=\; 0, \qquad j = 1, \ldots, N-1, \\ &u_0 \;=\; 1, \qquad u_N \;=\; 1. \end{aligned} \tag{6.4}$$

El sistema anterior expresado en forma vectorial queda:

$$A\,U \;=\; b, \tag{6.5}$$

donde la matriz del sistema es:

$$A \;=\; \begin{pmatrix} 1 & 0 & & \ldots & 0 \\ 1 & -2+\Delta x^2\pi^2 & & 1 & 0 \\ 0 & \ddots & \ddots & \ddots & 0 \\ 0 & 1 & -2+\Delta x^2\pi^2 & & 1 \\ 0 & \ldots & & 0 & 1 \end{pmatrix}, \tag{6.6}$$

con el vector de incógnitas U y el término independiente b

$$U \;=\; \begin{pmatrix} u_0 \\ u_1 \\ \vdots \\ u_{N-1} \\ u_N \end{pmatrix}, \qquad b \;=\; \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix}. \tag{6.7}$$

Generalmente, las discretizaciones de problemas lineales de contorno llevan asociados matrices que tienen formas mucho más complicadas que la del ejemplo anterior.

Para facilitar la construcción de la matriz del sistema, se define una función vectorial $L(U)$ cuyo argumento es un vector de incógnitas o puntos nodales U y la imagen es el conjunto de ecuaciones (6.4). Si la función depende linealmente de U, entonces $F(U) = A\ U + b$ donde A es la matriz del sistema y b es el término independiente.

Mediante sucesivas evaluaciones de la función vectorial $L(U)$ obtenemos tanto el término independiente como la matriz del sistema. En concreto, $F(0)$ coincide con el término independiente b. Además, si hacemos todas las componentes de U cero menos la componente j que valga 1, la imagen de la función vectorial L para este valor de U es la columna j de la matriz más el término independiente b. De esta forma, barriendo todos los posibles valores de U desde $j = 0, \ldots N$, obtenemos todas las columnas de la matriz A .

6.2. Dominios bidimensionales

Cuando el dominio de integración Ω es bidimensional o tridimensional, el operador diferencial $\mathcal{L}(u) = 0$ puede involucrar derivadas parciales en cada una de las direcciones del espacio. Desde el punto de vista conceptual, el problema de contorno es el mismo que en un dominio unidimensional salvo que en este caso es necesario calcular derivadas a lo largo de direcciones coordenadas diferentes. En el caso de que las direcciones coordenadas sean ortogonales, el problema se puede reducir a un interpolante unidimensional de forma muy simple. Se considera un interpolante bidimensional a trozos

$$I(x,y) \ = \ \sum_{k=0}^{q} \sum_{m=0}^{q} f_{km}\ \ell_k(x)\ \ell_m(y)\ , \tag{6.8}$$

donde,

$$\ell_k(x) \ = \ \prod_{\substack{n=i-q/2 \\ n\neq k}}^{i+q/2} \frac{(x-x_n)}{(x_k-x_n)}, \qquad \ell_m(y) \ = \ \prod_{\substack{n=j-q/2 \\ n\neq j}}^{j+q/2} \frac{(y-y_n)}{(y_m-y_n)}. \tag{6.9}$$

La expresión (6.8) es un interpolante que satisface los puntos nodales (x_k, y_m) con los valores de la función $f(x_k, y_m)$ que se representan por f_{km} para los siguientes puntos de la molécula computacional:

$$\begin{aligned} &\{x_{i-\frac{q}{2}}, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_{i+\frac{q}{2}}\}, \\ &\{y_{j-\frac{q}{2}}, \ldots, y_{j-1}, y_j, y_{j+1}, \ldots, y_{j+\frac{q}{2}}\}. \end{aligned} \tag{6.10}$$

De esta forma se construye una partición equiespaciada o no equiespaciada del dominio Ω con $N_x + 1$ puntos en la dirección x y $N_y + 1$ puntos en la dirección y.

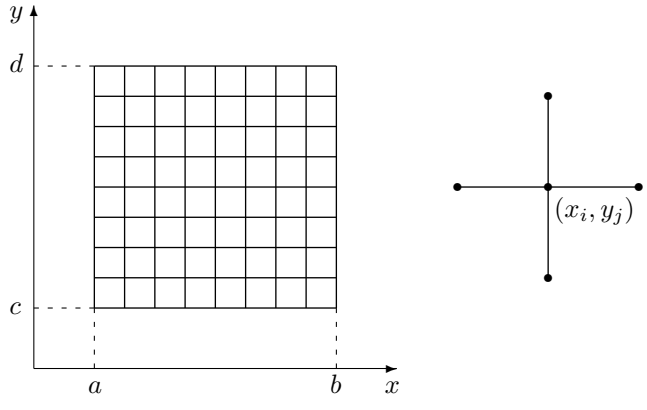


Figura 6.3: Partición equiespaciada en un dominio bidimensional

De igual forma que en el caso 1D, se fuerza al cumplimiento del operador diferencial en los puntos interiores o puntos de colocación:

$$\mathcal{L}(u)|_{x_i,y_j} = 0, \tag{6.11}$$

que constituyen un sistema de $(N_x - 1) \times (N_y - 1)$ ecuaciones que junto con las condiciones de contorno:

$$BC(u)|_{\partial\Omega} = 0, \tag{6.12}$$

que son $2(N_x - 1) + 2(N_y + 1) = 2N_x + 2N_y$ ecuaciones permiten obtener las $(N_x + 1)(N_y + 1)$ incógnitas de la función en los puntos nodales.

Capítulo 7

Problema de Cauchy en EDOS

Un problema de Cauchy en EDOS está formado por un conjunto de ecuaciones diferenciales de evolución junto con una condición inicial $U(t_0)$.

$$\begin{aligned} &\frac{dU}{dt} \;=\; F(U;\; t), \qquad U \in \mathbb{R}^N, \qquad F : \mathbb{R}^N \times \mathbb{R} \to \mathbb{R}^N, \\ &U(t_0) \;=\; U^0, \quad \forall\, t \;\in\; [t_0,\; +\infty). \end{aligned} \tag{7.1}$$

En los problemas de Cauchy, la variable independiente suele ser el tiempo. La nomenclatura adoptada en esta sección se muestra en la Figura 7.1. Con el superíndice n se indica la aproximación en el instante temporal t_n.

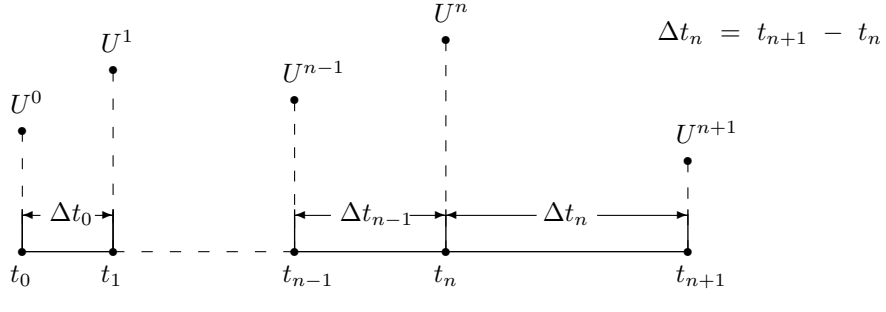


Figura 7.1: Partición de la malla temporal

En los problemas de Cauchy o problemas de evolución se trata de obtener una aproximación para U^{n+1} en función de los valores anteriores U^j y F^j. Los métodos numéricos para la discretización temporal se pueden dividir en dos grandes grupos: *(i)* métodos unipaso y *(ii)* métodos multipaso. Mientras que los métodos multipaso utilizan los valores U^j y F^j en pasos anteriores, los métodos unipaso utilizan la evaluación de $F(U;\ t)$ en etapas intermedias entre t_n y t_{n+1}.

7.1. Métodos multipaso. Métodos Adams

Para obtener los esquemas multipaso o esquemas Adams, se integra la ecuación diferencial de (7.1) entre t_n y t_{n+1}

$$U(t_{n+1}) \ = \ U(t_n) \ + \ \int_{t_n}^{t_{n+1}} \ F(U;\ t)\ dt. \tag{7.2}$$

para aproximar la integral de (7.2) se construye un interpolante para F de grado p que involucra los $p+1$ siguientes:

$$F^{n+1}, F^n, F^{n-1}, \ldots, F^{n+1-p}.$$

De esta forma, el interpolante de grado p se escribe:

$$I_p(t) \ = \ \sum_{j=0}^{p} \ F^{n+1-j}\ \ell_j(t), \qquad\qquad \ell_j(t) \ = \ \prod_{\substack{k=n+1-p \\ k\neq j}}^{k=n+1} \frac{(t-t_k)}{(t_j-t_k)}. \tag{7.3}$$

Cuando se lleva esta expresión del interpolante a la ecuación (7.2) y se integra, los esquemas resultantes o esquemas de Adams quedan:

$$U^{n+1} \ = \ U^n \ + \ \Delta t_n \ \sum_{j=0}^{p} \ \beta_j \ F^{n+1-j}, \qquad n = p-1, \ldots \tag{7.4}$$

donde los valores β_j son números que se obtienen al integrar los polinomios de Lagrange

$$\beta_j \ = \ \frac{1}{\Delta t_n} \ \int_{t_n}^{t_{n+1}} \ \ell_j(t)\ dt. \tag{7.5}$$

Si el interpolante (7.3) involucra F^{n+1} entonces, el esquema (7.4) no permite obtener de forma explícita U^{n+1} en función de los puntos anteriores y los esquemas resultantes son implícitos y se denominan esquemas Adams–Moulton. Por el contrario, cuando el interpolante (7.3) no involucra F^{n+1}, entonces los esquemas (7.4) son explícitos y se denominan esquemas Adams–Bashforth. La expresión (7.4) permite obtener la solución U^{n+1} en el instante t_{n+1} conocida la solución en p pasos

anteriores. Es decir, la ecuación en diferencias (7.4) requiere p condiciones iniciales para poder ser utilizado. Como solo disponemos de una condición inicial, las $p-1$ condiciones iniciales o de arranque de los métodos multipaso se deben calcular mediante otros esquemas unipaso o esquemas que involucren menos pasos. Una vez determinadas las p condiciones iniciales, la expresión (7.4) permite obtener de manera directa U^{n+1} si el esquema es explícito o $\beta_0 = 0$

$$U^{n+1} = U^n + \Delta t_n \left(\beta_1 F^n + \cdots + \beta_p F^{n+1-p}\right), \qquad n = p-1. \tag{7.6}$$

Sin embargo, cuando el esquema es implícito o β_0 es distinto de cero, la ecuación resultante (7.4) queda:

$$U^{n+1} = C + \Delta t_n \, \beta_0 \, F(U^{n+1}; t_{n+1}), \tag{7.7}$$

donde C se puede determinar a partir de los pasos anteriores y el sistema resultante (7.7) constituye un sistema vectorial no lineal de N ecuaciones con N incógnitas o componentes de U^{n+1}. Mientras que los esquemas explícitos permiten obtener U^{n+1} en función de los anteriores, los esquemas implícitos requieren resolver un sistema no lineal de ecuaciones cada vez que damos un paso temporal. La única razón para utilizar esquemas implícitos frente a los esquemas explícitos es que se comportan, en general, mejor en términos de estabilidad.

7.2. Métodos unipaso. Métodos Runge Kutta

Los métodos Runge-Kutta se basan en evaluar etapas intermedias entre t_n y t_{n+1}. Aunque son más robustos que los métodos multipaso exigen, en general, más carga computacional. La expresión general de un esquema Runge-Kutta de e etapas o número de evaluaciones intermedias entre t_n y t_{n+1} es:

$$U^{n+1} = U^n + \Delta t_n \sum_{i=1}^{e} b_i \, k_i, \tag{7.8}$$

donde

$$k_i = F\left(U^n + \Delta t \sum_{j=1}^{e} a_{ij} \, k_j; \; t_n + c_i \, \Delta t_n\right), \qquad i = 1, \ldots, e. \tag{7.9}$$

Los coeficientes b_i, c_i y a_{ij} están dados y son constantes propias del esquema. En general, los esquemas son implícitos y requieren resolver un sistema no lineal (7.9) de e N ecuaciones por cada paso de integración. Sin embargo, si la matriz a_{ij} es triangular inferior con ceros en la diagonal, los esquemas son explícitos.

Ejemplo:

Se considera el siguiente esquema Runge–Kutta de dos etapas:

$$U^{n+1} \;=\; U^n \;+\; \Delta t_n \;(b_1\; k_1 \;+\; b_2\; k_2)\,, \tag{7.10}$$

donde

$$k_1 = \; F\,(U^n \;+\; \Delta t \;(a_{11}k_1 \;+\; a_{12}k_2)\,;\; t_n \;+\; c_1\; \Delta t_n)\,, \tag{7.11}$$

$$k_2 \;=\; F\,(U^n \;+\; \Delta t \;(a_{21}k_1 \;+\; a_{22}k_2)\,;\; t_n \;+\; c_2\; \Delta t_n)\,. \tag{7.12}$$

Como F es una función vectorial de dimensión N, las ecuaciones vectoriales (7.11)–(7.12) son de dimensión N e involucran las incógnitas k_1 y k_2. Conocida la condición inicial U^0, la solución U^1 se determina resolviendo el sistema no lineal de dimensión $2N$ y llevando los valores obtenidos k_1 k_2 a la expresión (7.10).

Si $a_{11} = 0$, $a_{12} = 0$ y $a_{22} = 0$, el esquema es explícito. La pendiente k_1 se puede obtener a partir del paso anterior mediante:

$$k_1 \;=\; F\,(U^n;\; t_n \;+\; c_1\; \Delta t_n)\,, \tag{7.13}$$

y conocido k_1 determinar la evaluación intermedia k_2 mediante:

$$k_2 \;=\; F\,(U^n \;+\; \Delta t\; a_{21}k_1;\; t_n \;+\; c_2\; \Delta t_n)\,. \tag{7.14}$$

Finalmente, llevando los valores de k_1 y k_2 a la expresión (7.10) se puede obtener el valor en el paso siguiente.

7.3. Error global y error local de truncamiento

En esta sección se propone estudiar la relación que existe entre el error de la solución numérica y el error de truncamiento o aproximación asociado a un esquema numérico dado. Para poder obtener esta relación, se define el error global de la solución numérica y el error local de truncamiento del esquema.

Definición: El error global de la solución numérica en un instante t_n genérico se define como la diferencia entre la solución exacta evaluada en ese instante menos la solución numérica

$$E^n = U(t_n) - U^n. \tag{7.15}$$

Definición: Se dice que un esquema es de orden q cuando su error local de truncación es $O\left(\Delta t^{q+1}\right)$.

Debido al proceso de acumulación del error local, se demuestra que el error global es orden q o lo que es lo mismo su error es $O\left(\Delta t^q\right)$. De esta forma, para que

el error global sea al menos de primer orden o que el error tienda a cero cuando Δt tienda a cero, el error local debe ser al menos $O(\Delta t^2)$.

Definición: El error local de truncamiento de un esquema numérico en un instante t_n genérico se define como el residuo que deja la solución exacta cuando se introduce en el esquema numérico.

Para los métodos lineales multipaso, esta expresión queda:

$$T^{n+1} = U(t_{n+1}) - U(t_n) - \Delta t_n \sum_{j=0}^{p} \beta_j \; F(U(t_{n+1-j});\; t_{n+1-j}). \tag{7.16}$$

Para los métodos Runge–Kutta, esta expresión queda:

$$T^{n+1} = U(t_{n+1}) - U(t_n) - \Delta t_n \sum_{i=1}^{e} b_i \; k_i(U), \tag{7.17}$$

donde los valores k_i se determinan introduciendo la solución exacta en la ecuaciones (7.9). Una vez definido con precisión el error local y el error global , se procede a obtener la relación entre ambos.

Para los métodos Runge–Kutta, la relación se obtiene restando las ecuación (7.17) menos la ecuación (7.8)

$$T^{n+1} = E^{n+1} - E^n - \Delta t_n \sum_{i=1}^{e} b_i \; (k_i(U) - k_i). \tag{7.18}$$

Reordenando esta ecuación y considerando que k_i es una función suficientemente diferenciable de U,

$$E^{n+1} = E^n + \Delta t_n \sum_{i=1}^{e} b_i \; \frac{\partial k_i}{\partial U} E^n + \; T^{n+1}. \tag{7.19}$$

Esta ecuación en diferencias de primer orden o un paso expresa la relación entre el error local y el error global. El error local aparece como un término fuente en la acumulación del error global.

Para los métodos lineales multipaso, la relación se obtiene de forma similar restando la ecuación (7.16) menos la ecuación (7.4). Reordenando esta ecuación y considerando que F es una función suficientemente diferenciable de U, la ecuación del error global queda:

$$E^{n+1} = E^n + \Delta t_n \sum_{j=0}^{p} \beta_j \; \frac{\partial F}{\partial U} \; E^{n+1-j} + \; T^{n+1}. \tag{7.20}$$

El error global para los métodos lineales multipaso está gobernado por una ecuación en diferencias de p pasos en la que el error local aparece como un término fuente en la acumulación del error global.

En la siguiente sección, se analizará el comportamiento del error global a lo largo de la integración para esquemas Runge–Kutta y esquemas Adams. La teoría general se formula para sistemas de ecuaciones en diferencias de un paso para el error global. Por esta razón, la ecuación en diferencias (7.20) de p pasos se lleva a una ecuación vectorial de un paso mediante el cambio de variable independiente siguiente: $\mathbf{V}^n = (U^n, U^{n-1}, \ldots, U^{n+1-p})^T$.

7.4. Acotación del error y estabilidad numérica

Se considera que el error global está gobernado por una ecuación en diferencias de un paso del tipo (7.19) o de p pasos del tipo (7.20). Mediante la reducción a sistema de primer orden o un paso, la ecuación que gobierna el error global de un esquema numérico se escribe:

$$E^{n+1} = B\ E^n + \ T^{n+1}, \qquad n = 0, 1, \ldots \tag{7.21}$$

donde B es una matriz cuyos coeficientes dependen de la solución exacta $U(t)$ particularizada en los diferentes pasos de integración. Conocido el error en la condición inicial E^0, se puede determinar el error en instantes sucesivos.

En esta sección se analiza el comportamiento del error global cuando el sistema (7.1) es lineal de coeficientes constantes y se integra un paso de tiempo constante Δt. En este caso, el sistema se expresa:

$$\begin{aligned} &\frac{dU}{dt} \ = \ A\ U + b, \qquad U \in \mathbb{R}^N, \qquad F : \mathbb{R}^N \times \mathbb{R} \to \mathbb{R}^N, \\ &U(t_0) \ = \ U^0, \quad \forall\ t \ \in \ [t_0,\ +\infty), \end{aligned} \tag{7.22}$$

donde A es la matriz del sistema de coeficientes constantes y b es el término independiente de coeficientes constantes. En este caso, la matriz B del sistema (7.21) es de coeficientes constantes que son funciones de la matriz A y del paso de tiempo Δt y el sistema (7.21) se resuelve para dar:

$$E^n \ = \ B^n E^0 \ + \ \sum_{k=1}^{n} B^{n-k}\ T^k, \qquad n = 1, 2, \ldots \tag{7.23}$$

La norma del error global $\|E^n\|$ mediante la desigualdad triangular se expresa:

$$\|E^n\| \ \leq \ \|B^n\|\ \|E^0\| \ + \ \sup_{k\in[1,n]} \|T^k\| \ \sum_{k=1}^{n} \ \|B^{n-k}\|. \tag{7.24}$$

Si B es normal, $B\ B^T\ =\ B^T\ B$, existe un conjunto ortonormal de autovalores que diagonaliza B y la norma:

$$\|B^n\|\ =\ \rho^n, \qquad \text{con} \qquad \rho\ =\ \sup_{z\in\Lambda(B)}\ |z|. \tag{7.25}$$

En esta expresión ρ representa el radio espectral de B o el valor máximo del módulo de todos los autovalores de B o espectro $\Lambda(B)$ de B. De esta forma, la expresión (7.24) queda:

$$\|E^n\|\ \leq\ \rho^n\ \|E^0\|\ +\ \sup_{k\in[1,n]}\|T^k\|\ \sum_{k=1}^{n}\ \rho^{n-k}. \tag{7.26}$$

El sumatorio de la expresión anterior involucra una serie geométrica de razón ρ que se suma para dar:

$$\|E^n\|\ \leq\ \rho^n\ \|E^0\|\ +\ \sup_{k\in[1,n]}\|T^k\|\ \left|\frac{1\ -\ \rho^{n+1}}{1\ -\ \rho}\right|. \tag{7.27}$$

Si $\rho < 1$ y $n \to \infty$, la propagación del error de las condiciones iniciales tiende a cero y el error global está acotado por el error local de truncamiento

$$\|E^n\|\ \leq\ \frac{1}{1\ -\rho}\ \sup_{k\in[1,n]}\|T^k\|. \tag{7.28}$$

Si $\rho > 1$ y $n \to \infty$, tanto el error de las condiciones iniciales como el error local de truncamiento se amplifican invalidando la solución numérica.

La pregunta que surge ahora es determinar en qué situaciones ρ es menor que uno. Para poder analizar los autovalores de B en función de los autovalores de A y poder calcular el radio espectral, se utiliza el teorema de transformación espectral. El teorema de transformación espectral involucra el polinomio característico de estabilidad y la región de estabilidad absoluta del esquema numérico que se definen como:

Definición: El polinomio característico de estabilidad de un método lineal multipaso se define mediante:

$$\pi(r,\omega)\ =\ \sum_{j=0}^{p}\ (\alpha_j\ -\ \omega\ \beta_j)\ r^{p-j}, \tag{7.29}$$

donde ω es un parámetro complejo del que depende el polinomio.

Definición: El polinomio característico de estabilidad de un método lineal unipaso se define mediante:

$$\pi(r,\omega)\ =\ r\ -\ 1\ -\ \omega. \tag{7.30}$$

Definición: La región de estabilidad absoluta $\mathcal{R}_A$ de un esquema numérico es el conjunto de números $\omega \in \mathbb{C}$ para los cuales todas las raíces del polinomio característico de estabilidad satisfacen que su módulo es menor o igual a la unidad y aquellas que tienen módulo igual a la unidad son simples.

Si denominamos por $\lambda\Delta t$ los autovalores de $A\ \Delta t$ y utilizamos estas definiciones, el teorema de transformación espectral nos asegura que si todos los autovalores $\lambda\Delta t$ están dentro de la región de estabilidad absoluta $\mathcal{R}_A$, entonces todos los autovalores r de la matriz B tienen módulo menor que la unidad, el radio espectral es menor que la unidad y, en consecuencia, el error global está acotado.

Es importante hacer notar que el signo de la parte real de λ determina la estabilidad de las soluciones (7.22) y que el módulo de los autovalores de r de B determinan la estabilidad de las soluciones numéricas de un sistema lineal de coeficientes constantes.

La región de estabilidad absoluta incluye siempre parte del semiplano complejo con parte real negativa. Dependiendo de la forma de la región de estabilidad absoluta autovalores λ de la matriz A con parte real negativa se pueden meter dentro de la región de estabilidad eligiendo un paso temporal Δt pequeño. Otras veces o para otros esquemas esto se hace imposible y se invalida un determinado esquema numérico para un determinado problema.

Capítulo 8

Problema de condiciones iniciales y de contorno

8.1. Introducción

Un problema de condiciones iniciales y de contorno está caracterizado por un conjunto de ecuaciones de evolución que se deben verificar en un dominio espacial junto con condiciones de contorno para ese dominio. Generalmente, estas ecuaciones de evolución provienen de principios de conservación de energía, cantidad de movimiento y masa. Como ejemplos representativos de este tipo de problemas se encuentra la ecuación del calor que tiene carácter parabólico y la ecuación de ondas que tiene carácter hiperbólico. La ecuación del calor representa la conservación de energía térmica en un dominio genérico. Es decir, la diferencia de energía térmica que entra menos la energía térmica que sale de un determinado dominio se invierte en calentar o incrementar la temperatura de ese dominio. Las condiciones de contorno en este caso hacen referencia al flujo de energía térmica que se impone en el contorno del dominio. La condición de contorno más sencilla es imponer la temperatura del contorno. Sin embargo, en los problemas reales una condición de contorno muy usual es considerar que el flujo de energía que sale de un contorno es proporcional a la diferencia de energía entre el contorno y el espacio exterior. En este caso, como el flujo de calor se puede calcular mediante la ley de Fourier e involucra la gradientes normales al contorno, la condición de contorno es una relación entre las derivadas normales en el contorno y los valores de la temperatura exterior y del contorno.

Sea el siguiente problema de condiciones iniciales y de contorno:

$$\begin{aligned} \frac{\partial u}{\partial t} &= \mathcal{L}(u;t), \qquad u:\Omega\times[t_0,\ +\infty)\to\mathbb{R}, \qquad \Omega\subset\mathbb{R}^3, \\ BC(u)|_{\partial\Omega} &= 0, \\ u(x,t_0) &= f(x), \end{aligned} \tag{8.1}$$

donde $\mathcal{L}(u;t)$ representa el operador diferencial espacial que involucra derivadas a lo largo de direcciones de Ω y el primer término representa la variación frente al tiempo de la magnitud o magnitudes consideradas por el hecho de que existe un balance neto distinto de cero entre la entrada y la salida. Estas ecuaciones de evolución junto con las condiciones de contorno o relaciones $BC(u)$ impuestas en el contorno para todo instante y la condición inicial $u(x,t_0)=f(x)$ permiten determinar o predecir el comportamiento de u a lo largo del tiempo.

La característica fundamental que diferencia los problemas de evolución de los problemas de contorno es que mientras que en los problemas de contorno la información viaja a velocidad infinita, en los problemas de evolución la información viaja a la velocidad que determine el mecanismo físico en cuestión y la solución de puede integrar desde la condición inicial mediante sucesivos pasos temporales suficientemente pequeños. Esta característica permite abordar los problemas de evolución en ecuaciones en derivadas parciales mediante el método de las líneas. El método de la líneas consiste en dos pasos sucesivos: *(i)* discretización espacial del problema y *(ii)* discretización temporal del problema. Mediante el primer paso aproximamos una función como puede ser la temperatura de un dominio mediante un conjunto de puntos nodales finito. Es decir, pasamos de un espacio de dimensión infinita a un espacio de dimensión finita. Una vez realizada esta discretización el conjunto de los valores nodales de la función o de la temperatura constituye un vector de estado que está gobernado por un conjunto de ecuaciones de evolución junto con una condición inicial. Este segundo problema se trata con cualquiera de los esquemas numéricos analizados para la integración de problemas de Cauchy en ecuaciones diferenciales ordinarias. En la siguiente sección se describe el método de las líneas como algoritmo de cálculo para la resolución de estos problemas.

8.2. Discretización espacial y temporal

El operador $\mathcal{L}(u;t)$ es un operador diferencial que actúa sobre la variable $u(x,t)$ e incluye todas las variaciones espaciales de $u(x,t)$. Para abordar la integración de este tipo de problemas se utilizará el denominado método de las líneas que consiste en discretizar espacialmente el problema (8.1) para obtener un conjunto de ecuaciones diferenciales ordinarias. Posteriormente, mediante la discretización

temporal o mediante el uso de los esquemas estudiados para la solución numérica del problema de Cauchy se integra en el tiempo un problema de ecuaciones diferenciales ordinarias.

La discretización espacial se basa en pasar de un espacio de dimensión infinita donde está definida la función solución $u(x,t)$ a un espacio de dimensión finita que aproxima la función solución mediante los valores $u_i(t)$ en un conjunto de puntos nodales x_i o grados de libertad. Mediante estos puntos nodales se construye un interpolante global o continuo a trozos y se aproximan las derivadas espaciales que incluye el operador $\mathcal{L}(u;t)$. De esta manera, la evolución del grado de libertad $u_i(t)$ viene gobernando por una ecuación de la forma:

$$\frac{du_i}{dt} = L_i(u_0, u_1, \ldots, u_N), \qquad i = 0, \ldots, N, \tag{8.2}$$

donde L es una función vectorial que, en general, depende de todos los puntos nodales $u_i(t)$. Si expresamos este sistema en forma vectorial,

$$\frac{dU}{dt} = L(U;t), \tag{8.3}$$

donde U representa el vector de estado del sistema cuyas componentes son los valores nodales $u_i(t)$. Es importante hacer notar las condiciones de contorno al ser discretizadas aparecen en las componentes de $L(U;t)$.

La discretización temporal se realiza de la misma manera. Las funciones $u_i(t)$ se aproximan en un espacio de dimensión finita por un conjunto de pasos temporales u_i^n. Para los métodos Adams se construyen interpolantes y para los métodos Runge-Kutta se evaluan etapas intermedias. De cualquier forma, los esquemas temporales conducen a relaciones algebraicas que permiten determinar u_i^{n+1} en función de valores o pasos anteriores. Si los esquemas son explícitos, estas relaciones se expresan:

$$u_i^{n+1} = G_i(\ \{u_j^k\}\ ), \quad j = 0, \ldots, N, \quad k = n, \ldots, n+1-p, \quad i = 0, \ldots, N. \tag{8.4}$$

Si expresamos este sistema en forma vectorial,

$$U^{n+1} \ = \ G(U^n, U^{n-1}, \ldots, U^{n-1+p}), \tag{8.5}$$

donde el vector de estado $U^n = (u_0^n, u_1^n, \ldots, u_N^n)^T$ representa el valor aproximado de $U(t)$ en el instante t_n y G es una función vectorial que depende de los valores aproximados del vector de estado en instantes anteriores. De esta forma y mediante (8.5) el vector de estado aproximado U^{n+1} se determina en función de los valores aproximados de los vectores de estado en instantes anteriores.

En la figura (8.1) se representa el método de la líneas.

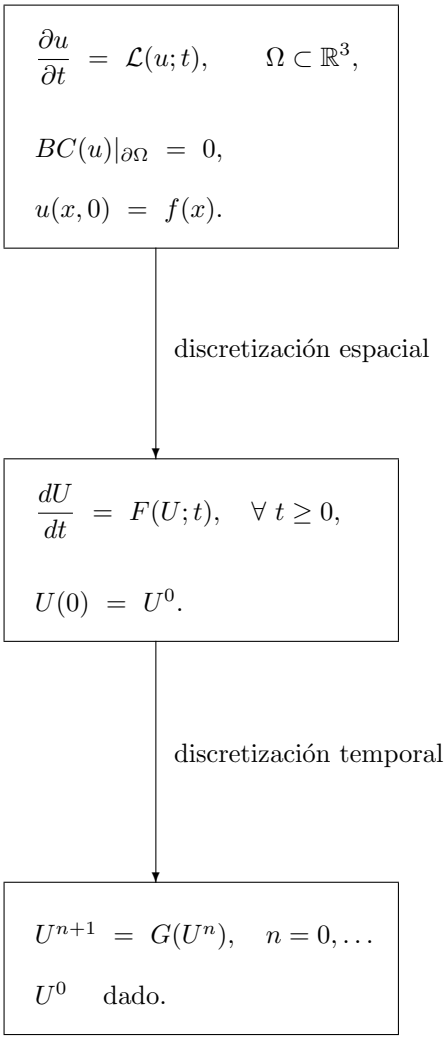


Figura 8.1: Método de las líneas para la discretización espacial y temporal de un problema de evolución

Ejemplo. *Ecuación del calor en un dominio unidimensional*

A continuación, se explica el método de las líneas para la ecuación del calor en un dominio unidimensional $\forall x \in [0,1]$ con condiciones de contorno adiabáticas o de flujo cero y considerando un perfil inicial de temperatura distinto de cero dado por $f(x)$.

$$\begin{aligned} \frac{\partial u}{\partial t} &= \frac{\partial^2 u}{\partial x^2}, \\ \left.\frac{\partial u}{\partial x}\right|_{(0,t)} &= 0, \\ \left.\frac{\partial u}{\partial x}\right|_{(1,t)} &= 0, \\ u(x,0) &= f(x). \end{aligned} \tag{8.6}$$

El algoritmo que se basa en el método de las líneas es el siguiente:

1. Se procede a hacer una partición equiespaciada del intervalo $[0,1]$.
2. Se impone la ecuación en derivadas parciales en el punto genérico x_j.
$$\left(\frac{\partial u}{\partial t}\right)_{x_j} = \left(\frac{\partial^2 u}{\partial x^2}\right)_{x_j}. \tag{8.7}$$
3. Se utiliza un interpolante continuo a trozos para $u(x,t)$ basado en tres puntos y se calculan las fórmulas en diferencias finitas para las derivadas.
4. Se plantea el sistema de ecuaciones diferenciales ordinarias resultante una vez realizada la semidiscretización espacial para los puntos interiores.
$$\frac{du_j}{dt} = \frac{1}{\Delta x^2}\left(u_{j+1} - 2u_j + u_{j-1}\right), \qquad j = 1, \ldots, N-1. \tag{8.8}$$
5. Se discretizan las condiciones de contorno.
$$\left.\frac{\partial u}{\partial x}\right|_{x=0} = \frac{1}{2\Delta x}\left(-3u_0 + 4u_1 - u_2\right) = 0, \tag{8.9}$$
$$\left.\frac{\partial u}{\partial x}\right|_{x=1} = \frac{1}{2\Delta x}\left(3u_N - 4u_{N-1} + u_{N-2}\right) = 0, \tag{8.10}$$
y se obtienen los puntos del contorno en función de los puntos interiores:
$$u_0 = \frac{1}{3}\left(4u_1 - u_2\right), \tag{8.11}$$
$$u_N = \frac{1}{3}\left(4u_{N-1} - u_{N-2}\right). \tag{8.12}$$

6. Se realiza la discretización temporal del sistema (8.8) mediante un esquema temporal. En este caso se utilizará un esquema Euler.

$$\begin{aligned} u_j^{n+1} &= u_j^n + \frac{\Delta t}{\Delta x^2}\left(u_{j+1}^n - 2u_j^n + u_{j-1}^n\right), \quad j = 1, \ldots, N-1, \quad n = 0, \ldots \\ u_0^n &= \frac{1}{3}\left(4u_1^n - u_2^n\right), \\ u_N^n &= \frac{1}{3}\left(4u_{N-1}^n - u_{N-2}^n\right), \\ u_j^0 &= f(x_j), \qquad j = 0, \ldots, N. \end{aligned} \tag{8.13}$$

A partir de la condición inicial $u_j^0 = f(x_j)$ se determinan los puntos interiores del siguiente paso temporal u_j^1,

$$u_j^1 = u_j^0 + \frac{\Delta t}{\Delta x^2}\left(u_{j+1}^0 - 2u_j^0 + u_{j-1}^0\right), \qquad j = 1, \ldots, N-1. \tag{8.14}$$

Las condiciones de contorno u_0^n y u_N^n aparecen cuando se terminan los puntos interiores próximos al contorno $j = 1$ y $j = N - 1$.

De esta forma, el problema se integra a lo largo del tiempo mediante dos bucles anidados. El bucle exterior se encarga de actualizar el instante temporal desde la condición inicial. Para cada instante temporal existe un bucle interno que permite determinar las variaciones de la temperatura con respecto al tiempo para cada punto nodal x_j. En cada instante temporal las condiciones de contorno u_0^n y u_N^n se imponen antes de terminar las variaciones de temperatura de los nodos interiores. Es decir, constituyen ligaduras o relaciones algebraicas que se deben satisfacer para todo instante t_n y se resuelven conocidos los valores de temperatura de los puntos interiores

Ejemplo. *Ecuación de ondas en un dominio unidimensional*

Consideramos las oscilaciones transversales de una cuerda unida entre dos puntos dados $x = 0$ y $x = 1$. Se considera como condiciones iniciales que la cuerda está elongada y parte del reposo. Como condiciones de contorno se supone que la oscilación es cero en los puntos en los que está fijada la cuerda. Así, la ecuación de ondas junto con sus condiciones iniciales y de contorno que permite modelar esta física son:

$$\begin{aligned} &\frac{\partial^2 u}{\partial t^2} = \frac{\partial^2 u}{\partial x^2}, \\ &u(0,t) = 0, \qquad u(1,t) = 0, \\ &u(x,0) = f(x), \qquad \frac{\partial u}{\partial t}(x,0) = 0. \end{aligned} \tag{8.15}$$

En este caso, la integración numérica del problema pasa por reducir inicialmente la ecuación de ondas a una ecuación de primer orden en el tiempo mediante un cambio de variables. Se realiza el cambio de variable $w = \partial u/\partial t$ y el sistema resultante de la ecuación de ondas queda:

$$\frac{\partial}{\partial t}\begin{pmatrix} u \\ \\ w \end{pmatrix} = \begin{pmatrix} w \\ \\ \dfrac{\partial^2 u}{\partial x^2} \end{pmatrix}. \tag{8.16}$$

La ecuación de ondas queda así reducida a un sistema de primer orden en el tiempo que puede ser integrado mediante el método de las líneas anteriormente descrito. Si se discretiza la derivada segunda mediante un esquema centrado de tres puntos, se obtiene el siguiente sistema de ecuaciones diferenciales ordinarias:

$$\begin{pmatrix} \dfrac{du_j}{dt} \\ \\ \dfrac{dw_j}{dt} \end{pmatrix} = \begin{pmatrix} w_j \\ \\ \dfrac{1}{\Delta x^2}\ (u_{j+1} - 2u_j + u_{j-1}) \end{pmatrix}, \tag{8.17}$$

donde se define el vector de estado $U = (u_0,\ w_0,\ \ u_1,\ w_1, \ldots,\ u_N,\ w_N)^T$. Finalmente, el sistema de ecuaciones diferenciales ordinarias así planteado se integra con cualquier esquema temporal.

8.3. Error espacial y error temporal

Cuando un problema se discretiza temporal y espacialmente aparecen dos tipos de problemas de errores: *(i)* el asociado a la aproximación de las derivadas en el espacio y *(ii)* el asociado a la aproximación de las derivadas en el tiempo o la integración temporal.

La solución exacta del problema evaluada en el nodo x_i se representa por $u(x_i, t)$. Cuando se realiza la semidiscretización espacial, $u_i(t)$ representa el valor aproximado en el nodo x_i. Finalmente, cuando se realiza la discretización temporal, u_i^n representa el valor numérico aproximado en el instante t_n y el nodo x_i.

Definición. *Error espacial*

Se define como la diferencia entre el valor exacto y el valor aproximado en cada nodo x_i,

$$E_i \;=\; u(x_i, t) \;-\; u_i(t). \tag{8.18}$$

Definición. *Error temporal*

Se define en cada nodo x_i como la diferencia entre el valor $u_i(t_n)$ y el valor aproximado u_i^n en el instante t_n,

$$E_i^n \;=\; u_i(t_n) \;-\; u_i^n. \tag{8.19}$$

Definición. *Error total*

Se define el error total asociado a las dos discretizaciones como la diferencia entre el valor exacto $u(x_i, t_n)$ evaluado en el nodo x_i y el instante t_n menos el valor aproximado u_i^n,

$$E_i^T \;=\; u(x_i, t_n) \;-\; u_i^n. \tag{8.20}$$

Con estas definiciones, se puede comprobar que el error total es la suma del error espacial más el error temporal

$$E_{T_i} \;=\; \underbrace{u(x_i, t_n) \;-\; u_i(t_n)}_{E_i(t_n)} \;+\; \underbrace{u_i(t_n) \;-\; u_i^n}_{E_i^n}, \tag{8.21}$$

o en notación vectorial:

$$E_T \;=\; E \;+\; E^n. \tag{8.22}$$

A continuación, se pasa a analizar el comportamiento del error espacial E y del error temporal E^n.

8.4. Error de la semidiscretización espacial

Para obtener la ecuación que gobierna el error de la semidiscretización espacial, se resta la ecuación (8.1) en el nodo genérico x_i menos la ecuación (8.2),

$$\frac{d}{dt}\left[u(x_i,t) \; - u_i\right] \; = \mathcal{L}(u;t) \mid_{x_i} \; - \; L_i(u_0,u_1,\ldots,u_N). \tag{8.23}$$

Se suma y resta el operador discreto aplicado a la solución exacta.

$$\begin{aligned} \frac{dE_i}{dt} \; = \; & \mathcal{L}(u;t) \mid_{x_i} \; - \; L_i(u_0,u_1,\ldots,u_N) \\ & +L_i(\; u(x_0;t),\; u(x_1;t),\ldots,\; u(x_N;t)\;) \\ & -L_i(\; u(x_0;t),\; u(x_1;t),\ldots,\; u(x_N;t)\;). \end{aligned} \tag{8.24}$$

Se define el error local de truncamiento de la semidiscretización espacial en cada nodo x_i mediante la diferencia entre el operador exacto y el operador discreto u operador en diferencias aplicado a la solución exacta:

$$R_i \; = \; \mathcal{L}(u;t) \mid_{x_i} \; - \; L_i(\; u(x_0;t),\; u(x_1;t),\ldots,\; u(x_N;t). \tag{8.25}$$

De esta forma, la ecuación para el error E_i queda:

$$\frac{dE_i}{dt} \; = \; R_i + L_i(\; u(x_0;t),\; u(x_1;t),\ldots,\; u(x_N;t)\;) - \; L_i(u_0,u_1,\ldots,u_N). \tag{8.26}$$

Utilizando el teorema del valor medio para la función vectorial L_i

$$\frac{dE_i}{dt} \; = \; R_i + \frac{\partial L_i}{\partial u_j}(\xi_0,\xi_1,\ldots,\xi_N)\;(\; u(x_j;t) - u_j\;), \tag{8.27}$$

donde $\partial L_i/\partial u_j$ representa la matriz jacobiana de la función vectorial L_i evaluada en un punto intermedio $(\xi_0,\ldots,\xi_N)$. Si denominamos por A la matriz jacobiana particularizada, el sistema anterior en forma vectorial se expresa:

$$\frac{dE}{dt} \; = A\;E \; + \; R. \tag{8.28}$$

Si la matriz A fuera de coeficientes constante, esta ecuación junto con la condición inicial $E(0)$ o error en las condiciones iniciales se puede integrar para dar:

$$E(t) \; = \; e^{At}\;E(0) \; + \; \int_0^t \; e^{A(t-\tau)}\;R(\tau)\;d\tau. \tag{8.29}$$

La acotación de la expresión anterior resulta:

$$\|E\| \; \leq \; \|e^{At}\|\;\|E(0)\| \; + \; \sup_{t\in[0,t_f]}\|R\| \; \left|\int_0^t \; \|e^{A(t-\tau)}\|\;d\tau\right|. \tag{8.30}$$

Si la matriz A es normal $A\,A^T = A^T\,A$, entonces existe un conjunto de autovectores ortogonales que diagonaliza la matriz A y

$$\|e^{At}\| = e^{\alpha(A)t}, \tag{8.31}$$

donde $\alpha(A)$ es la abscisa espectral de A que se define como:

$$\alpha(A) = \sup_{z\in\Lambda(A)} \mathrm{Re}(z), \tag{8.32}$$

con $\Lambda(A)$ como el espectro o conjunto de todos los autovalores de A. Es decir, la abscisa espectral es el valor máximo de la parte real de los autovalores de la matriz A.

Mediante la definición de la abscisa espectral (8.32), la integral del segundo término de la cota de error (8.30) queda:

$$\int_0^t \|e^{A(t-\tau)}\|\,d\tau = \int_0^t e^{\alpha(t-\tau)}\,d\tau = \left.\frac{e^{\alpha(t-\tau)}}{-\alpha}\right|_0^t = \frac{1}{\alpha}\left(e^{\alpha t}-1\right). \tag{8.33}$$

Finalmente, la acotación del error (8.30) de la semidiscretización espacial queda:

$$\|E\| \le e^{\alpha t}\,\|E(0)\| + \frac{1}{\alpha}\left(e^{\alpha t} - 1\right) \sup_{t\in[0,t_f]} \|R\|. \tag{8.34}$$

Si $\alpha < 0$, el error espacial con $t \to \infty$ está acotado por el error de truncamiento de la discretización espacial:

$$\|E\| \le \frac{1}{\alpha} \sup_{t\in[0,t_f]} \|R\|\,. \tag{8.35}$$

Si $\alpha > 0$, el error en las condiciones iniciales o el error de truncamiento, por pequeño que sea, se amplifica exponencialmente con $t \to \infty$.

Por esta razón es importante asegurar que la discretización espacial no cambia el carácter de estabilidad del problema diferencial. Generalmente, cuando se discretiza la ecuación de ondas mediante esquemas centrados de diferencias finitas de alto orden en mallas equiespaciadas, se obtienen autovalores que tienen parte real mayor que cero. Estos esquemas o fórmulas de diferencias finitas no son válidas para realizar la discretización espacial. Esta problemática está asociada al mal comportamiento de los interpolantes de alto orden en los contornos cuando la malla es equiespaciada. Este problema se suele resolver concentrando puntos en los contornos como lo hacen los métodos de Chebyshev.

8.5. Error en la ecuación del calor

El comportamiento del error espacial está gobernado por la ecuación (8.28) que incluye la matriz del sistema A y el error espacial de truncamiento R. De igual forma, el comportamiento del error temporal está gobernado por la ecuación (7.21) que incluye la matriz del sistema B y el error temporal de truncamiento T^{n+1}.

En esta sección y a modo de ejemplo, se vuelve a reproducir el procedimiento de obtención de la ecuación del error temporal y espacial para la integración de la ecuación del calor en un dominio unidimensional con condiciones de contorno homogéneas. La discretización espacial se lleva cabo mediante diferencias finitas centradas con tres puntos y la temporal mediante un esquema Euler. La semidiscretización espacial conduce al siguiente sistema de ecuaciones diferenciales ordinarias:

$$\frac{du_j}{dt} = \frac{1}{\Delta x^2}\ (u_{j+1}\ -\ 2u_j\ -\ u_{j-1})\,, \qquad j = 1, \ldots, N-1, \tag{8.36}$$

con $u_0 = 0$ y $u_N = 0$. Restando la ecuación del calor (8.7) menos la ecuación anterior (8.36) en cada punto x_j

$$\left(\frac{\partial u}{\partial t}\right)_{x_j} - \frac{du_j}{dt} = \left(\frac{\partial^2 u}{\partial x^2}\right)_{x_j} - \frac{1}{\Delta x^2}\ (u_{j+1}\ -\ 2u_j\ -\ u_{j-1})\,. \tag{8.37}$$

A continuación, se suma y se resta el operador discreto aplicado a la solución exacta

$$\begin{aligned} \frac{dE_j}{dt}\ =\ & \left(\frac{\partial^2 u}{\partial x^2}\right)_{x_j} - \frac{1}{\Delta x^2}\ (u_{j+1}\ -\ 2u_j\ -\ u_{j-1}) \\ & + \frac{1}{\Delta x^2}\ (u(x_{j+1})\ -\ 2u(x_j)\ -\ u(x_{j-1})) \\ & - \frac{1}{\Delta x^2}\ (u(x_{j+1})\ -\ 2u(x_j)\ -\ u(x_{j-1}))\,. \end{aligned} \tag{8.38}$$

Mediante la definición del error de truncamiento de la ecuación del calor

$$R_j = \left(\frac{\partial^2 u}{\partial x^2}\right)_{x_j} - \frac{1}{\Delta x^2}\ (u(x_{j+1})\ -\ 2u(x_j)\ -\ u(x_{j-1}))\,, \tag{8.39}$$

la ecuación (8.38) queda:

$$\frac{dE_j}{dt}\ =\ R_j + \frac{1}{\Delta x^2}\ (E_{j+1}\ -\ 2E_j\ -\ E_{j-1})\,. \tag{8.40}$$

que es el ejemplo de la ecuación del calor para el sistema general (8.28).

Para obtener la ecuación del error temporal se procede de la siguiente forma. Se considera el sistema (8.36) expresado en forma vectorial con el vector columna $U = (u_1, \dots u_{N-1})^T$

$$\frac{dU}{dt} \;=\; AU, \tag{8.41}$$

donde A es la matriz del sistema que es tridiagonal y está definida por (8.36). A continuación, se discretiza en el tiempo el sistema anterior mediante un esquema Euler y un paso temporal constante Δt

$$U^{n+1} \;=\; U^n \;+\; \Delta t\; A\; U^n. \tag{8.42}$$

La definición del error de truncamiento del esquema temporal (7.16)

$$T^{n+1} = U(t_{n+1}) - U(t_n) \;-\; \Delta t\; A\; U(t_n), \tag{8.43}$$

permite obtener la ecuación del error global de la semidiscretización espacial. Se resta la ecuación (8.43) menos la ecuación (8.42) y como $E^n = U(t_n) - U^n$ se obtiene:

$$E^{n+1} \;= T^{n+1} + \; E^n \;+\; \Delta t\; A\; E^n. \tag{8.44}$$

Si se define la matriz $B = I + \Delta t A$, el sistema anterior queda:

$$E^{n+1} \;= B\; E^n + T^{n+1}, \tag{8.45}$$

que es el ejemplo del esquema Euler para la ecuación del error global (7.21).

8.5.1. Acotación del error espacial

Conocida la matriz del sistema A y conocido el error de truncamiento R, la acotación del error de la semidiscretización espacial pasa por determinar la abscisa espectral de A. En este caso, la matriz A es normal y la determinación de los autovalores de A se puede realizar de forma analítica mediante el siguiente sistema de ecuaciones

$$\frac{1}{\Delta x^2}\;(u_{j+1} \;-\; 2u_j \;-\; u_{j-1}) \;=\; \lambda\; u_j, \qquad j = 1, \dots, N-1, \tag{8.46}$$

con condiciones de contorno $u_0 = 0$ y $u_N = 0$. Es fácil comprobar que existen soluciones no homogéneas de la forma

$$u_j \;=\; \operatorname{sen}(k\pi x_j), \tag{8.47}$$

con k un número entero. Si se introducen estas soluciones o autovectores en (8.46), se obtienen los autovalores correspondientes:

$$\lambda_k \;=\; -\frac{2}{\Delta x^2}\;\left(\,1 - \cos(k\pi\Delta x)\,\right) \;=\; -\frac{4}{\Delta x^2}\;\operatorname{sen}^2\left(\frac{k\pi\Delta x}{2}\right). \tag{8.48}$$

Como la matriz A es normal y cuadrada de dimensión $N-1$, deben existir $N-1$ autovectores y autovalores que se corresponden con los valores de $k = 1, \dots N-1$.

En este caso, la abscisa espectral vale:

$$\alpha \;=\; \sup_{\lambda\in\Lambda(A)} \mathrm{Re}(\lambda) \;=\; -\frac{4}{\Delta x^2}\;\mathrm{sen}^2\left(\frac{\pi\Delta x}{2}\right) \;\approx\; -\frac{4}{\Delta x^2}\;\frac{\pi^2\Delta x^2}{4} \;=\; -\pi^2, \quad (8.49)$$

y la ecuación para la acotación del error de semidiscretización espacial (8.34) se puede escribir como:

$$\|E(t)\| \;\le\; e^{-\pi^2 t}\;\|E(0)\| \;+\; \frac{1}{\pi^2}\;\left(1 \;-\; e^{-\pi^2 t}\right)\;\sup_{\forall\tau}\|R\|. \quad (8.50)$$

A continuación, se calcula el error de truncamiento R mediante (8.39) que coincide con el error de la derivada segunda de un interpolante con tres puntos equiespaciados

$$R_j \;=\; -2\Delta x^2\;\frac{u^{(4)}(\xi)}{3!}\;\left(\frac{d\xi}{dx}\right), \quad (8.51)$$

donde ξ es un punto intermedio que depende de x_j. De esta forma,

$$\|E(t)\| \;\le\; e^{-\pi^2 t}\;\|E(0)\| \;+\; \frac{1}{\pi^2}\;\left(1 \;-\; e^{-\pi^2 t}\right)\;\Delta x^2\;M, \quad (8.52)$$

donde M se obtiene mediante una cota de la derivada cuarta de la función. Se observa que el error en las condiciones iniciales decrece con el tiempo y el error espacial crece con el tiempo y permanece acotado por algo $O(\Delta x^2)$. Se dice que un esquema espacial es de orden q cuando su error de truncamiento es $O(\Delta x^q)$.

8.5.2. Acotación del error temporal

Conocida la matriz del sistema B y conocido el error de truncamiento T, la acotación del error de la semidiscretización temporal pasa por determinar el radio espectral de B. En este caso, la matriz $B = I + \Delta t\;A$ y los autovalores de B son:

$$r_k = 1 + \Delta t\lambda_k = 1 \;-\; \frac{4\Delta t}{\Delta x^2}\;\mathrm{sen}^2\left(\frac{k\pi\;\Delta x}{2}\right). \quad (8.53)$$

En este caso, el radio espectral vale:

$$\rho \;=\; \sup_{r\in\Lambda(B)} |r| \;=\; \left|1 \;-\; \frac{4\Delta t}{\Delta x^2}\right|. \quad (8.54)$$

Para que el error de semidiscretización temporal (7.27) esté acotado, el radio espectral debe ser menor que uno,

$$\frac{\Delta t}{\Delta x^2} < \frac{1}{2}. \quad (8.55)$$

Por otra parte, el error local de truncamiento del esquema Euler (8.43) desarrollando en serie de Taylor alrededor de t_n se puede poner como:

$$T^{n+1} = \frac{\Delta t^2}{2}\ A^2\ U^n + O(\Delta t^3). \tag{8.56}$$

De esta forma, la ecuación (7.27) queda:

$$\|E^n\| \ \leq\ \rho^n\ \|E^0\| \ +\ \left|\frac{1\ -\ \rho^{n+1}}{1\ -\ \rho}\right|\ C\ \Delta t^2, \tag{8.57}$$

donde C se obtiene mediante una cota de la derivada segunda de la función $U(t)$. Si ρ es menor que uno y $n \to \infty$, el error en las condiciones iniciales tiende a cero y el error de truncamiento está acotado por

$$\|E^n\| \ \leq\ C\ \Delta x^2\ \Delta t. \tag{8.58}$$

Así, el error de la discretización temporal del esquema Euler es $O(\Delta t)$. Es importante hacer notar cómo el denominador de la expresión (8.57) reduce un orden el error de truncamiento. Se dice que un esquema temporal para la discretización temporal es de orden q cuando su error local de truncamiento es de $O(\Delta t^{q+1})$.

Apéndice A

Nomenclatura

P_N : Polinomio de grado N

f : Función continua definida en $[a, b]$

ϕ_k : Funciones base.

$\hat{c}_k$: Coeficientes de la serie truncada.

$\tilde{c}_k$: Coeficientes de la serie discreta.

ω : Función de peso en el producto interno continuo.

E_N : Error de interpolación.

R_N : Error de truncamiento para una aproximación de grado N.

R_L : Error de redondeo de Lebesgue para una aproximación de grado N.

I_N : Interpolante polinómico de grado N.

α_j : Coeficientes de peso en el producto interno discreto.

ℓ_j : Polinomio de Lagrange asociado al punto x_j.

π_{N+1} : Función de pi de error de truncamiento de un conjunto de $N + 1$ puntos.

Λ_N : Constante de Lebesgue de un conjunto de $N+1$ puntos.

λ_N : Función de Lebesgue de un conjunto de $N+1$ puntos.

T_k : Polinomio de Chebyshev de grado k.

q : Orden del interpolante continuo a trozos.

Δx_j : Distancia $x_{j+1} - x_j$ o paso espacial.

$\frac{dU}{dt} = F(U,t)$: Sistema de EDOS. $U(t) \in \mathbb{R}^N, F : \mathbb{R}^N \times \mathbb{R} \to \mathbb{R}^N$.

U^n : Solución numérica en el instante t_n. $U^n \in \mathbb{R}^N$.

F^n : $F^n = F(U^n, t_n)$.

$u(x_j, t_n)$: Solución exacta en el punto x_j y en instante t_n.

u_j^n : Solución numérica en el punto x_j y en instante t_n.

Δt_n : Distancia $t_{n+1} - t_n$ o paso temporal.

R^n : Residuo del esquema numérico en el instante t_n.

T^n : Error local temporal de truncamiento en t_n.

$E^n = u_j(t_n) - u_j^n$: Error global temporal en t_n.

$E_j = u(x_j) - u_j$: Error global espacial en x_j.

$E_j^n = u(x_j, t_n) - u_j^n$: Error global espacial y temporal en (x_j, t_n).

R_j : Error de truncamiento espacial en x_j.

$\alpha(A)$: Abscisa espectral de la matriz A.

$\rho(A)$: Radio espectral de la matriz A.

$\mathcal{L}$: Operador diferencial de un problema de contorno.

BC : Operador diferencial de condiciones de contorno.

$\partial\Omega$: Puntos frontera de un dominio espacial Ω.

Bibliografía

[1] Boyd, J. P. 1999 Chebyshev and Fourier Spectral Methods. Dover Publications, Inc.

[2] Canuto, C., Hussaini, M.Y., Quarteroni, A. y Zang, T. A. 2006 Spectral Methods. Fundamentals in Single Domains. Springer.

[3] Hairer, W., Nørsett y Wanner, G. 1993 Solving Ordinary Differential Equations I. Nonstiff Problems. Springer–Verlag.

[4] Hairer, W. y Wanner, G. 1996 Solving Ordinary Differential Equations II. Stiff and Differential –Algebraic Problems. Springer–Verlag.

[5] Isaacson, E. and Keller, H.B. 1966 Analysis of Numerical Methods.

[6] Keller, H.B. 1976 Numerical Solution of Two Point Boundary Value Problems. CBMS–NSF 24. SIAM.

[7] Lambert, J.D. 1991 Numerical Methods for Ordinary Differential Systems. The Initial Value Problem. Wiley.

[8] Press, W. H., Teukolsky, S. A., Vetterling, W.T., Flannery, B.P. 1992 Numerical Recipes in FORTRAN. The Art of Scientific Computing. Cambridge University Press.

[9] shampine, L. F. y Gordon, M. K. 1974 Computer Solution of Ordinary differential Equations. The Initial Value Problem. W. H. Freeman and Company.

[10] Stetter, H. J. 1973 Analysis of Discretization Methods for Ordinary Differential Equations. Springer–Verlag.

[11] Stoer J. and Bulirsh R. 1980 Introduction to Numerical Analysis. Springer-Verlag.

[12] Trefethen, L. N. 2013 Approximation Theory and Approximation Practice. SIAM, Philadelphia.