¡¡APRUEBA TU EXAMEN CON SCHAUM!!

Problemas de Cálculo Numérico para ingenieros con aplicaciones Matlab

Schaum

Juan Miguel Sánchez 3 Antonio Souto

- REDUCE TU TIEMPO DE ESTUDIO
- 80 PROBLEMAS RESUELTOS, EXPLICADOS EN GRADO CRECIENTE DE DIFICULTAD
- EXTENSOS RESÚMENES TEÓRICOS POR CAPÍTULOS CON LAS DEFINICIONES, LOS TEOREMAS CLAVE Y LA DESCRIPCIÓN DETALLADA DE MÉTODOS NUMÉRICOS
- TUTORIAL MATLAB. ACCESO A CÓDIGOS MATLAB

Utilízalo para las siguientes asignaturas:

- ☑ CÁLCULO NUMÉRICO PARA INGENIEROS
- ☑ MÉTODOS NUMÉRICOS EN INGENIERÍA
- ☑ SIMULACIÓN NUMÉRICA DE MODELOS MATEMÁTICOS

PROBLEMAS DE CÁLCULO NUMÉRICO PARA INGENIEROS CON APLICACIONES MATLAB

PROBLEMAS DE CÁLCULO NUMÉRICO PARA INGENIEROS CON APLICACIONES MATLAB

Juán Miguel Sánchez
Antonio Souto

Escuela Técnica Superior de Ingenieros Navales
Universidad Politécnica de Madrid

MADRID • BUENOS AIRES • CARACAS • GUATEMALA • LISBOA • MÉXICO
NUEVA YORK • PANAMÁ • SAN JUAN • SANTAFÉ DE BOGOTÁ • SANTIAGO • SÃO PAULO

PROBLEMAS DE CÁLCULO NUMÉRICO PARA INGENIEROS CON APLICACIONES MATLAB

ISBN: 84-481-2951-2
Depósito legal:

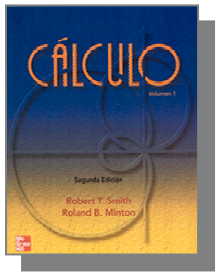
Editor: Jose Manuel Cejudo
Diseño cubierta: Luis Sanz

Impreso en:

IMPRESO EN ESPAÑA – PRINTED IN SPAIN

Otros títulos de interés relacionados

84-481-3861-9 **SMITH/MINTON** . CÁLCULO Y GEOMETRÍA ANALÍTICA.(2 VOLS)



- La segunda edición se ha mejorado mucho con respecto a la anterior. Las características fundamentales son: En la Sección 3.1 se introduce la Regla de L'Hôpital. El desarrollo completo se incluye en la Sección 7.6. Por este motivo, el método de Newton se ha trasladado a la Sección 3.2.
- En el Capítulo 2 se abordan las funciones trigonométricas, exponenciales y logarítmicas, con sus correspondientes reglas de derivación. La integración de estas funciones se desarrolla en el Capítulo 6.
- Mayor número de problemas resueltos
- Atractiva página Web. http://www.mhhe.com/smithminton
- Capítulo 8: Uso de una integral impropia para, acotar el resto de las series a las que se aplica el criterio integral y una sección nueva sobre aplicaciones de las series de Taylor. También se aplica la lista de criterios de convergencia con el criterio de la raíz

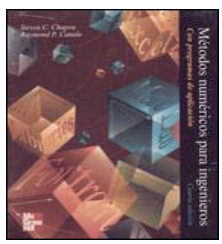
84-4812935-0 **WREDE /SPIEGEL**. CÁLCULO AVANZADO

El Cálculo avanzado no es una teoría única. Sin embargo, las diferentes sub-teorías, incluyendo el análisis vectorial, las series infinitas, y las funciones especiales, son la base de las nociones fundamentales del cálculo. Un importante objetivo de esta segunda edición, ha sido modernizar terminología y conceptos, para que sus interrelaciones sean muy claras. Por ejemplo, continuando con el uso actual de las funciones de una variable real se toman automáticamente las de una variable; la derivadas se definen como funciones lineales, y el carácter universal de la notación y teoría de vectores están muy enfatizadas. Se han incluido otras explicaciones y, en alguna ocasión, con la apropiada terminología que las acompaña

84-481-9840-9 **SPIEGEL/LIU/ABELLANAS.** FÓRMULAS Y TABLAS DE MATEMÁTICA APLICADA. 2º ED. Revisada

- Un libro de ayuda eficaz, de fácil acceso a fórmulas y datos.
- Alrededor de 3.000 fórmulas y tablas.
- Contiene unas amplias secciones sobre Estadística, Armónicos Esféricos y Métodos Numéricos.
- Muy apropiado para estudiantes de Ciencias e Ingeniería así como de Economía, LADE y Empresariales

97-010-3965-3 **CHAPRA**. MÉTODOS NUMÉRICOS PARA INGENIEROS CON APLICACIONES



La cuarta edición de Métodos numéricos para ingenieros ofrece una presentación accesible e innovadora que continúa la tradición de excelencia que ha establecido como ganador del premio Meriam-Wiley, otorgado por la American Society for Engineering Education al mejor librodetexto. La obra conserva la exitosa estructura didáctica que ha cracterizado las ediciones anteriores.

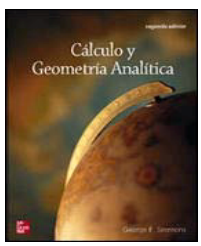
84-481-4212-8 **WARD/CHURCHILL**. VARIABLE COMPLEJA Y APLICACIONES

Este libro es una revisión de la sexta edición, publicada en USA en 1996. Esta edición, al igual que las anteriores, ha servido como libro de texto en cursos de introducción a la teoría y aplicaciones de las funciones de variable compleja. Esta nueva edición mantiene el contenido básico y el estilo de las que la precedieron. En esta edición, los cambios más relevantes aparecen en los nueve primeros capítulos, que constituyen el núcleo de un curso básico. Los tres capítulos restantes se dedican a algunas aplicaciones físicas, que admiten selección a gusto de cada cual y pueden ser estudiadas como complemento para el estudiante interesado.

97-010-2985-2 **SPIEGEL.** MATEMÁTICAS AVANZADAS

Contiene 950 problemas resueltos.Incluye cientos de ejercicios de práctica.Perfecto para estudiar por tu cuenta.Provee conceptos y métodos para ingenieros, matemáticas y disciplinas afines.Es un excelente complemento para cualquier llibro de texto.

84-481-3591-1 **SIMMONS**. CÁLCULO Y GEOMETRÍA ANALÍTICA



Texto del prestigioso autor y matemático George Simmons, diseñado para un primer curso de cálculo. Proporciona una aproximación intuitiva al Cálculo centrada en la aplicación de los métodos a problemas del mundo real. Características:Introduce la trigonometría al principio del texto de forma breve y a modo de repaso. Ecuaciones diferenciales no lineales, probabilidad elemental y funciones hiperbólicas. Completa cobertura de los tópicos empleados en Cálculo para ingeniería. Estilo elegante y comprensible para el estudiante, con un desarrollo creciente en los contenidos. Resolución de problemas de forma intuitiva, omitiendo los desarrollos teóricos innecesarios, así como las demostraciones de muchos teoremas proporcionando una visión realista del Cálculo. Biografías de matemáticos y anécdotas curiosas, que hacen el texto más ameno.

A Paula,

Juan Miguel Sánchez

A Carlos Pantaleón Prieto,
a José Luis Ramallo Olmos,

A. Souto Iglesias

Sobre los autores

Juan Miguel Sánchez Sánchez es Doctor Ingeniero Naval y profesor del área de Matemática Aplicada en la Escuela Técnica Superior de Ingenieros Navales (ETSIN) de la Universidad Politécnica de Madrid (UPM) desde el año 1976. De amplia experiencia docente en Álgebra y Análisis multilineal, Geometría diferencial, Ecuaciones diferenciales y Cálculo numérico, su investigación en Micromecánica computacional es fuente inagotable de problemas, experimentos y proyectos didácticos.

Antonio Souto Iglesias es Ingeniero Naval y Doctor Ingeniero Naval por la UPM. Es Profesor del área de Matemática Aplicada en la ETSIN de la UPM desde el año 1994, donde imparte docencia en Cálculo Numérico e Informática. Desarrolla su labor investigadora en Métodos Numéricos en Hidrodinámica dentro del grupo de investigación del Canal de Ensayos Hidrodinámicos de la ETSIN (*http://canal.etsin.upm.es*).

Contenido

Introducción

El objetivo de este libro es servir como texto de apoyo en los cursos de Cálculo Numérico que con un peso específico de entre 6 y 9 créditos forman parte de todos los estudios de Ingeniería. En cada capítulo existe una introducción teórica que creemos suficiente (desde luego lo es para los problemas aquí referidos), pero damos además referencias para completarla. El libro nace a partir de la enseñanza de una asignatura de este tipo en la Escuela Técnica Superior de Ingenieros Navales de la Universidad Politécnica de Madrid.

La orientación de los problemas implica que con esta colección no se pretenda cubrir todos los aspectos del Cálculo Numérico, pero sí se pretende que muestre de manera clara aquellos que puedan ser más importantes para los ingenieros, y en este sentido la selección de contenidos no es inocente. Hemos dejado fuera cosas que están en la mayoría de los textos enciclopédicos de Cálculo Numérico y que se han estudiado en este tipo de carreras, pero que creemos que ahora han perdido importancia, sustituidas por otras. Ello tiene que ver con que el estudiante dispone de medios de cálculo acorde con los tiempos, lo que le permite afrontar problemas de un modo que hasta hace poco era inviable. Así, parte de los mismos incorporan instrucciones Matlab, las cuales permiten resolver de modo eficiente y preciso la parte que tienen de cálculo puro, y visualizar claramente los resultados. Además, esta forma de hacer Cálculo Numérico hace innecesario el estudio de atajos para su aplicación a la resolución de problemas sencillos pero con cierta carga de cálculo, permitiendo al estudiante centrarse en la esencia de los métodos. De hecho, en la vida profesional, la realización de cálculos para el proyecto mediante aplicaciones complejas exige del usuario de estas aplicaciones un conocimiento de los conceptos básicos de discretización de problemas del continuo y de los errores que estas discretizaciones acarrean; este libro incide en estos conceptos.

Matlab es un programa de uso casi estándar en muchas ramas de la Ingeniería, y la tendencia es que su implantación será mayor en el futuro. Así, nos parece fundamental que el estudiante se encuentre cómodo con su utilización. No pretendemos llegar a los detalles de un usuario más avanzado del programa, pero sí que se intuya su potencia, y de hecho nuestra experiencia nos indica que es positivo permitir su utilización en los exámenes. Incorporamos como capítulo adicional un tutorial de Matlab que creemos que debe ser el capítulo inicial para aquellos estudiantes que no estén familiarizados con el mismo. Algunos de los problemas se completan con códigos con los que podemos aumentar la complejidad de los cálculos. Querer resolver con precisión suficiente un problema de Ingeniería exige iterar sobre cálculos elementales y eso lleva a la Programación de Ordenadores, que es una disciplina que permite extraer la utilidad real al Cálculo Numérico. Los códigos a los que nos referimos en el texto pueden ser descargados directamente de la web *http://canal.etsin.upm.es/libroftp/*. Podría parecer interesante tener unas nociones de utilización de parte simbólica de Matlab (que en realidad es Maple) para algunas simplificaciones, pero creemos que es sobrecargar un curso cuya esencia está más en la programación.

Aunque los problemas están agrupados en los capítulos habituales de un curso de introducción al Cálculo Numérico, en realidad su orientación ingenieril hace que en ellos se mezclen técnicas procedentes de diferentes partes de la teoría. Esto los hace adecuados también como proyectos de programación para trabajo en equipo; el orientar parcialmente la asignatura en esa dirección nos parece muy provechoso. Como requisitos previos, al estudiante se le supone haber pasado ya por los cursos de Física, Análisis Matemático y Álgebra Lineal y disponer de nociones básicas de algún lenguaje de programación. Respecto a la precisión con que realizamos los cálculos, nos gustaría decir que creemos que a este tema se le da una importancia demasiado grande en los cursos iniciales de Cálculo Numérico. En Ingeniería, los errores proceden en la mayoría de los casos más de las simplificaciones realizadas en los modelos que de los redondeos en los cálculos. Para presentar los resultados intermedios hemos truncado sin redondeo en el cuarto decimal, salvo que el número resultante no tuviese suficientes cifras significativas. Aun con esta información, es muy probable que los resultados numéricos puedan diferir un poco, dependiendo de cómo se vayan arrastrando estos errores de redondeo a lo largo de los diferentes apartados.

Asumimos que no es posible hacer un libro sin erratas y agradecemos comentarios relativos a éstas y a cualquier otro aspecto que enriquezca el presente trabajo. Rogamos se los haga llegar a *asouto@etsin.upm.es*.

Notación y abreviaturas

- Para indicar que una igualdad lo es por definición, utilizamos $:=$
- Cuando ponemos $i = 0, n$, significa que el índice i recorre todos los naturales entre 0 y n.
- e.v.n. por espacio vectorial normado.
- edos por ecuaciones diferenciales ordinarias.
- edps por ecuaciones diferenciales en derivadas parciales.
- edppo por ecuaciones diferenciales en derivadas parciales de primer orden.
- edpso por ecuaciones diferenciales en derivadas parciales de segundo orden.
- m.a. por mejor aproximación.
- pág. por página.
- ssi por si y sólo si.
- s.e.v. por subespacio vectorial.
- 2D, 3D por dos y tres dimensiones respectivamente.
- $\mathbb{Z}$ representa el conjunto de los enteros relativos.
- $\overline{\Omega}$ es la adherencia de Ω ($\overline{\Omega} = \Omega \cup \partial\Omega$).

CAPÍTULO 1

Resolución de ecuaciones no lineales

La búsqueda de raíces de ecuaciones no lineales es uno de los problemas más frecuentes en matemática aplicada.

La resolución de ecuaciones algebraicas

$$x^n + a_1 x^{n-1} + a_2 x^{n-2} + \cdots + a_{n-1} x + a_n = 0$$

ocupó buena parte de los esfuerzos de la matemática de los siglos XVI al XIX.

Las ecuaciones de primer y segundo grado habían sido ya resueltas en la antigüedad por lo que el objetivo central de los matemáticos del Renacimiento fue la resolución de las ecuaciones algebraicas de orden superior. La resolución de las ecuaciones de tercer y cuarto grado fueron los logros máximos de los algebristas italianos de este periodo [1]. El efecto que produjo este éxito fue enorme. Por primera vez la ciencia moderna superaba los logros de la matemática antigua y eso señaló un camino claro a seguir. No hubo matemático importante que no intentase, sin éxito, en los tres siglos siguientes, extender los resultados de los matemáticos italianos a la resolución de las ecuaciones de quinto, sexto y grados superiores de un modo análogo. Durante este periodo de tiempo, nadie dudaba de la posibilidad de poder expresar la solución de las ecuaciones algebraicas en función de sus coeficientes mediante fórmulas que implicasen sólo las operaciones de suma, resta, multiplicación, división y radicación con exponentes enteros positivos.

Fue una gran sorpresa la publicación en 1824 del trabajo del joven genio noruego Abel (1802-1829) en el que demostraba la imposibilidad de expresar las soluciones de cualquier ecuación algebraica de grado ≥ 5 mediante radicales.

A la vista de ello los matemáticos comenzaron a abrir nuevas direcciones en el estudio de las ecuaciones algebraicas utilizando el impresionante edificio de teoremas y métodos relacionados con el problema que se habían construido en los tres siglos anteriores.

Las tres más importantes fueron

- El estudio del problema de la existencia de una raíz.
- La obtención de propiedades de las raíces de la ecuación a partir de sus coeficientes sin resolverla. (¿Cuántas son?, ¿son reales o no?, etc.)
- El cálculo aproximado de las raíces de una ecuación (sólo se podían resolver por radicales unos pocos casos de poco valor práctico debido a la complejidad de las expresiones de sus raíces).

Esta última dirección es el objetivo de este capítulo.

Dos de las tres grandes estrategias para el diseño de algoritmos iterativos de busca de la raíz x^* de una ecuación $f(x) = 0$ aproximan en cada paso del proceso la función f en el entorno de la última estimación de x^*, eligiendo la siguiente estimación de entre los ceros de la función aproximante.

Una de ellas aproxima f con un polinomio de interpolación cuyo grado forma parte de un compromiso entre la precisión, el costo numérico y las demás características a considerar en el proceso. Una vez fijado

[1] Scipio del Ferro, Tartaglia, Cardano y Ferrari son los nombres más relevantes de esta historia que abarca un periodo de tiempo relativamente corto del siglo XVI.

dicho grado se sabe el número de estimaciones anteriores que se necesitan para determinarlo. Esta estrategia es la base de los métodos de interpolación (método de bisección de "regula falsi", de la secante y variantes).

La segunda estrategia aproxima f en el entorno de la última iterante mediante un polinomio de Taylor cuyo grado depende de la calidad de f y del compromiso antes comentado. Esta línea tiene su origen con el método de linealización de Newton que aproxima la gráfica de f en el punto con su recta tangente en dicho punto [2]. El método de Newton no sólo es el método más usado de todos los buscadores de raíces sino también el más influyente, ya que continuamente siguen apareciendo nuevos métodos que son modificaciones y variantes suyas.

Una tercera estrategia vincula la resolución de la ecuación dada con la resolución de una ecuación del tipo $x = T(x)$ (ecuación de punto fijo) que se debe construir a partir de la original. Una vez que se pruebe que ambas ecuaciones son equivalentes y que la función T cumple ciertas condiciones suficientes (teorema del punto fijo) [3] se resuelve la ecuación de punto fijo mediante el proceso iterativo $x^{(k+1)} = T(x^{(k)})$. Esta estrategia es la base del método de aproximaciones sucesivas y variantes cuya importancia es enorme.

1.1. Problema inverso no lineal

Un gran número de problemas que se plantean en matemática aplicada tienen como base una ecuación del tipo

$$R(\mathbf{x}) = \mathbf{y} \tag{1.1}$$

con $\mathbf{x} \in E$, $\mathbf{y} \in F$ donde E y F son dos espacios vectoriales reales o complejos y R es un operador de E en F no lineal en general. El **problema inverso** asociado a la ecuación (1.1) es encontrar $\mathbf{x}$, dados R e $\mathbf{y}$.

Nuestro objetivo es resolver este problema [4].

Por conveniencia, escribiremos (1.1) cuando las estructuras algebraicas de los espacios E y F lo permitan, en las formas

$$\mathbf{x} = T(\mathbf{x}) \tag{1.2}$$

y

$$f(\mathbf{x}) = \mathbf{0}_F \tag{1.3}$$

en cuyo caso, los datos, R e $\mathbf{y}$, del problema inverso estarán implícitos en la estructura de los operadores T y f.

Los tres problemas definidos por esas ecuaciones están muy relacionados y el paso de una a otra formulación forma parte de la estrategia para resolverlos.

El paso de la ecuación (1.1) a las ecuaciones (1.2) y (1.3) y de ellas entre sí, siempre es posible y de infinitas formas, siendo unas más útiles que otras para nuestro propósito.

En el primer caso, el problema inverso general equivale al problema de encontrar un $\mathbf{x} \in E$ tal que $\mathbf{x} = T(\mathbf{x})$, donde T es una función dada de E en sí mismo.

Las soluciones, si existen, se llaman puntos fijos de T (**problema de punto fijo**) y la ecuación (1.2) se llama de **punto fijo**.

En el segundo caso, el problema asociado al problema inverso general es el de la búsqueda del conjunto de las raíces o ceros de f (**cálculo de los ceros de** f).

En general, E y F serán dos espacios de Banach [5], pero aquí nos ocuparemos de los problemas asociados

[2] Ver en [13] el cálculo aproximado de la raíz próxima a 2 de la ecuación $x^3 - 2x - 5 = 0$ tratado por Newton en 1671 y por E. Halley en 1694 con un polinomio de Taylor aproximante de segundo grado.

[3] El teorema del punto fijo o de la aplicación contractiva de Banach (1.4.2) es uno de los más importantes teoremas de existencia. Se utiliza en demostraciones de existencia de soluciones para ecuaciones algebraicas, ecuaciones diferenciales e integrales y con su ayuda podemos construir soluciones aproximadas.

[4] Dependiendo de la estructura del operador R y de los espacios vectoriales E y F, se plantean así los problemas de resolución de ecuaciones lineales y no lineales, tanto algebraicas como diferenciales o integrales.

[5] Un espacio de Banach E es un espacio vectorial real o complejo, provisto de una norma con las siguientes propiedades

N1. $\|\mathbf{x}\| \geq \mathbf{0} \qquad (\forall \mathbf{x} \in E)$

N2. $\|\mathbf{x}\| = \mathbf{0} \qquad \Leftrightarrow \qquad \mathbf{x} = \mathbf{0}$

N3. $\|\lambda \mathbf{x}\| = |\lambda| \|\mathbf{x}\| \qquad (\forall \lambda \in I\!R (\text{o } \mathbb{C}), \mathbf{x} \in E)$

N4. $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\| \qquad (\forall \mathbf{x}, \mathbf{y} \in E)$

El cuerpo de los números reales $I\!R$ (resp: el de los números complejos) es un espacio de Banach con la norma definida por el valor absoluto (resp: por el módulo). Los espacios $I\!R^n$ (resp: $\mathbb{C}^n$) son espacios de Banach para cualquiera de sus normas al

al cálculo numérico de los ceros tanto de las funciones reales de variable real $f : \mathbb{R} \to \mathbb{R}$ como de las funciones vectoriales de varias variables reales $f : \mathbb{R}^n \to \mathbb{R}^m$.

1.2. Métodos iterativos de cálculo aproximado de raíces

En general no se puede determinar una raíz de la ecuación (1.3) o un punto fijo de T, explícitamente en un número finito de pasos, por ello se recurre a métodos aproximados iterativos.

La idea es construir una sucesión $\{\mathbf{x}^{(k)}\}_{k=0,1,..}$ de vectores de E mediante una fórmula de recurrencia, de modo que al crecer k sus términos se vayan aproximando en la norma de E cada vez más a la raíz de f o al punto fijo de T según sea el caso.

Una clase muy importante de métodos iterativos tienen la forma

$$\mathbf{x}^{(k+1)} = T(\mathbf{x}^{(k)}) \quad k = 0, 1, ... \tag{1.4}$$

Dado un valor inicial $\mathbf{x}^{(0)}$, la función de iteración T es independiente de k y sólo depende de la iterante anterior $\mathbf{x}^{(k)}$, luego caracteriza un método estacionario de un paso[6], el método de **aproximaciones sucesivas**, que define por recurrencia la sucesión (1.4) cuyo objetivo es resolver la ecuación (1.2), es decir, hallar un punto fijo $\mathbf{x}^*$ de T (**iteración de punto fijo**).

Definición 1.2.1 *El conjunto de todos lo puntos $\mathbf{x}^{(0)}$ tales que la sucesión (1.4) que definen converge a $\mathbf{x}^*$ se llama el dominio de atracción de $\mathbf{x}^*$.*

1.2.1. Órdenes de convergencia de un método iterativo

Definición 1.2.2 *Diremos que la sucesión $(\mathbf{x}^{(k)})$ converge a $\mathbf{x}^*$ con orden p al menos, si existen un rango k_0 y una constante $C > 0$ tales que*

$$\|\mathbf{x}^* - \mathbf{x}^{(k+1)}\| \leq C\|\mathbf{x}^* - \mathbf{x}^{(k)}\|^p$$

igual que cualquier espacio vectorial real (o complejo) de dimensión finita (los espacios de matrices por ejemplo, son de uso constante).

El espacio $\mathcal{C}_\infty([a,b])$ de las funciones reales y continuas en un intervalo cerrado y acotado $[a,b]$ provisto de la norma del máximo

$$\|f\|_\infty = \max_{x\in[a,b]} |f(x)|$$

es el ejemplo más simple de un espacio de Banach de funciones.

[6] En la estructura general de un método iterativo se conocen

- $(m+1)$ **valores iniciales** $\mathbf{x}^{(0)}, \mathbf{x}^{(1)}..., \mathbf{x}^{(m)}$ vectores de E que aproximan mejor o peor a la raíz que buscamos y que son los $(m+1)$ primeros términos de la sucesión iterante.
- una sucesión de funciones $(\phi_k)_{k\geq m+1}$ donde ϕ_{k+1} **es la función de iteración del paso k-ésimo** que actualiza $\mathbf{x}^{(k)}$ definiendo $\mathbf{x}^{(k+1)}$ mediante la ecuación recurrente

$$\mathbf{x}^{(k+1)} = \phi_{k+1}\left(\mathbf{x}^{(0)}, \mathbf{x}^{(1)}, ..., \mathbf{x}^{(k)}\right) \tag{1.5}$$

Conocidos los $m+1$ primeros términos de la sucesión, $\mathbf{x}^{(0)}, \mathbf{x}^{(1)}..., \mathbf{x}^{(m)}$ tendremos utilizando (1.5)

$$\begin{cases} \mathbf{x}^{(m+1)} = \phi_{m+1}\left(\mathbf{x}^{(0)}, \mathbf{x}^{(1)}, ..., \mathbf{x}^{(m)}\right) \\ \mathbf{x}^{(m+2)} = \phi_{m+2}\left(\mathbf{x}^{(0)}, \mathbf{x}^{(1)}, ..., \mathbf{x}^{(m)}, \mathbf{x}^{(m+1)}\right) \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ \mathbf{x}^{(k+1)} = \phi_{k+1}\left(\mathbf{x}^{(0)}, \mathbf{x}^{(1)}, ..., \mathbf{x}^{(k)}\right) \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \end{cases}$$

Si cualquiera que sea k, la función ϕ_{k+1} que define la iteración del paso k-ésimo, sólo depende de las s iterantes $\mathbf{x}^{(k-s+1)}, ..., \mathbf{x}^{(k)}$ previas a $\mathbf{x}^{(k+1)}$

$$\mathbf{x}^{(k+1)} = \phi_{k+1}\left(\mathbf{x}^{(k-s+1)}, ..., \mathbf{x}^{(k)}\right) \tag{1.6}$$

se dice que el esquema iterativo es de **s-pasos**.

Por último, cuando ϕ_k **no depende de** k, es decir, cuando la función de iteración es la misma en todos los pasos, el método iterativo se dice que es **estacionario** (1.6).

para todo $k \geq k_0$.

Definición 1.2.3 *Un método iterativo se dice que es de orden* p ($p \in \mathbb{R}$) *si genera una sucesión* $(\mathbf{x}^{(k)})$ *de iterantes de esas características.*

En particular, cuando existe

$$\lim_{k\to\infty} \frac{\|\mathbf{x}^{(k+1)} - \mathbf{x}^*\|}{\|\mathbf{x}^{(k)} - \mathbf{x}^*\|^p} = K \tag{1.7}$$

con $0 < K < \infty$ y $K < 1$ para $p = 1$, la sucesión $(\mathbf{x}^{(k)})$ tiene orden de convergencia al menos p.

Si $K \neq 0$ diremos que p es el orden de convergencia de $(\mathbf{x}^{(k)})$ y del método que la genera. Para $p = 1, 2$ y 3 hablaremos de convergencia al menos lineal, cuadrática y cúbica, respectivamente.

La pareja (p, K) formada por el orden de convergencia y la constante asintótica de error caracteriza el comportamiento de la sucesión iterante $(\mathbf{x}^{(k)})$. Cuanto mayor es p y menor es K, mayor es la rapidez de convergencia de $(\mathbf{x}^{(k)})$ a $(\mathbf{x}^*)$ al menos asintóticamente.

Definición 1.2.4 *Se dice que una sucesión* $(\mathbf{x}^{(k)})$ *converge superlinealmente a* $(\mathbf{x}^*)$ *si*

$$\|\mathbf{x}^{(k+1)} - \mathbf{x}^*\| \leq \beta_k \|\mathbf{x}^{(k)} - \mathbf{x}^*\| \tag{1.8}$$

para alguna sucesión infinitésima β_k *de números reales positivos.*

Es fácil probar las implicaciones

$$\text{convergencia superlineal} \implies \text{convergencia lineal}$$
$$\text{convergencia superlineal} \impliedby \text{convergencia cuadrática}$$

1.3. Ecuaciones no lineales de una variable real

Analizaremos aquí el problema del cálculo numérico de los ceros de funciones reales de variable real. Lógicamente, la dificultad del problema depende de la estructura de la función f que lo define. Si deseamos encontrar numéricamente una sóla raíz de la ecuación $f(x) = 0$, el proceso tiene dos fases distintas:

- Acotar la raíz enmarcándola en un intervalo que la contenga a ella sola. Se podrá asegurar lo anterior cuando, siendo la función monótona en ese intervalo, f tome signos opuestos en sus extremos.
- Tomando como aproximación inicial la estimación hallada en la fase anterior se procede a refinarla mediante un proceso sistemático hasta alcanzar la precisión deseada.

Si lo que se quiere es obtener numéricamente las raíces de la ecuación no lineal $f(x) = 0$ con una aproximación arbitraria, debemos comenzar localizando esas raíces, es decir, debemos definir una partición del conjunto de busca en intervalos parciales en los que se sepa que o bien no hay ninguna raíz o bien hay una sola[7].

Una vez **separadas** las raíces se les podrá dar un tratamiento individualizado.

1.3.1. Método "regula falsi" y variantes (Método de bisección. Método de la secante. Métodos Illinois y Pegasus)

Métodos de interpolación

Una de las estrategias básicas para diseñar algoritmos de cálculo de ceros de una función f es aproximarla en cada paso mediante un polinomio de interpolación de las aproximaciones anteriores, eligiendo la siguiente iterante de entre los ceros de dicho polinomio.

Si se quiere que el grado del polinomio de interpolación sea r, se deben conocer $r+1$ valores aproximados distintos $x^{(k-r)}, x^{(k-(r-1))}, \ldots, x^{(k-1)}, x^{(k)}$ de la raíz x^* de la ecuación $f(x) = 0$.

El método general de interpolación determina un polinomio Q de grado r tal que

$$Q(x^{(k-j)}) = f(x^{(k-j)}) \qquad j = 0, 1, \ldots, r$$

[7]Para obtener esa descomposición es necesario estudiar el signo de la derivada f' lo que obliga a hallar las raíces de la ecuación $f'(x) = 0$, problema del mismo orden de dificultad que la resolución de $f(x) = 0$.

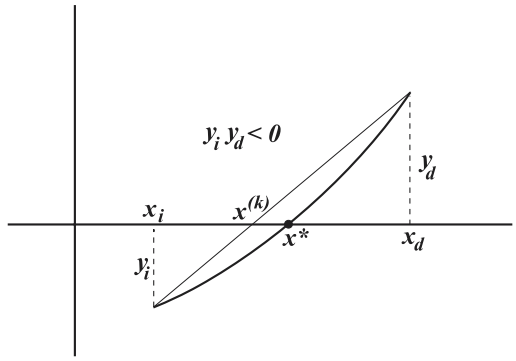


Figura 1.1: Un paso del método "regula falsi".

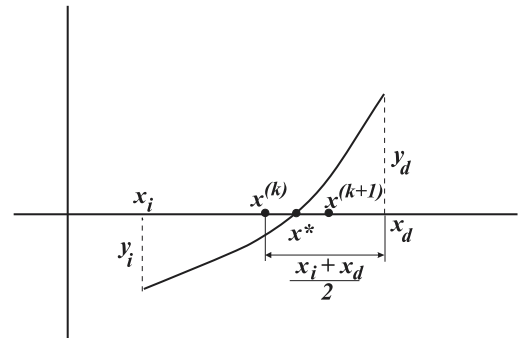


Figura 1.2: Dicotomía de Bolzano.

y se elige como siguiente iterante $x^{(k+1)}$, la raíz de Q más próxima a $x^{(k)}$.

Los métodos de interpolación tienen varias ventajas sobre los métodos obtenidos a través de los polinomios de Taylor de f. No es necesario evaluar las derivadas de f y en un sentido a precisar son más rápidos que éstos. En general, $r = 1$ o 2.

Interpolación lineal. Métodos "regula falsi" y variantes

El algoritmo del método "regula falsi" o de la posición falsa se puede describir del siguiente modo

- Se conoce un intervalo (x_i, x_d) que contiene sólo la raíz x^*
- Se calculan $y_d = f(x_d)$ e $y_i = f(x_i)$ (claramente $y_d y_i < 0$)
- Se aproxima x^* con la abscisa del punto de intersección con el eje Ox de la recta secante (**interpolación lineal**) que pasa por los puntos (x_i, y_i) y (x_d, y_d) (**paso de secante**), ver la Figura (7.1), cuya ecuación es $y = Ax + B$ con

$$A = \frac{y_d - y_i}{x_d - x_i} \qquad \text{y} \qquad B = \frac{x_d y_i - x_i y_d}{x_d - x_i}$$

  de donde la abscisa $x^{(k)}$ de la siguiente iterante es

$$x^{(k)} = -\frac{B}{A} = \frac{x_i y_d - x_d y_i}{y_d - y_i}; \qquad y^{(k)} = f(x^{(k)})$$

- Se evalúa $P = y_d y^{(k)}$ que se usa como herramienta para determinar en qué intervalo $(x_i, x^{(k)})$ o $(x^{(k)}, x_d)$ está la raíz.

  Si $P > 0$, $x^* \in (x_i, x^{(k)})$, luego $x_d = x^{(k)}$ e $y_d = y^{(k)}$.

  Si $P < 0$, $x^* \in (x^{(k)}, x_d)$, luego $x_i = x^{(k)}$ e $y_i = y^{(k)}$
- Se reitera.
- Se para el algoritmo al activarse algún test de parada.
- Se toma como solución

$$x^* = \frac{x_i - x_d}{2}$$

Es un método no estacionario de dos pasos cuyo orden de convergencia es lineal.

Variantes del método "regula falsi"

1. **Método de bisección o de dicotomía de Bolzano**

   Se dan x_i y x_d tales que f es continua en $[x_i, x_d]$ y $f(x_i)f(x_d) < 0$, luego sabemos que hay un cero x^* de f en (x_i, x_d) que aproximamos por dicotomía, es decir, se toma como siguiente iterante $x^{(k)}$ el punto medio del intervalo (x_i, x_d) (Figura 1.2)

   $$x^{(k)} = \frac{x_i + x_d}{2}$$

   En cada paso, la longitud del nuevo intervalo que contiene a x^* se divide entre dos y el proceso iterativo se continua hasta que x^* esté en un intervalo de longitud suficientemente pequeña. Si el intervalo inicial de búsqueda de la raíz es (a, b) con $b > a$, llamando $L_0 = b - a$ a su longitud entonces la sucesión $L_k = \frac{b-a}{2^{k+1}}$ tiende a cero. El punto medio $x^{(k)}$ del intervalo después de k bisecciones, es una aproximación de x^* con una estimación "a priori" del error,

   $$\left| x^{(k)} - x^* \right| \leq \frac{b-a}{2^{k+1}} \qquad k = 0, 1, 2, ...$$

   La cota de error disminuye como una sucesión geométrica de razón $(q = 1/2)$ (orden de convergencia $p = 1$).

2. **Método de la secante**

   El método de la secante tiene una mecánica similar al de la falsa posición, pero ahora x^* no tiene que estar necesariamente en $(x^{(k-1)}, x^{(k)})$ por lo que también se aplica cuando $f(x^{(k-1)})f(x^{(k)}) > 0$ aun cuando algunas veces no converja.

   La sucesión de las iterantes se genera por la fórmula estacionaria de dos pasos

   $$x^{(k+1)} = \frac{x^{(k-1)} f(x^{(k)}) - x^{(k)} f(x^{(k-1)})}{f(x^{(k)}) - f(x^{(k-1)})} \tag{1.9}$$

   Se recomienda usarlo sólo en entornos muy próximos a la raíz.

3. **Métodos Illinois y Pegasus**

   Estos métodos [7] son generalizaciones del método de "regula falsi" que utilizan un paso de secante modificado con el objetivo de acelerar la convergencia lineal de dicho método.

   El algoritmo asociado comparte los dos primeros pasos con el de la falsa posición.

   - Se conoce un intervalo (x_i, x_d) que contiene la raíz x^*.
   - Se calculan $y_d = f(x_d)$ e $y_i = f(x_i)$ (claramente $y_d y_i < 0$).
   - Se define $x^{(k)} = \frac{x_i y_d - x_d y_i}{y_d - y_i}$ de abscisa del punto de intersección con el eje Ox de la secante que pasa por los puntos (x_i, y_i) y (x_d, y_d).
   - Se evalúa $P = y_d y^{(k)}$.

   Si $P > 0$

   $$\begin{aligned} x_d &= x^{(k)}, & x_i &= x_i \\ y_d &= y^{(k)}, & y_i &= \alpha y_i \end{aligned}$$

   Si $P < 0$,

   $$\begin{aligned} x_i &= x^{(k)}, & x_d &= x_d \\ y_i &= y^{(k)}, & y_d &= \alpha y_d \end{aligned}$$

   donde $\alpha = \frac{1}{2}$ en el método Illinois y

   $$\begin{aligned} \alpha &= \frac{y_d}{y_d + y^{(k)}}, & P > 0 \\ \alpha &= \frac{y_i}{y_i + y^{(k)}}, & P < 0 \end{aligned}$$

en el método Pegasus que tiene orden de convergencia superlineal ≈ 1.642.

Ambos métodos se usan exhaustivamente en los problemas (1.11), (1.12) y (1.13). Allí se incluyen sus programas Matlab y se hacen comparaciones de su comportamiento con otros métodos.

1.3.2. Iteración de punto fijo. Métodos de Wegstein y de relajación. Aceleración de la convergencia. Método Δ^2 de Aitken. Método de Steffensen

Iteración de punto fijo

Consideramos aquí la iteración de punto fijo y su aplicación a la resolución numérica de la ecuación

$$f(x) = 0 \tag{1.10}$$

Para calcular numéricamente la raíz x^* de (1.10), buscamos una función T que nos permita reescribir esa ecuación en la forma $x = T(x)$ de modo que el cálculo de x^* sea equivalente al cálculo de un punto fijo de T, luego que

$$f(x^*) = 0 \qquad \Longleftrightarrow \qquad x^* = T(x^*)$$

Una vez definida T con esas características, hallamos x^* utilizando el esquema iterativo (1.4) del método de aproximaciones sucesivas con T como función de iteración.

Ahora bien, ¿es esto siempre posible?, es decir, ¿es siempre posible hallar una función de iteración T tal que la sucesión iterante asociada $(x^{(k)})$ converja a x^* para una estimación inicial $x^{(0)}$ de x^* suficientemente buena?

Contestaremos esta pregunta con dos teoremas que darán una condición suficiente para que la sucesión de las iterantes converja a x^* analizando su comportamiento y orden de convergencia en dos situaciones diferentes.

Teorema 1.3.1 *Sea $I = [a, b]$, si $T : I \to I$ es una función continua con derivada primera continua y no nula en I tal que $|T'(x)| \leq L < 1$, $(\forall x \in I)$* [8] *entonces cualquiera que sea $x^{(0)} \in I$, la sucesión (1.4) converge a un punto fijo x^* de T* [9]*. Además llamando $\epsilon_k = x^{(k)} - x^*$ se tiene*

$$\lim_{k\to\infty} \frac{\epsilon_{k+1}}{\epsilon_k} = T'(x^*) \tag{1.11}$$

Para un valor de k suficientemente grande, el error ϵ_k disminuye como una progresión geométrica de razón $K \sim T'(x^*)$

$$\epsilon_{k+1} \approx K\epsilon_k \qquad |K| < 1 \qquad k \gg 1$$

*La convergencia es lineal con $|T'(x^*)|$ el **factor de convergencia** como constante asintótica de error.*

En el caso en que no se satisfacen las condiciones del teorema 1.3.1 porque $T'(x^*) = 0$, la convergencia de la sucesión iterante se caracteriza en el teorema siguiente.

Teorema 1.3.2 *Sea $I = [a, b]$, si $T : I \to I$ es de clase $\mathcal{C}^2$ en I con $|T'(x)| < 1$ $(\forall x \in I - \{x^*\})$, $T'(x^*) = 0$ y $T''(x) \neq 0$ $(\forall x \in I)$ entonces el error $\epsilon_k = x^{(k)} - x^*$ cumple,*

$$\lim_{k\to\infty} \frac{\epsilon_{k+1}}{\epsilon_k^2} = \frac{1}{2}T''(x^*) \tag{1.12}$$

Luego para k suficientemente grande

$$\epsilon_{k+1} = K\epsilon_k^2 \quad con \quad K = \frac{1}{2}|T''(x^*)| \quad y \quad k \gg 1$$

ϵ_{k+1} es proporcional a ϵ_k^2 con constante de proporcionalidad K independiente de k.

La convergencia de (x_k) es cuadrática con $\frac{1}{2}|T''(x^*)|$ como constante asintótica de error.

[8] Esta condición implica a través del teorema del valor medio que T es una aplicación contractiva en I (definición 1.4.2).

[9] El enunciado de este teorema es un caso particular del teorema del punto fijo de Banach (teorema 1.4.2) de aplicación más amigable.

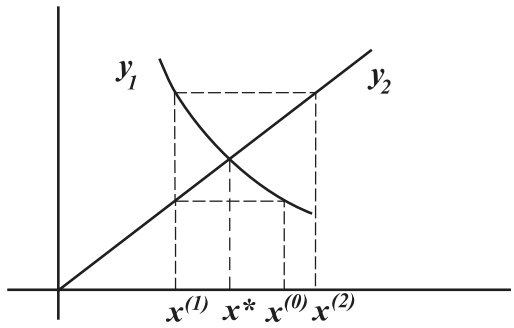


Figura 1.3: $T'(x^*) < -1$.

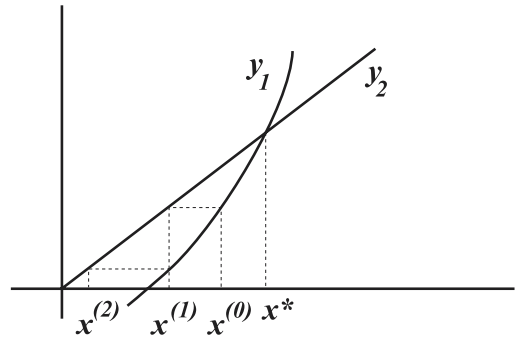


Figura 1.4: $T'(x^*) > 1$.

Método del promotor de convergencia de Wegstein

Una de las peculiaridades de la iteración de punto fijo en una variable es la posibilidad de razonar a través de una representación gráfica del proceso iterativo.

Considerando las dos funciones

$$y_1(x) = T(x) \quad \text{y} \quad y_2(x) = x$$

resolver la ecuación $x = T(x)$ equivale a hallar la intersección de sus gráficas.

En el paso k-ésimo del algoritmo se traza la vertical por $x^{(k)}$ hasta que corte a la curva y_1, y por ese punto se traza la horizontal hasta que corte a la recta y_2. La abscisa del punto de intersección es la nueva iterante $x^{(k+1)}$.

Se puede interpretar $x^{(k+1)}$ como una corrección que hemos hecho a $x^{(k)}$ en el paso k-ésimo mediante el esquema de cálculo

$$x^{(k+1)} = x^{(k)} + \Delta x^{(k)} \tag{1.13}$$

con $\Delta x^{(k)} = T(x^{(k)}) - x^{(k)}$. Por ejemplo, en la Figura 1.3, para pasar de $x^{(1)}$ a $x^{(2)}$ debemos sumar $\Delta x^{(1)} = T(x^{(1)}) - x^{(1)}$ a $x^{(1)}$. La corrección introducida es muy grande y el método diverge [10]. Si multiplicamos la corrección $\Delta x^{(1)}$ por un factor $\alpha \in (0, 1/2)$ podemos forzar la convergencia de la sucesión iterante. Del mismo modo, en el caso $T'(x) > 1$ de la figura 1.4, la corrección $\Delta x^{(k)}$ orienta la búsqueda en la dirección equivocada y el proceso aleja la iterante cada vez más del objetivo. Podríamos forzar la convergencia de la sucesión $(x^{(k)})$ introduciendo un factor $\alpha < 0$ que reoriente la busca en la dirección adecuada.

Si el método ya fuera convergente, se puede **acelerar la convergencia** mediante un factor de relajación α que se busca de modo que sea el mejor posible en cada paso del algoritmo. ¿Qué criterio se debería seguir para asegurar que el factor $\hat{\alpha}$ es el óptimo en el paso k-ésimo del proceso de convergencia? Parece razonable que el mejor α sea el que defina la corrección que nos dé directamente la raíz, $x^{(k+1)} = x^*$. Como x^* no se conoce, $\hat{\alpha}$ se debe estimar.

Refiriéndonos a la Figura 1.5, podemos escribir simultáneamente

$$\tan\theta = \frac{(\alpha - 1)\Delta x^{(k)}}{\alpha \Delta x^{(k)}} = \frac{\alpha - 1}{\alpha} \quad \text{y} \quad \tan\theta = \frac{T(x^*) - T(x^{(k)})}{x^* - x^{(k)}}$$

Aplicando el teorema del valor medio a T, existe un $s \in (x^{(k)}, x^*)$ tan desconocido como $\hat{\alpha}$ o x^*, tal que

$$\frac{T(x^*) - T(x^{(k)})}{x^* - x^{(k)}} = T'(s) \quad \Rightarrow \quad \alpha = \frac{1}{1 - T'(s)} \tag{1.14}$$

$T'(s)$ es también desconocido, pero se puede estimar interpolando

$$T'(s) = \frac{T(x^{(k)}) - T(x^{(k-1)})}{x^{(k)} - x^{(k-1)}} = \frac{T(x^{(k)}) - x^{(k)}}{x^{(k)} - x^{(k-1)}}$$

[10] La pendiente de la tangente a y_1 en el entorno de x^* negativa es menor que -1.

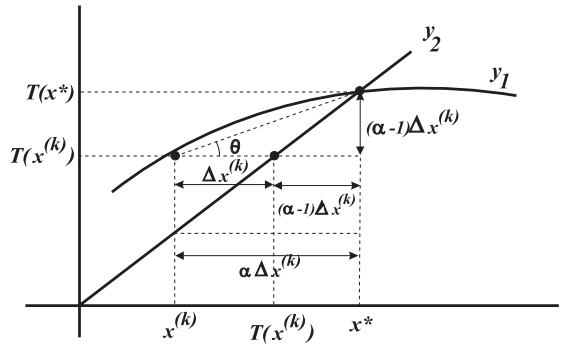


Figura 1.5: Factor óptimo de convergencia.

y se obtiene α sustituyendo $T'(s)$ en (1.14).

Se define así el **método de Wegstein o del promotor de convergencia** que alterna un paso del método de aproximaciones sucesivas con la corrección descrita según el siguiente algoritmo.

- Dado $x^{(0)}$.
  Al comienzo del paso $(k-1)$-ésimo se conoce $x^{(k-1)}$.
- Se actualiza $x^{(k-1)}$ por el método de aproximaciones sucesivas $x^{(k)} = T(x^{(k-1)})$.
- Se calculan

$$A = \frac{T(x^{(k)}) - x^{(k)}}{x^{(k)} - x^{(k-1)}} \qquad \text{y} \qquad \alpha = \frac{1}{1-A}$$

- Se corrige $x^{(k)}$ definiendo

$$x^{(k+1)} = x^{(k)} + \alpha \left(T(x^{(k)}) - x^{(k)} \right)$$

- Se reitera.

Método de relajación

Se puede reinterpretar el método de Wegstein como un proceso de relajación del esquema de aproximaciones sucesivas asociado a la ecuación de punto fijo $x = T(x)$.

Supongamos que T de clase $\mathcal{C}^1$ no cumple las condiciones suficientes del teorema 1.3.1 en el entorno del punto fijo x^*.

Reformulamos $x = T(x)$ en la forma de una combinación lineal convexa de x y de $T(x)$

$$x = h_\alpha(x) = (1-\alpha)x + \alpha T(x) \tag{1.15}$$

donde α es un parámetro de convergencia distinto de cero.

Es claro que si la iteración de punto fijo (1.15) es convergente, converge a x^* independientemente de α.

Investiguemos esta convergencia. Derivando $h_\alpha(x)$

$$h'_\alpha(x) = (1-\alpha) + \alpha T'(x) \tag{1.16}$$

¿Qué criterio deberíamos usar para que (1.15) sea convergente? De acuerdo con los teoremas 1.3.1 y 1.3.2 es suficiente que $|h'_\alpha(x)| < 1$ en un entorno de la raíz, pero la convergencia sería cuadrática si $|h'_\alpha(x^*)| = 0$, en cuyo caso

$$\alpha = \frac{1}{1 - T'(x^*)} \tag{1.17}$$

Como antes, no conocemos x^*, pero podemos estimar $T'(x^*)$ de un modo zafio poniendo

$$T'(x^*) \approx \frac{T(x^{(k)}) - x^{(k)}}{x^{(k)} - x^{(k-1)}}$$

Sustituyendo $T'(x^*)$ en (1.17) obtenemos de nuevo el promotor de convergencia de Wegstein

$$\alpha = \frac{1}{1 - \dfrac{T(x^{(k)}) - x^{(k)}}{x^{(k)} - x^{(k-1)}}}$$

No obstante, el proceso de relajación asociado a esta interpretación se puede usar independientemente de la aplicación del algoritmo de Wegstein ya que existen varios posibles criterios para estimar $T'(x^*)$. Una vez enmarcada la raíz x^* en un intervalo I, la habilidad del modelador puede seleccionar un punto ξ **adecuado**, en el que imponer la condición $|h'_\alpha(\xi)| = 0$. En un entorno sin precisar de ξ, la función h_α correspondiente, satisface las condiciones del teorema 1.3.1 y también sus conclusiones.

Combinando separación de raíces y relajación hemos resuelto varios de los ejercicios propuestos.

Aceleración de la convergencia. Método Δ^2 de Aitken

Cuando un método iterativo estacionario de un solo paso converge linealmente se pueden usar las iterantes $x^{(k)}$ para construir la sucesión

$$y^{(k)} = x^{(k)} - \frac{\left(x^{(k+1)} - x^{(k)}\right)^2}{x^{(k+2)} - 2x^{(k+1)} + x^{(k)}} \tag{1.18}$$

que converge a x^* más rápidamente [11] que la $x^{(k)}$.

Es importante destacar que el proceso de aceleración de la convergencia de Aitken construye la nueva sucesión $y^{(k)}$ usando exclusivamente la información dada por la sucesión original $x^{(k)}$. Necesitamos tres iterantes consecutivas $x^{(k)}, x^{(k+1)}, x^{(k+2)}$ del método iterativo que estemos aplicando para calcular el término $y^{(k)}$ de la nueva sucesión, término que no se usa en la iteración siguiente.

Método de Steffensen

La combinación de un esquema dado de punto fijo con convergencia lineal y del proceso Δ^2 de Aitken define el método de Steffensen.

Con base en el proceso de aceleración de Aitken, y dada una sucesión iterativa linealmente convergente, se define otro esquema iterativo de mayor orden de convergencia.

El valor $y^{(0)}$ que obtuvimos a partir de las tres iterantes consecutivas $x^{(0)}$, $x^{(1)}$ y $x^{(2)}$ por el proceso de Aitken es a menudo mejor aproximación de x^* que $x^{(2)}$, luego es natural tomar $y^{(0)}$ como valor inicial, para dar dos pasos consecutivos del esquema de punto fijo y aplicar después a esas tres iterantes la aceleración de Aitken reiterando a continuación el proceso.

Se obtiene así una sucesión que denotaremos $w^{(k)}$ en la que $w^{(0)} = x^{(0)}$, $w^{(1)}$ será el resultado de acelerar $x^{(0)}$, $x^{(1)}$ y $x^{(2)}$ y

$$w^{(k+1)} = w^{(k)} - \frac{\left[T(w^{(k)}) - w^{(k)}\right]^2}{T\left(T(w^{(k)})\right) - 2T(w^{(k)}) + w^{(k)}} = \overline{T}(w^{(k)}) \tag{1.20}$$

El esquema definido en (1.20) es de nuevo una iteración de punto fijo con función de iteración $\overline{T}$.

Tanto el proceso de aceleración de la convergencia de Aitken como el método de Steffensen se analizan con todo detalle en el problema 1.11, donde se incluyen códigos Matlab de ambos esquemas y se detalla el proceso asociado a cada paso.

Se demuestra que si el orden de convergencia de la sucesión de punto fijo original es uno, el orden de convergencia del método de Steffensen es al menos dos, lo que compensa del coste numérico extra que conlleva su aplicación [26].

[11] Usando el operador en diferencias $\Delta x^{(k)} = x^{(k+1)} - x^{(k)}$ y teniendo en cuenta que $\Delta^2 x^{(k)} = \Delta x^{(k+1)} - \Delta x^{(k)} = x^{(k+2)} - 2x^{(k+1)} + x^{(k)}$ podemos escribir la sucesión $y^{(k)}$ en la forma clásica

$$y^{(k)} = x^{(k)} - \frac{(\Delta x^{(k)})^2}{\Delta^2 x^{(k)}} \tag{1.19}$$

que da nombre a este proceso.

1.3.3. Método de Newton-Raphson. Método de von Mises

Estudiamos ahora los métodos iterativos que se construyen utilizando la estrategia, muy intuitiva, de aproximar en cada paso localmente la función f, por un desarrollo limitado de orden m en el entorno de la última iterante $x^{(k)}$. Se sustituye, por tanto, la función f por un polinomio de grado m que tiene las mismas derivadas $f^{(i)}(x^{(k)}) \quad i = 0, 1, ..., m$ que f en el punto $x^{(k)}$. Una de las raíces del polinomio aproximante se toma como nueva aproximación de la raíz x^* de f.

Sea x^* una raíz de la ecuación $f(x) = 0$ donde $f : I\!R \to I\!R$ es una función suficientemente diferenciable en un entorno V de x^*. El polinomio de Taylor de orden m alrededor de $x^{(k)}$ es:

$$f(x) = f(x^{(k)}) + (x - x^{(k)})f'(x^{(k)}) + \frac{(x - x^{(k)})^2}{2!}f''(x^{(k)}) + \cdots +$$
$$\cdots + \frac{(x - x^{(k)})^m}{m!}f^{(m)}(x^{(k)} + \theta(x - x^{(k)})) \quad \text{con} \quad 0 < \theta < 1$$

Ignorando las potencias de $(x - x^{(k)})$ superiores a la primera (resp: a la segunda) obtenemos la función lineal afín (resp: la función cuadrática)

$$L_k(x) = f(x^{(k)}) + (x - x^{(k)})f'(x^{(k)})$$

(resp: $Q_k(x) = f(x^{(k)}) + (x - x^{(k)})f'(x^{(k)}) + \frac{(x - x^{(k)})^2}{2!}f''(x^{(k)})$) y tomamos la raíz de la ecuación $L_k(x) = 0$ (resp: una de las raíces de la ecuación de segundo grado $Q_k(x) = 0$) como nueva estimación de la raíz x^* en el método iterativo correspondiente.

Este argumento produce los métodos iterativos de un paso $x^{(k+1)} = T(x^{(k)})$ con

$$T(x) = x - \frac{f(x)}{f'(x)} \tag{1.21}$$

(resp: con

$$T(x) = x - \frac{f'(x^{(k)}) \pm \sqrt{f'(x)^2 - 2f(x)f''(x)}}{f''(x)}) \tag{1.22}$$

El primero de ellos es el clásico método de Newton-Raphson y el segundo es una de sus extensiones naturales el método de Halley. Ambos métodos ya utilizados en el siglo XVII [12] [13].

[12] Joseph Raphson (1648, Middlesex, Inglaterra-1715). No hay mucha información sobre su vida. Se licenció en la Universidad de Cambridge en 1692, aunque entró en la Royal Society en 1691, un año antes de su licenciatura, lo cual era muy raro. Su elección para la Royal Society se basó en su libro ***Analysis aequationum universalis***, publicado en 1690, que contiene una discusión del método de Newton-Raphson para aproximar las raíces de una ecuación, atribuyendo su autoría a Newton.

Newton lo publicó en el ***Principia Mathematica*** mucho más tarde, como una herramienta para resolver una ecuación de Kepler pero ya en 1669, en su trabajo sobre ecuaciones infinitas, había discutido la cúbica $x^3 - 2x - 5 = 0$ [10] aplicando el método a la aproximación de la raíz que está entre 2 y 3.

No se sabe mucho de la relación entre Newton y Raphson, aunque parece que era importante. Se cree que Raphson era una de las pocas personas a las que Newton mostraba sus artículos, y participó en algunas de las disputas entre Newton y Leibniz, pero ésta es otra historia.

[13] Isaac Newton (1642-1727). Virtual creador de la Física moderna de influencia decisiva en el desarrollo de la humanidad. Una de las más relevantes inteligencias de todos los tiempos. Nació en Woolsthorpe Manor, una granja en Lincolnshire, al oeste de Inglaterra, el día de Navidad de 1642. Su padre había muerto dos meses antes y su madre pronto se volvió a casar dejando a Newton niño en Woolsthorpe Manor al cuidado de sus padres. Su infancia fue solitaria e influyó en su carácter introvertido y en la tendencia al secretismo que luego se mostró a lo largo de su vida, especialmente en la resistencia a publicar sus monumentales descubrimientos que guardó para sí mismo durante larguísimos periodos de tiempo.

En 1661 Newton dejó Lincolnshire para seguir sus estudios en Cambridge. El periodo 1661-1665 de sus estudios de grado fue irrelevante pero en 1665 regresó a Woolsthorpe Manor huyendo de la peste que había obligado a cerrar las universidades. Allí, en la soledad del campo, se produjo un arrebato de creatividad incomparable, de dos años de duración, entre los 22 y los 24 años, en el que descubrió el cálculo diferencial, la composición de la luz blanca y la ley de gravitación universal.

Ya anciano se refirió a este periodo milagroso de su juventud en los siguientes términos: "In the two plague years I was in the prime of my age for invention and minded Mathematics and Philosophy more than at any time since".

Fue un soltero de gustos simples, muy sensible a las críticas que le producían amargos resentimientos y enfados. Como reacción a las críticas de Robert Hooke a finales de 1670 escribió, en otro periodo de 18 meses de concentración increíble, su máximo

También se puede considerar el método de Newton como un caso particular del método de aproximaciones sucesivas.

Sea $g \in \mathcal{C}^1([a,b])$ que no se anula en $[a,b]$. Consideremos la ecuación

$$g(x)f(x) = 0$$

que tiene las mismas raíces que la dada, con la que formamos la ecuación de punto fijo

$$x = x + g(x)f(x) = T(x)$$

Queremos determinar g obligando a que $T'(x^*) = 0$. De

$$T'(x) = 1 + g'(x)f(x) + g(x)f'(x)$$

y suponiendo que x^* es raíz simple de la ecuación $f(x) = 0$, luego que $f(x^*) = 0$ y $f'(x^*) \neq 0$, tenemos:

$$T'(x^*) = 1 + g(x^*)f'(x^*) = 0 \quad \Rightarrow \quad g(x^*) = -\frac{1}{f'(x^*)}$$

condición que se satisface si

$$g(x) = -\frac{1}{f'(x)}$$

en un entorno de x^*. Ello exige que f' sea distinto de cero en un entorno V de x^* en el que f debe ser de clase $\mathcal{C}^2$.

Se obtiene así de nuevo el método de Newton

$$x^{(k+1)} = x^{(k)} - \frac{f(x^{(k)})}{f'(x^{(k)})} \tag{1.23}$$

donde la función de iteración es (1.21).

Reglas prácticas para aplicar el método de Newton

Para aplicar el metodo de Newton en $[a,b]$ se deben observar las reglas siguientes que se tienen que verificar en cada paso

a) $f(a)f(b) < 0$ (Se asegura la existencia de una raíz x^* en $[a,b]$).

b) f'' debe tener signo constante en $[a,b]$ (f tiene la misma convexidad/concavidad en $[a,b]$).

c) Se aplica al extremo del intervalo en el que f y f'' tienen el mismo signo.

Sólo se pueden dar los cuatro casos representados en las gráficas adjuntas (1.6), (1.7), (1.8) y (1.9).

Si se cumplen las condiciones a) y b), f es estrictamente monótona en $[a,b]$, ya que toma distintos signos en los extremos a y b, $f'(x) \neq 0$ y no cambia la concavidad en el intervalo.

Modificaciones del método de Newton

El mayor inconveniente práctico del método de Newton es que exige conocer y evaluar f'. Si la fórmula de f es muy complicada, el cálculo de f' puede tener un coste muy alto e incluso podría ser imposible obtener f' si por ejemplo no se conoce una expresión matemática de f.

---

trabajo, el ***Principia Mathematica***, uno de los logros supremos de la mente humana.

En 1696 dejó Cambridge para ser "Warden of the Royal Mint" (encargado de la Casa Real de la Moneda) y en el resto de su larga vida llegó a tomarle gusto a su posición de referente de la ciencia en Europa sobre todo a partir del final de la guerra de sucesión española, en 1714, cuando la paz en Europa permitió la transmisión definitiva de sus teorías científicas. Estos cambios en sus intereses y en su entorno social y físico no disminuyeron sus capacidades intelectuales. De regreso de un agotador día de trabajo en la Casa de la Moneda supo del reto de Johann Bernoulli (representante del cálculo de Leibniz) a los mejores matemáticos del mundo a resolver el problema de la braquistócrona y lo resolvió esa misma noche antes de acostarse.

Su genio se mostró en otros campos no científicos. Como muestra, uno de sus estudios en Teología fue la investigación de la forma y dimensiones del templo de Salomón en Jerusalén a partir de las descripciones de la Biblia.

A su muerte fue enterrado con gran pompa en la abadía de Westminster.

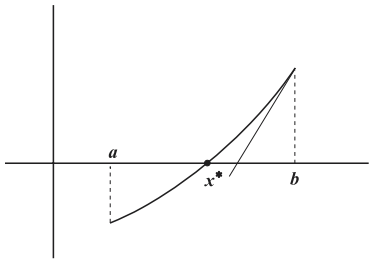


Figura 1.6: $f(b) > 0$ **y** $f''(b) > 0$**.**

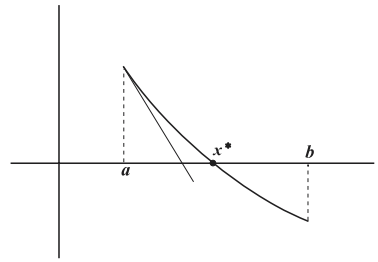


Figura 1.7: $f(a) > 0$ **y** $f''(a) > 0$**.**

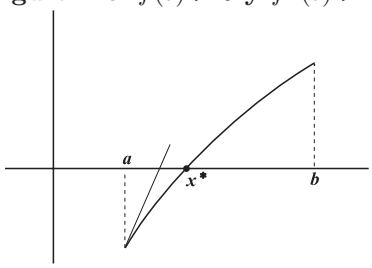


Figura 1.8: $f(a) < 0$ **y** $f''(a) < 0$**.**

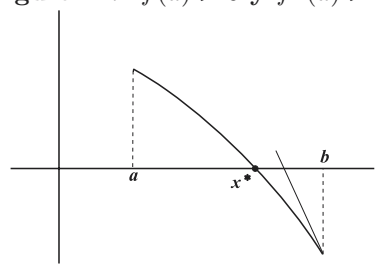


Figura 1.9: $f(b) < 0$ **y** $f''(b) < 0$**.**

Una modificación que evita la evaluación en cada paso de f' es el método de von Mises de convergencia lineal, en el que se evalúa f' sólo una vez para un buen valor inicial $x^{(0)}$, manteniéndose fija en las iteraciones siguientes.

En el caso de ecuaciones muy complicadas, puede interesar cambiar la dirección de la función aproximante para acelerar la convergencia del método de Newton, es decir, considerar en vez de la tangente a la gráfica de f en el punto $(x^{(k)}, f(x^{(k)}))$ una recta que pase por ese punto, que sea próxima a la tangente, pero que corte el eje Ox en un punto x^+ tal que $x^* < x^+ < x^{(k+1)}$ de modo que x^+ sea una estimación de x^* mejor que $x^{(k+1)}$. Dicha recta tendrá una ecuación

$$y - f(x^{(k)}) = \tau f'(x^{(k)})(x - x^{(k)})$$

donde τ es un parámetro que deberá ser próximo a 1.

Este esquema produce el método iterativo

$$x^{(k+1)} = x^{(k)} - \frac{f(x^{(k)})}{\tau f'(x^{(k)})} \tag{1.24}$$

Se simplifica la notación poniendo $\tau = \dfrac{1}{\omega}$, donde ω es el **factor de superrelajación**.

Se toma $0 < \omega < 2$, lo que equivale a $\dfrac{1}{2} < \tau < 1$ quedando definitivamente

$$x^{(k+1)} = x^{(k)} - \omega \frac{f(x^{(k)})}{f'(x^{(k)})} \tag{1.25}$$

1.3.4. Instrucciones de parada de las iteraciones

Sabemos que en general no podemos obtener la solución de la ecuación $f(x) = 0$ en un número finito de pasos de cualquiera de los métodos iterativos analizados aún cuando sean convergentes, por lo que debemos diseñar tests que paren las operaciones cuando hayamos alcanzado los objetivos previstos.

En la práctica los test de parada más usados son:

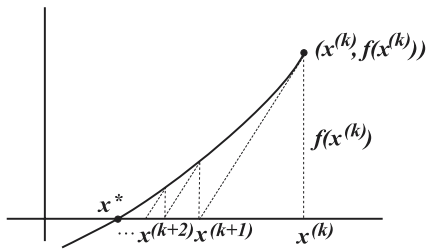


Figura 1.10: Método de von Mises.

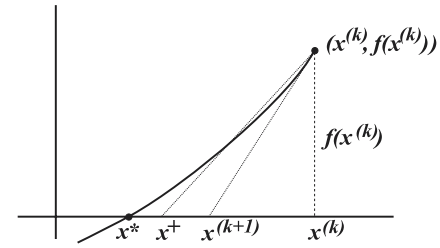


Figura 1.11: Modificación de la dirección de avance en el método de Newton.

(T1) $f(x^{(k)})$ es casi nula.

Se fija un número real arbitrariamente pequeño ϵ_1 y se paran las operaciones con la instrucción de parada

$$\|f(x^{(k)})\| \leq \epsilon_1 \tag{1.26}$$

(T2) La mejoría que vamos a conseguir corrigiendo $x^{(k)}$ mediante una iteración no justifica el esfuerzo de cálculo suplementario.

Se mide lo anterior fijando un número real arbitrariamente pequeño y se para el proceso utilizando bien el concepto de error absoluto,

$$|x^{(k)} - x^{(k-1)}| \leq \epsilon_2 \tag{1.27}$$

o bien el de error relativo

$$\frac{|x^{(k)} - x^{(k-1)}|}{|x^{(k)}|} \leq \epsilon_3 \tag{1.28}$$

que presenta varias variantes

$$\frac{2\left|x^{(k-1)} - x^{(k)}\right|}{|x^{(k-1)}| + |x^{(k)}|} \leq \epsilon_3 \tag{1.29}$$

e incluso

$$\frac{\left|x^{(k-1)} - x^{(k)}\right|}{|x^{(k)}| + \epsilon_3} \leq \epsilon_3 \tag{1.30}$$

Si $\hat{x}$ es una aproximación de x, el error relativo se puede traducir en una conclusión relativa al número de dígitos significativos correctos en x. Si, por ejemplo,

$$\frac{|\hat{\mathbf{x}} - \mathbf{x}|}{|\mathbf{x}|} \sim 10^{-p}$$

entonces $\hat{x}$ tendrá aproximadamente p cifras significativas correctas.

(T3) Se sobrepasa un número razonable de pasos del proceso iterativo.

Si la convergencia no se obtiene en un número de iteraciones kmax previamente fijado, se para el proceso antes que se alcance la precisión fijada ϵ_1 ϵ_2 o ϵ_3 si $k >$ kmax

1.4. Sistemas de ecuaciones no lineales de varias variables reales

Nos ocuparemos ahora del problema del cálculo numérico de los ceros de funciones vectoriales de varias variables reales.

La función $f : \mathbb{R}^n \to \mathbb{R}^m$ viene definida por sus m funciones componentes $f_i : \mathbb{R}^n \to \mathbb{R}$ con $i = 1, ..., m$ y la ecuación $f(\mathbf{x}) = \mathbf{0}_m$ se expresa

$$\begin{cases} f_1(x_1, x_2, ..., x_n) = 0 \\ f_2(x_1, x_2, ..., x_n) = 0 \\ \cdots\cdots\cdots\cdots\cdots \\ f_m(x_1, x_2, ..., x_n) = 0 \end{cases} \tag{1.31}$$

un sistema de m ecuaciones no lineales.

El salto de una a varias variables conlleva la introducción de nuevos objetos que generalicen los habituales en $\mathbb{R}$. El tiempo ha hecho muy intuitivo el concepto de norma como distancia compatible con las operaciones de la estructura de espacio vectorial, que generaliza el valor absoluto de $\mathbb{R}$, y con el que es fácil expresar el análisis en varias variables manteniendo una semejanza casi total con el caso real. La existencia de varias normas distintas que definen distancias distintas pero que son equivalentes para estudiar propiedades topológicas enriquece el marco de trabajo.

La dificultad es mayor. Las herramientas parecidas. Los métodos generalizan los que hemos estudiado en una variable.

1.4.1. Método de aproximaciones sucesivas

El marco natural para desarrollar el estudio general de la iteración de punto fijo es el de los espacios de Banach [14].

Sea $(E, \|.\|)$ un espacio vectorial normado.

Teorema 1.4.1 *Si $(\mathbf{x}^{(k)})$ es la sucesión de E definida por la recurrencia (1.4) con $\mathbf{x}^{(0)}$ arbitrario y si*

$$\mathbf{x}^{(k)} \xrightarrow[k\to\infty]{} \mathbf{x}^* \in E$$

con T continua en $\mathbf{x}^*$, entonces $\mathbf{x}^*$ es un punto fijo de T.

Definición 1.4.1 *Un operador T en E es L-lipchiciano en una parte cerrada A de E ($T(A) \subset A$), con $L > 0$, si*

$$\|T(\mathbf{x}) - T(\mathbf{x}')\| \leq L \cdot \|\mathbf{x} - \mathbf{x}'\|$$

*para todo $\mathbf{x}, \mathbf{x}' \in A$. L es una **constante de Lipschitz** de T en A.*

Definición 1.4.2 *Un operador T en E es una contracción en la bola cerrada $B_c(\mathbf{z}; r)$ de centro en $\mathbf{z} \in \mathbb{R}^n$ y radio $r > 0$ si es L-lipchiciano con $L \in (0, 1)$ (L es el factor de contracción en $B_c(\mathbf{z}; r)$) luego si,*

$$(\exists\, L \;:\; 0 < L < 1) : (\forall \mathbf{x}, \mathbf{x}' \in B_c(\mathbf{z}; r)) \qquad \|T(\mathbf{x}) - T(\mathbf{x}')\| \leq L \cdot \|\mathbf{x} - \mathbf{x}'\| \tag{1.32}$$

Teorema 1.4.2 ***(Teorema del punto fijo de Banach)***

Sea $(E, \|.\|)$ un espacio de Banach y $T : E \to E$ una aplicación contractiva en $B_c(\mathbf{x}^{(0)}; r)$ con

$$r \geq r_0 = \frac{1}{1 - L}\|\mathbf{x}^{(0)} - T(\mathbf{x}^{(0)})\|$$

- *Entonces existe un único punto fijo $\mathbf{x}^*$ de T en $B_c(\mathbf{x}^{(0)}; r)$.*
- *La sucesión $\mathbf{x}^{(n+1)} = T(\mathbf{x}^{(n)})$ converge a $\mathbf{x}^*$.*

[14] Ver la nota al pie 5.

Corolario 1.4.1 *Se tiene la siguiente estimación del error*

$$\|\mathbf{x}^{(k)} - \mathbf{x}^*\| \leq \frac{L^{k-m}}{1-L}\|\mathbf{x}^{(m+1)} - \mathbf{x}^{(m)}\| \tag{1.33}$$

para $0 \leq m \leq k$ *con* $k = 1, 2, ...$ *de la que se obtienen las estimaciones*

- *"a priori"*

$$\|\mathbf{x}^{(k)} - \mathbf{x}^*\| \leq L^k r_0 = \frac{L^k}{1-L}\|\mathbf{x}^{(1)} - T(\mathbf{x}^{(0)})\| \tag{1.34}$$

- *"a posteriori"*

$$\|\mathbf{x}^{(k)} - \mathbf{x}^*\| \leq \frac{L}{1-L}\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\| \tag{1.35}$$

Corolario 1.4.2 *El número* k *de iteraciones necesarias para que el error* $\|\mathbf{x}^{(k)} - \mathbf{x}^*\|$ *sea menor que un cierto número* ϵ *satisface la desigualdad*

$$k \geq \frac{\ln\left(\frac{\epsilon(1-L)}{\|\mathbf{x}^{(1)} - \mathbf{x}^{(0)}\|}\right)}{\ln L} \tag{1.36}$$

Si suponemos que T es $\mathcal{C}^1$ en A, luego que tiene derivadas parciales continuas en A, se puede generalizar el teorema 1.3.1 que enunciamos para el caso de una variable real dando de una forma cómoda criterios suficientes para que se cumpla la condición (1.32).

Representando por $\mathbf{T}(\mathbf{z})$ la matriz jacobiana de T en el punto $\mathbf{z} \in A$ ($\mathbf{T}(\mathbf{z}) = M(dT(\mathbf{z}); B_n) \in M_n(\mathbb{R})$) [15] se tiene el criterio suficiente siguiente:

Teorema 1.4.3 *Si para cualquier norma matricial* $\|,\|$ *se tiene* $\|\mathbf{T}(\mathbf{z})\| \leq L < 1$ *para todo* $\mathbf{z} \in A$ *entonces se satisface (1.32) para una norma vectorial compatible.*

Se prueba [16] que T cumple (1.32) ssi $\rho\mathbf{T}(\mathbf{z})$ el radio espectral de la matriz jacobiana es estrictamente menor que 1 ($\forall \mathbf{z} \in A$).

1.4.2. Método de Newton y modificaciones. Método de Broyden

Método de Newton

La misma idea que utilizamos en el caso de una variable se generaliza con facilidad al caso de un operador diferenciable $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^m$ siempre que se tenga la técnica suficiente.

Sean $\mathbf{x}^*$ una solución de la ecuación $\mathbf{f}(\mathbf{x}) = 0$ y $\mathbf{x}^{(0)}$ una estimación inicial de $\mathbf{x}^*$ que suponemos suficientemente próxima a $\mathbf{x}^*$. Desarrollando $\mathbf{f}$ en serie en el entorno de $\mathbf{x}^{(0)}$ se tiene,

$$0 = \mathbf{f}(\mathbf{x}^*) = \mathbf{f}(\mathbf{x}^{(0)}) + \mathbf{F}^{(0)}(\mathbf{x}^* - \mathbf{x}^{(0)}) + \eta(\mathbf{x}^{(0)}, \mathbf{x}^*)$$

con $\eta(\mathbf{x}^{(0)}, \mathbf{x}^*)$ pequeño y donde hemos denotado $d\mathbf{f} = \mathbf{F}$ y $d\mathbf{f}(\mathbf{x}^{(0)}) = \mathbf{F}^{(0)}$.

Despreciando $\eta(\mathbf{x}^{(0)}, \mathbf{x}^*)$, aproximamos $\mathbf{f}$ en el entorno de $\mathbf{x}^{(0)}$ con la aplicación lineal afín

$$L_0(\mathbf{x}) = \mathbf{f}(\mathbf{x}^{(0)}) + \mathbf{F}^{(0)}\left(\mathbf{x} - \mathbf{x}^{(0)}\right)$$

[15] Si $\mathbf{f} = f_1 \times \cdots \times f_n : \mathbb{R}^n \to \mathbb{R}^n$ con $\mathbf{x} = (x_1, \cdots, x_n)$ tendremos con las notaciones habituales

$$\mathbf{F}^{(k)} = \left(\frac{\partial f_i}{\partial x_j}(\mathbf{x}^{(k)})\right)_{i,j=1,\ldots,n}$$

matriz jacobiana de $\mathbf{f}$ en el punto $\mathbf{x}^{(k)}$ cuya fila i-ésima es

$$\mathbf{F}_i^{(k)} = \left(\frac{\partial f_i}{\partial x_1}, \cdots, \frac{\partial f_i}{\partial x_n}\right) = \text{grad} f_i(\mathbf{x}^{(k)})$$

[16] Ver la sección (2.1.3) del Capítulo 2.

que comparte con $\mathbf{f}$ la diferencial primera en el punto $\mathbf{x}^{(0)}$ y cuya gráfica "pasa" por el punto $(\mathbf{x}^{(0)}, \mathbf{f}(\mathbf{x}^{(0)}))$.

Tomamos la raíz de la ecuación lineal $L_0(\mathbf{x}) = \mathbf{0}$ como nueva aproximación $\mathbf{x}^{(1)}$ de $\mathbf{x}^*$.

El proceso descrito es la base del método iterativo de Newton.

Se construye para cada $k \geq 0$ la aproximación lineal afín

$$L_k(\mathbf{x}) = \mathbf{f}(\mathbf{x}^{(k)}) + \mathbf{F}^{(k)} \left(\mathbf{x} - \mathbf{x}^{(k)}\right) \tag{1.37}$$

a $\mathbf{f}$ en $\mathbf{x}^{(k)}$ y se toma la solución de la ecuación lineal $L_k(\mathbf{x}) = \mathbf{0}$ como nueva iterante $\mathbf{x}^{(k+1)}$, de modo que [17]

$$\mathbf{f}(\mathbf{x}^{(k)}) + \mathbf{F}^{(k)} \left(\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\right) = 0 \tag{1.38}$$

Llamando $\Delta\mathbf{x}^{(k)}$ al vector $\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}$, la nueva iterante $\mathbf{x}^{(k+1)}$ es suma de la anterior $\mathbf{x}^{(k)}$ y de la corrección $\Delta\mathbf{x}^{(k)}$ que es solución de la ecuación lineal (1.38).

Esta reflexión permite reescribir el método de Newton en la forma

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \Delta\mathbf{x}^{(k)} \tag{1.39}$$

Si $\mathbf{F}^{(k)}$ posee inversa, lo que no sucede en general, $\Delta\mathbf{x}^{(k)} = -\left(\mathbf{F}^{(k)}\right)^{-1} \mathbf{f}(\mathbf{x}^{(k)})$ y

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \left(\mathbf{F}^{(k)}\right)^{-1} \mathbf{f}(\mathbf{x}^{(k)}) \qquad k \geq 0 \tag{1.40}$$

Fórmula que se reduce, cuando $n = 1$, a (1.23) [18].

Convergencia del método de Newton

Teorema 1.4.4 *Sea $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^n$ y supongamos que $\mathbf{f}(\mathbf{x})$, $[\mathbf{F}(\mathbf{x})]^{-1}$ y $[Hess\,(\mathbf{f})(\mathbf{x})]$ están acotadas en una bola, $B(\mathbf{x}^*; r)$, con $\mathbf{f}(\mathbf{x}^*) = 0$, entonces el método de Newton tiene una convergencia de orden 2 en un entorno de $\mathbf{x}^*$.*

Elegido un estimador inicial $\mathbf{x}^{(0)}$, si el método de Newton converge lo suele hacer tan rápido que se nota enseguida cuando no converge, en cuyo caso es necesario cambiar el estimador inicial.

Tomando en el espacio de las matrices cuadradas de orden n la norma del máximo tenemos,

$$\|\mathbf{F}(\mathbf{x})\| = \max_i \sum_{j=1}^{n} \left|\frac{\partial f_i}{\partial x_j}\right| \qquad \text{y} \qquad \|\text{Hess}\,(\mathbf{f})(\mathbf{x})\| = \max_i \sum_{j,k=1}^{n} \left|\frac{\partial^2 f_i}{\partial x_j \partial x_k}\right|$$

Expresiones que debemos ser capaces de mayorar cerca de la solución para poden usar el teorema (1.4.4).

Modificaciones del método de Newton. Métodos de von Mises y de Broyden

La desventaja más seria del método de Newton es el tiempo que requiere evaluar $\mathbf{F}^{(k)}$ en cada paso.

Una de las estrategias para evitar la evaluación sucesiva de las n^2 derivadas parciales de $\mathbf{f}$, es usar en cada paso una aproximación lineal afín más fácil de evaluar que L_k.

En el método de von Mises se utiliza la misma en todos los pasos.

$$\ell(\mathbf{x}) = \mathbf{f}(\mathbf{x}^k) + \mathbf{D}\left(\mathbf{x} - \mathbf{x}^k\right) \qquad k \geq 0 \tag{1.41}$$

[17] Para cada función componente f_i de $\mathbf{f}$

$$0 = f_i(\mathbf{x}^*) = f_i(\mathbf{x}^{(k)}) + \text{grad} f_i(\mathbf{x}^{(k)}) \cdot (\mathbf{x}^* - \mathbf{x}^{(k)}) + (\mathbf{x}^* - \mathbf{x}^{(k)})^T \text{Hess}(f_i)(\mathbf{x}^{(k)})(\mathbf{x}^* - \mathbf{x}^{(k)}) + \cdots \qquad i = 1, ..., n$$

con $\text{Hess}(f_i)(\mathbf{x}^{(k)}) = \left(\dfrac{\partial^2 f_i}{\partial x_j \partial x_k}(\mathbf{x}^{(k)})\right)_{j,k=1,\ldots,n}$.

Linealizando,

$$\text{grad} f_i(\mathbf{x}^{(k)}) \cdot (\mathbf{x}^* - \mathbf{x}^{(k)}) = -f_i(\mathbf{x}^{(k)}) \qquad i = 1, ..., n$$

que escrita matricialmente reproduce (1.38).

[18] En la práctica, aun cuando $\mathbf{F}^{(k)}$ posea inversa, si $n > 2$, no se calcula $\left(\mathbf{F}^{(k)}\right)^{-1}$. Es menos costoso numéricamente utilizar (1.39) resolviendo el sistema lineal (1.38) para hallar la corrección.

y se elige $\mathbf{D} = \mathbf{F}^{(0)}$, para un estimador inicial $\mathbf{x}^{(0)}$ bien elegido [19].

Con este método se disminuye, a menudo considerablemente, el costo numérico de cada paso del método de Newton. Lógicamente la velocidad de convergencia debe resentirse y de hecho este método es de orden 1.

En el método de Broyden se define en cada paso una aproximación lineal afín

$$\ell_k(\mathbf{x}) = \mathbf{f}(\mathbf{x}^{(k)}) + \mathbf{D}^{(k)} \left(\mathbf{x} - \mathbf{x}^{(k)}\right) \tag{1.42}$$

donde $\mathbf{D}^{(k)}$ es en principio una matriz de M_n distinta de $\mathbf{F}^{(k)}$.

Ya que $\ell_k(\mathbf{x}^{(k)}) = \mathbf{f}(\mathbf{x}^{(k)})$, la aproximación lineal afín pasa por el punto $(\mathbf{x}^{(k)}, \mathbf{f}(\mathbf{x}^{(k)}))$ y de modo similar al método de Newton se llega a $\mathbf{x}^{(k+1)}$ resolviendo el sistema lineal $\ell_k(\mathbf{x}) = \mathbf{0}$.

$$\mathbf{D}^{(k)} \left(\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\right) = -\mathbf{f}(\mathbf{x}^{(k)}) \qquad k \geq 0 \tag{1.43}$$

Si $\mathbf{D}^{(k)}$ posee inversa

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - {\mathbf{D}^{(k)}}^{-1} \mathbf{f}(\mathbf{x}^{(k)}) \qquad k \geq 0 \tag{1.44}$$

La elección de $\mathbf{D}^{(0)}$ y su posterior actualización paso a paso se rige por el criterio de proximidad entre las aproximaciones lineales $\ell_k(\mathbf{x})$ y $L_k(\mathbf{x})$ eligiendo además ℓ_k la de cálculo más fácil.

Se llega al algoritmo siguiente

1. Se elige $\mathbf{D}^{(0)}$, por ejemplo, con una evaluación de $\mathbf{F}$ en $\mathbf{x}^{(0)}$ de modo que $\mathbf{F}^{(0)} = \mathbf{D}^{(0)}$.
2. Se resuelve el sistema lineal
$$\mathbf{D}^{(k)} \left(\mathbf{x} - \mathbf{x}^{(k)}\right) = -\mathbf{f}(\mathbf{x}^{(k)})$$
en $\mathbf{x}$ y se pone $\mathbf{x}^{(k+1)} = \mathbf{x}$.
3. Se calcula $\mathbf{f}(\mathbf{x}^{(k+1)})$ y se llama $\mathbf{s}^{(k)} = \mathbf{x}^{(k+1)} - \mathbf{x}(k)$
4. Se actualiza
$$\mathbf{D}^{(k+1)} = \mathbf{D}^{(k)} + \frac{\mathbf{f}(\mathbf{x}^{(k+1)}) \otimes \mathbf{s}^{(k)}}{\mathbf{s}^{(k)} \cdot \mathbf{s}^{(k)}} = \mathbf{D}^{(k)} + \frac{\mathbf{f}(\mathbf{x}^{(k+1)}) \otimes \mathbf{s}^{(k)}}{\|\mathbf{s}^{(k)}\|^2} \tag{1.45}$$
5. Se vuelve al punto 2 y se continúa iterando [20].

Ver en [8] un estudio exhaustivo de los métodos Quasi-Newton que evitan muchas de las desventajas del método de Newton.

Hemos usado este método en el problema 1.9 comparándolo con el método de Newton.

Se utilizan en el caso de varias variables los mismos tests de parada que en una variable, sustituyendo el valor absoluto por la norma que se seleccione en $\mathbb{R}^n$.

[19] Se evalúa $\mathbf{F}$ solamente una vez y se toma $\ell = L_0$ para todo k.

[20] Si $\mathbf{a}$ y $\mathbf{b}$ son dos vectores de $\mathbb{R}^n$, $\mathbf{a} \otimes \mathbf{b}$ representa la matriz cuadrada de orden n de rango 1 cuyo elemento intersección de la fila i columna j es $a_i b_j$.

PROBLEMAS

PROBLEMA 1.1 *Formulación de punto fijo para una ecuación de segundo grado.*

Dada la función $f(x) = x^2 - x - 2$.

1. ¿Converge la fórmula
$$x^{(n+1)} = \left(x^{(n)}\right)^2 - 2 \tag{1.46}$$
a una raíz de $f(x) = 0$?

2. Escribir una fórmula de Newton-Raphson que resuelva el problema del cálculo de los ceros de f.

Solución:

1. Llamemos g a la función $g(x) = x^2 - 2$. La sucesión (1.46) es la del método de aproximaciones sucesivas asociada a la ecuación de punto fijo $x = g(x)$. Una de las condiciones de aplicación del teorema 1.3.1

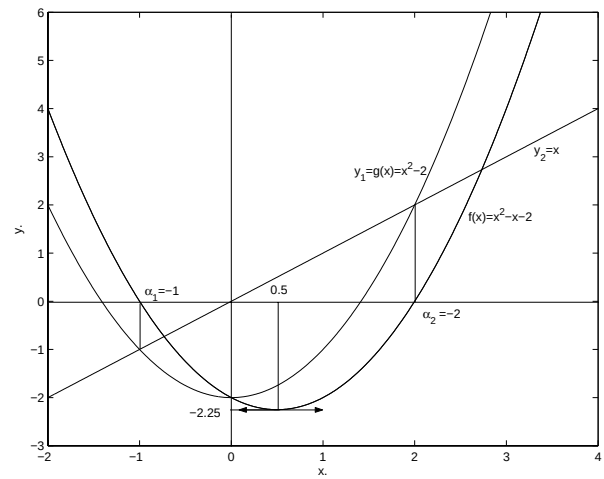


Figura 1.12: Gráficas de todas las funciones y raíces que intervienen en el problema 1.1.

en un intervalo $I = (a, b)$ exige que $g(I) \subset I$, luego que
$$\begin{aligned} a < g(a) = a^2 - 2 \quad &\Rightarrow \quad 0 < (a-2)(a+1) \quad \Rightarrow \quad a > 2 \quad \text{y} \quad a < -1 \\ g(b) = b^2 - 2 < b \quad &\Rightarrow \quad -1 < b < 2 \end{aligned}$$
de donde $g(a) < b_{\max} = 2 \quad \Rightarrow \quad |a| < \sqrt{2}$ y $g(b) > a_{\min} = -\sqrt{2} \quad \Rightarrow \quad |b| > \sqrt{2 - \sqrt{2}}$.

Representando gráficamente esos resultados se prueba que no existe ningún intervalo (a, b) tal que $g(a, b) \subset (a, b)$ y que la iteración de punto fijo (1.46) no es convergente.

2. El esquema de Newton-Raphson es
$$x^{(n+1)} = x^{(n)} - \frac{\left(\left(x^{(n)}\right)^2 - x^{(n)} - 2\right)}{2x^{(n)} - 1} = \frac{\left(x^{(n)}\right)^2 + 2}{2x^{(n)} - 1}$$
Tomando como estimador inicial $x^{(0)} = 1$, se obtiene sucesivamente $x^{(1)} = 3$, $x^{(2)} = 2.2$, $x^{(3)} = 2.0110$, $x^{(4)} = 2.0000$.

Tomando como estimador inicial $x^{(0)} = 0$, se obtiene $x^{(1)} = -2$, $x^{(2)} = -1.2$, $x^{(3)} = -1.0110$, $x^{(4)} = -1.0000$.

PROBLEMA 1.2 *Ecuación no lineal de una variable. Método de Newton Raphson.*

Se considera la ecuación no lineal de una variable

$$\ln(x+1) - x + 1 = 0$$

1. Representar gráficamente la función $f(x) = \ln(x+1) - x + 1$.

2. Localizar y separar los ceros de f, definiendo intervalos en los que sea recomendable la aplicación del método de Newton-Raphson.

   Aproximar dichas raíces por el método de Newton con un error menor que 10^{-5}.

Solución:

1. El dominio de definición de f es $(-1, \infty)$. Cuando $x \to -1$ por la derecha la función tiende a $-\infty$. La gráfica (1.13) muestra los dos ceros de f, $\alpha_1 < 0$ y $\alpha_2 > 0$. Ya que

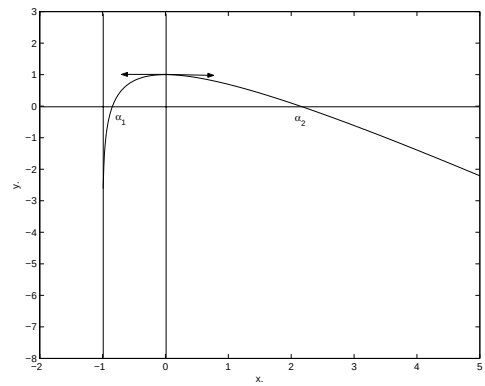


Figura 1.13: Representación gráfica de $x \to \ln(x+1) - x + 1$.

$$f'(x) = -\frac{x}{x+1}$$

   la tangente es horizontal en $x = 0$ y la función es estrictamente creciente en el intervalo abierto $(-1, 0)$ y estrictamente decreciente en $(0, \infty)$. Además

$$f''(x) = -\frac{1}{(x+1)^2} < 0 \qquad \forall x \in (-1, \infty)$$

2. Para obtener por ejemplo α_1 por el método de Newton buscamos un intervalo de acotación $[a, b]$ que satisfaga las reglas prácticas de la sección (1.3.3). Con todos los datos en la mano seleccionamos el intervalo $[-0.9, -0.7]$ para aplicar el esquema de Newton

$$x_{n+1} = \frac{1 + (1 + x_n)\ln(1 + x_n)}{x_n} \tag{1.47}$$

   y tomamos como estimador inicial $x_0 = -0.9$ el extremo del intervalo en el que f y f'' tienen el mismo signo negativo.

   El pequeño código Matlab que hemos escrito llega a la solución $\alpha_1 = -0.841406$ en 5 iteraciones con un error menor que 10^{-5}.

$x^{(0)}$	$x^{(1)}$	$x^{(2)}$	$x^{(3)}$	$x^{(4)}$
-0.900000	-0.855268	-0.842136	-0.841408	-0.841406

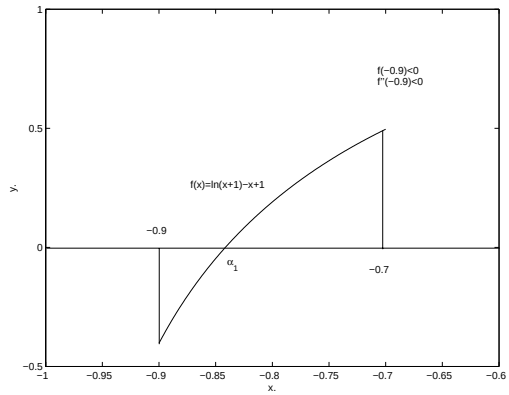


Figura 1.14: Aproximación de la raíz α_1.

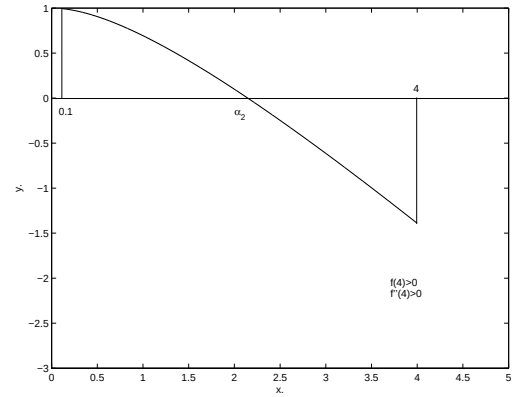


Figura 1.15: Aproximación de la raíz α_2.

Para la raíz positiva α_2, tomamos en principio el intervalo $[0.1, 4]$ y como estimador inicial el extremo superior del intervalo $x^{(0)} = 4$ en el que como antes f y f'' tienen el mismo signo.

El código alcanza la solución $\alpha_2 = 2.146193$ en 4 iteraciones con un error menor que 10^{-5}.

$x^{(0)}$	$x^{(1)}$	$x^{(2)}$	$x^{(3)}$
4.000000	2.261797	2.147121	2.146193

PROBLEMA 1.3 *Ecuación no lineal de una variable. Método de aproximaciones sucesivas.*

Se considera de nuevo la ecuación no lineal de una variable

$$\ln(x+1) - x + 1 = 0 \tag{1.48}$$

1. Utilizar el método de aproximaciones sucesivas para aproximar las raíces de (1.48). Se elegirán las correspondientes iteraciones de punto fijo de modo que se cumplan las condiciones de aplicación del teorema del punto fijo.
2. Hacer una predicción del número de términos que se deberán calcular para aproximar las raíces con un error menor que 10^{-5}.

Solución:

1. Reformulamos la ecuación (1.48) en la forma

$$x = g(x) = \ln(x+1) + 1 \tag{1.49}$$

Para estudiar la raíz $\alpha_1 < 0$ tomamos en principio el intervalo $[-0.9, 0]$.

Estudiemos si g es una aplicación contractiva en $[-0.9, 0]$

$$g'(x) = \frac{1}{x+1}$$

y $0 < 0.1 \leq x+1 \leq 1$ de donde $|g'(x)| = \dfrac{1}{|x+1|} \geq 1$ y la respuesta es negativa [21].

[21] No se cumple la condición de suficiencia del teorema 1.3.1 pero como se puede comprobar en la figura 1.17 la intersección de las gráficas de $y_1 = g(x)$ y de $y_2 = x$ definen ambas raíces.

Podemos forzar la convergencia por relajación (ver la sección 1.3.2). Partiendo de (1.49) consideramos el nuevo problema de punto fijo

$$x = h(x) = (1-\omega)x + \omega g(x)$$

Para obligar a que converja localmente de forma cuadrática, determinamos el factor de relajación ω de modo que $h'(\alpha_1) = 0$. Como desconocemos la raíz, hacemos una estimación seleccionando un valor $\xi \in [-0.9, 0]$ y obligando a que

$$h'(\xi) = (1-\omega) + \omega g'(\xi) = 0$$

Si, por ejemplo, $\xi = -0.8$ tenemos $g'(\xi) = \dfrac{1}{-0.8+1} = 5$ y

$$h'(\xi) = 1 - \omega + 5\omega = 0 \quad \Rightarrow \quad \omega = -\frac{1}{4}$$

Representemos la función derivada $h'(x) = (1/4)\,(5x+4)$.

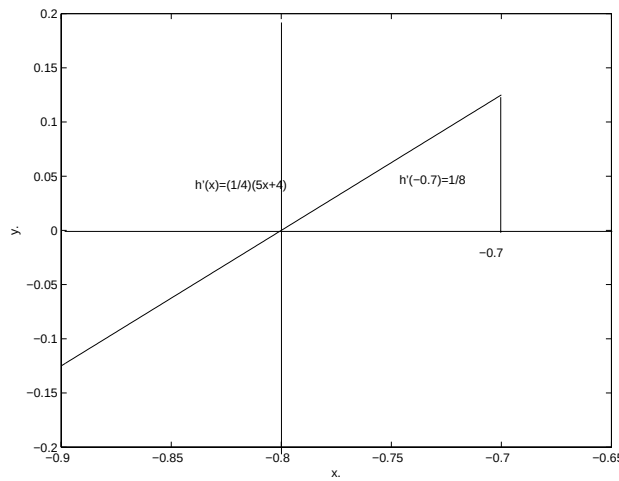


Figura 1.16: Representación gráfica de $h'(x)$ para $\omega = -\frac{1}{4}$.

Es evidente que $h'(x) < 1$ en $[-0.8, -0.7]$ y que al ser la función estrictamente creciente $|h'(x)| \leq h'(-0.7) = \dfrac{1}{8}$, lo que sugiere tomar ese valor como constante de Lipschitz [22] para estimar el error y el número de iteraciones.

Hemos relajado la función obteniendo el problema de punto fijo

$$x = h(x) = \frac{5}{4}x - \frac{1}{4}\left(\ln(x+1) + 1\right)$$

$$x^{(n+1)} = \frac{5}{4}x^{(n)} - \frac{1}{4}\left(\ln(x^{(n)}+1) + 1\right)$$

Iteramos, comenzando con $x^{(0)} = -0.6$

$x^{(0)}$	$x^{(1)}$	$x^{(2)}$	$x^{(3)}$	$x^{(4)}$	$x^{(5)}$	$x^{(6)}$	$x^{(7)}$
-0.6000	-0.7709	-0.8452	-0.8401	-0.8418	-0.8413	-0.8414	-0.8414

[22] Sería interesante estudiar la dependencia de L y ω. Ello precisaría la elección del mejor factor de convergencia relativo al menor valor de la constante de Lipschitz.

2. El error "a priori" tras siete iteraciones es (1.34)

$$|x^{(7)} - \alpha_1| \leq \frac{L^7}{1-L}|x^{(1)} - x^{(0)}| = 9.31 \cdot 10^{-8}$$

El número k de iteraciones necesarias para que el error sea menor que 10^{-5} satisface la desigualdad

$$k \geq \frac{\ln\left(\frac{10^{-5}(1-0.125)}{|\mathbf{x}^{(1)}-\mathbf{x}^{(0)}|}\right)}{\ln 0.125} \approx 4.75$$

luego $k = 5$.

En el caso de la raíz positiva α_2 podemos usar directamente la formulación (1.49) sin relajar la función g ya que $|g'(x)| < 1$ en cualquier intervalo $[a, b]$ con $0 < a < b < \infty$. En el intervalo $[2, 3]$, g verifica las

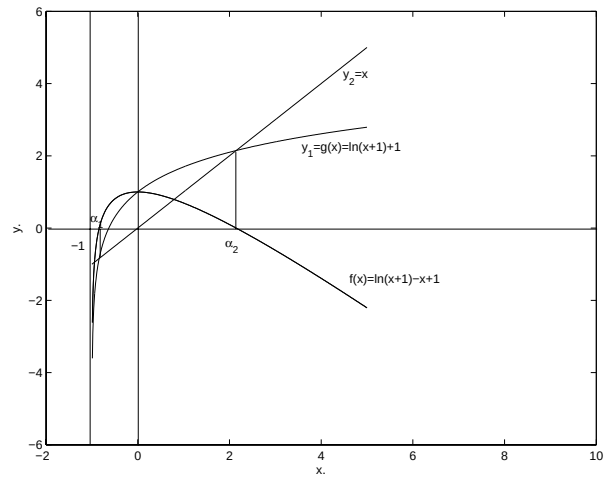


Figura 1.17: Representación gráfica de todas las funciones que definen la raíz $\alpha_2 > 0$.

dos condiciones del teorema de la aplicación contractiva.

En efecto, $g'(x) > 0$, luego g es estrictamente creciente y $g([2, 3]) = [2.09, 2.3863] \subset [2, 3]$.

Además g' es estrictamente decreciente con $g'(2) = \frac{1}{3}$, luego $|g'(x)| \leq \frac{1}{3}$ con $L = \frac{1}{3}$ como constante de Lipschitz.

Iteramos, comenzando con $x^{(0)} = 2.5 \in [2, 3]$. El método de aproximaciones sucesivas converge a α_2 en 9 iteraciones

$x^{(0)}$	$x^{(1)}$	$x^{(2)}$	$x^{(3)}$	$x^{(4)}$	$x^{(5)}$	$x^{(6)}$	$x^{(7)}$	$x^{(8)}$	$x^{(9)}$
2.5000	2.2528	2.1795	2.1567	2.1495	2.1472	2.1465	2.1463	2.1462	2.1462

siendo el error "a priori" tras ocho iteraciones (1.35)

$$|x^{(8)} - \alpha_2| \leq \frac{L^8}{1-L}|x^{(1)} - x^{(0)}| = 5.6516 \cdot 10^{-5}$$

El número k de iteraciones necesarias para que el error sea menor que 10^{-5} satisface la desigualdad

$$k \geq \frac{\ln\left(\frac{10^{-5}(1-0.333)}{|\mathbf{x}^{(1)}-\mathbf{x}^{(0)}|}\right)}{\ln 0.333} \approx 10.8383 \quad \Rightarrow \quad k = 11$$

PROBLEMA 1.4 *Newton Raphson 2D.*

Escribir las ecuaciones del esquema de Newton-Raphson para resolver el sistema

$$\begin{cases} x - 3x^2y = 0 \\ y - x^3 = 0 \end{cases}$$

Solución:
El método de Newton-Raphson para resolver el problema

$$\begin{pmatrix} g_1(x,y) \\ g_2(x,y) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

se escribe

$$\begin{pmatrix} x \\ y \end{pmatrix}_{n+1} = \begin{pmatrix} x \\ y \end{pmatrix}_n - \begin{pmatrix} \dfrac{\partial g_1}{\partial x} & \dfrac{\partial g_1}{\partial y} \\ \dfrac{\partial g_2}{\partial x} & \dfrac{\partial g_2}{\partial y} \end{pmatrix}_n^{-1} \begin{pmatrix} g_1 \\ g_2 \end{pmatrix}_n$$

y en este caso

$$\begin{aligned} \begin{pmatrix} x \\ y \end{pmatrix}_{n+1} &= \begin{pmatrix} x \\ y \end{pmatrix}_n - \begin{pmatrix} 1-6xy & -3x^2 \\ -3x^2 & 1 \end{pmatrix}_n^{-1} \begin{pmatrix} x - 3x^2y \\ y - x^3 \end{pmatrix}_n = \\ &= \begin{pmatrix} x \\ y \end{pmatrix}_n - \frac{1}{(1-6xy-9x^4)_n} \begin{pmatrix} 1 & 3x^2 \\ 3x^2 & 1-6xy \end{pmatrix}_n \begin{pmatrix} x - 3x^2y \\ y - x^3 \end{pmatrix}_n = \\ &= \begin{pmatrix} x \\ y \end{pmatrix}_n - \frac{1}{(1-6xy-9x^4)_n} \begin{pmatrix} x - 3x^2y + 3x^2y - 3x^5 \\ 3x^3 - 9x^4y + y - x^3 - 6xy^2 + 6x^4y \end{pmatrix}_n \end{aligned}$$

es decir

$$\begin{cases} x_{n+1} = x_n - \dfrac{(x - 3x^5)_n}{(1-6xy-9x^4)_n} \\ y_{n+1} = y_n - \dfrac{(2x^3 + y - 6xy^2 - 3x^4y)_n}{(1-6xy-9x^4)_n} \end{cases}$$

Se recomienda seguir el ejercicio con su aplicación práctica previo análisis de la existencia y unicidad de solución.

PROBLEMA 1.5 *Iteración de punto fijo 2D.*

El objetivo del ejercicio es hallar la solución del sistema no lineal

$$\left.\begin{aligned} x + \frac{y^2}{4} &= \frac{1}{16} \\ \frac{1}{3}\sin x + y &= \frac{1}{2} \end{aligned}\right\} \tag{1.50}$$

utilizando el método de aproximaciones sucesivas.

1. Escribir el sistema dado en la forma $\mathbf{z} = T(\mathbf{z})$ estudiando las propiedades de la aplicación T. En particular se determinará un conjunto abierto en el que el sistema tenga solución única $\mathbf{z}^*$ y que la sucesión iterativa $(\mathbf{z}^{(k)})$ del método de aproximaciones sucesivas definida por cualquier estimador inicial $\mathbf{z}^{(0)}$ de ese abierto converge a $\mathbf{z}^*$.

2. Elegir un estimador inicial $\mathbf{z}^{(0)}$ y determinar $\mathbf{z}^*$ con un error menor que 10^{-4}. ¿Cuántas iteraciones se deben hacer para conseguir que el error sea menor que 10^{-4}?

Solución:

1. El sistema viene dado en la forma $R(\mathbf{z}) = \mathbf{b}$ con $\mathbf{z} = (x, y)^T \in I\!R^2$, y lo escribiremos en la forma $\mathbf{z} = T(\mathbf{z})$ poniendo,

$$\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \frac{1}{16} - \frac{1}{4}y^2 \\ \frac{1}{2} - \frac{1}{3}\sin x \end{pmatrix} = T\begin{pmatrix} x \\ y \end{pmatrix}$$

Aun en un caso tan simple como éste, podemos escribir el sistema (1.50) en la forma $\mathbf{z} = T(\mathbf{z})$ de varias formas distintas y no todas ellas conducen a algoritmos convergentes. Para conseguir una aplicación T contractiva queremos que el segundo miembro de $\mathbf{z} = T(\mathbf{z})$ sea de algún modo "pequeño", lo que como veremos conseguimos en (2.31).

Eligiendo en $I\!R^2$ la norma del máximo, es fácil probar que T es una aplicación contractiva con $L = \frac{1}{2}$ en la bola cerrada unidad de origen $\mathbf{z}^{(0)} = (0, 0)$.

En efecto, tomemos $\mathbf{z}^{(i)}, \mathbf{z}^{(j)} \in B_c(0, 1)$ con $\mathbf{z}^{(i)} = (x^{(i)}, y^{(i)})$, $\mathbf{z}^{(j)} = (x^{(j)}, y^{(j)})$ y analicemos $\|T\mathbf{z}^{(i)} - T\mathbf{z}^{(j)})\|$,

$$T(\mathbf{z}^{(i)}) - T(\mathbf{z}^{(j)}) = \begin{pmatrix} \frac{1}{4}((y^{(j)})^2 - (y^{(i)})^2) \\ \frac{1}{3}(\sin x^{(j)} - \sin x^{(i)}) \end{pmatrix}$$

de donde

$$\|T(\mathbf{z}^{(i)}) - T(\mathbf{z}^{(j)})\| = \text{máx}\Big(\frac{1}{4}|(y^{(j)})^2 - (y^{(i)})^2|, \frac{1}{3}|\sin x^{(j)} - \sin x^{(i)}|\Big)$$

Como $(y^{(j)})^2 - (y^{(i)})^2 = (y^{(j)} - y^{(i)})(y^{(j)} + y^{(i)})$ se tiene $|(y^{(j)})^2 - (y^{(i)})^2| = |y^{(j)} - y^{(i)}||y^{(j)} + y^{(i)}|$ de donde ya que $\|\mathbf{z}^{(i)}\|$ y $\|\mathbf{z}^{(j)}\|$ son menores que uno, $|y^{(j)} + y^{(i)}| \leq 2$ y $|(y^{(j)})^2 - (y^{(i)})^2| \leq 2|y^{(j)} - y^{(i)}|$.

Además de

$$\sin\alpha - \sin\beta = \sin\frac{\alpha - \beta}{2}\cos\frac{\alpha + \beta}{2}$$
$$|\sin\alpha - \sin\beta| \leq 2\Big|\sin\frac{\alpha - \beta}{2}\Big| \leq |\alpha - \beta|$$

tendremos

$$|\sin x^{(i)} - \sin x^{(j)}| \leq |x^{(i)} - x^{(j)}|$$

con ello,

$$\|T(\mathbf{z}^{(i)}) - T(\mathbf{z}^{(j)})\| \leq \text{máx}\Big(\frac{1}{2}|y^{(i)} - y^{(j)}|, \frac{1}{3}||x^{(j)} - x^{(i)}|\Big)$$

y en cualquier caso

$$\|T(\mathbf{z}^{(i)}) - T(\mathbf{z}^{(j)})\| \leq \frac{1}{2}\|\mathbf{z}^{(i)} - \mathbf{z}^{(i)}\|$$

El sistema tiene solución única $\mathbf{z}^*$ en $[-1, 1]^2$ y la sucesión iterativa $(\mathbf{z}^{(k)})$ del método de aproximaciones sucesivas definida por cualquier estimador inicial $\mathbf{z}^{(0)}$ de esa bola (por ejemplo, $\mathbf{z}^{(0)} = (0, 0)$) converge a $\mathbf{z}^*$.

2. Con $\mathbf{z}^{(0)} = (0, 0)$, obtenemos las iterantes

$$\mathbf{z}^{(1)} = \begin{pmatrix} 0.0625 \\ 0.5000 \end{pmatrix}, \quad \mathbf{z}^{(2)} = \begin{pmatrix} 0 \\ 0.4792 \end{pmatrix}, \quad \mathbf{z}^{(3)} = \begin{pmatrix} 0.0051 \\ 0.5000 \end{pmatrix}, \ldots$$

que convergen a $\mathbf{z}^* = \big(0, \frac{1}{2}\big)$.

Debemos destacar que el estudio anterior siendo local no da ninguna información sobre la posibilidad de otras soluciones en el exterior de $B_c([0, 1])$.

Utilizando la desigualdad (1.36) que se establece en el corolario (1.4.2)

$$k \geq \frac{\ln\left(\frac{\epsilon(1-L)}{\|\mathbf{z}^{(1)} - \mathbf{z}^{(0)}\|}\right)}{\ln L} = \frac{\ln\left(\frac{10^{-4}\cdot 0.5}{0.5}\right)}{\ln 0.5} = 13.2886 \quad \Rightarrow \quad k = 14$$

PROBLEMA 1.6 *Teorema de la aplicación contractiva y dominio de atracción del método de Newton.*

Se trata de resolver el problema no lineal $f(x) = 0$ con

$$f(x) = x - \cos x$$

1. Reformular este problema en otro de la forma $x = T(x)$ del modo más sencillo posible.

2. Encontrar un intervalo compacto $[a, b]$ en el que la función T elegida cumpla las hipótesis del teorema 1.3.1, comprobando su verificación.

3. Se considera en $[a, b]$ la iteración de punto fijo asociada a T. Tomando como estimador inicial $(a+b)/2$, se pide iterar hasta que el residuo $r(x) = |x - T(x)|$ sea menor o igual que 0.1. ¿Cuántos pasos son necesarios? Valorar la velocidad de convergencia, utilizando los resultados del apartado anterior.

4. Aplicar el método de Newton para resolver el mismo problema, definiendo de modo preciso el esquema iterativo resultante.

5. Partiendo del mismo estimador inicial que en el apartado 3, iterar hasta que el residuo sea menor o igual que 0.1. ¿Cuántos pasos son necesarios?

6. Se define como en (1.2.1) el dominio de atracción de una raíz x^* para el método de Newton [23]. Encontrar algún punto que no pertenezca al dominio de atracción de la raíz cuya existencia se ha asegurado en el apartado 2.

7. Encontrar un punto que no pertenezca al dominio de atracción citado y que **no** sea múltiplo de $\pi/2$. Se podrá buscar un estimador inicial $x^{(0)}$ tal que algún término $x^{(k)}$ de la sucesión iterante del esquema de Newton que define sea un punto de tangente horizontal, por ejemplo $3\pi/2$.

   Escribir este problema en la forma $g(x) = 0$.

8. Encontrar un intervalo $[c, d]$ de longitud menor o igual que $\pi/2$ en el que exista al menos un cambio de signo de la función g, y tal que dicha función esté definida en los extremos de dicho intervalo.

9. Reformular este problema en la forma de una ecuación de punto fijo $x = T(x)$ del modo más sencillo posible.

10. Tomando como estimador inicial $(c+d)/2$, dar dos pasos en el esquema iterativo asociado a la ecuación de punto fijo del apartado anterior.

11. Tomando esta vez como medida de la convergencia el valor absoluto de la diferencia entre dos iteradas consecutivas, decidir teniendo en cuenta las dos iteraciones anteriores si hay o no convergencia.

12. Dar, a partir de los tres valores (estimador inicial y dos iteraciones), una estimación de la derivada de la función T en esa región.

13. ¿Justifica este valor el comportamiento encontrado en 11?

14. Supuesto que hayas decidido en el apartado 11 que hay convergencia, utiliza el valor encontrado en el apartado 12 para relajar el esquema iterativo del apartado 10 de modo que la convergencia sea lo más rápida posible. En el caso en que hayas decidido que no hay convergencia efectúa la misma operación para forzar la convergencia y que ésta sea lo más rápida posible.

    Justificar la selección del factor de relajación w.

15. Dar dos pasos con este esquema de relajación y verificar utilizando la misma medida de la convergencia que en 11 que las cosas han mejorado.

[23] Dicho dominio es el conjunto formado por los puntos $x^{(0)}$ que tomados como estimadores iniciales del esquema iterativo del método de Newton definen sucesiones que convergen a esa raíz.

Solución:

1. Lo más sencillo es poner $T(x) = \cos x$.

2. Dado que la imagen de T está acotada entre 0 y 1 en valor absoluto, podemos probar con el compacto $[0, 1]$. Como hay cambio de signo en los extremos

$$f(0) = -1 \quad \text{y} \quad f(1) = 0.4597$$

sabemos que hay al menos una raíz en $[0, 1]$.

Veamos ahora si T verifica las condiciones del teorema 1.3.1.

i. T es una función continua y $T[0, 1] \subset [0, 1]$.

En efecto, la función coseno es continua monótona decreciente entre 0 y π con $T(0) = 1 \in [0, 1]$ y $T[(1) = 0.5403 \in [0, 1]$, luego $T[0, 1] = [0.5403, 1.0000] \subset [0, 1]$.

ii. $T'(x)$ es continua y no nula en $(0, 1]$ ($T'(0) = 0$, pero 0 no es punto fijo de T) y

$$\exists L > 0, |T'(x)| \leq L < 1 \quad \forall x \in [0, 1]$$

$|T'(x)| = \sin x$, ya que el seno es positivo y creciente en el intervalo $[0, 1]$. Su máximo en ese intervalo lo alcanza por tanto en 1 donde vale $\sin(1) = 0.8415$. L existe y un posible valor es $L = 0.8415$.

T verifica las condiciones del teorema 1.3.1 y también sus conclusiones.

El operador T tiene un único punto fijo en $[0, 1]$ que se obtiene como límite de la sucesión $x^{(k+1)} = T(x^{(k)})$.

3. Construyamos la tabla correspondiente a partir del estimador inicial 0.5.

k	0	1	2	3	4
$x^{(k)}$	0.5000	0.8776	0.6390	0.8027	0.6947
$r(x^{(k)})$	0.3776	0.2386	0.1637	0.1079	0.0734

Como vemos, son necesarias cuatro iteraciones para que $r(x^{(k)}) < 0.1$. La velocidad de convergencia no es muy alta, ya que el valor de L encontrado es próximo a la unidad.

4.

$$x^{(k+1)} = x^{(k)} - \frac{f(x^{(k)})}{f'(x^{(k)})} = \frac{x^{(k)} \sin(x^{(k)}) + \cos(x^{(k)})}{1 + \sin(x^{(k)})}$$

5. Construyamos la tabla con los resultados de las sucesivas iteraciones a partir del estimador inicial $x^{(0)} = 0.5$.

k	$x^{(k)}$	$r(x^{(k)})$
0	0.5000	0.3776
1	0.7552	0.0271

Como era de esperar, la convergencia es más rápida. Sólo necesitamos un paso para que $r(x^{(k)}) < 0.1$.

6. De la interpretación geométrica del método de Newton, se deduce que en los valores de x donde la tangente es horizontal el esquema es divergente. Por eso buscamos puntos solución de la ecuación $f'(x) = 0$

$$1 + \sin(x) = 0 \quad \Rightarrow \quad x = \frac{3}{2}\pi + 2k\pi, \quad k \in \mathbb{Z}$$

Tomando por ejemplo $x^{(0)} = \frac{3}{2}\pi$, el método diverge.

7. Siguiendo la sugerencia del enunciado buscamos un $x^{(0)}$ tal que $x^{(1)} = \frac{3}{2}\pi$, en cuyo caso el esquema divergerá. Ese punto $x^{(0)}$ es tal que la tangente a la gráfica de f en dicho punto corta al eje x en $\frac{3}{2}\pi$, luego $x^{(0)}$ será, como podemos ver en la Figura 1.18, la solución de la ecuación no lineal

$$g(x) = x - \frac{x - \cos(x)}{1 + \sin(x)} - \frac{3}{2}\pi = 0$$

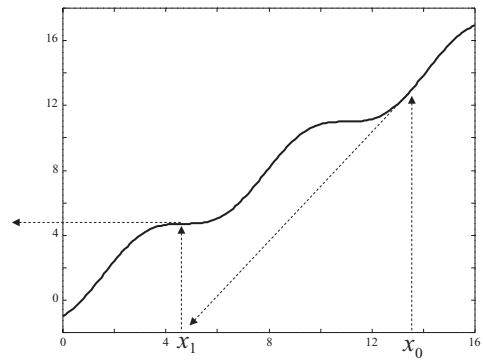


Figura 1.18: Función $f(x) = x - \cos(x)$.

8. Haciendo un estudio de signos de g, hallaremos un intervalo que contenga la raíz buscada.

   Construyamos la tabla correspondiente a partir de $\frac{4}{2}\pi$, que es la siguiente fracción de $\frac{1}{2}\pi$ a la derecha de $\frac{3}{2}\pi$.

x	$g(x)$	Signo
$\frac{4}{2}\pi$	$\frac{4}{2}\pi - \frac{\frac{4}{2}\pi - 1}{1+0} - \frac{3}{2}\pi = 1 - \frac{3}{2}\pi$	< 0
$\frac{5}{2}\pi$	$\frac{5}{2}\pi - \frac{\frac{5}{2}\pi - 0}{1+1} - \frac{3}{2}\pi = -\frac{1}{4}\pi$	< 0
$\frac{6}{2}\pi$	$\frac{6}{2}\pi - \frac{\frac{6}{2}\pi + 1}{1+0} - \frac{3}{2}\pi = -1 - \frac{3}{2}\pi$	< 0
$\frac{7}{2}\pi$	∞	
$\frac{8}{2}\pi$	$\frac{8}{2}\pi - \frac{\frac{8}{2}\pi - 1}{1+0} - \frac{3}{2}\pi = 1 - \frac{3}{2}\pi$	< 0
$\frac{9}{2}\pi$	$\frac{9}{2}\pi - \frac{\frac{9}{2}\pi - 0}{1+1} - \frac{3}{2}\pi = \frac{3}{4}\pi$	> 0

   de donde el intervalo buscado es $\left[\frac{8}{2}\pi, \frac{9}{2}\pi\right]$

9. Lo más natural es definir $T : D \to \mathbb{R}$ poniendo

$$T(x) = \frac{x - \cos(x)}{1 + \sin(x)} + \frac{3}{2}\pi$$

   con $D = \mathbb{R} - \left\{\frac{3}{2}\pi + 2k\pi\right\}, \quad k \in \mathbb{Z}$

10. El estimador inicial del esquema es $\frac{c+d}{2} = \frac{17}{4}\pi = 13.3518$.

    Disponemos en la misma tabla los resultados de las dos primeras iteraciones y el valor absoluto de la diferencia entre dos iteradas consecutivas

k	0	1	2
$x^{(k)}$	13.3518	12.1195	24.4681
$\lvert x^{(k)} - x^{(k-1)} \rvert$		1.2323	12.3487

11. Para tomar una decisión más razonada, habría que dar algunos pasos más pero los resultados obtenidos no son nada prometedores por lo que no parece que haya convergencia.

12.
$$T'(x^{(0)}) \approx \frac{T(x^{(1)}) - T(x^{(0)})}{x^{(1)} - x^{(0)}} = \frac{x^{(2)} - x^{(1)}}{x^{(1)} - x^{(0)}} = -10.0208$$

13. La derivada es en valor absoluto muy superior a la unidad. Esto significa que muy posiblemente no se verifique la condición de acotación de $|T'(x)|$ por un número menor que 1 del teorema 1.3.1 y que por tanto no se dé la convergencia.

14. Siguiendo la mecánica del método de relajación (sección 1.3.2) consideramos el nuevo esquema relajado
$$\mathcal{T}(x) = (1-w)x + wT(x)$$
y elegimos el factor de relajación w de modo que $\mathcal{T}'(x^{(0)}) = 0$ para acelerar la velocidad de convergencia, luego
$$\mathcal{T}'(x^{(0)}) = (1-w) + wT'(x^{(0)}) = 0 \quad \Rightarrow \quad w = \frac{1}{1 - T'(x^{(0)})} = 0.0907$$
Y por tanto el nuevo esquema relajado será:
$$x^{(k+1)} = 0.9093x^{(k)} + 0.0907\ T(x^{(k)})$$
También se recomienda aplicar el algoritmo de Wegstein y comparar sus comportamientos.

15. El estimador inicial del esquema vuelve a ser 13.3518 y disponemos en la misma tabla los resultados de las sucesivas iteraciones tanto del esquema original como del relajado y el valor absoluto de la diferencia entre dos iteradas consecutivas del esquema relajado

k	$\hat{x}^{(k)}$	$x^{(k)}$	$\lvert x^{(k)} - x^{(k-1)} \rvert$
0	13.3518		
1	12.1195	13.2400	0.1118
2	12.3848	13.1624	0.0776

donde $\hat{x}^{(k)}$ es el valor obtenido en cada paso sin relajar. Al menos en este punto del proceso y usando la misma medida de la convergencia que antes la cosa mejora.

De hecho, si seguimos iterando

k	$\hat{x}^{(k)}$	$x^{(k)}$	$\lvert x^{(k)} - x^{(k-1)} \rvert$
3	12.6126	13.1125	0.0499
4	12.7802	13.0823	0.0301
5	12.8902	13.0649	0.0174
6	12.9570	13.0551	0.0098
7	12.9956	13.0497	0.0054
...	...	...	...

Tomando este último valor 13.0497 como estimador inicial en el método de Newton, esperamos que el valor $x^{(1)}$ sea $\frac{3}{2}\pi$. El valor $x^{(1)}$ obtenido es 4.7449, que es muy similar a $4.7124 = \frac{3}{2}\pi$.

PROBLEMA 1.7 *Relajación de un esquema iterativo para resolver un problema físico.*

Se tienen una esfera de radio unidad cuya densidad es la cuarta parte de la del agua dulce y una pileta muy grande, de fondo plano, con un agujero de radio 0.5 unidades en el cual se ha quedado la esfera después de haber rodado por el fondo.

Se empieza a llenar de agua la pileta.

1. Obtener la ecuación que proporciona la altura de agua a la que hay que llenar la pileta para que la bola deje su agujero y empiece a flotar.
2. Utilizar el método de Newton con estimador inicial 0.1660 para encontrar dicha altura. Dar 3 pasos en el esquema iterativo detallando los valores intermedios obtenidos.
3. Convertir el problema en uno de punto fijo que se resolverá por el método de aproximaciones sucesivas.
4. Aplicar la iteración de punto fijo con el mismo estimador inicial y dando el mismo número de pasos que en el apartado 2. Valorar y justificar los resultados.

Se recuerda que el volumen correspondiente a un casquete esférico:

$$V(z) = \pi R z^2 - \frac{\pi}{3} z^3$$

Solución:

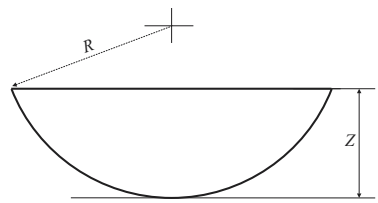


Figura 1.19: Casquete esférico.

1. Lo particular de este problema es que por debajo del agujero sólo actúa la presión atmosférica. Esto invalida el uso directo del principio de Arquímedes, pues para ello la ley de presiones tiene que estar relacionada linealmente con la profundidad siendo la pendiente ρg.

   Dividimos el cuerpo en tres partes, y planteamos el equilibrio global de fuerzas a partir del estudio de cada uno de estos cuerpos, en el momento justo en que la bola empieza a flotar. En ese instante no hay reacción en el borde del agujero, y las fuerzas de presión equilibran el peso de la bola (ver la Figura 1.20). El casquete superior está en equilibrio. En el casquete intermedio actúa el empuje vertical hacia

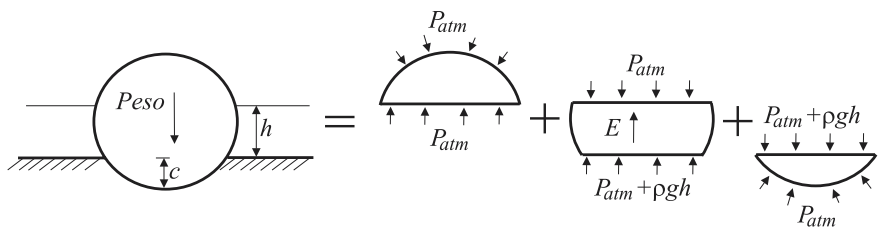


Figura 1.20: Modelo de equilibrio de fuerzas.

   arriba, ya que en ese casquete se verifica la distribución de presión creciente con la profundidad, lo que reflejamos haciendo que la presión en su parte inferior sea $P_{atm} + \rho g h$, siendo h la profundidad del agua. De hecho, ese empuje es consecuencia del crecimiento lineal de la ley de presiones.

   En el casquete inferior, la resultante va hacia abajo y su módulo es $\rho g h S$ siendo $S = \pi r^2$ la superficie del agujero. Por último actúa también hacia abajo, el peso de la bola.

   La ecuación global de equilibrio de fuerzas es:

$$Peso - E + \rho g h \pi r^2 = 0$$

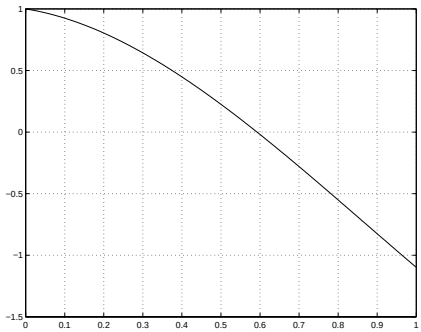


Figura 1.21: Función de equilibrio para h.

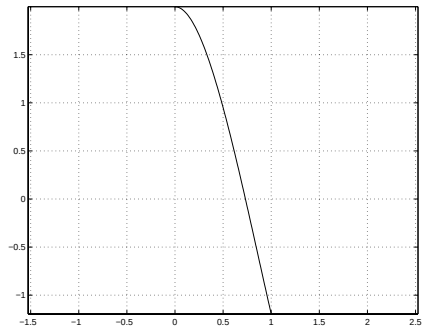


Figura 1.22: Función de aproximaciones sucesivas para h.

es decir,

$$\frac{4}{3}\pi R^3\frac{\rho}{4}g+\pi\left(\frac{h^3}{3}+(c-R)\,h^2+\left(c^2-2cR\right)h\right)\rho g+\rho g h\pi\left(\frac{R}{2}\right)^2=0$$

Operando

$$\frac{R^3}{3}+\frac{h^3}{3}+(c-R)\,h^2+\left(c^2-2cR\right)h+h\left(\frac{R}{2}\right)^2=0$$

$$h^3+3\,(c-R)\,h^2+3\left(c^2-2Rc+\frac{R^2}{12}\right)h+R^3=0$$

y ya que $c=1-\sin(\pi/3)=0.1340$.

$$h^3-2.5980h^2-0.5001h+1=0 \tag{1.51}$$

2. Llamando f a la función

$$f(h)=h^3-2.5980h^2-0.5001h+1$$

que representamos en la Figura 1.21, aplicamos el método de Newton con estimador inicial $h^{(0)}=0.1660$, para hallar la raíz de la ecuación $f(h)=0$. Los resultados de las distintas iteraciones son

k	0	1	2	3	4
$h^{(k)}$	0.1660	0.8301	0.5995	0.5924	0.5924

El valor buscado h^* es 0.5924.

3. Para utilizar el método de aproximaciones sucesivas, reformulamos el problema en la forma $h=T(h)$. Lo más simple es despejar h en la ecuación $f(h)=0$

$$h=1.9996h^3-5.1950h^2+1.9996$$

4. Iteremos sobre el esquema asociado a esa elección de T.

k	0	1	2	3	4
$h^{(k)}$	0.1660	1.8656	-3.0977	-107.2887	$-2.529\cdot10^6$

El proceso es claramente divergente.

El valor de la derivada de T en la zona de la raíz

$$T'(h)=3\cdot1.9996h^2-2\cdot5.1950h \quad\Rightarrow\quad T'(h^*)=-4.3894$$

es mucho mayor que uno en valor absoluto.

La raíz es un punto de repulsión.

Representando gráficamente la función T y las primeras iteraciones en esa zona (Figura 1.22), es más fácil comprender la divergencia.

Podemos utilizar las siguientes instrucciones Matlab para presentar esta figura:

```
h=0:0.01:1;
T=1.9996*h.^3-5.1950*h.^2+1.9996;
plot(h,T);
axis equal;
grid;
```

Se deja como ejercicio relajar este esquema para forzar la convergencia.

PROBLEMA 1.8 *Caída por un plano inclinado.*

Una partícula parte del reposo descendiendo sobre un plano inclinado empujada por su propio peso. El ángulo que forma el plano inclinado con la horizontal θ cambia con el tiempo con una velocidad constante ω, siendo $\theta = 0$ en el instante inicial. La partícula se encuentra en dicho instante sobre el eje de giro del plano. Planteando este problema en un sistema de coordenadas polares con origen en dicho eje (ver Figura 1.23) se llega a la ecuación diferencial

$$\frac{d^2r}{dt^2} - r\omega^2 + g\sin(\omega t) = 0$$

que describe la variación de la posición de la partícula sobre la rampa. Esta ecuación admite una integral

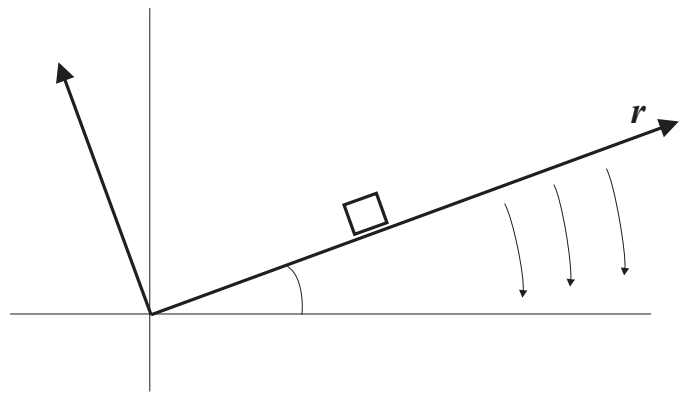


Figura 1.23: Plano inclinado correspondiente al problema 1.8.

analítica cuyas constantes se fijan utilizando las condiciones iniciales de posición y velocidad nulas

$$r(t) = -\frac{g}{2\omega^2}\left(\sinh(\omega t) - \sin(\omega t)\right) \tag{1.52}$$

Se tomará la gravedad $g = 10\ m/s^2$.

Se supone que en un segundo la partícula ha recorrido $0.5\ m$, lo que significa, con el sentido dado a los ejes, que $\omega = \frac{d\theta}{dt}$ es negativa.

Nuestro objetivo es calcular el valor de ω correspondiente a ese desplazamiento.

1. Formular el problema como la busca de la raíz de una ecuación $f(\omega) = 0$.

2. Variar ω empezando por $\omega = -1$, evaluando la función f en los distintos valores hasta detectar un cambio de signo que permita definir un intervalo que contenga a la raíz buscada.

3. Tomando como estimador inicial -0.375, iterar con el método de Newton hasta que la diferencia entre dos iteraciones consecutivas sea en valor absoluto menor que 10^{-2}.

4. Reformular el problema $f(\omega) = 0$ en una ecuación de punto fijo $\omega = T(\omega)$.

   Asumiendo que el valor de ω obtenido en el apartado 3 es la raíz buscada, estudiar la convergencia local del esquema $\omega = T(\omega)$.

5. Tomando como estimador inicial $\omega^{(0)} = -0.375$, dar 3 pasos en este esquema.

6. Caso de que no se observen indicios claros de convergencia, relajar el esquema 4 tomando como factor de relajación el óptimo correspondiente a la raíz obtenida en el apartado 3.

   Escribir el esquema resultante $\omega^{(i+1)} = \mathcal{T}(\omega^{(i)})$.

7. Tomando el mismo estimador inicial $\omega^{(0)} = -0.375$, iterar con este esquema relajado hasta que la diferencia entre dos iteraciones consecutivas sea en valor absoluto menor que 0.05.

Solución:

1. Para reformular el problema en la forma $f(\omega) = 0$ basta sustituir en (1.52) el valor $r = 0.5$ para $t = 1$.

$$10\,(\sinh\omega - \sin\omega) + \omega^2 = 0 \tag{1.53}$$

2. La tabla de valores

ω	-1	-2	-0.5	-0.25
$f(\omega)$	-2.33	-23.1756	-0.1666	0.0104

   detecta un cambio de signo en el intervalo $[-0.5, -0.25]$.

3.
$$\omega^{(i+1)} = \omega^{(i)} - \frac{f(\omega^{(i)})}{f'(\omega^{(i)})}$$
$$\omega^{(i+1)} = \omega^{(i)} - \frac{10\left(\sinh\omega^{(i)} - \sin\omega^{(i)}\right) + {\omega^{(i)}}^2}{10\left(\cosh\omega^{(i)} - \cos\omega^{(i)}\right) + 2\omega^{(i)}}$$

i	0	1	2	3
$\omega^{(i)}$	-0.375	-0.3214	-0.3025	-0.3000
$\lvert\omega^{(i)} - \omega^{(i-1)}\rvert$		0.0536	0.0189	0.0025

   La velocidad angular buscada es -0.3000 rad/sg.

4. Tomamos $T(\omega) = \omega + f(\omega)$. El esquema es localmente convergente si el valor absoluto de la derivada de la función T en la raíz es menor que la unidad.

$$T(\omega) = \omega + 10\,(\sinh\omega - \sin\omega) + \omega^2 \quad \Rightarrow \quad T'(\omega) = 1 + f'(\omega) = 1 + 10\,(\cosh\omega - \cos\omega) + 2\omega$$

   de donde $T'(-0.3000) = 1.3$ que es superior a la unidad en valor absoluto. No se dan las condiciones que garantizan la convergencia local. La raíz es un punto de repulsión.

5.
$$\omega^{(i+1)} = T(\omega^{(i)}) = \omega^{(i)} + 10\left(\sinh\omega^{(i)} - \sin\omega^{(i)}\right) + \left(\omega^{(i)}\right)^2\omega^{(i)} + f(\omega^{(i)})$$

   Podemos iterar sobre este esquema utilizando las dos líneas Matlab siguientes, repitiendo la segunda

```
w=-0.375
w=w+10*(sinh(w)-sin(w))+w^2
```

Se tiene con ello la siguiente tabla de resultados:

i	0	1	2	3	4	5	6
$\omega^{(i)}$	-0.375	-0.4102	-0.4719	-0.5996	-0.9588	-2.9807	-90.7476

Que como preveíamos en el apartado anterior resulta ser un esquema divergente.

6. Sea α el factor de relajación buscado. Siguiendo la mecánica del método de relajación (sección 1.3.2), definimos el esquema relajado
$$\omega = (1-\alpha)\omega + \alpha T(\omega) = \mathcal{T}(\omega)$$
obligando a que $\mathcal{T}'(-0.3000)(1-\alpha) + \alpha T'(-0.3000) = 0$ se tiene
$$\alpha = -3.3333$$
El esquema relajado es por tanto
$$\omega^{(i+1)} = 4.3333\omega^{(i)} - 3.3333T(\omega^{(i)}) = \mathcal{T}(\omega^{(i)})$$
es decir,
$$\omega^{(i+1)} = 4.3333\omega^{(i)} - 3.3333\left(\omega^{(i)} + f(\omega^{(i)})\right)$$
de donde por fin
$$\omega^{(i+1)} = \omega^{(i)} - 3.3333 f(\omega^{(i)})$$
Siguiendo las instrucciones del enunciado se completa la tabla de resultados

i	0	1	2	3	4
$\omega^{(i)}$	-0.375	-0.2578	-0.2890	-0.2992	-0.3000
$\lvert\omega^{(i)} - \omega^{(i-1)}\rvert$		0.1172	0.0312	0.0102	0.0008

donde es evidente la mejora de comportamiento del esquema relajado.

PROBLEMA 1.9 *Comparación de los métodos de Newton y Broyden para la resolución de sistemas de ecuaciones no lineales.*

Se considera el sistema no lineal de ecuaciones (S)

$$(S) \qquad \begin{cases} x^2+y^2+z^2 &= 3 \\ x^2+y^2-z &= 1 \\ x+y+z &= 3 \end{cases}$$

que representa la intersección de una esfera, un paraboloide y un plano y que posee la solución única $\mathbf{x}^* = (1,1,1)$.

El sistema (S) se puede expresar en la forma $\mathbf{g}(\mathbf{x}) = \mathbf{0}$ con

$$\mathbf{g}(\mathbf{x}) = \left(x^2+y^2+z^2-3, x^2+y^2-z-1, x+y+z-3\right)$$

1. Calcular los tres primeros términos de la sucesión producida por el método de Newton con estimador inicial $\mathbf{x}^{(0)} = (1,0,1)$ y analizar lo que sucede cuando $\mathbf{x}^{(0)} = (0,0,0)$.

2. Tomando de nuevo como estimador inicial $\mathbf{x}^{(0)} = (1,0,1)$ se considera el método de Broyden con matriz inicial $\mathbf{D}_0 = \text{grad}\ \mathbf{g}(1,0,1)$ que se irá actualizando en cada paso.

   Calcular dos términos más de la sucesión aproximante definida por este método y compararlos con los obtenidos en el apartado anterior.

Solución:

1. Se tiene

$$\mathbf{G} = \text{grad } \mathbf{g}(x,y,z) = \begin{pmatrix} 2x & 2y & 2z \\ 2x & 2y & -1 \\ 1 & 1 & 1 \end{pmatrix}$$

que no es simétrica.

Con el estimador inicial $\mathbf{x}^0 = (1,0,1)$ tenemos

$$\mathbf{G}^{(0)} = \begin{pmatrix} 2 & 0 & 2 \\ 2 & 0 & -1 \\ 1 & 1 & 1 \end{pmatrix} \; ; \qquad \mathbf{g}^{(0)} = \begin{pmatrix} -1 \\ -1 \\ -1 \end{pmatrix}$$

y $\det(\mathbf{G}^{(0)}) = 6 \neq 0$.

El sistema $\mathbf{G}^{(0)}\left(\mathbf{x} - \mathbf{x}^{(0)}\right) = \mathbf{G}^{(0)}\mathbf{s}^{(0)} = -\mathbf{g}^{(0)}$

$$\left\{ \begin{array}{ll} 2(x-1) & +2(z-1) = 1 \\ 2(x-1) & -(z-1) = 1 \\ (x-1)+y+(z-1) & = 1 \end{array} \right.$$

tiene solución única $\mathbf{x}^{(1)} = \left(\frac{3}{2}, \frac{1}{2}, 1\right)$.

Si se calcula la inversa de $\mathbf{G}^{(0)}$ tendremos

$$\left(\mathbf{G}^{(0)}\right)^{-1} = \begin{pmatrix} \frac{1}{6} & \frac{1}{3} & 0 \\ -\frac{1}{2} & 0 & 1 \\ \frac{1}{3} & -\frac{1}{3} & 1 \end{pmatrix}$$

y

$$\mathbf{s}^{(0)} = -\left(\mathbf{G}^{(0)}\right)^{-1}\mathbf{g}^{(0)} = \begin{pmatrix} \frac{1}{2} \\ \frac{1}{2} \\ 0 \end{pmatrix}$$

Creando una imagen dinámica del método de Newton en el que el punto $\mathbf{x}^{(k+1)}$ se obtiene moviéndose en $\mathbb{R}^n \times \mathbb{R}^n$ a partir del punto $\left(\mathbf{x}^{(k)}, \mathbf{g}^{(k)}\right)$ sobre la recta que pasa por dicho punto de vector director $\mathbf{s}^{(k)} = -{\mathbf{G}^{(k)}}^{-1}\mathbf{g}^{(k)}$ hasta interceptar al hiperplano coordenado $\mathbb{R}^n \times \{\mathbf{0}_n\}$, nos movemos aquí en el plano de ecuación $z = 1$ (ver la figura 1.24) de $\mathbf{x}^0 = (1,0,1)$ a $\mathbf{x}^{(1)} = \left(\frac{3}{2}, \frac{1}{2}, 1\right) = (1.5, 0.5, 1)$.

Calculemos la segunda iteración

$$\mathbf{G}^{(1)} = \begin{pmatrix} 3 & 1 & 2 \\ 3 & 1 & -1 \\ 1 & 1 & 1 \end{pmatrix} \; ; \qquad \mathbf{g}^{(1)} = \begin{pmatrix} \frac{1}{2} \\ \frac{1}{2} \\ 0 \end{pmatrix}$$

El sistema $\mathbf{G}^{(1)}\left(\mathbf{x} - \mathbf{x}^{(1)}\right) = \mathbf{G}^{(1)}\mathbf{s}^{(1)} = -\mathbf{g}^{(1)}$ tiene solución única

$$\mathbf{s}^{(1)} = \begin{pmatrix} -\frac{1}{4} \\ \frac{1}{4} \\ 0 \end{pmatrix} \quad \Longrightarrow \quad \mathbf{x}^{(2)} = \left(\frac{5}{4}, \frac{3}{4}, 1\right)$$

y de nuevo nos movemos en el plano de ecuación $z = 1$ de $\mathbf{x}^{(1)} = \left(\frac{3}{2}, \frac{1}{2}, 1\right)$ a $\mathbf{x}^{(2)} = \left(\frac{5}{4}, \frac{3}{4}, 1\right) = (1.25, 0.75, 1)$.

En la tercera iteración se obtiene

$$\mathbf{s}^{(2)} = \begin{pmatrix} -\frac{1}{8} \\ \frac{1}{8} \\ 0 \end{pmatrix} \quad \Longrightarrow \quad \mathbf{x}^{(3)} = \left(\frac{9}{8}, \frac{7}{8}, 1\right) = (1.125, 0.875, 1)$$

Calculemos una útima iteración

$$\mathbf{G}^{(3)} = \begin{pmatrix} -\frac{9}{4} & \frac{7}{4} & 2 \\ \frac{9}{4} & \frac{7}{4} & -1 \\ 1 & 1 & 1 \end{pmatrix} ; \qquad \mathbf{g}^{(3)} = \begin{pmatrix} \frac{1}{32} \\ \frac{1}{32} \\ 0 \end{pmatrix}$$

de donde

$$\mathbf{s}^{(3)} = \begin{pmatrix} -\frac{1}{16} \\ \frac{1}{16} \\ 0 \end{pmatrix} \quad \Longrightarrow \quad \mathbf{x}^{(4)} = \left(\frac{17}{16}, \frac{15}{16}, 1 \right) = (1.0625, 0.9375, 1.0000)$$

Se puede intuir una ley de formación de la sucesión aproximante.

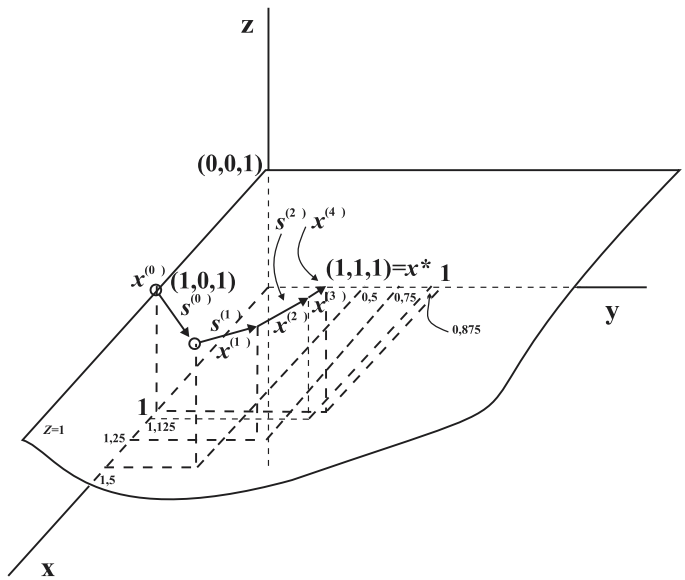


Figura 1.24: Sucesivas iteradas del método de Newton en el movimiento que se va aproximando paso a paso a la solución $\mathbf{x}^* = (1, 1, 1)$.

Comenzando en $\mathbf{x}^{(0)} = (1, 0, 1)$ la ley de recurrencia es

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \mathbf{s}^{(k)}$$

con

$$\mathbf{s}^{(k)} = \begin{pmatrix} -\frac{1}{2^{k+1}} \\ \frac{1}{2^{k+1}} \\ 0 \end{pmatrix}$$

Si tomamos como estimador inicial $\mathbf{x}^{(0)} = (0, 0, 0)$,

$$\mathbf{G}^{(0)} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -1 \\ 1 & 1 & 1 \end{pmatrix}$$

que no posee inversa, y por tanto la sucesión del método de Newton no está definida para esa elección de estimador inicial.

2. De acuerdo con el algoritmo que se incluye en la sección 1.4.2, tomamos $\mathbf{D}^{(0)} = \mathbf{G}^{(0)}$ luego

$$\mathbf{D}^{(0)} = \begin{pmatrix} 2 & 0 & 2 \\ 2 & 0 & -1 \\ 1 & 1 & 1 \end{pmatrix}$$

de modo que la primera iteración de Broyden coincide con la de Newton luego $\mathbf{x}^{(1)} = \left(\frac{3}{2}, \frac{1}{2}, 1\right)$.

Para definir la segunda iteración actualizamos $\mathbf{D}^{(0)}$ según el algoritmo adjunto, se tiene

$$\mathbf{s}^{(0)} = \left(\mathbf{x}^{(1)} - \mathbf{x}^{(0)}\right) = \left(\frac{1}{2}, \frac{1}{2}, 0\right) \quad \text{y} \quad \mathbf{g}^{(1)} = \left(\frac{1}{2}, \frac{1}{2}, 0\right)$$

luego

$$\mathbf{D}^{(1)} = \mathbf{D}^{(0)} + \frac{\mathbf{g}^{(1)} \otimes \mathbf{s}^{(0)}}{\|\mathbf{s}^{(0)}\|^2} = \begin{pmatrix} 2 & 0 & 2 \\ 2 & 0 & -1 \\ 1 & 1 & 1 \end{pmatrix} + 2 \begin{pmatrix} \frac{1}{2} \\ \frac{1}{2} \\ 0 \end{pmatrix} \cdot \left(\frac{1}{2}, \frac{1}{2}, 0\right) = \begin{pmatrix} \frac{5}{2} & \frac{1}{2} & 2 \\ \frac{5}{2} & \frac{1}{2} & -1 \\ 1 & 1 & 1 \end{pmatrix}$$

Para calcular $\mathbf{x}^{(2)}$ resolvemos el sistema

$$\mathbf{D}^{(1)} \left(\mathbf{x} - \mathbf{x}^{(1)}\right) = -\mathbf{g}^{(1)}$$

cuya solución es $\mathbf{x}^{(2)} = \left(\frac{5}{4}, \frac{3}{4}, 1\right)$, la misma aproximación que por el método de Newton.

Actualicemos $\mathbf{D}^{(1)}$

$$\mathbf{s}^{(1)} = \left(\mathbf{x}^{(2)} - \mathbf{x}^{(1)}\right) = \left(\frac{5}{4}, \frac{3}{4}, 1\right) - \left(\frac{3}{2}, \frac{1}{2}, 1\right) = \left(-\frac{1}{4}, \frac{1}{4}, 0\right) \quad \text{y} \quad \mathbf{g}^{(2)} = \left(\frac{1}{8}, \frac{1}{8}, 0\right)$$

luego

$$\mathbf{D}^{(2)} = \mathbf{D}^{(1)} + \frac{\mathbf{g}^{(2)} \otimes \mathbf{s}^{(1)}}{\|\mathbf{s}^{(1)}\|^2} = \begin{pmatrix} \frac{5}{2} & \frac{1}{2} & 2 \\ \frac{5}{2} & \frac{1}{2} & -1 \\ 1 & 1 & 1 \end{pmatrix} + 8 \begin{pmatrix} \frac{1}{8} \\ \frac{1}{8} \\ 0 \end{pmatrix} \cdot \left(-\frac{1}{4}, \frac{1}{4}, 0\right) =$$

$$= \begin{pmatrix} \frac{5}{2} & \frac{1}{2} & 2 \\ \frac{5}{2} & \frac{1}{2} & -1 \\ 1 & 1 & 1 \end{pmatrix} + \begin{pmatrix} -\frac{1}{4} & \frac{1}{4} & 0 \\ -\frac{1}{4} & \frac{1}{4} & 0 \\ 0 & 0 & 0 \end{pmatrix} = \begin{pmatrix} \frac{9}{4} & \frac{3}{4} & 2 \\ \frac{9}{4} & \frac{3}{4} & -1 \\ 1 & 1 & 1 \end{pmatrix}$$

Para calcular $\mathbf{x}^{(3)}$ resolvemos el sistema

$$\mathbf{D}^{(2)} \left(\mathbf{x} - \mathbf{x}^{(2)}\right) = -\mathbf{g}^{(2)}$$

cuya solución es $\mathbf{x}^{(3)} = \left(\frac{7}{6}, \frac{5}{6}, 1\right) = (1.1667, 0.8334, 1.0000)$.

Aquí se empiezan a separar los puntos de la sucesión aproximante de Broyden de los correspondientes de la de Newton aunque también nos movemos en la interpretación dinámica de aproximación a la raíz dentro del plano $z = 1$.

Representando como antes dichos puntos en $\mathbb{R}^3$ se observa en la Figura 1.25 que la convergencia de las sucesivas iteradas es más lenta.

Demos un paso más del algoritmo actualizando $\mathbf{D}^{(2)}$

$$\mathbf{s}^{(2)} = \left(\mathbf{x}^{(3)} - \mathbf{x}^{(2)}\right) = \left(\frac{7}{6}, \frac{5}{6}, 1\right) - \left(\frac{5}{4}, \frac{3}{4}, 1\right) - = \left(-\frac{1}{12}, \frac{1}{3}, 0\right) \quad \text{y} \quad \mathbf{g}^{(3)} = \left(\frac{1}{18}, \frac{1}{18}, 0\right)$$

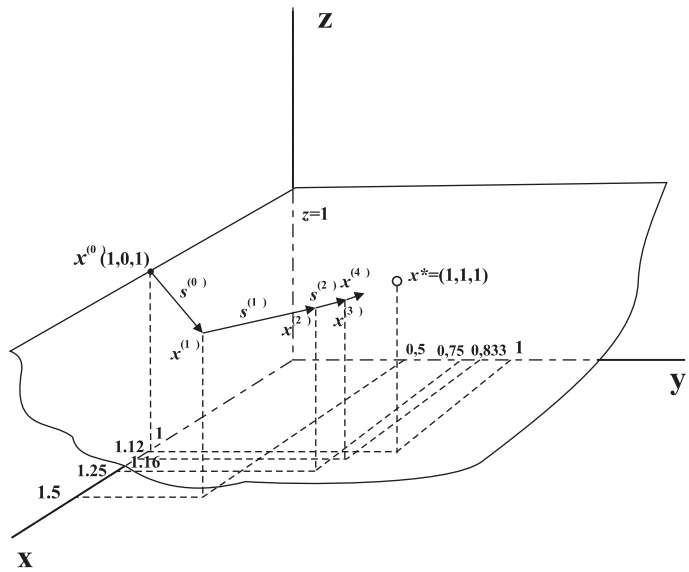


Figura 1.25: Distintas aproximaciones sucesivas de la raíz $\mathbf{x} = (1, 1, 1)$ obtenidas por el método de Broyden correspondientes al problema 1.9.

luego

$$\mathbf{D}^{(3)} = \mathbf{D}^{(2)} + \frac{\mathbf{g}^{(3)} \otimes \mathbf{s}^{(2)}}{\|\mathbf{s}^{(2)}\|^2} = \begin{pmatrix} \frac{9}{4} & \frac{3}{4} & 2 \\ \frac{9}{4} & \frac{3}{4} & -1 \\ 1 & 1 & 1 \end{pmatrix} + \frac{144}{17} \begin{pmatrix} \frac{1}{18} \\ \frac{1}{18} \\ 0 \end{pmatrix} \cdot \left(-\frac{1}{12}, \frac{1}{3}, 0 \right) =$$

$$= \begin{pmatrix} \frac{9}{4} & \frac{3}{4} & 2 \\ \frac{9}{4} & \frac{3}{4} & -1 \\ 1 & 1 & 1 \end{pmatrix} + \begin{pmatrix} -\frac{2}{51} & \frac{8}{51} & 0 \\ -\frac{2}{51} & \frac{8}{51} & 0 \\ 0 & 0 & 0 \end{pmatrix} = \begin{pmatrix} \frac{451}{204} & \frac{185}{204} & 2 \\ \frac{451}{204} & \frac{185}{204} & -1 \\ 1 & 1 & 1 \end{pmatrix}$$

Para calcular $\mathbf{x}^{(4)}$ resolvemos el sistema

$$\mathbf{D}^{(3)} \left(\mathbf{x} - \mathbf{x}^{(3)} \right) = -\mathbf{g}^{(3)}$$

cuya solución es $\mathbf{x}^{(4)} = \left(\frac{299}{266}, \frac{233}{266}, 1 \right) = (1.1240, 0.8760, 1.0000)$.

A la vista de los valores obtenidos está claro que la obtenida por el método de Newton está más próxima a la solución que la obtenida por el método de Broyden.

PROBLEMA 1.10 *Resolución de un sistema no lineal mediante aproximaciones sucesivas.*

El objetivo de este ejercicio es utilizar el esquema iterativo asociado al método de aproximaciones sucesivas para resolver el sistema de ecuaciones no lineales (E)

$$(E) \qquad \begin{cases} x + 0.25y^2 &= 1.25 \\ 0.25x^2 + y &= 1.25 \end{cases}$$

Sabemos desde el principio que el sistema (E) tiene dos soluciones $\mathbf{x}_1^* = (1, 1)$ y $\mathbf{x}_2^* = (-5, -5)$.

1. Escribir (E) en la forma $\mathbf{z} = \mathbf{T}(\mathbf{z})$ con $\mathbf{z} = (x, y)$ del modo más sencillo posible. Demostrar que la aplicación $\mathbf{T}$ es contractiva en el entorno de $(1, 1)$ precisando el radio de una bola de centro en ese punto en la que $\mathbf{T}$ posee dicha propiedad así como una constante de Lipschitz asociada.

   Utilizaremos el esquema iterativo asociado al método de aproximaciones sucesivas $\mathbf{z}^{(k+1)} = \mathbf{T}(\mathbf{z}^{(k)})$ con $\mathbf{z}^{(k)} = (x^{(k)}, y^{(k)})$ para aproximar el punto fijo de $\mathbf{T}$.

2. Tomando como estimador inicial $\mathbf{z}^{(0)}$ un punto cualquiera de la bola determinada en 1.

   a) Efectuar una estimación a priori del error cometido al tomar el término $\mathbf{z}^{(20)}$ de la sucesión aproximante como solución de (E).

   b) Hacer una estimación del número de iteraciones k necesarias para conseguir que el error $\|\mathbf{x}_1^* - \mathbf{z}^{(k)}\|$ sea menor o igual que 10^{-6}.

   c) Determinar $\mathbf{z}^{(k)}$ efectuando una estimación a posteriori del error cometido y comparando el resultado con el obtenido en a).

3. Tomando ahora la estimación inicial que se desee, incluida, claro es, en la bola determinada en 1. y efectuando una aceleración de la convergencia por el método Δ^2 de Aitkens determinar la solución $\mathbf{x}_1^*$ con error menor que 10^{-6}.

4. Demostrar que $\mathbf{T}$ no es una aplicación contractiva en el entorno de $\mathbf{x}_2^* = (-5, -5)$.

Comentarios y cuestiones suplementarias

1. Es habitual considerar en $\mathbb{R}^2$ la norma de máximo $\|(x, y)\|_\infty = \max(|x|, |y|)$. Sería interesante elegir alguna otra, todas son equivalentes, y ver en qué cambian los razonamientos o los resultados.

2. Por consideraciones de simetría del problema parece posible **rebajar su dimensionalidad** en el sentido siguiente:

   Se hace $y = x$, es decir, se avanza por la diagonal principal de $\mathbb{R}^2$ y se considera el problema **1D** de la resolución de la ecuación $x = f(x)$ con

   $$f(x) = 1.25 - 0.25x^2$$

   Analizar la equivalencia entre el sistema (E) y el problema de una variable así enunciado. Responder a todas las cuestiones propuestas en el ejercicio utilizando esta equivalencia. ¿Qué se observa?

3. Analizar la siguiente posibilidad aparentemente menos natural que la anterior.

   Considerando el sistema (E) escrito en la forma

   $$(E) \qquad \begin{cases} x &= f(y) \\ y &= f(x) \end{cases}$$

   se obtiene eliminando una de las variables, por ejemplo, y, la ecuación $x = f^2(x)$.

4. El dominio de atracción de la "otra" raíz $\mathbf{x}_2^* = (-5, -5)$ es el conjunto de puntos $\{(\pm 5, \pm 5)\}$. Si tomamos uno cualquiera de ellos como estimador inicial nos plantamos en la raíz $\mathbf{x}_2^*$ en una iteración. Para cualquier otro estimador inicial o bien la sucesión $(\mathbf{z}_n)$ converge a $\mathbf{x}_1^*$ o bien no converge en lo absoluto.

   Como muestra de lo anterior analizar las sucesiones aproximantes de estimadores iniciales $(3, 3)$, $(-4, -4)$ $(6, 6)$, $(-5, -0.45)$.

Solución:
El esquema iterativo asociado al método de aproximaciones sucesivas viene definido por:

$$\begin{pmatrix} x^{(k+1)} \\ y^{(k+1)} \end{pmatrix} = \begin{pmatrix} 1.25 - 0.25{y^{(k)}}^2 \\ 1.25 - 0.25{x^{(k)}}^2 \end{pmatrix} = \mathbf{T} \begin{pmatrix} x^{(k)} \\ y^{(k)} \end{pmatrix}$$

1. Veamos que se define así una contracción en un entorno de $(1, 1)$.

   Dotamos a $\mathbb{R}^2$ de la norma del máximo

   $$\|(x, y)\| = \max(|x|, |y|)$$

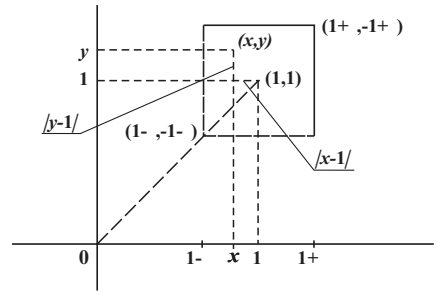


Figura 1.26: Bola $B((1,1);\epsilon)$ de $\mathbb{R}^2$ para la norma del máximo.

y tomamos como entorno de $(1,1)$ la bola $B((1,1);\epsilon)$

$$B((1,1);\epsilon) = \{(x,y) \in \mathbb{R}^2 \quad \|(x-1,y-1)\| = \max(|x-1|,|y-1|) < \epsilon\}$$

representada geométricamente por un cuadrado de centro $(1,1)$ y lado 2ϵ (ver Figura 1.26). Sean (x_1,y_1) y (x_2,y_2) dos puntos cualesquiera de $\mathbb{R}^2$,

$$\|\mathbf{T}(x_1,y_1) - \mathbf{T}(x_2,y_2)\| = \max\left\{0.25|y_2^2 - y_1^2|, 0.25|x_2^2 - x_1^2|\right\} =$$
$$= 0.25\max\{|y_1 - y_2||y_1 + y_2|, |x_1 - x_2||x_1 + x_2|\}$$

Si ambos puntos están en la $B((1,1);\epsilon)$ podemos mayorar [24] tanto $|y_1 + y_2|$ como $|x_1 + x_2|$ por $2 + 2\epsilon$, luego

$$\|\mathbf{T}(x_1,y_1) - \mathbf{T}(x_2,y_2)\| \leq \frac{1+\epsilon}{2}\|(x_1,y_1) - (x_2,y_2)\|$$

Con ello, si $\epsilon = \frac{1}{2}$ por ejemplo, una constante de Lipschitz sería

$$L = \frac{1+\epsilon}{2} = \frac{3}{4} = 0.75 < 1$$

y la aplicación $\mathbf{T}$ es contractiva en la bola $B\left((1,1);\frac{1}{2}\right)$ con $L = \dfrac{3}{4}$. Si tomamos $\epsilon = \frac{1}{4}$, una constante de Lipschitz en la bola $B\left((1,1);\frac{1}{4}\right)$ sería

$$L = \frac{1+\epsilon}{2} = \frac{5}{8} = 0.625$$

2. Podemos aplicar el método de aproximaciones sucesivas con un estimador inicial $\mathbf{z}^{(0)}$ que pertenezca a una de las bolas estudiadas. Por consideraciones de simetría, parece razonable tomar el estimador inicial en la diagonal principal de $\mathbb{R}^2$, con lo que se avanzará siempre a lo largo de esa diagonal y el problema se convierte en unidimensional [25]. Tomamos por tanto en la bola $B\left((1,1);\frac{1}{2}\right)$ y sobre la diagonal principal el elemento $\mathbf{z}^{(0)} = (0.8, 0.8)$.

   a) Sabemos que

   $$\|\mathbf{x}_1^* - \mathbf{z}^{(k)}\| \leq \frac{L^k}{1-L}\|\mathbf{z}^{(1)} - \mathbf{z}^{(0)}\| \qquad k = 1,2,...$$

   luego con $L = \dfrac{3}{4}$ y $k = 20$ obtenemos la siguiente estimación a priori del error

   $$\|\mathbf{x}_1^* - \mathbf{z}^{(20)}\| \leq \frac{(\frac{3}{4})^{20}}{1-\frac{3}{4}}\|\mathbf{z}^{(1)} - \mathbf{z}^{(0)}\|$$

[24] El caso más desfavorable se produce cuando y_1 e y_2 o x_1 y x_2 sean $1 \pm \epsilon$, en cuyo caso se tiene la mayoración hallada.
[25] Ver el apartado 2. de las cuestiones suplementarias.

con $\mathbf{z}^{(1)} = \mathbf{T}\mathbf{z}^{(0)} = (1.09, 1.09)$, luego $\|\mathbf{z}^{(1)} - \mathbf{z}^{(0)}\| = |1.09 - 0.8| = 0.29$ y

$$\|\mathbf{x}_1^* - \mathbf{z}^{(20)}\| < \frac{0.0031712}{0.25} 0.29 \approx 3.6786 \times 10^{-3}$$

Si hubiéramos tomado la bola $B\left((1,1); \frac{1}{4}\right)$ con el mismo estimador inicial $\mathbf{z}^{(0)} = (0.8, 0.8)$ se refinan las mayoraciones anteriores,

$$\|\mathbf{x}_1^* - \mathbf{z}^{(20)}\| < \frac{0.000082718}{0.375} 0.29 \approx 6.3969 \times 10^{-5}$$

Como se ve, el estudio anterior es muy pesimista y ni siquiera esta última estimación mejora esa impresión.

b) El número de iteraciones k necesarias para conseguir que el error $\|\mathbf{x}_1^* - \mathbf{z}^{(k)}\|$ sea menor o igual que un cierto número ϵ es

$$k \geq \frac{\ln\left(\frac{\epsilon(1-L)}{\|\mathbf{z}^{(1)} - \mathbf{z}^{(0)}\|}\right)}{\ln L}$$

y en nuestro caso, con $\epsilon = 10^{-6}$ y $L = \frac{3}{4}$ se tiene

$$k \geq \frac{-13.9639}{-0.2877} = 48.5394 \approx 49 \text{ iteraciones}$$

Con $L = 0.625$

$$k \geq \frac{-13.5585}{-0.4700} = 28.8476 \approx 29 \text{ iteraciones}$$

y de nuevo constatamos la gran mejora que se obtiene en el segundo caso.

c) Se obtiene la siguiente tabla de iteradas en la que sólo se ha escrito la primera proyección de $\mathbf{z}^{(k)}$ ya que todos los puntos $\mathbf{z}^{(k)}$ están sobre la diagonal principal

$\mathbf{z}^{(0)} = (0.8, .)$	$\mathbf{z}^{(8)} = (0.99927693, .)$	$\mathbf{z}^{(16)} = (0.99999718, .)$
$\mathbf{z}^{(1)} = (1.09, .)$	$\mathbf{z}^{(9)} = (1.00036140, .)$	$\mathbf{z}^{(17)} = (1.00000141, .)$
$\mathbf{z}^{(2)} = (0.952975, .)$	$\mathbf{z}^{(10)} = (0.99981927, .)$	$\mathbf{z}^{(18)} = (0.999997176, .)$
$\mathbf{z}^{(3)} = (1.02295966, .)$	$\mathbf{z}^{(11)} = (1.00009036, .)$	$\mathbf{z}^{(19)} = (1.00000035, .)$
$\mathbf{z}^{(4)} = (0.98838832, .)$	$\mathbf{z}^{(12)} = (0.99995482, .)$	$\mathbf{z}^{(20)} = (0.99999982, .)$
$\mathbf{z}^{(5)} = (1.00577210, .)$	$\mathbf{z}^{(13)} = (1.00002259, .)$	$\mathbf{z}^{(21)} = (1.00000009, .)$
$\mathbf{z}^{(6)} = (0.99710562, .)$	$\mathbf{z}^{(14)} = (0.99998870, .)$	$\mathbf{z}^{(22)} = (0.9999999, .)$
$\mathbf{z}^{(7)} = (1.00144510, .)$	$\mathbf{z}^{(15)} = (1.00000565, .)$	

La estimación del error cometido al tomar $\mathbf{z}^{(20)}$ como solución de (E) a posteriori es

$$\|\mathbf{x}_1^* - \mathbf{z}^{(20)}\| \leq \frac{L}{1-L}\|\mathbf{z}^{(20)} - \mathbf{z}^{(19)}\| < \frac{0.75}{0.25}|0.99999982 - 1.00000035| \approx 1.59 \times 10^{-6}$$

que resalta el gran pesimismo de la estimación a priori $\|\mathbf{x}_1^* - \mathbf{z}^{(20)}\| < 3.678592 \times 10^{-3}$.
Con $L = 0.625$ se tiene

$$\|\mathbf{x}_1^* - \mathbf{z}_{20}\| \leq \frac{0.625}{0.375}|0.99999982 - 1.00000035| \approx 8.825 \times 10^{-7}$$

La conclusión llegados a este punto, es que esta hipótesis $L = 0.625$ es bastante realista. De hecho, con $\mathbf{z}^{(0)} = (0.0, 0.0)$ que está claramente fuera de los entornos que hasta ahora hemos considerado, se consigue con $\mathbf{z}^{(20)}$ una estimación a posteriori $\|\mathbf{x}_1^* - \mathbf{z}^{(20)}\| \approx 4 \cdot 10^{-7}$ como se puede comprobar.

3. Ya que el proceso es prácticamente lineal, es oportuno aplicar el algoritmo Δ^2 de Aitkens para acelerar la convergencia del proceso iterativo.

   Partiremos de la sucesión $\{\mathbf{z}^{(k)}\}$ obtenida con la estimación inicial $\mathbf{z}^{(0)} = (0.8, 0.8)$.

   Lo aplicaremos tan sólo a la primera proyección $x^{(k)}$ de $\mathbf{z}^{(k)} = (x^{(k)}, y^{(k)})$ con la que obtenemos la sucesión $\{w^{(k)}\}$ mediante el algoritmo

$$w^{(k)} = x^{(k)} - \frac{(x^{(k+1)} - x^{(k)})^2}{x^{(k+2)} - 2x^{(k+1)} + x^{(k)}}$$

   para lo que necesitamos tres iteradas consecutivas de la sucesión $\{x^{(k)}\}$, así obtenemos

$$\begin{aligned}
w^{(0)} &= x^{(0)} - \frac{(x^{(1)} - x^{(0)})^2}{x^{(2)} - 2x^{(1)} + x^{(0)}} \approx 0.99694398 \\
w^{(1)} &= x^{(1)} - \frac{(x^{(2)} - x^{(1)})^2}{x^{(3)} - 2x^{(2)} + x^{(1)}} \approx 0.99929658 \\
w^{(2)} &= x^{(2)} - \frac{(x^{(3)} - x^{(2)})^2}{x^{(4)} - 2x^{(3)} + x^{(2)}} \approx 0.99981927 \\
w^{(3)} &= x^{(3)} - \frac{(x^{(4)} - x^{(3)})^2}{x^{(5)} - 2x^{(4)} + x^{(3)}} \approx 0.99995557 \\
w^{(4)} &= x^{(4)} - \frac{(x^{(5)} - x^{(4)})^2}{x^{(6)} - 2x^{(5)} + x^{(4)}} \approx 0.99998900 \\
w^{(5)} &= x^{(5)} - \frac{(x^{(6)} - x^{(5)})^2}{x^{(7)} - 2x^{(6)} + x^{(5)}} \approx 0.99999705 \\
w^{(6)} &= x^{(6)} - \frac{(x^{(7)} - x^{(6)})^2}{x^{(8)} - 2x^{(7)} + x^{(6)}} \approx 0.99999914 \\
w^{(7)} &= x^{(7)} - \frac{(x^{(8)} - x^{(7)})^2}{x^{(9)} - 2x^{(8)} + x^{(7)}} \approx 1.00000012
\end{aligned}$$

   y con nueve iteradas $\mathbf{z}^{(k)}$ obtenemos siete iteradas de la sucesión $\{w^{(k)}\}$ de Aitkens con las que aproximamos la solución x_1^* con un error igual al que cometíamos en el método de aproximaciones sucesivas con $\mathbf{z}^{(21)}$. La aceleración de la convergencia es muy importante.

4. Estudiemos si $\mathbf{T}$ es una aplicación contractiva en el entorno de $\mathbf{x}_2^* = (-5, -5)$.

   De nuevo buscamos un número real $\epsilon > 0$ tal que cualesquiera que sean los puntos (x_1, y_1) y (x_2, y_2) pertenecientes a la bola $B((-5, -5); \epsilon)$ se tenga

$$\|\mathbf{T}(x_1, y_1) - \mathbf{T}(x_2, y_2)\| \leq L\|(x_1, y_1) - (x_2, y_2)\|$$

   con $L < 1$.

   Un razonamiento análogo al efectuado en el apartado 1 nos da

$$\|\mathbf{T}(x_1, y_1) - \mathbf{T}(x_2, y_2)\| \leq \frac{10 + 2\epsilon}{4}\|(x_1, y_1) - (x_2, y_2)\|$$

   Las desigualdades

$$10 + 2\epsilon < 4 \qquad \text{y} \qquad \epsilon > 0$$

   son incompatibles, luego la respuesta es negativa.

Comentarios y cuestiones suplementarias

1. Siguiendo la sugerencia del enunciado exploramos otras normas de $\mathbb{R}^2$

Utilizando la norma euclídea $\|(x,y)\|_2 = \sqrt{x^2+y^2}$ llegamos a

$$\begin{aligned}\|\mathbf{T}(x_1,y_1) - \quad & \mathbf{T}(x_2,y_2)\|_2 = 0.25\sqrt{(y_2^2-y_1^2)^2+(x_2^2-x_1^2)^2} = \\ & = 0.25\sqrt{(y_2+y_1)^2(y_2-y_1)^2+(x_2+x_1)^2(x_2-x_1)^2}\end{aligned}$$

Si r es un número estrictamente positivo consideremos el cuadrado $\mathcal{C}$ centrado en el origen $(0,0)$ de lado $4r$. Cualesquiera que sean (x_1,y_1) y (x_2,y_2) en $\mathcal{C}$, se tiene que $|x_1|<2r,\ |x_2|<2r,\ |y_1|<2r,\ |y_2|<2r$ luego

$$\|\mathbf{T}(x_1,y_1) - \mathbf{T}(x_2,y_2)\|_2 =\leq r\|(x_1,y_1)-(x_2,y_2)\|_2$$

La mayor bola euclídea centrada en $(1,1)$ contenida en $\mathcal{C}$ tiene radio $|2r-1|$.

Cuando $r<1$, la aplicación $\mathbf{T}$ es contractiva en la bola $B_2((1,1),|2r-1|)$. Si, por ejemplo, $r=0.75$ (luego también L) $|2r-1|=0.5$ y el problema se continuaría sin grandes cambios utilizando la norma euclídea.

Tomando ahora como estimador inicial $\mathbf{z}^{(0)} = (1.1,1.1) \in B_2((1,1),\frac{1}{2})$, se tiene $\mathbf{z}^{(1)} = \mathbf{T}(\mathbf{z}^{(0)}) = (0.9475,0.9475) \in B_2((1,1),\frac{1}{2})$.

El error a priori al tomar $\mathbf{z}^{(20)}$ como aproximación de la solución $\mathbf{x}_1^*$ es

$$\|\mathbf{x}_1^* - \mathbf{z}^{(20)}\|_2 \leq \frac{(0.75)^{20}}{1-0.75}\|\mathbf{z}^{(1)}-\mathbf{z}^{(0)}\|_2 \approx 2.7357\cdot 10^{-3}$$

El número de iteraciones k necesarias para conseguir que el error $\|\mathbf{x}_1^* - \mathbf{z}^{(k)}\|_2$ sea menor o igual que 10^{-6} es

$$k \geq \frac{-13.66778751}{-0.287682} \approx 47.7977 \quad \Rightarrow \quad 48 \text{ iteraciones}$$

Se obtiene $\mathbf{z}^{(20)} = 1.000000098$ y el error a posteriori es aproximadamente $1.25\cdot 10^{-6}$. Si se toma la iterada $\mathbf{z}^{(21)} = 0.99999995$ el error es $6.245\cdot 10^{-7} < 10^{-6}$. Las conclusiones son similares a las ya obtenidas con la norma del máximo.

2. En efecto, haciendo $y=x$ avanzamos por la diagonal principal de $\mathbb{R}^2$ y el problema se convierte en uno de una variable ligado a la resolución de la ecuación $x=f(x)$ con

$$f(x) = 1.25 - 0.25x^2$$

ecuación que define una parábola $y = 1.25-0.25x^2$ en $\mathbb{R}^2$ cuyas raíces son $x=1$ y $x=-5$ (Figura 1.27) y el problema se ha terminado. Como $f'(x) = -0.50\cdot x$, es claro que f será contractiva en $B(1;\epsilon)$

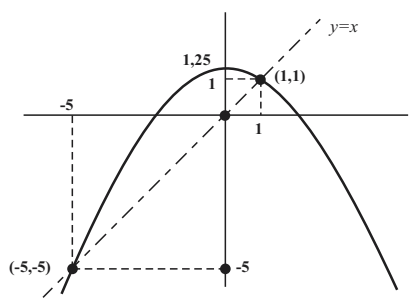


Figura 1.27: Puntos fijos de la ecuación $x = 1.25 - 0.25x^2$

ssi $\|f'\| = \sup_{x\in(1-\epsilon,1+\epsilon)} 0.50\cdot x = 0.50(1+\epsilon) < 1$, luego si $1+\epsilon<2$, lo que, por ejemplo, se cumple si $\epsilon<1$.

¿Qué sucede en el entorno de -5?

$$\|f'\| = \sup_{x\in(-5-\epsilon,-5+\epsilon)} 0.50\cdot x = 0.50(5+\epsilon) < 1$$

y obviamente no existe ningún $\epsilon>0$ tal que $5+\epsilon<2$.

3. La ecuación $x = f \circ f(x) = f^2(x)$ de cuarto grado [26], hereda los puntos fijos de f.

   Como la derivada de $f \circ f(x)$ es $(f \circ f)'(x) f' \circ f(x) \cdot f'(x)$ se tiene en la bola de centro 1 y radio ϵ que

   $$\|(f \circ f)'\| = \sup_{x \in B(1;\epsilon)} |f'(f(x)||f'(x)| = \sup_{x \in B(1;\epsilon)} 0.5|f(x)|0.5|x|$$

   Si $x \in (1-\epsilon, 1+\epsilon)$; $|x-1| < \epsilon \Rightarrow |x| < 1+\epsilon$ y $|f(x)-1| = 0.25|1-x^2| = 0.25|x-1||x+1| < 0.25\epsilon(1+\epsilon)$ luego de $|f(x)| - 1 < |f(x) - 1| \Rightarrow |f(x)| < 0.25\epsilon(1+\epsilon) + 1$.

   Con ello,

   $$\|(f \circ f)'\| \leq 0.5\,(0.25\epsilon(1+\epsilon) + 1)\,0.5(1+\epsilon)$$

   que debe ser menor que 1 para que $f \circ f$ sea contractiva.

   Operando, $\epsilon^3 + 2\epsilon^2 + 5\epsilon - 12 < 0$, cúbica siempre creciente que toma signo negativo para $0 \leq \epsilon \leq \alpha$ con $1 < \alpha < 1.5$, luego en toda bola de centro 1 donde f es contractiva también lo es $f \circ f$, siendo la recíproca falsa y al menos teóricamente la constante de Lipschitz es mejor, más pequeña que para f en esa bola.

   Si $\epsilon = \frac{1}{2}$, por ejemplo

   $$\begin{aligned}\|(f \circ f)'\| =& \sup_{x \in B(1;\frac{1}{2})} 0.5|f(x)|0.5|x| \leq 0.25\left(1.25 - 0.25^2\right)(1 + 0.5) = \\ &= 0.25 \cdot 1.5 \cdot 1.1875 = 0.4453125 < 1\end{aligned}$$

4. La conclusión relativa al dominio de atracción de $\mathbf{x}_2^* = (-5, -5)$ es correcta. Las sucesiones aproximantes de los estimadores iniciales $(3,3), (-4,-4), (6,6), (-5,-0.45)$ son

   $(3,3);\ (1,-1);\ (1,1)$
   $(-4,-4);\ (-2.75,-2.75);\ (-0.640,-0.640);\ (1.147,1.147); \ldots \rightarrow \quad (1,1)$
   $(6,6);\ (-7.75,-7.75);\ (-13.8,-13.8) \qquad$ que no converge.
   $(-5,-0.45);\ (1.12,-5);\ (-5,0.89);\ (1.05,-5) \quad$ que no converge.

PROBLEMA 1.11 *Coeficiente de pérdida de carga lineal en una tubería.*

Se considera un flujo turbulento en una tubería y sea $\mathcal{R} \geq 3500$ el número de Reynolds asociado a su diámetro D. Se desea calcular el coeficiente λ de pérdida de carga lineal utilizando la relación [27] de Colebrook [5]

$$(E) \qquad \lambda^{-\frac{1}{2}} = -2\log_{10}\left(\frac{\epsilon}{3.71 \cdot D} + \frac{2.51}{\mathcal{R}\lambda^{\frac{1}{2}}}\right)$$

donde

- λ es un parámetro adimensional el coeficiente de pérdida de carga lineal o factor de fricción de Moody.
- ϵ es la rugosidad de la tubería.

Para obtener una estimación inicial de λ se usará el valor suministrado por la fórmula empírica de Hermann

$$\lambda^{(0)} = 0.0054 + 0.395\mathcal{R}^{-0.3}$$

El objetivo del ejercicio es calcular aproximadamente λ para los siguientes valores de $\mathcal{R}$ y de ϵ/D

[26] ¡Cuidado! la función considerada es $f \circ f = f^2$ y no $f \cdot f = f^2$ que por un claro abuso de notación se denotan del mismo modo. En el primer caso $f^2(x) = 1.25 - 0.25(f(x))^2 = 1.25 - 0.25(1.25 - 0.25x^2)^2 = -0.015625x^4 + 0.15625x^2 + 0.76172$ y en el segundo, $f^2(x) = (f(x))^2 = 0.0625x^4 - 0.625x^2 + 1.95312$. En nuestro caso, los puntos fijos de f son también puntos fijos de $f \circ f$. En efecto, si $x^* = f(x^*) \Rightarrow f \circ f(x^*) = f(x^*) = x^*$, lo que no sucede en el segundo caso. En suma, son funciones distintas que sólo tienen en común la notación.

[27] Esta ecuación se basa sobre los datos empíricos obtenidos por Nikuradse (1933) a través de una serie muy extensa de experimentos en los que hallaba λ en función de $\mathcal{R}$ y de la rugosidad relativa de la tubería ϵ/D.

$\mathcal{R}$	ϵ/D
10^4	0.05
10^5	0.003
10^6	0.003

Se utilizarán para ello los siguientes métodos:

1. Método de las aproximaciones sucesivas.
   - 1.1 Efectuar un análisis previo de la existencia y unicidad de la solución y de la convergencia de la sucesión aproximante.
   - 1.2 Hacer una estimación a priori del número de iteraciones necesarias para obtener la solución con un error $e < 10^{-6}$ para cada pareja de valores $(\mathcal{R}, \epsilon/D)$.
   - 1.3 Resolver la ecuación.
   - 1.4 Efectuar estimaciones "a priori" y "a posteriori" del error en cada caso comparando los resultados.
   - 1.5 Efectuar una aceleración de la convergencia por el método Δ^2 de Aitkens en el caso $(\mathcal{R}, \epsilon/D) = (10^4, 0.05)$.
   - 1.6 Aplicar el método de Steffensen al caso $(\mathcal{R}, \epsilon/D) = (10^4, 0.05)$.
2. Método de Wegstein con estimadores iniciales sugeridos en cada caso, por el apartado anterior y test de parada $\epsilon < 10^{-6}$ comparando el número de iteraciones.
3. Método de Newton-Raphson.
   - 3.1 Hallar un intervalo que contenga a la raíz λ^* y en el que el método de Newton-Raphson converja independientemente del estimador inicial
   - 3.2 Determinar la raíz λ^* con el estimador inicial $\lambda^{(0)}$ obtenido mediante la fórmula de Hermann, y test de parada $\epsilon < 10^{-6}$.
   - 3.3 Utilizar el método de von Mises para determinar λ^* ahorrándonos así las evaluaciones de la derivada primera de f. Analizar ambos métodos.
4. Con el mismo estimador inicial que en el apartado 3.2, determinar la raíz λ^* en el caso $(\mathcal{R}, \epsilon/D) = (10^5, 0.003)$ utilizando
   - 4.1 el método Illinois.
   - 4.2 el método Pegasus.

Solución:

1. Tomando como nueva variable $x = \dfrac{1}{\sqrt{\lambda}} > 0$ la ecuación (E) se escribe

$$(E) \qquad x = -2\log_{10}\left(\frac{\epsilon}{3.71 \cdot D} + \frac{2.51}{\mathcal{R}}x\right)$$

- $(\mathcal{R}, \epsilon/D) = (10^4, 0.05)$
  En este caso,

$$(E) \qquad x = -2\log_{10}\left(0.013477 + 0.000251\, x\right) = F_1(x)$$

  - 1.1 Estudiaremos F_1 en su dominio aunque luego restringiremos el estudio a los valores de x estrictamente positivos.
    La función F_1 no existe si el paréntesis argumento del $\log_{10}$ es negativo. Su dominio de definición es por tanto $x > -53.69322709$.
    En el punto $x = -53.69322709$, $\lim_{x \to -53.69322709} F_1(x) = +\infty$ y F_1 tiene una asíntota vertical.

F_1 se anula para $x = 3930.370518$.
Derivando,

$$F_1'(x) = -\frac{0.000218015}{0.013477 + 0.000251\ x} < 0 \quad \forall x > -53.69322709$$

F_1 es siempre decreciente. Para $x = -52.82464143$ se tiene $F_1'(x) = -1$ de modo que cualquiera que sea $x > -52.82464143$, $|F_1'(x)| < 1$.
La aplicación F_1 satisface el teorema del punto fijo de Banach en cualquier conjunto cerrado contenido en el intervalo $[-52.82464143, +\infty)$ y el método de aproximaciones sucesivas converge al único punto fijo de F_1 cualquiera que sea el estimador inicial escogido en dicho cerrado.
De acuerdo con el enunciado, la búsqueda de esa raíz se encauza eligiendo una estimación inicial suministrada por la fórmula empírica de Hermann

$$\lambda^{(0)} = 0.0054 + 0.39510^{-1.2} = 0.03032281$$

de donde

$$x^{(0)} = \frac{1}{\sqrt{\lambda^{(0)}}} = 5.742688256$$

Ese punto pertenece al intervalo $[-52.82464143, +\infty)$, luego es un estimador inicial válido. Además, $F_1(5.742688256) = 3.65255464 < 5.742688256$, y $F_1(5.742688256) - 5.742688256 = -2.09013361 < 0$. Por otro lado, $F_1(3) = 3.6793602 > 3$ y $F_1(3) - 3 = 0.6793602 > 0$, luego x^* la única raíz de la ecuación $F_1(x) - x = 0$ y único punto fijo de F_1 está en el intervalo cerrado $I_1 = [3, 5.742688256]$ que tomaremos como el conjunto cerrado en el que se cumplen las condiciones y conclusiones del teorema de Banach citado y que usaremos también en apartados posteriores.
En la Figura 1.28 se sintetizan los resultados anteriores.

1.2 El número k de iteraciones necesarias para conseguir que el error $|x^* - x^{(k)}|$ sea menor o igual que $\epsilon = 10^{-6}$ basado en una estimación "a priori" del error, viene dado por 1.36. La constante de Lipschitz de F_1 en I_1 es

$$L = \max_{x \in I_1} |F_1'(x)| = |F_1'(3)| = 0.015320801$$

Con ello, tomando como estimador inicial $x^{(0)} = 5.742688256$, $x^{(1)} = F_1(x^{(0)}) = 3.652554646$ y

$$k \geq \frac{\ln(0.0000004711)}{ln(0.015320801)} = \frac{-6.3268869}{-1.8147185} \sim 3.48643 \quad \Rightarrow \quad k = 4$$

Necesitamos al menos cuatro iteraciones del método de aproximaciones sucesivas para obtener la precisión pedida.

1.3 Usando el código Matlab con el estimador inicial ya considerado $x^{(0)} = 5.742688256$ y limitando el número de iteraciones a 4, se tiene:

$$\begin{aligned} x^{(1)} &= 3.65255464607588 \\ x^{(2)} &= 3.68364960220285 \\ x^{(3)} &= 3.68317874950901 \\ x^{(4)} &= 3.68318587745245 \end{aligned}$$

Se aproxima la raíz x^* por el valor $x^{(4)} = 3.68318573506278$ con los indicadores de convergencia absoluto $|x^{(4)} - x^{(3)}| = 0.00000712794343 > 10^{-6}$ y relativo $\frac{|x^{(4)} - x^{(3)}|}{|x^{(4)}| + 10^{-6}} = 0.00000193526521 > 10^{-6}$ y no se consigue la precisión deseada.
Con 5 iteraciones, aproximamos la raíz x^* por el valor $x^{(5)} = 3.68318576954655$ con un error absoluto $|x^{(5)} - x^{(4)}| = 1.0790590 \cdot 10^{-7} < 10^{-6}$ y un error relativo $\frac{|x^{(5)} - x^{(4)}|}{|x^{(5)}| + 10^{-6}} = 2.929688 \cdot 10^{-8} < 10^{-6}$ y $\lambda^* = 0.07371447078918$.

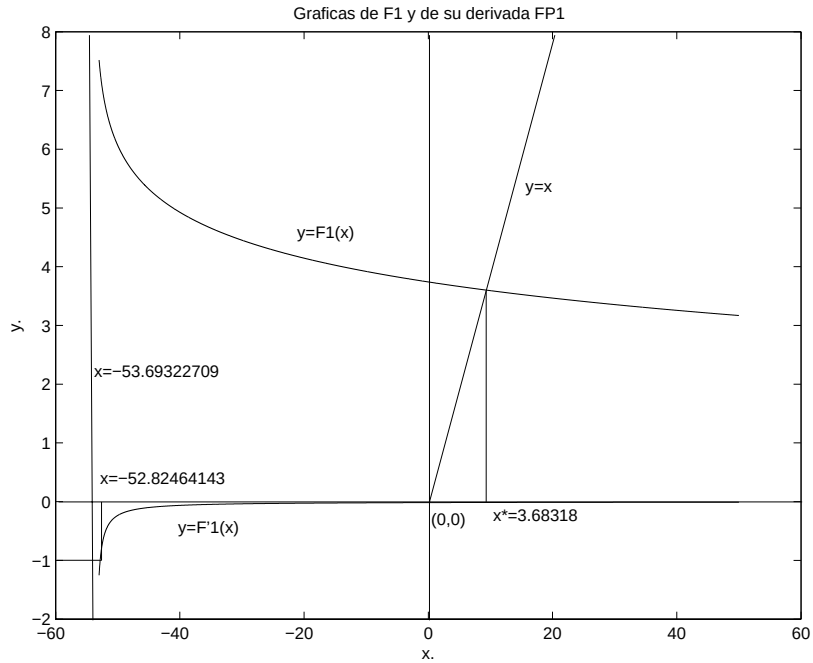


Figura 1.28: Representación gráfica de F_1 y de su derivada F_1'. Se representan la asíntota vertical $x = -53.69322709$ de ambos grafos, el extremo inferior $x = -52.82464143$ del intervalo en el que F_1 es contractiva y el punto fijo $x^* \sim 3.68317875$ de F_1 objetivo del ejercicio.

1.4 La estimación "a priori" del error cometido al tomar $x^{(5)}$ como solución del problema es

$$|x^* - x^{(5)}| \leq \frac{L^5}{1-L}|x_1 - x_0| \approx (8.5726006893 \cdot 10^{-10})2.09013361 = 1.79179 \cdot 10^{-9}$$

y la estimación "a posteriori" del error es

$$|x^* - x^{(5)}| \leq \frac{L}{1-L}|x^{(5)} - x^{(4)}| < 0.01555918 \cdot 0.0000001079059 \approx 1.67893 \cdot 10^{-9}$$

verdaderamente ajustados.

1.5 Hemos programado el proceso de construcción de la sucesión $z^{(k)}$ de Aitken en un código Matlab que incluimos a continuación ya aplicado a nuestro caso

```
function [k,err,X,Z]=Aitkens(f,x0,epsilon,max1)
x0=5.742688256;
epsilon=1e-7;
max1=6;
X=zeros(1,max1+1);
Z=zeros(1,max1+1);
X(1)=x0
X(2)=feval('Colebrook1',X(1))
X(3)=feval('Colebrook1',X(2))
Z(3)=X(1)-(X(2)-X(1))^2/(X(3)-2*X(2)+X(1))
for k=4:max1+1
        X(k)=feval('Colebrook1',X(k-1));
        denom=X(k)-2*X(k-1)+X(k-2);
        if denom=0
        '<¡Ojo!, division por cero en Aitkens'
```

```
        break
    else
        Z(k)=X(k-2)-(X(k-1)-X(k-2))^2/denom
end
    err=abs(Z(k-1)-Z(k));
    relerr=err/(abs(Z(k-1))+epsilon);
    if (err<epsilon) | (relerr<epsilon)
        break
    end
end
[X' Z']
[k  err relerr]
```

donde Colebrook1 es la función

```
function f=Colebrook1(x)
R=0.013477+(0.000251).*x;
y=log10(R);
f=(-2).*y-x;
```

El resultado de correr dicho programa se explica a continuación. En la quinta fila del código definimos una matriz fila X de 7 ceros. El elemento $X(1)$ es el estimador inicial $x^{(0)}$ y como no hemos puesto "punto y coma" después del comando se muestra en pantalla

$$X = \quad (5.74268825600000, \quad 0, \quad 0, \quad 0, \quad 0, \quad 0, \quad 0)$$

En la siguiente línea del código calcula $x^{(1)} = f(x^{(0)})$, lo pone en X(2) y lo muestra en pantalla

$$X = \quad (5.74268825600000, \quad 3.65255464607588, \quad 0, \quad 0, \quad 0, \quad 0, \quad 0)$$

Hace lo mismo con $x^{(2)} = f(x^{(1)})$ que pone en X(3),

$$X = (5.74268825600000\ ,\ 3.65255464607588\ ,\ 3.68364960220285\ ,\ 0\ ,\ 0\ ,\ 0)$$

En la línea siguiente ya calcula el primer elemento no nulo de la matriz fila de 7 ceros Z, definida en la línea 6, que es el $Z(3)$ usando el proceso de aceleración de Aitken y lo muestra en pantalla

$$Z = \quad (0, \quad 0, \quad 3.68319378319682, \quad 0, \quad 0\ ,\ 0\ ,\ 0)$$

Lo que sucede a partir de ahora es importante. Como hemos puesto punto y coma hemos anulado el eco en pantalla de X y de "denom", y sin embargo muestra en pantalla para los valores de $k = 4, ..., 7$, el vector fila Z en donde va colocando los siguientes elementos de la sucesión de Aitken

$$Z = (0\ ,\ 0\ , 3.68319378319682\ , 3.68318577300271\ ,\quad 0\ ,\ 0\ ,\ 0)$$
$$Z = (0\ ,\ 0\ , 3.68319378319682\ , 3.68318577300271\ , 3.68318577115614,\ 0\ ,\ 0)$$

Y ya da el resultado. En la primera columna representa X y en la segunda Z

$$\begin{array}{lll} x^{(0)} = & 5.74268825600000 & 0 \\ x^{(1)} = & 3.65255464607588 & 0 \\ x^{(2)} = & 3.68364960220285 & z^{(0)} = \quad 3.68319378319682 \\ x^{(3)} = & 3.68317874950901 & z^{(1)} = \quad 3.68318577300271 \\ x^{(4)} = & 3.68318587745245 & z^{(2)} = \quad 3.68318577115614 \\ & 0 & 0 \\ & 0 & 0 \end{array}$$

y también da el número de pasos $k = 5$ que ha dado para obtener las tres iterantes de Aitken así como la medidas de convergencia absoluta $1.84657 \cdot 10^{-9}$ y relativa $5.0135 \cdot 10^{-10}$.
Se deja la aplicación del código a los otros casos del enunciado como ejercicio.

1.6 Denotamos $w^{(n)}$, como en la sección 1.3.2 del resumen teórico, la sucesión definida por el método de Steffensen en la que $w^{(0)} = x^{(0)}$.
Hemos utilizado para estudiar el caso $(\mathcal{R}, \epsilon/D) = (10^4, 0.05)$ el código Matlab del método de Steffensen que se incluye en la página web vinculada al libro. Limitando el número de pasos a dos, el resultado del programa es

$$w^{(1)} = 3.68364960220285 \quad w^{(2)} = 3.68318577299185$$

con $|w^{(2)} - w^{(1)}| = 8.01204 \cdot 10^{-6}$ y $\dfrac{|w^{(2)} - w^{(1)}|}{|w^{(2)}| + 10^{-11}} = 2.1752972 \cdot 10^{-6}$. El término siguiente $w^{(3)}$ vuelve a ser de nuevo $w^{(2)}$.

- $(\mathcal{R}, \epsilon/D) = (10^5, 0.003)$ En este caso,

$$(E) \qquad x = -2\log_{10}(0.000080862 + 0.0000251\ x) = F_2(x)$$

1.1 Estudiemos la función F_2 en su dominio $x > -32.21593625$, aunque sólo consideremos valores de x estrictamente positivos.
En el extremo inferior del dominio $\lim_{x \to -32.21593625} F_2(x) = +\infty$ luego $F_2(x)$ tiene una asíntota vertical de ecuación $x = -32.21593625$.
Derivando,

$$F_2'(x) = -\frac{0.0000218015}{0.013477 + 0.0000251\ x}$$

siempre negativa y F_2 es decreciente en su dominio.
Para $x = -31.34734725$ se tiene $F_2'(x) = -1$ de modo que $|F_2'(x)| < 1$ cualquiera que sea $x \in [-31.34734725, +\infty)$.
La aplicación F_2 satisface en cualquier conjunto cerrado contenido en dicho intervalo el teorema del punto fijo de Banach y sus conclusiones.
La estimación inicial suministrada por la fórmula empírica de Hermann es aquí

$$\lambda^{(0)} = 0.0054 + 0.39510^{-1.5} = 0.017891 \quad \Rightarrow \quad x^{(0)} = 7.476231341$$

Este punto pertenece al intervalo $[-31.34734725, +\infty)$, luego es un estimador inicial válido. Como $F_2(7.476231431) = 7.142061574$, tendremos $F_2(7.476231431) - 7.476231431 =$
$= -0.3341698 < 0$. Por otro lado, $F_2(5) = 7.370740 > 5$ y $F_2(5) - 5 = 2.370740 > 0$, luego el único punto fijo x^* de F_2 está en el intervalo cerrado $I_2 = [5, 7.476231431]$. En la Figura 1.29 se sintetizan los resultados anteriores.

1.2 Una constante de Lipschitz de F_2 en I_2 es

$$L = \max_{x \in I_2} |F_2'(x)| = |F_2'(5)| = 0.023338543$$

Con ello, tomando como estimador inicial $x^{(0)} = 7.476231431$, $x^{(1)} = F_2(x^{(0)}) = 6.003242922$ y el número de iteraciones k necesarias para conseguir que el error $|x^* - x^{(k)}|$ sea menor o igual que $\epsilon = 10^{-6}$ es

$$k \geq \frac{-6.1784865}{-1.63192626} \sim 3.78600 \ \Rightarrow \ k = 4$$

1.3 Usando el código Matlab del método de aproximaciones sucesivas con $x^{(0)} = 7.476231431$, tenemos:

$$\begin{aligned} x^{(1)} &= 6.00324292216879 \\ x^{(2)} &= 6.03608984584707 \\ x^{(3)} &= 6.03534367015773 \\ x^{(4)} &= 6.03536061373841 \\ x^{(5)} &= 6.03536022899301 \end{aligned}$$

que da el valor aproximado $x^* = 6.03536022899301$ luego $\lambda^* = 0.02745323999234$, con $|x^{(5)} - x^{(4)}| =< 10^{-6}$ y $\dfrac{|x^{(5)} - x^{(4)}|}{|x^{(5)}| + 10^{-6}} =< 10^{-7}$.

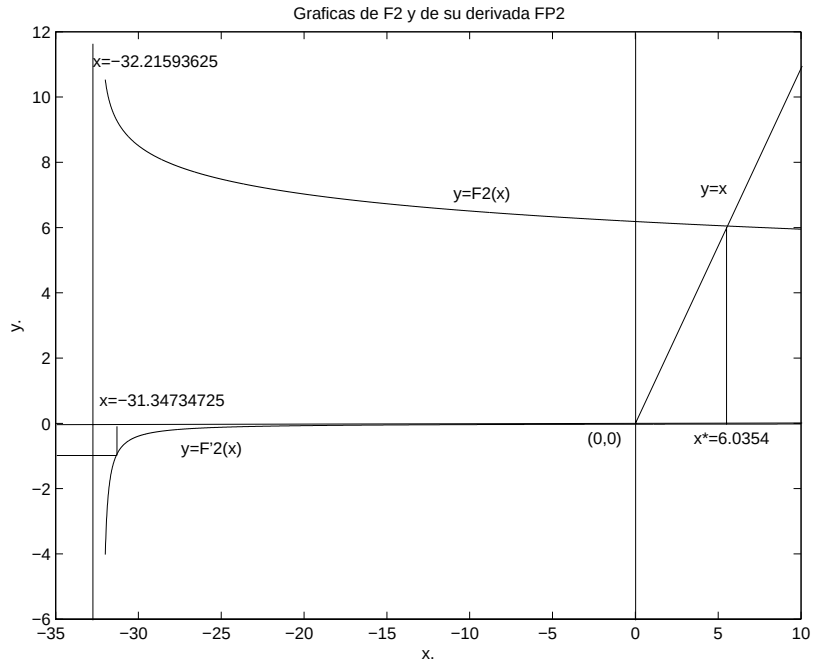


Figura 1.29: Representación gráfica de F_2 y de su derivada F_2'. Se representan la asíntota vertical $x = -32.21593625$ de ambos grafos, el extremo inferior $x = -31.34734725$ del intervalo en el que F_2 es contractiva y el punto fijo $x^* \sim 6.035360614$ de F_2 objetivo del ejercicio.

1.4 La estimación "a priori" del error cometido al tomar $x^{(5)}$ como solución del problema es

$$|x^* - x^{(5)}| \leq 1.044297 \cdot 10^{-8}$$

y la estimación "a posteriori"

$$|x^* - x^{(5)}| \leq 0.9193970 \cdot 10^{-10}$$

En este caso la estimación "a priori" era bastante pesimista.

- $(\mathcal{R}, \epsilon/D) = (10^6, 0.003)$

$$(E) \qquad x = -2\log_{10}(0.000080862 + 0.00000251\ x) = F_3(x)$$

Se deja este último caso como ejercicio.

2. - $(\mathcal{R}, \epsilon/D) = (10^4, 0.05)$
Con el estimador inicial $x^{(0)} = 5.742688256$, épsilon $= 10^{-7}$ y un máximo de 2 iteraciones, el resultado es $x^* \approx 3.68318564986598$, luego $\lambda^* \approx 0.07371447557970$ con un error absoluto $6.064487 \cdot 10^{-8}$ y $F_1(3.68318564986598) \approx 3.68318577299185$.
Si aumentamos el número de iteraciones, manda un mensaje en pantalla de división por cero en la línea 8, es decir, al calcular Δ (véase el código Matlab del método Wegstein en el apartado 3 b) del problema 1.12).
- $(\mathcal{R}, \epsilon/D) = (10^5, 0.003)$
El proceso es aquí una repetición del caso anterior. Con $x^{(0)} = 7.476231431$ y manteniendo épsilon, en tres pasos el resultado fue $x^{(1)} = 6.00324292216879$, $x^{(2)} = 6.03535993971825$ y $x^{(3)} = 6.03536023753561$ con el mismo mensaje de antes. El código da como respuesta $x^* = 6.03536023753561$ luego $\lambda^* \approx 0.02745323991462$ sin haber podido calcular el error.

- $(\mathcal{R}, \epsilon/D) = (10^6, 0.003)$
  De nuevo el proceso es un calco de los casos anteriores. Con $x^{(0)} = 9.260717036$, épsilon $= 10^{-7}$ y 2 iteraciones el resultado es $x^* = 6.16803829943892$, luego $\lambda^* \approx 0.02628487260934$ con un error $1.3303 \cdot 10^{-10}$ y $F_3(6.16803829943892) = 6.16803829970567$.

3. De acuerdo con las reglas prácticas que se deben observar para aplicar eficientemente el método de Newton en el intervalo $[a, b]$ (sección 1.3.3), $f(a)f(b) < 0$, f'' debe tener signo constante en ese intervalo y se aplica el método al extremo del intervalo en el que f y f'' tienen el mismo signo.

   - $(\mathcal{R}, \epsilon/D) = (10^4, 0.05)$

     3.1 En el apartado 1.1 probamos que la raíz única de la ecuación $f_1(x) = F_1(x) - x = 0$, estaba en el intervalo $I_1 = [3, 5.742688256]$ y que $f_1(5.742688256) = -2.09013361 < 0$ y $f_1(3) = 0.6793602 > 0$. De otro lado,

     $$f_1'(x) = -\frac{0.000218015}{0.013477 + 0.000251\ x} - 1 \quad \text{y} \quad f_1''(x) = \frac{5.4721765 \cdot 10^{-8}}{(0.013477 + 0.000251 \cdot x)^2} > 0$$

     El extremo del intervalo en el que f_1 y f_1'' tienen el mismo signo es 3 que tomaremos como primer estimador inicial.

     3.2 Aplicamos la iteración de punto fijo de Newton Raphson con el estimador inicial $x^{(0)} = 3$, tolerancia para la raíz, delta $= 10^{-6}$, tolerancia para los valores de la función f_1, épsilon $= 10^{-10}$ y máximo número de iteraciones 10. Se obtiene el resultado $x^* \approx 3.68318577115872$ ($\lambda^* \approx 0.07371447072465$) en dos pasos con un valor de la función $|f_1(x^{(2)})| = 0.05 \cdot 10^{-12} < 10^{-10}$ razón por la que el programa se paró cuando todavía la tolerancia para la raíz era inaceptable.
     Corrimos de nuevo el programa después de eliminar ese control del código, manteniendo el resto de datos y controles. El resultado fue $x^* = 3.68318577115572$ ($\lambda^* = 0.07371447072477$) en tres pasos

     $$\begin{aligned} x^{(0)} &= 3.00000000000000 \\ x^{(1)} &= 3.68312418997359 \\ x^{(2)} &= 3.68318577115872 \\ x^{(3)} &= 3.68318577115572 \end{aligned}$$

     con $|x^{(3)} - x^{(2)}| = 3.00 \cdot 10^{-12}$, $2\dfrac{|x^{(3)} - x^{(2)}|}{|x^{(3)}| + 10^{-6}} = 1.63 \cdot 10^{-12}$ y $|f_1(x^{(3)})| = 0$.

     3.3 Aunque no parece que en este ejemplo el cálculo de la derivada f_1' sea particularmente costoso numéricamente, hemos querido ensayar el método de von Mises en el que se evalúa f_1' para un buen valor inicial $x^{(0)}$, valor que se mantiene fijo en las iteraciones siguientes.
     El resultado es el esperado aunque hay que saber "leerlo"

     $$\begin{aligned} x^{(0)} &= 3.00000000000000 \\ x^{(1)} &= 3.68312418997359 \\ x^{(2)} &= 3.68318576009521 \\ x^{(3)} &= 3.68318577115373 \end{aligned}$$

     en 3 iteraciones se llega a $x^* \approx x_3 = 3.68318577115373$ con $|x^{(3)} - x^{(2)}| = 1.105852 \cdot 10^{-8}$, $2\dfrac{|x^{(3)} - x^{(2)}|}{|x^{(3)}| + 10^{-9}} = 6.00487 \cdot 10^{-9}$ y un valor de la función $|f_1(x^{(3)})| > 2.02 \cdot 10^{-12}$.
     La primera iterante coincide lógicamente con la de Newton-Raphson, pero en las sucesivas se va perdiendo paulatinamente precisión como lo prueban los valores de los errores absoluto y relativo y el valor de la función para el mismo número de iteraciones.

   - $(\mathcal{R}, \epsilon/D) = (10^5, 0.003)$
     Se deja como ejercicio.
   - $(\mathcal{R}, \epsilon/D) = (10^6, 0.003)$
     Se deja como ejercicio.

4. Por último consideramos en este apartado el caso $(\mathcal{R}, \epsilon/D) = (10^5, 0.003)$.

   La ecuación a resolver es

   $$f_2(x) = F_2(x) - x = -2\log_{10}(0.000080862 + 0.0000251\ x) - x = 0$$

   cuya raíz única está en el intervalo $I_2 = [5, 7.476231431]$ según probamos en el apartado 1.1 relativo a este caso.

   Usaremos los métodos de interpolación lineal Illinois y Pegasus cuyos códigos Matlab se incluyen en el problema 1.13.

   4.1 Con los controles delta= 10^{-9}, épsilon= 10^{-8}, y un número máximo de iteraciones max1= 30, el código Illinois da en 18 iteraciones como resultado $x^* \approx 6.03536023261592$ con un error $7.37948 \cdot 10^{-9}$ que mide la mitad de la longitud del último intervalo de búsqueda y un valor de la función $x \to F_2(x) - x$ igual a $.0314 \cdot 10^{-9}$.

   4.2 Con los mismos datos y controles que en el caso anterior el código Pegasus no presenta ventajas en este caso.

   En 17 iteraciones se consigue el resultado $x^* \approx 6.03536024632427$ con un error $1.318312 \cdot 10^{-8}$ y con un valor de la función $x \to F_2(x) - x$ igual a $-8.98822 \cdot 10^{-9}$. Una explicación del comportamiento similar entre los dos métodos está en los valores que el factor alpha de convergencia toma en el código Pegasus. Salvo en el primer paso en el que alpha = 0.55330220043780, en todos los demás pasos alpha coincide con 0.5 al menos en cuatro dígitos siendo el valor en el último paso 0.50000000109212.

PROBLEMA 1.12 *Coeficiente de empuje para ángulo de astilla muerta cero.* [28]

Método de Savitsky

En el cálculo de la resistencia hidrodinámica de lanchas planeadoras por el método de Savitsky [25] [29], es crucial la estimación del ángulo de inclinación del barco en la dirección de la marcha τ. Esto es así porque existen fórmulas empíricas que relacionan τ con todas las variables que intervienen en el diseño [30].
La ecuación (E), base del método y cuya resolución es el objetivo del correspondiente algoritmo, es

$$(E) \qquad \Delta\left\{\frac{[1 - \sin\tau\sin(\tau+\epsilon)]c}{\cos\tau} - f\sin\tau\right\} + D_f(a - f) = 0$$

Las variables c, D_f dependen de τ, dependencia que no es expresable mediante funciones elementales pero que se puede describir a través de las siguientes relaciones empíricas.

$$(c) \qquad c = LCG - C_p\lambda b$$

$$(D_f) \qquad D_f = \frac{\rho V M^2 \lambda b^2 (C_F + \Delta C_F)}{\cos\beta}$$

donde

$$C_p = 0.75 - \frac{1}{5.21\frac{C_V^2}{\lambda^2} + 2.39}$$

$$C_V = \frac{V}{\sqrt{gb}}$$

y λ y C_F se obtienen mediante las ecuaciones $(E1)$, $(E2)$ y $(E3)$.

[28] La astilla muerta es un ángulo que mide a "grosso modo" lo que se desvía la sección transversal de un buque de un rectángulo. Si el fondo es plano, como en la planeadora casera de la figura 1.31, la astilla muerta vale cero.

[29] Dan Savitsky, que fue director del Davidson Laboratory, es en la actualidad consultor y profesor emérito en "The Center for Maritime Systems" del "Steven's Institute".

[30] Ver detrás una lista con las definiciones de todas las variables que intervienen en el algoritmo y la Figura 1.32 con su significado físico.

- $(E1)$ Coeficiente de empuje para ángulo de astilla muerta cero.

$$(E1) \qquad C_{L0} - 0.0065\beta\,(C_{L0})^{0.6} - C_{L\beta} = 0 \quad \Rightarrow \quad C_{L0}$$

donde $C_{L\beta}$ y C_{L0} son coeficientes de empuje adimensionalizados relativos a un ángulo de astilla muerta igual a β y 0 respectivamente y donde

$$C_{L\beta} = \frac{\Delta}{\frac{1}{2}\rho V^2 b^2}$$

- $(E2)$ Línea de fricción de Schoenherr para flujo turbulento.

$$(E2) \qquad F(\mathcal{R}, C_F) = \frac{0.242}{\sqrt{C_F}} - \log_{10}(\mathcal{R} \cdot C_F) \quad \Rightarrow \quad C_F$$

donde $\mathcal{R}$ es el número de Reynolds.

- $(E3)$ Relación entre la eslora mojada y la manga.

$$(E3) \quad \frac{\tau^{1.1}}{C_{L0}}\left(0.012\sqrt{\lambda} + 0.0055\frac{\lambda^{0.25}}{C_V^2}\right) - 1 = 0 \quad \Rightarrow \quad \lambda$$

Existen además una ecuación que relaciona el número de Reynolds $\mathcal{R}$ con las otras variables

$$\mathcal{R} = \frac{V_M \lambda b}{\nu}$$

y una última ecuación $(E4)$ que relaciona V_M con V

$$(E4) \qquad V_M = kV$$

donde $k \in [0.8, 1.0]$ se obtiene mediante unas curvas en función de τ, β y λ.

El resto de las variables se incluye a continuación con una pequeña definición que permita situarse correctamente en el problema global y una figura donde se aclara su significado físico.

Δ Peso del buque (libras)

D_f Componente viscosa de la resistencia (libras). Se supone que actúa paralelamente a la línea de la quilla

τ Ángulo de trimado (grados)

ϵ Inclinación de la línea de empuje relativa a la quilla (grados)

CG Centro de gravedad

LCG Distancia longitudinal del centro de gravedad desde la popa medida a lo largo de la quilla (pies)

a Distancia entre D_f y CG, medida normalmente a D_f (pies)

T Empuje del propulsor (libras)

N Resultante de las fuerzas de presión actuando normalmente a la base (libras)

f Distancia entre T y CG medida normalmente a la línea de ejes (pies)

c Distancia entre N y CG medida normalmente a N (pies)

β Ángulo transversal a que da lugar la astilla muerta (grados)

b Manga (pies)

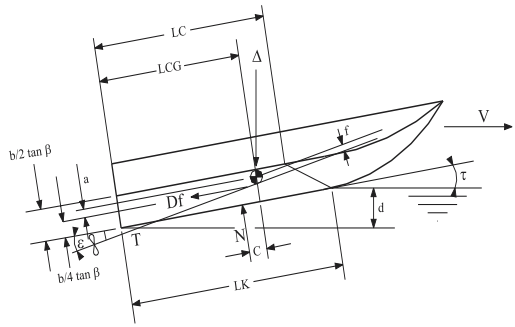


Figura 1.30: Principales características de una planeadora.

Figura 1.31: Planeadora casera de astilla muerta cero.

V Velocidad (pies/segundo)

d Calado de la quilla en la popa (pies)

El objetivo del algoritmo es resolver (E) previo análisis de la existencia de solución y debe incluir subrutinas de resolución de las ecuaciones $(E1)$, $(E2)$ y $(E3)$, previa demostración de la existencia y unicidad de sus respectivas soluciones. Nosotros propondremos aquí la resolución de la ecuación $(E1)$ con el análisis previo de la existencia de solución y en el problema 1.13 la correspondiente de la ecuación $(E2)$.

Enunciado:

En un momento del proceso numérico asociado al método de Savitsky se debe resolver la ecuación

$$(E1) \qquad C_{L0} - 0.0065\beta \left(C_{L0}\right)^{0.6} - C_{L\beta} = 0$$

donde $C_{L0} > 0$ es la incógnita y $C_{L\beta} > 0$ y $0° < \beta < 20°$, como hemos comentado.
El objetivo de este ejercicio es analizar la existencia y unicidad de la solución de $(E1)$ determinándola después por varios métodos cuyos comportamientos se compararán.

1. Buscar una escritura conveniente de $(E1)$ del tipo

   $$C_{L0} = T(C_{L0})$$

   que permita demostrar la existencia y unicidad de la solución de $(E1)$ para valores estrictamente positivos de $C_{L\beta}$ y β.

2. ¿Para qué valores de C_{L0} será necesariamente convergente el método de las aproximaciones sucesivas asociado a la ecuación de punto fijo de 1.?

3. Se supone que $C_{L\beta} = 0.07$ y $\beta = 20°$

   a) Resolver $(E1)$ por el método de las aproximaciones sucesivas seleccionando un estimador inicial que cumpla con los valores obtenidos en el apartado 2 con test de parada $\epsilon < 10^{-7}$.

   b) Usar el método de Wegstein con el mismo estimador inicial y test de parada y comparar el número de iteraciones.

   c) Determinar un intervalo que contenga a la raíz de $(E1)$ y usar para determinarla

      - el método Illinois.
      - el método Pegasus.

   d) Comprobar el mejor comportamiento del segundo método respecto al primero y compararlo con el de Wegstein.

Solución:

1. Escribimos $(E1)$ según sugiere el enunciado

$$C_{L0} = C_{L\beta} + 0.0065\beta\,(C_{L0})^{0.6} = T(C_{L0})$$

ecuación que podemos expresar de forma más cómoda

$$x = \alpha + k_1\beta x^{\frac{3}{5}} = T_{\alpha,\beta}(x)$$

con $C_{L0} = x > 0$, $C_{L\beta} = \alpha > 0$, $\beta > 0$ y $k_1 = 0.0065$.

Estudiemos la familia biparamétrica de funciones $T_{\alpha,\beta}$.

Comencemos con las intersecciones con el eje x

$$T_{\alpha,\beta}(x) = 0 \quad \Rightarrow \quad \alpha + k_1\beta x^{\frac{3}{5}} = 0 \quad \Rightarrow \quad x = \left(-\frac{\alpha}{k_1\beta}\right)^{\frac{5}{3}}$$

Este valor de x es estrictamente negativo para $\alpha, \beta, k_1 \in I\!R_+$. Por tanto, ningún miembro de la familia se anula en $I\!R_+$ y $T_{\alpha,\beta}(0) = \alpha$. Además,

$$\lim_{x\to+\infty} T_{\alpha,\beta}(x) = +\infty$$

Derivando

$$T'_{\alpha,\beta}(x) = \frac{3}{5}k_1\beta x^{-\frac{2}{5}} = 0.0039\ \beta\ \frac{1}{x^{\frac{2}{5}}}$$

función que es siempre estrictamente positiva y

$$\lim_{x\to0^+} T'_{\alpha,\beta}(x) = +\infty$$

Derivando de nuevo

$$T''_{\alpha,\beta}(x) = -0.00156\ \beta\ \frac{1}{x^{\frac{7}{5}}}$$

estrictamente negativa para todo valor de x.

Todos los miembros de la familia son funciones estrictamente crecientes y cóncavas hacia abajo, que arrancan del punto $(0, \alpha)$ con tangente vertical.

En la Figura 1.32 representamos gráficamente el miembro de la familia que es objeto de estudio en el apartado 3. También hemos representado en ella la recta $y = x$ (obsérvese la diferencia de escala en los dos ejes) y se hace evidente la existencia y unicidad de la raíz x^* abscisa del punto de intersección de las gráficas de la recta diagonal $y = x$ y de $T_{0.07,20}$. Dicha gráfica sería prácticamente igual cualesquiera que fueran α y $\beta > 0$.

2. Para demostrar la existencia y unicidad de la raíz x^* analizamos para qué valores de x se verifica la condición suficiente de convergencia del método de aproximaciones sucesivas (teorema 1.3.1).

$$\left|T'_{\alpha,\beta}(x)\right| = 0.0039\beta\frac{1}{x^{\frac{2}{5}}} < 1$$

luego

$$0.0039\beta < \frac{1}{x^{\frac{2}{5}}} \quad \Rightarrow \quad (0.0039\beta)^{\frac{5}{2}} < x \quad \Rightarrow \quad 0.949864195\cdot10^{-6}\beta^{\frac{5}{2}} < x$$

En el intervalo $I = [0.949864195\cdot10^{-6}\beta^{\frac{5}{2}}, +\infty)$, la función $T_{\alpha,\beta}$ es contractiva de modo que el teorema de Banach del punto fijo asegura la existencia y unicidad del punto fijo $x^*_{\alpha,\beta}$ de $T_{\alpha,\beta}$ en cualquier conjunto cerrado contenido en I y asegura también la convergencia de la sucesión $x^{(k)} = T_{\alpha,\beta}(x^{(k-1)})$ a dicho punto fijo cualquiera que sea el estimador inicial $x^{(0)}$ elegido en ese cerrado.

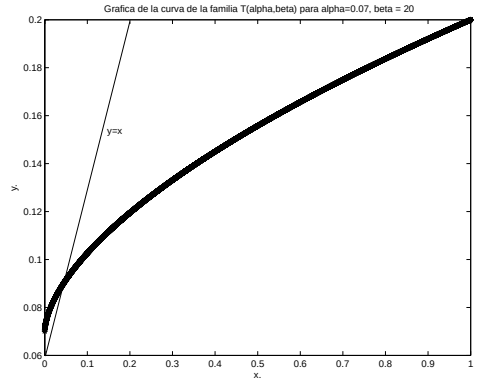


Figura 1.32: Representación gráfica de la igualdad $x = T_{\alpha,\beta}(x)$ para el miembro de la familia relativo a los valores $(\alpha, \beta) = (0.07, 20)$.

3. La sucesión del método de aproximaciones sucesivas es aquí

$$x^{(k+1)} = T_{0.07,20}(x^{(k)}) = 0.07 + (0.13)(x^{(k)})^{\frac{3}{5}}$$

Con $\beta = 20$, $0.949864195 \cdot 10^{-6} 20^{\frac{5}{2}} = 1.699169 \cdot 10^{-3}$, luego debemos tomar un estimador inicial $x^{(0)} > 1.699169 \cdot 10^{-3}$, por ejemplo, $x^{(0)} = 0.02 > 0.0017$.

a) Una vez definida la sucesión aproximante usaremos primero en nuestros cálculos una calculadora normal; se obtiene así

$$\begin{array}{cccc} & x_1 = 0.082432585 & x_2 = 0.099080555 & x_3 = 0.102474046 \quad ... \\ ... & x_{10} = 0.103296124 & x_{11} = 0.103296130 & x_{12} = 0.103296132 \end{array}$$

Usemos ahora el código Matlab del método de aproximaciones sucesivas *mas.m* que llama a la función *Savitsky1.*

Dicho código utiliza dos medidas de la convergencia, una absoluta $|x^{(k)} - x^{(k-1)}|$ y otra relativa $\dfrac{|x^{(k)} - x^{(k-1)}|}{|x^{(k)}| + \epsilon}$ donde ϵ es la tolerancia impuesta en el código. Con una tolerancia $\epsilon = 10^{-7}$ y limitando a 10 el número de iteraciones max1= 10, la primera vez que lo hemos corrido, el programa se paró en $x^{(10)} = 0.10329608920873$ por exceso en el número de iteraciones dando el correspondiente mensaje, con $|x^{(10)} - x^{(9)}| = 1.7977867 \cdot 10^{-7}$ y $\dfrac{|x^{(10)} - x^{(9)}|}{|x^{(10)}| + 10^{-7}} = 1.74041922 \cdot 10^{-6}$, inaceptable.

Aumentando el número máximo de iteraciones obtenemos en 11 iteraciones

$$\begin{array}{llllll} x^{(1)} = & 0.08243258249727 & x^{(2)} = & 0.09908055534169 & x^{(3)} = & 0.10247404675137 \\ x^{(4)} = & 0.10313688530197 & x^{(5)} = & 0.10313688530197 & x^{(6)} = & 0.10326532412006 \\ x^{(7)} = & 0.10329017359292 & x^{(8)} = & 0.10329497987304 & x^{(9)} = & 0.10329590943006 \\ x^{(10)} = & 0.10329608920873 & x^{(11)} = & 0.10329612397831 & & \end{array}$$

El valor aproximado de la raíz es ahora $C^*_{L0} = x^* = 0.10329612397831$ con $|x^{(11)} - x^{(10)}| = 3.476957 \cdot 10^{-8} < 10^{-7}$ y $\dfrac{|x^{(11)} - x^{(10)}|}{|x^{(11)}| + 10^{-7}} = 3.3660063 \cdot 10^{-7}$.

b) Con el mismo estimador inicial aceleraremos la convergencia por el método de Wegstein utilizando el programa Matlab *Wegstein.m* que incluimos en la página web vinculada al libro. Dicho código llama a la función Savitsky1 que se escribe en el programa *SavitskyA11.m*

En sólo tres iteraciones aproximamos la raíz con el error exigido.

Los resultados parciales que se obtienen son

Iteración	x1	y1
1	0.0824325824972	0.09908055534169
2	0.10365032684131	0.10336458734513
3	0.10329619013218	0.10329614349710

donde x1 e y1 tienen el significado que se define en el código y el resultado final es

error	$C^*_{L0} = x^*$	$T_{0.07,20}(x^*)$
$2.890849 \cdot 10^{-8}$	0.10329619013218	0.10329614349710

c) Por último usaremos los métodos Illinois y Pegasus que resuelven la ecuación

$$f_{\alpha,\beta}(x) = T_{\alpha,\beta}(x) - x = 0.07 + (0.13)x^{\frac{3}{5}} - x = 0$$

En ambos casos tomamos como intervalo que enmarca a la raíz x^* el $(0.02, 1)$ y las tolerancias impuestas han sido[31] usando la notación de los códigos Matlab, *illinois.m* y *pegasus.m*, delta $= 10^{-7}$ para la raíz y epsilon $= 10^{-6}$ para el valor de $f_{\alpha,\beta}$ limitando el número de iteraciones a 30 (max1 = 30).

- El programa *illinois.m* se paró en la iteración 17 porque el error en el valor de la función epsilon $= 5.0820698 \cdot 10^{-7} < 10^{-6}$ aunque todavía delta $= 9.4507571 \cdot 10^{-7} > 10^{-7}$ siendo el resultado $C^*_{L0} = x^* = 0.10329550225270$.
  Disminuyendo ambas tolerancias delta $= 10^{-9}$ para la raíz y epsilon $= 10^{-8}$ para el valor de la función, obtenemos en 23 pasos $C^*_{L0} = x^* = 0.10329612247055$ con epsilon $= 7.94066 \cdot 10^{-9} < 10^{-8}$ y delta $= 1.476694 \cdot 10^{-8} > 10^{-9}$.
- El programa *pegasus.m* se paró en el paso 6 porque el error en el valor de la función epsilon $= 5.0232315 \cdot 10^{-7} < 10^{-6}$ aunque todavía delta $= 1.14669613 \cdot 10^{-6} > 10^{-7}$ siendo el resultado $C^*_{L0} = x^* = 0.10329550954734$.
  Disminuyendo ambas tolerancias delta $= 10^{-9}$ para la raíz y epsilon $= 10^{-8}$ para el valor de la función, obtenemos $C^*_{L0} = x^* = 0.10329612169824$ en 12 pasos con epsilon $= 8.56361 \cdot 10^{-9} < 10^{-8}$ y delta $= 1.597167 \cdot 10^{-8} > 10^{-9}$.

Está claro que el método Pegasus es más rápido que el método Illinois. La causa está en el factor alpha, que es variable en el primero y 0.5 fijo en el segundo. En el programa *pegasus.m* incluimos en los resultados los valores variables de alpha que son en los últimos pasos muy cercanos a 0.5.

d) Ni Pegasus ni Illinois son comparables al método del promotor de convergencia de Wegstein en cuanto a velocidad de convergencia. Este método requiere además menor precisión en la definición del estimador inicial.

PROBLEMA 1.13 *Línea de fricción de Schoenherr para flujo turbulento.*

En un momento del proceso numérico asociado al método de Savitsky para el diseño hidrodinámico de lanchas planeadoras, aparece la ecuación

$$\frac{0.242}{\sqrt{C_F}} - \log_{10}(\mathcal{R} \cdot C_F) = 0 \qquad (E2)$$

donde $\mathcal{R}$ es el número de Reynolds, un dato que es una medida de la velocidad, y la incógnita C_F mide la resistencia al avance de la planeadora a esa velocidad.

1. Reescribir la ecuación $(E2)$ en la forma

$$x = f_\lambda(x) \qquad (\forall \lambda \in I) \qquad (E2)'$$

donde x será una incógnita estrictamente positiva y $\lambda = \sqrt{\mathcal{R}}$ aparece como un parámetro. Supondremos que $10^4 < \mathcal{R} < 10^8$ por lo que $I = \left(10^2, 10^4\right)$.

[31] Ver la sección 1.3.4 del resumen teórico.

2. Estudiar la familia uniparamétrica de funciones $(f_\lambda)_{\lambda\in I}$ y deducir del estudio la posible existencia y unicidad de las raíces de $(E2)'$.

3. Determinar un intervalo (k,∞) independiente de λ, en el que $(E2)'$ defina por el método de las aproximaciones sucesivas una sucesión convergente independientemente de la selección del estimador inicial.

4. Utilizar el método de las aproximaciones sucesivas con un estimador inicial adecuado para determinar aproximadamente la raíz x^* de $(E2)'$ para $\lambda = 10^2$ con un tolerancia de 10^{-7}.

5. Comprobar que $\dfrac{1}{\mathcal{R}} < C_F^* < 1$.

6. Usar el método Pegasus con el intervalo determinado en el apartado anterior para aproximar C_F^* cuando $\mathcal{R} = 10^4$.

7. Usar el método de Wegstein con un estimador inicial adecuado para aproximar C_F^* cuando $\mathcal{R} = 10^8$.

Solución:

1. Multiplicando los dos miembros de $(E2)$ por $-\dfrac{1}{2}$

$$-\frac{0.242}{2\sqrt{C_F}} = -\frac{1}{2}\log_{10}(\mathcal{R}\cdot C_F) = \log_{10}\left(\frac{1}{\sqrt{\mathcal{R}\cdot C_F}}\right)$$

y tomando como nueva variable $x = \dfrac{1}{\sqrt{C_F}} > 0$ obtenemos

$$-0.121x = \log_{10}\left(\frac{x}{\sqrt{\mathcal{R}}}\right) \Rightarrow \quad x = -8.264463\,\log_{10}\left(\frac{x}{\sqrt{\mathcal{R}}}\right)$$

luego

$$(E2)' \quad x = f_\lambda(x) = k_1\,\log_{10}\left(\frac{x}{\lambda}\right) = k_2\,\ln\left(\frac{x}{\lambda}\right)$$

con $k_1 = -8.264463$ y $k_2 = \dfrac{k_1}{\ln 10} = -3.589211$.

2. El dominio de definición de f_λ es $(0^+,\infty)$ y se anula para $x = \lambda$.

   Derivando, $f'_\lambda(x) = k_2\dfrac{1}{x} = -3.589211\dfrac{1}{x}$ que es siempre negativa independientemente de λ.

   Además

$$\lim_{x\to 0^+} f_\lambda(x) = +\infty; \quad \lim_{x\to 0^+} f'_\lambda(x) = -\infty$$

   Se obtiene la gráfica de f_λ (ver Figura 1.33) que ilustra la existencia y unicidad de la raíz x^* abscisa del punto de intersección de las gráficas de la recta diagonal $y = x$ y de f_λ cualquiera que sea λ, luego también abscisa de la única raíz C_F^* de la ecuación $(E2)$.

3. $|f'_\lambda| < 1$ cuando $x > |k_2|$, de modo que en cualquier conjunto cerrado contenido en el intervalo $[|k_2|,\infty)$ el método de aproximaciones sucesivas converge al punto fijo único x^* de f_λ cualquiera que sea el estimador inicial que seleccionemos, siendo esa convergencia más rápida cuanto mayor es el valor del estimador que tomemos.

4. Utilizando el método de las aproximaciones sucesivas con un estimador inicial $x^{(0)} = 5 > 3.589211$, obtenemos $x^* = 8.74555809784666$, es decir, $C_F^* = \dfrac{1}{(x^*)^2} = 0.01307449555014$ en 20 iteraciones con

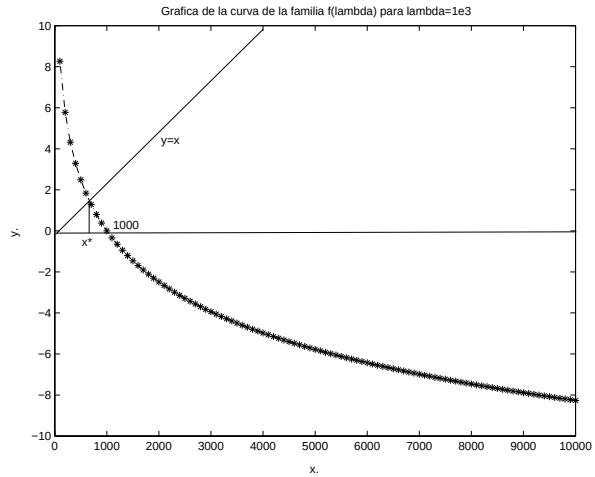


Figura 1.33: Representación gráfica de la curva de la familia f_λ relativa al valor $\lambda = 1000$. Se representa también la primera bisectriz $y = x$ y la raíz x^* de la ecuación $(E2)'$.

las medidas absoluta y relativa de convergencia $|x^{(20)} - x^{(19)}| = 6.9886349 \cdot 10^{-7}$ y $\dfrac{|x^{(20)} - x^{(19)}|}{|x^{(20)}| + 10^{-7}} = 7.991068 \cdot 10^{-8}$ respectivamente.

Hemos ensayado también el estimador inicial $x^{(0)} = 1 < 3.589211$, es decir, fuera del intervalo en el que f_λ es contractiva, y obtuvimos $x^* = 8.74555800820974$ luego $C_F^* = 0.01307449581815$ en 22 iteraciones con$|x^{(22)} - x^{(21)}| = 3.9081508 \cdot 10^{-7}$ y $\dfrac{|x^{(22)} - x^{(21)}|}{|x^{(22)}| + 10^{-7}} = 4.468727 \cdot 10^{-8}$.

Se observa gráficamente en la Figura 1.33 que la primera iterante $x^{(1)} \approx 16.52893 \in [|k_2|, \infty)$ lo que justifica el resultado.

5. Llamemos

$$F(\mathcal{R}, C_F) = \frac{0.242}{\sqrt{C_F}} - \log_{10}(\mathcal{R} \cdot C_F)$$

cuando $C_F = \dfrac{1}{\mathcal{R}}$, $F(\mathcal{R}, C_F) = 0.242\sqrt{\mathcal{R}} > 0$, bastaría encontrar otro valor de C_F de modo que $F(\mathcal{R}, C_F) < 0$ para tener un intervalo que contenga a la raíz C_F^* de $(E2)$.

Para $C_F = 1$, $F(\mathcal{R}, 1) = 0.242 \leq \log_{10}(\mathcal{R})$ que se anula cuando $\mathcal{R} = 10^{0.242} \sim 1.745822$ luego para $\mathcal{R} > 1.745822$, lo que siempre sucede en nuestro estudio, se tiene $F(\mathcal{R}, 1) < 0$, y $\dfrac{1}{\mathcal{R}} < C_F^* < 1$.

6. Hemos corrido los programas Matlab de los métodos Illinois y Pegasus en el intervalo determinado en el apartado 5 y el resultado con el método Pegasus, tras 20 iteraciones, es

error	$F(\mathcal{R}, C_F^*)$	CF^*
$4.354059 \cdot 10^{-8}$	$3.31271303 \cdot 10^{-6}$	0.01307446756497

Los códigos usados para correr el método Pegasus *pegasus.m* se incluyen con líneas de comentarios en la página web vinculada al libro el caso del método Illinois es suficiente sustituir alpha= 0.5.

7. El método de Wegstein permite una convergencia muy rápida que no depende tan crucialmente del estimador inicial elegido como los métodos Illinois o Pegasus.

Tomando como estimador inicial $x^{(0)} = 3.7$, epsilon $= 10^{-7}$, max1 $= 10$, hemos obtenido para $\mathcal{R} = 10^4$ en cuatro iteraciones los valores siguientes

Iteración	$x1$	$y1$
1	11.83304387491211	7.66035988851281
2	8.61273780814712	8.80048593398860
3	8.74526232123806	8.74567920094078
4	8.74555789303570	8.74555789508560

y el resultado es

error	$x*$	$f_\lambda(x*)$	CF^*
$7.2670 \cdot 10^{-10}$	8.74555789303570	8.74555789508560	0.01307449616252

Para $\mathcal{R} = 10^8$, con $x^{(0)} = 5$, la misma tolerancia e igual límite del número de iteraciones,

Iteración	$x1$	$y1$
1	27.28124026105513	21.19120618698574
2	21.88303495099136	21.98258270395608
3	21.96855586507904	21.96858307651373
4	21.96857925506979	21.96857925507183

y el resultado es

error	$x*$	$f_\lambda(x*)$	CF^*
$8.7 \cdot 10^{-13}$	21.96857925507183	21.96857925506979	0.00207203008767

Este método es rapidísimo. Lo hemos usado en varios contextos y el resultado ha sido siempre espectacular. Se exige un conocimiento razonable de la localización de la raíz que suele ser consecuencia del estudio de un método iterativo, casi siempre el de aproximaciones sucesivas, cuya convergencia se desea acelerar.

CAPÍTULO 2

Resolución de sistemas lineales

En la etapa final de la resolución numérica del modelo matemático de un problema físico, una vez terminado el proceso de discretización en el que se sustituye el modelo continuo por una versión en dimensión finita, se tiene que resolver un sistema de ecuaciones que en la mayoría de los casos son lineales.

Ejemplos típicos son la interpolación y aproximación con familias lineales de funciones y la resolución de ecuaciones diferenciales en derivadas parciales por métodos en diferencias en cuyo caso los sistemas suelen tener una estructura que los hace especialmente interesantes.

Por esta razón, es importante disponer de métodos eficientes para resolver los problemas asociados a sistemas lineales.

Los dos problemas básicos que se presentan son: la busca de la solución de un sistema $Ax = b$ de igual número de ecuaciones que de incógnitas y el cálculo de valores-vectores propios de una matriz cuadrada [1].

Dentro de las técnicas para resolver sistemas lineales, se distinguen dos grandes familias de métodos, los **métodos directos** y los **métodos iterativos**, basados en dos filosofías diferentes de abordar el problema. En los métodos directos se obtiene la solución mediante un número finito de operaciones y esta solución sería exacta si las operaciones pudieran efectuarse con aritmética infinita. En los métodos iterativos se construye mediante una relación de recurrencia una sucesión infinita, que a partir de una estimación inicial y bajo ciertas condiciones converge a la solución buscada.

Un elemento importante en el estudio de los sistemas lineales es su **número de condición** o **condicionamiento**. Este valor es una medida de la influencia de las perturbaciones en los datos iniciales, errores en los coeficientes de la matriz A y en las componentes del vector b constante, en la solución del sistema lineal, y está asociado al concepto de **estabilidad**.

Para el estudio de estos métodos recomendamos como referencias [19] y los textos de G. Golub, en particular el [11].

Por último un comentario que pensamos oportuno: Matlab es una herramienta espectacular en la resolución de problemas numéricos lineales pero ello no excusa conocer los algoritmos más importantes sobre los que basa sus códigos así como sus limitaciones.

2.1. Complementos de álgebra y análisis matricial

En este tema se utiliza un lenguaje de álgebra y análisis lineales muy preciso que en algunos casos sobrepasa el nivel habitual del alumno, por lo que es necesario recordar conceptos básicos que hemos estudiado, complementándolos con aquellos que sean imprescindibles para entender los problemas que vamos a resolver.

2.1.1. Matrices

Denotaremos $M_{m\times n}(\mathbb{K})$, el espacio vectorial de las matrices $m \times n$ (m-filas, n-columnas) de coeficientes en el cuerpo $\mathbb{K}$ con $\mathbb{K} = \mathbb{R}$ o $\mathbb{C}$, aunque casi siempre los coeficientes serán reales. Denotaremos por

[1] El caso en que el número de ecuaciones es superior al de incógnitas conduce a la busca de la solución en el sentido de los mínimos cuadrados. Se busca un vector x que haga mínima la suma de cuadrados $\sum_{i=1,n} (b_i - (Ax)_i)^2$.

A^T la matriz $n \times m$ traspuesta de A, A^* su matriz **conjugada** y por A^+ su matriz **adjunta** (traspuesta de la conjugada). Si $A = (a_{ij})$, $A^T = (a_{ji})$, $A^* = (\overline{a_{ij}})$ y $A^+ = (A^*)^T = (\overline{a_{ji}})$.

Es evidente que esos conceptos coinciden si los elementos de A son reales.

Definición 2.1.1 *La matriz A se llama **hermítica** si $A = A^+$ y **simétrica** si $A = A^T$.*

Si $A \in M_{n\times n}(\mathbb{C})$, AA^+ es hermítica.

Definición 2.1.2 *Se dice que $A \in M_{n\times n}(\mathbb{K})$ es **unitaria** si $A^{-1} = A^+$.*
*Si $\mathbb{K} = \mathbb{R}$, $A^T = A^+$ y A se llama **ortogonal**.*

Definición 2.1.3 *La matriz $A \in M_{n\times n}(\mathbb{K})$ se llama **normal** si $A^+A = AA^+$*

Las matrices hermíticas, unitarias y sus correspondientes simétricas y ortogonales en el caso real son normales.

Supondremos en lo que sigue que A es una matriz cuadrada $n \times n$.

Teorema 2.1.1 *Para $A \in M_{n\times n}(\mathbb{K})$, las siguientes afirmaciones son equivalentes:*

1. *A^{-1} existe.*
2. *$\det A \neq 0$.*
3. *El sistema lineal $A\mathbf{x} = \mathbf{0}$ tiene solamente la solución $\mathbf{x} = \mathbf{0}$.*
4. *Para cualquier vector $\mathbf{b}$, el sistema lineal $A\mathbf{x} = \mathbf{b}$ tiene solución única.*
5. *Las filas y columnas de A son linealmente independientes.*
6. *El rango de la matriz A es n.*

Definición 2.1.4 *La matriz A es de **diagonal dominante** si*

$$|a_{ii}| \geq \sum_{j=1, j\neq i}^{j=n} |a_{ij}|, \quad i = 1, \cdots n \tag{2.1}$$

Definición 2.1.5 *Se dice que A es de **diagonal estrictamente dominante** si*

$$|a_{ii}| > \sum_{j=1, j\neq i}^{j=n} |a_{ij}|, \quad i = 1, \cdots n \tag{2.2}$$

Teorema 2.1.2 *Si A es de diagonal estrictamente dominante, entonces es regular.*

2.1.2. Valores y vectores propios

Sea A una matriz cuadrada $n \times n$.

Definición 2.1.6 *El **espectro** de A es el conjunto $Esp(A) \subset \mathbb{K}$ descrito por los valores propios o autovalores de A.*

Definición 2.1.7 *El **radio espectral** de A es el número real positivo $\rho(A)$*

$$\rho(A) = \max_{\lambda_i \in Esp(A)} |\lambda_i| \tag{2.3}$$

Definición 2.1.8 *Una pareja $(\lambda, \mathbf{x})$ es un **elemento propio** de A si $\mathbf{x}$ es un vector propio de A asociado al valor propio λ.*

Teorema 2.1.3 *Si $(\lambda, \mathbf{x})$ es un elemento propio de A, entonces, para cualquier entero positivo m, $(\lambda^m, \mathbf{x})$ es un elemento propio de A^m.*

Teorema 2.1.4 *Si $(\lambda, \mathbf{x})$ es un elemento propio de A regular, entonces $(\lambda^{-1}, \mathbf{x})$ es un elemento propio de su inversa A^{-1}.*

Teorema 2.1.5 *Si $(\lambda, \mathbf{x})$ es un elemento propio de A, $(\lambda, \mathbf{x})$ es un elemento propio de A^T.*

Teorema 2.1.6 *Si λ es un valor propio de A, $\overline{\lambda}$ es un valor propio de A^+.*

Si $\mathbf{x}$ es un vector propio de A^+ asociado al valor propio $\overline{\lambda}$, $A^+\mathbf{x} = \overline{\lambda}\mathbf{x}$ tomando traspuestas conjugadas se tiene $\mathbf{x}^* A = \lambda \mathbf{x}^*$.

Definición 2.1.9 *Se llama a $\mathbf{x}$ vector propio* ***por la izquierda*** *de A asociado a λ.*

Teorema 2.1.7 *Los valores propios de una matriz simétrica son reales.*

Definición 2.1.10 *Se dice que $A \in M_{n\times n}(\mathbb{K})$ es definida positiva si $\mathbf{x}^T A\mathbf{x} > 0 \ \forall \mathbf{x} \neq 0$.*

Teorema 2.1.8 *Una matriz A hermítica es definida positiva ssi todos sus valores propios son estrictamente positivos.*

En particular, los autovalores de una matriz simétrica son positivos ssi la matriz es definida positiva.

Semejanza de matrices y valores y vectores propios

Definición 2.1.11 *Se dice que dos matrices cuadradas A y B de orden n son semejantes si existe una matriz regular P tal que:*

$$B = PAP^{-1} \tag{2.4}$$

Teorema 2.1.9 *Los espectros de dos matrices semejantes A y B son iguales.*

Si $(\lambda, \mathbf{x})$ es un elemento propio de A, $(\lambda, P^{-1}\mathbf{x})$ es un elemento propio de B.

Teorema 2.1.10 *La matriz A es semejante a una matriz diagonal ssi tiene n vectores propios linealmente independientes.*

Teorema 2.1.11 *Si la matriz A tiene n autovalores distintos, entonces es semejante a una matriz diagonal.*

Teorema 2.1.12 *Toda matriz normal es diagonalizable y sus vectores propios son ortogonales.*

Teorema 2.1.13 *Una matriz hermítica es diagonalizable, sus valores propios son reales y sus vectores propios son ortogonales.*

Las matrices reales y simétricas son hermíticas, luego cumplen las mismas conclusiones.

Descomposición en valores singulares

Definición 2.1.12 *Sea $A \in M_{m\times n}(\mathbb{K})$.*

Se llaman ***valores singulares*** *μ de A, las raíces cuadradas positivas de los valores propios de la matriz cuadrada A^+A de $n \times n$.*

Teorema 2.1.14 *Sea $A \in M_{m\times n}(\mathbb{K})$ y μ_i $i = 1, ..., n$ el conjunto de los valores singulares de A con elementos eventualmente nulos.*

Existen dos matrices unitarias U, de orden m y V, de orden n, tales que $U^+AV = \Sigma$ donde Σ es la matriz $m \times n$

$$\Sigma = \begin{pmatrix} \Sigma_1 \\ O \end{pmatrix}$$

donde

$$\Sigma_1 = \begin{pmatrix} \mu_1 & 0 & \dots & 0 \\ 0 & \mu_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \mu_n \end{pmatrix}$$

El rango de A es igual al número de valores singulares no nulos.

2.1.3. Normas matriciales

Definición 2.1.13 *Dadas normas* $\| \ \|_{\mathbb{K}^n}$, $\| \ \|_{\mathbb{K}^m}$, *definimos en* $M_{m\times n}(\mathbb{K})$ *una* ***norma inducida***[2],

$$\| \ \| : \begin{array}{ccc} M_{m\times n}(\mathbb{K}) & \to & \mathbb{R} \\ A & \mapsto & \|A\| \end{array}, \quad \|A\| = \max_{\|\mathbf{x}\|=1} \|A\mathbf{x}\|$$

Las normas inducidas por una norma vectorial son compatibles con el producto de matrices

$$\|AB\| \leq \|A\| \cdot \|B\|, \quad A \in M_{m\times n},\ B \in M_{n\times p},$$

y con la norma vectorial $\| \ \|$

$$\|A\mathbf{x}\| \leq \|A\| \cdot \|\mathbf{x}\|, \quad A \in M_{m\times n},\ \mathbf{x} \in \mathbb{K}^n.$$

Las diferentes normas vectoriales en los espacios $\mathbb{K}^n$ y $\mathbb{K}^m$ definen distintas normas inducidas en $M_{m\times n}(\mathbb{K})$, todas ellas equivalentes, ya que éste es un espacio de dimensión finita.

Es habitual tomar en $\mathbb{K}^n$ las normas $\| \ \|_p$

$$\|\mathbf{x}\|_p := \left\{ \sum_{j=1}^{n} |x_j|^p \right\}^{\frac{1}{p}}$$

fundamentalmente $p = 1, 2, \infty$.

Sus normas inducidas en $M_{m\times n}(\mathbb{K})$, son

$$\|A\|_p := \max_{\|\mathbf{x}\|_p=1} \|A\mathbf{x}\|_p.$$

Se puede demostrar que la norma $\| \ \|_1$ es el máximo de las sumas de las columnas de la matriz:

$$\|A\|_1 = \max_{1\leq j\leq n} \left\{ \sum_{i=1}^{m} |a_{ij}| \right\}, \quad A = (a_{ij}) \in M_{m\times n}(\mathbb{K}).$$

y que la norma euclídea, $\| \ \|_2$, de una matriz

$$\|A\|_2 = \sqrt{\rho(A^+A)} = \mu_{\max},$$

donde $\mu_{\max}$ es el mayor valor singular de A.

En el caso de una matriz simétrica o hermítica A, luego diagonalizable teorema (2.1.13), $\|A\|_2 = \sqrt{\rho(A^+A)} = \rho(A)$.

Si A es unitaria u ortogonal, $\|A\|_2 = \sqrt{\rho(A^+A)} = \sqrt{\rho(I_n)} = 1$.

La norma $\| \ \|_\infty$, por su parte, es el máximo de las sumas de las filas de la matriz

$$\|A\|_\infty = \max_{1\leq i\leq m} \left\{ \sum_{j=1}^{n} |a_{ij}| \right\}, \quad A = (a_{ij}).$$

Una propiedad interesante de las normas inducidas es que la matriz identidad tiene norma unidad

$$\|\mathbb{I}\| = \max_{\|\mathbf{x}\|=1} \|\mathbb{I}\mathbf{x}\| = \max_{\|\mathbf{x}\|=1} \|\mathbf{x}\| = 1,$$

También existen normas matriciales, por ejemplo, la **norma de Schur**, $\| \ \|_S$

$$\|A\|_S = \sqrt{\mathrm{Tr}\,(A^+A)} = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} |a_{ij}|^2}, \quad A = (a_{ij}) \in M_{m\times n}(\mathbb{K}),$$

[2]Es lícito sustituir el supremo por el máximo, ya que, al ser $\mathbb{K}^n$ un espacio de dimensión finita, $\{\mathbf{x} \in \mathbb{K}^n :\ \|\mathbf{x}\| = 1\}$ es un compacto y, por tanto, la función $\|A\mathbf{x}\|$, continua, presenta máximo y mínimo.

que no son inducidas por ninguna norma en $\mathbb{K}^n$, ya que $\|I_n\|_S = \sqrt{n}$.

No obstante, la norma de Schur es compatible con el producto $\|AB\|_S \leq \|A\|_S \cdot \|B\|_S$ y es equivalente a $\|A\|_2$

$$\|A\|_2 \leq \|A\|_S = \sqrt{n}\|A\|_2$$

Desde un punto de vista práctico es importante observar que mientras las normas $\|A\|_1$, $\|A\|_\infty$ y $\|A\|_S$ son fáciles de calcular, la norma $\|A\|_2$ no lo es, lo que hace la norma de Schur muy conveniente [3].

Teorema 2.1.15 *Sea $\|\ \|$ una norma compatible con el producto de matrices en $M_n(\mathbb{K})$. Entonces, $\rho(A) \leq \|A\|$, para toda $A \in M_n(\mathbb{K})$.*

Teorema 2.1.16 *Sea $A \in M_n(\mathbb{K})$ una matriz fija. Sea $\varepsilon > 0$. Entonces existe una norma inducida $\|\ \|_\varepsilon$ tal que $\|A\|_\varepsilon \leq \rho(A) + \varepsilon$.*

Corolario 2.1.1 $\rho(A) = \text{ínf}_{\|\ \|} \|A\| \quad (\forall A \in M_n(\mathbb{K}))$.

En efecto, $\rho(A)$ es una minorante por el teorema 2.1.15 y es ínfimo, ya que, por el teorema 2.1.16, hay valores de $\|A\|$ tan próximos a $\rho(A)$ como queramos.

Conviene recordar también el siguiente teorema relativo a la exponenciación de matrices.

Teorema 2.1.17 *Para una matriz cuadrada A de $n \times n$, las siguientes afirmaciones son equivalentes:*

1. $\lim_{k\to\infty} A^k = 0$.
2. $\lim_{k\to\infty} \| A^k \| = 0$.
3. $\rho(A) < 1$.

2.2. Condicionamiento de un sistema lineal

Consideremos el sistema lineal

$$A\mathbf{x} = \mathbf{b} \tag{2.5}$$

con

$$A = \begin{pmatrix} 10 & 7 & 8 & 7 \\ 7 & 5 & 6 & 5 \\ 8 & 6 & 10 & 9 \\ 7 & 5 & 9 & 10 \end{pmatrix} \quad \text{y} \quad \mathbf{b} = \begin{pmatrix} 32 \\ 23 \\ 33 \\ 31 \end{pmatrix}$$

cuya solución es el vector $\mathbf{x} = (1, \ldots, 1)$.

En general cuando se resuelve un sistema lineal (2.5) los errores que afectan al resultado final pueden provenir de dos causas.

- Errores en los datos iniciales $(A, \mathbf{b})$.
- Errores de redondeo en el proceso de cálculo.

Los errores en los datos iniciales son consecuencia de un lado de la aproximación de los elementos de la matriz A y de las componentes del vector $\mathbf{b}$ por números máquina de un número finito de dígitos.

Por otra parte, puede ocurrir también que esos datos $(A, \mathbf{b})$ no se conozcan exactamente por ser cantidades obtenidas en experimentos sujetas a errores de observación o que son resultado de mediciones que podrían estar afectadas de errores del aparato de medida.

Por tanto, el sistema lineal que en realidad se resuelve no será (2.5), sino un sistema lineal perturbado tanto en la matriz del sistema como en el término independiente.

$$(A + \Delta A)\,\mathbf{x} = \mathbf{b} + \Delta\mathbf{b} \tag{2.6}$$

[3] Ver una aplicación exhaustiva de esta propiedad en los problemas 2.9 y 2.11.

En nuestro caso, el sistema perturbado (2.6) podría tener el siguiente aspecto

$$\begin{pmatrix} 10 & 7 & 8.1 & 7.2 \\ 7.08 & 5.04 & 6 & 5 \\ 8 & 5.98 & 9.89 & 9 \\ 6.99 & 4.99 & 9 & 9.98 \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 32.01 \\ 22.99 \\ 33.01 \\ 30.99 \end{pmatrix}$$

y por tanto ΔA y $\Delta \mathbf{b}$, las perturbaciones tendrían los siguientes valores:

$$\Delta A = \begin{pmatrix} 0 & 0 & 0.1 & 0.2 \\ 0.08 & 0.04 & 0 & 0 \\ 0 & -0.02 & -0.11 & 0 \\ -0.01 & -0.01 & 0 & -0.02 \end{pmatrix} \quad \text{y} \quad \Delta \mathbf{b} = \begin{pmatrix} 0.01 \\ -0.01 \\ 0.01 \\ -0.01 \end{pmatrix}$$

Estudiemos cómo varía la solución del sistema lineal cuando perturbamos la matriz del sistema y cuando perturbamos sólo el término independiente.

Resolvamos primero el sistema lineal

$$A(\mathbf{x} + \Delta \mathbf{x}) = \mathbf{b} + \Delta \mathbf{b} \tag{2.7}$$

en el que se ha perturbado sólo el término independiente y hemos denotado $\mathbf{x} + \Delta \mathbf{x}$ su solución, destacando la perturbación $\Delta \mathbf{x}$. Se tiene

$$\mathbf{x} + \Delta \mathbf{x} = \begin{pmatrix} 1.82 \\ -0.36 \\ 1.35 \\ 0.79 \end{pmatrix} \rightarrow \Delta \mathbf{x} = \begin{pmatrix} 0.82 \\ -1.36 \\ 0.35 \\ -0.21 \end{pmatrix} \tag{2.8}$$

La perturbación relativa del término independiente es

$$\frac{\| \Delta \mathbf{b} \|_\infty}{\| \mathbf{b} \|_\infty} = 3.03\ 10^{-4} \tag{2.9}$$

y esta perturbación induce una variación en la solución:

$$\frac{\| \Delta \mathbf{x} \|_\infty}{\| \mathbf{x} \|_\infty} = 1.36 \tag{2.10}$$

El error se ha amplificado alrededor de 4.500 veces. Partiendo de una perturbación relativa del orden de $3.03\ 10^{-4}$ hemos obtenido un error relativo del orden de 1.36.

Estudiemos este fenómeno desde el punto de vista teórico:

$$A(\mathbf{x} + \Delta \mathbf{x}) = \mathbf{b} + \Delta \mathbf{b} \quad \Rightarrow \quad A\mathbf{x} + A\Delta \mathbf{x} = \mathbf{b} + \Delta \mathbf{b} \tag{2.11}$$

como $A\mathbf{x} = \mathbf{b}$ se tiene

$$A\Delta \mathbf{x} = \Delta \mathbf{b} \quad \Rightarrow \quad \Delta \mathbf{x} = A^{-1} \Delta \mathbf{b} \tag{2.12}$$

Como las normas que utilizamos son inducidas, se verifica que

$$\| \Delta \mathbf{x} \| = \| A^{-1} \Delta \mathbf{b} \| \leq \| A^{-1} \| \| \Delta \mathbf{b} \| \tag{2.13}$$

Aplicando similar desigualdad a $A\mathbf{x} = \mathbf{b}$ e invirtiendo ambos factores

$$\| A\mathbf{x} \| = \| \mathbf{b} \| \leq \| A \| \| \mathbf{x} \| \quad \Rightarrow \quad \frac{1}{\| \mathbf{x} \|} \leq \frac{\| A \|}{\| \mathbf{b} \|} \tag{2.14}$$

multiplicando término a término estas dos desigualdades entre sí tendremos

$$\frac{\| \Delta \mathbf{x} \|}{\| \mathbf{x} \|} \leq \| A \| \| A^{-1} \| \frac{\| \Delta \mathbf{b} \|}{\| \mathbf{b} \|} \tag{2.15}$$

Por tanto, el error relativo en la solución se mayora respecto al error relativo en la perturbación mediante el factor $\| A \| \| A^{-1} \|$. Este número se denomina **condicionamiento** o **número de condición** de la matriz A, y se denota $K(A)$ o $\text{cond}(A)$.

Cuanto más pequeño sea este número, menos afectarán a la solución final las perturbaciones en el término independiente.

Comprobemos esta desigualdad en el ejemplo anterior

$$\text{cond}_\infty(A) = \| A \|_\infty \| A^{-1} \|_\infty = 33 \cdot 136 = 4488 \tag{2.16}$$

y, por tanto, se debe dar que

$$1.36 \leq 4488 \cdot 3.0303\ 10^{-4} = 1.36$$

Si ahora perturbamos sólo la matriz del sistema tendremos el sistema perturbado

$$(A + \Delta A)(\mathbf{x} + \Delta\mathbf{x}) = \mathbf{b} \tag{2.17}$$

cuya solución es

$$\mathbf{x} + \Delta\mathbf{x} = \begin{pmatrix} -81 \\ 137 \\ -34 \\ 22 \end{pmatrix} \quad \Rightarrow \quad \Delta\mathbf{x} = \begin{pmatrix} -82 \\ 136 \\ -35 \\ 21 \end{pmatrix} \tag{2.18}$$

Una perturbación de la matriz del sistema

$$\frac{\| \Delta A \|_\infty}{\| A \|_\infty} = \frac{0.3}{33} \approx 0.01 \tag{2.19}$$

induce una variación en la solución:

$$\frac{\| \Delta\mathbf{x} \|_\infty}{\| \mathbf{x} \|_\infty} = 136 \tag{2.20}$$

El error se ha amplificado unas 13.600 veces. De una perturbación relativa de A del orden de 0.01 obtenemos un error relativo del orden de 136.

Analicemos este fenómeno teóricamente.

De (2.17) se tiene

$$A\Delta\mathbf{x} + \Delta A(\mathbf{x} + \Delta\mathbf{x}) = \mathbf{0} \quad \Rightarrow \quad \Delta\mathbf{x} = A^{-1}\Delta A(\mathbf{x} + \Delta\mathbf{x})$$

Tomando normas, mayorando y multiplicando y dividiendo por $\| A \|$

$$\| \Delta\mathbf{x} \| \leq \| A^{-1} \| \| \Delta A \| \| \mathbf{x} + \Delta\mathbf{x} \| \quad \Rightarrow \quad \frac{\| \Delta\mathbf{x} \|}{\| \mathbf{x} + \Delta\mathbf{x} \|} \leq \| A \| \| A^{-1} \| \frac{\| \Delta A \|}{\| A \|} \tag{2.21}$$

Con otra medida del error relativo, el condicionamiento de la matriz A vuelve a ser el factor que mayora los errores en las perturbaciones respecto a los errores en la solución.

Por su definición es evidente que el condicionamiento de una matriz depende de la norma elegida.

Propiedades

- El condicionamiento de una matriz tiene como cota inferior a la unidad.

En efecto, ya que la norma inducida de la matriz identidad es siempre 1.

$$1 = \| I \| = \| AA^{-1} \| \leq \| A \| \| A^{-1} \| = \text{cond}(A) \tag{2.22}$$

- Si la matriz A es simétrica,

$$\text{cond}_2(A) = \| A \|_2 \| A^{-1} \|_2 = \rho(A)\rho(A^{-1}) = \frac{|\lambda_{max}|}{|\lambda_{min}|} \tag{2.23}$$

consecuencia de la propiedad $\|A\|_2 = \rho(A)$.

Este resultado tiene una versión general que se expresa en función de los valores singulares y que nos permite asegurar que una matriz estará mejor condicionada cuanto más parecidos en módulo sean entre sí los elementos de su espectro de autovalores o de su espectro de valores singulares.

De hecho, la matriz A del ejemplo anterior es simétrica. Su espectro de autovalores es:

$$\text{Esp}(A) = \{0.0102,\ 0.8431,\ 3.8581,\ 30.2887\ \} \tag{2.24}$$

Como vemos, los autovalores son muy diferentes entre sí. Al ser simétrica, su número de condición en la norma 2 será (2.23)

$$\text{cond}_2(A) = \frac{30.2887}{0.0102} = 2984 \tag{2.25}$$

2.3. Métodos directos

Queda claro que, aunque sepamos resolver nuestro sistema lineal, la solución que obtengamos puede tener errores significativos respecto al resultado esperado. Procede ahora estudiar cómo podemos resolver del modo más rápido y preciso el sistema en cuestión y comenzaremos con los métodos directos.

2.3.1. Eliminación gaussiana

Entre los métodos directos, el más popular para resolver sistemas lineales es la **eliminación gaussiana**, que también se utiliza para calcular determinantes e invertir matrices. La idea básica del método, que ilustramos con un ejemplo, consiste en utilizar transformaciones elementales de fila y/o columna para eliminar sucesivamente las variables empezando por la primera ecuación y la primera variable y continuando con el resto. De esta manera, tras $(n-1)$ eliminaciones se llega a un sistema equivalente al dado, de matriz triangular superior que se resuelve directamente por sustitución hacia atrás [4].

Veamos cómo funciona

$$\left(\begin{array}{ccc|c} 1 & 2 & 3 & 6 \\ 2 & 3 & 4 & 9 \\ -1 & 0 & -1 & -2 \end{array}\right) \tag{2.26}$$

haciendo transformaciones elementales utilizando la primera fila, hacemos cero todos los elementos de la primera columna excepto el diagonal.

$$\left(\begin{array}{ccc|c} 1 & 2 & 3 & 6 \\ 0 & -1 & -2 & -3 \\ 0 & 2 & 2 & 4 \end{array}\right) \tag{2.27}$$

[4] Carl Friedrich Gauss. Conocido como el *príncipe de las matemáticas*. El más grande matemático del siglo XIX y junto a Arquímedes y Newton uno de los tres matemáticos más grandes de todos los tiempos.

Original, profundo, imaginativo, con visión de futuro, sus logros abarcan un espectro muy amplio dando siempre muestras de genialidad.

Nació en Brunswick, en el norte de Alemania, cerca de Hannover, en 1777. Fue un niño prodigio con una excepcional habilidad para la aritmética que mostró muy temprano. A los tres años le corrigió a su padre un error aritmético que tenía en las nóminas. A los diez años su maestro en la escuela pública propuso a los alumnos, para tenerlos ocupados un buen rato, que sumaran los cien primeros números enteros. Casi de inmediato Carl dejó el resultado correcto, 5.050, sin ninguna operación adicional, en la mesa del maestro. Había sumado mentalmente la progresión aritmética $1 + 2 + 3 + ... + 99 + 100$ tras observar que $100 + 1 = 99 + 2 = 98 + 3 = ... = 101$. Ya que hay 50 parejas el resultado es $50 \times 101 = 5.050$.

Estas habilidades no pasaron desapercibidas y le ganaron la tutela del duque de Brunswick, que sufragó su educación primero en Brunswick y luego en Göttingen.

Como se puede comprender, la biografía de una figura de este tamaño necesita más espacio que una nota al pie, así que nos centraremos en los aspectos más relacionados con el cálculo numérico. Con catorce o quince años inventó el método de los mínimos cuadrados que aplicó después con gran éxito en astronomía.

Los astrónomos buscaban a finales del siglo XVIII un planeta nuevo entre las órbitas de Marte y Júpiter cuya existencia era sugerida por la ley de Bode (1772). El mayor de estos asteroides se descubrió en 1801 y se llamó Ceres. Se hicieron entonces, unas pocas mediciones de su posición antes de que se alejara del Sol. ¿Cómo calcular su órbita con tan pocos datos? Gauss aceptó el desafío y utilizó el método de los mínimos cuadrados para determinar su órbita, y precisó a los astrónomos dónde debían enfocar sus telescopios para encontrar a Ceres, y allí estaba. Este logro le dio fama, un contrato de profesor de astronomía y la dirección del nuevo observatorio de Göttingen.

En 1820 "abandonó" las matemáticas, y se dedicó a la astronomía aplicada, encargándose de supervisar los trabajos cartográficos del reino, un trabajo rutinario y tedioso que le tuvo ocupado bastantes años. Hasta en este entorno, tan poco motivador, se mostró la genialidad de Gauss. En 1827 publicó su obra maestra de la teoría de superficies, las ***Disquisitiones generales circa superficies curvas*** en la que fundó la geometría intrínseca sobre una superficie y demostró el teorema ***egregium***.

Gauss murió en su casa del observatorio de Göttingen en 1855.

Fijándonos en el elemento diagonal de la segunda fila, hacemos cero todos los elementos de la segunda columna por debajo del diagonal.

$$\left(\begin{array}{ccc|c} 1 & 2 & 3 & 6 \\ 0 & -1 & -2 & -3 \\ 0 & 0 & -2 & -2 \end{array}\right) \tag{2.28}$$

Con esto tenemos ya el sistema triangular superior equivalente que resolvemos por sustitución hacia atrás comenzando por la última ecuación y acabando en la primera.

$$\begin{aligned} -2x_3 = -2 \quad &\Rightarrow \quad x_3 = 1 \\ -x_2 - 2x_3 = -3 \quad &\Rightarrow \quad x_2 = 1 \\ x_1 + 2x_2 + 3x_3 = 6 \quad &\Rightarrow \quad x_1 = 1 \end{aligned} \tag{2.29}$$

En la eliminación gausiana tal y como la acabamos de exponer se asume que en la k-ésima eliminación el coeficiente de la variable que se desea eliminar que se llama el **pivote** y que ocupa la posición (k, k) de la matriz del sistema en ese momento es distinto de cero. A menudo no sucede así, por lo que se aconseja reordenar las ecuaciones e incluso los términos en cada una de ellas para lograr, por razones de estabilidad numérica, que el pivote sea el mayor posible en valor absoluto [5].

Existen diferentes estrategias para la elección del pivote, según que la reordenación afecte sólo a las filas **pivotación parcial** o tanto a las filas como a las columnas **pivotación total** [6].

En principio podría parecer más conveniente la pivotación total, pero su alto costo numérico, ya que exige en cada eliminación la comparación de todos los elementos de la matriz, hace preferible en la práctica la pivotación parcial con **equilibrado** de filas y columnas. Este equilibrado tiene como objeto normalizarlas, lo que se consigue multiplicando todos sus elementos por números convenientes.

2.3.2. Descomposición LU

Esta variante del método de Gauss consiste en factorizar la matriz A del sistema lineal (2.5) como producto de una triangular inferior L y otra U triangular superior.

Una vez obtenida la descomposición se resuelve el sistema lineal

$$(LU)\mathbf{x} = L(U\mathbf{x}) = \mathbf{b} \tag{2.30}$$

resolviendo sucesivamente los dos sistemas lineales triangulares

$$\begin{aligned} L\mathbf{y} &= \mathbf{b} \\ U\mathbf{x} &= \mathbf{y} \end{aligned} \tag{2.31}$$

Existen muchas formas de realizar esta descomposición. Explicaremos la más sencilla, el **algoritmo de Crout**, que funciona siempre que todos los menores principales de A sean distintos de cero[7]. En esta descomposición se supone que la matriz triangular superior U tiene elementos unidad en la diagonal principal.

Resolvamos el sistema lineal del apartado anterior con este método.

$$\left(\begin{array}{ccc} 1 & 2 & 3 \\ 2 & 3 & 4 \\ -1 & 0 & -1 \end{array}\right) = \left(\begin{array}{ccc} l_{11} & 0 & 0 \\ l_{21} & l_{22} & 0 \\ l_{31} & l_{32} & l_{33} \end{array}\right) \left(\begin{array}{ccc} 1 & u_{12} & u_{13} \\ 0 & 1 & u_{23} \\ 0 & 0 & 1 \end{array}\right) \tag{2.32}$$

Para calcular estos coeficientes procedemos sucesivamente por identificación:

$$\begin{array}{lll} l_{11} = 1 & & \\ l_{11}u_{12} = 2 & \Rightarrow & u_{12} = 2 \ ; \\ l_{11}u_{13} = 3 & \Rightarrow & u_{13} = 3 \end{array} \qquad \begin{array}{lll} l_{21} = 2 & & \\ l_{21}u_{12} + l_{22} = 3 & \Rightarrow & u_{12} = 2 \ ; \\ l_{21}u_{13} + l_{22}u_{23} = 4 & \Rightarrow & u_{23} = 2 \end{array}$$

[5] Hay un mundo de pequeñas sutilezas que mejoran el comportamiento numérico del algoritmo reduciendo los errores de redondeo, aunque su estudio detallado escapa al contenido del curso.

[6] Ver en el problema 2.9 una descripción muy detallada del algoritmo de eliminación gaussiana con estrategia de pivotación parcial, así como comentarios sobre el mejor almacenamiento de la información en cada eliminación para su posterior uso.

[7] Recordemos que los menores principales de A son los determinantes de las submatrices principales y que la submatriz principal A_i de A es la que tiene como elementos a_{jk} con $1 \leq j, k \leq i$.

$$\begin{array}{rcl} l_{31} = -1 & & \\ l_{31}u_{12} + l_{32} = 3 & \Rightarrow & l_{32} = 2 \\ l_{31}u_{13} + l_{32}u_{23} + l_{33} = -1 & \Rightarrow & l_{33} = -2 \end{array}$$

Una vez obtenida la factorización resolvemos los dos sistemas triangulares (2.31).
El primero de ellos $L\mathbf{y} = \mathbf{b}$

$$\begin{pmatrix} 1 & 0 & 0 \\ 2 & -1 & 0 \\ -1 & 2 & -2 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} 6 \\ 9 \\ -2 \end{pmatrix} \tag{2.33}$$

se resuelve por sustitución hacia adelante comenzando por la primera ecuación y acabando en la última.

$$\begin{array}{rcl} y_1 = 6 & & \\ 2y_1 - y_2 = 9 & \Rightarrow & y_2 = 3 \\ -y_1 + 2y_2 - 2y_3 = -2 & \Rightarrow & y_3 = 1 \end{array} \tag{2.34}$$

El segundo $U\mathbf{x} = \mathbf{y}$ por sustitución hacia atrás.

$$\begin{pmatrix} 1 & 2 & 3 \\ 0 & 1 & 2 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 6 \\ 3 \\ 1 \end{pmatrix} \tag{2.35}$$

$$\begin{array}{rcl} x_3 = 1 & & \\ x_2 + 2x_3 = 3 & \Rightarrow & x_2 = 1 \\ x_1 + 2x_2 + 3x_3 = 6 & \Rightarrow & x_1 = 1 \end{array} \tag{2.36}$$

Si la matriz A del sistema no tiene ninguna estructura especial (simétrica, con muchos ceros, etc.), y no converge para los métodos iterativos disponibles, la descomposición LU es el método más aconsejable.

2.3.3. Descomposición de Cholesky, $A = LL^T$

A menudo, la matriz A del sistema es simétrica y definida positiva (2.1.10). Para este tipo de matrices existe una descomposición LU especial, la **descomposición de Cholesky** [8] en la que $U = L^T$. Se determina de nuevo la matriz L por identificación, igual que en el algoritmo de Crout.

Se deja como ejercicio resolver de este modo el sistema lineal (2.5).

2.3.4. Método de Gauss-Jordan

El método de Gauss-Jordan es el método directo óptimo para encontrar la inversa de una matriz cuando ésta no tiene ninguna estructura particular [9]. En el método de Gauss-Jordan se dispone la matriz cuadrada A de $n \times n$ que se desea invertir a la izquierda y la matriz unidad I_n a su derecha. Se realizan sucesivas eliminaciones gausianas mediante transformaciones elementales de fila y columna hasta tener a la izquierda la matriz unidad I_n, en cuyo caso, a la derecha tendremos la inversa A^{-1}.

Veámoslo con el mismo ejemplo.

$$\left(\begin{array}{ccc|ccc} 1 & 2 & 3 & 1 & 0 & 0 \\ 2 & 3 & 4 & 0 & 1 & 0 \\ -1 & 0 & -1 & 0 & 0 & 1 \end{array}\right) \rightarrow \left(\begin{array}{ccc|ccc} 1 & 2 & 3 & 1 & 0 & 0 \\ 0 & -1 & -2 & -2 & 1 & 0 \\ 0 & 2 & 2 & 1 & 0 & 1 \end{array}\right)$$

[8] Andrè-Louis Cholesky (1875-1918) Ex-alumno de l'Ecole Polytechnique, comandante de artillería del ejército francés adscrito a la sección geodésica del servicio geográfico. Trabajó en Creta y en el norte de África antes de la primera guerra mundial. Desarrolló el método que lleva su nombre para calcular las soluciones de las ecuaciones normales que aparecen en los problemas de ajuste de datos por el método de los mínimos cuadrados. Lo publicó después de su muerte su compañero el comandante Benoit en el ***Bulletin Geodesique***. La conexión entre Cholesky y la eliminación gausiana se descubrió más tarde.

[9] Es un error de concepto resolver un sistema lineal mediante la matriz inversa, ya que el cálculo de la inversa equivale a resolver n sistemas lineales.

Normalizamos la segunda fila dividiendo por el elemento diagonal, hacemos cero el resto de los elementos de la segunda columna.

$$\left(\begin{array}{ccc|ccc} 1 & 2 & 3 & 1 & 0 & 0 \\ 0 & 1 & 2 & 2 & -1 & 0 \\ 0 & 2 & 2 & 1 & 0 & 1 \end{array}\right) \rightarrow \left(\begin{array}{ccc|ccc} 1 & 0 & -1 & -3 & 2 & 0 \\ 0 & 1 & 2 & 2 & -1 & 0 \\ 0 & 0 & -2 & -3 & 2 & 1 \end{array}\right)$$

Normalizamos la tercera fila dividiendo por el elemento diagonal, hacemos cero el resto de los elementos de la tercera columna.

$$\left(\begin{array}{ccc|ccc} 1 & 0 & -1 & -3 & 2 & 0 \\ 0 & 1 & 2 & 2 & -1 & 0 \\ 0 & 0 & 1 & 1.5 & -1 & -0.5 \end{array}\right) \rightarrow \left(\begin{array}{ccc|ccc} 1 & 0 & 0 & -1.5 & 1.0 & -0.5 \\ 0 & 1 & 0 & -1.0 & 1.0 & 1.0 \\ 0 & 0 & 1 & 1.5 & -1.0 & -0.5 \end{array}\right)$$

Ver en el apartado 2.c) del problema 2.9 el cálculo detallado de la inversa de una matriz por el método de Gauss-Jordan con aritmética de 6 dígitos.

2.4. Métodos iterativos

2.4.1. Convergencia

Los métodos iterativos responden a la misma filosofía que utilizamos para resolver de modo iterativo problemas no lineales. Transformar el problema propuesto en uno de punto fijo para después aplicar el teorema de la aplicación contractiva.

En general, los esquemas iterativos a que haremos referencia serán todos del tipo:

$$\mathbf{x}^{(k+1)} = B\mathbf{x}^{(k)} + \mathbf{c} \tag{2.37}$$

donde $B \in M_n(\mathbb{K})$ y $\mathbf{c} \in \mathbb{K}^n$.

Supuesto que la sucesión recurrente así definida fuera convergente a $\mathbf{x}$, pasando al límite en (2.37)

$$\mathbf{x} = B\mathbf{x} + \mathbf{c} \tag{2.38}$$

y por tanto necesariamente

$$\mathbf{c} = (I - B)A^{-1}\mathbf{b} \tag{2.39}$$

Una vez que $\mathbf{c}$ verifica esa condición, la convergencia depende del radio espectral de la matriz B.

Teorema 2.4.1 *Un esquema iterativo para resolver sistema lineales del tipo (2.37) converge ssi el radio espectral de la matriz B es estrictamente menor que la unidad.*

Demostración

En efecto, si el esquema converge, la sucesión $\mathbf{x}^{(k)} - \mathbf{x}$ debe tender a cero. Como

$$\mathbf{x}^{(k+1)} - \mathbf{x} = B\mathbf{x}^{(k)} + c - (B\mathbf{x} + \mathbf{x}) = B(\mathbf{x}^{(k)} - \mathbf{x}) = B^{k+1}(\mathbf{x}^{(0)} - \mathbf{x})$$

$B^{k+1}(\mathbf{x}^{(0)} - \mathbf{x})$ ha de tender a cero cualquiera que sea $\mathbf{x}^{(0)}$.

Esto sólo es posible si la sucesión B^k de las potencias sucesivas de B tiende a cero, es decir, si el radio espectral de la matriz B es menor que uno (teorema 2.1.17).

La recíproca es una consecuencia directa del teorema de la aplicación contractiva aplicada a la ecuación $\mathbf{x} = T(\mathbf{x}) = B\mathbf{x} + \mathbf{c}$. Si $\rho(B) < 1$, existe alguna norma matricial inducida en la que $\| B \| < 1$. Por tanto, ya que $\| B(\mathbf{x} - \mathbf{y}) \| \leq \| B \| \| \mathbf{x} - \mathbf{y} \|$ tendremos que la aplicación $T(\mathbf{x}) = B\mathbf{x} + \mathbf{c}$ verificará la condición de Lipschitz y, por tanto, será una contracción de $\mathbb{K}^n$, con lo que la sucesión asociada (2.37) convergerá a un punto fijo que será la solución del sistema lineal.

Esta convergencia es además independiente del estimador inicial, al ser una contracción en todo el espacio, aunque será más rápida si el estimador inicial está correctamente elegido y sobre todo si $\rho(B)$ es bastante menor que la unidad.

2.4.2. Esquema general

Todos los métodos iterativos que estudiaremos para resolver el sistema lineal $A\mathbf{x} = \mathbf{b}$ se basan en una descomposición de la matriz A del tipo $A = M - N$ con M invertible.

El sistema lineal se escribe entonces

$$M\mathbf{x} = N\mathbf{x} + \mathbf{b} \quad \Rightarrow \quad \mathbf{x} = (M^{-1}N)\mathbf{x} + M^{-1}\mathbf{b} \tag{2.40}$$

en la que se reconoce la escritura $\mathbf{x} = T(\mathbf{x}) = B\mathbf{x} + \mathbf{c}$ con $B = M^{-1}N$ y $\mathbf{c} = M^{-1}\mathbf{b}$.

Se define así el esquema iterativo de punto fijo.

$$M\mathbf{x}^{(k+1)} = N\mathbf{x}^{(k)} + \mathbf{b} \quad \Rightarrow \quad \mathbf{x}^{(k+1)} = (M^{-1}N)\mathbf{x}^{(k)} + M^{-1}\mathbf{b} \tag{2.41}$$

que verifica la condición necesaria de convergencia (2.39)

$$\begin{aligned}(I - B)A^{-1}\mathbf{b} = (I - M^{-1}N)A^{-1}\mathbf{b} = (M^{-1}M - M^{-1}N)A^{-1}\mathbf{b} \\ = M^{-1}(M - N)A^{-1}\mathbf{b} = M^{-1}(M - N)(M - N)^{-1}\mathbf{b} = M^{-1}\mathbf{b} = \mathbf{c}\end{aligned} \tag{2.42}$$

y que de acuerdo con el teorema 2.4.1 será convergente ssi $\rho(M^{-1}N) < 1$.

Las dos formulaciones del esquema iterativo (2.41) plantean estrategias distintas. En la primera, se obtiene la nueva iterada $\mathbf{x}^{(k+1)}$ resolviendo un sistema de ecuaciones. Se elige M de forma que ese sistema lineal en cada paso sea sencillo de resolver. En la segunda se invierte M desde la salida y el proceso de actualización requiere sólo multiplicaciones y sumas. Se elige M fácilmente invertible. Ambos criterios de selección están obviamente relacionados aunque son distintos.

2.4.3. Método iterativo de Jacobi

Se basa en una descomposición de la matriz A del tipo $A = M - N$, con $M = D$, parte diagonal de la matriz A cuyos elementos se suponen no nulos, y $N = L + U$, siendo L, U las partes triangulares inferior y superior respectivamente de la matriz A cambiadas de signo.

La descomposición propuesta satisface los criterios de selección expuestos.

Es posible definir estas matrices D, L y U con los mismos criterios, a partir de una partición de la matriz A en bloques de varios elementos

$$A = \begin{pmatrix} A_{11} & \dots & A_{1s} \\ A_{21} & \dots & A_{2s} \\ \vdots & \ddots & \vdots \\ A_{s1} & \dots & A_{ss} \end{pmatrix} \tag{2.43}$$

donde A_{ij} es una matriz de $n_i \times n_j$ con $\sum_{i=1,s} n_i = n$ y todas las matrices diagonales A_{ii} son invertibles. Todas las fórmulas que obtengamos son válidas para este caso sustituyendo los productos entre elementos por productos matriciales entre bloques. La mayor ventaja de esta formulación se consigue en los grandes sistemas tridiagonales que se obtienen en la resolución numérica de ecuaciones diferenciales en derivadas parciales. (Ver el problema 2.10 en el que se resuelve un sistema tridiagonal por métodos iterativos (Gauss-Seidel y relajación) con descomposición en bloques de la matriz).

Veamos con un ejemplo cómo funciona el esquema 2.41 en la descomposición de Jacobi [10] de la matriz A.

Se trata de resolver el sistema lineal:

$$\begin{pmatrix} 4 & 1 & 0 \\ 1 & 4 & 1 \\ 0 & 1 & 4 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} -3 \\ 10 \\ 1 \end{pmatrix} \tag{2.44}$$

[10] Carl Gustav Jacob Jacobi nació en 1804 en Potsdam, Prusia y murió en 1851 en Berlín. De padres judíos, en 1825 se hizo cristiano para facilitar su entrada como profesor en la universidad. En el año académico 1825-26 daba clase en la Universidad de Berlín y en 1826 Jacobi se trasladó a la universidad de Königsberg. En 1832 obtuvo la posición de profesor vitalicio en la universidad de Königsberg. Su fama de excelente profesor atrajo a numerosos estudiantes. Introdujo la técnica del seminario para enseñar a los alumnos los últimos avances en matemáticas. Jacobi tuvo un impacto enorme en sus estudiantes y creó "escuela". C. W. Borchardt, E. Heine, L. O. Hesse, y P. L. von Seidel pertenecieron a este círculo; contribuyendo tanto a extender las creaciones matemáticas de Jacobi como la nueva actitud de la enseñanza de la matemática orientada a la investigación. La terna formada por Bessel, Jacobi, y Franz Neumann fue el núcleo de la revitalización de las matemáticas en las universidades alemanas.

cuya solución es el vector $\mathbf{x} = (-1.5,\ 3,\ -0.5)$.

Se tiene

$$M = D = \begin{pmatrix} 4 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 4 \end{pmatrix} \qquad L = \begin{pmatrix} 0 & 0 & 0 \\ -1 & 0 & 0 \\ 0 & -1 & 0 \end{pmatrix} \qquad U = \begin{pmatrix} 0 & -1 & 0 \\ 0 & 0 & -1 \\ 0 & 0 & 0 \end{pmatrix} \tag{2.45}$$

con

$$N = L + U \tag{2.46}$$

La elección del estimador inicial no influye en la convergencia, pero sí en el número de iteraciones para llegar a una solución aceptable. Si se conoce una aproximación de la solución, se usará como estimador inicial. Caso de que no sea así, se puede tomar como estimador inicial bien el término independiente, o bien un vector de componentes todas iguales a 1.

Nosotros tomaremos aquí como estimador inicial un vector que satisfaga la primera ecuación del sistema lineal $\mathbf{x}^{(0)} = (-1,\ 1,\ -1)$.

Para comprobar si estamos convergiendo a la solución es necesario disponer de un buen criterio de convergencia.

Demostraremos más adelante un resultado (2.4.4) que sugiere tomar la sucesión $\| \mathbf{r}^{(k)} \|$ de la norma del vector residuo $\mathbf{r}^{(k)} = \mathbf{b} - A\mathbf{x}^{(k)}$ como indicador de convergencia.

Aquí, tomamos por comodidad la norma $\| \ . \ \|_\infty$ y

$$\| \mathbf{r}^{(0)} \|_\infty = 8 \tag{2.47}$$

Demos el primero paso del esquema

$$D\mathbf{x}^{(1)} = N\mathbf{x}^{(0)} + \mathbf{b} = (L+U)\mathbf{x}^{(0)} + \mathbf{b} \tag{2.48}$$

$$\begin{pmatrix} 4 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 4 \end{pmatrix} x^{(1)} = \begin{pmatrix} 0 & -1 & 0 \\ -1 & 0 & -1 \\ 0 & -1 & 0 \end{pmatrix} \begin{pmatrix} -1 \\ 1 \\ -1 \end{pmatrix} + \begin{pmatrix} -3 \\ 10 \\ 1 \end{pmatrix} \tag{2.49}$$

es decir,

$$\begin{pmatrix} 4 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 4 \end{pmatrix} \mathbf{x}^{(1)} = \begin{pmatrix} -4 \\ 10 \\ 0 \end{pmatrix} \quad \Rightarrow \quad \mathbf{x}^{(1)} = \begin{pmatrix} -1 \\ 3 \\ 0 \end{pmatrix} \tag{2.50}$$

con

$$\| \mathbf{r}^{(1)} \|_\infty = 2 \tag{2.51}$$

Por tanto, el residuo ha disminuido. Si seguimos iterando:

$$\mathbf{x}^{(2)} = \begin{pmatrix} -1.5 \\ 2.75 \\ -0.5 \end{pmatrix}; \quad \| \mathbf{r}^{(2)} \|_\infty = 1 \tag{2.52}$$

$$\mathbf{x}^{(5)} = \begin{pmatrix} -1.4922 \\ 3.0000 \\ -0.4922 \end{pmatrix}; \quad \| \mathbf{r}^{(5)} \|_\infty = 0.0313 \tag{2.53}$$

$$\mathbf{x}^{(9)} = \begin{pmatrix} -1.4999 \\ 3.0000 \\ -0.4999 \end{pmatrix}; \quad \| \mathbf{r}^{(9)} \|_\infty = 4.8828 \cdot 10{-}4 < 10^{-3} \tag{2.54}$$

etcétera.

Enunciamos a continuación un teorema de convergencia relativo a sistemas lineales de matriz A de diagonal estrictamente dominante (2.1.5) muy frecuentes al resolver numéricamente ecuaciones diferenciales en derivadas parciales.

Teorema 2.4.2 *Si la matriz A del sistema es de diagonal estrictamente dominante, el método de Jacobi es convergente.*

2.4.4. Método iterativo de Gauss-Seidel

Se basa en una descomposición de la matriz A del tipo $A = M - N$, con $M = D - L$, y $N = U$.

En cada iteración se tiene que resolver un sistema triangular por sustitución hacia adelante.

La mecánica de cada paso del método de Gauss-Seidel [11]es por tanto, más complicada que en el método de Jacobi, pero la velocidad de convergencia es superior [12].

Ilustremos el algoritmo con el mismo ejemplo 2.44.

$$M\mathbf{x}^{(1)} = N\mathbf{x}^{(0)} + \mathbf{b} = U\mathbf{x}^{(0)} + \mathbf{b} \tag{2.55}$$

$$\begin{pmatrix} 4 & 0 & 0 \\ 1 & 4 & 0 \\ 0 & 1 & 4 \end{pmatrix} x^{(1)} = \begin{pmatrix} 0 & -1 & 0 \\ 0 & 0 & -1 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} -1 \\ 1 \\ -1 \end{pmatrix} + \begin{pmatrix} -3 \\ 10 \\ 1 \end{pmatrix} \tag{2.56}$$

Y tendremos, por tanto, que resolver el sistema triangular superior:

$$\begin{pmatrix} 4 & 0 & 0 \\ 1 & 4 & 0 \\ 0 & 1 & 4 \end{pmatrix} \mathbf{x}^{(1)} = \begin{pmatrix} -4 \\ 11 \\ 1 \end{pmatrix} \tag{2.57}$$

que se resuelve por sustitucíón hacia adelante:

$$\mathbf{x}^{(1)} = \begin{pmatrix} -1.0000 \\ 3.0000 \\ -0.5000 \end{pmatrix} \tag{2.58}$$

con

$$\| \mathbf{r}^{(1)} \|_\infty = 2 \tag{2.59}$$

En el paso siguiente se obtiene el resultado con cuatro decimales

$$\mathbf{x}^{(2)} = \begin{pmatrix} -1.5000 \\ 3.0000 \\ -0.5000 \end{pmatrix} \tag{2.60}$$

Se constata un comportamiento mucho mejor que el del método de Jacobi (ver en el problema 2.7, un estudio comparativo de la convergencia de ambos métodos).

Enunciamos a continuación un teorema de convergencia similar al teorema 2.4.2 relativo al método de Jacobi.

Teorema 2.4.3 *Si A es una matriz de diagonal estrictamente dominante, el método de Gauss-Seidel es convergente.*

2.4.5. Test de parada

Fijada una tolerancia ϵ el proceso se debe parar en la iteración k-ésima tal que el error

$$\| \mathbf{e}^{(k)} \| = \| \mathbf{x} - \mathbf{x}^{(k)} \| < \epsilon \tag{2.61}$$

[11] Philipp Ludwig von Seidel nació en Zweibrücken, Alemania en 1821 y murió en Munich en 1896.

En vez de entrar directamente en la universidad, pasó un tiempo de formación privada en matemáticas con L. C. Schnürlein, un buen matemático que había estudiado con Gauss.

Seidel entró en la universidad de Berlín en 1840 y estudió con Dirichlet y Encke. En 1842 se trasladó a Könisberg donde estudió con Bessel, Jacobi y Frank Neumann. Cuando Jacobi dejó Könisberg, por motivos de salud, Bessel aconsejó a Seidel que fuera a Munich para estudiar el doctorado. Seidel obtuvo su doctorado en Munich en 1846 y a los seis meses se cualificó como profesor en Munich. Su carrera en esta universidad progresó rápidamente y en 1855 obtuvo la posición equivalente a catedrático. Los resultados que nos afectan aparecen en sus trabajos sobre el método de los mínimos cuadrados.

[12] En cada iteración del método de Jacobi una componente cualquiera del vector $\mathbf{x}^{(k+1)}$ se calcula en función únicamente de las componentes del vector $\mathbf{x}^{(k)}$ resultado de la iteración anterior. El método de Gauss-Seidel está basado, por el contrario, en la idea de emplear en el cálculo de una componente de $\mathbf{x}^{(k+1)}$ las componentes ya calculadas de dicho vector en vez de sus correspondientes de $\mathbf{x}^{(k)}$. Esta estrategia es responsable en parte del mejor comportamiento del método de Gauss-Seidel.

El método se para cuando la diferencia en norma, entre la solución $\mathbf{x}$ y la estimación de ese valor en la iteración k-ésima, es menor que un valor ϵ preasignado.

Dado que no conocemos $\mathbf{x}$, tendremos que imponer esa condición de un modo indirecto.

Uno de los test de parada más convenientes utiliza el residuo

$$\mathbf{r}^{(k)} = \mathbf{b} - A\mathbf{x}^{(k)} \tag{2.62}$$

basado en la conclusión del siguiente teorema:

Teorema 2.4.4 *Si* $\dfrac{\| \mathbf{r}^{(k)} \|}{\| \mathbf{b} \|} < \epsilon$, *entonces* $\dfrac{\| \mathbf{e}^{(k)} \|}{\| \mathbf{x} \|} < \epsilon \; cond(A)$

Es interesante "ver" este pequeño teorema desde el punto de vista del condicionamiento de la matriz.

Cuánto mayor sea cond(A)*, peores son las mayoraciones en el residuo como indicador del error.*

Si la matriz está bien condicionada, acotar el residuo es casi equivalente a acotar el error total.

En el ejemplo que hemos utilizado para ilustrar los métodos de Jacobi y Gauss-Seidel,

$$\begin{aligned} \text{cond}_\infty(A) &= 2.5714 \\ \frac{\| \mathbf{r}^{(5)} \|}{\| \mathbf{b} \|} &= 0.0031 \\ \| \mathbf{e}^{(5)} \| &= \| \mathbf{x} - \mathbf{x}^{(5)} \| = 0.0078 \rightarrow \frac{\| \mathbf{e}^{(5)} \|}{\| \mathbf{x} \|} = 0.0026 \end{aligned}$$

y $0.0026 \leq 2.5714 \cdot 0.0031 = 0.0080$, con lo que se verifica la cota dada por el teorema.

La selección del valor de ϵ depende en cada caso del problema real que estemos resolviendo, así como de las estimaciones que podamos tener del condicionamiento de la matriz del sistema.

Otro posible test de parada es establecer que la diferencia en norma entre dos iteraciones consecutivas no supere un valor preasignado.

Combinando ambos tests se obtiene uno que nos garantiza que estamos suficientemente cerca de la solución y que no necesitamos continuar trabajando.

Como regla fija se debe imponer un límite al número de iteraciones que puede efectuar el algoritmo. El método se para cuando se supera dicho número. Será el momento de analizar el porqué.

2.5. Cálculo de valores y vectores propios

El problema del cálculo de los autovalores y autovectores de una matriz cuadrada está resuelto a nivel teórico en álgebra lineal satisfactoriamente. Desde el punto de vista numérico esta solución no es útil por su elevado costo numérico, así que es necesario desarrollar algoritmos que resuelvan el problema.

Estos algoritmos son de dos tipos:

- Métodos de transformaciones.

Se efectúan transformaciones del tipo $P^{-1}AP$, rotaciones en los métodos de Jacobi y Givens, transformaciones ortogonales de Householder, etc., con objeto de reducir la matriz A a una forma diagonal o tridiagonal, que permita calcular más fácilmente los valores propios.

- Métodos iterativos.

Nosotros sólo estudiaremos aquí el método de la potencia y variantes que son los métodos iterativos más simples que resuelven el problema de la determinación de uno o varios valores propios y de los vectores propios asociados.

2.5.1. Método de la potencia y variantes

Se supone que la matriz A es diagonalizable y que tiene un valor propio dominante que denotaremos λ_1, luego

$$|\lambda_1| > |\lambda_2| \geq |\lambda_3| \geq \cdots \geq |\lambda_n| \tag{2.63}$$

Sea $\{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\}$ la base de vectores propios asociada, de modo que $A\mathbf{x}_i = \lambda_i \mathbf{x}_i$ para $i = 1, .., n$.

Este método determina λ_1 y un vector propio asociado $\mathbf{x}_1$.

Veamos su fundamento e ilustremos su aplicación con un ejemplo.

Cualquier vector arbitrario $\mathbf{x} \in I\!K^n$ se expresa de una única forma como combinación lineal de los vectores propios de la base.

$$\mathbf{x} = C_1\mathbf{x}_1 + C_2\mathbf{x}_2 + \cdots + C_n\mathbf{x}_n \tag{2.64}$$

con $C_i \in I\!K$ para $i = 1, .., n$.

Multiplicando los dos miembros de (2.65) por la potencias A^k de A con $k \in I\!N$

$$\begin{aligned} A^k\mathbf{x} =& A^k\left(C_1\mathbf{x}_1 + C_2\mathbf{x}_2 + \cdots + C_n\mathbf{x}_n\right) = C_1\lambda_1^k\mathbf{x}_1 + C_2\lambda_2^k\mathbf{x}_2 + \cdots + C_n\lambda_n^k\mathbf{x}_n = \\ =& \lambda_1^k\left[C_1\mathbf{x}_1 + C_2\left(\frac{\lambda_2}{\lambda_1}\right)^k\mathbf{x}_2 + \cdots + C_n\left(\frac{\lambda_n}{\lambda_1}\right)^k\mathbf{x}_n\right] \end{aligned} \tag{2.65}$$

Ya que $|\lambda_1| > |\lambda_i|$ para $i \geq 2$, los cocientes $(\lambda_i/\lambda_1)^k \to 0$ cuando $k \to \infty$ y $A^k\mathbf{x} \xrightarrow[k\to\infty]{} \lambda_1^k C_1\mathbf{x}_1$ luego a partir de un cierto rango en adelante es razonable escribir

$$A^k\mathbf{x} \approx \lambda_1^k C_1\mathbf{x}_1 \tag{2.66}$$

igualdad que es la base del método iterativo.

Elegido un estimador inicial $\mathbf{x}^{(0)}$ construimos la sucesión

$$\mathbf{x}^{(k+1)} = A\mathbf{x}^{(k)} \quad \Rightarrow \quad \mathbf{x}^{(k)} = A^k\mathbf{x}^{(0)} \tag{2.67}$$

Si se quiere usar (2.66) para el cálculo de λ_1 y $\mathbf{x}_1$ hay que observar que

- Si $|\lambda_1| < 1$, $|\lambda_1|^k \xrightarrow[k\to\infty]{} 0$ y $\mathbf{x}^{(k)} \to \mathbf{0}$.
- Si $|\lambda_1| > 1$, $|\lambda_1|^k \xrightarrow[k\to\infty]{} \infty$ y (2.67) tiende a infinito.

Luego, si no se pone remedio, el esquema iterativo (2.67) será inútil.

Como buscamos realmente un vector cualquiera de la recta de vectores propios asociada a λ_1 se puede jugar con el factor C del vector propio $C\mathbf{x}_1$, es decir, podemos introducir en cada paso un **factor de escala** que corrija el problema.

En el caso más simple, elegimos que el vector propio tenga una de sus componentes, por ejemplo la primera, igual a 1 y por tanto en cada paso escalamos el vector $\mathbf{x}^{(k)}$ para que sea $\mathbf{x}_1^{(k)} = 1$.

Al final del paso k-ésimo

$$\mathbf{x}_1^{(k)} \approx \lambda_1^k C_1\mathbf{x}_1^{(0)} = \lambda_1^k C_1 \tag{2.68}$$

Si damos un paso en el proceso iterativo (de k a $k+1$) tendremos

$$\mathbf{x}_1^{(k+1)} \approx \lambda_1^{k+1} C_1 \tag{2.69}$$

Dividiendo (2.69) entre (2.68) obtenemos

$$\frac{\mathbf{x}_1^{(k+1)}}{\mathbf{x}_1^{(k)}} = \lambda_1 \tag{2.70}$$

luego si $\mathbf{x}_1^{(k)} = 1$, entonces $\mathbf{x}_1^{(k+1)} = \lambda_1$. Si a continuación escalamos $\mathbf{x}^{(k+1)}$ para que $\mathbf{x}_1^{(k+1)} = 1$, entonces $\mathbf{x}_1^{(k+2)} = \lambda_1$, etc.

Cuando $k \to \infty$ el factor de escala aproxima a λ_1, y el vector escalado a un vector propio asociado $\mathbf{x}_1$.

Apliquemos el algoritmo al cálculo del valor propio λ y del vector propio correspondiente $\mathbf{x}$ de la matriz simétrica

$$A = \begin{pmatrix} 1 & 1 & 2 \\ 2 & 1 & 1 \\ 1 & 1 & 3 \end{pmatrix} \tag{2.71}$$

asumiendo que la segunda componente de $\mathbf{x}$ es igual a la unidad.

Tomamos como estimador inicial $\mathbf{x}^{(0)} = (0,\ 1,\ 0)$ y aplicamos el esquema iterativo $\mathbf{x}^{(k+1)} = A\mathbf{x}^{(k)}$ con $k \geq 0$ manteniendo la notación del vector $\mathbf{x}^{(k)}$ antes y después del escalado de su segunda componente.

Dispongamos las sucesivas iterantes en la tabla siguiente

	$\mathbf{x}^{(0)}$	$\mathbf{x}^{(1)}$	$\mathbf{x}^{(2)}$		$\mathbf{x}^{(3)}$		$\mathbf{x}^{(4)}$		$\mathbf{x}^{(5)}$	
	0	1	4	1.00	4.50	0.9444	4.6503	1.0963	4.9268	1.0692
	1	1	4	1.00	4.25	1.0000	4.2418	1.0000	4.6079	1.0000
	0	1	5	1.25	5.75	1.3529	6.0033	1.4153	6.3421	1.3547
λ			4		4.25		4.2418		4.6079	

	$\mathbf{x}^{(6)}$		$\mathbf{x}^{(7)}$		$\mathbf{x}^{(8)}$		$\mathbf{x}^{(9)}$		$\mathbf{x}^{(10)}$	$\approx \mathbf{x}$
	4.7785	1.0635	4.7936	1.0671	4.8091	1.0675	4.8110	1.0675	4.8113	1.0675
	4.4931	1.0000	4.4921	1.0000	4.5052	1.0000	4.5067	1.0000	4.5070	1.0000
	6.1332	1.3650	6.1586	1.3710	6.1801	1.3718	6.1828	1.3719	6.1833	1.3719
λ	4.4921		4.4921		4.5052		4.5067		4.5070	

el resultado es con cuatro decimales, $\lambda = 4.5070$ y $\mathbf{x} = (1.0675,\ 1.0000,\ 1.3719)$.

Método de la potencia inversa

Se puede obtener el valor propio de menor valor absoluto de A y su vector propio asociado aplicando el método de la potencia a A^{-1} cuyos valores propios son, como sabemos, (2.1.4) los recíprocos de los valores de A.

El recíproco del menor valor propio de A en valor absoluto es el de mayor absoluto de A^{-1}.

En la práctica, se utiliza la descomposición $A = LU$ para resolver este problema en vez de calcular A^{-1}.

Una vez elegido el estimador inicial $\mathbf{x}^{(0)}$ se calcula $\mathbf{x}^{(1)}$ resolviendo el sistema $A\mathbf{x}^{(1)} = (LU)\mathbf{x}^{(1)} = \mathbf{x}^{(0)}$ como hicimos en la sección 2.3.2 seguido del escalado de la componente que hayamos decidido que sea la unidad.

Si A^{-1} no existe, 0 es el valor propio de menor valor absoluto y cualquier vector del núcleo de A se puede tomar como vector propio asociado.

El resto de los valores propios y de los vectores propios asociados se pueden obtener aplicando reiteradamente la siguiente idea. Una vez conocido el elemento propio $(\lambda_1, \mathbf{x}_1)$, se selecciona un estimador inicial que sea ortogonal a $\mathbf{x}_1$ y aplicando el método de las potencias se obtiene λ_2 y un vector propio asociado $\mathbf{x}_2$. Para obtener λ_3 se elige un vector inicial que sea ortogonal tanto a $\mathbf{x}_1$ como a $\mathbf{x}_2$ y se sigue el proceso (método de deflación).

PROBLEMAS

PROBLEMA 2.1 *Método de Gauss.*

Utilizar con detalle el método de eliminación de Gauss con estrategia de pivote parcial (cambio de filas) si fuera menester, para resolver el sistema lineal $A \cdot \mathbf{x} = \mathbf{b}$ con

$$A = \begin{pmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & -1 & -1 \\ 0 & -1 & 1 & 0 \\ 1 & -1 & 0 & 0 \end{pmatrix} \qquad \mathbf{b} = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \end{pmatrix}$$

Solución:

Escribimos la matriz aumentada

$$\left(\begin{array}{c|c} A & \mathbf{b} \end{array} \right) = A^{(1)} = \left(\begin{array}{cccc|c} 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & -1 & -1 & 0 \\ 0 & -1 & 1 & 0 & 1 \\ 1 & -1 & 0 & 0 & 0 \end{array} \right)$$

Eliminación de x_1 en las ecuaciones 2, 3 y 4 lo que se consigue con la operación elemental [13] de fila $F_{41}(-1)$

$$A^{(2)} = \left(\begin{array}{cccc|c} 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & -1 & -1 & 0 \\ 0 & -1 & 1 & 0 & 1 \\ 0 & -1 & 0 & -1 & 0 \end{array}\right)$$

Eliminación de x_2 en las ecuaciones 3 y 4 se consigue con las operaciones elementales de fila $F_{32}(1)$ y $F_{42}(1) = F_4 + F_2$

$$A^{(3)} = \left(\begin{array}{cccc|c} 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & -1 & -1 & 0 \\ 0 & 0 & 0 & -1 & 1 \\ 0 & 0 & -1 & -2 & 0 \end{array}\right)$$

Nos encontramos un pivote nulo en la fila F_3 por lo que la intercambiamos con la fila F_4 (operación elemental F_{34}) y cambiamos de signo ($F_3(-1)$ y $F_4(-1)$)

$$A^{(4)} = \left(\begin{array}{cccc|c} 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & -1 & -1 & 0 \\ 0 & 0 & 1 & 2 & 0 \\ 0 & 0 & 0 & 1 & -1 \end{array}\right)$$

que ya se puede resolver por sustitución hacia atrás.

$$\begin{aligned} x_4 = -1 & \\ x_3 + 2x_4 = 0 \quad &\Rightarrow \quad x_3 = -2 \\ x_2 - x_3 - x_4 = 0 \quad &\Rightarrow \quad x_2 = 1 \\ x_1 + x_4 = 0 \quad &\Rightarrow \quad x_1 = 1 \end{aligned}$$

PROBLEMA 2.2 *Herramientas básicas. Matrices de rotación elemental.*

Se llama matriz de **rotación elemental** de Jacobi, a la matriz real de tipo n, n ortogonal

$$\Omega_{(pq)} = \begin{pmatrix} 1 & & & & & & & & \\ & \ddots & & & & & & & \\ & & 1 & & & & & & \\ & & & \omega_{pp} & \cdots & \omega_{pq} & & & \\ & & & \vdots & 1 & \vdots & & & \\ & & & \omega_{qp} & \cdots & \omega_{qq} & & & \\ & & & & & & \ddots & & \\ & & & & & & & 1 & \end{pmatrix}$$

[13] Las transformaciones elementales de fila y columna son

1. Intercambio de las filas (resp. columnas) F_i y F_j (resp. C_i y C_j) que denotamos F_{ij} (resp. C_{ij}).
2. Producto de los elementos de la fila (resp. columna) i-ésima por un escalar k operación que denotamos $F_i(k)$ (resp. $C_i(k)$).
3. Suma a la fila (resp. columna) i-ésima del producto por k de la fila (resp. columna) j-ésima, operación que denotamos $F_{ij}(k)$ (resp. $C_{ij}(k)$). Se tiene $F_{ij}(k) = F_i + F_j(k)$.

Las matrices elementales de fila (resp. columna) son el resultado de aplicar a I_n las operaciones elementales de fila (resp. columna). Son matrices regulares y se cumplen las igualdades, $F_{ij}^{-1} = F_{ij}$, $F_i^{-1}(k) = F_i(1/k)$ $F_{ij}^{-1}(k) = F_{ij}(-k)$ y sus correspondientes para columna. Se obtienen las transformaciones elementales de fila (resp. columna) premultiplicando (resp. postmultiplicando) por las correspondientes matrices elementales.

donde $\omega_{pp} = \cos\theta$, $\omega_{pq} = -\sin\theta$, $\omega_{qp} = \sin\theta$, $\omega_{qq} = \cos\theta$ con $\theta \in [0, \pi]$ y en la que todos los elementos fuera de la diagonal principal, que no aparecen en la matriz son ceros.
Representaremos por simplicidad esa matriz en la forma

$$\Omega_{(pq)} = \begin{pmatrix} c & -s \\ s & c \end{pmatrix}$$

con $c = \cos\theta$ y $s = \sin\theta$.

1. - Comprobar que $\Omega_{(pq)}$ es ortogonal.
   - Dada una matriz A cuadrada de orden n simétrica, hallar un valor de θ que anule el elemento b_{pq} de la matriz $B = {\Omega_{(pq)}}^T A \Omega_{(pq)}$ transformada de A por $\Omega_{(pq)}$.
2. Aplicar una rotación elemental de Jacobi para eliminar el término $a_{12} = -1$ de la matriz

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

   calculando la matriz transformada.
3. Utilizar las rotaciones de Jacobi para hallar los valores propios de

$$A = \begin{pmatrix} 2 & \sqrt{3}/2 \\ \sqrt{3}/2 & 1 \end{pmatrix}$$

Solución:

1. - Multiplicando $\Omega_{(pq)}$ por su traspuesta se tiene

$$\Omega_{(pq)} \cdot \Omega_{(pq)}^T = \begin{pmatrix} c & -s \\ s & c \end{pmatrix} \cdot \begin{pmatrix} c & s \\ -s & c \end{pmatrix} = \begin{pmatrix} c^2+s^2 & cs-cs \\ sc-cs & c^2+s^2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = I_n$$

   - Sólo nos interesa el elemento b_{pq} de B, luego

$$b_{pq} = c(sa_{pp} + ca_{pq}) - s(a_{qp}s + ca_{qq}) = cs(a_{pp} - a_{qq}) + \\ + a_{pq}(c^2 - s^2) = (a_{pp} - a_{qq})\sin 2\theta/2 + a_{qp}\cos 2\theta$$

     y la condición $b_{pq} = 0$ implica que

$$(a_{qq} - a_{pp})\sin 2\theta = 2a_{qp}\cos 2\theta \quad \Rightarrow \quad \tan 2\theta = \frac{2a_{qp}}{a_{qq} - a_{pp}}$$

     Se utiliza una estrategia basada en una sucesión de este tipo de rotaciones planas para llevar A a la forma diagonal con los valores propios en la diagonal [19], [26], ver también la sección 2.1 del resumen teórico.

2. Queremos anular el elemento b_{12} y lo haremos calculando directamente el producto

$$B = {\Omega_{(12)}}^T A \Omega_{(12)} =$$

$$= \begin{pmatrix} c & s & 0 \\ -s & c & 0 \\ 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} 4 & -1 & 0 \\ -1 & 4 & -1 \\ 0 & -1 & 4 \end{pmatrix} \cdot \begin{pmatrix} c & -s & 0 \\ s & c & 0 \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 4-2sc/2 & s^2-c^2 & -s \\ s^2-c^2 & 4-2sc & -c \\ -s & -c & 4 \end{pmatrix}$$

   luego $s^2 - c^2 = 0 \quad \Rightarrow \quad \theta = \dfrac{\pi}{4}$ y $s = c = \sqrt{2}/2$ de donde

$$B = \begin{pmatrix} 3/2 & 0 & -\sqrt{2}/2 \\ 0 & 5 & -\sqrt{2}/2 \\ -\sqrt{2}/2 & -\sqrt{2}/2 & 4 \end{pmatrix}$$

3. Una sola rotación es suficiente para diagonalizar la matriz dada.
   Con $p = 1$, $q = 2$ tendremos

$$\tan 2\theta = \frac{2a_{12}}{a_{22} - a_{11}} = \frac{2\left(\sqrt{3}/2\right)}{2-1} = \sqrt{3} \quad \Rightarrow$$
$$\Rightarrow \quad 2\theta = \frac{\pi}{3} \quad \text{y} \quad \theta = \frac{\pi}{6}$$

   Con ello,

$$\begin{pmatrix} \sqrt{3}/2 & 1/2 \\ -1/2 & \sqrt{3}/2 \end{pmatrix} = \begin{pmatrix} 2 & \sqrt{3}/2 \\ \sqrt{3}/2 & c \end{pmatrix} \cdot \begin{pmatrix} \sqrt{3}/2 & -1/2 \\ -1/2 & \sqrt{3}/2 \end{pmatrix} = \begin{pmatrix} 5/2 & 0 \\ 0 & 1/2 \end{pmatrix}$$

   que son los valores propios pedidos.

PROBLEMA 2.3 *Métodos de Jacobi y Gauss-Seidel.*

Se considera el sistema de ecuaciones $(S) \quad A\mathbf{x} = \mathbf{b}$ con

$$A = \begin{pmatrix} 2 & -1 & 0 \\ -1 & 3 & -2 \\ 0 & -2 & 3 \end{pmatrix} \quad \text{y} \quad \mathbf{b} = \begin{pmatrix} 3 \\ 2 \\ 1 \end{pmatrix}$$

1. Convertir (S) en la iteración de punto fijo

$$\mathbf{x}^{(k+1)} = J\mathbf{x}^{(k)} + \mathbf{c}$$

   donde J es la matriz de iteración de Jacobi.

2. Convertir (S) en la iteración de punto fijo

$$\mathbf{x}^{(k+1)} = G\mathbf{x}^{(k)} + \mathbf{d}$$

   donde G es la matriz de iteración de Gauss-Seidel.

3. Estudiar la convergencia de ambos métodos.

4. Tomando como estimador inicial el vector

$$\mathbf{x}^{(0)} = \begin{pmatrix} 1 & 1 & 1 \end{pmatrix}^T$$

   Resolver (S) por ambos métodos haciendo un pequeño análisis comparativo de sus comportamientos respectivos.

Solución:
Sean D la matriz diagonal diag$(2, 3, 3)$, L y U como en la sección (2.4)

$$L = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 2 & 0 \end{pmatrix} \quad \text{y} \quad U = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 2 \\ 0 & 0 & 0 \end{pmatrix}$$

1. En el método de Jacobi la sucesión iterante es

$$D\mathbf{x}^{(k+1)} = (L + U)\mathbf{x}^{(k)} + \mathbf{b}$$

   de donde $J = D^{-1} \cdot (L + U)$ y $\mathbf{c} = D^{-1}\mathbf{b}$

$$\Rightarrow \quad \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}^{(k+1)} = \begin{pmatrix} 0 & 0.5000 & 0 \\ 0.3333 & 0 & 0.6667 \\ 0 & 0.6667 & 0 \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}^{(k)} + \begin{pmatrix} 1.5000 \\ 0.6667 \\ 0.3333 \end{pmatrix}$$

2. En el método de Gauss-Seidel

$$(D-L)\mathbf{x}^{(k+1)} = U\mathbf{x}^{(k)} + \mathbf{b}$$

de donde $G = (D-L)^{-1}\cdot U$ y $\mathbf{d} = (D-L)^{-1}\mathbf{b}$

$$\Rightarrow \quad \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}^{(k+1)} = \begin{pmatrix} 0 & 0.5000 & 0 \\ 0 & 0.1667 & 0.6667 \\ 0 & 0.1111 & 0.4444 \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}^{(k)} + \begin{pmatrix} 1.5000 \\ 1.6667 \\ 1.1111 \end{pmatrix}$$

3. Es fácil calcular los valores propios de ambas matrices de iteración y su radio espectral.

   En el caso de Jacobi se tiene $\rho_J = 0.7817$ y en el Gauss-Seidel $\rho(G) = 0.6111$ ambos estrictamente menores que 1, luego los dos métodos son convergentes siendo mayor la velocidad de convergencia en el de Gauss-Seidel como es habitual.

4. Hemos escrito un código Matlab para cada una de las sucesiones iterantes. Incluimos aquí el de Gauss-Seidel

```
G=[ 0   0.5000         0;
    0   0.1667       0.6667;
    0   0.1111       0.4444];
C= [1.5000;
   1.16667;
   1.1111];
S=[1 1 1]';
delta=1e-6;
eps=1e-6;
max1=40;
X=zeros(1,3);
N=length(C);
for j=1:N
    X(j)=G(j,1:N)*S(1:N)+C(j)
    for k=2:max1
    for j=1:N
    X(j)=G(j,1:N)*(X(1:N))'+C(j)
    end
end
err=abs(norm(X'-S));
relerr=err/(norm(X)+eps);
S=X';
   if (err<delta)|(relerr<delta)
       break
   end
end
X=X'
[k err relerr]
```

   y ponemos ambos *gseid1.m* y *jacobi.m* en la página web vinculada al libro.

5. Los resultados finales de ambos códigos en *format long* son:

   • Jacobi

$$\mathbf{x}^* = \begin{pmatrix} 3.28571428571429 \\ 3.57142857142859 \\ 2.71428571428573 \end{pmatrix}$$

   tras 40 iteraciones siendo las medidas de convergencia definidas err $= 2.6\cdot 10^{-8}$ y relerr $= 4.7\cdot 10^{-9}$.

• Gauss-Seidel

$$\mathbf{x}^* = \begin{pmatrix} 3.28571428571429 \\ 3.57142857142858 \\ 2.71428571428568 \end{pmatrix}$$

tras 40 iteraciones siendo las medidas de convergencia definidas err $= 5.3 \cdot 10^{-10}$ y relerr $= 9.6 \cdot 10^{-11}$.

A igual número de iteraciones la precisión del método de Gauss-Seidel es mayor.

PROBLEMA 2.4 *Método de la potencia.*

Determinar el mayor valor propio λ y el vector propio correspondiente $\mathbf{x}$ de la matriz simétrica considerada en el resumen teórico asumiendo que la primera componente de $\mathbf{x}$ es igual a la unidad.

$$A = \begin{pmatrix} 1 & 1 & 2 \\ 2 & 1 & 1 \\ 1 & 1 & 3 \end{pmatrix}$$

Solución:

Tomamos como estimador inicial $\mathbf{x}^{(0)} = (1,\ 0,\ 0)$ y aplicamos el esquema iterativo $\mathbf{x}^{(k+1)} = A\mathbf{x}^{(k)}$ con $k \geq 0$ manteniendo la notación del vector $\mathbf{x}^{(k)}$ antes y después del escalado de su primera componente.

Dispongamos las sucesivas iterantes en la tabla siguiente

	$\mathbf{x}^{(0)}$	$\mathbf{x}^{(1)}$	$\mathbf{x}^{(2)}$		$\mathbf{x}^{(3)}$		$\mathbf{x}^{(4)}$		$\mathbf{x}^{(5)}$		$\mathbf{x}^{(6)}$		$\mathbf{x}^{(7)}$
	1	1	5	1.00	4.40	1.0000	4.5000	1.0000	4.5050	1.0000	4.5068	1.0000	4.5069
	0	2	5	1.00	4.20	0.9545	4.2273	0.9394	4.2222	0.9372	4.2220	0.9368	4.2219
	0	1	6	1.20	5.60	1.2727	5.7727	1.2828	5.7879	1.2848	5.7915	1.2851	5.7920
λ			5		4.40		4.5000		4.5050		4.5068		4.5069

el resultado es con cuatro decimales, $\lambda = 4.5069$ y $\mathbf{x} = (1.0000,\ 0.9368,\ 1.2851)$.

Los vectores obtenidos son proporcionales siendo la constante de proporcionalidad 0.9368 con un error menor que 0.6910^{-4}. Obsérvese que la convergencia ha sido mucho más rápida que en el caso de la teoría.

PROBLEMA 2.5 *Resolución de un sistema lineal mediante esquemas iterativos.*

Se pretende resolver el sistema lineal $A \cdot \mathbf{x} = \mathbf{b}$ con

$$A = \begin{pmatrix} -108810.539 & 1 & 0 & 0 \\ -27412.782 & 1 & 0 & 0 \\ -1 & 0 & -1 & 0 \\ -2 & 0 & 0 & -1 \end{pmatrix} \qquad \mathbf{b} = \begin{pmatrix} 0 \\ 44.644 \\ 0 \\ 0 \end{pmatrix}$$

Dar un paso con el método iterativo que se considere más interesante, dada la forma del sistema. No se podrán realizar transformaciones elementales en el sistema lineal, y se tomará como estimador inicial un vector cuyas componentes son todas la unidad. Valorar la convergencia, a partir del resultado obtenido.

Solución:

En principio es más conveniente elegir el método de Gauss-Seidel porque sus propiedades de convergencia son "a priori" mejores que las del método de Jacobi. Además, en este caso, la parte de la matriz por encima de la diagonal principal sólo tiene un elemento no nulo, lo que facilita la decisión.

Descompongamos A de acuerdo con el método de Gauss-Seidel

$$A = (D - L) - U = M - N$$

con

$$N = U = \begin{pmatrix} 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \quad \text{y} \quad M = D - L = \begin{pmatrix} -108810.539 & 0 & 0 & 0 \\ -27412.782 & 1 & 0 & 0 \\ -1 & 0 & -1 & 0 \\ -2 & 0 & 0 & -1 \end{pmatrix}$$

El esquema iterativo tiene la forma

$$M\mathbf{x}^{(k+1)} = N\mathbf{x}^{(k)} + \mathbf{b}$$

El primer paso será

$$\begin{pmatrix} -108810.539 & 0 & 0 & 0 \\ -27412.782 & 1 & 0 & 0 \\ -1 & 0 & -1 & 0 \\ -2 & 0 & 0 & -1 \end{pmatrix} \begin{pmatrix} x_1^{(1)} \\ x_2^{(1)} \\ x_3^{(1)} \\ x_4^{(1)} \end{pmatrix} = \begin{pmatrix} 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} + \begin{pmatrix} 0 \\ 44.644 \\ 0 \\ 0 \end{pmatrix} \quad \Rightarrow$$

$$\Rightarrow \quad \begin{pmatrix} -108810.539 & 0 & 0 & 0 \\ -27412.782 & 1 & 0 & 0 \\ -1 & 0 & -1 & 0 \\ -2 & 0 & 0 & -1 \end{pmatrix} \begin{pmatrix} x_1^{(1)} \\ x_2^{(1)} \\ x_3^{(1)} \\ x_4^{(1)} \end{pmatrix} = \begin{pmatrix} -1 \\ 44.644 \\ 0 \\ 0 \end{pmatrix}$$

Por sustitución hacia adelante, obtenemos:

$$x^{(1)} = \begin{pmatrix} 9.19\ 10^{-6} \\ 44.8959 \\ -9.19\ 10^{-6} \\ -1.838\ 10^{-5} \end{pmatrix}$$

Estudiamos simultáneamente la sucesión $\{\mathbf{r}^{(k)}\}$ de los residuos para ir controlando paso a paso la convergencia del esquema. El esquema será convergente si la sucesión $\{\|\mathbf{r}^{(k)}\|_\infty\} \to 0$ y cuanto menor sea el tamaño $\|\mathbf{r}^{(k)}\|_\infty$ del residuo, más cerca estamos de la solución, de residuo nulo. O sea, calculamos

$$\mathbf{r}^{(k)} = \mathbf{b} - A\mathbf{x}^{(k)}$$

$$\mathbf{r}^{(0)} = \begin{pmatrix} 0 \\ 44.644 \\ 0 \\ 0 \end{pmatrix} - A \cdot \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} -108809.539 \\ -27456.42 \\ -2 \\ -3 \end{pmatrix} \quad \Rightarrow \quad \| \mathbf{r}^{(0)} \|_\infty = 1.088\ 10^5$$

$$\mathbf{r}^{(1)} = \begin{pmatrix} 0 \\ 44.644 \\ 0 \\ 0 \end{pmatrix} - A \cdot \begin{pmatrix} 9.19\ 10^{-6} \\ 44.8959 \\ -9.19\ 10^{-6} \\ -1.838\ 10^{-5} \end{pmatrix} = \begin{pmatrix} 43.8959313 \\ 0 \\ 0 \\ 0 \end{pmatrix} \quad \Rightarrow \quad \| \mathbf{r}^{(1)} \|_\infty = 43.8959313$$

Como vemos, el tamaño del residuo disminuye. De hecho, si seguimos iterando sobre este esquema, podremos comprobar que convergemos a la solución del mismo:

$$\mathbf{x}^{(2)} = \begin{pmatrix} 4.1260\ 10^{-4} \\ 55.9546 \\ -4.1260\ 10^{-4} \\ -8.2521\ 10^{-4} \end{pmatrix} \quad ; \quad \| \mathbf{r}^{(2)} \|_\infty = 11.058$$

$$\mathbf{x}^{(3)} = \begin{pmatrix} 5.1423\ 10^{-4} \\ 58.7407 \\ -5.1423\ 10^{-4} \\ -1.0284\ 10^{-3} \end{pmatrix} \quad ; \quad \| \mathbf{r}^{(3)} \|_\infty = 2.786$$

.....

La solución es:

$$\mathbf{x} = \begin{pmatrix} 5.4846\ 10^{-4} \\ 59.6790 \\ -5.4846\ 10^{-4} \\ -1.0969\ 10^{-3} \end{pmatrix}$$

El estudio teórico de la convergencia se hace analizando el radio espectral de la matriz de iteración $M^{-1}N = (D-L)^{-1}U$. Tenemos

$$(D-L)^{-1}U = \begin{pmatrix} 0 & 9.19\ 10^{-6} & 0 & 0 \\ 0 & 0.251931 & 0 & 0 \\ 0 & -9.19\ 10^{-6} & 0 & 0 \\ 0 & -1.8380\ 10^{-5} & 0 & 0 \end{pmatrix}$$

cuyo radio espectral es obviamente 0.25193, que es bastante menor que 1, de donde la rápida convergencia del método. Existe una descomposición por bloques admisible de la matriz A, que conduce a la solución en el primer paso, al ser el bloque superior nulo:

$$U = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} \quad \text{y} \quad M = D - L = \begin{pmatrix} -108810.539 & 1 & & \\ -27412.782 & 1 & & \\ -1 & 0 & -1 & 0 \\ -2 & 0 & 0 & -1 \end{pmatrix}$$

Repitamos el ejemplo utilizando ahora el método de Jacobi. Se tiene sucesivamente

$$A = D - (L+U) = M - N$$

$$N = L + U = \begin{pmatrix} 0 & -1 & 0 & 0 \\ 27412.782 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 2 & 0 & 0 & 0 \end{pmatrix} \quad \text{y} \quad M = D = \begin{pmatrix} -108810.539 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix}$$

$$M\mathbf{x}^{(1)} = N\mathbf{x}^{(0)} + \mathbf{b}$$

$$\begin{pmatrix} -108810.539 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix} \begin{pmatrix} x_1^{(1)} \\ x_2^{(1)} \\ x_3^{(1)} \\ x_4^{(1)} \end{pmatrix} = \begin{pmatrix} 0 & -1 & 0 & 0 \\ 27412.782 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 2 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} + \begin{pmatrix} 0 \\ 44.644 \\ 0 \\ 0 \end{pmatrix} \quad \Rightarrow$$

$$\Rightarrow \quad \begin{pmatrix} -108810.539 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix} \begin{pmatrix} x_1^{(1)} \\ x_2^{(1)} \\ x_3^{(1)} \\ x_4^{(1)} \end{pmatrix} = \begin{pmatrix} -1 \\ 27457.426 \\ 1 \\ 2 \end{pmatrix}$$

Resolviendo este sistema diagonal, obtenemos:

$$\mathbf{x}^{(1)} = \begin{pmatrix} 9.19\ 10^{-6} \\ 27457.426 \\ -1 \\ -2 \end{pmatrix}$$

Repitamos el cálculo de la sucesión $\{\|\mathbf{r}^{(k)}\|_\infty\}$ de los tamaños de los residuos. La norma del residuo del estimador inicial es por supuesto la misma, $\| \mathbf{r}^{(0)} \|_\infty = 1.088\ 10^5$, y la del primer vector calculado $\mathbf{x}^{(1)}$

$$\| \mathbf{r}^{(1)} \|_\infty = 2.7456\ 10^4$$

menor que $\| \mathbf{r}^{(0)} \|_\infty$, aunque mucho mayor que el correspondiente término de la sucesión asociada al método de Gauss-Seidel. El radio espectral de la matriz de Jacobi $M^{-1}N = D^{-1}(L+U)$

$$D^{-1}(L+U) = \begin{pmatrix} 0 & 9.19\ 10^{-6} & 0 & 0 \\ 2.7412\ 10^{4} & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ -2 & 0 & 0 & 0 \end{pmatrix}$$

es 0.50192, bastante menor que 1, pero mayor que 0.2519 que era el radio espectral de la matriz de Gauss-Seidel, lo que justifica la inferior velocidad de convergencia.

PROBLEMA 2.6 *Condicionamiento de un sistema lineal.*

Se considera el número real α, con $\alpha \neq \pm 1$. Sea la matriz:

$$A(\alpha) = \begin{pmatrix} \alpha & 1 \\ 1 & \alpha \end{pmatrix}$$

Se pide

1. Calcular el condicionamiento de la matriz A en la norma 2 utilizando sus propiedades de simetría y representar la curva que tiene como abscisa a α y como ordenada a $\text{cond}_2(A(\alpha))$.
2. Resolver en función de α el sistema lineal $A(\alpha)\mathbf{x} = \mathbf{b}$ con $\mathbf{b} = (1, -1)^T$.
3. Se supone que el término independiente del sistema lineal del apartado anterior transporta un error $\delta\mathbf{b}$ cuyas componentes en valor absoluto pertenecen al intervalo $[0.03, 0.04]$. Sea $\mathbf{x} + \delta\mathbf{x}$ la solución del sistema perturbado $A(\alpha)(\mathbf{x} + \delta\mathbf{x}) = \mathbf{b} + \delta\mathbf{b}$. Acotar el valor $\|\delta\mathbf{x}\|_2$ en función del parámetro α.
4. Se supone que $\alpha = 0.5$.
   a) Verificar en este caso que se cumple la cota dada en el apartado 3 tomando $\delta\mathbf{b} = (0.03, -0.04)^T$.
   b) Estudiar la posible convergencia del método de Jacobi para resolver un sistema lineal cuya matriz fuera $A(0.5)$.

Solución:

1. Ya que $A(\alpha)$ es una matriz simétrica, el condicionamiento en la norma 2 es (ver (2.23) en la sección 2.2.)

$$\text{cond}_2(A(\alpha)) = \frac{|\lambda_{max}|}{|\lambda_{min}|}$$

   Para calcular los autovalores de $A(\alpha)$ resolvemos la ecuación

$$\begin{vmatrix} \alpha - \lambda & 1 \\ 1 & \alpha - \lambda \end{vmatrix} = 0 \quad \Rightarrow \quad \begin{array}{l} \lambda_1 = \alpha - 1 \\ \lambda_2 = \alpha + 1 \end{array}$$

   y por tanto

$$\begin{aligned} \text{cond}_2(A(\alpha)) &= \frac{|\lambda_{max}|}{|\lambda_{min}|} = \text{máx}\left\{ \frac{|\alpha+1|}{|\alpha-1|}, \frac{|\alpha-1|}{|\alpha+1|} \right\} = \\ &= \begin{cases} |\alpha-1|/|\alpha+1|, & \alpha \leq 0 \\ |\alpha+1|/|\alpha-1|, & \alpha > 0 \end{cases} \end{aligned}$$

   Representamos en la Figura 2.1 uno de los dos cocientes y en la Figura 2.2 la curva de los máximos, que tiene dos asíntotas verticales en los puntos correspondientes a $\alpha = \pm 1$.

   Incluimos el código Matlab que genera cada uno de esos gráficos

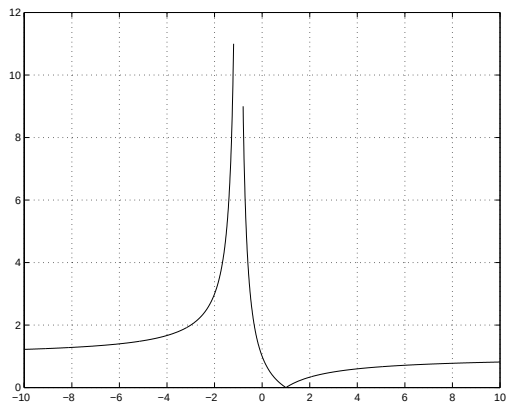


Figura 2.1: Gráfico de $\alpha \to \frac{|\alpha-1|}{|\alpha+1|}$.

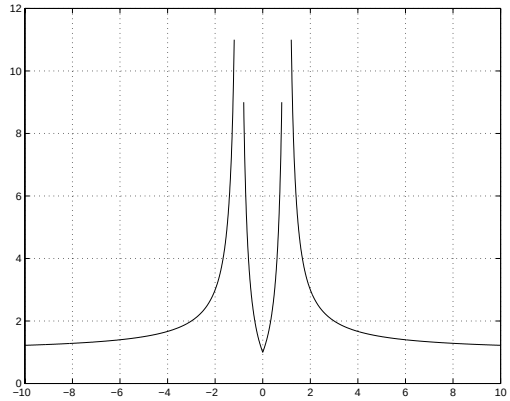


Figura 2.2: Condicionamiento en norma 2 de la matriz A.

```
%%%%%%%%%%%% |lambda1|/|lambda2|
clear
hold off;
alpha1=-10:0.01:-1.2;
alpha2=-0.8:0.01:10;
plot(alpha1,abs(alpha1-1)./abs(alpha1+1));
hold on;
plot(alpha2,abs(alpha2-1)./abs(alpha2+1));
grid;
pause;
%%%%%% cond2(A)
hold off;
clear;
alpha1=-10:0.01:-1.2;
alpha2=-0.80:0.01:0.80;
alpha3=1.2:0.01:10;
cond2_1=max(abs(alpha1+1)./abs(alpha1-1),abs(alpha1-1)./abs(alpha1+1));
cond2_2=max(abs(alpha2+1)./abs(alpha2-1),abs(alpha2-1)./abs(alpha2+1));
cond2_3=max(abs(alpha3+1)./abs(alpha3-1),abs(alpha3-1)./abs(alpha3+1));
plot(alpha1,cond2_1,'b',alpha2,cond2_2,'b',alpha3,cond2_3,'b');
grid;
pause;
%%%%%%% norma2 (delta x)
deltax2_1=[0.05*cond2_1./abs(1-alpha1)];
deltax2_2=[0.05*cond2_2./abs(1-alpha2)];
deltax2_3=[0.05*cond2_3./abs(1-alpha3)];
plot(alpha1,deltax2_1,'b',alpha2,deltax2_2,'b',alpha3,deltax2_3,'b');
grid;
```

2. Se tiene
$$\begin{pmatrix} \alpha & 1 \\ 1 & \alpha \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 \\ -1 \end{pmatrix} \quad \Rightarrow \quad \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \frac{1}{1-\alpha} \begin{pmatrix} -1 \\ 1 \end{pmatrix}$$

3. Consideremos ahora el sistema lineal perturbado
$$A(\alpha)(\mathbf{x} + \delta\mathbf{x}\mathbf{x}) = \mathbf{b} + \delta\mathbf{b}$$
Sabemos que (ver la sección 2.2)
$$\frac{\|\delta\mathbf{x}\|_2}{\|\mathbf{x}\|_2} \leq \mathrm{cond}_2(A) \frac{\|\delta\mathbf{b}\|_2}{\|\mathbf{b}\|_2}$$

de donde

$$\|\delta\mathbf{x}\|_2 \leq \mathrm{cond}_2(A)\frac{\|\delta\mathbf{b}\|_2}{\|\mathbf{b}\|_2}\|\mathbf{x}\|_2$$

En nuestro caso,

$$\mathbf{x} = \frac{1}{1-\alpha}\begin{pmatrix} -1 \\ 1 \end{pmatrix} \quad \text{luego} \quad \|\mathbf{x}\|_2 = \frac{\sqrt{2}}{|1-\alpha|}$$

$\|\delta\mathbf{b}\|_2 \leq \|(0.03, 0.04)\|_2 = 0.05$ y $\|\mathbf{b}\|_2 = \sqrt{2}$. Con todo ello,

$$\|\delta\mathbf{x}\|_2 \leq \frac{0.05}{|1-\alpha|}\,\text{máx}\left\{\frac{|\alpha+1|}{|\alpha-1|}, \frac{|\alpha-1|}{|\alpha+1|}\right\}$$

En el gráfico de la Figura 2.3 correspondiente a la variación con α de esta mayorante del valor $\|\delta\mathbf{x}\|_2$, se observa que en las zonas donde el condicionamiento es bajo, esa cota es pequeña, y por tanto $\|\delta\mathbf{x}\|_2$ es bajo y que debido al denominador $|1-\alpha|$, la mayoración es peor en el entorno de la asíntota correspondiente a $\alpha = 1$.

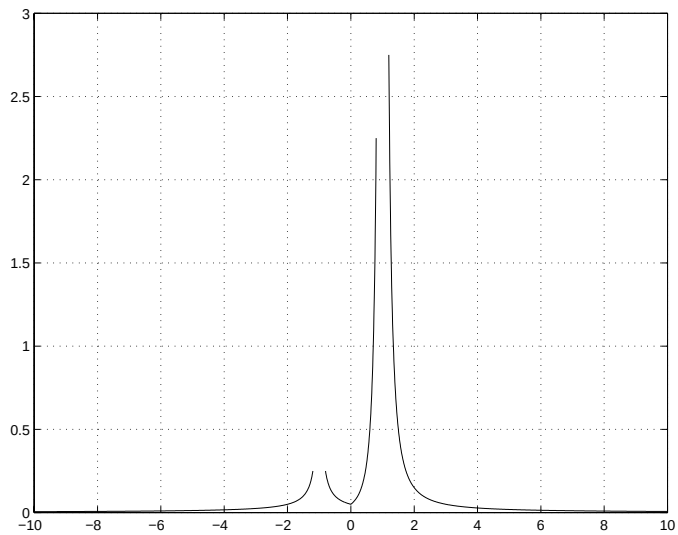


Figura 2.3: Cota correspondiente a $\|\delta\mathbf{x}\|_2$.

4. La solución del sistema para $\alpha = 0.5$ es $\mathbf{x} = (-2\ \ 2)^T$.

   a) Con los datos del enunciado el sistema perturbado es

   $$\begin{pmatrix} 0.5 & 1 \\ 1 & 0.5 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1.03 \\ -1.04 \end{pmatrix}$$

   cuya solución es $(-2.0733, 2.0667)$. Comparándola a la solución $(-2, 2)$ del sistema no perturbado se observa que $\delta\mathbf{x} = (-0.0733, 0.0667)$ y $\|\delta\mathbf{x}\|_2 = 0.0991$. La cota que obtendríamos para $\alpha = 0.5$ es 0.3, que no se alcanza.

   b) El método iterativo de Jacobi se enfrenta a graves problemas en este caso, pues el radio espectral de la matriz de iteración B es 2, como se puede comprobar calculando sus autovalores

   $$B = M^{-1}N = D^{-1}(L+U)\begin{pmatrix} 0.5 & 0 \\ 0 & 0.5 \end{pmatrix}^{-1}\begin{pmatrix} 0 & -1 \\ -1 & 0 \end{pmatrix} = \begin{pmatrix} 0 & -2 \\ -2 & 0 \end{pmatrix}$$

PROBLEMA 2.7 *Convergencia de esquemas iterativos para una matriz tridiagonal.*

Se considera una matriz tridiagonal de 3×3 del tipo siguiente:

$$A = \begin{pmatrix} 1 & a_{12} & 0 \\ a_{21} & 1 & a_{23} \\ 0 & a_{32} & 1 \end{pmatrix}$$

Se pide:

1. Estudiar si los métodos de Jacobi y Gauss-Seidel para A convergen o divergen simultáneamente.
2. Encontrar una condición necesaria y suficiente, expresada mediante los elementos de la matriz A, para que ambos métodos converjan.
3. En el caso de que ambos métodos converjan, ¿cuál lo hace más rápidamente? Justificar la respuesta.
4. Suponiendo que $a_{12} = a_{21} = a_{32} = a_{23} = 0.5$ y que el término independiente del sistema lineal es $\mathbf{b} = (1.5, 2.0, 1.5)^T$, y tomando como estimador inicial $\mathbf{b}$, comprobar que se verifica el apartado 3 utilizando el residuo de la primera iteración para ambos métodos.
5. Si $a_{12} = a_{21} = 0.5$ y $a_{32} = e^{a_{23}}$, ¿cuánto debe valer a_{23} para que la velocidad de convergencia del método de Gauss-Seidel sea máxima?

   Si en la respuesta fuera necesario resolver una ecuación no lineal, se hará mediante el método de Newton.
6. Para ese valor de a_{23} y tomando el mismo $\mathbf{b}$ y estimador inicial que en 4, ¿cuántos pasos son necesarios para tener un error menor que 0.001?
7. Si $a_{12} = a_{21} = 0.5$ y $a_{32} = e^{a_{23}}$, ¿cuánto debe valer a_{23} para que la velocidad de convergencia de Jacobi sea máxima?

   En caso de necesidad se usará como antes, el método de Newton.

Solución:

1. En la descomposición de la matriz A asociada al método de Jacobi (ver la sección 2.4)

   $$M = I; \quad M - N = A; \quad N = I - A$$

   de donde la matriz de iteración de Jacobi $J = M^{-1}N$ es

   $$M^{-1}N = I(I - A) = I - A = \begin{pmatrix} 0 & -a_{12} & 0 \\ -a_{21} & 0 & -a_{23} \\ 0 & -a_{32} & 0 \end{pmatrix}$$

   Estudiemos sus autovalores:

   $$\begin{vmatrix} -\lambda & -a_{12} & 0 \\ - & -\lambda & -a_{23} \\ 0 & -a_{32} & -\lambda \end{vmatrix} = 0 \quad \Rightarrow \quad \lambda_1 = 0, \quad \lambda_2 = -\lambda_3 = \sqrt{a_{12}a_{21} + a_{23}a_{32}}$$

   y por tanto, el radio espectral vale:

   $$\rho(J) = |\sqrt{a_{12}a_{21} + a_{23}a_{32}}|$$

   En la descomposición de la matriz A asociada al método de Gauss-Seidel

   $$M = \begin{pmatrix} 1 & 0 & 0 \\ a_{21} & 1 & 0 \\ 0 & a_{32} & 1 \end{pmatrix} \quad \Rightarrow \quad M^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ -a_{21} & 1 & 0 \\ a_{21}a_{32} & -a_{32} & 1 \end{pmatrix} \text{ y } N = \begin{pmatrix} 0 & -a_{12} & 0 \\ 0 & 0 & -a_{23} \\ 0 & 0 & 0 \end{pmatrix}$$

de modo que la matriz de iteración de Gauss-Seidel, $GS = M^{-1}N$. es

$$M^{-1}N = \begin{pmatrix} 0 & -a_{12} & 0 \\ 0 & a_{21}a_{12} & -a_{23} \\ 0 & -a_{21}a_{12}a_{32} & a_{32}a_{23} \end{pmatrix}$$

Calculando sus autovalores:

$$\begin{vmatrix} -\lambda & -a_{12} & 0 \\ 0 & a_{21}a_{12} - \lambda & -a_{23} \\ 0 & -a_{21}a_{12}a_{32} & a_{32}a_{23} - \lambda \end{vmatrix} = 0 \quad \Rightarrow \quad \lambda_1 = \lambda_2 = 0, \quad \lambda_3 = a_{12}a_{21} + a_{23}a_{32}$$

se tiene su radio espectral

$$\rho(GS) = |a_{12}a_{21} + a_{23}a_{32}|$$

es decir,

$$\rho(J) = \sqrt{\rho(GS)}$$

La condición necesaria y suficiente para la convergencia es que el radio espectral de la matriz de iteración sea menor que la unidad. Dada la relación entre ellos, es claro que las dos serán mayores o menores que la unidad simultáneamente y por tanto convergerán o divergerán simultáneamente.

2. La condición necesaria y suficiente pedida es

$$|a_{12}a_{21} + a_{23}a_{32}| < 1$$

3. Caso de que haya convergencia, los dos radios espectrales son menores que la unidad, luego $\rho(J) \geq \rho(GS)$ y por tanto el método de Jacobi convergerá más lentamente.

4. Si $a_{12} = a_{21} = a_{32} = a_{23} = 0.5$,

$$\rho(J) = 0.7071, \quad \rho(GS) = 0.5$$

En el caso de Jacobi, como la matriz M es directamente la matriz unidad

$$\mathbf{x}^{(k+1)} = Nx^{(k)} + b$$

luego

$$\mathbf{x}^{(1)} = N\mathbf{x}^{(0)} + \mathbf{b} = (N + I)\mathbf{b} = \begin{pmatrix} 1 & -0.5 & 0 \\ -0.5 & 1 & -0.5 \\ 0 & -0.5 & 1 \end{pmatrix} \begin{pmatrix} 1.5 \\ 2.0 \\ 1.5 \end{pmatrix} = \begin{pmatrix} 0.5 \\ 0.5 \\ 0.5 \end{pmatrix}$$

Evaluemos el residuo:

$$\mathbf{r}^{(1)} = \mathbf{b} - Ax^{(1)} = \begin{pmatrix} -0.75 \\ -1.00 \\ -0.75 \end{pmatrix}, \quad \|r^{(1)}\|_\infty = 1.0$$

En el caso de Gauss-Seidel:

$$M\mathbf{x}^{(k+1)} = N\mathbf{x}^{(k)} + \mathbf{b}$$

$$M\mathbf{x}^{(1)} = N\mathbf{x}^{(0)} + \mathbf{b} = (N + I)\mathbf{b} = \begin{pmatrix} 1 & -0.5 & 0 \\ 0 & 1 & -0.5 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1.5 \\ 2.0 \\ 1.5 \end{pmatrix} = \begin{pmatrix} 0.5 \\ 1.25 \\ 1.5 \end{pmatrix} \quad \Rightarrow$$

$$\Rightarrow \quad \begin{pmatrix} 1 & 0 & 0 \\ 0.5 & 1 & 0 \\ 0 & 0.5 & 1 \end{pmatrix} \mathbf{x}^{(1)} = \begin{pmatrix} 0.5 \\ 1.25 \\ 1.5 \end{pmatrix}, \quad \Rightarrow \quad \mathbf{x}^{(1)} = \begin{pmatrix} 0.5 \\ 1.0 \\ 1.0 \end{pmatrix}$$

Evaluemos el residuo:

$$\mathbf{r}^{(1)} = \mathbf{b} - A\mathbf{x}^{(1)} = \begin{pmatrix} 0.50 \\ -0.25 \\ 0.00 \end{pmatrix}, \quad \|\mathbf{r}^{(1)}\|_\infty = 0.5$$

Vemos que el residuo de GS es menor que el de Jacobi al final de la primera iteración, como era de esperar dada la diferencia de la velocidad de convergencia en ambos métodos.

5. En el caso de que $a_{12} = a_{21} = 0.5$ y que $a_{32} = e^{a_{23}}$, tendremos que

$$\rho(GS) = |0.25 + a_{23}e^{a_{23}}|$$

La velocidad de convergencia óptima se consigue cuando el radio espectral es lo más pequeño posible.

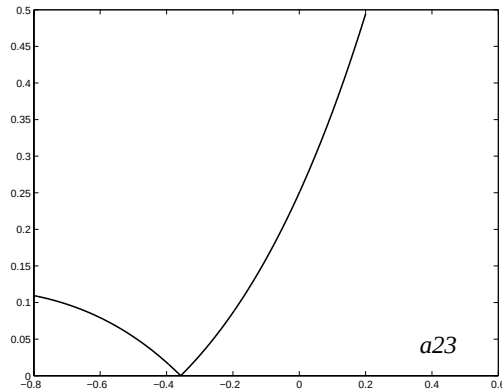


Figura 2.4: Apartado 5, problema 2.7.

Hay que buscar, por tanto, el valor de a_{23} que hace mínimo el radio espectral [14], luego tal que

$$0.25 + a_{23}e^{a_{23}} = 0$$

Expresión a la que podemos aplicar directamente el método de Newton:

$$a_{23}^{(k+1)} = a_{23}^{(k)} - \frac{0.25 + a_{23}^{(k)}e^{a_{23}{}^{(k)}}}{e^{a_{23}{}^{(k)}} + a_{23}^{(k)}e^{a_{23}{}^{(k)}}}$$

Podemos utilizar $a_{23}^{(0)} = 0$ como estimador inicial, ya que en este punto se anula la función $x \to xe^x$ próxima a la nuestra.

Los valores que vamos obteniendo son

k	0	1	2	3	4	5
$a_{23}^{(k)}$	0.000000	-0.250000	-0.344675	-0.357199	-0.357402	-0.357402

Por tanto, $a_{23} = -0.357402$. Para hacer estas iteraciones en Matlab, basta hacer un pequeño fichero con las siguientes sentencias, y repetirla tantas veces como iteraciones queramos hacer.

```
format long;
x=0;
x=x-(0.25+x*exp(x))/(exp(x)+x*exp(x))
x=x-(0.25+x*exp(x))/(exp(x)+x*exp(x))
x=x-(0.25+x*exp(x))/(exp(x)+x*exp(x))
```

6. Al ser el radio espectral 0, la convergencia se produce directamente en una sola iteración.

7. Dada la relación que existe entre los radios espectrales de Jacobi y Gauss-Seidel, el valor que obtendremos para a_{23} será el mismo en este caso.

[14] El valor mínimo de $0 \leq \rho(GS) = |0.25 + a_{23}e^{a_{23}}|$ se obtiene para el valor de a_{23} que anula el valor absoluto. Representando gráficamente la función $x \to xe^x + 0.25$ dicho valor es la abscisa del punto de corte con el eje x (ver Figura 2.4).

PROBLEMA 2.8 *Valores propios de una matriz perturbada.*

Dada una matriz cuadrada $A \in \mathcal{M}_n(\mathbb{R})$ llamaremos **perturbación de** A una matriz de la forma ϵB donde ϵ es un escalar arbitrariamente pequeño y B una matriz de $\mathcal{M}_n(\mathbb{R})$.
El objetivo del problema es hacer en varios casos un estudio comparativo de los valores propios de A y de su **matriz perturbada** $A^\epsilon = A + \epsilon B$.

1. Se definen aquí

$$A_1 = \begin{pmatrix} 1 & 2 \\ 0 & 2 \end{pmatrix} \qquad \text{y} \qquad B_1 = \begin{pmatrix} 1 & \frac{1}{2} \\ \frac{1}{2} & 1 \end{pmatrix}$$

   a) Demostrar que los valores propios de A_1^ϵ son

$$\begin{cases} 1 + C_1\epsilon + O(\epsilon) \\ 2 + C_2\epsilon + O(\epsilon) \end{cases}$$

   para ϵ suficientemente pequeño [15].
   Se hará un desarrollo limitado de esos valores propios en el entorno de $\epsilon = 0$, determinando las constantes C_1 y C_2.

   b) Comprobar el resultado para $\epsilon = 10^{-4}$ efectuando los cálculos con cinco dígitos. Hallar la distancia entre los valores propios de A_1 y de A_1^ϵ.

2. Sean ahora

$$A_2 = \begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix} \qquad \text{y} \qquad B_2 = \begin{pmatrix} 1 & \frac{1}{2} \\ \frac{1}{2} & 1 \end{pmatrix} \tag{2.72}$$

   a) Demostrar que los valores propios de A_2^ϵ son

$$\begin{cases} 1 + C_1\sqrt{\epsilon} + O(\sqrt{\epsilon}) \\ 1 + C_2\sqrt{\epsilon} + O(\sqrt{\epsilon}) \end{cases}$$

   para ϵ suficientemente pequeño, determinando las constantes C_1 y C_2.

   b) Comprobar el resultado para $\epsilon = 10^{-4}$ efectuando los cálculos con cinco dígitos.

3. Sea

$$A_3 = \begin{pmatrix} n & n-1 & n-2 & \dots & 3 & 2 & 1 \\ n-1 & n-1 & n-2 & \dots & 3 & 2 & 1 \\ 0 & n-2 & n-2 & \dots & 3 & 2 & 1 \\ 0 & 0 & n-3 & \dots & 3 & 2 & 1 \\ \vdots & \vdots & \vdots & \dots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 2 & 2 & 1 \\ 0 & 0 & 0 & \dots & 0 & 1 & 1 \end{pmatrix}$$

   de término general a_{ij}

$$a_{ij} \begin{cases} = & n-j+1, & \text{cuando} & j \geq i \\ = & n-j, & \text{cuando} & j = i-1 \\ = & 0, & \text{cuando} & j < i-1 \end{cases}$$

   a) Calcular $\det A_3$ mediante una recurrencia sobre n.

[15] $O(\epsilon^p)$ representa a una función de ϵ que cumple $\lim_{\epsilon \to 0} \frac{O(\epsilon^p)}{\epsilon^p} = 0$.

b) Sea

$$B_3 = \begin{pmatrix} 0 & 0 & \dots & 0 & 1 \\ 0 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \dots & \vdots & \vdots \\ 0 & 0 & \dots & 0 & 0 \end{pmatrix}$$

$b_{1n} = 1$ siendo todos los demás elementos nulos.
Calcular $\det A_3^\epsilon$.

c) A la vista de los resultados obtenidos, y suponiendo que n es suficientemente grande, ¿qué se podría comentar sobre los valores propios A_3^ϵ en comparación con los de A_3?

d) Estudiar el caso $n = 3$. Hallar los valores propios de A_3 en este caso. Escribir el polinomio característico de A_3^ϵ. Estudiar sus valores propios en función de ϵ. Analizar en particular el caso $\epsilon = 0.35$.

Solución:

1. a) Los polinomios característicos de A_1 y de su perturbada A_1^ϵ son respectivamente

$$\det(A_1 - \lambda I) = \det \begin{pmatrix} 1-\lambda & 2 \\ 0 & 2-\lambda \end{pmatrix} = (1-\lambda)(2-\lambda)$$

con lo que los valores propios de A_1 son $\lambda_1 = 1 \quad \text{y} \quad \lambda_2 = 2$, y

$$\det(A_1^\epsilon - \lambda I) = \det \begin{pmatrix} 1+\epsilon-\lambda & 2+\frac{\epsilon}{2} \\ \frac{\epsilon}{2} & 2+\epsilon-\lambda \end{pmatrix} =$$
$$= (1+\epsilon-\lambda)(2+\epsilon-\lambda) - \frac{\epsilon}{2}\left(2+\frac{\epsilon}{2}\right) = 0$$

y los valores propios de la matriz perturbada son raíces del trinomio

$$\lambda^2 - \lambda(3+2\epsilon) + 2 + 2\epsilon + \frac{3}{4}\epsilon^2 = 0$$

de discriminante

$$\Delta = (3+2\epsilon)^2 - 4\left(2 + 2\epsilon + \frac{3}{4}\epsilon^2\right) = 1 + 4\epsilon + \epsilon^2$$

Un desarrollo limitado de $\sqrt{\Delta}$ en el entorno de $\epsilon = 0$ (con lo que $4\epsilon + \epsilon^2 \to 0$) es [16]

$$\sqrt{\Delta} = \left(1 + 4\epsilon + \epsilon^2\right)^{1/2} =$$
$$= 1 + \frac{1}{2}\left(4\epsilon + \epsilon^2\right) + \binom{\frac{1}{2}}{2}\left(4\epsilon + \epsilon^2\right)^2 + \cdots =$$
$$= 1 + 2\epsilon + \frac{\epsilon^2}{2} - 2\epsilon^2 - \epsilon^3 - \frac{\epsilon^4}{8} + \cdots =$$
$$= 1 + 2\epsilon - \frac{3}{2}\epsilon^2 + \epsilon^2 p(\epsilon)$$

En el último término hemos sacado ϵ^2 factor común de un polinomio $p(\epsilon)$ cuyo término de menor grado [17] es $k\epsilon$ y por tanto $\lim_{\epsilon \to 0} p(\epsilon) = 0$.

---

[16]
$$(1+x)^{\frac{1}{2}} = 1 + \frac{1}{2}x + \binom{\frac{1}{2}}{2}x^2 + \cdots + \binom{\frac{1}{2}}{p}x^p + O(x^p)$$
en el entorno de $x = 0$.

[17] Si se escribe un desarrollo limitado de $\sqrt{\Delta}$ de orden 3, se comprueba que $k = 3$.

Entrando con ese desarrollo en la fórmula de las raíces de la ecuación de segundo grado obtenemos

$$\lambda_1^\epsilon = \frac{3+2\epsilon+\sqrt{\Delta}}{2} = \frac{3}{2}+\epsilon+\frac{1}{2}\left(1+2\epsilon-\frac{3}{2}\epsilon^2+\epsilon^2 p(\epsilon)\right) = 2+2\epsilon-\frac{3}{4}\epsilon^2+\frac{p(\epsilon)}{2}\epsilon^2$$
$$\lambda_2^\epsilon = \frac{3+2\epsilon-\sqrt{\Delta}}{2} = \frac{3}{2}+\epsilon-\frac{1}{2}\left(1+2\epsilon-\frac{3}{2}\epsilon^2+\epsilon^2 p(\epsilon)\right) = 1+\quad -\frac{3}{4}\epsilon^2-\frac{p(\epsilon)}{2}\epsilon^2$$

con lo que las constantes del enunciado son $C_1 = 2$ y $C_2 = 0$.

b) Para $\epsilon = 10^{-4}$

$$A_1^\epsilon = \begin{pmatrix} 1.00010 & 2.00005 \\ 0.00050 & 2.00010 \end{pmatrix}$$

y

$$\det\left(A_1^\epsilon - \lambda I\right) = (1.00010-\lambda)(2.00010-\lambda) - 0.00010 = \lambda^2 - 3.00020\lambda + 2.00020 = 0$$

cuyas raíces [18] son $\lambda_1^\epsilon \approx 2.00020 = 2+2\cdot 10^{-4}$ y $\lambda_2^\epsilon \approx 1$.
Las distancias pedidas son $|\lambda_1^\epsilon - \lambda_1| = |0.0002|$ y $|\lambda_2^\epsilon - \lambda_2| = 0$.

2. a)

$$\det\left(A_2^\epsilon - \lambda I\right) = \det\begin{pmatrix} 1+\epsilon-\lambda & 2+\dfrac{\epsilon}{2} \\ \dfrac{\epsilon}{2} & 1+\epsilon-\lambda \end{pmatrix} = (1+\epsilon-\lambda)^2 - \frac{\epsilon}{2}\left(2+\frac{\epsilon}{2}\right) = 0$$

Podríamos seguir el mismo proceso que antes, pero es mejor aquí despejar $1+\epsilon-\lambda$ que aparece al cuadrado en ese trinomio

$$1+\epsilon-\lambda = \pm\sqrt{\epsilon}\left(1+\frac{\epsilon}{4}\right)^{1/2}$$

y desarrollando como antes el segundo factor del segundo miembro

$$\left(1+\frac{\epsilon}{4}\right)^{1/2} = 1+\frac{1}{2}\frac{\epsilon}{4}-\frac{1}{8}\left(\frac{\epsilon}{4}\right)^2+\cdots = 1+\frac{\epsilon}{8}-\frac{\epsilon^2}{128}+\cdots$$

luego

$$\lambda^\epsilon = 1\pm\sqrt{\epsilon}+\epsilon+\frac{\epsilon^{3/2}}{8}-\frac{\epsilon^{5/2}}{128}+\cdots$$

con lo que las constantes del enunciado son ahora $C_1 = 1$ y $C_2 = -1$ y el resto del desarrollo es un polinomio cuya indeterminada es $\sqrt{\epsilon}$ y cuyo término de menor grado es $\epsilon = (\sqrt{\epsilon})^2$.

b) Para $\epsilon = 10^{-4}$

$$A_2^\epsilon = \begin{pmatrix} 1.00010 & 2.00005 \\ 0.00050 & 1.00010 \end{pmatrix}$$

y

$$\det\left(A_1^\epsilon - \lambda I\right) = (1.00010-\lambda)^2 - 0.00010 \quad \Rightarrow \quad 1.00010-\lambda = \pm 0.00010$$

cuyas raíces calculadas con cinco dígitos son $\lambda^\epsilon \approx 1.0001 \pm 0.01 = 1 \pm 0.01 + 0.0001$ como queríamos comprobar.

[18] El resto de los términos de los respectivos desarrollos se perdió para la precisión con la que estamos trabajando, en el producto $2.00005 \cdot 0.00005 \approx 0$.

3. a) Siguiendo la mecánica del método de recurrencia estudiamos valores particulares.

• Para $n = 2$

$$\Delta_2 = \det \begin{pmatrix} 2 & 1 \\ 2 & 1 \end{pmatrix} = 1$$

• Para $n = 3$

$$\Delta_3 = \begin{pmatrix} 3 & 2 & 1 \\ 2 & 2 & 1 \\ 0 & 1 & 1 \end{pmatrix} = 3 - 2 = 1$$

Una vez intuida una posible ley, suponemos que $\Delta_{n-1} = \ldots \Delta_3 = 1$ y probamos que $\Delta_n = 1$. Desarrollando Δ_n por los elementos de su primera columna se obtiene

$$n\Delta_{n-1} - (n-1)\Delta_{n-1} = \Delta_{n-1}$$

de donde el resultado.

b)

$$A_3^\epsilon = \begin{pmatrix} n & n-1 & n-2 & \ldots & 3 & 2 & 1+\epsilon \\ n-1 & n-1 & n-2 & \ldots & 3 & 2 & 1 \\ 0 & n-2 & n-2 & \ldots & 3 & 2 & 1 \\ 0 & 0 & n-3 & \ldots & 3 & 2 & 1 \\ \vdots & \vdots & \vdots & \ldots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & 2 & 2 & 1 \\ 0 & 0 & 0 & \ldots & 0 & 1 & 1 \end{pmatrix}$$

Desarrollando de nuevo por los elementos de la primera columna

$$\det(A_3^\epsilon) = n\Delta_{n-1} - (n-1)\det \begin{pmatrix} n-1 & n-2 & \ldots & 3 & 2 & 1+\epsilon \\ n-2 & n-2 & \ldots & 3 & 2 & 1 \\ 0 & n-3 & \ldots & 3 & 2 & 1 \\ \vdots & \vdots & \ldots & \vdots & \vdots & \vdots \\ 0 & 0 & \ldots & 2 & 2 & 1 \\ 0 & 0 & \ldots & 0 & 1 & 1 \end{pmatrix}$$

utilizando la propiedad de linealidad de la forma multilineal alternada det respecto del último vector columna

$$\det(A_3^\epsilon) = n\Delta_{n-1} - (n-1)\left[\Delta_n + (-1)^{n-1}\epsilon \det \begin{pmatrix} n-2 & n-2 & \ldots & 3 & 2 \\ 0 & n-3 & \ldots & 3 & 2 \\ \vdots & \vdots & \ldots & \vdots & \vdots \\ 0 & 0 & \ldots & 2 & 2 \\ 0 & 0 & \ldots & 0 & 1 \end{pmatrix}\right]$$

es decir,

$$\det A_3^\epsilon = n\Delta_{n-1} - (n-1)\left[\Delta_n + (-1)^{n-1}\epsilon(n-2)!\right]$$

y sustituyendo los datos conocidos,

$$\det(A_3^\epsilon) = n + (-1)^{n-1}(n-1)!\epsilon$$

c) Para n grande y ϵ fijo, $\det(A_3^\epsilon)$ que es el producto de los valores propios de A_3^ϵ, es muy grande en relación a los de A_3.

d) Si, por ejemplo, $n = 3$, el polinomio característico de A_3 es $f(\lambda) = \lambda^3 - 6\lambda^2 + 6\lambda - 1 = 0$ cuyas raíces son 0.2085, 1, 4.7915. En su matriz perturbada A_3^ϵ el único coeficiente del polinomio característico que cambia en relación con el de A_3 es el término constante, ya que $\det A_3^\epsilon = 1 + 2\epsilon$, luego $\lambda^3 - 6\lambda^2 + 6\lambda - (1 + 2\epsilon) = 0$.

Del estudio de las variaciones de los miembros de la familia de cúbicas $g_\epsilon(\lambda) = \lambda^3 - 6\lambda^2 + 6\lambda - (1 + 2\epsilon) = f(\lambda) - 2\epsilon$ se deduce que los valores de tangente horizontal en los que cambia el sentido de crecimiento o decrecimiento de la función son $2 \pm \sqrt{2}$, los mismos independientemente de ϵ, luego el grafo de $g_\epsilon(\lambda)$ se obtiene mediante una traslación de 2ϵ en el sentido negativo del eje vertical. Ya que $f(0.586) \approx 0.6$, los miembros de la familia relativas a valores $\epsilon > 0.35$ sólo cortan una vez el eje λ, luego las correspondientes matrices perturbadas poseen un valor propio real y dos valores propios complejos conjugados.

PROBLEMA 2.9 *Estimación del número de condición de una matriz. Sistema mal condicionado. Influencia de los errores de redondeo en la solución calculada numéricamente.*

Sean una matriz cuadrada $A \in \mathcal{M}_n(\mathbb{R})$, $\|,\|$ una norma vectorial en $\mathbb{R}^n$ y $\|,\|$ su norma inducida en $\mathcal{M}_n(\mathbb{R})$. Se desea resolver el sistema de ecuaciones lineales

$$(S) \qquad A\,\mathbf{x} = \mathbf{b}$$

usando eliminación gaussiana y aritmética de d dígitos en todas las operaciones salvo en las operaciones de cálculo del residuo que se realizarán con doble precisión, es decir, con aritmética de $2d$ dígitos.
Se puede probar que en este caso [19], el vector residuo $\mathbf{r}$ relativo al valor aproximado $\overline{\mathbf{x}}$ de la solución de (S) calculada por cualquier método, verifica la ecuación aproximada

$$\|\mathbf{r}\| = 10^{-d}\|A\|\|\overline{\mathbf{x}}\| \qquad (*)$$

1. Sea $\overline{\mathbf{y}}$ la solución del sistema

   $$(S') \qquad A\,\mathbf{y} = \mathbf{r}$$

   Utilizar $(*)$ para demostrar que se obtiene en este caso la siguiente estimación del número de condición de la matriz A

   $$\mathrm{cond}(A) \approx \frac{\|\overline{\mathbf{y}}\|}{\|\overline{\mathbf{x}}\|} 10^d \qquad (**)$$

2. Supongamos $n = 3$ y A y $\mathbf{b}$ definidos por

   $$A = \begin{pmatrix} 3.02 & -1.05 & 2.53 \\ 4.33 & 0.56 & -1.78 \\ -0.83 & -0.54 & 1.47 \end{pmatrix} \quad \text{y} \quad \mathbf{b} = \begin{pmatrix} -1.61 \\ 7.23 \\ -3.38 \end{pmatrix}$$

   El sistema (S) asociado tiene la solución exacta

   $$\mathbf{x} = (1, 2, -1)^T$$

   a) Resolver (S) usando eliminación gaussiana con estrategia de pivote y aritmética de redondeo de $d = 6$ dígitos [20]. Sea $\overline{\mathbf{x}}$ la solución aproximada obtenida.

   b) Determinar la estimación del número de condición de A objeto del apartado 1, considerando en $\mathbb{R}^3$ la norma $\|\ \|_\infty$ del máximo.

   c) Determinar $\mathrm{cond}_\infty(A)$, número de condición exacto de la matriz A para la norma matricial inducida por la norma $\|\ \|_\infty$ de $\mathbb{R}^3$ comparando su valor con el antes obtenido.

   d) Determinar un intervalo que enmarque $\mathrm{cond}_2(A)$ número de condición de la matriz A para la norma matricial inducida por la norma euclídea $\|\ \|_2$ de $\mathbb{R}^3$.

[19] La importancia del número de condición de A para mayorar el error de una solución aproximada $\overline{\mathbf{x}}$ o para decidir sobre la necesidad de un refinamiento iterativo de dicha solución es enorme (ver la sección 2.2 del resumen teórico).

El conocimiento de $\mathrm{cond}(A)$ exige conocer la inversa A^{-1} de A, o bien el mayor y el menor de los valores singulares de $A^T A$. Estos cálculos pueden ser más costosos que la propia resolución del sistema.

Un método para estimar con un coste numérico razonable $\mathrm{cond}(A)$ que no requiere el cálculo de la inversa de A, se incluye en el libro de Forsythe y Moler (1967) pp. 49-51. En ese método se basan la aproximación $(*)$ y su consecuencia $(**)$.

[20] Ver Apéndice B.

e) Efectuar estimaciones del error absoluto y relativo cometido al tomar $\overline{\mathbf{x}}$ solución de (S) y compáralo con el error real.

f) Efectuar un refinamiento iterativo de la solución $\overline{\mathbf{x}}$ usando doble precisión, poniendo como test de parada que la norma $\|\mathbf{y}_{(k)}\|_\infty$ del vector corrector en la iteración k-ésima, sea $\leq 10^{-4}$.

Comentarios y cuestiones suplementarias

1. La precisión aritmética variable pondrá al descubierto el mal condicionamiento del problema en estudio.

   Si se resuelve (S) usando eliminación gaussiana, estrategia de pivote y tres dígitos significativos, se concluye con un resultado que en nada se parece a la solución del sistema.

2. Perturbando ligeramente la matriz A sustituyendo el elemento $a_{11} = 3.02$ por 3.00 y resolviendo el sistema perturbado con 6 cifras significativas se observa que las soluciones enormemente inestables varían mucho con las más pequeñas variaciones de los coeficientes.

3. Tener una idea ajustada de las características técnicas de la máquina de cálculo que se utilice es importante. No siempre se tiene acceso a ordenadores de gran potencia y conocer el tipo de aritmética que se puede usar es fundamental como prueba este ejercicio.

Solución:

1. Del sistema lineal (S')

$$\overline{\mathbf{y}} \sim A^{-1}\mathbf{r} = A^{-1}\left(\mathbf{b} - A\overline{\mathbf{x}}\right) = A^{-1}\mathbf{b} - A^{-1}A\overline{\mathbf{x}} = \mathbf{x} - \overline{\mathbf{x}}$$

de modo que $\overline{\mathbf{y}}$ es una estimación del error cometido al aproximar la solución del sistema original.
Tomando normas en las igualdades anteriores y usando la relación $(*)$ tendremos

$$\|\overline{\mathbf{y}}\| \sim \|\mathbf{x} - \overline{\mathbf{x}}\| = \|A^{-1}\mathbf{r}\| = \|A^{-1}\|\|\mathbf{r}\| \sim \|A^{-1}\|\|A\|10^{-d}\|\overline{\mathbf{x}}\| = \mathrm{cond}(A)10^{-d}\|\overline{\mathbf{x}}\|$$

de donde

$$\mathrm{cond}(A) \approx \frac{\|\overline{\mathbf{y}}\|}{\|\overline{\mathbf{x}}\|}10^{d}$$

2.

a) Partiendo de la matriz aumentada

$$\left(\begin{array}{c|c} A & \mathbf{b} \end{array}\right) = A^{(1)} = \left(\begin{array}{ccc|c} 3.02 & -1.05 & 2.53 & -1.61 \\ 4.33 & 0.56 & -1.78 & 7.23 \\ -0.83 & -0.54 & 1.47 & -3.38 \end{array}\right)$$

y utilizando eliminación gaussiana con estrategia de pivote parcial y aritmética de redondeado a 6 dígitos comenzamos intercambiando la primera y la segunda fila $F_{12}A^{(1)}$ matriz que seguimos denotando $A^{(1)}$

$$A^{(1)} = \left(\begin{array}{ccc|c} 4.33 & 0.56 & -1.78 & 7.23 \\ 3.02 & -1.05 & 2.53 & -1.61 \\ -0.83 & -0.54 & 1.47 & -3.38 \end{array}\right)$$

El pivote es $a_{11}^{(1)} = 4.33$ y los multiplicadores $m_{21}^{(1)} = \frac{a_{21}^{(1)}}{a_{11}^{(1)}} = \frac{3.02}{4.33} \sim 0.697459$ y $m_{31}^{(1)} = \frac{a_{31}^{(1)}}{a_{11}^{(1)}} = -\frac{0.83}{4.33} \sim -0.191686$ con ello se obtiene

$$A^{(2)} = \left(\begin{array}{ccc|c} 4.33 & 0.56 & -1.78 & 7.23 \\ 0 & -1.440578 & 3.77148 & -6.65263 \\ -0 & -0.432656 & 1.12880 & -1.99411 \end{array}\right)$$

El pivote es ahora $a_{22}^{(2)} = -1.44058$ y los multiplicadores $m_{32}^{(2)} = \frac{a_{32}^{(2)}}{a_{22}^{(2)}} = -\frac{-0.432656}{-1.44058} \sim 0.300334$ luego

$$A^{(3)} = \left(\begin{array}{ccc|c} 4.33 & 0.56 & -1.78 & 7.23 \\ 0 & -1.440578 & 3.77148 & -6.65263 \\ -0 & 0 & -0.0039036 & 0.0039009 \end{array}\right)$$

La solución obtenida [21] es $\overline{\mathbf{x}} = (1.00005, 2.00180, -0.999308)$.

b) El vector residuo correspondiente a $\overline{\mathbf{x}}$ es

$$\mathbf{r} = \begin{pmatrix} 7.23 \\ -1.61 \\ -3.38 \end{pmatrix} - \begin{pmatrix} 4.33 & 0.56 & -1.78 \\ 3.02 & -1.05 & 2.53 \\ -0.83 & -0.54 & 1.47 \end{pmatrix} \begin{pmatrix} 1.00005 \\ 2.00180 \\ -0.999308 \end{pmatrix} = \begin{pmatrix} 0.0000073 \\ -0.0000118 \\ -0.0000038 \end{pmatrix}$$

de modo que $\|\mathbf{r}\|_\infty = 0.0000118$.

Utilizando $(*)$, podemos estimar $\|\mathbf{r}\|_\infty$. Como $\|A\|_\infty = \max_i \sum_j^n |a_{ij}| = 6.67$, $\|\mathbf{r}\|_\infty \sim 10^{-6}(6.67)2.00180 = 0.0000133$ bastante ajustado.

Para aplicar la estimación $(**)$ debemos resolver el sistema (S'); para lo que utilizamos la matriz que hemos almacenado en el ordenador con los multiplicadores por debajo de la diagonal principal, sustituyendo el segundo miembro por los elementos correspondientes al residuo. Se tiene

$$\left(\begin{array}{ccc|c} 4.33 & 0.56 & -1.78 & 0.0000073 \\ 0.697459 & -1.440578 & 3.77148 & -0.0000168 \\ -0.191686 & 0.300334 & -0.0039036 & 0.0000026 \end{array}\right)$$

de donde [22] $\overline{\mathbf{y}} = (-0.0000481, -0.0017319, -0.000666)$.

Con todo ello,

$$\mathrm{cond}_\infty(A) \approx 10^6 \frac{\|\overline{\mathbf{y}}\|_\infty}{\|\overline{\mathbf{x}}\|_\infty} == 10^6 \frac{0.0017319}{2.00180} = 865.1$$

que como veremos en el apartado siguiente es una pobre estimación aunque revela claramente un sistema mal condicionado.

c) Calculamos A^{-1} por Gauss-Jordan. Se obtiene sucesivamente

$$\left(\begin{array}{c|c} A & I \end{array}\right) = \left(\begin{array}{ccc|ccc} 3.02 & -1.05 & 2.53 & 1 & 0 & 0 \\ 4.33 & 0.56 & -1.78 & 0 & 1 & 0 \\ -0.83 & -0.54 & 1.47 & 0 & 0 & 1 \end{array}\right) \Rightarrow$$

$$\Rightarrow \left(\begin{array}{ccc|ccc} 1 & 0.12933 & -0.411085 & 0.230947 & 0 & 1.66975 \\ 0 & -1.44058 & 3.77148 & 1 & -0.69746 & 0 \\ 0 & -0.432656 & 1.12880 & 0 & 0.191686 & 1 \end{array}\right) \Rightarrow$$

$$\Rightarrow \left(\begin{array}{ccc|ccc} 1 & 0 & -0.072495 & 0.0897764 & 0.168332 & 0 \\ 0 & 1 & -2.61803 & -0.694165 & 0.484152 & 0 \\ 0 & 0 & -0.00391 & -0.300335 & 0.401157 & 1 \end{array}\right) \Rightarrow$$

$$\Rightarrow \left(\begin{array}{ccc|ccc} 1 & 0 & 0 & 5.65827 & -7.26951 & -18.5409 \\ 0 & 1 & 0 & 200.402 & -268.321 & -669.572 \\ 0 & 0 & 1 & 76.812 & -102.598 & -255.754 \end{array}\right) = \left(\begin{array}{c|c} I & A^{-1} \end{array}\right)$$

[21] Por cuestiones de economía de almacenamiento y por su posible uso posterior, se guardan los valores de los cocientes m_{ij} por debajo de la diagonal principal de la última matriz del algoritmo $A^{(3)}$ en vez de los inútiles ceros. El resultado es la matriz

$$\left(\begin{array}{ccc|c} 4.33 & 0.56 & -1.78 & 7.23 \\ 0.697459 & -1.440578 & 3.77148 & -6.65263 \\ -0.191686 & 0.300334 & -0.0039036 & 0.0039009 \end{array}\right)$$

[22] El vector $\overline{\mathbf{y}}$ es una aproximación del vector error absoluto $\mathbf{e} = \mathbf{x} - \overline{\mathbf{x}}$. En efecto, $A\mathbf{e} = A\mathbf{x} - A\overline{\mathbf{x}} = \mathbf{b} - A\overline{\mathbf{x}} = \mathbf{r}$.

y la matriz bloque a la derecha de la línea vertical de puntos es A^{-1} obtenida con aritmética de 6 dígitos.

Tenemos $\|A\|_\infty = 6.67$ y $\|A^{-1}\|_\infty = 1138.09$ de donde $\text{cond}_\infty(A) = \|A\|_\infty \|A^{-1}\|_\infty = (6.67)(1138.09) = 7591.06 \gg 1$.

d) Utilizaremos la norma de Schur para enmarcar $\|A\|_2$.

$$\|A\|_S = \left[\sum_{i,j}^{n} |a_{ij}|^2\right]^{1/2} = \sqrt{41.9961} \approx 6.48$$

de donde

$$\frac{1}{\sqrt{3}}\|A\|_S \leq \|A\|_2 \leq \|A\|_S \quad \Rightarrow \quad 3.74134 \leq \|A\|_2 \leq 6.48$$

e) Ya que conocemos la solución exacta $x(1, 2, -1)$, tendremos

$$\|\mathbf{x} - \overline{\mathbf{x}}\|_\infty = 0.0018 \qquad \text{y} \qquad \frac{\|\mathbf{x} - \overline{\mathbf{x}}\|_\infty}{\|\mathbf{x}\|_\infty} = \frac{0.0018}{2} = 0.0009$$

f) Partimos de $\overline{\mathbf{x}} = \mathbf{x}^{(1)}$ que corregimos con $\overline{\mathbf{y}} = \mathbf{y}^{(1)}$, de modo que

$$\mathbf{x}^{(2)} = \mathbf{x}^{(1)} + \mathbf{y}^{(1)} = (1.0000019, 2.0000681, -0.999974)$$

un resultado mejor que el obtenido en 2.a) para ello el cálculo del residuo se ha hecho con doble precisión.

Es fácil comprobar que si se utiliza la aritmética de seis dígitos en las operaciones del refinamiento iterativo no se consigue mejora en la solución.

En el paso siguiente del proceso iterativo,

$$\mathbf{r}_{(2)} = (0, 0, (-1) \cdot 10^{-7})$$

Resolviendo el sistema $A\mathbf{y} = \mathbf{r}_{(2)}$

$$\left(\begin{array}{ccc|c} 4.33 & 0.56 & -1.78 & 0 \\ 0.697459 & -1.440578 & 3.77148 & 0 \\ -0.191686 & 0.300334 & -0.0039036 & -0.0000001 \end{array}\right)$$

de modo que

$$\mathbf{y}^{(2)} = (0.0000964, 0.0000669, 0.0000256)$$

y $\|\mathbf{y}^{(2)}\|_\infty = 0.0000964 \leq 10^{-4}$.

El vector $\mathbf{x}^{(3)}$ obtenido en este paso con la precisión que permite una calculadora de 8 dígitos, no mejora la solución $\mathbf{x}^{(2)}$. Sería necesario usar Matlab en este punto del proceso para rehacer los cálculos de $\mathbf{r}_{(2)}$, $\mathbf{y}^{(2)}$ y $\mathbf{x}^{(3)}$.

Comentarios y cuestiones suplementarias

1. Partiendo de la matriz aumentada

$$\left(\begin{array}{c|c} A & \mathbf{b} \end{array}\right) = A^{(1)} = \left(\begin{array}{ccc|c} 3.02 & -1.05 & 2.53 & -1.61 \\ 4.33 & 0.56 & -1.78 & 7.23 \\ -0.83 & -0.54 & 1.47 & -3.38 \end{array}\right)$$

y utilizando eliminación gaussiana y estrategia de pivote parcial, comenzamos como antes, intercambiando la primera y la segunda fila de $A^{(1)}$

$$A^{(1)} = \left(\begin{array}{ccc|c} 4.33 & 0.56 & -1.78 & 7.23 \\ 3.02 & -1.05 & 2.53 & -1.61 \\ -0.83 & -0.54 & 1.47 & -3.38 \end{array}\right)$$

El pivote es $a_{11}^{(1)} = 4.33$ y los multiplicadores $m_{21}^{(1)} = \frac{a_{21}^{(1)}}{a_{11}^{(1)}} = \frac{3.02}{4.33} \sim 0.697$ y $m_{31}^{(1)} = \frac{a_{31}^{(1)}}{a_{11}^{(1)}} = \frac{-0.83}{4.33} \sim -0.192$ con ello se obtiene

$$A^{(2)} = \left(\begin{array}{ccc|c} 4.33 & 0.56 & -1.78 & 7.23 \\ 0 & -1.44 & 3.77 & -6.65 \\ -\ 0 & -0.432 & 1.13 & -1.99 \end{array}\right)$$

El pivote es ahora $a_{22}^{(2)} = -1.44$ y los multiplicadores $m_{32}^{(2)} = \frac{a_{32}^{(2)}}{a_{22}^{(2)}} = \frac{-0.432}{-1.44} \sim 0.3$ y con detalle los elementos de la última matriz $A^{(3)}$ son $a_{33}^{(3)} = 1.13 - (0.3)(3.77) = 1.13 - 1.131 = -0.001$ y $a_{34}^{(3)} = 1.99 - (0.3)(-6.65) = -1.99 + 1.995 \approx 0.005$

$$A^{(3)} = \left(\begin{array}{ccc|c} 4.33 & 0.56 & -1.78 & 7.23 \\ 0 & -1.44 & 3.77 & -6.66 \\ -\ 0 & 0 & -0.001 & 0.005 \end{array}\right)$$

cuya solución es $(0.709, -8.46, -5)$ claramente errónea. Es fácil comprobar que

$$\begin{pmatrix} 4.33 & 0.56 & -1.78 \\ 3.02 & -1.05 & 2.53 \\ -0.83 & -0.54 & 1.47 \end{pmatrix} \cdot \begin{pmatrix} 0.709 \\ -8.46 \\ -5 \end{pmatrix} = \begin{pmatrix} 7.23 \\ -1.58 \\ -3.37 \end{pmatrix}$$

extraordinariamente próxima al segundo miembro $\mathbf{b}$ de (S).

La norma del vector residuo es $\|\mathbf{r}\|_\infty = 0.03$ y la del vector error absoluto $\|\mathbf{e}\|_\infty = \|\mathbf{x} - \overline{\mathbf{x}}\|_\infty = 10.5$.

La norma del vector residuo no es una buena medida de la norma del vector error absoluto en un sistema mal condicionado.

La aritmética variable pone al descubierto la mala condición del sistema. Un ingeniero que aborda un problema concreto tantea muchas veces una solución zafia operando con baja precisión. Aquí la información que obtendría sería fatal.

2. Partiendo de la matriz aumentada

$$\left(\begin{array}{ccc|c} 3 & -1.05 & 2.53 & -1.61 \\ 4.33 & 0.56 & -1.78 & 7.23 \\ -0.83 & -0.54 & 1.47 & -3.38 \end{array}\right)$$

Intercambiamos la primera y la segunda fila

$$A^{(1)} = \left(\begin{array}{ccc|c} 4.33 & 0.56 & -1.78 & 7.23 \\ 3 & -1.05 & 2.53 & -1.61 \\ -0.83 & -0.54 & 1.47 & -3.38 \end{array}\right)$$

El pivote es $a_{11}^{(1)} = 4.33$ y los multiplicadores $m_{21}^{(1)} = \frac{a_{21}^{(1)}}{a_{11}^{(1)}} = \frac{3.00}{4.33} \sim 0.692841$ y $m_{31}^{(1)} = \frac{a_{31}^{(1)}}{a_{11}^{(1)}} = \frac{-0.83}{4.33} \sim -0.191686$ con ello se obtiene

$$A^{(2)} = \left(\begin{array}{ccc|c} 4.33 & 0.56 & -1.78 & 7.23 \\ 0 & -1.43799 & 3.76325 & -6.61924 \\ -\ 0 & -0.432656 & 1.1288 & -1.99411 \end{array}\right)$$

El pivote es ahora $a_{22}^{(2)} = -1.43799$ y el multiplicador $m_{32}^{(2)} = \frac{a_{32}^{(2)}}{a_{22}^{(2)}} = \frac{-0.432656}{-1.43799} \sim 0.300875$, luego

$$A^{(3)} = \left(\begin{array}{ccc|c} 4.33 & 0.56 & -1.78 & 7.23 \\ 0 & -1.43799 & 3.76325 & -6.61924 \\ -0 & 0 & -0.0034678 & -0.0025462 \end{array}\right)$$

cuya solución es $(1.12775, 6.52464, 0.734241)$. Si comparamos este resultado con el obtenido con la misma aritmética de 6 dígitos en el apartado 2.a) vemos el enorme cambio que se ha producido en la solución como consecuencia de una perturbación muy pequeña $\epsilon = -0.02$ de un elemento tan sólo de la matriz del sistema.

Para un ingeniero esa pequeña perturbación puede ser consecuencia de algún error de medida en los datos del problema que esté estudiando y le conducirá a graves errores si no sabe "a priori" que el sistema está mal condicionado, lo que le exigirá extremar los controles.

3. Es interesante observar el tratamiento que hace Matlab de este problema.

   Si se utiliza la resolución del sistema (S) mediante eliminación gaussiana [23] con el formato normal de 4 decimales, se obtiene el resultado exacto $\mathbf{x} = (1, 2, -1)$ y no detecta el mal condicionamiento del problema lógicamente porque sus rutinas internas trabajan en doble precisión y al final dan el resultado redondeando a 4 decimales. Esto se hace evidente si se resuelve (S) en *>>format long* en cuyo caso la solución es $\mathbf{x} = (1, 1.99999999999985, -1.00000000000006)$.

   En el caso de la matriz perturbada y también en *>>format long* se obtiene $\mathbf{x} = (1.12767729102095, 6.52208909654403, 0.73326548549736)$ y redondeado a 6 dígitos $\mathbf{x} = (1.12768, 6.52209, 0.733265)$ que al compararla con la obtenida antes revela la diferencia que hay entre operar con máxima precisión y redondear al final y operar de modo que todas las operaciones se vayan realizando con un cierto tipo de aritmética.

PROBLEMA 2.10 *Resolución de un sistema de ecuaciones lineales de matriz tridiagonal simétrica.*

Resolver el sistema tridiagonal siguiente

$$A\mathbf{x} = \mathbf{b}$$

$$(S) \quad \begin{pmatrix} 4 & -1 & 0 & 0 & 0 \\ -1 & 4 & -1 & 0 & 0 \\ 0 & -1 & 4 & -1 & 0 \\ 0 & 0 & -1 & 4 & -1 \\ 0 & 0 & 0 & -1 & 4 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{pmatrix} = \begin{pmatrix} 100 \\ 200 \\ 200 \\ 200 \\ 100 \end{pmatrix}$$

1. Mediante una descomposición LU.

2. Utilizando el *método de Gauss-Seidel* con estimador inicial

$$\mathbf{x}^{(0)} = (25, 50, 50, 50, 25)^T$$

3. Este sistema está especialmente adaptado para utilizar el *método de relajación*.

   Resolver el sistema mediante superrelajación de Gauss-Seidel, variando el factor ω de superrelajación hasta su valor óptimo.

Comentarios

- Obsérvese lo rápido que se opera en el apartado 1 en este caso particular.
- Utilizar en el último apartado el mismo estimador inicial $\mathbf{x}^{(0)}$ que en 2. Se comenzará la búsqueda del factor de superrelajación óptimo con $\omega = 1.00$ incrementando su valor en dos centésimas hasta $\omega = 1.14$. Se

[23] Ver A.5 del Tutorial de Matlab.

comprobará que para $\omega = 1.06$ se alcanza el resultado con un test de parada $\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\| \leq \epsilon \|\mathbf{x}^{(k)}\|$ con $\epsilon = 10^{-7}$, en siete iteraciones [24].

- El sistema propuesto es un clásico en la resolución de ecuaciones diferenciales en derivadas parciales.

Solución:

1. La descomposición de $A = LU$ con L triangular inferior bidiagonal con unos en la diagonal principal y U triangular superior también bidiagonal es particularmente simple en este caso. Se tiene

$$L = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ -\frac{1}{4} & 1 & 0 & 0 & 0 \\ 0 & -\frac{4}{15} & 1 & 0 & 0 \\ 0 & 0 & -\frac{15}{56} & 1 & 0 \\ 0 & 0 & 0 & -\frac{56}{209} & 1 \end{pmatrix} \quad U = \begin{pmatrix} 4 & -1 & 0 & 0 & 0 \\ 0 & \frac{15}{4} & -1 & 0 & 0 \\ 0 & 0 & \frac{56}{15} & -1 & 0 \\ 0 & 0 & 0 & \frac{209}{56} & -1 \\ 0 & 0 & 0 & 0 & \frac{780}{209} \end{pmatrix}$$

La resolución del sistema (S) asociada a esa descomposición se hace resolviendo por sustitución adelante el sistema $L\mathbf{y} = \mathbf{b}$ y una vez obtenida su solución se resuelve por retrosustitución el sistema $U\mathbf{x} = \mathbf{y}$.

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ -\frac{1}{4} & 1 & 0 & 0 & 0 \\ 0 & -\frac{4}{15} & 1 & 0 & 0 \\ 0 & 0 & -\frac{15}{56} & 1 & 0 \\ 0 & 0 & 0 & -\frac{56}{209} & 1 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \end{pmatrix} = \begin{pmatrix} 100 \\ 200 \\ 200 \\ 200 \\ 100 \end{pmatrix} \Rightarrow \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \end{pmatrix} = \begin{pmatrix} 100 \\ 225 \\ 260 \\ 269.643 \\ 172.249 \end{pmatrix}$$

$$\begin{pmatrix} 4 & -1 & 0 & 0 & 0 \\ 0 & \frac{15}{4} & -1 & 0 & 0 \\ 0 & 0 & \frac{56}{15} & -1 & 0 \\ 0 & 0 & 0 & \frac{209}{56} & -1 \\ 0 & 0 & 0 & 0 & \frac{780}{209} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{pmatrix} = \begin{pmatrix} 100 \\ 225 \\ 260 \\ 269.643 \\ 172.249 \end{pmatrix} \Rightarrow \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{pmatrix} = \begin{pmatrix} 46.15384708887603 \\ 84.61538835550414 \\ 92.30770633314054 \\ 84.61543697705802 \\ 46.15389871794871 \end{pmatrix}$$

Se observa que $x_1 = x_5$ y $x_2 = x_4$ propiedad que posee el término independiente $\mathbf{b}$ y que respeta la matriz simétrica y tridiagonal A. Ello sugiere tomar el estimador inicial con la misma propiedad en los métodos iterativos de los apartados siguientes.

2. Resolveremos ahora (S) por el método de Gauss-Seidel pero haciendo una partición Π de la matriz A por bloques que sea admisible (las matrices diagonales son invertibles) y que respete la estructura tridiagonal.

$$A = \begin{pmatrix} A_1 & B_1 & O \\ C_2 & A_2 & B_2 \\ O & C_3 & A_3 \end{pmatrix} = \begin{pmatrix} \begin{pmatrix} 4 & -1 \\ -1 & 4 \end{pmatrix} & \begin{pmatrix} 0 & 0 \\ -1 & 0 \end{pmatrix} & \begin{pmatrix} 0 \\ 0 \end{pmatrix} \\ \begin{pmatrix} 0 & -1 \\ 0 & 0 \end{pmatrix} & \begin{pmatrix} 4 & -1 \\ -1 & 4 \end{pmatrix} & \begin{pmatrix} 0 \\ -1 \end{pmatrix} \\ 0 & \begin{pmatrix} 0 & -1 \end{pmatrix} & 4 \end{pmatrix}$$

El vector solución $\mathbf{x}$, el estimador inicial $\mathbf{x}^{(0)}$ y el término independiente se escriben también por

[24] Este test de parada, cómodo desde el punto de vista de cálculo, tiene como inconveniente que algunas veces se verifica sin que $\mathbf{x}^{(k)}$ esté cerca de la solución.

bloques de modo coherente con la partición Π

$$\mathbf{x} = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} = \begin{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \\ \begin{pmatrix} x_3 \\ x_4 \end{pmatrix} \\ x_5 \end{pmatrix} \quad \mathbf{x}^{(0)} = \begin{pmatrix} \mathbf{X}_1^{(0)} \\ \mathbf{X}_2^{(0)} \\ \mathbf{X}_3^{(0)} \end{pmatrix} = \begin{pmatrix} \begin{pmatrix} 25 \\ 50 \end{pmatrix} \\ \begin{pmatrix} 50 \\ 50 \end{pmatrix} \\ 25 \end{pmatrix} \quad \text{y}$$

$$\mathbf{b} = \begin{pmatrix} \mathbb{b}_1 \\ \mathbb{b}_2 \\ \mathbb{b}_3 \end{pmatrix} = \begin{pmatrix} \begin{pmatrix} 100 \\ 200 \end{pmatrix} \\ \begin{pmatrix} 200 \\ 200 \end{pmatrix} \\ 100 \end{pmatrix}$$

El método iterativo de Gauss-Seidel asociado a Π es

$$(D - L)\mathbf{x}^{(k+1)} = U\mathbf{x}^{(k)} + \mathbf{b}$$

donde

$$D = \begin{pmatrix} A_1 & O & O \\ O & A_2 & O \\ O & O & A_3 \end{pmatrix};\ L = \begin{pmatrix} O & O & O \\ C_2 & O & O \\ O & C_3 & O \end{pmatrix};\ U = \begin{pmatrix} O & B_1 & O \\ O & O & B_2 \\ O & O & O \end{pmatrix}$$

luego

$$\begin{cases} A_1\mathbf{X}_1^{(k+1)} = -B_1\mathbf{X}_2^{(k)} + \mathbb{b}_1 \\ A_2\mathbf{X}_2^{(k+1)} = -C_2\mathbf{X}_1^{(k+1)} - B_2\mathbf{X}_3^{(k)} + \mathbb{b}_2 \\ A_3\mathbf{X}_3^{(k+1)} = -C_3\mathbf{X}_2^{(k+1)} + \mathbb{b}_3 \end{cases}$$

Cada paso de esta iteración nos lleva a resolver por un método directo sistemas de la forma $A_i\mathbf{X}_i = \mathbf{y}_i$. En este caso,

$$A_1^{-1} = A_2^{-1} = \frac{1}{15}\begin{pmatrix} 4 & 1 \\ 1 & 4 \end{pmatrix} \quad \text{y} \quad A_3^{-1} = \frac{1}{4}$$

de donde

$$\begin{cases} \mathbf{X}_1^{(k+1)} = -A_1^{-1}B_1\mathbf{X}_2^{(k)} + A_1^{-1}\mathbb{b}_1 \\ \mathbf{X}_2^{(k+1)} = -A_2^{-1}C_2\mathbf{X}_1^{(k+1)} - A_2^{-1}B_2\mathbf{X}_3^{(k)}A_2^{-1} + \mathbb{b}_2 \\ \mathbf{X}_3^{(k+1)} = -A_3^{-1}C_3\mathbf{X}_2^{(k+1)} + A_3^{-1}\mathbb{b}_3 \end{cases}$$

Hemos realizado en Matlab de un modo poco sofisticado los cálculos anteriores y se incluye aquí el conjunto de sentencias usadas para comprobación (*tridiagonal1.m*)

```
%Definimos los datos del problema
A1=[1/15 0
    4/15 0];
A2=[0 4/15
    0 1/15];
A3=[1/15 4/15]';
A4=[0 1/4];
x10=[25 50];
%Este elemento es innecesario en este algoritmo pero es necesario en el
%algoritmo del metodo de relajacion
x20=[50 50];
x30=25;
b1=[40 60];
b2=[200/3 200/3];
b3=100;
```

```
%Una vez definida la estructura del sistema y las condiciones iniciales
%comienza el proceso iterativo.
%Partiendo de los resultados obtenidos en el paso anterior se comienza la
%actualizacion de las matrices componentes de la solucion. Se comienza por
%x10
x10=A1*x20+b1;
%matriz que ya usamos para actualizar x20
x20=A2*x10+A3*x30+b2;
%y de nuevo el elemento actualizado x20 se usa para actualizar x30
x30=A4*x20+25;
%Una vez terminado el paso tenemos definido el vector actualizado y se
%reiteran las sentencias de calculo en el orden expuesto para obtener
%la actualizacion siguiente
```

Los resultados en *format long* son

$$\mathbf{x}^{(1)} = \begin{pmatrix} 41.66666666666 \\ 66.66666666667 \\ 86.11111111111 \\ 77.77777777779 \\ 44.44444444444 \end{pmatrix} \quad \mathbf{x}^{(2)} = \begin{pmatrix} 45.74074074074 \\ 82.96296296296 \\ 91.75308641975 \\ 84.04938271605 \\ 46.01234567901 \end{pmatrix} \quad \mathbf{x}^{(3)} = \begin{pmatrix} 46.11687242798 \\ 84.46748971193 \\ 92.25882030178 \\ 84.56779149520 \\ 46.14194787380 \end{pmatrix} ....$$

$$....\mathbf{x}^{(8)} = \begin{pmatrix} 46.15384596861 \\ 84.61538387446 \\ 92.30769206769 \\ 84.61538439628 \\ 46.15384609907 \end{pmatrix} \quad \mathbf{x}^{(9)} = \begin{pmatrix} 46.15384613785 \\ 84.61538455138 \\ 92.30769228697 \\ 84.61538459651 \\ 46.15384614913 \end{pmatrix} \quad \mathbf{x}^{(10)} = \begin{pmatrix} 46.15384615246 \\ 84.61538460986 \\ 92.30769230590 \\ 84.61538461376 \\ 46.15384615344 \end{pmatrix}$$

Además $\|\mathbf{x}^{(10)} - \mathbf{x}^{(9)}\|_\infty = (0.584768)10^{-7}$. También hemos calculado la norma intermedia $\|\mathbf{x}^{(7)} - \mathbf{x}^{(6)}\|_\infty = (0.903395)10^{-4}$ que después compararemos con la correspondiente del método con superrelajación.

En este caso podemos hallar una estimación del radio espectral de la matriz de Gauss-Seidel $G(A;\Pi) = (D-L)^{-1}U$ para k suficientemente grande por el cociente

$$\frac{\|\mathbf{x}^{(10)} - \mathbf{x}^{(9)}\|_\infty}{\|\mathbf{x}^{(9)} - \mathbf{x}^{(8)}\|_\infty} = \frac{(0.584768)10^{-7}}{(0.676923)10^{-6}} \sim 0.0864 \ll 1$$

Si efectuamos un cálculo directo de ese radio espectral utilizando Matlab tendremos:

$$(D-L)^{-1}U = \begin{pmatrix} 0.2667 & 0.0667 & 0 & 0 & 0 \\ 0.0667 & 0.2667 & 0 & 0 & 0 \\ -0.0178 & -0.0711 & 0.2667 & 0.0667 & 0 \\ -0.0044 & -0.0178 & 0.0667 & 0.2667 & 0 \\ 0.0011 & 0.0044 & -0.0167 & -0.0667 & 0.2500 \end{pmatrix}.$$

$$\begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 0 & -0.0667 & 0 & 0 \\ 0 & 0 & -0.2667 & 0 & 0 \\ 0 & 0 & 0.0711 & 0 & -0.0667 \\ 0 & 0 & 0.0178 & -1 & -0.2667 \\ 0 & 0 & -0.0044 & 0 & 0.0667 \end{pmatrix}$$

y usando el comando *eig(B)* obtenemos un vector con los valores propios de la matriz de Gauss-Seidel

$$\begin{pmatrix} 0 & 0 & 0 & 0.0862 & 0.0515 \end{pmatrix}^T$$

de donde se deduce lo ajustado de nuestra estimación anterior.

3. Se busca acelerar la covergencia del método anterior introduciendo un factor de superrelajación ω y se define el método iterativo relativo a la partición admisible Π por

$$\begin{cases} A_1\overline{\mathbf{X}}_1^{(k+1)} = -B_1\mathbf{X}_2^{(k)} + \mathbb{b}_1 \\ A_2\overline{\mathbf{X}}_2^{(k+1)} = -C_2\mathbf{X}_1^{(k+1)} - B_2\mathbf{X}_3^{(k)} + \mathbb{b}_2 \\ A_3\overline{\mathbf{X}}_3^{(k+1)} = -C_3\mathbf{X}_2^{(k+1)} + \mathbb{b}_3 \\ \mathbf{X}_i^{(k+1)} = \mathbf{X}_i^{(k)} + \omega\left(\overline{\mathbf{X}}_i^{(k+1)} - \mathbf{X}_i^{(k)}\right) \quad (i=1,2,3) \end{cases}$$

Se calcula $\overline{\mathbf{X}}_i^{(k+1)}$ por el método de Gauss-Seidel y luego se corrige mediante el factor de convergencia ω para obtener $\mathbf{X}_i^{(k+1)}$.

Escribamos con detalle las distintas instrucciones de un paso de esta iteración

$$\begin{cases} \overline{\mathbf{X}}_1^{(k+1)} = -A_1^{-1}B_1\mathbf{X}_2^{(k)} + A_1^{-1}\mathbb{b}_1 \\ \mathbf{X}_1^{(k+1)} = \mathbf{X}_1^{(k)} + \omega\left(\overline{\mathbf{X}}_1^{(k+1)} - \mathbf{X}_1^{(k)}\right) \\ \overline{\mathbf{X}}_2^{(k+1)} = -A_2^{-1}C_2\mathbf{X}_1^{(k+1)} - A_2^{-1}B_2\mathbf{X}_3^{(k)}A_2^{-1} + \mathbb{b}_2 \\ \mathbf{X}_2^{(k+1)} = \mathbf{X}_2^{(k)} + \omega\left(\overline{\mathbf{X}}_2^{(k+1)} - \mathbf{X}_2^{(k)}\right) \\ \overline{\mathbf{X}}_3^{(k+1)} = -A_3^{-1}C_3\mathbf{X}_2^{(k+1)} + A_3^{-1}\mathbb{b}_3 \\ \mathbf{X}_3^{(k+1)} = \mathbf{X}_3^{(k)} + \omega\left(\overline{\mathbf{X}}_3^{(k+1)} - \mathbf{X}_3^{(k)}\right) \end{cases}$$

Manteniendo el significado y el valor de los datos del apartado anterior, habría que añadir la definición de ω.

Los resultados en *format long* para el valor de $\omega = 1.06$ óptimo son

$$\mathbf{x}^{(1)} = \begin{pmatrix} 44.43333333333334 \\ 74.73333333333333 \\ 90.55795555555557 \\ 80.01448888888891 \\ 46.20383955555556 \end{pmatrix} \quad \mathbf{x}^{(2)} = \begin{pmatrix} 46.13342885925926 \\ 84.71371543703705 \\ 92.44400422546174 \\ 84.91251853858766 \\ 46.22958703939240 \end{pmatrix} \quad .....$$

$$..... \mathbf{x}^{(6)} = \begin{pmatrix} 46.15384671676641 \\ 84.61538919986565 \\ 92.30769579116469 \\ 84.61538763490647 \\ 46.15384612043506 \end{pmatrix} \quad \mathbf{x}^{(7)} = \begin{pmatrix} 46.15384636623632 \\ 84.61538532497728 \\ 92.30769229690111 \\ 84.61538447491365 \\ 46.15384611862601 \end{pmatrix} \quad ....$$

$$..... \mathbf{x}^{(9)} = \begin{pmatrix} 46.15384615361501 \\ 84.61538461395612 \\ 92.30769230823317 \\ 84.61538461581019 \\ 46.15384615390768 \end{pmatrix} \quad \mathbf{x}^{(10)} = \begin{pmatrix} 46.15384615389824 \\ 84.61538461562321 \\ 92.30769230773166 \\ 84.61538461539334 \\ 46.15384615384478 \end{pmatrix}$$

de donde

$$\mathbf{x}^{(7)} - \mathbf{x}^{(6)} = \begin{pmatrix} -0.00000035053009 \\ -0.00000387488837 \\ -0.00000349426358 \\ -0.00000315999282 \\ -0.00000000180905 \end{pmatrix} \quad \mathbf{x}^{(10)} - \mathbf{x}^{(9)} = \begin{pmatrix} 0.00000000028323 \\ 0.00000000166709 \\ -0.00000000050151 \\ -0.00000000041685 \\ -0.0000000000629 \end{pmatrix}$$

con lo que $\|\mathbf{x}^{(7)} - \mathbf{x}^{(6)}\|_\infty = (0.387488837)10^{-5}$ y $\|\mathbf{x}^{(10)} - \mathbf{x}^{(9)}\|_\infty = (0.166709)10^{-8}$ que son bastante mejores que las correspondientes del método sin factor de convergencia.

El proceso iterativo se para, de acuerdo con nuestro test de parada, en la séptima iteración ya que

$$\frac{\|\mathbf{x}^{(7)}-\mathbf{x}^{(6)}\|_\infty}{\|\mathbf{x}^{(6)}\|_\infty}=\frac{0.387488837)10^{-5}}{92.30769579116469}=4.1978010^{-8}\leq 10^{-7}$$

Comentarios

a) Existen otras posibles descomposiciones $A=LU$. Por ejemplo

$$U=\begin{pmatrix} 1 & -\frac{1}{4} & 0 & 0 & 0 \\ 0 & 1 & -\frac{4}{15}0 & 0 & 0 \\ 0 & 0 & 1 & -\frac{15}{56} & 0 \\ 0 & 0 & 0 & 1 & -\frac{56}{209} \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \quad L=\begin{pmatrix} 4 & 0 & 0 & 0 & 0 \\ -1 & \frac{15}{4} & 0 & 0 & 0 \\ 0 & -1 & \frac{56}{15} & 0 & 0 \\ 0 & 0 & -1 & \frac{209}{56} & 0 \\ 0 & 0 & 0 & -1 & \frac{780}{209} \end{pmatrix}$$

b Ahora ya con más detalle veamos de qué modo programaríamos en Matlab el método de Gauss-Seidel que hemos usado en este problema.

Lo primero será definir en un archivo m. una función GaussSeidel especial para este problema (*tridiagonal2.m*)

```
function[Xout, Yout, Zout, error]=GaussSeidel(A1, A2, A3, A4, Xin, Yin, Zin, b1, b2, b3);
Xout=A1*Xin+b1;
Yout=A2*Xout+A3*Zin+b2;
Zouy=A4*Yout+b3;
return;
```

Los datos son los ya definidos y tan sólo hemos cambiado el nombre de las variables x10, x20 y x30 que llamamos aquí con mejor criterio Xin, Yin y Zin cuando son datos que entran en el programa y con "out" a la salida.

En otro archivo m. incluimos los datos del problema y el motor de inferencia (*tridiagonal3.m*)

```
%Definimos los datos del problema
A1=[1/15 0
    4/15 0];
A2=[0 4/15
    0 1/15];
A3=[1/15 4/15]';
A4=[0 1/4];
x10=[25 50];
%Este elemento es innecesario en este algoritmo pero es necesario en el
%algoritmo del metodo de relajacion
x20=[50 50];
x30=25;
b1=[40 60];
b2=[200/3 200/3];
b3=100;
error=1;
i=0;
while (error>1e-5)
    [X10, X20, X30, error]=GaussSeidel(A1, A2, A3, A4, X10, X20, X30, b1, b2, b3);
    i=i+1;
```

```
    [i, error]
end;
X10
X20
X30
```

PROBLEMA 2.11 *Resolución de un sistema de ecuaciones lineales por el método de aproximaciones sucesivas.*

Se transforma el sistema de ecuaciones lineales

$$(S) \qquad \begin{cases} 8x_1 + x_2 - 2x_3 - 8 = 0 \\ x_1 + 18x_2 - 6x_3 + 10 = 0 \\ 2x_1 + x_2 + 16x_3 + 2 = 0 \end{cases}$$

en el siguiente problema de punto fijo $\mathbf{x} = T(\mathbf{x})$

$$\begin{cases} x_1 = 0.2x_1 - 0.1x_2 + 0.2x_3 + 0.8 \\ x_2 = -0.05x_1 + 0.1x_2 + 0.3x_3 - 0.5 \\ x_3 = -0.1x_1 - 0.05x_2 + 0.2x_3 - 0.1 \end{cases}$$

donde T es la función lineal afín

$$T(\mathbf{x}) = A\mathbf{x} + \mathbf{b}$$

con

$$A = \begin{pmatrix} 0.2 & -0.1 & 0.2 \\ -0.05 & 0.1 & 0.3 \\ -0.1 & -0.05 & 0.2 \end{pmatrix} \quad \text{y} \quad \mathbf{b} = \begin{pmatrix} 0.8 \\ -0.5 \\ -0.1 \end{pmatrix}$$

1. a) Verificar la equivalencia de los dos problemas planteados.
   b) Probar que T es contractiva en $\mathbb{R}^3$ entero, determinando una constante de Lipschitz L
   c) Hallar el radio espectral de su matriz jacobiana.

2. a) Utilizar la estructura de diagonal estrictamente dominante de la matriz asociada al sistema (S) para hallar un estimador inicial adecuado $\mathbf{x}^{(0)}$ del método de aproximaciones sucesivas.
   b) Realizar al menos 10 iteraciones de dicho método.
   c) Hallar estimaciones "a priori" y "a posteriori" del error cometido.
   d) Número de iteraciones k necesarias para conseguir que el error sea menor o igual que $\epsilon = 10^{-6}$.
   e) Verificar el factor o cociente de covergencia de los sucesivos vectores calculados.

Solución:

1. a) Escribiendo los coeficientes de la ecuación de punto fijo como cocientes y operando se obtienen las ecuaciones de (S). Como muestra, tomemos la última ecuación

$$x_3 = -\frac{1}{10}x_1 - \frac{5}{100}x_2 + \frac{2}{10}x_3 - \frac{1}{10} \quad \Rightarrow$$
$$\Rightarrow \quad 20x_3 + 2x_1 + x_2 - 4x_3 + 2 = 16x_3 + 2x_1 + x_2 + 2 = 0$$

que es la tercera ecuación de (S).

b) Para demostrar que T es contractiva en $\mathbb{R}^3$ será suficiente probar que para alguna norma matricial la norma de su matriz jacobiana $\mathbf{T}(\mathbf{x})$ está mayorada por una constante L estrictamente menor que 1 para todo $\mathbf{x} \in \mathbb{R}^3$, en cuyo caso T es L-lipchiciana para la norma vectorial compatible.

Como T es lineal afín, su diferencial en cualquier punto es constante e igual a la aplicación lineal asociada, luego $(\forall \mathbf{x} \in \mathbb{R}^3) \quad \mathbf{T}(\mathbf{x}) = A$ y el problema se limita a probar que $\|A\| \leq L < 1$ para alguna norma matricial inducida.

Con objeto de aprovechar bien el ejercicio, utilizaremos varias de esas normas y también la norma de Schur que no es inducida por ninguna norma vectorial pero es fácil de calcular y permite enmarcar la norma $\|A\|_2$ (ver la sección 2.1.3 donde se encuentran las definiciones y algunos comentarios sobre las normas habituales).

Comencemos con la norma $\|A\|_1$ asociada a la norma vectorial $\|\|_1$ de $\mathbb{R}^3$

$$\|A\|_1 = \max_j \sum_i^n |a_{ij}| = \max\{0.35, 0.25, 0.7\} = 0.7$$

Con este resultado ya estaría contestado este apartado y $L = 0.7 < 1$ es una constante de Lipschitz.

Asociada a la norma vectorial $\|, \|_\infty$ se define la norma matricial $\|A\|_\infty$ y

$$\|A\|_\infty = \max_i \sum_j^n |a_{ij}| = \max\{0.5, 0.45, 0.35\} = 0.5$$

hemos mejorado la mayoración anterior.

La norma de Schur está definida por

$$\|A\|_S = \left[\sum_{i,j}^n |a_{ij}|^2\right]^{1/2} = \sqrt{0.245} \approx 0.495$$

y cumple las desigualdades (sección 2.1.3)

$$\frac{1}{\sqrt{3}}\|A\|_S \leq \|A\|_2 \leq \|A\|_S \quad \Rightarrow \quad 0.28561 \leq \|A\|_2 \leq 0.495$$

mayoraciones muy interesantes en el estudio del problema.

c) El radio espectral de A es el máximo de los valores absolutos de los valores propios de A y es siempre inferior a $\|A\|$ para cualquier norma matricial.

Para estimarlo estudiaremos el polinomio característico de A

$$\det(A - \lambda I) = -\lambda^3 + 0.5\lambda^2 - 0.11\lambda + 0.0115 = 0$$

La cúbica $f(\lambda) = -\lambda^3 + 0.5\lambda^2 - 0.11\lambda + 0.0115$ se representa en la Figura 2.5 Como su trinomio derivada tiene raíces imaginarias y es siempre negativa, f es decreciente. Cambia su concavidad en $\lambda = \dfrac{1}{6}$ con una tangente muy horizontal de pendiente -0.026. Sólo tiene una raíz real que es fácil enmarcar estudiando el cambio de signo de f, en el intervalo $(0.16667, 0.4)$. Utilizaremos para aproximarla el método de Newton con estimador inicial $\lambda^0 = 0.4$, se tiene sucesivamente

$$\lambda^1 = \lambda^0 - \frac{f(\lambda^0)}{f'(\lambda^0)} = 0.4 - \frac{0.0165}{0.19} = 0.313158$$

$$\lambda^2 = 0.313158 - -\frac{0.0046242}{0.091046} = 0.26237$$

$$\lambda^3 = 0.26237 - \frac{0.001003}{0.054144} = 0.2438454 \quad \text{etc.}$$

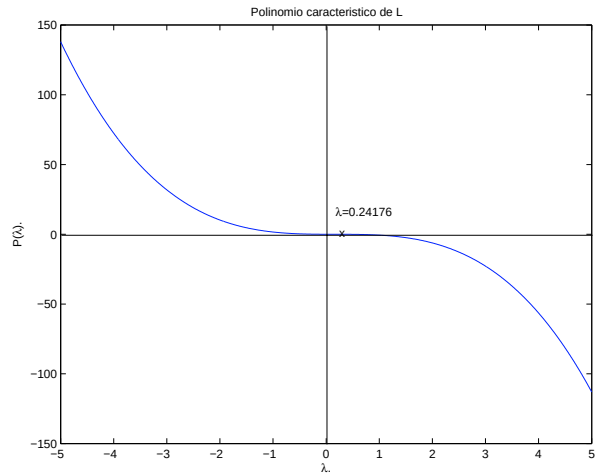


Figura 2.5: Grafo del polinomio característico de A.

Para este último valor se tiene $f(0.2438454) = -0.0000917$, nos vamos acercando a la raíz por la derecha, luego

$$\rho(A) = 0.243845$$

en concordancia con todas las mayoraciones obtenidas [25].

2. a) Eliminando del sistema (S) los elementos extradiagonales obtenemos el sistema diagonal

$$\begin{cases} 8x_1 & -8 & = 0 \\ 18x_2 & +10 & = 0 \\ 16x_3 & +2 & = 0 \end{cases}$$

cuya solución $x_1 = 1$, $x_2 = -\dfrac{5}{9} = -0.55\ldots$ y $x_3 = -\dfrac{1}{8} = 0.125$ debe presumiblemente estar próxima a la solución de (S) por lo que será un buen estimador inicial. Para simplificar, tomaremos como estimador $\mathbf{x}^{(0)} = (1, -0.5, -0.2)$.

b) Utilizando ese estimador inicial y programando en Matlab la sucesión recurrente del método de aproximaciones sucesivas

$$\mathbf{x}^{(n)} = T(\mathbf{x}^{(n-1)}) = A\mathbf{x}^{(n-1)} - \mathbf{b}$$

se obtiene sucesivamente

$$\mathbf{x}^{(0)} = \begin{pmatrix} 1 \\ -0.5 \\ -0.2 \end{pmatrix} \quad \mathbf{x}^{(1)} = \begin{pmatrix} -0.5900 \\ 0.3400 \\ -0.0150 \end{pmatrix} \quad \mathbf{x}^{(2)} = \begin{pmatrix} -0.9550 \\ 0.5590 \\ 0.1390 \end{pmatrix}$$

$$\mathbf{x}^{(3)} = \begin{pmatrix} -1.0191 \\ 0.6453 \\ 0.1954 \end{pmatrix} \quad \mathbf{x}^{(4)} = \begin{pmatrix} -1.0293 \\ 0.6741 \\ 0.2087 \end{pmatrix} \quad \mathbf{x}^{(5)} = \begin{pmatrix} -1.0315 \\ 0.6815 \\ 0.2110 \end{pmatrix}$$

$$\mathbf{x}^{(6)} = \begin{pmatrix} -1.0315 \\ 0.6830 \\ 0.2113 \end{pmatrix} \quad \mathbf{x}^{(7)} = \begin{pmatrix} -1.0325 \\ 0.6833 \\ 0.2113 \end{pmatrix} \quad \mathbf{x}^{(8)} = \begin{pmatrix} -1.0326 \\ 0.6834 \\ 0.2114 \end{pmatrix}$$

La iterada siguiente $\mathbf{x}^{(9)} = \mathbf{x}^{(8)}$ por lo que paramos el proceso.

[25] La matriz A con valores propios distintos $\lambda_1 = \alpha + i\beta$, $\lambda_2 = \overline{\lambda_1} = \alpha - i\beta$ y $\lambda_3 \approx 0.2438454$ es diagonalizable en $\mathbb{C}$, luego es semejante a $\mathrm{diag}\,(\alpha + i\beta, \alpha - i\beta, \lambda_3)$ y comparten el mismo polinomio característico. Este razonamiento nos permitirá hallar los valores propios complejos de A. En efecto, $\mathrm{Tr}A = 2\alpha + \lambda_3 = 0.5 \quad \Rightarrow \quad \alpha \approx 0.1280773$ y $\det A = |\lambda_1|^2\lambda_3 = 0.0115 \quad \Rightarrow$ $|\lambda_1| = \sqrt{\dfrac{0.0115}{0.2438454}} \approx 0.2173$. Obsérvese que su módulo es menor que $|\lambda_3|$ y por tanto el radio espectral es $|\lambda_3|$. Por último,$\alpha^2 + \beta^2 = 0.047161$ de donde $\beta \approx 0.1754$ y $\lambda_1 = 0.1281 + i0.1754$.

c) Llamando a la solución exacta del problema $\mathbf{x}^*$, sabemos que

$$\|\mathbf{x}^* - \mathbf{x}^{(k)}\| \leq \frac{L^k}{1-L}\|\mathbf{x}^{(1)} - \mathbf{x}^{(0)}\| \qquad k = 1, 2, ...$$

luego con $L = 0.25$ y $k = 8$ y tomando la norma $\|,\|_\infty$ obtenemos la siguiente estimación "a priori" del error

$$\|\mathbf{x}^* - \mathbf{x}^{(8)}\|_\infty \leq \frac{0.25^8}{0.75}\|\mathbf{x}^{(1)} - \mathbf{x}^{(0)}\|_\infty =$$
$$= (0.0000152588)(1.59) = 2 \cdot 10^{-5} < 10^{-4}$$

La estimación del error cometido al tomar $\mathbf{x}_8$ como solución del problema "a posteriori" es

$$\|\mathbf{x}^* - \mathbf{x}^{(8)}\|_\infty \leq \frac{L}{1-L}\|\mathbf{x}^{(8)} - \mathbf{x}^{(7)}\|_\infty < \frac{0.25}{0.75}0.0001 \approx 3 \cdot 10^{-5}$$

es curioso que el error "a posteriori" sea mayor que el error "a priori" que debería ser más pesimista.

d) El número de iteraciones k necesarias para conseguir que el error $\|\mathbf{x}^* - \mathbf{x}^{(k)}\|$ sea menor o igual que un cierto número ϵ por ejemplo $\epsilon = 10^{-6}$ es

$$k \geq \frac{\ln\left(\frac{\epsilon\,(0.75)}{\|\mathbf{x}^{(1)} - \mathbf{x}^{(0)}\|}\right)}{\ln 0.25} = 15.9532 \sim 16$$

e) Sabemos que la sucesión $\{\mathbf{x}^{(k)}\}$ converge linealmente con factor de convergencia $\|T'(\mathbf{x}^*)\|$ (Capítulo 1, sección 1.3.2, teorema 1.3.1).
En este tipo de convergencia, después de m iteraciones se añade otro decimal correcto al valor calculado. De

$$\|\epsilon^{(k+m)}\| \sim \|T'(\mathbf{x}^*)\|\|\epsilon^{(k)}\|$$

se deduce que

$$m \geq \frac{-1}{\log\|T'(\mathbf{x}^*)\|}$$

Se requiere para k suficientemente grande el mismo número m de pasos de iteración para reducir la magnitud de $\|\epsilon^{(k)}\|$ en 0.1.
En nuestro caso, ya que no conocemos $\|T'(\mathbf{x}^*)\|$, lo estimamos con la constante de Lipschitz $L = 0.25$ con lo que

$$m \geq \frac{-1}{\log(0.25)} = -1.6609 \sim 2$$

Aproximadamente a partir de un valor de k suficientemente grande cada dos iteraciones se consolida un decimal del valor calculado, por ejemplo

$$\mathbf{x}^{(5)} = \begin{pmatrix} -1.03\cdots \\ 0.68\cdots \\ 0.21\cdots \end{pmatrix} \qquad \mathbf{x}^{(7)} = \begin{pmatrix} -1.032\cdots \\ 0.683\cdots \\ 0.211\cdots \end{pmatrix}$$
$$\mathbf{x}^{(9)} = \begin{pmatrix} -1.0325\cdots \\ 0.6833\cdots \\ 0.2113\cdots \end{pmatrix} \qquad \mathbf{x}^{(11)} = \begin{pmatrix} -1.03258\cdots \\ 0.68337\cdots \\ 0.21136\cdots \end{pmatrix}$$
$$\mathbf{x}^{(13)} = \begin{pmatrix} -1.032581\cdots \\ 0.683375\cdots \\ 0.211361\cdots \end{pmatrix} \quad \mathbf{x}^{(15)} = \begin{pmatrix} -1.03258145\cdots \\ 0.68337510\cdots \\ 0.21136173\cdots \end{pmatrix} \cdots$$

y se comprueba que el error absoluto es ya $< 10^{-6}$ en la iterada $\mathbf{x}^{(15)}$.

Comentario

Como es fácil ver, el problema está resuelto a mano con una calculadora normal excepto en los procesos iterativos que se programaron en Matlab (ver en la página web asociada el código *metaproxsuc1.m*).

PROBLEMA 2.12 *Estudio del polinomio característico y de los valores propios de una matriz de orden 4 que estudió Leverrier.*

En los cálculos que condujeron a Leverrier [26] al descubrimiento del planeta Neptuno en 1847, una etapa muy importante fue la determinación de los valores propios de la matriz

$$L = \begin{pmatrix} -5.509882 & 1.870086 & 0.422908 & 0.008814 \\ 0.287865 & -11.811654 & 5.711900 & 0.058717 \\ 0.049099 & 4.308033 & -12.970687 & 0.229326 \\ 0.006235 & 0.269851 & 1.397369 & -17.596207 \end{pmatrix}$$

Efectuaremos este cálculo utilizando una variante del método que utilizó originalmente Leverrier y se darán los resultados al menos con seis cifras significativas.

1. Sea L' la matriz obtenida anulando en L aquellos elementos cuyo valor absoluto sea menor que 1. Determinar utilizando un cálculo directo, los valores propios de L' que denotaremos $\lambda'_1, \lambda'_2, \lambda'_3, \lambda'_4$, dispuestos en orden decreciente.

2. Sean $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ los valores propios de L que admitiremos que son negativos y distintos [27], y numerados de modo que $\lambda_1 > \lambda_2 > \lambda_3 > \lambda_4$.

   Se pone
   $$s_n = \lambda_1^n + \lambda_2^n + \lambda_3^n + \lambda_4^n \qquad (\forall n \geq 0)$$

   2.1 Demostrar que s_n es la traza de L^n potencia n-ésima de L.

   2.2 Estudiar el límite cuando $n \to \infty$ de las sucesiones
   $$\sigma_n = \frac{s_n}{s_{n-1}} \qquad \text{y} \qquad \tau_n = |s_n|^{1/n} \quad (n \in I\!N \ )$$

3. Determinar L^2. Deducir los números s_1, s_2, s_3, s_4.

4. Sea $P(\lambda) = \lambda^4 + c_1\lambda^3 + c_2\lambda^2 + c_3\lambda + c_4$ el polinomio característico de L.

   4.1 Utilizar las relaciones de Cardano-Vieta [28] para demostrar la relación
   $$s_4 + c_1 s_3 + c_2 s_2 + c_3 s_1 + c_4 s_0 = 0$$

[26] Urbain Jean Joseph Leverrier (1811-1871), codescubridor con John Couch Adams (1819-1892) independientemente y casi de modo simultáneo del planeta Neptuno utilizando métodos puramente matemáticos. Uno de los más claros ejemplos del verdadero valor de las teorías matemáticas en las ciencias físicas. El descubrimiento del planeta Neptuno fue el resultado de un análisis de las perturbaciones del planeta Urano de acuerdo con la teoría de la gravedad de Newton.

[27] Hipótesis que implica que L es diagonalizable.

[28] Profesor de física y matemática de Milán, Girolamo Cardano (1501-1576) pertenece al grupo de algebristas del Renacimiento italiano del siglo XVI al que también pertenecieron Scipio del Ferro, Tartaglia y Ferrari. Centraron su trabajo en la resolución de las ecuaciones algebraicas de tercer y cuarto grado con total éxito, obteniendo resultados que superaban por primera vez en nuestra era los logros de la matemática antigua.

François Viète (1540-1603) fue el mejor matemático francés del siglo XVI y estableció para los polinomios de grado ≤ 5 las fórmulas que expresan sus coeficientes como funciones simétricas de sus raíces.

Si $P_n(x) = x^n + c_1 x^{n-1} + \cdots + c_n$ es un polinomio de grado n y $\{\lambda_i\}$ $(i = 1, ..., n)$ son las n raíces de P_n iguales o distintas (en el caso de una raíz múltiple con orden de multiplicidad m, esa raíz aparece m veces en la lista) las fórmulas de Vieta son

$$\begin{cases} -c_1 = \lambda_1 + \lambda_2 + \cdots + \lambda_n \\ c_2 = \lambda_1\lambda_2 + \lambda_1\lambda_3 + \cdots + \lambda_{n-1}\lambda_n \\ -c_3 = \lambda_1\lambda_2\lambda_3 + \lambda_1\lambda_2\lambda_4 + \cdots\cdots \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ (-1)^n c_n = \lambda_1\lambda_2\cdots\lambda_n \end{cases}$$

4.2 Escribir el polinomio característico de L.

4.3 Comprobar que se verifica la igualdad

$$\sum_{i=1,4} (\lambda_i)^{n-4} P(\lambda_i) = 0$$

4.4 Determinar utilizando la igualdad anterior la relación

$$s_n + c_1 s_{n-1} + c_2 s_{n-2} + c_3 s_{n-3} + c_4 s_{n-4} = 0$$

válida para todo entero $n \geq 4$.

Calcular usando esta relación los números σ_n y τ_n para $1 \leq n \leq 16$. Se dispondrá el resultado en una tabla del tipo

n	s_n	σ_n	τ_n

Utilizamos ahora una variante del método de Leverrier para hacer el cálculo aproximado simultáneo de los dos valores propios de mayor módulo.

5. Sean a_n y b_n dos sucesiones de números y

$$(x - a_n)(x - b_n) = x^2 - \lambda_n x + \mu_n \qquad \text{con} \quad a_n \geq b_n,$$

comprobar que si $\lambda_n \to a + b$ y $\mu_n \to ab$ entonces $a_n \to a$ y $b_n \to b$.

6. Se definen dos sucesiones de números a_n y b_n mediante las relaciones

$$a_n + b_n = \sigma_n + \delta_n \sigma_{n-2}$$
$$a_n b_n = \sigma_{n-1} \sigma_{n-2} \delta_n$$

con $a_n \geq b_n$ y $\delta_n = \dfrac{\sigma_n - \sigma_{n-1}}{\sigma_{n-1} - \sigma_{n-2}}$.

Teniendo en cuenta que las tres sucesiones σ_n, σ_{n-1} y σ_{n-2} tienen el mismo límite, utilizar la conclusión del apartado anterior para demostrar que cada una de las sucesiones a_n y b_n converge hacia uno de los valores propios de L de mayor módulo.

Calcular los términos a_n y b_n con $3 \leq n \leq 16$.

7. Lo anterior permite calcular un valor aproximado de dos de las raíces de $P(\lambda)$. ¿Cómo se pueden calcular aproximadamente las otras dos?

8. Para evaluar la fiabilidad del método, aplicarlo al cálculo aproximado de los valores propios de L' ya realizado en 1 y que son suficientemente cercanos.

Comentarios y cuestiones suplementarias

1. El algoritmo de cálculo directo de los coeficientes $c_i \quad i = 1, ..., n$ del polinomio característico de una matriz cuadrada A de orden n y del valor propio dominante λ_n descrito en los apartados 2 a 5 del enunciado, proviene directamente de Leverrier (1840). Se trata de un método muy costoso numéricamente. El algoritmo requiere $n^3(n-1)$ multiplicaciones, para calcular las potencias sucesivas de A; $A^2, A^3, ..., A^n$ y $(n-1)(n+2)/2$ para calcular los c_i. ¡Demasiadas! Según hemos propuesto aquí, es suficiente calcular las potencias $A^2, A^3, ..., A^p$ con $p = n/2$ (resp. $p = (n+1)/2$) si n es par (resp. si n es impar) y los términos diagonales del resto de las potencias $A^{p+1}, ..., A^n$, lo que reduce a la mitad el número de operaciones, aunque este planteamiento sólo es posible en el cálculo a mano.

2. Para probar la bondad de Matlab como herramienta de álgebra numérica, una vez terminado el ejercicio, haremos los cálculos que se piden usando los comandos que el código Matlab posee en esta área.

   Estas instrucciones ponen en marcha la selección y ejecución de una serie de algoritmos de manejo muy engorroso, como comprobamos en este enunciado, que provienen de las mejores librerías de rutinas de álgebra lineal numérica (Linpack y Eispack).

2.1 Utilizando directamente los comandos Matlab obtener el polinomio característico y los valores y vectores propios de L.
Representar gráficamente el polinomio característico de L que obtuvimos en el apartado 4.1 y hallar sus raíces utilizando la función *roots*. Comparar el resultado obtenido con el del comando $eig(L)$.

2.2 Efectuar un escalado de la matriz L mediante la función *balance* y calcular posteriormente sus valores propios. Comparar L con $balance(L)$ y los dos conjuntos de valores propios.
¿Qué hubiera pensado Leverrier si hubiera tenido esta herramienta en su época?

3. Aprovechemos la ocasión para verificar las diferentes mayorantes de los valores propios de una matriz que se incluyen en el resumen teórico. Comprobar que todo valor propio λ de L satisface la desigualdad

$$|\lambda| \leq \|L\|$$

para cualquier norma matricial natural $\| \|$.

Este enunciado es una modificación del problema propuesto en 1985 en la prueba práctica del Concurso de Álgebra de las "Ecoles normales supèrieures d'Ulm et de Sèvres" en Francia.

Solución:

1.
$$L' = \begin{pmatrix} -5.509882 & 1.870086 & 0 & 0 \\ 0 & -11.811654 & 5.711900 & 0 \\ 0 & 4.308033 & -12.970687 & 0 \\ 0 & 0 & 1.397369 & -17.596207 \end{pmatrix}$$

La factorización de $\det(L' - \lambda I)$ es inmediata.

$$\det(L' - \lambda I) = (5.509882 - \lambda)(-17.596207 - \lambda)(\lambda^2 + 24.782341128\lambda + 128.598213293598)$$

Se obtienen los valores propios de L' resolviendo una ecuación de segundo grado

$$\lambda' = \begin{cases} -7.39688436349626054 \\ -17.3854567645037395 \end{cases}$$

Ordenándolos como sugiere el enunciado tendremos:

$$\lambda'_1 = -5.509882\ ; \quad \lambda'_2 = -7.396884\ ; \quad \lambda'_3 = -17.385457\ ; \quad \lambda'_4 = -17.596207$$

2. 2.1 Como L es diagonalizable, en una base de vectores propios se representa mediante la matriz diagonal de sus valores propios $\mathrm{diag}(\lambda_1, \lambda_2, \lambda_3, \lambda_4)$, y sus potencias sucesivas también diagonales tienen como elementos las potencias correspondientes de esos valores, de donde $s_n = \sum_{i=1}^{4} \lambda_i^n$.

2.2 Sacando factor común $(\lambda_4)^n$ en s_n,

$$s_n = \lambda_4^n \left(1 + \left(\frac{\lambda_3}{\lambda_4}\right)^n + \left(\frac{\lambda_2}{\lambda_4}\right)^n + \left(\frac{\lambda_1}{\lambda_4}\right)^n\right)$$

Como todos los λ_i son negativos y λ_4 es el de mayor valor absoluto, $\lambda_i/\lambda_4 < 1 \ \ (i = 1, 2, 3)$ luego $(\lambda_i/\lambda_4)^n \to 0$ cuando $n \to \infty$ y $s_n \simeq (\lambda_4)^n$, por tanto,

$$\lim_{n\to\infty} \sigma_n = \lim_{n\to\infty} \frac{s_n}{s_{n-1}} = \frac{\lambda_4^n}{\lambda_4^{n-1}} = \lambda_4$$

Ya que $|s_n| = \sum_{i=1,4} |\lambda_i|^n$ se obtiene de modo análogo,

$$\lim_{n\to\infty} \tau_n = \lim_{n\to\infty} |s_n|^{\frac{1}{n}} = |\lambda_4| = -\lambda_4$$

3. Multiplicando L por sí misma obtenemos con ocho cifras significativas

$$L^2 = \begin{pmatrix} 30.917951 & -30.568482 & 2.878480 & 0.003133 \\ -4.705449 & 164.676401 & -141.350463 & -0.414317 \\ 0.334184 & -106.609440 & 193.186992 & -6.756396 \\ 0.002224 & -1.904170 & -41.169231 & 309.962854 \end{pmatrix}$$

de modo que $s_1 = -47.888430$, $s_2 = 698.744198$.

Utilizando la igualdad

$$\mathrm{tr}(AB) = \sum_{i,j=1,n} a_{ij} b_{ji} \qquad AB \in M_n(\mathbb{R})$$

sólo necesitamos calcular los elementos diagonales de L^3 y L^4 y tendremos llamando $M = L^2$

$$s_3 = \mathrm{tr}(L^3) = \mathrm{tr}(L \cdot L^2) = \sum_{i,j=1}^{4} l_{ij} m_{ji} = -11329.701$$

y

$$s_4 = \mathrm{tr}(L^4) = \mathrm{tr}(L^2 \cdot L^2) = \sum_{i,j=1}^{4} m_{ij} m_{ji} = 192458.50$$

después de calcular esas dos tediosas sumas de 16 términos cada una [29].

4. 4.1 A partir de las fórmulas de Vieta se establecen sucesivamente las igualdades de Newton.

$$\begin{aligned} c_1 &= -s_1 \\ 2c_2 &= -s_2 - c_1 s_1 \\ 3c_3 &= -s_3 - c_1 s_2 - c_2 s_1 \\ 4c_4 &= -s_4 - c_1 s_3 - c_2 s_2 - c_3 s_1 \end{aligned}$$

por ejemplo,

$$\begin{aligned} 2c_2 = &- s_2 - c_1 s_1 = -\left(\lambda_1^2 + \lambda_2^2 + \lambda_3^2 + \lambda_4^2\right) + \left(\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4\right)^2 \quad \Rightarrow \\ &\Rightarrow c_2 = (\lambda_1\lambda_2 + \lambda_1\lambda_3 + \lambda_1\lambda_4 + \lambda_2\lambda_3 + \lambda_2\lambda_4 + \lambda_3\lambda_4) \end{aligned}$$

que es la segunda fórmula de Vieta.

4.2 Con ellas se obtiene

c_1	c_2	c_3	c_4
47.888430000	797.278764779	5349.45551533	12296.5505661

de modo que

$$P(\lambda) = \lambda^4 + 47.888430000\lambda^3 + 797.278764779\lambda^2 + 5349.45551533\lambda + 12296.5505661$$

[29] En el cálculo del producto de dos matrices de orden 4, necesitamos 4 multiplicaciones y 3 sumas por cada uno de los 16 elementos del producto, luego un total de 64 multiplicaciones y 48 sumas. Para hallar, por tanto, L^2, L^3 y L^4 necesitamos 192 multiplicaciones y 144 sumas. Si sólo calculamos L^2 y los elementos diagonales de L^3 y L^4, necesitamos 96 multiplicaciones y 72 sumas. ¡Nos ahorramos la mitad de las operaciones! En ambos casos necesitamos además otras 12 sumas para calcular los cuatro s_i y 9 multiplicaciones para hallar los c_i usando las fórmulas de Newton.

Este tipo de cuentas eran frecuentes no hace mucho tiempo. Cuando el cálculo era manual o con máquinas mecánicas, un ahorro como el expuesto era importante. Después, por motivos de capacidad de almacenamiento de información y de coste temporal, era obligatorio analizar a fondo el coste de los algoritmos. Se medía con el *flop* "floating point operation" que es la cantidad de tiempo de ordenador asociado a la sentencia $s := s + a_{ik} b_{kj}$.

Hoy en día la enorme capacidad y velocidad de nuestros ordenadores y los nuevos algoritmos cada vez más rápidos nos permiten ignorar estas cuentas.

4.3 Desarrollando la expresión del enunciado teniendo en cuenta que los λ_i son las raíces de P

$$\sum_{i=1,4}(\lambda_i)^{n-4}P(\lambda_i) = (\lambda_1)^{n-4}P(\lambda_1) + (\lambda_2)^{n-4}P(\lambda_2) + (\lambda_3)^{n-4}P(\lambda_3) + (\lambda_4)^{n-4}P(\lambda_4) =$$

$$\begin{aligned}
&=(\lambda_1)^{n-4}\left(\lambda_1^4 + c_1\lambda_1^3 + c_2\lambda_1^2 + c_4\lambda_1 + c_5\right) + \cdots = \\
&= (\lambda_1^n + \lambda_2^n + \lambda_3^n + \lambda_4^n) + c_1\left(\lambda_1^{n-1} + \lambda_2^{n-1} + \lambda_3^{n-1} + \lambda_4^{n-1}\right) + \cdots \quad \Rightarrow \\
&\Rightarrow s_n + c_1 s_{n-1} + c_2 s_{n-2} + c_3 s_{n-3} + c_4 s_{n-4} = 0
\end{aligned}$$

4.4 Utilizando esa fórmula recurrente se calculan los s_n en función de los c_i conocidos y de los tres valores s_i anteriores.

Hemos escrito un programa script *Leverrier2.m* para efectuar dicho cálculo que se incluye en la página web vinculada al libro. Con sus resultados se completa la tabla

n	s_n	σ_n	τ_n
1	-47.888430	-11.97210750000000	47.88843000000000
2	698.744198	-14.59108594706488	26.43377003002031
3	-11329.701	-16.21437577933205	22.45981369977575
4	192458.50	-16.98707671102706	20.94518854489397
5	$-3.0\cdot10^{6}$	-17.31616888487335	20.16312218579679
6	$5.8\cdot10^{7}$	-17.45388902250155	19.68400741466914
7	$-1.019\cdot10^{9}$	-17.51388534146537	19.35825612861988
8	$1.7871\cdot10^{10}$	-17.54259468159412	19.12139977133451
9	$-3.13796\cdot10^{11}$	-17.55858914248475	18.94110143117381
10	$5.513182\cdot10^{12}$	-17.56933301532567	18.79923748653306
11	$-9.6909960\cdot10^{13}$	-17.57786293302885	18.68478211382226
12	$1.70420323\cdot10^{15}$	-17.58542902700179	18.59060180287879
13	$-2.9981280934\cdot10^{16}$	-17.59255022336702	18.51185816600348
14	$5.27653825316\cdot10^{17}$	-17.59944234788991	18.44514518253941
15	$-9.289975793367\cdot10^{18}$	-17.60619434116677	18.38799197289179
16	$1.63622872847139\cdot10^{20}$	-17.61284167865894	18.33856085776775

Un comentario interesante sobre el resultado obtenido tiene que ver con la gran diferencia en magnitud entre los valores de s_n. El código ajusta la magnitud del conjunto al valor mayor, es decir, expresa todos los valores como producto de un número de 15 dígitos por 10^{20} de modo que los cuatro primeros elementos de la columna aparecen como ceros con signo en el resultado, s_5 se define mediante un solo dígito y conforme se progresa en la columna el valor aumenta su precisión. Para resolver este inconveniente deberíamos fraccionar la suma en varias. Una, por ejemplo, de 4 a 9 y la otra de 9 a 16 donde ya la diferencia en magnitud de esos valores es menor. Se obtiene entonces

$s_5 = -0.00003332643889 \cdot 10^{11}$
$s_6 = 0.00058167596596 \cdot 10^{11}$
$s_7 = -0.01018740617369 \cdot 10^{11}$
$s_8 = 0.17871353736177 \cdot 10^{11}$
$s_9 = -3.13795757673536 \cdot 10^{11}$

valores con los que reharíamos la primera columna de la tabla de arriba.

Se obtiene como conclusión que el valor propio de mayor valor absoluto λ_4 está en el intervalo $(-18.33856085776775, -17.61284167865894)$. La convergencia es muy lenta.

5. El lema es trivial. Pasando al límite en la relación entre los coeficientes y las raíces

$$a_n = \frac{1}{2}\left(\lambda_n + \sqrt{\lambda_n^2 - 4\mu_n}\right), \quad b_n = \frac{1}{2}\left(\lambda_n - \sqrt{\lambda_n^2 - 4\mu_n}\right)$$

se tiene por ejemplo

$$\lim_{n\to\infty} a_n = \frac{1}{2}\left(a + b + \sqrt{(a+b)^2 - 4ab}\right) = \frac{1}{2}\left(a + b + |a - b|\right)$$

Suponiendo que $a > b$, se tiene $|a-b| = a-b$ y $\lim_{n\to\infty} a_n = a$, de modo que $\lim_{n\to\infty} b_n = b$, coherente con la propiedad $a_n \geq b_n \ \ (\forall n)$.

6. Las tres sucesiones σ_n, σ_{n-1} y σ_{n-2} tienen el mismo límite λ_4 cuando $n \to \infty$.
   Si $\delta_n \to \delta$ cuando $n \to \infty$ se tiene llamando $\lambda_n = a_n + b_n$ y $\mu_n = a_n b_n$ como en el apartado anterior,

$$\lim_{n\to\infty} \lambda_n = \lim_{n\to\infty} (\sigma_n + \delta_n \sigma_{n-2}) = \lambda_4(1+\delta)$$
$$\lim_{n\to\infty} \mu_n = \lim_{n\to\infty} (\sigma_{n-1}\sigma_{n-2}\delta_n) = (\lambda_4)^2\delta$$

   luego

$$a + b = \lambda_4(1+\delta) \quad \text{y} \quad ab = (\lambda_4)^2\delta$$

   Es claro que si $\delta = \dfrac{\lambda_3}{\lambda_4}$ entonces $a = \lambda_3$ y $b = \lambda_4$. Veamos que en efecto $\lim_{n\to\infty} \delta_n = \dfrac{\lambda_3}{\lambda_4}$. Denotando por comodidad $p = -\lambda_4$, $q = -\lambda_3$, $r = -\lambda_2$ y $s = -\lambda_1$ los valores absolutos de las raíces se tiene

$$\sigma_n - \sigma_{n-1} = -\frac{p^n + q^n + r^n + s^n}{p^{n-1} + q^{n-1} + r^{n-1} + s^{n-1}} + \frac{p^{n-1} + q^{n-1} + r^{n-1} + s^{n-1}}{p^{n-2} + q^{n-2} + r^{n-2} + s^{n-2}}$$
$$\sigma_{n-1} - \sigma_{n-2} = -\frac{p^{n-1} + q^{n-1} + r^{n-1} + s^{n-1}}{p^{n-2} + q^{n-2} + r^{n-2} + s^{n-2}} + \frac{p^{n-2} + q^{n-2} + r^{n-2} + s^{n-2}}{p^{n-3} + q^{n-3} + r^{n-3} + s^{n-3}}$$

   Centrémosnos en la primera igualdad. Llamemos $A_i = p^i + q^i + r^i + s^i$. Con ello,

$$\sigma_n - \sigma_{n-1} = -\frac{A_n}{A_{n-1}} + \frac{A_{n-1}}{A_{n-2}} \quad \Rightarrow \quad A_{n-1}A_{n-2}(\sigma_n - \sigma_{n-1}) = -A_n A_{n-2} + A_{n-1}^2$$

   con

$$\begin{aligned}
&A_{n-1}A_{n-2} = A_{2n-3} + p^{n-1}\left(q^{n-2} + r^{n-2} + s^{n-2}\right) + q^{n-1}\left(p^{n-2} + r^{n-2} + s^{n-2}\right) + \\
&+ r^{n-1}\left(p^{n-2} + q^{n-2} + s^{n-2}\right) + s^{n-1}\left(p^{n-2} + q^{n-2} + r^{n-2}\right) = \\
&A_{2n-3} + (pq)^{n-2}\left[p\left(1 + \left(\frac{r}{q}\right)^{n-2} + \left(\frac{s}{q}\right)^{n-2}\right) + q\left(1 + \left(\frac{r}{p}\right)^{n-2} + \left(\frac{s}{p}\right)^{n-2}\right) + \right. \\
&\left. + r^{n-1}\left(\left(\frac{1}{q}\right)^{n-2} + \left(\frac{1}{p}\right)^{n-2} + + \left(\frac{s}{pq}\right)^{n-2}\right) + s^{n-1}\left(\left(\frac{1}{q}\right)^{n-2} + \left(\frac{1}{p}\right)^{n-2} + \left(\frac{r}{pq}\right)^{n-2}\right)\right]
\end{aligned}$$

   y cuando $n \to \infty$

$$A_{2n-3} = p^{2n-3}\left\{1 + \left(\frac{q}{p}\right)^{2n-3} + \left(\frac{r}{p}\right)^{2n-3} + \left(\frac{s}{p}\right)^{2n-3}\right\} \simeq p^{2n-3}$$

   y

$$\begin{aligned}
&(pq)^{n-2}\left[p\left(1 + \left(\frac{r}{q}\right)^{n-2} + \left(\frac{s}{q}\right)^{n-2}\right) + q\left(1 + \left(\frac{r}{p}\right)^{n-2} + \left(\frac{s}{p}\right)^{n-2}\right) + r^{n-1}\left(\left(\frac{1}{q}\right)^{n-2} + \right.\right. \\
&\left.\left. + \left(\frac{1}{p}\right)^{n-2} + + \left(\frac{s}{pq}\right)^{n-2}\right) + s^{n-1}\left(\left(\frac{1}{q}\right)^{n-2} + \left(\frac{1}{p}\right)^{n-2} + \left(\frac{r}{pq}\right)^{n-2}\right)\right] \simeq (pq)^{n-2}(p+q)
\end{aligned}$$

   luego

$$A_{n-1}A_{n-2} \simeq p^{2n-3} + (pq)^{n-2}(p+q) = p^{2n-3}\left[1 + \left(\frac{q}{p}\right)^{n-2} + \left(\frac{q}{p}\right)^{n-1}\right]$$

y por fin

$$A_{n-1}A_{n-2} \simeq p^{2n-3}$$

de un modo análogo se obtienen las equivalencias cuando $n \to \infty$

$$\begin{aligned} A_{n-1}^2 &=\simeq A_{2n-2} + (pq)^{n-2}2pq \\ A_nA_{n-2} &\simeq A_{2n-2} + (pq)^{n-2}(p^2+q^2) \end{aligned}$$

de donde

$$A_{n-1}^2 - A_nA_{n-2} \simeq (pq)^{n-2}(p-q)^2$$

Con ello,

$$\begin{aligned} p^{2n-3}\left(\sigma_n - \sigma_{n-1}\right) &\simeq -(p-q)^2(pq)^{n-2} \\ p^{2n-5}\left(\sigma_{n-1} - \sigma_{n-2}\right) &=\simeq -(p-q)^2(pq)^{n-3} \end{aligned}$$

luego

$$\frac{\sigma_n - \sigma_{n-1}}{\sigma_{n-1} - \sigma_{n-2}} \sim \frac{\dfrac{-(p-q)^2(pq)^{n-2}}{p^{2n-3}}}{\dfrac{-(p-q)^2(pq)^{n-3}}{p^{2n-5}}} = \frac{(pq)^{n-2}p^{2n-5}}{(pq)^{n-3}p^{2n-3}} = \frac{q}{p} = \frac{\lambda_3}{\lambda_4} = \delta$$

como queríamos probar.

Utilizando el programa *Leverrier3.m* que encontraréis en la página web vinculada al libro obtenemos los valores de ambas sucesiones a_n y b_n hasta $n = 16$.

n	a_n	b_n
1	0	0
2	0	0
3	2.12801186441248	-16.00776295480073
4	-6.94549148379272	-16.98707671102707
5	-6.69177405700441	-17.53007327848992
6	-7.01504951607498	-17.54767626544418
7	-7.49841657426211	-17.55907429606818
8	-8.32599018975802	-17.56861083440383
9	-9.73732929763942	-17.57854276673312
10	-11.76199178456966	-17.59113362286948
11	-13.90794616647291	-17.61026421732841
12	-15.52683825171721	-17.64270650317528
13	-16.44187109578596	-17.69493750912579
14	-16.86091521600477	-17.75827499814482
15	-17.03214123086371	-17.80891014698804
16	-17.10143637332802	-17.83805658643178

A partir de estos resultados se obtienen los valores aproximados $\lambda_3 \sim -17.10143637332802$ y $\lambda_4 \sim -17.83805658643178$ que como veremos más tarde son poco ajustados.

7. Una vez calculados los valores aproximados de λ_3 y λ_4 obtendremos los otros dos resolviendo la ecuación de segundo grado

$$\frac{P(\lambda)}{(\lambda - a_{16})(\lambda - b_{16})} = 0$$

En Matlab se utiliza la función *deconv* para dividir dos polinomios. El pequeño programa que hemos escrito efectúa la división pedida y calcula además las raíces del cociente.

```
p=[1 47.8884300000 797.278764779 5349.45551533 12296.5505661];
b=[1 34.9394929597598 305.0563897366879];
[q,r]=deconv(p,b);
r=roots(q);
```

Se obtienen como raíces de este trinomio los números -7.93243358470824 y -5.01650345553196 que identificamos como λ_2 y λ_1 respectivamente. Con todo ello, los valores propios aproximados de L son
$\lambda_1 = -5.01650345553196$
$\lambda_2 = -7.93243358470824$
$\lambda_3 = -17.10143637332802$
$\lambda_4 = -17.83805658643178$
que compararemos después con los obtenidos mediante los comandos de Matlab.

8. Repitiendo para la matriz L' los cálculos que hemos hecho para L obtenemos
$s_1 = -47.8884300000,\ s_2 = 696.953299294,\ s_3 = -11275.063304118,\ s_4 = 191141.3779865$
$c_1 = 47.8884300000,\ c_2 = 798.1742142855,\ c_3 = 5374.1246687440,\ c_4 = 12467.9856877910$
y a partir de estos valores obtenemos la tabla

n	s_n	a_n	b_n
1	-47.888430	0	0
2	696.953299294	0	0
3	-11275.063304118	2.12597547609799	-15.97065724283309
4	191141.3779865	-6.94483580347031	-16.95257692404168
5	$-3.302437365\cdot10^{6}$	-6.57640582586358	-17.48336927419297
6	$5.748841346\cdot10^{7}$	-6.72207617298993	-17.49092713551179
7	$-1.003749782\cdot10^{9}$	-6.87644140785800	-17.49409528927410
8	$1.754679407\cdot10^{10}$	-7.06290553775708	-17.49564939568054
9	$-3.068963850\cdot10^{11}$	-7.34186687655928	-17.49663126262993
10	$5.368899\cdot10^{12}$	-7.84498457472685	-17.49743468632786
11	$-9.3935295\cdot10^{13}$	-8.80215707657165	-17.49823521316212
12	$1.643623129\cdot10^{15}$	-10.44232246325441	-17.49917943861433
13	$-2.8760554953\cdot10^{16}$	-12.62897310203212	-17.50052130968592
14	$5.03280850451\cdot10^{17}$	-14.69585031329833	-17.50282646494417
15	$-8.807248114326\cdot10^{18}$	-16.09579950577157	-17.50737859734548
16	$1.541296256565\cdot10^{20}$	-16.83915348893440	-17.51682154923292

Se obtiene $\lambda'_3 \sim -16.839153489$ y $\lambda'_4 \sim -17.516821549$. Comparándolos con los valores $\lambda'_3 = -17.385457$ y $\lambda'_4 = -17.596207$ que obtuvimos en el apartado 1. se ve lo poco ajustado del método que estamos usando.

Si hallamos los otros dos valores propios por el método del apartado anterior tendremos
$$\lambda'_2 = -9.50424713345013 \text{ y } \lambda'_1 = -4.02820782838255.$$

Podemos hallar la distancia entre ambos conjuntos. Se llega a $\|\Lambda' - \Lambda\| = \max_i |\lambda'_i - \lambda_i| = 2.10736313345013$. Se produce el máximo en $|\lambda'_2 - \lambda_2|$. La distancia entre los valores propios de mayor absoluto $|\lambda'_4 - \lambda_4| = 0.0793854507670$ es razonable no así las demás.

Cuestiones complementarias

2 Utilizando el programa *Leverrier0.m*

```
L=[-5.509882  1.870086   0.422908     0.008814;
    0.287865 -11.811654  5.711900   0.058717 ;
    0.049099  4.308033   -12.970687  0.229326;
    0.006235  0.269851   1.397369  -17.596207];
X=eig(L)
```

```
[V,D]=eig(L)
p=poly(L)
r=roots(p)
[A,Z]=balance(L)
T=polyeig(Z)
```

obtenemos todos los resultados relevantes del ejercicio.

2.1 El comando *eig* da como resultado los valores y vectores propios de una matriz cuadrada.

- E = *eig*(L) es un vector que contiene los valores propios de L.
- [V,D] = *eig*(L) produce una matriz diagonal D con los valores propios y una matriz V cuyas columnas son los correspondientes vectores propios de modo que L*V = V*D.

Se obtiene aquí

$$V = \begin{pmatrix} 0.993206519261 & 0.639811277546 & 0.048059102167 & -0.023454814208 \\ 0.0981483336120 & -0.598336985014 & -0.381329657937 & 0.200407254049 \\ 0.061758485688 & -0.475303246154 & 0.345329749111 & -0.220952922090 \\ 0.009674923814 & -0.081982959719 & 0.856169013137 & 0.954183740396 \end{pmatrix}$$

$$D = \text{diag}\,(-5.298698068962\ -7.574043430621\ -17.152427162920 - 17.863261337497)$$

Obtenemos así los valores propios de L y las rectas de vectores propios asociadas, que están generadas por los vectores columna de V.

Hemos representado gráficamente el polinomio característico de L que obtuvimos en el apartado 4.2. (ver Figura 2.6). Se hallan sus raíces usando la función *roots*; para ello se introduce directa-

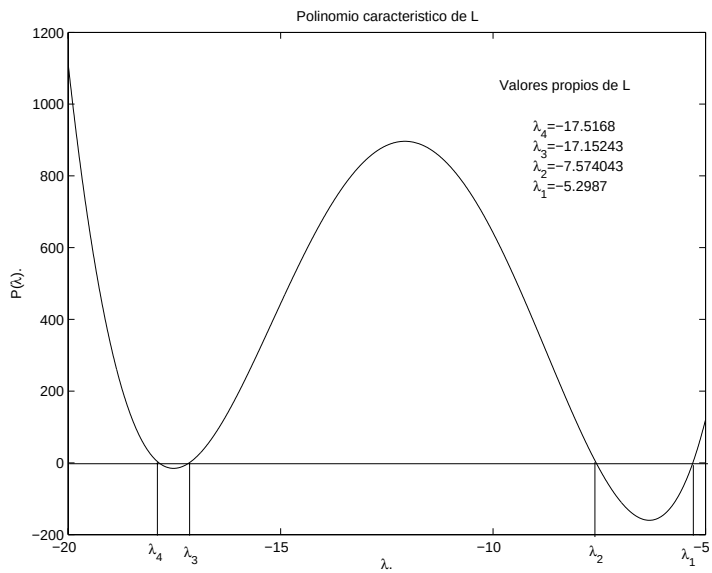


Figura 2.6: Representación gráfica del polinomio característico de la matriz L.

mente en la ventana de comandos el vector fila de los coeficientes de dicho polinomio, ordenado por potencias decrecientes, y la función *roots*

```
>> p=[1 47.8884300000 797.278764779 5349.45551533 12296.5505661];
>> r=roots(p)
```

se obtiene

```
r= -17.86326133806126
   -17.15242716239645
    -7.57404343044235
    -5.29869806909996
```

Matlab considera por convenio las raíces del polinomio como un vector columna y las ordena por valores crecientes. Comparado este vector de valores propios con el que hemos obtenido antes midiendo su distancia mediante la norma del máximo de la diferencia se obtiene $5.6501 \cdot 10^{-10}$.

Se comprende que la pérdida de precisión del método objeto del enunciado se produce en el cálculo de los valores propios. Comparando los valores aproximados obtenidos en 7. con cualquiera de los anteriores usando la misma distancia tenemos $3.58390 \cdot 10^{-1}$ que se produce en λ_2, siendo $2.5204 \cdot 10^{-2}$ la distancia en λ_4.

2.2 Hay un comando *balance*(L) que hace un escalado diagonal de la matriz L que mejora la precisión en el cálulo de los valores propios. La matriz que se obtiene es semejante a L. La matriz T de la transformación se busca de modo que *balance*$(L) = T^{-1}LT$ tenga las normas de sus filas y de sus columnas aproximadamente iguales, lógicamente no se usa con matrices simétricas que ya cumplen esa propiedad.

T es una permutación de una matriz diagonal cuyos elementos son potencias de dos de modo que el escalado no introduzca errores de redondeo. En nuestro caso se tiene

$$T = \begin{pmatrix} 2 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 2 \end{pmatrix}$$

y

$$balance(L) = \begin{pmatrix} -5.509882000 & 0.935043000 & 0.211454000 & 0.008814000 \\ 0.575730000 & -11.811654000 & 5.711900000 & 0.117434000 \\ 0.098198000 & 4.308033000 & -12.970687000 & 0.458652000 \\ 0.006235000 & 0.134925500 & 0.698684500 & -17.596207000 \end{pmatrix}$$

Los valores propios obtenidos con esta matriz escalada se diferencian de los antes calculados en menos de $1 \cdot 10^{-14}$.

Como vemos, todas las posibles alternativas ofrecidas por el código en este problema son de enorme eficacia.

Desde luego que Leverrier hubiera estado encantado con Matlab aunque es seguro que no existiría ningún método que llevara su nombre para el cáculo del polinomio característico de una matriz.

3. Calculamos las tres normas matriciales inducidas

- *norm*(L,1), la norma-1 de L, se tiene

$$\|L\|_1 = \max_{1 \le i \le 4} \left\{ \sum_{j=1}^{4} |\, l_{ji}| \right\} = 20.50286400000000$$

- *norm*(L,inf), la norma $\| \ \|_\infty$

$$\|L\|_\infty = \max_{1 \le i \le 4} \left\{ \sum_{j=1}^{4} |\, l_{ij}| \right\} = 19.26966200000000$$

- *norm*(L), la norma-2 de L, el mayor valor singular de la matriz

$$\|L\|_2 = 18.08187998940853$$

Como podemos comprobar, la desigualdad del enunciado se cumple de modo evidente.

CAPÍTULO 3

Interpolación lineal

Los métodos de interpolación lineal proporcionan la posibilidad de reconstruir funciones que simulen de modo exacto el comportamiento de un sistema bajo ciertas condiciones y que lo hagan, aunque no de modo exacto, sí razonablemente bien, para condiciones más generales. La idea es buscar dentro de un espacio vectorial de funciones aquella que verifique una serie de propiedades. En interpolación lineal, encontrarla conducirá al planteamiento de un sistema lineal, una vez que se escribe la hipotética solución en una base de dicho espacio vectorial.

En algunos manuales, la interpolación lineal y la aproximación de funciones se estudian juntas dado que el caso más importante de la interpolación (el problema de Lagrange) puede verse como un caso particular de aquélla. En este texto creemos que es más enriquecedor el considerarlas de modo separado dado el enfoque algebraico con el que afrontamos los problemas de interpolación lineal. De hecho, es conveniente tener una base previa en álgebra lineal en los aspectos relativos a espacios vectoriales y dualidad.

Utilizaremos básicamente técnicas de interpolación polinomial y haremos un uso exhaustivo de la interpolación polinomial a trozos y de su caso particular más interesante, el de los splines, de uso muy extendido en ingeniería. De hecho, los splines constituyen la base de los programas de diseño asistido o CAD.

Como referencias recomendadas para este capítulo tenemos el texto de Hammerlin et al. [15], que cubre perfectamente la parte de los splines; el libro de Linz [20], muy orientado a los aspectos generales de la teoría general de interpolación lineal y que contiene demostraciones de algunos teoremas aquí sólo enunciados y el libro de Sanz-Serna [24], una referencia interesante para la parte de interpolación polinomial de Lagrange.

3.1. El problema general de interpolación

En los cursos de Cálculo Numérico, la interpolación de funciones se suele estudiar tomando como elemento motivador el problema del polinomio interpolador de Lagrange. La idea de este problema es reconstruir una función de la que sabemos su valor en una serie de puntos, mediante un polinomio que pase por todos esos puntos. El problema tiene solución única si ajustamos el grado del polinomio del modo adecuado suponiendo que todos los puntos tienen diferentes abscisas (luego veremos este caso con todo detalle). Creemos que a efectos de tener una perspectiva más global del problema, es conveniente enfocarlo de un modo más general, asumiendo que el lector tiene una base de álgebra lineal suficiente.

En este sentido, ya que la función solución de nuestro problema es un elemento de un espacio vectorial real de funciones E de dimensión finita n, la elección de una u otra base en E para expresar esa solución condiciona que la resolución del problema sea más o menos difícil.

En efecto, una vez escrita la función buscada como combinación lineal de los elementos de una base de E, encontrar los coeficientes de esa expansión lineal conduce a la resolución de un sistema lineal[1] y, dependiendo de la base escogida, ese sistema será más o menos fácil de resolver. Un concepto abstracto de gran importancia en este problema es el de espacio dual.

[1] Si no es así, el problema de interpolación no será lineal. En ese caso, su resolución requiere encontrar la solución a un sistema de ecuaciones no lineales, para lo cual se usarán las técnicas estudiadas en el Capítulo 1, como veremos en algún ejemplo y sobre todo en el problema 3.9.

Definición 3.1.1 *Las aplicaciones lineales definidas en E con valores en $\mathbb{R}$ estructurado como $\mathbb{R}$-espacio vectorial se llaman formas lineales en E.*

Definición 3.1.2 *El espacio dual E^* de un espacio vectorial real E es el espacio de las formas lineales definidas en E.*

Ejemplo 3.1.1 *Sea $E = \mathbb{R}^3$. Se define la función f*

$$f(x, y, z) = 2y, \quad \forall (x, y, z) \in \mathbb{R}^3$$

Es fácil ver que f es una forma lineal, por ser una aplicación lineal entre $\mathbb{R}^3$ y $\mathbb{R}$. f es un elemento de $(\mathbb{R}^3)^*$.

Ejercicio 3.1.1 *Sea $E = P_2(\mathbb{R})$, espacio vectorial de los polinomios de segundo grado de coeficientes reales. Se define la función F*

$$\begin{array}{rccc} F: & P_2(\mathbb{R}) & \to & \mathbb{R} \\ & p & \mapsto & p(2.37) \end{array}$$

O sea, $F(p)$ es el valor que toma el polinomio para la abscisa $x = 2.37$. Se pide demostrar que $F \in P_2(\mathbb{R})^*$.

Se demuestra que en dimensión finita las dimensiones de E y E^* coinciden y que se puede definir un isomorfismo canónico entre E y su bidual $E^{**} = (E^*)^*$ que permite identificarlos, de modo que $(E^*)^* = E$ y ambos espacios E y E^* se pueden considerar como las dos caras de una misma moneda. Dada una base en uno cualquiera de ellos, se define una base del dual, *la base dual*, mediante las relaciones de dualidad, o sea, cada forma lineal de la base dual toma valor uno en el elemento correspondiente (de igual índice) de la base dada y cero en los demás. Veremos innumerables ejemplos de parejas de bases duales en los problemas de este capítulo.

El concepto de espacio dual es muy interesante y poderoso, aunque difícil para el estudiante. Está íntimamente ligado con el de interpolación lineal a través de la idea *qué información de la función es necesario conocer para que ésta esté completamente definida*. Si se encuentran muchas dificultades en comprender lo que sigue, se puede pasar directamente a la sección 3.2 con promesa de volver de nuevo a esta sección al final del mismo, ya que esta perspectiva enriquece mucho el problema de interpolación.

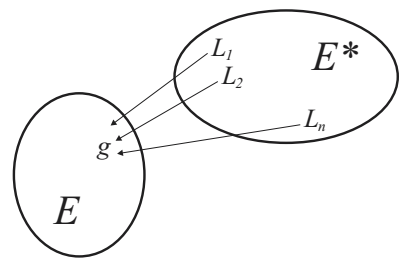


Figura 3.1: Problema de interpolación.

Describamos de modo formal el problema:

Definición 3.1.3 *Dados*

- *un $\mathbf{R}$-e.v. de funciones E de dimensión finita n de dual E^*,*
- *una familia $L_i \in E^*$, $i = 1, n$ de n formas lineales en E,*
- *y una familia w_i, $i = 1, n$ de n números reales,*

el problema general de interpolación consiste en determinar un vector $g \in E$ tal que:

$$L_i(g) = w_i \quad i = 1, ..., n \tag{3.1}$$

En el problema de interpolación se debe encontrar una función $g \in E$ que verifique ciertas propiedades que se pueden expresar obligando a que una serie de formas lineales de E^*, bien elegidas, $L_i = 1, n$ tomen en g valores preasignados.

Ejemplo 3.1.2 *Sea $E = P_2(\mathbb{R})$, espacio vectorial de los polinomios de segundo grado de coeficientes reales. Se definen las tres formas lineales sobre $P_2(\mathbb{R})$ siguientes:*

$$L_1(p) = p(2.37), \quad L_2(p) = p(3.12), \quad L_3(p) = p(3.76)$$

Se definen los tres números reales siguientes: $w_1 = 0.13$, $w_2 = 0.27$, $w_3 = -0.04$. El problema de interpolación se plantea cómo encontrar un elemento $p \in P_2(\mathbb{R})$ tal que $L_i(p) = w_i$, $i = 1, 3$. El resultado es el polinomio p que podemos ver en la Figura 3.2, generada con las siguientes líneas Matlab, cuyo cálculo comentamos más adelante:

```
x=[2.37 3.12 3.76];
y=[0.13 0.27 -0.04];
p=polyfit(x,y,2);
t=2:0.01:4;
pdet=polyval(p,t);
plot(x,y,'x',t,pdet);
```

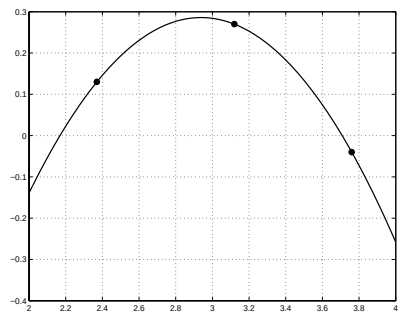


Figura 3.2: Solución correspondiente al ejemplo 3.1.2.

Estudiemos ahora cuándo el problema general de interpolación tiene solución y si esa solución es única o no. Para ello, enunciemos primero el lema:

Lema 3.1.1 *Sean E un $\mathbf{R}$-e.v. de dimensión n y $B = \{e_i\}_{i=1,n}$, una de sus bases. El conjunto de n formas lineales $(L_i)_{i=1,n}$ es linealmente independiente en E^* ssí*

$$\det\left(L_i(e_j)\right) \neq 0$$

Teorema 3.1.1 *El problema de interpolación general (3.1) tiene solución única si y sólo si las n formas lineales (L_i) son linealmente independientes en E^*.*

Demostración

En efecto, sea $B = \{e_j\}$, $j = 1, ..., n$ una base de E y escribamos la solución buscada g en esa base

$$g = \sum_{j=1}^{n} c_j e_j$$

Imponiendo a g las condiciones del problema

$$w_i = L_i(g) = L_i\left(\sum_{j=1}^{n} c_j e_j\right) = \sum_{j=1}^{n} c_j L_i\left(e_j\right), \quad i = 1, n$$

de donde obtenemos el sistema lineal:

$$\begin{pmatrix} L_1(e_1) & \dots & L_1(e_n) \\ \vdots & \ddots & \vdots \\ L_n(e_1) & \dots & L_n(e_n) \end{pmatrix} \begin{pmatrix} c_1 \\ \vdots \\ c_n \end{pmatrix} = \begin{pmatrix} w_1 \\ \vdots \\ w_n \end{pmatrix} \tag{3.2}$$

La matriz del sistema es la misma que en el lema anterior. El sistema lineal (3.2) tendrá solución única ssi el determinante de la matriz es distinto de cero. Por el lema 3.1.1 esto será así ssí las formas lineales son linealmente independientes, lo que cierra la demostración.

La demostración anterior es constructiva y suministra una forma de resolver el problema. Eligiendo una base cualquiera de E, la resolución del sistema (3.2) suministra las componentes únicas de la solución respecto de esa base.

Ejemplo 3.1.3 *Polinomio interpolador de Lagrange.*

Se trata de encontrar un polinomio $P \in P_n(\mathbb{R})$ que tome en $n+1$ puntos distintos $x_0, \ldots, x_n$ de $\mathbb{R}$, los $n+1$ valores $y_0, \ldots, y_n$ previamente asignados. Gráficamente se interpreta como la busca de un polinomio cuya gráfica pase por los $n+1$ puntos $(x_i, y_i)_{i=0,n}$.

El primer paso para hallar la solución es seleccionar una base de $E = P_n(\mathbb{R})$. Si elegimos la base canónica $\{x^i \quad i = 0, 1, ..., n\}$ a la que nos referiremos en adelante como el sistema de monomios, tendremos

$$P(x) = \sum_{k=0}^{n} c_k x^k$$

Si ahora imponemos las condiciones

$$P(x_i) = \sum_{k=0}^{n} c_k x_i^k = y_i, \quad i = 1, n$$

y llegamos al sistema lineal

$$\begin{pmatrix} 1 & x_0 & x_0^2 & \ldots & x_0^n \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & \ldots & x_n^n \end{pmatrix} \begin{pmatrix} c_0 \\ \vdots \\ c_n \end{pmatrix} = \begin{pmatrix} y_0 \\ \vdots \\ y_n \end{pmatrix}$$

La matriz del sistema es la de Vandermonde, cuyo determinante es distinto de cero si las abscisas de los puntos son distintas entre sí. Por tanto, el problema de interpolación de Lagrange[2] *propuesto tiene solución única en el espacio de los polinomios de grado n. El interpolador de Lagrange se puede calcular utilizando correctamente la función* polyfit *de Matlab.*

[2]Lagrange, Joseph-Louis, 1736-1813. El siglo XVIII, el de la Ilustración, el siglo del Despotismo Ilustrado de *todo para el pueblo pero sin el pueblo.* El siglo por excelencia de los franceses (de Diderot, Voltaire, Rousseau, Montesquieu, Condorcet, D'Alembert...). El siglo de Carlos III en España, Federico II el Grande en Prusia, José II en Austria, Catalina II la Grande en Rusia; con sus ostentosas cortes rivalizando en *glamour*, con sus espléndidos bailes en engalanados salones, sus banquetes pantagruélicos, sus tertulias eruditas moderadas en algunos casos por el propio monarca. El Siglo de las Luces, con la defensa encendida del conocimiento analítico e inductivo frente al oscuro sistema metafísico de los siglos anteriores; el triunfo definitivo de la razón frente a la fe, la superstición, la magia. El siglo que tenía que desembocar por ley en las revoluciones americana y francesa.

También el siglo en el que las cortes más poderosas de Europa se rifaban a los más grandes artistas, filósofos y científicos..., entre los que solía haber un matemático destacado, pues éstos no sólo podían desempeñar importantes funciones dentro del Estado en cargos directos o de asesoramiento, sino que su sola presencia derramaba lustre sobre el monarca que los adoptaba. Como en el caso de Joseph-Louis Lagrange.

El superdotado Lagrange nace en Turín en el seno de una familia aristocrática venida a menos: *Si hubiera heredado una fortuna, probablemente no me habría dedicado a las matemáticas.* A los 19 años (profesor ya en la escuela de artillería) alcanza su primer logro científico: ser el cofundador, junto a Euler, de una nueva rama de las matemáticas, el cálculo de variaciones, mediante un impecable procedimiento analítico (el turinés defendió durante toda su vida que las matemáticas no necesitaban de las intuiciones geométricas y que todos los problemas tenían que ser atacados desde el análisis puro). Lagrange, ferviente y devoto admirador-seguidor de Euler, conseguía el respeto y el reconocimiento de su maestro y de toda la ilustre comunidad matemática de la época. Isaac Newton podía descansar tranquilo en su tumba: su *reinvención* del mundo tenía unos geniales continuadores.

El rey más grande de Europa debe tener en su corte al matemático más grande. Con esta frase *made in* Federico, el rey prusiano da la bienvenida a Lagrange como flamante director de la Academia de las Ciencias de Berlín, sucediendo en el cargo a Euler, caído en desgracia y desplazado a la corte de Catalina la Grande (después de 25 años de servicios). Lagrange, que sólo tiene 30 años, podía ser el hijo de más de un conspicuo matemático de la Academia. Su natural modestia y sobriedad en el trato, su intuitiva prudencia envasada en una elegancia formal impecable, sus vastos conocimientos en disciplinas como la filosofía, muy pronto le granjearon el respeto y el cariño de sus colegas y el aprecio del monarca. Su producción científica fue ciclópea: más de ciento cincuenta memorias que profundizaban decisiva e innovadoramente en todas las ramas de las matemáticas conocidas hasta entonces. Ganó las cinco veces que se presentó el concurso bianual que organizaba su amigo D'Alembert en la Academia de las Ciencias de París resolviendo problemas de mecánica celeste. Era considerado por sus colegas ni más ni menos como el matemático (siempre junto a Euler) más relevante de un siglo que la Historia, con el tiempo, se encargaría de dictaminar como el más fructífero para la ciencia matemática.

No sé lo que haré en matemáticas dentro de diez años o *la mina matemática es ya demasiado profunda y a no ser que se descubran nuevas vetas tendrá que ser abandonada.* Cuando Lagrange escribe esto tiene 45 años y está en un profundo estado

En el problema 3.1 tenemos un ejemplo en el que las funciones de base son trigonométricas y en el problema 3.2 tenemos un ejemplo sin solución única o con infinitas soluciones.

3.2. Interpolación polinomial

En los problemas de interpolación polinomial, el espacio E donde buscamos la solución al problema de interpolación es un espacio vectorial de polinomios o de polinomios a trozos. La interpolación polinomial es el caso más interesante y el único que estudiaremos con detalle.

3.2.1. Interpolación de Lagrange

Ya hemos introducido este problema en el ejemplo 3.1.3, viendo cómo se resuelve cuando buscamos su solución en la base de los monomios, ahora lo estudiaremos con más detalle.

Se dan $(n+1)$ abscisas distintas (nodos) x_i y $(n+1)$ valores reales y_i, $(i = 0, ..., n)$. Se busca un polinomio P de grado n tal que

$$P(x_i) = y_i, \quad i = 0, ..., n$$

A menudo los datos y_i son las imágenes de una cierta funcion f, función que se pretende interpolar, en los nodos x_i.

Polinomios de Lagrange

Como vimos en el ejemplo 3.1.3, el problema de interpolación de Lagrange tiene solución única. Esto quiere decir que las $n + 1$ formas lineales $L_i(P) := P(x_i)$, $(i = 0, n)$ son linealmente independientes y constituyen por tanto una base B^* de $P_n(\mathbb{R})^*$. Su base dual B en $P_n(\mathbb{R})$ está formada por los polinomios $\{l_j\}_{i=0,...,n}$, llamados de Lagrange, asociados a los puntos $x_0, \ldots, x_n$ y que están definidos por las relaciones $L_i(l_j) = \delta_{ij}$, es decir,

$$l_j(x_i) = \begin{cases} 0 & \text{si} \quad i \neq j \\ 1 & \text{si} \quad i = j \end{cases}$$

luego l_j tiene n raíces x_i $(i \neq j)$ y toma el valor 1 cuando $x = x_j$.

Si lo factorizamos con sus raíces, tendremos que

$$l_j(x) = \prod_{\substack{i=0 \\ i \neq j}}^{n} (x - x_i) \tag{3.3}$$

depresivo. El 17 de marzo de 1783 muere Daniel Bernoulli. El 18 de marzo fallece repentinamente Euler. D'Alambert el 29 de octubre. Todos maestros y amigos. De ser el benjamín de una generación se había convertido en el decano de la siguiente (Laplace, Legendre, Monge, Gauss). En agosto de ese mismo año muere su amada esposa, de la que se había enamorado después de casarse. El 17 de agosto de 1786 muere su protector, Federico el Grande, sucedido por un Federico Guillermo II que no contaba ni por asomo con el talante ilustrado de su tío. En 1787 llegó la hora para un cansado Lagrange de hacer la mudanza, después de 21 años en Berlín. El destino elegido, entre el ramillete selecto de cortes que se le pusieron a sus pies (por ejemplo la española), fue París.

En su primera etapa francesa se dedica al estudio exhaustivo de la química, medicina y filosofía, como si las matemáticas hubieran dejado de interesarle (aunque en esta época se publicará por primera vez su colosal y decisiva *Mecánica analítica*, redactada en Berlín).

Yo creo que, en general, uno de los primeros principios que debe tener todo hombre prudente es el de conformar su conducta estrictamente a las leyes del país que vive, incluso cuando éstas no le parecen razonables. Gracias a esta filosofía de vida salvó con éxito el escollo de la Revolución Francesa (con la que no comulgaba), a diferencia de otros ilustres científicos que perdieron la cabeza en la guillotina por tomar excesivo partido en uno u otro bando (Lavoisier, Condorcet...). En esto emerge la figura de Napoleón: *El progreso y el perfeccionamiento de las matemáticas están íntimamente ligadas a la prosperidad del Estado*; o *Lagrange es la excelsa pirámide de las ciencias matemáticas.* En el año 1808 el emperador le concede al venerado matemático el título de conde del Imperio.

Los éxitos y reconocimientos nunca se le subirán a la cabeza, y, como su mentor y amigo Euler, hará gala de esa rara virtud que escasos genios poseen: reconocer y aplaudir el talento ajeno. En una carta fechada en mayo de 1804, tras leer sus *Disquisitiones*, le escribe a Gauss: *Vuestras Disquisitiones os han elevado de golpe al primer rango de los matemáticos (...) Creed, señor, que nadie aplaude vuestro éxito más sinceramente que yo.*

El 10 de abril de 1813, a la edad de 77 años, Lagrange, con el espíritu calmo, muere en París. Por delante todo el XIX, en la que la tan denostada Geometría se convierte en el motor de unas renovadas matemáticas. El prudente Lagrange, sin duda, esbozará una sonrisa irónica en su mausoleo.

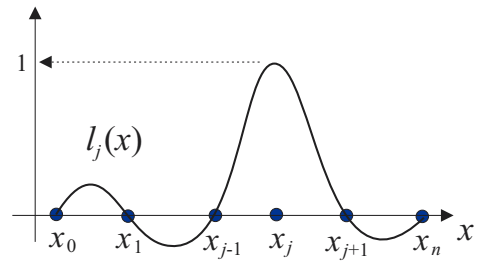


Figura 3.3: Esquema del elemento $l_j(x)$ de la base de Lagrange.

Si queremos normalizarlo para que tenga valor 1 en $x = x_j$, tendremos que dividir el anterior por su valor en $x = x_j$

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

En la Figura 3.3 representamos la gráfica de un elemento de la base de Lagrange.

Descomponiendo el polinomio de interpolación solución P respecto de esta base

$$P = \sum_{j=0}^{n} c_j l_j$$

de donde para $i = 0, 1, \ldots, n$

$$y_i = P(x_i) = \sum_{j=0}^{n} c_j l_j(x_i) = \sum_{j=0}^{n} c_j \delta_{ij} = c_i$$

por tanto, las componentes del polinomio solución en la base de Lagrange son directamente las imágenes que ese polinomio toma en las abscisas $\{x_i\} i = 0, ..., n$, es decir,

$$P = \sum_{j=0}^{n} y_j l_j \quad \Rightarrow \quad P(x) = \sum_{j=0}^{n} y_j l_j(x) \tag{3.5}$$

Esta idea es fundamental para entender la teoría de interpolación. Cada uno de los elementos de la base de Lagrange es nulo en todos los nodos menos en su propio nodo. Por tanto, multiplicar este elemento por cualquier factor *sólo* afecta al polinomio P en su valor en ese nodo y no en los demás, aunque sí en cualquier otro punto que no sea nodo.

Si quisiésemos construir un editor gráfico que jugara con el valor del polinomio de interpolación en los nodos $(x_i)_{i=0,n}$, elegiríamos la base de Lagrange, pues en esta base la escritura del polinomio de interpolación es directa a partir de los valores y_i en los nodos.

Ejercicio 3.2.1 *Calcular los elementos de la base de Lagrange para $x_0 = -2.1$, $x_1 = -1.1$ y $x_2 = 1.4$. Hacer una gráfica de los mismos. Calcular la parábola que pasa por los puntos $(-2.1, -0.3)$, $(-1.1, 0.2)$ y $(1.4, -1.0)$ en la base de Lagrange y en la de los monomios, desarrollando la primera para comprobar que coinciden.*

3.2.2. Estimaciones del error en la interpolación de Lagrange

En general, los valores y_i son las imágenes en los puntos x_i de una cierta función f. Es importante saber cuál es la separación entre el polinomio interpolante P y f. Es fácil ver que sólo se puede mayorar ese error si suponemos que f verifica determinadas propiedades y/o pertenece a un tipo restringido de funciones. En concreto, exigiremos que las funciones sean de clase $\mathcal{C}^{n+1}$ en un compacto que contenga a todos los nodos y al punto donde evaluamos el error. Presentamos aquí el teorema correspondiente y nos remitimos a la referencia [24] para su demostración.

Teorema 3.2.1 *Sea $f \in \mathcal{C}^{n+1}$ y sea P el polinomio de $P_n(\mathbb{R})$ interpolador de Lagrange de f en la nube $x_0, \cdots, x_n$ con $x_i \in [a,b]$, $i = 0, ..., n$. Entonces a cada punto x en $[a,b]$ le corresponde un punto $\xi(x)$ con $\min(x_0, \cdots, x_n, x) < \xi(x) < \max(x_0, \cdots, x_n, x)$ para el que*

$$f(x) - P(x) = \frac{H(x)}{(n+1)!} f^{(n+1)}(\xi(x)) \tag{3.6}$$

con

$$H(x) = \prod_{i=0}^{n} (x - x_i) \tag{3.7}$$

Corolario 3.2.1 *Veamos algunas estimaciones del error. Utilizando la norma del máximo tendremos*

$$|f(x) - P(x)| \leq \frac{|H(x)|}{(n+1)!} \|f^{(n+1)}\|_\infty \tag{3.8}$$

de la que se deduce la fórmula general de estimación del error en interpolación

$$\|f - P\| \leq \frac{\|f^{(n+1)}\|_\infty}{(n+1)!} \|H\| \tag{3.9}$$

válida para todas las normas funcionales $\|\,.\,\|_p$, $1 \leq p \leq n$ y en particular para la $\|\quad\|_\infty$, que es la más interesante en interpolación.

Ejemplo 3.2.1 *Sea el intervalo $[a,b] = [-\pi, \pi]$. Se considera la función $f(t) = \sin(10t)$. Calculemos la norma del máximo de f y sus derivadas:*

$$\begin{aligned} \|f\|_\infty &:= \max_{x \in [-\pi,\pi]} |\sin(10x)| = 1 \\ \|f'\|_\infty &:= \max_{x \in [-\pi,\pi]} |10\cos(10x)| = 10 \\ \|f^{(n)}\|_\infty &:= 10^n \end{aligned}$$

Corolario 3.2.2 *Si queremos encontrar una cota global al error en un intervalo, o sea, en el caso en que $\|\quad\| = \|\quad\|_\infty$, y si nos restringimos $x \in [x_0, x_n]$, se puede deducir una estimación muy útil del error de interpolación de $\|f - P\|_\infty$. Para ello utilizamos la mayoración*

$$\|H\|_\infty = \max_{x \in [x_0, x_n]} |(x - x_0)(x - x_1) \cdots (x - x_n)| \leq \frac{n!}{4} h^{n+1}$$

donde h es la máxima de las distancias entre puntos x_i y x_{i+1} adyacentes. Con ello

$$\|f - P\|_\infty \leq \frac{\|f^{(n+1)}\|_\infty}{4(n+1)} h^{n+1} \tag{3.10}$$

Si mantenemos el número de puntos pero los cogemos cada vez más cerca e interpolamos, el error tiende a cero con h^{n+1}; por tanto, el orden de método es $O(h^{n+1})$. Esto no significa que fijada la zona de interpolación, el error disminuya por utilizar más puntos; eso depende también de las derivadas sucesivas de la función f las cuales pueden crecer en valor absoluto, como ya hemos visto en el ejemplo 3.2.1 y veremos en otro ejemplo más tarde.

Ejemplo 3.2.2 *Se dispone de una tabla que proporciona la raíz cuadrada de los números enteros. Se trata de utilizar esta tabla para estimar con interpolación lineal el valor de la raíz de 19.58, dando una cota del error asociado.*

Por un lado tenemos que

$$\sqrt{19} = 4.3589 \qquad y \qquad \sqrt{20} = 4.4721$$

La recta que interpola estos dos puntos es

$$y - \sqrt{19} = \frac{\sqrt{20} - \sqrt{19}}{20 - 19}(x - 19) \quad \Rightarrow \quad y = 4.3589 + 0.1132(x - 19)$$

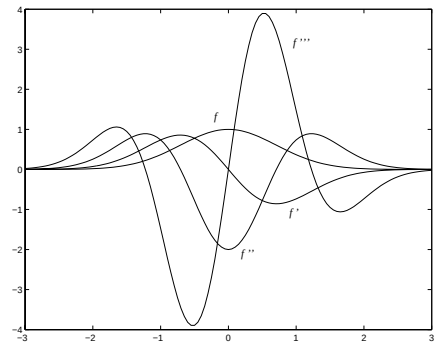


Figura 3.4: Función $f(x) := e^{-x^2}$ y derivadas sucesivas.

Particularizada en el punto en cuestión nos da el valor estimado

$$y = 4.3589 + 0.1132(19.58 - 19) = 4.4246$$

El error cometido es, por tanto, $|\sqrt{19.58} - 4.4246| = 0.00033$*, que no puede superar la acotación del error 3.10*

$$\|\sqrt{x} - y\|_\infty \leq \frac{\|f''\|_\infty}{8} 1^2 = \frac{\left\| \frac{-1}{4x\sqrt{x}} \right\|_\infty}{8} = \frac{max_{x\in[19,20]} \left| \frac{-1}{4x\sqrt{x}} \right|}{8} = \frac{1}{32 \cdot 19\sqrt{19}} = 0.00038$$

como así sucede. El máximo para la norma del máximo se alcanza entre 19 y 20 con $n = 1$*.*

Ejercicio 3.2.2 *Mejorar esta acotación del error utilizando la expresión 3.8.*

Ejercicio 3.2.3 *Acotar el error cuando utilizamos la recta obtenida en el ejemplo 3.2.2 para extrapolar, empleando esa fórmula para estimar la raíz de 24.37.*

Como hemos comentado antes, vamos a estudiar un ejemplo de interpolación de Lagrange en el que el error no disminuye cuando aumentamos el número de puntos, debido a que aún cuando el término $|H(x)|$ en 3.6 es más pequeño al utilizar más puntos, no sucede lo mismo con las derivadas sucesivas. Si consideramos, por ejemplo, $f(x) := e^{-x^2}$, y observamos sus derivadas en la Figura 3.4, vemos que el máximo de estas derivadas crece con el orden de derivación. Esta figura la hemos obtenido con el siguiente código Matlab en el que hacemos uso de sus posibilidades de cálculo simbólico.

```
%%%%% emx2yderivadas.m
syms x
f=exp(-x^2)
f1=diff(f) % deriva f y asigna esa derivada a f1
f2=diff(f1) % deriva f1 y asigna esa derivada a f2
f3=diff(f2) % etc.
xx=-3:0.05:3;
for i=1:length(xx),
    x=xx(i);
    ff(i)=eval(f);
    ff1(i)=eval(f1);
    ff2(i)=eval(f2);
    ff3(i)=eval(f3);
end
plot(xx,ff,'k',xx,ff1,'k',xx,ff2,'k',xx,ff3,'k');
shg;
```

Si ahora construimos los polinomios de interpolación de diferentes grados de la función f y los representamos superpuestos a la función en la Figura 3.5, vemos que no hay convergencia hacia la función f a

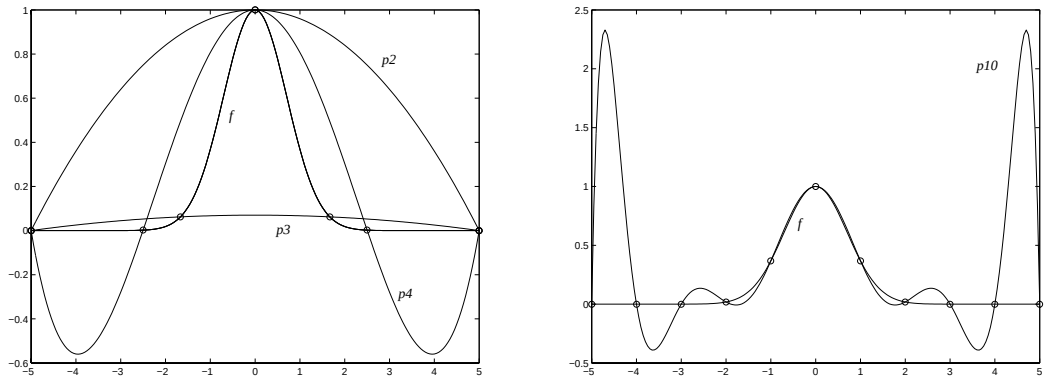


Figura 3.5: Polinomios interpoladores de grados 2, 3, 4 y 10 de $f(x) := e^{-x^2}$.

medida que aumentamos el grado del polinomio, sino todo lo contrario. En los bordes del intervalo hay una divergencia creciente a medida que aumentamos el grado. Hemos utilizado las siguientes líneas Matlab para obtener estos gráficos (sin utilizar *polyfit*).

```
%%%%% interpolaemx2.m
clear
n=10; % grado del polinomio de interpolacion
x=-5:10/n:5; % puntos q utilizamos para interpolar [-5,5]
A=vander(x); % matriz de Vandermonde
y=exp(-x.^2)'; % funcion a interpolar evaluada en los puntos utilizados
c=A\y;         % c = coeficientes del polinomio interpolador
xeval=-5:0.05:5;
yeval=exp(-xeval.^2); % evaluamos la funcion a interpolar en muchos
                      % puntos para dibujarla.
p=polyval(c,xeval);   % idem con el polinomio interpolador
plot(x,y,'o',xeval,p,'k-',xeval,yeval,'k-')
shg;
```

El mensaje que transmite este ejemplo es que aumentar el grado del polinomio interpolante sin más no significa que consigamos mejorar el modelo con el que pretendemos reproducir la función f. Con el polinomio de interpolación de Lagrange, ese *más* supone elegir los puntos de interpolación de un modo determinado (soporte de Tchebychev), de tal modo que el incremento en los valores de las derivadas sucesivas se amortigüe con una disminución del valor de $|H(x)|$.

No estudiaremos esta técnica, muy interesante, sino que propondremos una solución diferente a este problema, la interpolación utilizando polinomios a trozos, que se utiliza mucho en Ingeniería.

3.2.3. Diferencias divididas

Hay dos problemas importantes cuando usamos la base de Lagrange para interpolar una tabla de valores. El primero es el alto coste numérico. El segundo y más importante, es que si necesitamos añadir o quitar un punto al conjunto que se ha usado para construir el polinomio, tenemos que empezar los cálculos desde el principio. Con el método de las diferencias divididas este problema desaparece.

El problema sigue siendo el mismo: se trata de interpolar una función f de la que se conoce su valor en

una nube de puntos $(x_i, f_i)_{i=0,n}$ [3].

i	x_i	f_i
0	x_0	f_0
1	x_1	f_1
	$\cdots$	
n	x_n	f_n

En esta tabla no es necesario suponer que las abscisas están equiespaciadas; ni siquiera que están dadas en un cierto orden. Consideremos la base de los polinomios de grado n formada por los siguientes polinomios

$$B = \{1, (x - x_0), (x - x_0)(x - x_1), \ldots, (x - x_0)(x - x_1)\cdots(x - x_{n-1})\}$$

Ejercicio 3.2.4 *Demostrar que B es base de $P_n(\mathbb{R})$.*

Hallemos el polinomio interpolador de Lagrange P, buscando su expresión en esta base:

$$P(x) = a_0 + a_1(x - x_0) + a_2(x - x_0)(x - x_1) + \cdots + a_n(x - x_0)(x - x_1)\cdots(x - x_{n-1})$$

Vamos a ver que los coeficientes a_i se determinan fácilmente usando las diferencias divididas de los valores tabulados. Usamos una notación especial para las diferencias divididas.

Definición 3.2.1 *Se llama diferencia dividida de primer orden de la función f correspondiente a $[x_i, x_j]$ y se nota $f[x_i, x_j]$ al siguiente valor:*

$$f[x_i, x_j] := \frac{f_j - f_i}{x_j - x_i} \tag{3.11}$$

Es importante darse cuenta de que $f[x_i, x_j] = f[x_j, x_i]$. Esta conmutatividad se mantiene aunque el orden de las diferencias sea más alto. Las diferencias divididas de segundo orden y de órdenes superiores se obtienen a partir de las diferencias divididas de órdenes anteriores

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

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

Estamos ya en posición de hallar los coeficientes a_i en función de esas diferencias. Obtengamos la imagen de cada punto x_i por dicho polinomio.

$$\begin{aligned}
&P(x_0) = a_0 = f_0\\
&P(x_1) = a_0 + a_1(x_1 - x_0)\\
&P(x_2) = a_0 + a_1(x_2 - x_0) + a_2(x_2 - x_0)(x_2 - x_1)\\
&\ldots\ldots\ldots\ldots\ldots\\
&P(x_n) = a_0 + a_1(x_n - x_0) + a_2(x_n - x_0)(x_n - x_1) + \cdots +\\
&+a_n(x_n - x_0)(x_n - x_1)\cdots(x_n - x_{n-1})
\end{aligned}$$

Si hacemos $a_1 = f[x_0, x_1]$,

$$P(x_1) = f_0 + \frac{f_1 - f_0}{x_1 - x_0}(x_1 - x_0) = f_1$$

Si hacemos $a_2 = f[x_0, x_1, x_2]$,

$$P(x_2) = f_0 + \frac{f_1 - f_0}{x_1 - x_0}(x_2 - x_0) + \frac{\frac{f_2 - f_1}{x_2 - x_1} - \frac{f_1 - f_0}{x_1 - x_0}}{x_2 - x_0}(x_2 - x_0)(x_2 - x_1) = f_2$$

Podríamos ver de modo similar que cada $P(x_i) = f_i$ si $a_i = f[x_0, x_1, \ldots, x_i]$.

[3] Utilizamos en este apartado la notación $f_i := f(x_i)$, en vez de $y_i = f(x_i)$ por homogeneidad con el tratamiento que reciben las diferencias divididas en otros textos.

Ejemplo 3.2.3 *Una típica tabla de diferencias divididas podría ser la siguiente*

x_i	f_i	$f[x_i, x_{i+1}]$	$f[x_i, \ldots, x_{i+2}]$	$f[x_i, \ldots, x_{i+3}]$	$f[x_0, \ldots, x_5]$
3.2	22.0				
2.7	17.8	8.400			
1.0	14.2	2.118	2.856		
4.8	38.3	6.342	2.012	−0.528	
5.6	51.7	16.750	2.263	0.0865	0.256

Ejercicio 3.2.5 *Validar la tabla de diferencias divididas anterior y obtener el polinomio de interpolación en esta base.*

Ejercicio 3.2.6 *Utilizar Matlab para calcular este polinomio de interpolación (función* polyfit*) y hallar su valor para* $x = 3.8$*. Calcular también este valor con la expresión correspondiente a la base de diferencias divididas. Comprobar que coinciden.*

3.2.4. Interpolación simple de Hermite

Dados $n+1$ puntos distintos $x_0, \ldots, x_n$ se busca el polinomio P que cumpla las $2n+2$ condiciones

$$\begin{aligned} P(x_i) &= f(x_i) \quad i = 0, n \\ P'(x_i) &= f'(x_i) \quad i = 0, n \end{aligned}$$

Las $2n+2$ formas lineales L_i y M_i son en este caso

$$f \longrightarrow L_i(f) := f(x_i) \quad \text{y} \quad f \longrightarrow M_i(f) := f'(x_i)$$

La solución del problema de interpolación simple de Hermite es única si todos los nodos $x_0, \ldots, x_n$ son distintos[4]. La idea es que se llega a un determinante similar al de Vandermonde pero un poco más general, que es siempre distinto de cero si los nodos son distintos entre sí. Un ejemplo de aplicación directa de este tipo de interpolación la tenemos en el problema 3.3.

Dado que este problema de interpolación tiene solución única, las formas lineales $\{L_i, M_i\}$, $i = 0, n$ son linealmente independientes y tiene sentido, igual que hicimos con Lagrange, buscar su base dual $B = \{q_i, m_i\}$, $i = 0, n$, que nos permitirá escribir el polinomio solución en la forma (ver ref.[20])

$$P(x) = \sum_{i=0}^{n} f(x_i) q_i(x) + \sum_{i=0}^{n} f'(x_i) m_i(x)$$

Los polinomios de la base B se pueden poner en función de la base de Lagrange l_i, $i = 0, n$, como

$$\begin{aligned} q_i(x) &= [1 - 2l_i'(x_i)(x - x_i)] l_i^2(x) \\ m_i(x) &= (x - x_i) l_i^2(x) \end{aligned}$$

De este modo, el polinomio interpolador de Hermite se puede escribir en la forma

$$P(x) = \sum_{i=0}^{n} [1 - 2l_i'(x_i)(x - x_i)] l_i^2(x) f(x_i) + \sum_{i=0}^{n} (x - x_i) l_i^2(x) f'(x_i) \tag{3.12}$$

Ejercicio 3.2.7 *Comprobar que este polinomio verifica las propiedades que se le exigen particularizándolo primero y derivándolo y particularizándolo después en uno de los nodos.*

Ejercicio 3.2.8 *Resolver el problema 3.3 en la base* B *que acabamos de estudiar y comprobar que los valores de ambas expresiones coinciden en el punto* $x = 1$*.*

[4]Ver Davis, P. J., Interpolation and Approximation, Blaisdell, 1963.

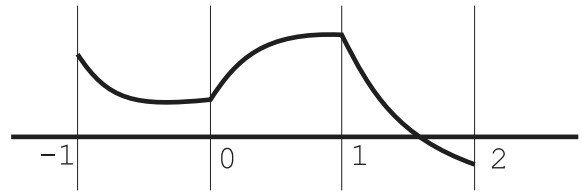


Figura 3.6: Polinomio a trozos de grado 2 y clase 0.

3.2.5. Diferencias divididas para la interpolación simple de Hermite

Es interesante ver en 3.11 cómo la diferencia dividida de primer orden es una estimación de la derivada de la función f en un punto cualquiera x_0. En efecto, cuando x_1 tiende a x_0

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

y podemos definir:

$$f[x_0, x_0] := f'(x_0)$$

Si conocemos alguna derivada de la función, podemos duplicar ese nodo en la lista y sustituir la diferencia dividida correspondiente a esos dos nodos (que son el mismo) por la derivada en ese punto, sin variar el resto del esquema-proceso. Esta idea es muy importante y práctica, y la utilizaremos a menudo en los problemas, por ejemplo, en el 3.3.

3.3. Interpolación polinomial a trozos

La construcción de polinomios de interpolación de grado alto aunque justificable teóricamente plantea muchos problemas. Por un lado, la forma de la función polinómica de grado alto a menudo no responde al fenómeno debido al gran número de extremos e inflexiones. Por otro lado, su cálculo es muy complicado, tedioso y a veces con grandes errores de redondeo por mal condicionamiento de algunas matrices. Ello limita su utilidad en análisis numérico. Es a menudo más conveniente dividir el intervalo de interés en subintervalos más pequeños y usar en cada subintervalo polinomios de grado relativamente bajo, tratando de que la función a trozos definida de este modo tenga un aspecto final adecuado al fenómeno que estamos representando.

Definición 3.3.1 *Polinomios a trozos de grado k.*

Sea $[a, b]$ un intervalo finito y $\{x_0 = a, x_1, \ldots, x_n = b\}$ los nodos de una subdivisión o partición Ω de $[a, b]$ estrictamente creciente $(x_0 < x_1 < \cdots < x_n)$.

Un polinomio a trozos de grado k en Ω es una función cuya restricción a cada uno de los subintervalos $[x_i, x_{i+1}]_{i=0,n-1}$ es un polinomio de grado k. Al espacio vectorial de estos polinomios a trozos se les denota $P_k^p(\Omega)$, siendo p su clase.

Ejemplo 3.3.1 *El ejemplo más sencillo y uno de los más interesantes es el de la interpolación a trozos de grado 1, o sea, la construcción de una poligonal que una los puntos de una nube.*

Otra posibilidad menos interesante a nivel práctico, es la representada en la Figura 3.6, correspondiente a un polinomio a trozos de grado 2 continuo pero con derivada discontinua en los nodos de enganche de los diferentes tramos. Es un tipo de polinomio a trozos que utilizaremos en algunos problemas por su interés teórico, aunque a nivel práctico, es fácil ver que este polinomio a trozos, a pesar de poder interpolar cualquier nube de puntos, no tiene necesariamente derivada continua, lo que es a menudo poco deseable.

Trabajamos más abajo con un polinomio a trozos que sí la tiene.

3.3.1. Interpolación a trozos de grado 3 y clase $\mathcal{C}^1$

Sea $[a, b]$ un intervalo de $\mathbb{R}$. Dada la subdivisión $\Omega = \{x_0 = a, \ldots, x_n = b\}$ de $[a, b]$ y los valores correspondientes $(f(x_i))$, $(f'(x_i))$, $i = 0, n$, se llama función de interpolación cúbica a trozos de clase $\mathcal{C}^1$ de la función f a un polinomio C de grado 3 a trozos cuyas restricciones a los subintervalos $[x_i, x_{i+1}]_{i=0,n-1}$ son cúbicas, y tal que cada una de esas cúbicas ajusta los valores de f y de su derivada en los nodos.

Construyamos en cada uno de los subintervalos $[x_i, x_{i+1}]_{i=0,n-1}$ un polinomio C_i de grado 3, que ajuste los valores de una función y de su derivada en cada uno de los dos nodos x_i y x_{i+1}, utilizando la fórmula de Hermite simple. Tendremos para cada $i = 0, n-1$ que

$$\begin{array}{ll} C_i(x_i) = f(x_i), & C_i(x_{i+1}) = f(x_{i+1}) \\ C_i'(x_i) = f'(x_i) & C_i'(x_{i+1}) = f'(x_{i+1}) \end{array}$$

$$\begin{aligned} C_i(x) &= [1 - 2l_i'(x_i)(x - x_i)]l_i^2(x)f(x_i) + (x - x_i)l_i^2(x)f'(x_i) \\ &+ [1 - 2l_{i+1}'(x_{i+1})(x - x_{i+1})]l_{i+1}^2(x)f(x_{i+1}) + (x - x_{i+1})l_{i+1}^2(x)f'(x_{i+1}) \end{aligned} \tag{3.13}$$

con $h_i := x_{i+1} - x_i$ y

$$\begin{aligned} l_i(x) = \frac{x - x_{i+1}}{x_i - x_{i+1}} = -\frac{x - x_{i+1}}{h_i} &\Rightarrow l_i'(x) = -\frac{1}{h_i} \\ l_{i+1}(x) = \frac{x - x_i}{x_{i+1} - x_i} = \frac{x - x_i}{h_i} &\Rightarrow l_{i+1}'(x) = \frac{1}{h_i} \end{aligned}$$

y sustituyendo en 3.13

$$\begin{aligned} C_i(x) &= \left[1 + \frac{2}{h_i}(x - x_i)\right]\left(\frac{x - x_{i+1}}{h_i}\right)^2 f(x_i) + (x - x_i)\left(\frac{x - x_{i+1}}{h_i}\right)^2 f'(x_i) \\ &+ \left[1 - \frac{2}{h_i}(x - x_{i+1})\right]\left(\frac{x - x_i}{h_i}\right)^2 f(x_{i+1}) + (x - x_{i+1})\left(\frac{x - x_i}{h_i}\right)^2 f'(x_{i+1}) \end{aligned} \tag{3.14}$$

Como consecuencia, la definición del polinomio a trozos es:

$$C(x) = C_i(x), \quad x \in [x_i, x_{i+1}), \quad i = 0, n-1 \tag{3.15}$$

cerrando por la derecha en el último punto. Un ejemplo de este tipo de polinomios lo tenemos en el problema 3.4.

Al ajustar simultáneamente valores y derivadas, hemos construido un polinomio a trozos y de clase $\mathcal{C}^1$. Para ello, hemos necesitado conocer los valores de la derivada de la función en los nodos. A menudo esto no es posible y complica innecesariamente el problema. ¿Es posible construir polinomios de interpolación a trozos de una determinada clase sin necesidad de imponer en los nodos de enganche los valores de las derivadas correspondientes? La respuesta afirmativa la dan los *Splines*.

3.4. Interpolación polinomial a trozos: Splines

Como acabamos de ver, los splines surgen como respuesta a la pregunta de si es posible construir polinomios interpolantes a trozos de cierta clase sin necesidad de asignar explícitamente en los nodos los valores de las derivadas hasta la clase correspondiente. La notación spline se debe a Schoenberg, que la introdujo en un artículo [5] muy famoso en 1946, aunque el concepto ya se había utilizado anteriormente.

Spline es la traducción inglesa de junquillo. Antes de que toda la delineación se hiciese con programas de CAD, los delineantes, sobre todo en los astilleros, ajustaban las formas de los barcos mediante unas varillas de vidrio flexibles, los junquillos, que hacían pasar por los puntos por los que se quería que pasase una curva, fijándolo, una vez conseguido el objetivo, con unas pesas para luego dibujar apoyándose en ellas la curva en cuestión (ver la Figura 3.7).

Los splines cumplen ahora la misma función en los programas de CAD.

La idea es por tanto buscar polinomios interpolantes a trozos P_k^p de grado k y de clase $\mathcal{C}^p$ en $[a, b]$ con $p > 0$. La clase debe ser estrictamente inferior al grado $p < k$ ya que si $p = k$, el polinomio que se obtiene es el mismo en todos los subintervalos. Es especialmente importante entender esta afirmación. Si en un determinado nodo imponemos que el trozo de la izquierda enganche con el de la derecha con continuidad hasta el orden k, la única posibilidad es que los dos polinomios sean el mismo. Veámoslo en la Figura 3.8 y analíticamente con una cúbica a trozos C.

[5] Schoenberg, I. J., "Contributions to the problem of approximation of equidistant data by analytic functions", *A,B. Quart. Appl. Math.*, 4, 1946.

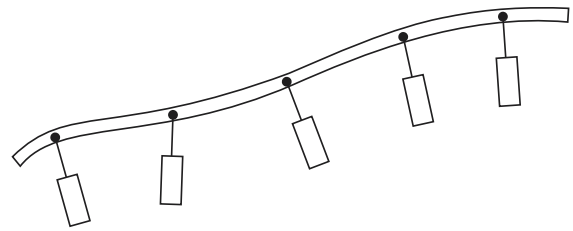


Figura 3.7: Junquillo.

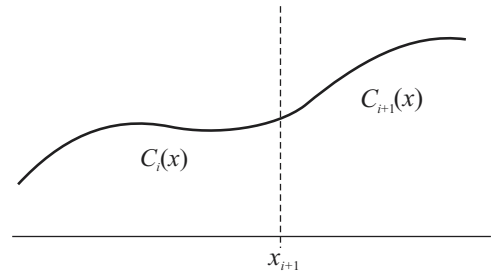


Figura 3.8: Enganche entre dos tramos de una cúbica a trozos.

Supongamos que estamos en el nodo $i+1$ y que tenemos para el polinomio a trozos C por la izquierda de x_{i+1} la cúbica C_i y por la derecha la cúbica C_{i+1}.

$$C_i(x) = a_i x^3 + b_i x^2 + c_i x + d_i$$

$$C_{i+1}(x) = a_{i+1} x^3 + b_{i+1} x^2 + c_{i+1} x + d_{i+1}$$

Impongamos que el valor y las tres derivadas sucesivas de estas dos cúbicas coinciden en x_{i+1}

$$\begin{aligned} a_i x_{i+1}^3 + b_i x_{i+1}^2 + c_i x_{i+1} + d_i &= a_{i+1} x_{i+1}^3 + b_{i+1} x_{i+1}^2 + c_{i+1} x_{i+1} + d_{i+1} \\ 3a_i x_{i+1}^2 + 2b_i x_{i+1} + c_i &= 3a_{i+1} x_{i+1}^2 + 2b_{i+1} x_{i+1} + c_{i+1} \\ 6a_i x_{i+1} + 2b_i &= 6a_{i+1} x_{i+1} + 2b_{i+1} \\ 6a_i &= 6a_{i+1} \end{aligned}$$

De la última ecuación obtenemos la igualdad de los coeficientes a_i y a_{i+1}. Entrando con esta igualdad en la tercera obtenemos la igualdad de los coeficientes b_i y b_{i+1}, y así sucesivamente, luego las dos cúbicas son la misma. Hay que dejar algún grado de libertad y, por tanto, lo máximo que podemos imponer es continuidad hasta la segunda derivada. O sea, que si el grado es 3, la continuidad máxima es $\mathcal{C}^2$. En general, si el grado es k, la clase p será como máximo $k-1$, y de hecho este es el caso más interesante, por ser los de mayor clase posible.

Definición 3.4.1 *Sean $[a,b]$ un intervalo de $\mathbb{R}$, $\Omega = \{x_0 = a, x_1, \ldots, x_n = b\}$ una partición de $[a,b]$ estrictamente creciente $x_0 < x_1 < \ldots < x_n$. Se llama spline de grado k asociado a la partición Ω a cualquier polinomio a trozos de grado k y de clase $\mathcal{C}^{k-1}$, o sea, con derivadas continuas hasta el orden $k-1$.*

Denotaremos $S_k(\Omega)$ al espacio vectorial de los splines de grado k asociados a la partición Ω.

Ejercicio 3.4.1 *Demostrar que $S_k(\Omega)$ es un espacio vectorial.*

Ejercicio 3.4.2 *Demostrar que $P_k(\mathbb{R})$, polinomios de grado k, son un subespacio vectorial de $S_k(\Omega)$, cuando se les considera restringidos a $[a,b]$.*

Una vez definidos los splines, tenemos que estudiar sus propiedades en lo que se refiere a interpolar valores de una función en los nodos de la partición Ω, discutiendo en qué condiciones esto es posible y en qué condiciones las soluciones son únicas. En este sentido, nos limitaremos al caso $k = 3$ de los splines cúbicos, el más importante en la práctica, y dejaremos para los problemas el estudio de los splines de grado 2, que también son muy interesantes desde un punto de vista teórico.

3.4.1. Splines cúbicos

Definición 3.4.2 *Sean $[a,b]$ un intervalo finito, $\Omega = \{x_0 = a, x_1, \ldots, x_n = b\}$ una partición equiespaciada*[6] *de $[a,b]$ estrictamente creciente $x_0 < x_1 < \ldots < x_n$. Se llama spline de interpolación cúbica a una función de clase $\mathcal{C}^2$ tal que la restricción a cada tramo es un polinomio de grado 3.*

Ejemplo 3.4.1 *Dada la partición $\Omega = \{x_0 = a, x_1, x_2 = b\}$ del intervalo $[a,b]$, se define el polinomio a trozos S*

$$S(x) := \begin{cases} S_0(x) = a_0x^3 + b_0x^2 + c_0x + d_0, & x_0 \le x < x_1 \\ S_1(x) = a_1x^3 + b_1x^2 + c_1x + d_1, & x_1 \le x \le x_2 \end{cases}$$

Impongamos que S sea un spline cúbico que interpole una función f en esos nodos.
Llamemos $y_i := f(x_i)$, $i = 0, 2$.
Primeramente, el valor en los nodos ha de coincidir con el de la función

$$S_0(x_0) = y_0, \quad S_1(x_1) = y_1, \quad S_1(x_2) = y_2$$

Además, la función ha de ser continua, con derivadas primera y segunda continuas en x_1, el único nodo intermedio

$$S_0(x_1) = S_1(x_1), \quad S'_0(x_1) = S'_1(x_1), \quad S''_0(x_1) = S''_1(x_1)$$

Tenemos seis condiciones y ocho parámetros a determinar, los 8 coeficientes de las dos cúbicas S_0 y S_1. Por tanto, habrá infinitas soluciones, y para eliminar la ambigüedad tendremos que añadir dos condiciones.

La situación que se da en este ejemplo es semejante a la que se produce en general con los splines cúbicos, ya que si el spline tiene n tramos, tenemos que definir $4n$ números, pues cada tramo es una cúbica con 4 coeficientes. Si imponemos los valores en todos los nodos, $n+1$ condiciones, y la continuidad de la función y de sus dos primeras derivadas en todos los nodos interiores, $3(n-1)$ condiciones, tendremos $4n - (n+1) - 3(n-1) = 2$ parámetros libres. O sea, que en general también tenemos 2 grados de libertad y necesitamos por tanto dos condiciones adicionales.

Ejercicio 3.4.3 *Se tiene la partición $\Omega = \{x_0, x_1, \ldots, x_n\}$ del intervalo $[a,b]$. Calcular la dimensión del espacio vectorial $S_3(\Omega)$, razonando de igual modo. La solución es $n+3$.*

El objetivo ahora es buscar esas dos condiciones adicionales que hagan que todo el problema tenga solución única. Vamos a demostrar que definir las derivadas en el primer y último punto de la partición son dos condiciones adecuadas.

Teorema 3.4.1 *Sean $[a,b]$ un intervalo de $\mathbb{R}$, $\Omega = \{x_0 = a, x_1, \ldots, x_n = b\}$ una partición equiespaciada de $[a,b]$, $\{y_i\}_{i=0,n}$, s_0, s_n, $(n+3)$ números reales. Existe un spline cúbico único S tal que*

$$S(x_i) = y_i, \quad i = 0, n, \quad S'(x_0) = s_0, \quad S'(x_n) = s_n$$

Demostración:
Consideremos el tramo i, S_i, del spline S; el que define el valor del spline cuando $x_i \le x < x_{i+1}$ (con $S(x_n) := S_{n-1}(x_n)$).

Si escribimos ese tramo en la base de Hermite en función de los valores y_i e y_{i+1} conocidos en los nodos y de los valores de las derivadas en estos nodos s_i y s_{i+1}, que todavía desconocemos, tendremos

$$\begin{aligned} S_i(x) &= \left[\frac{(x-x_{i+1})^2}{h^2} + \frac{2(x-x_i)(x-x_{i+1})^2}{h^3}\right] y_i + \left[\frac{(x-x_i)^2}{h^2} - \frac{2(x-x_{i+1})(x-x_i)^2}{h^3}\right] y_{i+1} \\ &+ \frac{(x-x_i)(x-x_{i+1})^2}{h^2} s_i + \frac{(x-x_i)^2(x-x_{i+1})}{h^2} s_{i+1} \end{aligned} \tag{3.16}$$

con $h = x_{i+1} - x_i$.

Esta cúbica se ha construido usando la misma filosofía que la usada para definir los polinomios a trozos de grado 3 y clase C^1, obtenidos mediante la base de Hermite. Fueron estos polinomios los que utilizamos

[6] El que sea equiespaciada obedece a que permite una mayor simplicidad en los desarrollos, pero no introduce ningún aspecto teórico nuevo.

en la sección 3.3.1 para llegar a cada uno de los tramos de 3.15. Con esta definición, no hay ninguna duda de que nuestro spline va a tener derivada continua, ya que ésta coincide por la derecha y por la izquierda en todos los nodos interiores (ver problema 3.4):

$$\begin{aligned} S_i(x_{i+1}) &= S_{i+1}(x_{i+1}) = y_{i+1} \\ S'_i(x_{i+1}) &= S'_{i+1}(x_{i+1}) = s_{i+1} \end{aligned}$$

Como s_0 y s_n son datos y sabemos los valores y_i en todos los nodos, sólo necesitamos encontrar las derivadas s_i en los nodos interiores $x_i \;\; i = 1, n-1$, que son en los que se producen los enganches. Recordemos que todavía nos falta imponer la continuidad de la segunda derivada, para lo cual diferenciamos dos veces la expresión 3.16

$$S''_i(x_{i+1}) = \frac{6}{h^2}y_i - \frac{6}{h^2}y_{i+1} + \frac{2}{h}s_i + \frac{4}{h}s_{i+1} \tag{3.17}$$

$$S''_{i+1}(x_{i+1}) = -\frac{6}{h^2}y_{i+1} + \frac{6}{h^2}y_{i+2} - \frac{4}{h}s_{i+1} - \frac{2}{h}s_{i+2} \tag{3.18}$$

Si igualamos ambas expresiones para forzar la continuidad de la segunda derivada, obtenemos el sistema

$$s_i + 4s_{i+1} + s_{i+2} = \frac{3}{h}(y_{i+2} - y_i) \quad i = 0, n-2 \tag{3.19}$$

que podemos escribir

$$\begin{pmatrix} 4 & 1 & 0 & \cdots & \cdots & 0 & 0 \\ 1 & 4 & 1 & 0 & \cdots & \cdots & 0 \\ 0 & 1 & 4 & 1 & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & 0 & 1 & 4 \end{pmatrix} \begin{pmatrix} s_1 \\ s_2 \\ s_3 \\ \vdots \\ s_{n-1} \end{pmatrix} = \frac{3}{h} \begin{pmatrix} y_2 - y_0 - \frac{h}{3}s_0 \\ y_3 - y_1 \\ y_4 - y_2 \\ \vdots \\ y_n - y_{n-2} - \frac{h}{3}s_n \end{pmatrix} \tag{3.20}$$

Un sistema tridiagonal y de diagonal estrictamente dominante de $n-1$ ecuaciones con $n-1$ incógnitas. El que sea estrictamente dominante implica que la matriz es regular y que por tanto el sistema lineal tiene solución única. De este modo hemos definido las derivadas en todos los nodos. De cada tramo conocemos los valores en los nodos y las derivadas en los nodos, y por tanto, la cúbica de cada tramo es única. Y si la cúbica de cada tramo es única, el spline que los engloba también lo es.

Ejercicio 3.4.4 *Calcular el spline cúbico correspondiente a la función $f(x) = \sin(1/x)$ (la misma que en el problema 3.3) asociado a la partición $\Omega = \{0.20, 0.40, 0.60, 0.80\}$ del intervalo $[0.20, 0.80]$, y que interpole también las derivadas en el primer y último nodo. Será suficiente para definirlo con calcular las derivadas en los nodos intermedios. Se pide el valor de dicho spline en el punto 0.45.*

3.4.2. Splines cúbicos naturales

En vez de las condiciones adicionales que se refieren a las derivadas en los extremos, se pueden fijar las segundas derivadas en los extremos. Si a éstas se les asigna el valor cero, se obtienen los splines cúbicos naturales.

$$S''(x_0) = S''(x_n) = 0 \tag{3.21}$$

Ejercicio 3.4.5 *Demostrar que existe un spline natural cúbico único. (Indicación: se trata de plantear el mismo sistema lineal en las derivadas con dos líneas adicionales correspondientes a las segundas derivadas en los nodos extremos, obtenidas de 3.17 y 3.18. Queda otro sistema de diagonal estrictamente dominante con el que obtenemos las derivadas en todos los nodos. Como además conocemos los valores en los nodos, ya tenemos completamente definido el spline.)*

Ejercicio 3.4.6 *Calcular el spline cúbico natural correspondiente a la función $f(x) = \sin(1/x)$ (la misma que en el ejemplo 3.3) asociado a la partición $\Omega = \{0.20, 0.40, 0.60, 0.80\}$ del intervalo $[0.20, 0.80]$. Será suficiente para definirlo con calcular las derivadas en los nodos intermedios. Se pide el valor de dicho spline en el punto 0.45 comparándolo con el obtenido en el ejercicio 3.4.4.*

Al comienzo de esta sección explicamos el origen del nombre spline sinónimo de junquillo, y de qué modo se hacían pasar los junquillos por unos puntos sujetándolos con pesas como aparece en la Figura 3.7. Antes y después del primer y último punto, cuando ya no hay pesas, la varilla de vidrio permanece recta, y las rectas verifican que su segunda derivada es cero. Por eso el spline cúbico natural es que el más se parece al junquillo.

3.4.3. Bases de splines asociadas a un problema de interpolación

Los splines de un determinado grado forman un espacio vectorial de dimensión finita (ver ejercicios 3.4.1 y 3.4.3). Podemos intentar calcular bases de este espacio para tener una forma más general de escribir sus elementos y así utilizar esta descripción en otros problemas.

Para los splines cúbicos, aplicamos la misma idea que ya hemos utilizado con éxito en varios problemas de interpolación de solución única. Las formas lineales asociadas al problema definirán una base del dual del espacio de polinomios considerado con cuya base dual habremos resuelto el problema.

Por ejemplo, sea $S_3(\Omega)$ el espacio vectorial de los splines cúbicos correspondientes a una partición Ω. Sea $S_3^*(\Omega)$ su dual. Definimos las $n+3$ formas lineales:

$$\begin{aligned} L_i(S) &= S(x_i), \quad i = 0, n \\ L_{n+1}(S) &= S'(x_0) \\ L_{n+2}(S) &= S'(x_n) \end{aligned}$$

Como ya hemos visto, estas formas lineales definen un problema de interpolación de solución única, y son por tanto una base de $S_3^*(\Omega)$. Podemos encontrar su base dual $(c_i)_{i=0,n+2}$ resolviendo el sistema 3.19 pues sabemos que:

$$L_j(c_i) = \delta_{ij}$$

De cada elemento c_j sabemos sus valores en los nodos y sus derivadas en el primero y último, y por tanto está completamente definido.

En el problema 3.6 se encuentra una aplicación directa de estas ideas.

Cualquier spline del que sepamos sus valores en los nodos, y las derivadas en los nodos extremos, puede ser escrito como combinación lineal de esta base de modo sencillo, lo que a efectos computacionales es muy práctico. Controlando esos valores podemos controlar muy bien el aspecto de la curva. Así funcionan los paquetes de diseño CAD.

$$S(x) = \sum_{j=0}^{n} y_j c_j(x) + s_0 c_{n+1}(x) + s_n c_{n+2}(x) \tag{3.22}$$

3.5. Interpolación spline con bases de soporte mínimo: B-splines

3.5.1. Introducción

Sea Ω una partición del intervalo $[a,b]$, $\Omega = \{t_0 = a, t_1, \ldots, t_{n-1}, t_n = b\}$ [7].

Un polinomio de interpolación a trozos de grado k y clase $\mathcal{C}^p$ está definido por n polinomios de grado k en cada uno de los intervalos $[t_i, t_{i+1}]_{i=0,n-1}$. Como tenemos n tramos, el número de parámetros a definir es $n\cdot(k+1)$. El número de condiciones se refiere a la continuidad de orden p en los $(n-1)$ enlaces, $(p+1)\cdot(n-1)$. La dimensión de este espacio vectorial de funciones polinómicas a trozos de clase $\mathcal{C}^p$ y grado k será

$$dim(P_k^p(\Omega)) = n \cdot (k+1) - (p+1) \cdot (n-1)$$

Si lo que tenemos es un spline de grado k, su clase es $p = k-1$. En consecuencia, la dimensión de ese espacio de splines $S_k(\Omega)$ será:

$$dim(S_k(\Omega)) = n \cdot (k+1) - (k-1+1) \cdot (n-1) = n + k$$

[7] Usaremos aquí la letra t para la variable independiente en vez de la x por ser la notación habitual en los libros que tratan el tema. La notación se debe a que se suelen construir curvas en el plano o en el espacio dependientes de un parámetro a partir de splines. En el plano, por ejemplo, cada una de las coordenadas de esa curva $(x(t), y(t))$ es a su vez un spline.

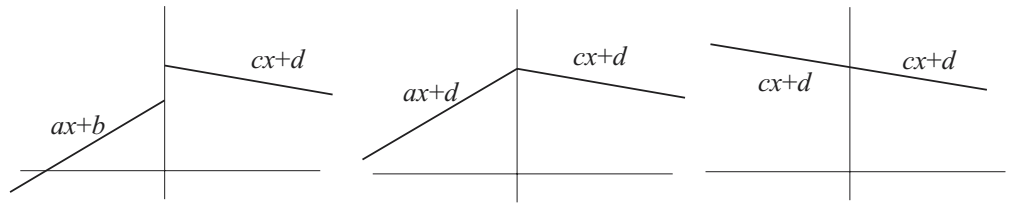


Figura 3.9: Dimensión de un espacio de polinomios a trozos.

Cuando fijamos una condición de continuidad de cualquier derivada de cualquier orden en un punto de enganche de dos tramos estamos reduciendo en uno la dimensión del espacio de funciones. El ejemplo que incluimos a continuación nos va a permitir ver muy bien este punto.

Consideremos el siguiente tipo de funciones a trozos

$$r(x) = \begin{cases} ax + b & x < 0 \\ cx + d & x \geq 0 \end{cases}$$

Los entes de este tipo forman un espacio vectorial E de dimensión 4, que son los grados de libertad que debemos fijar para definir un elemento genérico de ese espacio (ver Figura 3.9).

Ejercicio 3.5.1 *Definir una base de E. Indicación: Hacer 1 uno de los coeficientes y 0 los otros tres y ver qué pasa con $r(x)$.*

Si imponemos que la función r debe ser continua en $x = 0$, el límite por la derecha y por la izquierda coinciden en el punto. Dado que el valor correspondiente a $x = 0$ es $r(0) = d$, el límite por la izquierda que es directamente b debe ser igual a d (ver Figura 3.9).

El nuevo aspecto de la función r es

$$r(x) = \begin{cases} ax + d & x < 0 \\ cx + d & x \geq 0 \end{cases}$$

Las funciones de este tipo describen un subespacio E' de E de dimensión 3.

Ejercicio 3.5.2 *Definir una base de E' y demostrar que E' es un subespacio vectorial de E.*

Si imponemos que la función r debe tener derivada continua en $x = 0$, sus pendientes por la derecha y por la izquierda de $x = 0$ coincidirán. Dado que el valor correspondiente a $x = 0$ es $r'(0) = c$, esa continuidad exige que $a = c$ lo que se traduce en que la parte de la derecha y la de la izquierda son la misma; lo que tenemos es una recta y la dimensión de este espacio no es 3 sino 2.

Con este ejemplo hemos visto que cada restricción lineal reduce la dimensión del espacio en 1.

3.5.2. Soporte de un polinomio a trozos

Es importante entender que para manejar los espacios de splines y extraer de ellos todo lo que pueden ofrecer es bueno contemplarlos como espacios vectoriales convencionales, apoyándonos en bases para trabajar en ellos. Hay muchas bases del espacio $S_k(\Omega)$ que son útiles en la resolución de nuestro problema. En la sección 3.4.3 hemos estudiado cómo construirlas asociadas a problemas de interpolación con solución única. Disponer de una base de este tipo facilita la escritura del elemento buscado en dicha base, pues a menudo sus componentes son los propios valores a interpolar como ya hemos visto.

En estas bases normalmente el polinomio solución se puede escribir del siguiente modo

$$P_{n,k}(t) = \sum_{i=0}^{n} f_i \cdot e_i(t) + \sum_{i=n+1}^{n+k-1} a_i \cdot e_i(t) \tag{3.23}$$

siendo los f_i los valores de la función a interpolar y los demás coeficientes a_i se refieren a las condiciones adicionales, como derivadas en los extremos, segundas derivadas nulas, etc.

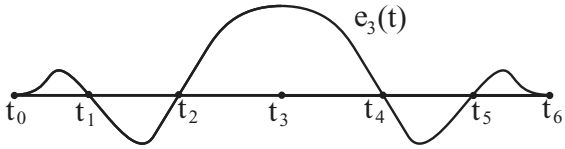


Figura 3.10: Dual de L_3 tal que $L_3(S) = S(t_3)$.

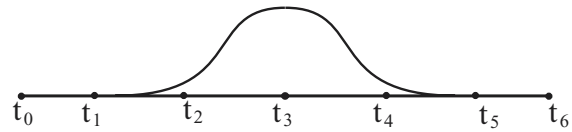


Figura 3.11: Spline con soporte pequeño.

No todas las bases de $S_k(\Omega)$ son convenientes a la hora de evaluar el polinomio de interpolación descompuesto respecto de dicha base. Interesa que los elementos de la base sean nulos en la mayor parte del dominio de definición para no tener que calcular su valor en ellos sabiendo de antemano que ahí son nulos. Para manejar mejor esta idea, introducimos un nuevo concepto:

Definición 3.5.1 *Se denomina soporte de una función, al subconjunto del dominio de definición de la función donde esa función es no nula.*

El planteamiento de bases duales que es fenomenal para calcular los coeficientes de la escritura del spline en esa base, tiene el inconveniente de que las funciones $e_i(t)$ de la base dual tienen un soporte muy grande. Ello obliga a que, independientemente del punto t donde queramos evaluar $P_{n,k}(t)$, haya que extender los sumatorios desde $i = 0$ hasta $i = n+k-1$, con el consiguiente consumo de recursos, como podemos observar en la Figura 3.10. Lo ideal sería que las funciones $e_i(t)$, siendo splines de grado k, tuvieran soporte mínimo, es decir, que sean nulas en el mayor número posible de tramos $[t_i, t_{i+1}]$ (ver Figura 3.11).

Ejercicio 3.5.3 *Demostrar que el soporte de un spline está directamente relacionado con los tramos del mismo, en el sentido de que, salvo en un conjunto finito de puntos, cada tramo completo pertenecerá o no al soporte.*

3.5.3. Splines de soporte mínimo: B-splines

Hemos visto en la sección anterior que es conveniente que los elementos de la base del espacio de splines tengan soporte mínimo. A los splines de soporte mínimo se les llama B-splines. Si los elegimos adecuadamente formarán una base de $S_k(\Omega)$. Un concepto importante para estudiar estos splines es el de orden.

Definición 3.5.2 *El orden r de un B-spline es el número de tramos en los que el B-spline es no nulo.*

Calculemos el orden[8] r de un B-spline en función de su grado k. Si el spline es no nulo en r tramos, tendremos que definir por tanto $r(k+1)$ coeficientes. Si aplicamos las k restricciones de continuidad de función y las $r-1$ de la derivada en los $r-1$ nodos interiores del soporte nos quedarán $r(k+1) - k(r-1) = r+k$ parámetros libres (ver Figura 3.11).

En los extremos del soporte el B-spline se hace nulo y engancha con tramos nulos; por tanto, además de ser nulo, sus derivadas hasta la $p = k-1$ también lo deben ser. Por consiguiente, tenemos $2k$ restricciones más, con lo que el número de parámetros se reduce de $r+k$ a $r-k$.

Pero qué pasa si tomamos el número de tramos r como el grado k, o sea, si $r-k=0$. En este caso, el B-spline sería idénticamente nulo[9], por tanto, el mínimo r con sentido es $r = k+1$.

[8] Cuando hablamos de un B-spline manejamos siempre tres características. Las dos primeras, el grado (k) y su clase ($\mathcal{C}^p$), son comunes a todos los spline; la tercera es el orden.

[9] Por ejemplo, si lo que tenemos es un spline parabólico, $k = 2$, no podemos conseguir con sólo dos tramos que sea distinto de cero, y luego vuelva a ser cero con continuidad en la función y en su derivada.

En el primer tramo, tenemos una parábola, que viene definida por tres coeficientes a, b, y c. El primer punto de esa parábola vale cero y tiene derivada nula. Si además queremos que la imagen del último punto sea distinta de cero y valga un determinado valor, ya tendremos completamente definida la parábola y por tanto su derivada en ese punto valdrá un valor que no podemos controlar. Una vez fijado el valor de la función en ese punto de enganche, a su vez el segundo tramo tendrá en su punto final derivada nula y valor nulo, con lo que sólo quedará un parámetro libre: el valor de la función en el enganche (continuidad). Si además queremos que haya continuidad en la derivada, que ya viene fijada por el otro tramo, pedimos demasiado.

En la Figura 3.12 da la sensación de que lo hemos conseguido, pero si nos fijamos bien en cada uno de los tramos, ¿desde cuándo una parábola tiene un punto de inflexión?

Si aplicamos este resultado a los splines de grado 2, veremos que su orden es 3, y que para los B-splines cúbicos, su orden es 4; necesitaremos para poder construir un B-spline cúbico al menos un soporte de 4 tramos (ver otra vez la Figura 3.11).

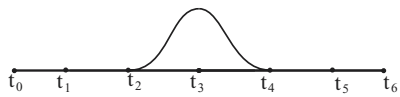


Figura 3.12: Ejemplo de un spline que no puede ser spline.

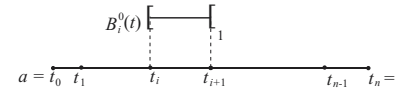


Figura 3.13: B-spline de grado 0.

Definición 3.5.3 *Dado el conjunto de nodos*

$$t_{-r+1} < \cdots < t_{-1} < a = t_0 < t_1 < t_2 < \cdots < t_{n-1} < t_n = b < \cdots < t_{n-1+r}$$

un B-spline de grado k asociado a los mismos es una función polinómica a trozos de grado k, de clase $\mathcal{C}^{k-1}$ en $[a, b]$ y de soporte mínimo.

Añadimos nodos virtuales a la izquierda de t_0 y a la derecha de t_n para que las expresiones generales que demos a los B-splines sirvan para todos los nodos y no haya que hacer ninguna consideración especial con los nodos iniciales y finales. Esto ayuda mucho en computación, que es donde más se utilizan estos entes, para representar curvas y superficies.

Construiremos una base de este espacio vectorial $S_k(\Omega)$ de polinomios a trozos formada por elementos que notaremos $B_i^k(t)$, siguiendo la siguientes convenciones [10]

$$\begin{array}{rcl} t \in [t_i, t_{i+r}) & \Rightarrow & B_i^k(t) \geq 0 \\ t \notin [t_i, t_{i+r}) & \Rightarrow & B_i^k(t) = 0 \end{array}$$

o sea, diciendo que el elemento $B_i^k(t)$ empieza en el nodo $i-$ésimo y acaba donde le obliga su orden, o sea, en el nodo $i+r$. Como i toma los valores $-r+1, -r+2, 0, 1, \ldots, n-1$, la base B de $S_k(\Omega)$ tendrá los siguientes elementos

$$B = \{B_{-r+1}^k, \cdots, B_{n-1}^k\} \tag{3.24}$$

Demostrar que estas funciones son una base, no es nada fácil. Cuando el grado es bajo resulta fácil intuirlo viendo la gráfica de las funciones que no vamos a poder escribir como combinación lineal de los demás, pero la demostración general no es sencilla (ver [15]).

Ejercicio 3.5.4 *Comprobar que el número de elementos así definidos coincide con la dimensión de $S_k(\Omega)$.*

3.5.4. B-splines de grado 0

$$B_i^0(t) = \begin{cases} 0, & t \notin [t_i, t_{i+1}) \\ 1 & t \in [t_i, t_{i+1}) \end{cases} \tag{3.25}$$

Los B-splines de grado 0 tienen orden 1 y clase $\mathcal{C}^{-1}$, siendo por tanto tramos constantes pero discontinuos en los enganches. Tenemos en la colección un problema interesante en esta base, el 5.3.

Ejercicio 3.5.5 *Sea Ω una partición del intervalo $[a, b]$, $\Omega = \{t_0 = a, t_1, \ldots, t_{n-1}, t_n = b\}$. Calcular la dimensión de $S_0(\Omega)$, y definir una base formada por B-splines.*

[10] Esta forma de construcción de una base de B-splines no es la única posible. Sin embargo, las expresiones que damos son de uso común, y permiten abordar de modo sencillo los problemas de interpolación spline con este tipo de funciones de base.

3.5.5. B-splines de grado 1

La clase es 0 y el orden es 2, lo que significa que el soporte de cada elemento de la familia es de dos tramos. La base estará formada por $n+k = n+1$ splines. Añadimos un punto antes de t_0, el t_{-1} cuya abscisa debe ser arbitraria pero menor que t_0, para completar los nodos en los cuales se van a apoyar los splines que formarán la base (ver la Figura 3.14). La dimensión de $S_1(\Omega)$ es $n+1$ y, por tanto, particularizando la configuración general de la base, 3.24, la base tendrá $n+1$ elementos:

$$B = \{B^1_{-1}, \cdots, B^1_{n-1}\}$$

Los B-splines de grado 1 coinciden en cierta medida con una base dual de splines. Es fácil ver que $B^1_i(t_j) = \delta_j^{i+1}$. Para grados más altos, esto no pasa, como luego veremos. Que el valor máximo sea 1 es en principio arbitrario, pero consistente con la expresión general utilizada para construir los B-splines de grados superiores, que estudiaremos más adelante.

La expresión general de uno de estos elementos cuando el grado es 1 es

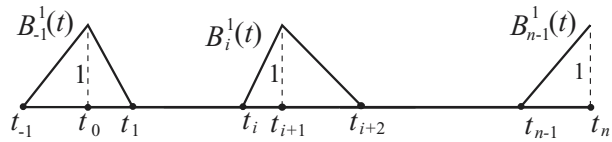


Figura 3.14: B-spline de grado 1.

$$B^1_i(t) = \begin{cases} 0, & t \notin [t_i, t_{i+2}) \\ \frac{t-t_i}{t_{i+1}-t_i}, & t \in [t_i, t_{i+1}) \\ \frac{t_{i+2}-t}{t_{i+2}-t_{i+1}}, & t \in [t_{i+1}, t_{i+2}) \end{cases} \tag{3.26}$$

3.5.6. Interpolación con B-splines de grado 1

Sea Ω una partición del intervalo $[a,b]$, $\Omega = \{t_0 = a, t_1, \ldots, t_{n-1}, t_n = b\}$. Se plantea el problema de encontrar un spline de grado uno que interpole a una función en esa partición. Sean $y_j = f(t_j)$, $j = 0, n$ los valores a interpolar. El número de condiciones, $(n+1)$, coincide con la dimensión del espacio $S_1(\Omega)$. En realidad estamos construyendo la poligonal que une todos esos puntos (t_j, y_j), $j = 0, n$, y este problema tiene solución única. Escribamos esa poligonal en la base de B-splines:

$$P_1(t) = \sum_{i=-1}^{n-1} a_i B^1_i(t)$$

Apliquemos las condiciones:

$$y_j = P_1(t_j) = \sum_{i=-1}^{n-1} a_i B^1_i(t_j) = \sum_{i=-1}^{n-1} a_i \delta_j^{i+1} = a_{j-1}$$

Y por tanto,

$$a_i = y_{i+1},\ i = -1, n-1 \quad \Rightarrow \quad P_1(t) = \sum_{i=-1}^{n-1} y_{i+1} B^1_i(t)$$

Ejercicio 3.5.6 *Sea Ω una partición del intervalo $[0.2, 0.8]$, $\Omega = \{0.20, 0.40, 0.60, 0.80\}$. Encontrar un spline de grado uno que interpole a la función $f(x) = \sin(1/x)$ en esa partición escribiéndolo en la base de los B-splines. Hallar su valor para $t = 0.45$.*

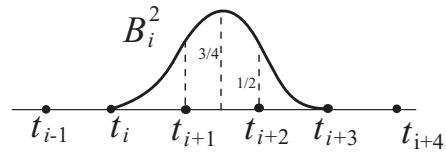


Figura 3.15: B-spline de grado 2.

3.5.7. B-splines de grado k

El orden será $r = k+1$. Se dispone de una formulación recurrente (ref. [18]) que nos permite obtener la expresión de una base de B-splines de grado k en función de la correspondiente a $k-1$, cuando la partición es equiespaciada.

$$B_i^k(t) = \left(\frac{t - t_i}{t_{i+k} - t_i}\right) B_i^{k-1}(t) + \left(\frac{t_{i+k+1} - t}{t_{i+k+1} - t_{i+1}}\right) B_{i+1}^{k-1}(t) \tag{3.27}$$

Aunque la expresión es válida, no la usaremos nunca en la práctica debido a su complejidad. Damos en la siguiente sección la expresión general de los de grado 2 obtenida de este modo pero en realidad la presentamos más para mostrar su estructura que por razones prácticas.

3.5.8. B-splines de grado 2 en una partición equiespaciada

Usando la fórmula recurrente 3.27, y teniendo en cuenta que la partición consta de nodos equiespaciados entre sí, $h = t_{i+1} - t_i$, podemos obtener fácilmente la expresión correspondiente a un elemento de un sistema de B-splines de este tipo.

En los B-splines de grado 2 y 3 siempre utilizaremos particiones equiespaciadas por dos razones: la primera para no complicar innecesariamente los problemas y la segunda porque la realidad en Ingeniería y Computación es que los métodos se estructuran para funcionar con esa restricción, ya que simplifica mucho las cosas.

$$B_i^2(t) = \begin{cases} 0, & t \notin [t_i, t_{i+3}) \\ \frac{1}{2h^2}(t - t_i)^2, & t \in [t_i, t_{i+1}) \\ \frac{1}{2h^2}[h^2 + 2h(t - t_{i+1}) - 2(t - t_{i+1})^2], & t \in [t_{i+1}, t_{i+2}) \\ \frac{1}{2h^2}(t_{i+3} - t)^2, & t \in [t_{i+2}, t_{i+3}) \end{cases} \tag{3.28}$$

En la práctica habitual utilizaremos su gráfica (ver la Figura 3.15) y los valores en los nodos, olvidándonos de la expresión anterior, que es necesaria al programar estos métodos en el ordenador pero muy poco cómoda para realizar los ejercicios y problemas.

La dimensión de $S_2(\Omega)$ es $n+2$ y por tanto, particularizando la configuración general de la base, 3.24, la base tendrá $n+2$ elementos, lo que significa que necesitamos dos nodos virtuales a la izquierda.

$$B = \{B_{-2}^2, \cdots, B_{n-1}^2\}$$

3.5.9. Interpolación con B-splines de grado 2

En grado uno ya hemos visto que el problema de escribir la solución del problema de interpolación en la base de B-splines es sencillo. A medida que aumentamos el grado, el problema se complica y parte de esa complejidad es consecuencia de la resolución de sistemas lineales con muchos elementos nulos.

Si Ω es una partición del intervalo $[a, b]$, $\Omega = \{t_0 = a, t_1, \ldots, t_{n-1}, t_n = b\}$ y se plantea el problema de encontrar un spline de grado k, P_k, que interpole a una función en esa partición siendo $y_j = f(t_j)$, $j = 0, n$ los valores a interpolar.

$$y_j = P_k(t_j) = \sum_{i=-r+1}^{n-1} a_i B_i^k(t_j) = \sum_{i=j-r+1}^{j-1} a_i B_i^k(t_j), \quad j = 0, \ldots, n \tag{3.29}$$

Se tienen $(n+k)$ incógnitas, los coeficientes a_i de la expansión y $(n+1)$ ecuaciones. Si $k=1$, el sistema es determinado, y la solución es única como hemos visto. Si $k>1$ hay que imponer condiciones adicionales hasta completar un conjunto completo de condiciones que definan una solución única, como ya vimos anteriormente. En el caso de grado 2, basta con conocer la derivada en un extremo, como se muestra en el problema 3.5, que es una aplicación directa de este apartado.

3.5.10. B-splines de grado 3 en una partición equiespaciada

Se puede usar otra vez la fórmula recurrente para obtener la expresión analítica de cada uno de los cuatro tramos del B-spline cúbico genérico, pero nos limitaremos a dar en la Figura 3.16 sus valores en los nodos. En cada caso, la expresión analítica de cada uno estos 4 tramos se obtiene fácilmente.

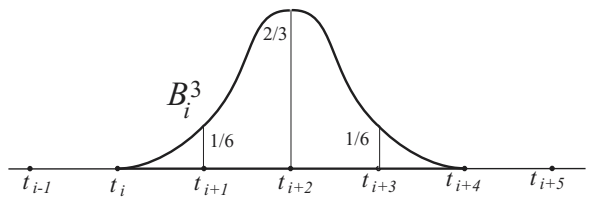


Figura 3.16: B-spline de grado 3.

3.5.11. Partición de la unidad

Se puede demostrar que el sistema de B-splines que hemos definido cumple una propiedad muy importante independientemente de su grado. Esa propiedad es que son una partición de la unidad, lo que significa que:

$$\sum_{i=-r+1}^{n-1} B_i^k(t) = 1 \tag{3.30}$$

Usaremos esta propiedad en alguno de los problemas para simplificar cálculos.

3.6. Interpolación en varias variables

La generalización a varias variables se puede hacer de modo natural teniendo en cuenta la teoría general. Por ejemplo, el espacio E donde se busca la solución al problema puede ser un espacio de polinomios en varias variables, y se le imponen determinadas condiciones a un polinomio que pueden dar lugar o no a una solución única.

3.6.1. Interpolación en recintos rectangulares

Podemos hacer un análisis interesante en dos variables planteando el problema de interpolación de Lagrange en una malla rectangular en cuyos nodos se sabe el valor de la función a interpolar f.

Sea

$$G = \{(x,y) \in \mathbb{R}^2 | a \le x \le b, c \le y \le d\}$$

y sean dos familias de nodos

$$a = x_0 < x_1 < \cdots < x_m = b, \qquad c = y_0 < y_1 < \cdots < y_n = d$$

Esto define $(m+1)(n+1)$ nodos (x_i, y_j) que están dentro del recinto rectangular. El problema de interpolación de Lagrange consiste en buscar un polinomio en dos variables $P \in P_{mn}$ tal que:

$$P(x_i, y_j) = f(x_i, y_j), \quad 0 \le i \le m, 0 \le j \le n$$

Se puede demostrar bastante fácilmente que este problema tiene solución, que esa solución es única, y que se escribe además en función de la base de Lagrange para cada una de las variables y su familia de nodos correspondiente.

$$P(x,y) = \sum_{0 \le i \le m, 0 \le j \le n} f(x_i, y_j) l_i(x) l_j(y) \tag{3.31}$$

Cuando el recinto no es rectangular o las condiciones son más complicadas, el tratamiento de cada problema es diferente.

También existen técnicas para problemas más generales basadas en triangularizaciónes, pero, a pesar de su interés, el tratamiento de las mismas escapan al contenido de este texto. Se recomienda el estudio del problema 3.13, centrado en estas cuestiones.

Ejercicio 3.6.1 *Sea*

$$G = \{(x, y) \in I\!R^{\ 2} | -1 \leq x \leq 1, 0 \leq y \leq 1\}$$

y sean dos familias de nodos

$$x_0 = -1, x_1 = 0, x_2 = 1, \quad y_0 = 0, y_1 = 1$$

Esto define 6 *nodos* (x_i, y_j) *que están dentro del recinto rectangular. Se pide buscar un polinomio en dos variables* $P \in P_{21}$ *tal que:*

$$P(x_i, y_j) = \sin\left(\frac{\pi x_i}{2}\right) \cos\left(\frac{\pi y_j}{2}\right), \quad 0 \leq i \leq 2, \ \ 0 \leq j \leq 1$$

Se pide dibujar la superficie solución en Matlab, y también la superficie original.

PROBLEMAS

PROBLEMA 3.1 *Interpolación trigonométrica.*

Sea E el espacio de dimensión 3 engendrado por las funciones $\{\mathbf{1}, \sin, \cos\}$, $E = L\{\mathbf{1}, \sin, \cos\}$, y deseamos determinar $f \in E$ tal que

$$f(-\pi/2) = y_0, \quad f(0) = y_1, \quad f(\pi/2) = y_2$$

Estudiar si este problema tiene solución y si esa solución es única.

Solución:

Si escribimos la solución buscada en la base que sugiere la forma de definir el espacio E, tendremos que:

$$f(t) = a_0 + a_1 \sin t + a_2 \cos t$$

Planteando las tres condiciones llegamos al siguiente sistema lineal:

$$\begin{pmatrix} 1 & \sin\left(-\frac{\pi}{2}\right) & \cos\left(-\frac{\pi}{2}\right) \\ 1 & \sin(0) & \cos(0) \\ 1 & \sin\left(\frac{\pi}{2}\right) & \cos\left(\frac{\pi}{2}\right) \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} y_0 \\ y_1 \\ y_2 \end{pmatrix}, \quad \begin{pmatrix} 1 & -1 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} y_0 \\ y_1 \\ y_2 \end{pmatrix}$$

cuya solución es única, dado que el determinante de la matriz es -2, distinto de 0.

PROBLEMA 3.2 *Problema de interpolación sin solución.*

Se pide hallar la expresión general del polinomio de segundo grado cuyos valores en ± 1 sean dados y cuya derivada en 0 también lo sea.

Solución:

No todo problema de interpolación es necesariamente resoluble. Si las hipótesis del teorema 3.1.1 no se dan, pueden pasar cosas interesantes. Planteando el problema en la base de los monomios, el sistema lineal tendrá por matriz:

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

cuyo determinante es nulo, y el problema puede no tener solución o puede tener infinitas dependiendo del rango de la matriz ampliada con los valores de la función en ± 1 y con su derivada en 0.

Si pensamos en las infinitas parábolas del tipo $p(x) = k(1 - x^2)$, con k real, esas parábolas tienen el mismo valor nulo en ± 1 y su derivada en cero también vale cero, verificando por tanto las tres condiciones

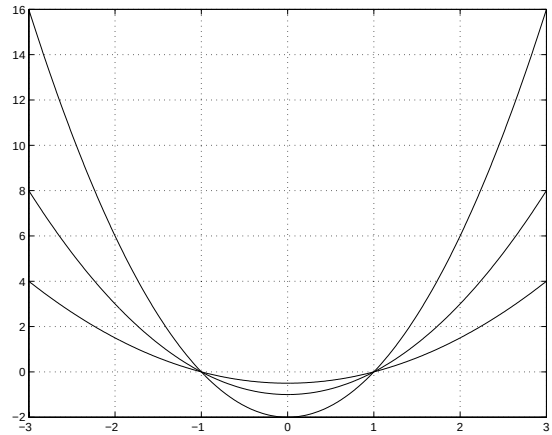


Figura 3.17: Problema 3.2, interpolación con infinitas soluciones.

del problema de interpolación planteado y formando un sistema de infinitas soluciones a un problema de interpolación (Figura 3.17).

Sin embargo, si obligamos a que la derivada en 0 sea -1, no hay ninguna solución.

PROBLEMA 3.3 *Interpolación simple de Hermite.*

Definir la cúbica de Hermite correspondiente a la función $f(x) = \sin(1/x)$ en los puntos 0.15, 0.8, en la base de los monomios, y en la base de diferencias divididas de Newton.

Solución:
La derivada de la función f es $f'(x) = -1/x^2 \cos(1/x)$. Los valores de f y de f' en los puntos dados son

$$f(0.15) = 0.3742,\ f(0.8) = 0.9490,\ f'(0.15) = -41.2163,\ f'(0.8) = -0.4927$$

1. En la base de los monomios, la cúbica solución será:

   $$C(x) = a_0 + a_1 x + a_2 x^2 + a_3 x^3$$

   Imponiendo las 4 condiciones llegamos al siguiente sistema lineal:

   $$\begin{pmatrix} 1.0000 & 0.1500 & 0.0225 & 0.0034 \\ 0.0000 & 1.0000 & 0.3000 & 0.0675 \\ 1.0000 & 0.8000 & 0.6400 & 0.5120 \\ 0.0000 & 1.0000 & 1.6000 & 1.9200 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \end{pmatrix} = \begin{pmatrix} 0.3742 \\ -41.2163 \\ 0.9490 \\ -0.4927 \end{pmatrix}$$

   cuya solución es $(9.8663, -87.6601, 177.9663, -102.9056)$.

   La cúbica buscada es entonces:

   $$C(x) = 9.8663 - 87.6601x + 177.9663x^2 - 102.9056x^3$$

   En la Figura 3.18 podemos observar la función original y la cúbica con la que la hemos interpolado.

2. En la base de diferencias divididas, tenemos que construir la tabla, aprovechando la relación entre derivadas y diferencias divididas de primer orden. Los datos disponibles son:

x_i	f_i	$f[x_i, x_{i+1}]$	$f[x_i, \ldots, x_{i+2}]$	$f[x_0, x_1, x_2, x_3]$
0.15	0.3742			
0.15	0.3742	-41.2163		
0.80	0.9490			
0.80	0.9490	-0.4927		

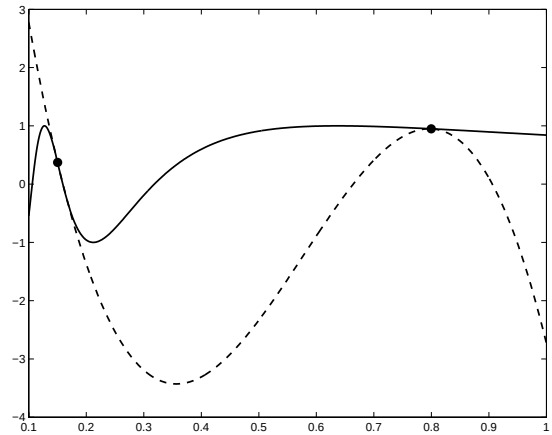


Figura 3.18: Curvas correspondientes al problema 3.3; en trazo discontinuo, el polinomio de Hermite.

A partir de estos valores construimos el resto de la tabla:

x_i	f_i	$f[x_i, x_{i+1}]$	$f[x_i, \ldots, x_{i+2}]$	$f[x_0, x_1, x_2, x_3]$
0.15	0.3742			
0.15	0.3742	-41.2163		
0.80	0.9490	0.8844	64.7702	
0.80	0.9490	-0.4927	-2.1186	-102.9058

Utilizando esta tabla ya podemos construir la expresión de la cúbica correspondiente:

$$C(x) = 0.3742 - 41.2163(x - 0.15) + 64.7702(x - 0.15)^2 - 102.9058(x - 0.15)^2(x - 0.80)$$

En la Figura 3.18 podemos observar la función original y la cúbica con la que la hemos interpolado, que debe ser la misma que la obtenida en el apartado anterior. La comprobación de esto, desarrollando la ecuación de $C(x)$ hasta tenerla en la base de los monomios se deja como ejercicio.

PROBLEMA 3.4 *Interpolación de Hermite a trozos.*

Definir la cúbica a trozos que interpola también la derivada, correspondiente a la función $f(x) = \sin(1/x)$ (la misma que en el problema 3.3) en los puntos 0.15, 0.25, 0.5, 0.8.

Solución:
Esta es la tabla de la función

i	0	1	2	3
x_i	0.15	0.25	0.50	0.80
$f(x_i)$	0.3742	-0.7568	0.9093	0.9490
$f'(x_i)$	-41.2163	10.4583	1.6646	-0.4927

Vamos a obtener cada uno de los tramos, empezando por el primero, $i = 0$, $h_0 = 0.1$, para lo cual usamos la teoría estudiada en la sección 3.3.1.

$$\begin{aligned} C_{0,3}(x) &= \left[1 + \frac{2}{0.1}(x - 0.15)\right]\left(\frac{x - 0.25}{0.1}\right)^2 0.3742 - (x - 0.15)\left(\frac{x - 0.25}{0.1}\right)^2 41.2163 \\ &- \left[1 - \frac{2}{0.1}(x - 0.25)\right]\left(\frac{x - 0.15}{0.1}\right)^2 0.7568 + (x - 0.25)\left(\frac{x - 0.15}{0.1}\right)^2 10.4583 \end{aligned}$$

En el segundo tramo, $i = 1$, $h_1 = 0.25$.

$$\begin{aligned} C_{1,3}(x) &= -\left[1 + \frac{2}{0.25}(x - 0.25)\right]\left(\frac{x - 0.50}{0.25}\right)^2 0.7568 + (x - 0.25)\left(\frac{x - 0.50}{0.25}\right)^2 10.4583 \\ &+ \left[1 - \frac{2}{0.25}(x - 0.50)\right]\left(\frac{x - 0.25}{0.25}\right)^2 0.9093 + (x - 0.50)\left(\frac{x - 0.25}{0.25}\right)^2 1.6646 \end{aligned}$$

En el tercero, $i = 2$, $h_2 = 0.30$.

$$\begin{aligned} C_{2,3}(x) &= \left[1 + \frac{2}{0.30}(x - 0.50)\right]\left(\frac{x - 0.80}{0.30}\right)^2 0.9093 + (x - 0.50)\left(\frac{x - 0.80}{0.30}\right)^2 1.6646 \\ &+ \left[1 - \frac{2}{0.30}(x - 0.80)\right]\left(\frac{x - 0.50}{0.30}\right)^2 0.9490 - (x - 0.80)\left(\frac{x - 0.50}{0.30}\right)^2 0.4927 \end{aligned}$$

El polinomio a trozos queda:

$$C_3(x) = \begin{cases} C_{0,3}(x), & 0.15 \le x < 0.25 \\ C_{1,3}(x), & 0.25 \le x < 0.50 \\ C_{2,3}(x), & 0.50 \le x < 0.80 \end{cases}$$

En la Figura 3.19 podemos observar la función original y la cúbica con la que la hemos interpolado. Es fácil ver que la aproximación es mejor que en el ejemplo correspondiente a una única cúbica (ver problema 3.3) y también es fácil intuir que hemos podido hacer todos estos cálculos gracias a Matlab. Los polinomios a trozos se han constituido como una herramienta de uso común a partir de la generalización de la Informática.

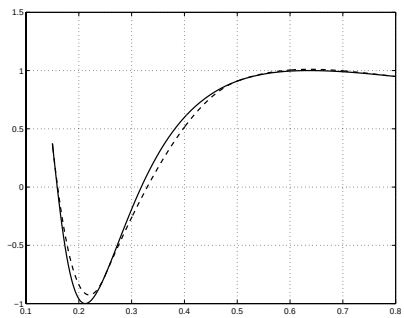


Figura 3.19: Curvas correspondientes al problema 3.4; en trazo discontinuo, la cúbica a trozos.

PROBLEMA 3.5 *Interpolación con B-splines de grado 2.*

Dada la siguiente tabla de valores, de abscisas equiespaciadas, se pide encontrar los coeficientes a_i que permiten escribir de modo único el spline parabólico de interpolación de esa tabla como combinación lineal de la base de B-splines de grado 2 estudiada.

t_0	t_1	t_2	t_3	t_4	t_4
y_0	y_1	y_2	y_3	y_4	y_4'

Solución:
Escribamos el spline solución P_2 en la base de los B-splines:

$$P_2(t) = \sum_{i=-2}^{3} a_i B_i^2(t)$$

Impongamos las diferentes condiciones, teniendo en cuenta la longitud del soporte de cada uno de los elementos de la base (ver Figura 3.15)

$$\begin{aligned}
y_0 = P_2(t_0) &= \sum_{i=-2}^{3} a_i B_i^2(t_0) = 0.5a_{-2} + 0.5a_{-1} \\
y_1 = P_2(t_1) &= \sum_{i=-2}^{3} a_i B_i^2(t_1) = 0.5a_{-1} + 0.5a_0 \\
y_2 = P_2(t_2) &= \sum_{i=-2}^{3} a_i B_i^2(t_2) = 0.5a_0 + 0.5a_1 \\
y_3 = P_2(t_3) &= \sum_{i=-2}^{3} a_i B_i^2(t_3) = 0.5a_1 + 0.5a_2 \\
y_4 = P_2(t_4) &= \sum_{i=-2}^{3} a_i B_i^2(t_4) = 0.5a_2 + 0.5a_3
\end{aligned}$$

Impongamos la condición adicional de derivada

$$y_4' = P_2'(t_4) = \sum_{i=-2}^{3} a_i \left(B_i^2(t_4)\right)' = a_2 \left(B_2^2(t_4)\right)' + a_3 \left(B_2^2(t_3)\right)'$$

El primer tramo del soporte y su derivada correspondiente particularizada en el nodo intermedio valen:

$$\begin{aligned}
B_i^2(t) &= \frac{1}{2h^2}(t - t_i)^2 \\
\left(B_i^2(t)\right)' = \frac{1}{h^2}(t - t_i) &\Rightarrow \left(B_i^2(t_{i+1})\right)' = 1/h
\end{aligned}$$

Por simetría, la derivada en el otro nodo vale lo mismo pero con distinto signo. Por tanto, volviendo a la condición de derivada, tendremos:

$$y_4' = -a_2 \frac{1}{h} + a_3 \frac{1}{h}$$

Y tenemos el siguiente sistema lineal:

$$\begin{pmatrix}
0.5 & 0.5 & 0 & 0 & 0 & 0 \\
0 & 0.5 & 0.5 & 0 & 0 & 0 \\
0 & 0 & 0.5 & 0.5 & 0 & 0 \\
0 & 0 & 0 & 0.5 & 0.5 & 0 \\
0 & 0 & 0 & 0 & 0.5 & 0.5 \\
0 & 0 & 0 & 0 & -1/h & 1/h
\end{pmatrix}
\begin{pmatrix} a_{-2} \\ a_{-1} \\ a_0 \\ a_1 \\ a_2 \\ a_3 \end{pmatrix} =
\begin{pmatrix} y_0 \\ y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_4' \end{pmatrix}$$

cuya solución es:

$$\begin{aligned}
a_{-2} &= 2(y_0 - y_1 + y_2 - y_3) + y_4 - hy_4'/2 \\
a_{-1} &= 2(y_1 - y_2 + y_3) - y_4 + hy_4'/2 \\
a_0 &= 2(y_2 - y_3) + y_4 - hy_4'/2 \\
a_1 &= 2y_3 - y_4 + hy_4'/2 \\
a_2 &= y_4 - hy_4'/2 \\
a_3 &= y_4 + hy_4'/2
\end{aligned}$$

PROBLEMA 3.6 *Bases de splines asociadas a un problema de interpolación.*

Se considera el problema de interpolación consistente en buscar un spline cúbico asociado a la partición $\Omega = \{0.20, 0.40, 0.60, 0.80\}$ del intervalo $[0.20, 0.80]$, que interpole la función en esos nodos y que interpole también las derivadas en el primer y último nodo. Se pide calcular el primer elemento de la base dual asociada a este problema de interpolación, identificándolo mediante su valor y el de su derivada en todos los nodos. Dibujar un esquema del mismo.

Solución:
Acondicionemos el sistema lineal 3.20 (pág. 136) a este problema:

$$\begin{pmatrix} 4 & 1 \\ 1 & 4 \end{pmatrix} \begin{pmatrix} s_1 \\ s_2 \end{pmatrix} = \frac{3}{h} \begin{pmatrix} y_2 - y_0 - \frac{h}{3} s_0 \\ y_3 - y_1 - \frac{h}{3} s_3 \end{pmatrix}$$

El primer elemento c_0 de la base dual, vale 0 en todas las formas lineales menos en la primera, que valdrá 1. Por tanto, como las 6 formas lineales son

$$\begin{aligned} L_0(S) &= S(0.2), & L_1(S) &= S(0.4) \\ L_2(S) &= S(0.6), & L_3(S) &= S(0.8) \\ L_4(S) &= S'(0.2), & L_5(S) &= S'(0.8) \end{aligned}$$

tendremos

$$\begin{aligned} L_0(c_0) &= c_0(0.2) = 1, & L_1(c_0) &= c_0(0.4) = 0 \\ L_2(c_0) &= c_0(0.6) = 0, & L_3(c_0) &= c_0(0.8) = 0 \\ L_4(c_0) &= c_0'(0.2) = 0, & L_5(c_0) &= c_0'(0.8) = 0 \end{aligned}$$

que, traducido a la notación del sistema lineal, significa (con $h = 0.2$)

$$\begin{aligned} y_0 &= 1, & y_1 &= 0 \\ y_2 &= 0, & y_3 &= 0 \\ s_0 &= 0, & s_3 &= 0 \end{aligned}$$

$$\begin{pmatrix} 4 & 1 \\ 1 & 4 \end{pmatrix} \begin{pmatrix} s_1 \\ s_2 \end{pmatrix} = \begin{pmatrix} -15 \\ 0 \end{pmatrix}$$

cuya solución es $s_1 = -4$ y $s_2 = 1$, valores coherentes con el esquema de la Figura 3.20. Se deja como

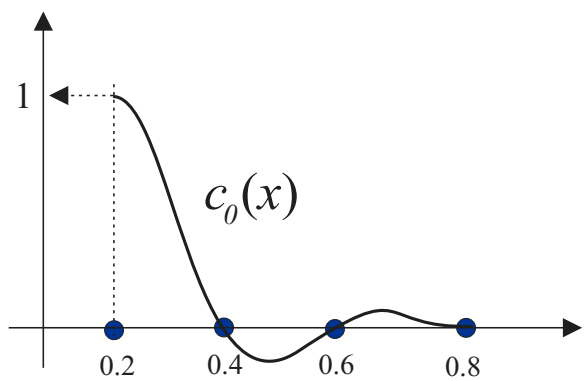


Figura 3.20: Esquema de la solución del problema 3.6.

ejercicio calcular $c_0(0.45)$ y comprobar si el valor obtenido es compatible con el esquema del mismo que hemos presentado en la Figura 3.20.

PROBLEMA 3.7 *Splines de segundo grado.*

Se consideran la partición $\Omega = \{-1, 0, 1, 2\}$ del intervalo compacto $[-1, 2]$ y la función a trozos definida en $[-1, 2]$ por

$$h(x) = \begin{cases} x^2 - 1 & x \in [-1, 0] \\ 2x^2 - 1 & x \in (0, 1] \\ -2x^2 + 8x - 5 & x \in (1, 2] \end{cases}$$

1. Comprobar que $h \in S_2(\Omega)$, espacio vectorial de los splines de segundo grado asociados a Ω.
2. Dibujar de modo esquemático los elementos de la base de B-Splines, asociados a Ω.
3. Dar las componentes de h respecto de la base de B-Splines.
4. Definir un vector vh resultado de "muestrear" la función h en los puntos $-1, -0.5, 0.0, 0.5, 1.0, 1.5, 2.0$.
5. Estimar
$$\int_{-1}^{2} h(x)dx$$
mediante el método compuesto de los trapecios tomando un paso de 0.5 unidades.
6. Calcular esa integral de modo exacto. Comprobar si se verifica la cota de error asociada al método compuesto de los trapecios.

Solución:

1. h cumple las condiciones necesarias para pertenecer a $S_2(\Omega)$, ya que
   - Su restricción a cada uno de los tramos es un polinomio de segundo grado;
   - En los nodos interiores, 0 y 1, los tramos enganchan con continuidad
   $$s(0)^- = s(0)^+ = -1 \qquad \text{y} \qquad s(1)^- = s(1)^+ = 1$$
   - En los nodos interiores, 0 y 1, los tramos enganchan con continuidad en la derivada. Se pueden utilizar estas líneas Matlab para dibujar la función h.

```
x=-1:0.01:2;
h=(x<=0).*(x.^2-1)+(x>0).*(x<=1).*(2*x.^2-1)+(x>1).*(-2*x.^2+8*x-5);
plot(x,h);
grid;
```

y obtendremos la Figura 3.21.

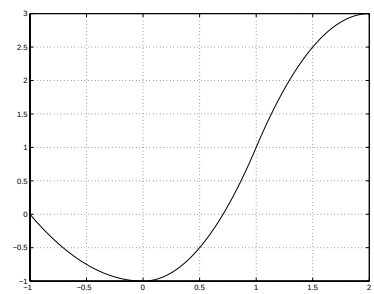


Figura 3.21: Gráfica de $h(x)$.

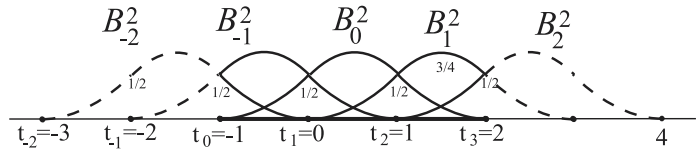


Figura 3.22: Elementos de la base de B-splines de $S_2(\Omega)$.

2. Ω define tres subintervalos, por tanto la dimensión del espacio $S_2(\Omega)$ es 5. Tenemos elementos de base que "nacen" en los nodos t_0, t_1 y t_2 y dos elementos adicionales que parten de los nodos t_{-1} y t_{-2}, ver Figura 3.40.

3. Se resuelve el problema de modo sencillo planteando en la nueva base un problema de interpolación que tenga solución única y dando a h los valores correspondientes a ese problema. Se puede, por ejemplo, suponer conocidos los valores en los nodos y la derivada en el primer nodo. Los valores en los nodos de h son 0, -1, 1 y 3. La derivada en el primer nodo es -2. Así, si nuestro spline se escribe de modo único en la forma

$$h(t) = \sum_{j=-2}^{2} a_j B_j(t)$$

tendremos que:

$$\begin{pmatrix} 1/2 & 1/2 & 0 & 0 & 0 \\ 0 & 1/2 & 1/2 & 0 & 0 \\ 0 & 0 & 1/2 & 1/2 & 0 \\ 0 & 0 & 0 & 1/2 & 1/2 \\ -1 & 1 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} a_{-2} \\ a_{-1} \\ a_0 \\ a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} 0 \\ -1 \\ 1 \\ 3 \\ -2 \end{pmatrix}$$

sistema lineal que nos da las componentes buscadas:

$$\begin{pmatrix} a_{-2} \\ a_{-1} \\ a_0 \\ a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} 1 \\ -1 \\ -1 \\ 3 \\ 3 \end{pmatrix}$$

4. $vh^T = (0.0000, -0.7500, -1.0000, -0.5000, 1.0000, 2.5000, 3.0000)$.

5. $I = 1.3750$.

6. La cota del error para el método compuesto de los trapecios (ver Theodor[29]):

$$Error(h) \leq \frac{h^2}{12}(b-a) \max_{x\in[-1,2]} |h''(x)| = \frac{0.5^2}{12} \cdot 3 \cdot 4 = 0.25$$

Como la integral real es 1.3333, tendremos que el $Error(0.5) = 1.3750 - 1.3333 = 0.0417$, que es menor que la cota, 0.25, como debía suceder.

PROBLEMA 3.8 *Splines de grado 1.*

Sea $\Omega = \{0.20, 0.40, 0.60, 0.80\}$ una partición del intervalo $[0.2, 0.8]$. Se plantea el problema de encontrar un spline de grado 1 que interpole a la función $f(x) = \sin(1/x)$ en todos los puntos de esa partición menos en el último, y que ajuste el área de f entre 0.2 y 0.8. Se sabe que el valor de este área es $A = 0.3266$.

1. Demostrar que este problema tiene solución y que ésta es única.

2. Calcular el valor del spline solución para $t = 0.8$.

3. Dibujar un esquema de los elementos de la base de $S_1(\Omega)$ dual de la base que este problema de interpolación define en $S_1(\Omega)^*$, indicando el valor en los nodos cuando este valor no sea nulo.

4. Descomponer el spline del apartado 1 respecto de esta base.

Sea ahora $[a, b]$ un intervalo compacto cualquiera y Ω una partición equiespaciada del mismo, $\Omega = \{t_0 = a, t_1, \ldots, t_{n-1}, t_n = b\}$. Se plantea el problema de encontrar un spline de grado 1 que interpole a una función f dada en todos los puntos de la partición menos en el último. Sean $f_j = f(t_j)$, $j = 0, n-1$ los valores a interpolar. Se busca también que el spline solución tenga una integral entre t_0 y t_n de valor A.

5. Demostrar que este problema tiene solución única.

6. Calcular el valor del spline solución para $t = (t_i + t_{i+1})/2$, $i = 0, n-2$.

7. Calcular el valor del spline solución para $t = (t_{n-1} + t_n)/2$.

Solución:

1. Si evalúamos la función f en la nube dada, menos el último punto, tendremos:

```
>>t=0.2:0.2:0.6;
>>f=sin(1./t)
f = -0.9589    0.5985    0.9954
```

   La poligonal que une los puntos $(t_i, f_i)_{i=0,2}$ es única. Bastará ajustar el valor f_3 en el último punto para que el área sea la pedida. El área del hipotético spline solución P_1 será:

$$A = \int_{0.2}^{0.8} P_1(t)dt = \frac{0.2}{2}(f_0 + 2f_1 + 2f_2 + f_3)$$

   De aquí deducimos que

$$f_3 = \frac{A}{0.1} - f_0 - 2f_1 - 2f_2$$

   Por tanto ese valor es único, así como el spline buscado.

2. Su valor en el último nodo es:

$$f_3 = 10A - f_0 - 2f_1 - 2f_2 = 1.0372$$

```
>>A=0.3266;
>>f(4)=10*A-f(1)-2f(2)-2f(3);
>>f(4)
ans = 1.0372
```

   Hay que tener cuidado con los índices de los vectores, que en Matlab empiezan en 1. Podemos representar la función original y el spline solución (Figura 3.23) con las siguientes líneas:

```
t=0.2:0.2:0.8;
P_1=[-0.9589    0.5985    0.9954    1.0372];
tt=0.2:0.001:0.8;
ff=sin(1./tt);
plot(tt,ff,t,P_1);
axis([0.15 0.85 -1.2 1.2]);
shg;
```

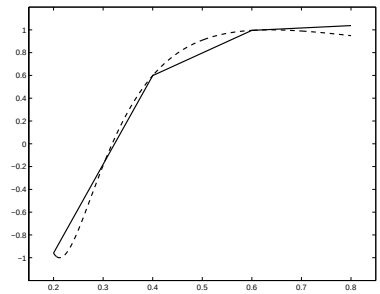


Figura 3.23: Solución del apartado 1 del problema 3.8.

3. Las 4 formas líneales con las que estamos definiendo el problema son:

$$L_0(P) = P(0.2), \quad L_1(P) = P(0.4), \quad L_2(P) = P(0.6), \quad L_3(P) = \int_{0.2}^{0.8} P(t)dt$$

Los elementos de la base dual e_i $i = 0, 1, 2, 3$ de $S_1(\Omega)$ vienen definidos sucesivamente por

$$\begin{array}{llll} e_0(0.2) = 1, & e_0(0.4) = 0, & e_0(0.6) = 0, & \int_{0.2}^{0.8} e_0(t)dt = 0 \\ e_1(0.2) = 0, & e_1(0.4) = 1, & e_1(0.6) = 0, & \int_{0.2}^{0.8} e_1(t)dt = 0 \\ e_2(0.2) = 0, & e_2(0.4) = 0, & e_2(0.6) = 1, & \int_{0.2}^{0.8} e_2(t)dt = 0 \\ e_3(0.2) = 0, & e_3(0.4) = 0, & e_3(0.6) = 0, & \int_{0.2}^{0.8} e_3(t)dt = 1 \end{array}$$

El valor en 0.8 de todos los elementos, se obtiene con la expresión obtenida en el apartado 2. Sus gráficas las representamos en la Figura 3.24.

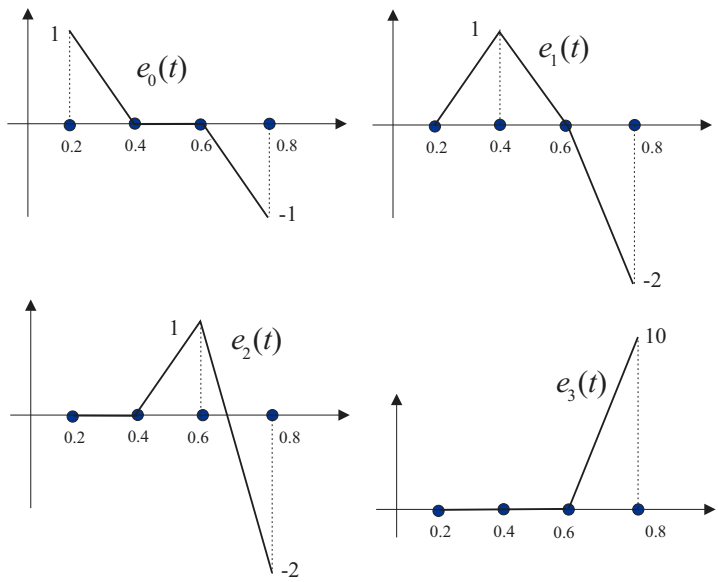


Figura 3.24: Solución del apartado 3 del problema 3.8.

4. En esta base, la solución del problema viene dada por los valores que esperamos tome la función en los nodos y por su área:

$$P_1(t) = -0.9589e_0(t) + 0.5985e_1(t) + 0.9954e_2(t) + 0.3266e_3(t)$$

Se deja como ejercicio comprobar que el valor que se obtiene para $t = 0.8$ es el adecuado.

5. Con un número arbitrario de nodos, el problema se plantea de modo similar. La solución es única porque la poligonal está completamente definida hasta el penúltimo nodo, y el valor del área nos permite ajustar el valor del último. Llamemos h al paso y P_1 al spline buscado:

$$A = \int_{t_0}^{t_n} P_1(t)dt = \frac{h}{2}\left(f_0 + 2\sum_{i=1}^{n-1} f_i + f_n\right)$$

De aquí obtenemos f_n:

$$f_n = \frac{2A}{h} - \left(f_0 + 2\sum_{i=1}^{n-1} f_i\right)$$

6.

$$P_1\left(\frac{t_i + t_{i+1}}{2}\right) = \frac{f_i + f_{i+1}}{2} \quad i = 0, n-2$$

7. Con el valor de f_n obtenido en un apartado anterior:

$$P_1\left(\frac{t_n + t_{n-1}}{2}\right) = \frac{f_n + f_{n-1}}{2}$$

PROBLEMA 3.9 *Interpolación no lineal.*

Se trata de encontrar una función de la forma $f(x) = ax^2 + e^{bx} + cos(cx)$ que interpole los puntos $(1, 1.6756)$, $(2, 3.6022)$, $(3, 8.0125)$. Se pide:

1. Escribir el problema no lineal resultante.
2. Utilizar el método de Newton-Raphson (NR) con estimador inicial $(a_0, b_0, c_0)^T = (0.5, -0.5, 0.5)^T$ para aproximar la solución del problema del apartado anterior. Dar un paso con dicho método.
3. En el caso de que se quisiese resolver el sistema lineal correspondiente a la primera iteración de NR, mediante Jacobi o Gauss-Seidel, estudiar su convergencia y decidir cuál de ellos tendría una convergencia más rápida.

Solución:

1. Si obligamos a que la función pase por esos puntos llegamos al sistema no lineal de tres ecuaciones con tres incógnitas.

$$\begin{aligned} a + e^b + \cos(c) &= 1.6756 \\ 4a + e^{2b} + \cos(2c) &= 3.6022 \\ 9a + e^{3b} + \cos(3c) &= 8.0125 \end{aligned}$$

2. Es un problema de interpolación no lineal. Para resolverlo, debemos recurrir a técnicas de resolución de sistemas no lineales, y una de las más poderosas es el método de Newton-Raphson. Para utilizarlo, ponemos nuestro problema en la forma $F(a, b, c) = 0$ con

$$F(a,b,c) = \begin{pmatrix} a + e^b + \cos(c) - 1.6756 \\ 4a + e^{2b} + \cos(2c) - 3.6022 \\ 9a + e^{3b} + \cos(3c) - 8.0125 \end{pmatrix}$$

El proceso iterativo de NR exige evaluar el jacobiano J_F de F y resolver en cada paso el sistema lineal

$$\begin{aligned} J_F(x_n)\Delta x &= -F(x_n) \\ x_{n+1} &= x_n + \Delta x \end{aligned}$$

con $x = (a, b, c)^T$.

$$J_F(a,b,c) = \begin{pmatrix} 1 & e^b & -\sin(c) \\ 4 & 2e^{2b} & -2\sin(2c) \\ 9 & 3e^{3b} & -3\sin(3c) \end{pmatrix}$$

En el primer paso,

$$J_F(a,b,c)^{(0)} = \mathbf{F}(0.5,-0.5,0.5) = \begin{pmatrix} 1 & 0.6065 & -0.4794 \\ 4 & 0.7358 & -1.6829 \\ 9 & 0.6694 & -2.9925 \end{pmatrix} \text{ y } F(a,b,c)^{(0)} = \begin{pmatrix} 0.3085 \\ -0.6940 \\ -3.2186 \end{pmatrix}$$

que definen la corrección $\Delta x = (0.7034, -1.0281, 0.8107)^T$. Por tanto,

$$\begin{pmatrix} a \\ b \\ c \end{pmatrix}^{(1)} = \begin{pmatrix} a \\ b \\ c \end{pmatrix}^{(0)} + \begin{pmatrix} 0.7034 \\ -1.0281 \\ 0.8107 \end{pmatrix} = \begin{pmatrix} 1.2034 \\ -1.5281 \\ 1.3101 \end{pmatrix}$$

Podemos utilizar estas líneas Matlab para realizar estos cálculos.

```
a=0.5;
b=-0.5;
c=0.5;
J=[1 exp(b) -sin(c)
   4 2*exp(2*b) -2*sin(2*c)
   9 3*exp(3*b) -3*sin(3*c)];
F=[a+exp(b)+cos(c)-1.6756
   4*a+exp(2*b)+cos(2*c)-3.6022
   9*a+exp(3*b)+cos(3*c)-8.0125];
delta=-J\F;
a=a+delta(1)
b=b+delta(2)
c=c+delta(3)
```

Usaremos un gráfico para ver el grado de aproximación de esta estimación de (a, b, c) comparada con la inicial y también respecto a la nube de puntos a interpolar (ver Figura 3.25). Para ello utilizamos las líneas Matlab

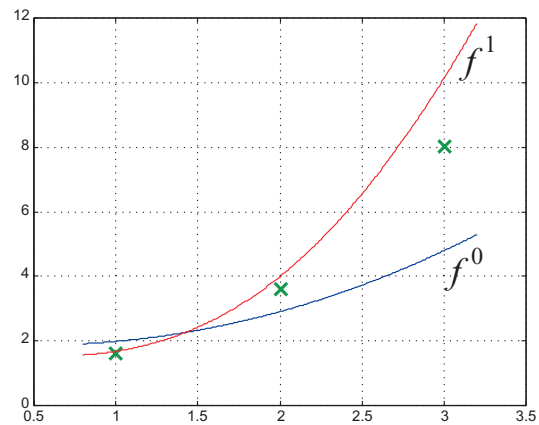


Figura 3.25: Curvas correspondientes al problema 3.9.

```
t=1:1:3;
ft=[1.6756 3.6022 8.0125];
x=0.8:0.01:3.2;
a=0.5; b=-0.5; c=0.5;
f=a*x.^2+exp(b*x)+cos(c*x);
```

a las que añadimos, después de haber calculado la primera iteración, las siguientes

```
f1=a*x.^2+exp(b*x)+cos(c*x);
plot(x,f,t,ft,'x',x,f1);
```

Se observa en la gráfica como la primera iteración mejora la estimación inicial de un modo sensible. Se deja como ejercicio realizar una iteración adicional, y evaluar el residuo del estimador inicial y de la primera iteración para valorar de este modo la convergencia.

3. Se quiere aquí resolver el sistema lineal de matriz J_F mediante un método iterativo. Lo primero que se observa es que esa matriz no tiene una estructura diagonalmente dominante. ¡Esto nos debe poner en guardia! Tenemos que estudiar el espectro de la matriz de iteración para decidir sobre la convergencia. En el caso de Jacobi, la matriz de iteración correspondiente al primer paso del NR vale

$$B = M^{-1}N = D^{-1}(L+U) = \begin{pmatrix} 0 & -0.2169 & 0.9662 \\ -42.4953 & 0 & 10.5848 \\ -4.2292 & -0.0144 & 0 \end{pmatrix}$$

cuyo radio espectral es $\rho(B) = 2.9175$ que es mayor que uno, y por lo tanto, no se da la convergencia. Si repetimos este proceso con Gauss-Seidel (GS), tenemos

$$B = M^{-1}N = (D-L)^{-1}U = \begin{pmatrix} 0 & -0.2169 & 0.9662 \\ 0 & 9.2190 & -30.4741 \\ 0 & 0.7848 & -3.6476 \end{pmatrix}$$

y cuyo radio espectral es $\rho(B) = 6.9656$, que también es mayor que uno. No hay convergencia con este método iterativo.

PROBLEMA 3.10 *Base de las parábolas.*

Se plantea el problema de encontrar una parábola p tal que:

$$p(-1) = -1.9 \quad p(1) = 2.3 \quad p'(1) = -1.1$$

1. Demostrar que este problema tiene solución única, dando su solución en la base de los monomios B.
2. Encontrar la base B' de $P_2(\mathbb{R})$ dual de la que en $P_2(\mathbb{R})^*$ define el problema de interpolación planteado, expresando cada uno de sus elementos en la base B.
3. Hacer un dibujo esquemático de la gráfica de los elementos de B'.
4. ¿Cuál es la matriz de cambio de la base B de los monomios a la B'?
5. Utilizar esa matriz para encontrar la expresión del polinomio solución de 1 en la base B'.

Solución:

1. Para demostrar que el problema tiene solución única, escribamos la hipotética solución en la base de los monomios $p(x) = a_0 + a_1x + a_2x^2$ y apliquémosle las condiciones:

$$\begin{pmatrix} 1 & -1 & 1 \\ 1 & 1 & 1 \\ 0 & 1 & 2 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} -1.9 \\ 2.3 \\ -1.1 \end{pmatrix}, \quad \begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} 1.8000 \\ 2.1000 \\ -1.6000 \end{pmatrix}$$

El determinante del sistema lineal vale 4, y por tanto, el problema de interpolación tiene solución única.

2. Al plantear el problema de interpolación, hemos definido tres formas lineales:

$$\begin{aligned} L_i : P_2(\mathbb{R}) &\rightarrow \mathbb{R}, \quad i = 1,2,3 \\ L_1(p) &:= p(-1) \\ L_2(p) &:= p(1) \\ L_3(p) &:= p'(1) \end{aligned}$$

Demostrar que el problema de interpolación tiene solución única, equivale a demostrar la independencia lineal de estas tres formas lineales, que definen por tanto una base del dual de $P_2(\mathbb{R})$. Es sensato por tanto buscar su base dual, descrita por aquellos polinomios l_j tales que $L_i(l_j) = \delta_i^j$.

Si definimos cada uno de estos elementos en la base de los monomios como:

$$l_j(x) = l_j^0 + l_j^1 x + l_j^2 x^2, \quad j = 1,2,3$$

las componentes de estos polinomios en la base de los monomios se obtienen resolviendo los tres sistemas lineales:

$$\begin{pmatrix} 1 & -1 & 1 \\ 1 & 1 & 1 \\ 0 & 1 & 2 \end{pmatrix} \begin{pmatrix} l_1^0 \\ l_1^1 \\ l_1^2 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \quad \begin{pmatrix} 1 & -1 & 1 \\ 1 & 1 & 1 \\ 0 & 1 & 2 \end{pmatrix} \begin{pmatrix} l_2^0 \\ l_2^1 \\ l_2^2 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$$

$$\begin{pmatrix} 1 & -1 & 1 \\ 1 & 1 & 1 \\ 0 & 1 & 2 \end{pmatrix} \begin{pmatrix} l_3^0 \\ l_3^1 \\ l_3^2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$$

Y por tanto, tendremos

$$\begin{pmatrix} l_1^0 & l_2^0 & l_3^0 \\ l_1^1 & l_2^1 & l_3^1 \\ l_1^2 & l_2^2 & l_3^2 \end{pmatrix} = \begin{pmatrix} 1 & -1 & 1 \\ 1 & 1 & 1 \\ 0 & 1 & 2 \end{pmatrix}^{-1} = \begin{pmatrix} 0.2500 & 0.7500 & -0.5000 \\ -0.5000 & 0.5000 & 0.0000 \\ 0.2500 & -0.2500 & 0.5000 \end{pmatrix} =: M$$

$$\begin{aligned} l_1(x) &= 0.2500 - 0.5000x + 0.2500x^2 \\ l_2(x) &= 0.7500 + 0.5000x - 0.2500x^2 \\ l_3(x) &= -0.5000 + 0.5000x^2 \end{aligned}$$

3. A partir de las condiciones que deben verificar los elementos de la base, es sencillo hacer un esquema de su gráfica (Figura 3.26).

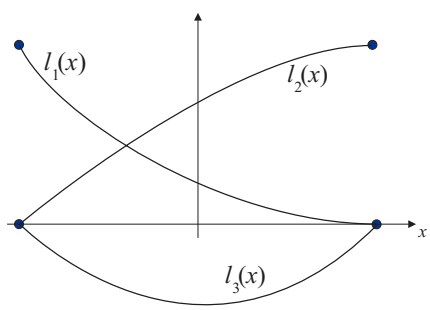


Figura 3.26: Gráfica de los elementos de la base B'.

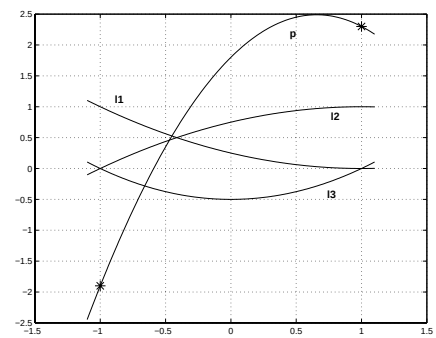


Figura 3.27: Gráfica de los elementos de la base B'.

4. En la matriz M tenemos como columnas las componentes de los elementos de la base B' en la base de los monomios, luego es la matriz de cambio de base buscada.

5. Premultiplicando un vector en la base de los monomios por la matriz M^{-1} se obtiene el vector expresado en la base B'. En efecto,

$$M^{-1} = \begin{pmatrix} 1 & -1 & 1 \\ 1 & 1 & 1 \\ 0 & 1 & 2 \end{pmatrix} \quad \text{y} \quad M^{-1} \begin{pmatrix} 1.8000 \\ 2.1000 \\ -1.6000 \end{pmatrix} = \begin{pmatrix} -1.9 \\ 2.3 \\ -1.1 \end{pmatrix}$$

Y las componentes en B' del polinomio solución son precisamente $p(-1)$, $p(1)$ y $p'(1)$, como debía suceder por la dualidad de las bases. En la figura representamos los tres elementos de la base B', el polinomio solución p y los dos puntos por los que tiene que pasar, los correspondientes a -1 y 1. También se observa que la pendiente es la correcta en 1, o sea, -1.1

$$p(x) = 1.800 + 2.1000x - 1.6000x^2 = -1.9l_1(x) + 2.3l_2(x) - 1.1l_3(x)$$

Presentamos el código Matlab que permite obtener este gráfico:

```
x=-1.1:0.01:1.1
xaux=[-1 1];
yaux=[-1.9 2.3]
l1 = 0.2500 -0.5000* x +0.2500* x.^2
l2 = 0.7500 +0.5000* x -0.2500 *x.^2
l3 = -0.5000  +0.5000 *x.^2
p=-1.9*l1+2.3*l2-1.1*l3;
plot(x,l1,x,l2,x,l3,x,p,xaux,yaux,'*');
grid;
shg;
```

PROBLEMA 3.11 *Polinomios a trozos.*

Se considera el problema de interpolación consistente en determinar un polinomio de segundo grado P_i, del cual conozcamos los siguientes datos:

$$\begin{aligned} P_i(x_i) &= w_i \\ P_i(x_{i+1}) &= w_{i+1} \\ P_i'(x_i) &= s_i \end{aligned}$$

con $x_i < x_{i+1}$.

1. Demostrar que el problema admite solución única.
2. Encontrar la expresión general de ese polinomio en función de los valores x_i, x_{i+1}, w_i, w_{i+1}, s_i.
3. Se extiende el problema para suponer que tenemos un polinomio a trozos de grado 2 que interpola la nube de puntos (x_i, w_i), $i = 0, n$ con continuidad en la función y en su derivada, o sea, un spline cuadrático. Se pide deducir el algoritmo que proporcione splines cuadráticos que interpolen esos datos (x_i, w_i), $i = 0, n$, a través de la definición de cada uno de sus tramos con la formulación del apartado anterior.

Solución:

1. El espacio en el que buscamos la solución es $E = P_2(\mathbb{R})$ cuya base canónica esta formada por los monomios $\{1, x, x^2\}$. Escribamos el determinante de Gramm asociado a las tres formas lineales que definen nuestro problema de interpolación (ver Linz[20], Capítulo 2).

$$\begin{aligned} L_0(P_i) &= P_i(x_i) \\ L_1(P_i) &= P_i(x_{i+1}) \\ L_2(P_i) &= P_i'(x_i) \end{aligned}$$

$$\begin{aligned} det\left(\langle L_i, x^j\rangle\right) &= \begin{vmatrix} L_0(1) & L_0(x) & L_0\left(x^2\right) \\ L_1(1) & L_1(x) & L_1\left(x^2\right) \\ L_2(1) & L_2(x) & L_2\left(x^2\right) \end{vmatrix} = \begin{vmatrix} 1 & x_i & x_i^2 \\ 1 & x_{i+1} & x_{i+1}^2 \\ 0 & 1 & 2x_i \end{vmatrix} \\ &= -\left(x_{i+1}-x_i\right)^2 \neq 0, \quad \text{pues} \quad x_{i+1} \neq x_i \end{aligned}$$

Por tanto, el problema tiene solución única.

2. Se puede hallar la expresión general de ese polinomio en función de esos valores de varias maneras. La más simple, pero más liosa, es plantear el sistema lineal en la base del apartado anterior y resolverlo. Vamos a hacerlo de otras dos formas. En la primera de ellas suponemos que la parábola P_i tiene segunda derivada constante z_i en $[x_i, x_{i+1}]$.

$$P_i''(x) = z_i$$

Integrando entre x_i y un valor cualquiera x,

$$P_i'(x) - P_i'(x_i) = z_i(x - x_i) \rightarrow P_i'(x) = s_i + z_i(x - x_i)$$

Integrando de nuevo,

$$P_i(x) - P_i(x_i) = s_i(x - x_i) + z_i\frac{(x - x_i)^2}{2}$$

que es lo mismo que:

$$P_i(x) = w_i + s_i(x - x_i) + z_i\frac{(x - x_i)^2}{2}$$

Para despejar z_i ya que conocemos w_{i+1}

$$P_i(x_{i+1}) = w_{i+1} = w_i + s_i h_i + z_i\frac{h_i^2}{2}, \quad \text{con} \quad h_i = x_{i+1} - x_i$$

Entrando en 2 con el valor de z_i obtenido de la ecuación anterior tenemos

$$P_i(x) = w_i + s_i(w - w_i) + \frac{w_{i+1} - w_i - s_i h_i}{h_i^2}(x - x_i)^2$$

Otra forma de hacerlo, más rápida, es usar diferencias divididas de Newton.

Construyamos la tabla de diferencias divididas

x_i	w_i		
x_i	w_i	s_i	
x_{i+1}	w_{i+1}	$f[x_i, x_{i+1}]$	$f[x_i, x_i, x_{i+1}]$

con

$$f[x_i, x_{i+1}] = \frac{w_{i+1} - w_i}{x_{i+1} - x_i}$$

$$f[x_i, x_i, x_{i+1}] = \frac{f[x_i, x_{i+1}] - s_i}{x_{i+1} - x_i}$$

A partir de esta tabla construimos la parábola en la base de las diferencias divididas.

$$\begin{aligned} P_i(x) &= w_i + s_i(x - x_i) + f[x_i, x_i, x_{i+1}](x - x_i)^2 \\ &= w_i + s_i(x - x_i) + \frac{w_{i+1} - w_i - s_i h_i}{h_i^2}(x - x_i)^2 \end{aligned}$$

3. Ya que es un spline parabólico, su derivada debe ser continua en los nodos donde enganchan los diferentes tramos; esto lo podemos expresar como

$$P_i'(x_{i+1}) = P_{i+1}'(x_{i+1}) \quad i = 0, n-2$$

Con la expresión que hemos obtenido de una parábola a partir de los valores de la misma en los nodos y la derivada en el primero, es fácil plantear esa igualdad, en todos los nodos interiores

$$\begin{aligned} P_i'(x_{i+1}) &= s_i + 2\frac{w_{i+1} - w_i - s_i h_i}{h_i} \\ P_{i+1}'(x_{i+1}) &= s_{i+1} \end{aligned}$$

De donde

$$s_i + s_{i+1} = 2\frac{w_{i+1} - w_i}{h_i} \quad i = 0, n-2$$

Por tanto, conociendo la derivada en cualquiera de los nodos, se obtienen todas las demás sucesivamente mediante esa fórmula, y una vez conocidas las derivadas y los valores en los nodos, quedan definidas todas las parábolas que forman el spline en cuestión.

PROBLEMA 3.12 *Splines con una condición adicional de área.*

Se considera la nube de puntos siguiente:

i	x_i	f_i	f_i'
0	0.0	1.0000	0.0000
1	0.1	0.9741	?
2	0.2	0.9041	?
3	0.3	0.7828	−1.2843

1. Calcular las derivadas en los dos nodos interiores ajustando esa nube mediante un spline cúbico asociado a la partición definida por el conjunto de abscisas de dichos nodos y que ajuste también las derivadas en el primer y último nodo.

2. Calcular el área de la curva definida por esa nube mediante el método compuesto de los trapecios.

3. Se considera la misma nube de puntos pero no se considera conocida ninguna de las derivadas. Estudiar la existencia y unicidad de un spline de grado 2 asociado a la misma partición y cuya área sea la del apartado anterior. Caso de que exista y sea único, se pide calcular su expresión en la base de los B-Splines asociada a esa partición obteniendo finalmente sus derivadas en todos los nodos.

Solución:

1. Las derivadas en los nodos en un spline cúbico se relacionan con los valores en los nodos mediante un sistema lineal que las tiene como incógnitas (ver Linz[20], Capítulo 2).

$$\begin{pmatrix} 4 & 1 \\ 1 & 4 \end{pmatrix} \begin{pmatrix} s_1 \\ s_2 \end{pmatrix} = \begin{pmatrix} \frac{3}{0.1}(0.9041 - 1.0000) \\ \frac{3}{0.1}(0.7828 - 0.9741) + 1.2843 \end{pmatrix} = \begin{pmatrix} -2.8770 \\ -4.4547 \end{pmatrix} \Rightarrow \begin{pmatrix} s_1 \\ s_2 \end{pmatrix} = \begin{pmatrix} -0.4702 \\ -0.9961 \end{pmatrix}$$

2.

$$I = 0.1\left(\frac{1}{2}1 + 0.9741 + 0.9041 + \frac{1}{2}0.7828\right) = 0.2770$$

3.

$$S(t) = \sum_{i=-2}^{2} a_i B_i^2(t)$$

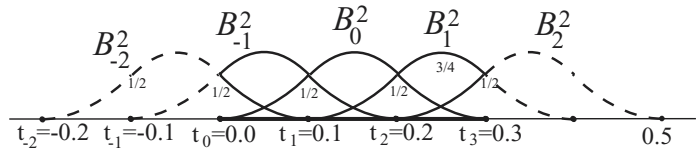


Figura 3.28: Base de splines para el apartado 3.

$$\begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 \\ 0 & \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{h}{6} & \frac{5h}{6} & h & \frac{5h}{6} & \frac{h}{6} \end{pmatrix} \begin{pmatrix} a_{-2} \\ a_{-1} \\ a_0 \\ a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} 1.0000 \\ 0.9741 \\ 0.9041 \\ 0.7828 \\ 0.2770 \end{pmatrix} \Rightarrow \begin{pmatrix} a_{-2} \\ a_{-1} \\ a_0 \\ a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} 1.0374 \\ 0.9626 \\ 0.9856 \\ 0.8226 \\ 0.7430 \end{pmatrix}$$

Obtengamos las derivadas teniendo en cuenta que en los nodos de cada uno de los elementos de la base su valor es $\pm\frac{1}{h}$.

$$S'(t) = \sum_{i=-2}^{2} a_i (B_i^2)'(t)$$

$$S'(0.1) = -\frac{1}{0.1}a_{-1} + \frac{1}{0.1}a_0 = 0.23 \quad \text{y} \quad S'(0.2) = -\frac{1}{0.1}a_0 + \frac{1}{0.1}a_1 = -1.630$$

Dejamos como ejercicio el cálculo en los otros nodos y representar gráficamente con Matlab este spline para explicar la gran diferencia entre las derivadas para el spline cúbico y el cuadrático.

PROBLEMA 3.13 *Interpolación multidimensional.*

Sea E el espacio de los polinomios de dos variables y grado 2. Cada elemento de E se escribe de modo único como:

$$P(x,y) = \sum_{i,j=0}^{2} a_{ij} x^i y^j$$

1. Estudiar la existencia y unicidad de un elemento P de E que verifique las igualdades

$$\begin{array}{c|c|c} P(0,0) = P_{00} & P(1,0) = P_{10} & P(2,0) = P_{20} \\ P(0,1) = P_{01} & P(1,1) = P_{11} & P(2,1) = P_{21} \\ P(0,2) = P_{02} & P(1,2) = P_{12} & P(2,2) = P_{22} \end{array}$$

   con $P_{ij} \in I\!R$.

2. Si el resultado del apartado 1 lo hace procedente, se pide encontrar un elemento P de E que verifique que:

$$\begin{array}{c|c|c} P(0,0) = 0 & P(1,0) = 0 & P(2,0) = 0 \\ P(0,1) = 1 & P(1,1) = 1 & P(2,1) = 1 \\ P(0,2) = 0 & P(1,2) = 0 & P(2,2) = 0 \end{array}$$

   Si en el proceso de busca se plantea un sistema lineal, se justificará la elección de un método para la resolución del mismo, y se resolverá, indicando los pasos intermedios.

3. Si procede de acuerdo con el resultado del apartado 1, encontrar al menos un elemento de la base de E dual de la que define en E^* el problema de interpolación de dicho apartado.

4. Relacionar si es posible dicha base con la de Lagrange en x y en y.

Solución:

1. El problema es claramente de interpolación lineal. Decir que cada elemento de E se escribe de modo único como

$$P(x,y) = \sum_{i,j=0}^{2} a_{ij} x^i y^j$$

equivale a afirmar que

$$\left\{1, x, x^2, y, xy, x^2y, y^2, xy^2, x^2y^2\right\} = \{e_j\}_{j=1,9}$$

es una base de E. El problema planteado se puede interpretar como la busca de un elemento P de E que tome imágnes preasignadas para una serie de formas lineales L_i, por ejemplo

$$L_1(P) = P(0,0) = P_{00}$$

Para que el problema tenga solución y ésta sea única,

$$det\langle L_i, e_j\rangle \neq 0$$

Se llega al mismo resultado imponiendo a un elemento genérico P de E las condiciones expresadas en el enunciado. Ello conduce al planteamiento de un sistema lineal cuya matriz es:

	1	x	x^2	y	xy	x^2y	y^2	xy^2	x^2y^2
(0,0)	1	0	0	0	0	0	0	0	0
(1,0)	1	1	1	0	0	0	0	0	0
(2,0)	1	2	4	0	0	0	0	0	0
(0,1)	1	0	0	1	0	0	1	0	0
(1,1)	1	1	1	1	1	1	1	1	1
(2,1)	1	2	4	1	2	4	1	2	4
(0,2)	1	0	0	2	0	0	4	0	0
(1,2)	1	1	1	2	2	2	4	4	4
(2,2)	1	2	4	2	4	8	4	8	16

Podemos demostrar que este determinante es distinto de cero bien calculando su valor con Matlab (sale ± 64 dependiendo del orden en que se tomen las filas), o bien mediante desarrollos o reducciones.

Otra forma interesante de ver que la matriz del sistema es regular es consecuencia de su estructura. Si definimos un bloque B como

$$B = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 2 & 4 \end{pmatrix}$$

con $\det(B) = 2 \neq 0$. Resulta que la gran matriz del sistema se puede escribir por bloques como:

$$A = \begin{pmatrix} B & 0 & 0 \\ B & B & B \\ B & 2B & 4B \end{pmatrix}$$

Mediante transformaciones elementales de filas, es fácil convertir esta matriz en una matriz diagonal por bloques múltiplos de B que será regular si B lo es.

2. Sea P_2 el elemento buscado en este apartado y sea b_2 el término independiente del sistema lineal cuando buscamos P_2

$$b_2 = \begin{pmatrix} 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \end{pmatrix}^T$$

En el apartado 3 se pide encontrar un elemento c_j de la base dual de la que en E^* define el problema de interpolación planteado en el apartado 1. Podemos hallar por ejemplo, c_7, el correspondiente a la forma lineal $L_7(P) = P(0,2)$ y que verifica

$$L_i\left(c_7\right) = \delta_{i7} \quad i = 1, ..., 7$$

O sea, llamando c_7 a ese vector y b_7 a su término independiente correspondiente que será

$$b_7 = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}^T$$

tendremos que encontrar p_2 y c_7 tales que:

$$A\,(p_2 \quad c_7) = (b_2 \quad b_7)$$

Resolviendo este doble sistema lineal simultáneamente sabremos de paso si el problema general tiene solución única o no. Para resolverlo podemos elegir un método directo o iterativo.

La descomposición por bloques antes descrita permitiría atacar el problema mediante un esquema iterativo. Sin embargo, la estructura de la matriz (de momento no es diagonalmente dominante) no induce a pensar que el esquema vaya a converger.

La matriz tiene una estructura cuasi triangular inferior que sugiere la conveniencia de convertirla en triangular inferior mediante eliminación gaussiana que realizaremos en su expresión por bloques.

$$A = \begin{pmatrix} B & 0 & 0 \\ B & B & B \\ B & 2B & 4B \end{pmatrix}$$

Podemos eliminar el bloque 2,3 multiplicando por 4 la fila 2 y restándole la tercera. Se obtiene así el sistema

$$\begin{pmatrix} B & 0 & 0 \\ 3B & 2B & 0 \\ B & 2B & 4B \end{pmatrix} (p_2 \quad c_7) = \begin{pmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 4 & -1 \\ 4 & 0 \\ 4 & 0 \\ 0 & 1 \\ 0 & 0 \\ 0 & 0 \end{pmatrix}$$

Procedemos por bloques, con sustitución hacia adelante

$$B \begin{pmatrix} p_2^1 & c_7^1 \\ p_2^2 & c_7^2 \\ p_2^3 & c_7^3 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{pmatrix}$$

De donde

$$p_2^1 = c_7^1 = p_2^2 = c_7^2 = p_2^3 = c_7^3 = 0$$

Sustituimos hacia adelante

$$2B \begin{pmatrix} p_2^4 & c_7^4 \\ p_2^5 & c_7^5 \\ p_2^6 & c_7^6 \end{pmatrix} = \begin{pmatrix} 4 & -1 \\ 4 & 0 \\ 4 & 0 \end{pmatrix}$$

y obtenemos

$$\begin{pmatrix} p_2^4 & c_7^4 \\ p_2^5 & c_7^5 \\ p_2^6 & c_7^6 \end{pmatrix} = \begin{pmatrix} 2 & -0.5 \\ 0 & 0.75 \\ 0 & -0.25 \end{pmatrix}$$

Para terminar

$$4B \begin{pmatrix} p_2^7 & c_7^7 \\ p_2^8 & c_7^8 \\ p_2^9 & c_7^9 \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \\ 0 & 0 \end{pmatrix} - 2B \begin{pmatrix} p_2^4 & c_7^4 \\ p_2^5 & c_7^5 \\ p_2^6 & c_7^6 \end{pmatrix} = \begin{pmatrix} -4 & 2 \\ -4 & 0 \\ -4 & 0 \end{pmatrix}$$

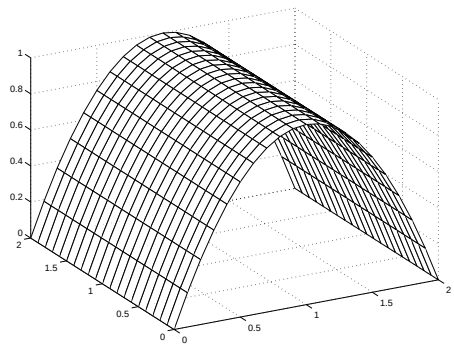


Figura 3.29: $p_2(x, y)$.

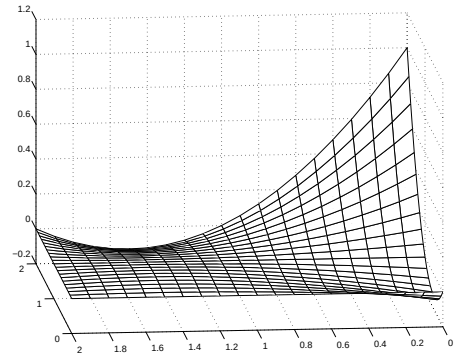


Figura 3.30: $c_7(x, y)$.

De donde

$$\begin{pmatrix} p_2^7 & c_7^7 \\ p_2^8 & c_7^8 \\ p_2^9 & c_7^9 \end{pmatrix} = \begin{pmatrix} -1 & 0.5 \\ 0 & -0.75 \\ 0 & 0.25 \end{pmatrix}$$

Dibujamos los resultados en las Figuras 3.29 y 3.30

$$p_2 = \begin{pmatrix} 0 & 0 & 0 & 2 & 0 & 0 & -1 & 0 & 0 \end{pmatrix}^T = 2y - y^2,$$
$$c_7 = \begin{pmatrix} 0 & 0 & 0 & -0.5 & 0.75 & -0.25 & 0.5 & -0.75 & 0.25 \end{pmatrix}^T$$

3. La relación pedida es fácil de ver.

   En el eje x

   $$l_i^x(x_j) = \delta_{ij}$$

   En el eje y

   $$l_i^y(y_j) = \delta_{ij}$$

   En el plano xy

   $$c_{ij}(x_k, y_m) = \delta_{ik} \cdot \delta_{jm} = l_i^x(x_k)\, l_j^y(y_m)$$

   Por tanto $c_{ij}(x, y)$ y $l_i^x(x) l_j^y(y)$ coinciden en los nodos. De la unicidad de la solución de este problema de interpolación se sigue que son el mismo elemento de E y podemos definir

   $$c_{ij}(x, y) := l_i^x(x)\, l_j^y(y)$$

PROBLEMA 3.14 *Splines paramétricos.*

Se quiere construir una representación mediante splines cúbicos de la caja de la guitarra española de la Figura 3.31. Para ello se utilizará un spline cúbico paramétrico $\mathbf{s}(t) = (x(t), y(t))$, consistente en dos splines cúbicos, uno para la coordenada x y otro para la coordenada y. Vamos a trabajar sobre la coordenada y, utilizando los puntos que nos da la figura, y utilizando como variable independiente t el índice de cada punto.

Las dos condiciones adicionales que necesitamos se refieren al valor de las derivadas en el primer y último punto. Éstas proceden de fijar las tangencias verticales en ambos. Por tanto, tomaremos $y'(0) = 1.0$ e $y'(4) = -1.0$.

1. Calcular el valor de las derivadas en los nodos interiores.

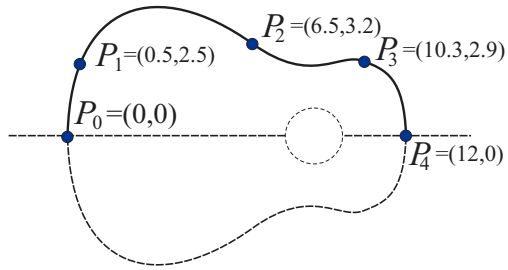


Figura 3.31: Representación mediante splines cúbicos de la caja de una guitarra.

i	t_i	y_i
0	0	0.0
1	1	2.5
2	2	3.2
3	3	2.9
4	4	0.0

Cuadro 3.1: Tabla de la figura 3.31.

2. Hacer una representación esquemática a partir de estos valores del gráfico de $(t, y(t))$ comprobando si, como se pretende, se ajusta con el de la Figura 3.31.

3. Utilizando diferencias divididas para calcular la cúbica de Hermite correspondiente al tercer tramo, calcular el valor correspondiente a $t = 2.5$ verificando que se corresponde aproximadamente con el esquema que se ha dibujado.

4. Escribir un código Matlab que resuelva el problema tanto para la variable x como para la y, representado gráficamente la curva $(x(t), y(t))$ resultado.

Solución:

1. En el apartado 3.4 de la teoría ya estudiamos el problema de encontrar un spline conocidas sus derivadas en el primero y último nodos. Este problema se reduce a resolver el sistema lineal 3.19, que tiene como incógnitas las derivadas s_1, s_2, s_3 de ese spline en los nodos interiores. Cada línea de este sistema lineal tiene el siguiente aspecto:

$$s_i + 4s_{i+1} + s_{i+2} = \frac{3}{h}(y_{i+2} - y_i) \quad i = 0, 1, 2.$$

En este caso, $h = 1$, y el sistema lineal queda

$$\begin{pmatrix} 4 & 1 & 0 \\ 1 & 4 & 1 \\ 0 & 1 & 4 \end{pmatrix} \begin{pmatrix} s_1 \\ s_2 \\ s_3 \end{pmatrix} = \begin{pmatrix} 3(y_2 - y_0) - s_0 \\ 3(y_3 - y_1) \\ 3(y_4 - y_2) - s_4 \end{pmatrix}$$

$$\begin{pmatrix} 3(y_2 - y_0) - s_0 \\ 3(y_3 - y_1) \\ 3(y_4 - y_2) - s_4 \end{pmatrix} = \begin{pmatrix} 3(3.2 - 0.0) - 1.0 \\ 3(2.9 - 2.5) \\ 3(0.0 - 3.2) + 1.0 \end{pmatrix} = \begin{pmatrix} 8.6 \\ 1.2 \\ -8.6 \end{pmatrix}$$

Sabemos que este tipo de sistemas lineales de diagonal estrictamente dominante siempre son regulares. La solución se obtiene directamente en Matlab

$$\begin{pmatrix} s_1 \\ s_2 \\ s_3 \end{pmatrix} = \begin{pmatrix} 2.0643 \\ 0.3429 \\ -2.2357 \end{pmatrix}$$

2. En la Figura 3.32 presentamos un gráfico esquemático del spline $y(t)$. Comparando con la curva correspondiente a la guitarra (Figura 3.31) destaca que entre los puntos 1 y 2 de la Figura 3.32 debería haber un máximo en la y, y entre los puntos 2 y 3 de la misma figura debería haber un mínimo en la y, mientras que en la curva obtenida no aparecen. La representación de la guitarra va a quedar solo regular como luego veremos.

3. Planteemos el problema de diferencias divididas, duplicando los nodos porque en ellos sabemos también la derivada.

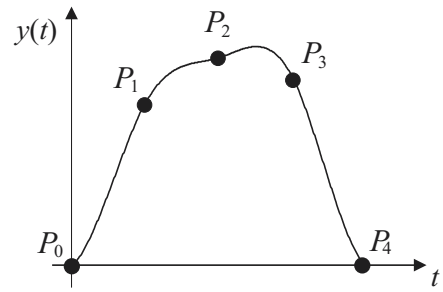


Figura 3.32: Esquema del spline cúbico $y(t)$.

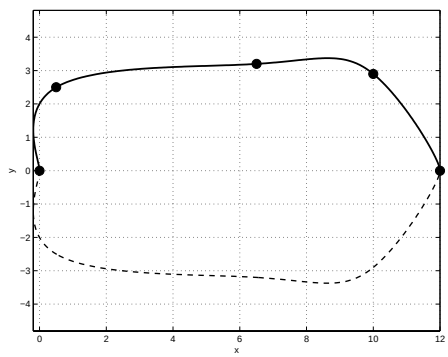


Figura 3.33: Spline paramétrico $(x(t), y(t))$.

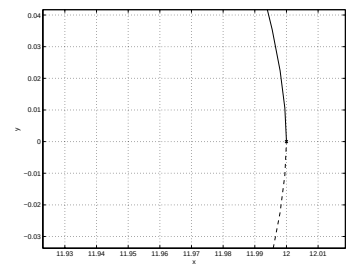


Figura 3.34: Detalle de la tangente vertical en el último punto.

j	u_j	y_j	$y[u_j, u_{j+1}]$	$y[u_j, u_{j+1}, u_{j+2}]$	$y[u_0, u_1, u_2, u_3]$
0	2	3.2			
1	2	3.2	0.3429		
2	3	2.9	-0.3000	-0.6429	
3	3	2.9	-2.2357	-1.9357	-1.2928

La parte de nuestro spline $y(t)$ correspondiente al tramo $t \in [2, 3]$ es

$$y(t) = 3.2 + 0.3429(t-2) - 0.6429(t-2)^2 - 1.2928(t-3)(t-2)^2$$

El valor obtenido para $t = 2.5$ es $y(2.5) = 3.3723$, que se ajusta al de la Figura 3.32 aunque no al dibujo original de la guitarra.

4. Podemos también interpolar con otro spline cúbico la curva $(t, x(t))$, tomando derivadas nulas en el primer y último punto, debido a las tangencias verticales. Con ello, ya estamos en condiciones de representar la curva $(x(t), y(t))$, que interpola los nodos que hemos elegido en la guitarra original (Figura 3.33). Para ello, usamos el siguiente código matlab, que utiliza la expresión de cada cúbica de Hermite entre dos nodos igual que hicimos para deducir en la teoría el sistema lineal (3.19) que nos permitía obtener las derivadas correspondientes a los splines cúbicos. Si nos fijamos en la Figura 3.33, podría parecer que la tangente no es vertical en el último punto, pero si hacemos un *zoom* de esa zona veremos que no es así (Figura 3.34).

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% guitarra.m
%
% Generar el spline S(t)=(x(t),y(t)) que ajuste la curva correspondiente
% a la caja exterior de la guitarra.
```

```
%
clear
t=0:4;
h=1;
A=[4 1 0
   1 4 1
   0 1 4];
sy1=1; sy5=-1; %derivadas de y(t) en primer y ultimo nodo
x=[0 0.5 6.5 10 12];
y=[0 2.5 3.2 2.9 0];
b=[3*(y(3)-y(1))-sy1; 3*(y(4)-y(2)); 3*(y(5)-y(3))-sy5];
aux=A\b;
sy=[sy1 aux' sy5];
sx1=0; sx5=0; %derivadas de x(t) en primer y ultimo nodo
b=[3*(x(3)-x(1))-sx1; 3*(x(4)-x(2)); 3*(x(5)-x(3))-sx5];
aux=A\b;
sx=[sx1 aux' sx5];
vt=0:0.01:4;
for j=1:length(vt)-1
    tt=vt(j);
    i=floor(tt)+1; % con tt barremos el rango y con esta instruccion descubrimos el tramo
    X(j)=[1+2/h*(tt-t(i))]*((tt-t(i+1))/h)^2*x(i)+(tt-t(i))*((tt-t(i+1))/h)^2*sx(i);
    X(j)=X(j)+[1-2/h*(tt-t(i+1))]*((tt-t(i))/h)^2*x(i+1)+(tt-t(i+1))*((tt-t(i))/h)^2*sx(i+1);
    Y(j)=[1+2/h*(tt-t(i))]*((tt-t(i+1))/h)^2*y(i)+(tt-t(i))*((tt-t(i+1))/h)^2*sy(i);
    Y(j)=Y(j)+[1-2/h*(tt-t(i+1))]*((tt-t(i))/h)^2*y(i+1)+(tt-t(i+1))*((tt-t(i))/h)^2*sy(i+1);
end
Y=[Y y(5)];
X=[X x(5)];
plot(X,Y,x,y,'x',X,-Y,':'); xlabel('x'); ylabel('y');
axis equal;
grid;
shg;
```

Si superponemos las gráficas correspondientes a la guitarra original y la curva que hemos obtenido, comprobamos que no es suficiente con ajustar la tangente en el primer y último nodo e interpolar esos nodos para tener la guitarra buscada (Figura 3.35).

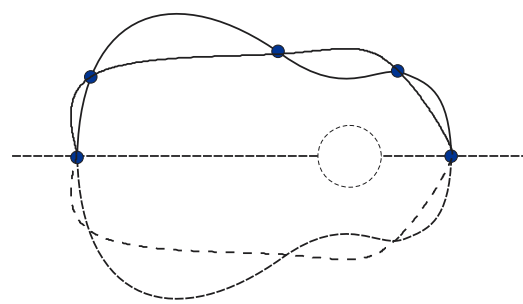


Figura 3.35: Comparación de guitarra y su spline de interpolación.

Las razones para este *desastre* escapan al contenido del texto y se explican desde una rama emergente de la matemática que es la Geometría Computacional. Cuando modelamos la figura en la aplicación gráfica correspondiente, además de controlar los puntos de paso y las tangencias en esos puntos, debemos también manipular la *intensidad* de esa tangencia. Eso es algo que no podemos hacer con el spline cúbico sobre el que hemos trabajado. Al estudiante interesado lo remitimos a Farin [11].

[11] Farin, G., *Curves and Surfaces for CAGD*, Morgan Kaufmann, Elsevier, 2002.

PROBLEMA 3.15 *Splines cíclicos.*

Este problema estudia el fundamento de los splines cíclicos, que son muy importantes en diseño gráfico, pues corresponden a aquellas curvas splines que toman los mismos valores y tienen la misma derivada en los nodos inicial y final como si se tratase de un nodo interior.

Vamos a construir una curva paramétrica cerrada $(x(t), y(t))$ en el plano, en la cual $x(t)$ e $y(t)$ son splines de grado 2 cíclicos asociados a la partición $\Omega = \{-1, 0, 1, 2\}$ del compacto $[-1, 2]$. Por tanto, $x(t) \in S_2(\Omega)$ e $y(t) \in S_2(\Omega)$. Cada uno de los nodos de esa partición se corresponde con un punto de la curva spline paramétrica cerrada cíclica en $\mathbb{R}^2$. En la curva que vamos a construir, representada en la Figura 3.36, el

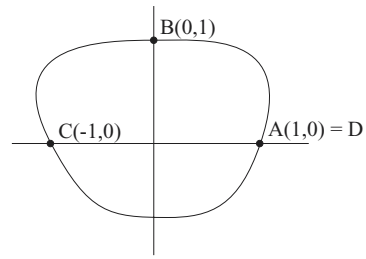


Figura 3.36: Spline cíclico.

primer nodo A es el punto $(1, 0)$, el segundo $B = (0, 1)$, el tercero $C = (-1, 0)$, y el cuarto se deduce fácilmente de la naturaleza de la curva.

Se pide definir de modo preciso uno de los dos splines, el $x(t)$ o el $y(t)$.

Solución:
Calculemos por ejemplo $x(t)$.

$$x(t) = \begin{cases} a_0 + b_0 t + c_0 t^2 & -1 \le t \le 0 \\ a_1 + b_1 t + c_1 t^2 & 0 \le t \le 1 \\ a_2 + b_2 t + c_2 t^2 & 1 \le t \le 2 \end{cases}$$

Apliquemos las diferentes condiciones:

i. Valor para $t = -1$, correspondiente al primer punto $A(1, 0)$, por tanto,

$$x(-1) = 1 = a_0 - b_0 + c_0$$

ii. Valor para $t = 0$, segundo punto $B(0, 1)$, y continuidad de $x(t)$ en ese punto

$$x(0)^- = 0 = a_0, \qquad x(0)^+ = 0 = a_1$$

iii. Ídem para $t = 1$, tercer punto $C(-1, 0)$, y continuidad de $x(t)$ en ese punto

$$x(1)^- = -1 = a_1 + b_1 + c_1, \qquad x(1)^+ = -1 = a_2 + b_2 + c_2$$

iv. Valor para $t = 2$, que como es un spline cíclico debe coincidir con el primero $A(1, 0)$. Por tanto:

$$x(2) = 1 = a_2 + 2b_2 + 4c_2$$

v. Ya que es un spline de grado 2, debe haber continuidad de la primera derivada en los nodos interiores, $t = 0$ y $t = 1$.

$$x'(0)^- = x'(0)^+ \;\Rightarrow\; b_0 = b_1, \qquad x'(1)^- = x'(1)^+ \;\Rightarrow\; b_1 + 2c_1 = b_2 + 2c_2$$

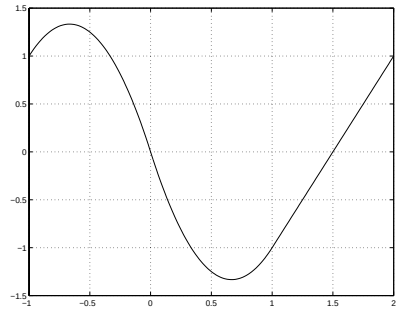


Figura 3.37: $x(t)$**.**

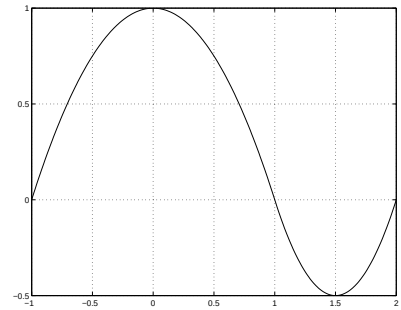


Figura 3.38: $y(t)$**.**

vi. Como es un spline cíclico, debe haber "continuidad" en la primera derivada entre el primero y el último nodo.

$$x'(-1)^+ = x'(2)^- \quad \Rightarrow \quad b_0 - 2c_0 = b_2 + 4c_2$$

Resolviendo este sistema de ecuaciones se tiene

$$\begin{array}{lll} a_0 = 0, & a_1 = 0, & a_2 = -3 \\ b_0 = -4\ , & b_1 = -4\ , & b_2 = 2 \\ c_0 = -3\ , & c_1 = 3\ , & c_2 = 0 \end{array}$$

Por tanto el spline buscado, que dibujamos en la Figura 3.37, es:

$$x(t) = \begin{cases} -4t - 3t^2 & -1 \leq t \leq 0 \\ -4t + 3t^2 & 0 \leq t \leq 1 \\ -3 + 2t & 1 \leq t \leq 2 \end{cases}$$

Análogamente obtenemos

$$y(t) = \begin{cases} 1 - t^2 & -1 \leq t \leq 0 \\ 1 - t^2 & 0 \leq t \leq 1 \\ 4 - 6t + 2t^2 & 1 \leq t \leq 2 \end{cases}$$

que representamos en la Figura 3.38. Finalmente representando x "versus" y se tiene la curva resultado, como aparece en la Figura 3.39.

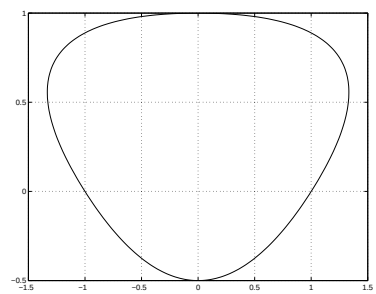


Figura 3.39: $(x(t), y(t))$**.**

Repitamos el problema eligiendo otra base del espacio $S_2(\Omega)$, por ejemplo la base de B-splines de segundo grado asociados a esa partición. Indexando correctamente los nodos obtendremos el elemento de la base asociado a cada nodo.

$$\begin{array}{ll} t_0 = -1, & t_1 = 0 \\ t_2 = 1\ , & t_3 = 2 \end{array}$$

De momento, los tres elementos de base: B_0^2, B_1^2, B_2^2. Como la dimensión es 5, $n+k = 3+2 = 5$, necesitamos dos elementos más, que conseguimos añadiendo dos nodos adicionales a la izquierda de nuestro dominio, B_{-1}^2 y B_{-2}^2 asociados a esos nodos:

$$t_{-1} = -2, \quad t_{-2} = -3$$

Presentamos en la Figura 3.40 los soportes de los elementos de la base. En función de esa base, el spline $x(t)$ se escribirá

$$x(t) = \sum_{i=-2}^{2} a_i B_i^2(t)$$

Apliquemos cada una de las condiciones:

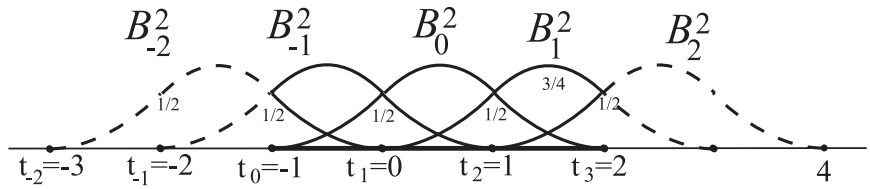


Figura 3.40: Soportes de los elementos de la base de B-splines de $S_2(\Omega)$.

i. Valor para $t = -1$, correspondiente al primer punto $A(1, 0)$

$$x(-1) = 1 = \sum_{i=-2}^{2} a_i B_i^2(-1) = \frac{1}{2}a_{-2} + \frac{1}{2}a_{-1}$$

ii. Valor para $t = 0$, segundo punto $B(0, 1)$

$$x(0) = 0 = \sum_{i=-2}^{2} a_i B_i^2(0) = \frac{1}{2}a_{-1} + \frac{1}{2}a_0$$

iii. Ídem para $t = 1$, tercer punto $C(-1, 0)$.

$$x(1) = -1 = \sum_{i=-2}^{2} a_i B_i^2(1) = \frac{1}{2}a_0 + \frac{1}{2}a_1$$

iv. Valor para $t = 2$, que como es un spline cíclico debe coincidir con el primero $A(1, 0)$. Por tanto:

$$x(2) = 1 = \sum_{i=-2}^{2} a_i B_i^2(2) = \frac{1}{2}a_1 + \frac{1}{2}a_2$$

v. Como es un spline cíclico, debe haber "continuidad" en la primera derivada entre el primero y el último nodo.

$$x'(-1)^+ = x'(2)^- \quad \leftrightarrow \quad x'(-1)^+ - x'(2)^- = 0$$

$$\sum_{i=-2}^{2} a_i B_i^{2'}(-1) - \sum_{i=-2}^{2} a_i B_i^{2'}(2) = 0 \quad \Rightarrow \quad -a_{-2} + a_{-1} + a_1 - a_2 = 0$$

Disponiendo los resultados anteriores como un sistema lineal tendremos, multiplicando por 2 todas las filas menos la última:

$$\begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ -1 & 1 & 0 & 1 & -1 \end{pmatrix} \begin{pmatrix} a_{-2} \\ a_{-1} \\ a_0 \\ a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} 2 \\ 0 \\ -2 \\ 2 \\ 0 \end{pmatrix}$$

Y por tanto, el spline solución en esta base es:

$$x(t) = 2B_{-1}^2(t) - 2B_0^2(t) + 2B_2^2(t)$$

De la expresión polinómica sabemos que, por ejemplo, $x(1.5) = 0$. Comprobémoslo con la expresión B-spline.

$$x(1.5) = 2B_{-1}^2(1.5) - 2B_0^2(1.5) + 2B_2^2(1.5) = 2 \cdot 0 - 2 \cdot 0.125 + 2 \cdot 0.125 = 0$$

Otra posibilidad muy interesante es calcular directamente las derivadas en los nodos, y construir utilizando esa información y los valores en los nodos, que son conocidos las diferentes parábolas. En ese sentido, merece la pena recordar cómo es una parábola que pasa por los puntos (t_0, x_0), (t_1, x_1), y cuya derivada en t_0 es s_0:

$$P_0(t) = x_0 + s_0\,(t - t_0) + \frac{x_1 - x_0 - s_0 h_0}{h_0^2}\,(t - t_0)^2$$

con $h_0 = t_1 - t_0$.
En general, suponiendo que esa parábola es uno de los tramos de un spline parabólico −el tramo que va entre los nodos de abscisas, t_i y t_{i+1}− la expresión de ese tramo parabólico P_i sería

$$P_i(t) = x_i + s_i\,(t - t_i) + \frac{x_{i+1} - x_i - s_i h_i}{h_i^2}\,(t - t_i)^2$$

con $h_i = t_{i+1} - t_i$.
Como estamos ante un spline parabólico, su derivada debe ser continua en los nodos donde enganchan los diferentes tramos; condición que podemos expresar así

$$P_i'\,(t_{i+1}) = P_{i+1}'\,(t_{i+1})$$

Imponiendo esta condición en cada uno de los nodos interiores, se llega a una relación que vincula las derivadas en dichos nodos (ver problema 3.11). Dicha condición es la siguiente:

$$s_i + s_{i+1} = 2\frac{x_{i+1} - x_i}{h_i}$$

Como la diferencia en t entre dos nodos consecutivos es 1, tendremos:

$$s_i + s_{i+1} = 2\,(x_{i+1} - x_i)$$

Apliquémoslo:

$$\begin{aligned} s_0 + s_1 &= 2\,(x_1 - x_0) = 2\,(0 - 1) = -2 \\ s_1 + s_2 &= 2\,(x_2 - x_1) = 2\,(-1 - 0) = -2 \\ s_2 + s_3 &= 2\,(x_3 - x_2) = 2\,(1 - (-1)) = 4 \end{aligned}$$

La última ecuación se refiere a la "continuidad" de la derivada entre el último y el primer nodo:

$$s_3 = s_0$$

Resolviendo este sistema se tiene

$$\begin{aligned} s_0 = 2, \quad s_1 = -4 \\ s_2 = 2\ , \quad s_3 = 2 \end{aligned}$$

Podemos ahora construir cada uno de los tramos de nuestro spline solución

$$X(t)|_{[t_i,t_{i+1}]} = x_i + s_i\,(t-t_i) + (x_{i+1} - x_i - s_i)\,(t-t_i)^2$$

$$X(t) = \begin{cases} 1 + 2(t+1) + (0-1-2)(t+1)^2 = 1 + 2(t+1) - 3(t+1)^2 & -1 \le t \le 0 \\ 0 - 4(t-0) + (-1-0-(-4))(t-0)^2 = -4t + 3t^2 & 0 < t \le 1 \\ -1 + 2(t-1) + (1-(-1)-2)(t-1)^2 = -1 + 2(t-1) & 1 < t \le 2 \end{cases}$$

Se comprueba con facilidad que coinciden con los obtenidos al principio.

PROBLEMA 3.16 *Polinomios a trozos de grado 2 y clase 0.*

Dada la partición $\Omega = \{-1, 0, 1, 2\}$ del compacto $[-1, 2]$. Se considera el espacio vectorial $P_{2,0}(\Omega)$ de los polinomios a trozos de grado 2 y de clase $\mathcal{C}^0$ en Ω, luego tales que su restricción a cada intervalo es un polinomio de grado 2, y que el "enganche" entre intervalos se produce con continuidad. Se representa uno de estos polinomios en la Figura 3.41.

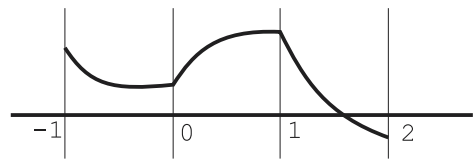


Figura 3.41: Polinomio a trozos de grado 2 y clase $\mathcal{C}^0$.

Se pide:

1. Hallar la dimensión de este espacio vectorial.
2. Definir una metodología que permita calcular bases en este espacio vectorial y aplicarla para definir cada uno de los elementos de una base determinada. Se pide también representar gráficamente estos elementos.

Solución:

1. Sea P un polinomio a trozos cualquiera asociado a Ω

$$P = \begin{cases} P_{-1}(x) & -1 \le x \le 0 \\ P_0(x) & 0 \le x < 1 \\ P_1(x) & 1 \le x \le 2 \end{cases}$$

Si P no tuviese que satisfacer las condiciones de $\mathcal{C}^0$ continuidad en los extremos de los intervalos de la partición, la dimension buscada sería 9, igual a la suma de los grados de libertad de los tres polinomios de segundo grado P_i, $i = -1, 0, 1$. Al imponer dichas condiciones, se establecerán ciertas relaciones entre los coeficientes de esos polinomios, por lo que $\dim P_{2,0} < 9$.

Una forma elegante y adecuada de interpretar las condiciones de continuidad permite definir $P_{2,0}$ como la intersección de los núcleos de dos formas lineales definidas en el espacio vectorial $P_{2,-1}(\Omega)$ de los polinomios a trozos en Ω no necesariamente continuos en 0 y 1.

En efecto, dado un polinomio a trozos $P \in P_{2,-1}(\Omega)$ definimos las 6 formas lineales

$$\begin{array}{lll} L_0^-(P) = P^-(0), & L_1^-(P) = P^-(1), & M(P) = L_0^-(P) - L_0^+(P) \\ L_0^+(P) = P^+(0), & L_1^+(P) = P^+(1), & N(P) = L_1^-(P) - L_1^+(P) \end{array}$$

y es evidente que $P_{2,0}(\Omega) = \text{Ker } M \bigcap \text{Ker } N$ de donde $\dim P_{2,0} = 3 \cdot 3 - 2 = 7$.

2. Se interpreta una base de $P_{2,0}(\Omega)$ como una base del bidual, es decir, la base dual de una base del dual $P_{2,0}(\Omega)^*$. La metodología asociada define primero una base de $P_{2,0}(\Omega)^*$ descrita por 7 formas lineales en $P_{2,0}(\Omega)$ linealmente independientes y calcula después su dual.

   Supongamos, por ejemplo, que conocemos en cada tramo los valores de la función en los nodos y la derivada en el primer nodo información que se puede escribir en el lenguaje del dual mediante las formas lineales

$$\begin{array}{cccc} L_0(P) = P(t_0), & L_1(P) = P(t_1), & L_2(P) = P(t_2) & L_3(P) = P(t_3) \\ L_4(P) = P'(t_0)^+, & L_5(P) = P'(t_1)^+, & L_6(P) = P'(t_2)^+ & \end{array}$$

   Este problema tiene solución única, ya que tiene solución única en cada tramo, que se calcula fácilmente usando diferencias divididas.

   De hecho, dado que $(t_{i+1} - t_i) = 1$, cada parábola P_i se escribe en la forma

$$P_i(t) \quad = \quad f_i + s_i(t - t_i) + (f_{i+1} - f_i - s_i)(t - t_i)^2$$

   Por definición, los elementos l_j de la base dual de L_i cumplen

$$L_i(l_j) \quad = \quad \delta_{ij}$$

   Se obtiene por ejemplo para l_0 que $L_0(l_0) = 1$ y $L_i(l_0) = 0$ si $i > 0$

$$\left.\begin{array}{l} f_0 = 1 \\ f_1 = 0 \\ f_2 = 0 \\ f_3 = 0 \\ s_0 = 0 \\ s_1 = 0 \\ s_2 = 0 \end{array}\right\} \quad l_0 = \begin{cases} 1 - (t+1)^2 & t \in [-1, 0) \\ 0 & t \in [0, 1) \\ 0 & t \in [1, 2] \end{cases}$$

   Y sucesivamente:

$$l_1(t) = \begin{cases} (t+1)^2 & t \in [-1, 0) \\ 1 - t^2 & t \in [0, 1) \\ 0 & t \in [1, 2] \end{cases} \quad l_2(t) = \begin{cases} 0 & t \in [-1, 0] \\ t^2 & t \in [0, 1] \\ 1 - (t-1)^2 & t \in [1, 2] \end{cases}$$

$$l_3(t) = \begin{cases} 0 & t \in [-1, 0) \\ 0 & t \in [0, 1) \\ (t-1)^2 & t \in [1, 2] \end{cases} \quad l_4(t) = \begin{cases} (t+1) - (t+1)^2 & t \in [-1, 0) \\ 0 & t \in [0, 1) \\ 0 & t \in [1, 2] \end{cases}$$

$$l_5(t) = \begin{cases} 0 & t \in [-1, 0) \\ t - t^2 & t \in [0, 1) \\ 0 & t \in [1, 2] \end{cases} \quad l_6(t) = \begin{cases} 0 & t \in [-1, 0) \\ 0 & t \in [0, 1) \\ (t-1) - (t-1)^2 & t \in [1, 2] \end{cases}$$

Podemos observar las gráficas de estas funciones de base en las Figuras 3.42, 3.43 y 3.44.

Otra posibilidad

Definimos una base del espacio vectorial $P_{2,-1}(\Omega)$ de dimensión 9.

$$e_1 = \begin{cases} 1 \\ 0 \\ 0 \end{cases} \quad e_2 = \begin{cases} x+1 \\ 0 \\ 0 \end{cases} \quad e_3 = \begin{cases} (x+1)^2 \\ 0 \\ 0 \end{cases} \quad e_4 = \begin{cases} 0 \\ 1 \\ 0 \end{cases} \quad e_5 = \begin{cases} 0 \\ x \\ 0 \end{cases}$$

$$e_6 = \begin{cases} 0 \\ x^2 \\ 0 \end{cases} \quad e_7 = \begin{cases} 0 \\ 0 \\ 1 \end{cases} \quad e_8 = \begin{cases} 0 \\ 0 \\ x-1 \end{cases} \quad e_9 = \begin{cases} 0 & -1 \le x < 0 \\ 0 & 0 \le x < 1 \\ (x-1)^2 & 1 \le x \le 2 \end{cases}$$

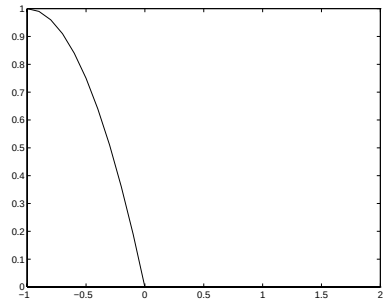


Figura 3.42: l_0.

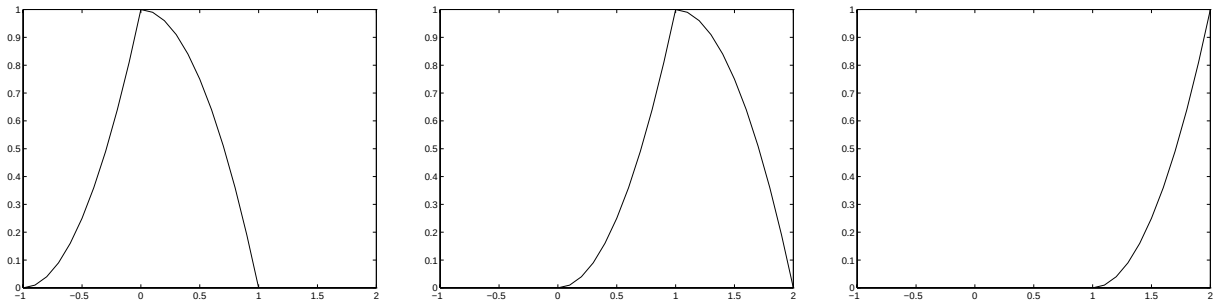


Figura 3.43: l_1, l_2 **y** l_3.

Impongamos ahora a un elemento genérico $C \in P_{2,0}(\Omega)$ cuya descomposición única como elemento de $P_{2,-1}(\Omega)$ es $C = \sum_{i=1}^{9} \lambda_i e_i$ la continuidad en 0 y 1, luego en esos puntos deben coincidir los valores por la derecha y por la izquierda de C.

$$\left.\begin{array}{l} C(0) = \lambda_1 + \lambda_2 + \lambda_3 = \lambda_4 \\ C(1) = \lambda_4 + \lambda_5 + \lambda_6 = \lambda_7 \end{array}\right\} \quad \Rightarrow \quad \lambda_7 = \lambda_1 + \lambda_2 + \lambda_3 + \lambda_5 + \lambda_6$$

Por tanto,

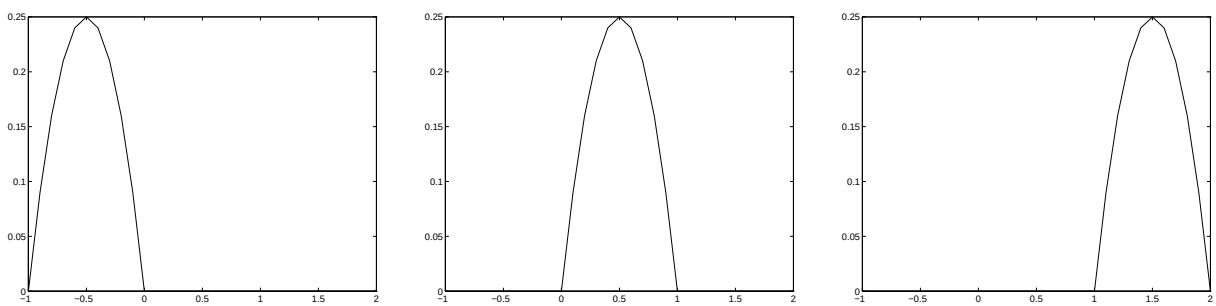


Figura 3.44: l_4, l_5 **y** l_6.

$$\begin{aligned} C &= \lambda_1(e_1 + e_4 + e_7) + \lambda_2(e_2 + e_4 + e_7) + \lambda_3(e_3 + e_4 + e_7) + \\ &\quad +\lambda_5(e_5 + e_7) + \lambda_6(e_6 + e_7) + \lambda_8(e_8) + \lambda_9 e_9 \end{aligned}$$

Llamando

$$\begin{array}{cccc} b_1 = e_1 + e_4 + e_7, & b_2 = e_2 + e_4 + e_7, & b_3 = e_3 + e_4 + e_7 & , \\ b_4 = e_5 + e_7, & b_5 = e_6 + e_7, & b_6 = e_8, & b_7 = e_9 \end{array}$$

se observa que la familia $\{b_i\}_{i=1,..,7}$ es un sistema generador de $P_{2,0}(\Omega)$ de 7 elementos luego una base

$$b_1 = \begin{cases} 1 \\ 1 \\ 1 \end{cases} \quad b_2 = \begin{cases} x+1 \\ 1 \\ 1 \end{cases} \quad b_3 = \begin{cases} (x+1)^2 \\ 1 \\ 1 \end{cases} \quad b_4 = \begin{cases} 0 \\ x \\ 1 \end{cases}$$

$$b_5 = \begin{cases} 0 \\ x^2 \\ 1 \end{cases} \quad b_6 = \begin{cases} 0 \\ 0 \\ x-1 \end{cases} \quad b_7 = \begin{cases} 0 \\ 0 \\ (x-1)^2 \end{cases}$$

Estudiemos el rango de esta familia analizando el rango de la matriz de las componentes de sus elementos respecto de la base $\{e_i\}_{i=1,...,9}$ de $P_{2,-1}(\Omega)$

$$(rgB) = \begin{pmatrix} 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

Quitando a B las columnas 4 y 7 nos queda la submatriz I_7 cuyo rango es 7.
En la dirección *http://canal.etsin.upm.es/ftp/p20int.zip* tenemos el fichero Matlab que resuelve y generaliza este problema para cualquier número de nodos.
Este problema tiene continuación en el problema de aproximación 4.6.

CAPÍTULO 4

Aproximación de funciones

El objetivo fundamental en este capítulo es buscar dentro de un espacio de funciones de dimensión finita aquellas que están más cerca (en el sentido que proporciona una distancia definida de modo riguroso en ese espacio) a una función dada que queramos aproximar. Los casos más interesantes surgen cuando esa distancia se deduce de un producto escalar porque las ideas de ortogonalidad permiten utilizar bases en las que la solución buscada se escribe más fácilmente.

Un estudio independiente merecen los problemas de aproximación cuando la función a aproximar está definida de modo discreto (**mínimos cuadrados**). Este problema tiene una relación directa con el concepto estadístico de regresión lineal, y por eso merece una atención especial tanto para los ingenieros que tienen como parte de sus tareas el simular sistemas como para los ingenieros dedicados a las finanzas que deben predecir el comportamiento de determinadas variables econométricas.

Dentro de los problemas de mínimos cuadrados, destacan los que tienen como espacio de funciones de base las exponenciales complejas (será el único caso en que utilicemos números complejos). Este caso se corresponde con la transformada de Fourier de funciones definidas de modo discreto. Los ejercicios de este último tipo son muy básicos, dado que este tema constituye una materia en sí misma, el **Tratamiento Digital de Señales**, que cae fuera del ámbito de este libro de problemas. Saber manejar bien las técnicas de aproximación por mínimos cuadrados es fundamental en ingeniería, pues así se construyen modelos continuos a partir de resultados discretos obtenidos de experimentos o de otros cálculos numéricos.

Un estudio exhaustivo de la parte correspondiente a las transformadas de Fourier de funciones definidas de modo discreto se puede encontrar en el texto de Oppenheim et al. [22]. Respecto al problema general, el texto de Hammerlin et al. [15] hace un estudio muy bueno partiendo del problema en su formulación más general.

4.1. Introducción

Nuestro objetivo fundamental será investigar la aproximación de funciones esencialmente complicadas y de las que a veces no sabemos mucho (normalmente sabremos que son continuas y poco más) mediante funciones más simples y que pertenecen a espacios de funciones de dimensión finita, como por ejemplo los polinomios de un cierto grado.

Ejemplo 4.1.1 *Se quiere encontrar una recta con la que aproximar-sustituir localmente a la función seno entre 0 y $\pi/2$. Se plantean 4 alternativas para su elección:*

1. *Que sea la recta de mínimos cuadrados correspondiente a evaluar la función seno en los puntos* $\{0.14, 0.32, 0.86, 1.37, 1.48\}$. *Aunque más adelante estudiaremos con detalle la aproximación por mínimos cuadrados, que es lo más importante de este capítulo, seguro que al lector no le resulta ajeno este concepto, quizá a raíz de analizar experimentos en cursos previos. La solución correspondiente a esta posibilidad es la recta* $r_1(x) = 0.6407x + 0.1032$. *En la gráfica 4.1 superponemos la solución con la nube de puntos.*
2. *Que sea la recta r_2 que minimice la norma del máximo (ver 3.2.1 sobre esta norma) de la diferencia entre el seno y dicha recta, entre 0 y $\pi/2$. O sea, que $\|\sin - r_2\|_\infty$ sea mínima. En este caso, la solución es la recta* $r_2(x) = 0.6366x + 0.1053$.

   Observando ambas ecuaciones, podemos comprobar que r_1 y r_2 son bastante parecidas aunque no iguales.

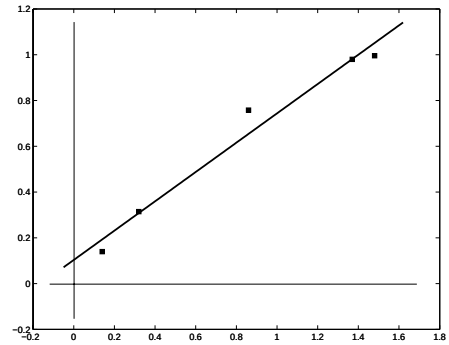


Figura 4.1: Aproximación por mínimos cuadrados: ejemplo 4.1.1.

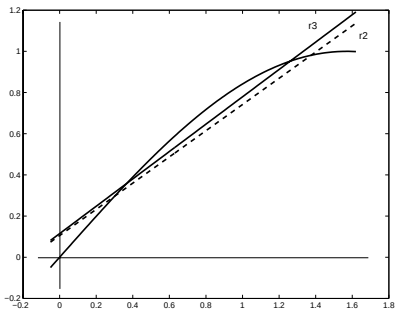


Figura 4.2: Aproximaciones 2 y 3: ejemplo 4.1.1.

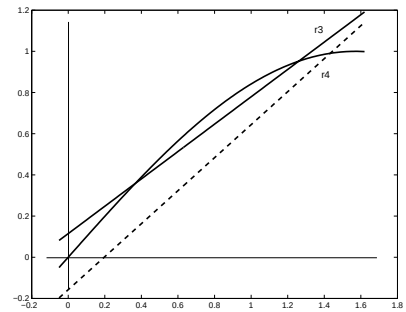


Figura 4.3: Aproximaciones 3 y 4: ejemplo 4.1.1.

3. *Que sea la recta que minimice el valor* $\int_0^{\pi/2}(\sin(x)-r(x))^2dx$. *En este caso, la solución es la recta* $r_3(x) = 0.6644x + 0.1148$, *muy similar también a las anteriores.*
   En la gráfica 4.2 superponemos las rectas r_2 *y* r_3 *con la función seno.*

4. *Puede ser interesante elegir una aproximación en la que se tenga también en cuenta la distancia entre las derivadas. Podemos buscar una recta que minimice el valor*

$$\int_0^{\pi/2}(\sin(x)-r(x))^2dx + \int_0^{\pi/2}(\sin'(x)-r'(x))^2dx$$

   En este caso, la recta obtenida difiere sustancialmente de las anteriores, $r_4(x) = 0.8014x - 0.1576$.
   En la Figura 4.3 superponemos las rectas r_3 *y* r_4 *con la función seno.*

Podríamos mejorar las aproximaciones anteriores aumentando el grado de los polinomios. Que las aproximaciones mejoren al aumentar la dimensión del espacio donde las buscamos es una de las características que debemos pedir a las funciones de aproximación.

Otra característica deseable de esas funciones es que sea fácil derivarlas e integrarlas y realizar en general las operaciones elementales.

Como respuesta, los espacios de aproximación más utilizados son espacios de polinomios y espacios de funciones trigonométricas. En estos espacios existen además innumerables teoremas, que garantizan que los resultados obtenidos verifican ciertas propiedades de convergencia y de estimación del error.

Utilizaremos también, sobre todo en los problemas, los espacios de polinomios a trozos por su uso generalizado en Ingeniería, aunque en este caso, debido a su dificultad teórica, no estudiaremos sus propiedades.

Un teorema fundamental en el sentido de dar solidez a la aproximación de funciones mediante polinomios es el de Weierstrass. Este teorema garantiza que para cualquier función continua siempre existe un polinomio [1] que está tan cerca de ella en el sentido de la norma del máximo como nosotros queramos. Su enunciado

[1] Existe un resultado similar para funciones trigonométricas.

riguroso es:

Teorema 4.1.1 *Teorema de aproximación de Weierstrass.*

- *Sea $f \in C[a,b]$ con $a, b \in \mathbb{R}$, $a < b$ una función continua cualquiera.*
- *Cualquiera que sea $\epsilon > 0$, existen un entero $n \in \mathrm{N}$ y un polinomio p de grado n tales que $\|f-p\|_\infty < \epsilon$, con $\|f-p\|_\infty = \max_{x\in[a,b]} |f(x)-p(x)|$*

Pasamos ahora a plantear de modo general el problema de aproximación.

4.2. El problema general de aproximación

Para hablar de aproximación sólo necesitamos un espacio métrico dotado de una distancia que permita dar sentido al concepto de proximidad. Sin embargo, lo más adecuado para el planteamiento general de este libro es restringir el problema a espacios vectoriales normados, en los que la distancia se define a partir de una norma. Como consecuencia, los conjuntos donde vamos a plantear y a resolver nuestros problemas serán los espacios vectoriales normados de funciones de los tipos antes mencionados.

4.2.1. Normas más habituales en espacios de funciones

El espacio vectorial de funciones más general con el que vamos a trabajar es el de las funciones continuas definidas sobre un intervalo cerrado $[a,b]$ de $\mathbb{R}$.

Para normar este espacio disponemos de varias posibilidades. La más común es la norma del máximo, que ya hemos visto con detalle al estudiar el error en la interpolación de Lagrange (pág. 126) y que se define como:

$$\|f\|_\infty := \max_{x\in[a,b]} |f(x)| \tag{4.1}$$

Existen otras dos posibilidades muy interesantes. La norma 1, o integral del valor absoluto de una función:

$$\|f\|_1 := \int_a^b |f(x)|dx \tag{4.2}$$

y la norma 2

$$\|f\|_2 := \sqrt{\langle f, f\rangle} = \left(\int_a^b f^2(x)dx\right)^{1/2}$$

que se deduce del producto escalar [2]

$$\langle f, g\rangle = \int_a^b f(x)g(x)dx$$

Aunque la norma 1 es interesante, no existen resultados que permitan aproximar de modo cómodo en esta norma, y por tanto, salvo para ejemplos muy sencillos, no nos volveremos a referir a ella.

4.2.2. Normas estrictas

Hay un cierto tipo de normas que son especialmente importantes, pues a ellas se refieren algunos teoremas generales de unicidad. Estas normas son las llamadas *estrictas*.

Definición 4.2.1 *Sea $(E, \|\ \|)$ un espacio vectorial normado. Sean $f, g \in E$, $f, g \neq 0$ con $\|f+g\| = \|f\|+\|g\|$. La norma $\|\ \|$ se llama estricta, si existe $\lambda \in \mathbb{R}$ tal que $g = \lambda f$.*

Para una norma estricta, si la norma de la suma de dos vectores es la suma de las normas, los vectores son combinación lineal uno del otro.

Éste es un concepto que cuesta asimilar, así que pondremos algunos ejemplos.

Ejemplo 4.2.1 *$(C([a,b]), \|\ \|_\infty)$ no está estrictamente normado. Veamos un contraejemplo para demostrarlo: $f(x) = 1$, $g(x) = x$ satisfacen:*

$$\|f+g\|_\infty = 1 + b = \|f\|_\infty + \|g\|_\infty$$

y sin embargo, son linealmente independientes.

[2] La norma asociada a un producto escalar es siempre la raíz cuadrada del producto de un elemento por sí mismo.

Ejemplo 4.2.2 *$(R^3, \| \ \|_\infty)$ no está estrictamente normado. $x = (1,0,0)$, $y = (1,1,0)$ que son linealmente independientes, cumplen*

$$\|x\|_\infty = 1, \|y\|_\infty = 1 \quad y \quad \|x+y\|_\infty = 2$$

Ejemplo 4.2.3 *La norma euclídea $\| \ \|_2$ es una norma estricta en $\mathbb{R}^n$. Trata de encontrar dos elementos de $\mathbb{R}^2$ tales que la norma de su suma sea la suma de sus normas y no sean linealmente dependientes. Vas a poder hacerlo con la norma del máximo pero no con la norma euclídea:*

$$\|\mathbf{x}\|_2 = \|(x_1, x_2)\|_2 = \sqrt{x_1^2 + x_2^2}$$

Ejemplo 4.2.4 *$(C([a,b]), \| \ \|_1)$ tampoco está estrictamente normado. Con las mismas f y g de antes*

$$\|1+x\|_1 = \int_0^1 (1+x)dx = \frac{3}{2} = \|1\|_1 + \|x\|_1 = \int_0^1 dx + \int_0^1 xdx = 1 + \frac{1}{2}$$

Es fácil demostrar que una norma no es estricta buscando un contraejemplo. No es sencillo demostrar que lo es.

Se dispone de un teorema que garantiza que ciertas normas son estrictas.

Teorema 4.2.1 *Toda norma asociada a un producto escalar es estricta.*

No vamos a demostrar este teorema, sólo comentaremos que es consecuencia de la desigualdad de Schwarz

$$\langle f, g \rangle \leq \|f\| \cdot \|g\|$$

4.3. Mejor aproximación

Sea $(E, \| \ \|)$ un espacio vectorial normado, T una parte arbitraria de E y v un elemento de E que se desea aproximar mediante elementos de T. Parece razonable decir que $u \in T$ es una buena aproximación de v si la distancia entre ambos, $d(v,u) = \|v-u\|$ es pequeña. La aproximación será la mejor posible si esa distancia es la menor posible, frase que lleva implícita un proceso de minimización de las posibles distancias.

Definición 4.3.1 *Un elemento $\hat{u} \in T$ es una mejor aproximación (m.a.) de v si*

$$\|v - \hat{u}\| \leq \|v - u\| \quad (\forall u \in T)$$

Ejemplo 4.3.1 *Supongamos $E = R^2$ con la distancia euclídea. Sea $T = \{u \in E : \|u\| \leq 1\}$, o sea, el círculo de radio unidad cerrado. Para todo $v \in R^2 - T$ existe un solo $\hat{u} \in T$ tal que*

$$\|v - \hat{u}\|_2 \leq \|v - u\|_2 \quad (\forall u \in T)$$

que es precisamente la intersección de la recta que une v con el origen y la circunferencia unidad como se puede apreciar en la Figura 4.4.

Ejemplo 4.3.2 *Supongamos $E = R^2$ otra vez con la distancia euclídea. Sea $T = \{u \in E : \|u\| < 1\}$, o sea, el círculo de radio unidad abierto; le hemos quitado la circunferencia unidad. Para todo $v \in R^2 - T$ no existe ningún $\hat{u} \in T$ tal que*

$$\|v - \hat{u}\|_2 \leq \|v - u\|_2 \quad (\forall u \in T)$$

pues precisamente se lo hemos quitado.

Ejemplo 4.3.3 *Supongamos $E = R^2$ ahora con la distancia que se deduce de la norma del máximo:*

$$d(u,v) = \|v-u\|_\infty = \text{máx}(|v_1 - u_1|, |v_2 - u_2|)$$

Sea $T = \{u \in E : \|u\|_\infty \leq 1\}$. T es el cuadrado de lado 2 centrado en el origen, como se muestra en la Figura 4.5. Sea $v = (0,2)$. La mejor aproximación a v dentro de T no es única. Está formada por el segmento $\hat{u} = (\hat{u}_1, \hat{u}_2)$, $-1 \leq \hat{u}_1 \leq 1$, $\hat{u}_2 = 1$, pues todos estos puntos están a distancia 1 de v.

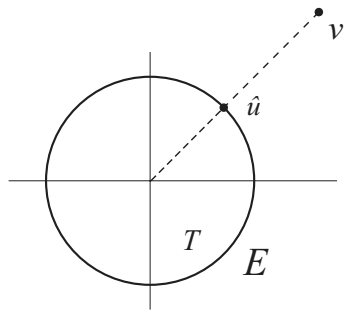


Figura 4.4: Solución del ejemplo 4.3.1.

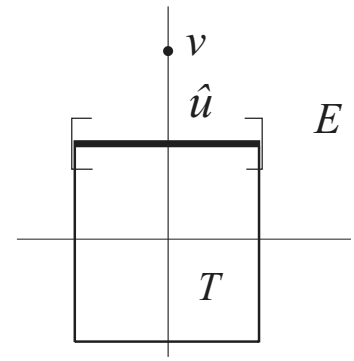


Figura 4.5: Solución del ejemplo 4.3.3.

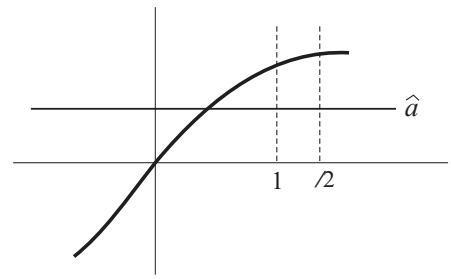


Figura 4.6: Solución del ejemplo 4.3.4.

Ejemplo 4.3.4 *Sean* $(E, \|\ \|) = (C([0,1]), \|\ \|_2)$ *y* T *el conjunto de las funciones de valor constante* $0 \leq a \leq 1$ *en ese intervalo. Hallar una m.a.* $\hat{u} \in T$ *de* v *equivale a hallar* $\hat{a} \in \mathbb{R}$ *tal que*

$$\epsilon(a) = \|v - a\|_2 = \int_0^1 (v(x) - a)^2 dx$$

sea mínimo. Apliquemos el ejemplo a $v(x) = \sin(x)$.

$$\epsilon(a) = \|\sin(x) - a\|_2 = \int_0^1 (\sin(x) - a)^2 dx$$

Si derivamos $\epsilon(a)$ *obtendremos un punto que hace esa derivada cero y que será siempre un mínimo*[3]. *Ese valor es* $\hat{a} = 0.4597$, *que como podemos ver no es la media entre los valores que el seno toma en 0 y 1 (ver Figura 4.6) como la intuición nos podría haber sugerido.*

Como acabamos de ver en los ejemplos anteriores, la mejor aproximación puede o no existir y si existe puede no ser única.

4.4. Aproximación lineal

Aunque los ejemplos que hemos expuestos son interesantes para introducir el problema y para corregir algunas intuiciones apresuradas, en el caso más importante y único que vamos a considerar a partir de ahora, T será un subespacio vectorial de E de dimensión finita n, que denominaremos U por analogía con la notación habitual en álgebra lineal.

Elegida una base B de U, $B = \{u_1, \ldots, u_n\}$, el problema de definir las m.a. de $v \in E$ en U, se reduce a hallar las componentes de sus elementos en la base B. Se trata por tanto, de encontrar

$$\hat{u} = \sum_{i=1}^{n} \alpha_i u_i$$

[3]Se deja como cuestión por qué ese punto de derivada nula será siempre un mínimo y no un máximo.

tal que

$$d(v,\hat{u}) = \left\| v - \sum_{i=1}^{n} \alpha_i u_i \right\|$$

sea mínima para todo $(\alpha_1, \ldots, \alpha_n) \in R^n$.

En los subespacios vectoriales de dimensión finita, las sucesiones convergentes lo hacen dentro del espacio, y esto es suficiente para fundamentar el siguiente teorema de existencia.

Teorema 4.4.1 *Si U es un subespacio de E de dimensión finita, para cada elemento $v \in E$ existe al menos una mejor aproximación $\hat{u} \in U$.*

Ejercicio 4.4.1 *Demostrar que U no es un conjunto acotado (pista: por reducción al absurdo).*

Ejercicio 4.4.2 *Supongamos $E = R^2$ con la distancia euclídea. Sea $U = \{u \in E : u_2 = 0\}$, o sea, el eje x. U es un subespacio de E. Se pide definir el conjunto de las mejores aproximaciones al elemento $v = (0,1)$.*

La respuesta es sencilla. Es un conjunto de un único elemento, el $(0,0)$. Pero ¿qué sucede si cambiamos la norma euclídea por la norma del máximo?

Ejercicio 4.4.3 *Supongamos $E = R^2$ con la distancia que se deduce de la norma del máximo. Sea $U = \{u \in E : u_2 = 0\}$, o sea, el eje x. U es un subespacio de E. Se pide definir el conjunto de las mejores aproximaciones al elemento $v = (0,1)$.*

Lo que sucede es similar a lo que sucedía en el ejemplo 4.3.3; todos los puntos de U entre -1 y 1 están a la misma distancia de v, y por tanto, la m.a. no es única. La única diferencia entre los dos ejemplos es la norma, la primera es estricta y la segunda no. Esto nos lleva al teorema de unicidad de la mejor aproximación en subespacios vectoriales.

Teorema 4.4.2 *Si E está estrictamente normado, la mejor aproximación de $v \in E$ desde un subespacio arbitrario de dimensión finita es única.*

A pesar de su importancia y como vimos en el ejemplo 4.2.1, $(C([a,b]), \| \quad \|_\infty)$ no está estrictamente normado. Por tanto, no podemos establecer conclusiones de unicidad basadas exclusivamente en las propiedades de las normas. Sin embargo, existen subespacios vectoriales de dimensión finita de $(C([a,b]), \| \quad \|_\infty)$ para los que la mejor aproximación es única, por ejemplo, los polinomios de un cierto grado. En el teorema 4.2.1 vimos otras normas que sí eran estrictas y que justifican el siguiente corolario.

Corolario 4.4.1 *En un subespacio U de un espacio E en el que la norma se deduce de un producto escalar(se dice entonces que E es prehilbertiano), el problema de la determinación de una mejor aproximación $\hat{u} \in U$ de $v \in E$ tiene solución única.*

Este corolario se deduce de modo inmediato de los teoremas 4.4.2 y 4.2.1. El primero relaciona unicidad con normas estrictas y el segundo indica que las normas que se deducen de un producto escalar son estrictas.

4.5. Aproximación en espacios prehilbertianos

4.5.1. General

No se pretende asustar a nadie con el título del apartado. Los espacios prehilbertianos[4], como comentábamos en 4.4 son aquellos en los que la norma se deduce de un producto escalar. Podemos ver con dos ejercicios la diferencia entre lo que supone buscar la mejor aproximación en esta norma o en la norma del máximo.

[4]Hilbert, David, 1862-1943. Nació cerca de Kaliningrado (Königsberg), que ahora pertenece a Rusia, pero que en su día formaba parte del imperio prusiano. Contribuyó de modo fundamental al intento de establecer las matemáticas como un conjunto de axiomas y reglas formales, aunque Gödel demostró más tarde que esa tarea tenía limitaciones intrínsecas. En 1900, en un congreso importante propuso una serie de 23 cuestiones que en aquel momento estaban abiertas en matemáticas. La mayoría han sido resueltas a lo largo del siglo XX, pero alguna de las que no lo ha sido, se ha incorporado a los premios del Milenio del millón de dólares en matemáticas, financiados por el millonario americano L. T. Clay.

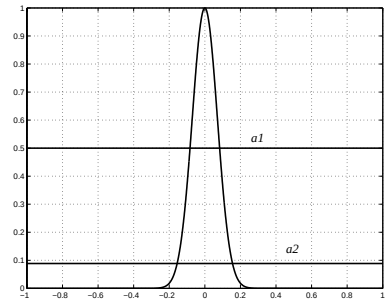


Figura 4.7: Aproximaciones correspondientes a los ejercicios 4.5.1 y 4.5.2.

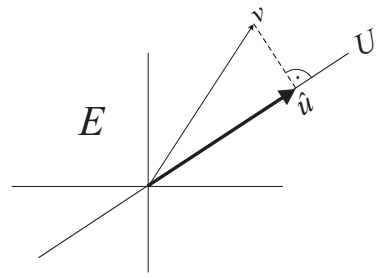


Figura 4.8: Mejor aproximación como proyección ortogonal.

Ejercicio 4.5.1 *Sea E el espacio vectorial $(C([-1,1]), \| \ \|_\infty)$. Sea $v(x) = \exp(-(x/0.1)^2)$. Definir, dentro del subespacio U de E formado por los polinomios de grado 0 (constantes), el conjunto de las m.a. del elemento v.*

Ejercicio 4.5.2 *Sean E el espacio vectorial $(C([-1,1]), \| \ \|_2)$, con $\|f\|_2 = \int_{-1}^{1}(f(x))^2 dx$ y $v(x) = \exp(-(x/0.1)^2)$. Hallar en el subespacio U de E formado por los polinomios de grado 0, el conjunto de las m.a. del elemento v.*

En ambos ejercicios el conjunto pedido se reduce a un único elemento, a_1 y a_2 respectivamente, que se muestran en la Figura 4.7.

Si tuviésemos que sustituir la función v por una de estas constantes, ¿cuál sería la más adecuada?. La respuesta no es única. Si queremos minimizar el error máximo cometido tendríamos que elegir a_1, pero si buscásemos un modelo más global y suavizado de la función v habría que elegir a_2 pues realmente v es casi nula en todo el intervalo.

Ejercicio 4.5.3 *Escribir el código Matlab para dibujar las curvas de la Figura 4.7.*

Las dos aproximaciones son importantes en ingeniería pero quizá la segunda sea más utilizada. El ejemplo es una exageración y si las funciones de aproximación fuesen un poco más elaboradas, la diferencia no sería tan acusada, y estaría más en la línea de lo que vimos en los ejemplos iniciales del capítulo. En el segundo ejemplo, la norma deriva de un producto escalar y aunque el estudiante haya sido capaz de obtener el valor buscado simplemente minimizando la distancia, se pueden elaborar estrategias más generales, a las cuales dedicaremos esta sección.

4.5.2. La mejor aproximación como proyección ortogonal

Hemos visto en el corolario del teorema 4.4.2, que si la norma deriva de un producto escalar $\langle \, , \, \rangle$ la m.a. existe y es única; falta ver cuáles son sus componentes en la base de U, subespacio vectorial de dimensión finita de E.

En la Figura 4.8, se plantea un problema de aproximación donde el subespacio de $\mathbb{R}^2$ en el que se busca la m.a. es una de las rectas que pasan por el origen. Si la norma que se utiliza es la euclídea, la Geometría elemental nos dice que el punto que está más cerca es el pie de la perpendicular desde el punto v a la recta.

El vector $\hat{u}$ verifica entonces que $\langle v - \hat{u}, u\rangle = 0$ para todos los elementos de U. Esta intuición gráfica es generalizable.

Dado que a partir de ahora nuestros espacios lo serán de funciones, cambiaremos la notación para referirnos a los elementos de los espacios por la habitual cuando se trata con funciones. Enunciamos el siguiente teorema y nos remitimos a Hammerlin et al. [15] para su demostración.

Teorema 4.5.1 *Sea E un espacio prehilbertiano real, U un s.e.v. de E. La mejor aproximación $\hat{f}$ de $f \in E$ en U se caracteriza por ser la proyección ortogonal de f sobre U, o sea, $\langle f - \hat{f}, g\rangle = 0\ \forall g \in U$.*

4.5.3. Componentes de la mejor aproximación

Sea B una base de U formada por las funciones (vectores) $\{g_1, \ldots, g_n\}$. Sean $\{c_j\}_{j=1,n}$ n escalares a priori desconocidos. La expresión de $\hat{f}$ en la base B es

$$\hat{f} = \sum_{j=1}^{n} c_j g_j \tag{4.3}$$

el objetivo es determinar los $\{c_j\}_{j=1,n}$.

En la Figura 4.5.3 expusimos las ideas que caracterizaban $\hat{f}$ mediante la proyección ortogonal. Veamos que ello es suficiente para calcular sus componentes. En efecto, $c = (c_1, \ldots, c_n)$ es la solución del sistema de ecuaciones lineales:

$$\langle \hat{f} - f, g_k\rangle = 0 \quad k = 1, n$$

con las que obligamos a que $\hat{f} - f$ sea ortogonal a todos los elementos de la base B, luego a todos los elementos de U. Sustituyendo (4.3) y utilizando las propiedades del producto escalar,

$$\langle \sum_{j=1}^{n} c_j g_j - f, g_k\rangle = 0 \quad k = 1, n \quad \Rightarrow \quad \sum_{j=1}^{n} c_j \langle g_j, g_k\rangle = \langle f, g_k\rangle \quad k = 1, n$$

que expresamos matricialmente

$$\begin{pmatrix} \langle g_1, g_1\rangle & \cdots & \cdots & \cdots & \langle g_n, g_1\rangle \\ \vdots & \cdots & \cdots & \cdots & \vdots \\ \vdots & \cdots & \cdots & \cdots & \vdots \\ \vdots & \cdots & \langle g_j, g_k\rangle & \cdots & \vdots \\ \vdots & \cdots & \cdots & \cdots & \vdots \\ \langle g_1, g_n\rangle & \cdots & \cdots & \cdots & \langle g_n, g_n\rangle \end{pmatrix} \begin{pmatrix} c_1 \\ \vdots \\ c_j \\ \vdots \\ \vdots \\ c_n \end{pmatrix} = \begin{pmatrix} \langle f, g_1\rangle \\ \vdots \\ \vdots \\ \langle f, g_k\rangle \\ \vdots \\ \langle f, g_n\rangle \end{pmatrix} \tag{4.4}$$

La solución de este sistema lineal es siempre única, ya que la independencia lineal de los vectores $g_1, \ldots, g_n$ implica que la matriz de Gram asociada a esa base y a ese producto escalar, que es la matriz $G = \langle g_j, g_k\rangle$ del sistema, es regular. Además, G es simétrica y definida positiva, lo que implica que este sistema lineal esté muy bien condicionado y, aunque el número de incógnitas sea grande, se resuelve muy rápidamente mediante un método iterativo.

Ejercicio 4.5.4 *Demostrar que $\langle f - \hat{f}, g\rangle = 0\ \forall g \in U$ ssi $\langle f - \hat{f}, g_k\rangle = 0\ k = 1, n$, siendo g_k los elementos de la base B.*

Ejemplo 4.5.1 *Retomamos el ejemplo 4.1.1 para hacer el apartado 3, atendiendo a las ideas aquí expresadas. El objetivo de dicho apartado era encontrar una recta r_3 con la que aproximar-sustituir localmente a la función seno entre 0 y $\pi/2$ eligiendo aquella que minimizase el valor $\int_0^{\pi/2}(\sin(x) - r_3(x))^2 dx$.*

Si definimos el marco general en el que estamos trabajando, $E = C([0, \pi/2])$, y el subespacio vectorial donde estamos buscando la m.a. es $U = P_1(\mathbb{R})$, polinomios de grado 1 de coeficientes reales. Tomamos en U la base (canónica) de los monomios, $B = \{1, x\}$ y definimos en E el producto escalar

$$\langle f, g\rangle = \int_0^{\pi/2} f(x)g(x)dx$$

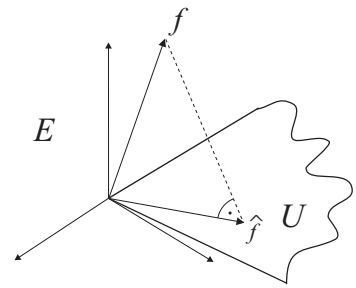


Figura 4.9: Proyección ortogonal.

cuya norma asociada es la norma 2:

$$\|f\|_2 = \sqrt{\langle f, f\rangle} = \left(\int_0^{\pi/2} (f(x))^2 dx\right)^{1/2}$$

Dos elementos de E se parecen *entre sí cuando su distancia (la norma de su diferencia) es pequeña. Se trata de encontrar la recta que más se parece al seno en este intervalo minimizando esa distancia. Como hemos estudiado, si escribimos la recta en función de su base (4.3)*

$$r_3(x) = c_1 \cdot 1 + c_2 \cdot x$$

tendremos

$$\begin{pmatrix} \langle 1,1\rangle & \langle x,1\rangle \\ \langle 1,x\rangle & \langle x,x\rangle \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \begin{pmatrix} \langle \sin,1\rangle \\ \langle \sin,x\rangle \end{pmatrix} \quad \Rightarrow \quad \begin{pmatrix} \pi/2 & \pi^2/8 \\ \pi^2/8 & \pi^3/24 \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

Resolviendo este sistema lineal, llegamos a la recta buscada $r_3(x) = 0.1148 + 0.6644x$.

Este sistema lineal está muy mal condicionado. A medida que subimos el grado del polinomio el condicionamiento del sistema crece de modo exponencial. Para este caso, la matriz tiene un condicionamiento del orden de 15. Si buscásemos una parábola, sería del orden de 450; sería 13600 para una cúbica y del orden de 400000 para una cuártica. Esto quiere decir que los errores de redondeo se amplifican mucho, haciendo imposible resolver este problema a partir de quinto grado utilizando la base de los monomios.

La solución a este inconveniente la veremos en el apartado siguiente.

Ejercicio 4.5.5 *Utilizar la función de Matlab* condest *para estimar el condicionamiento del sistema lineal anterior y del que se obtendría para las parábolas.*

4.5.4. Bases ortogonales

Un caso especialmente favorable se produce cuando los elementos de la base B son ortogonales entre sí respecto al producto escalar utilizado. En ese caso

$$\langle g_j, g_k\rangle = 0 \quad \text{si} \quad j \neq k$$

y las líneas del sistema lineal se reducen a

$$c_k\langle g_k, g_k\rangle = \langle f, g_k\rangle \quad k = 1, n$$

es decir, un sistema lineal diagonal cuya solución es

$$c_k = \frac{\langle f, g_k\rangle}{\langle g_k, g_k\rangle} \quad k = 1, n$$

Si además de ortogonal, nuestra base es ortonormal, $\langle g_j, g_k\rangle = 1$ y

$$c_k = \langle f, g_k\rangle \quad k = 1, n$$

En estos dos últimos casos, los coeficientes c_k son independientes entre sí y añadir más elementos de base no complica apenas el problema de aproximación. Veamos algunos ejemplos de bases ortogonales.

4.5.5. Bases ortogonales: Polinomios de Legendre

En $C([-1,1])$ consideramos el producto escalar habitual

$$\langle f,g\rangle = \int_{-1}^{1} f(x)g(x)dx$$

La familia de polinomios que se obtienen cuando aplicamos el método de ortonormalización de Gram-Schmidt a la base de los monomios es la base de Legendre[5] $\{L_j\}$. Los de grado más bajo son

$$L_0(t) = \frac{1}{\sqrt{2}}, \quad L_1(t) = \sqrt{\frac{3}{2}}t, \quad L_2(t) = \frac{1}{2}\sqrt{\frac{5}{2}}(3t^2-1), \quad L_3(t) = \frac{1}{2}\sqrt{\frac{7}{2}}(5t^3-3t)$$

[5]Legendre, Adrien-Marie, 1752-1835. El vetusto y poco funcional edificio de la Educación de inspiración medieval tenía las horas contadas cuando Adrien Marie Legendre vino al mundo en París en el año 1752. Entonces, la Universidad parisiense se dividía en cuatro facultades, las *tres mayores* (Teología, Derecho y Medicina) y la facultad *menor*, de Artes o de Filosofía, donde se impartían los conocimientos indispensables para ingresar en las *superiores* (desempeñando un rol similar al de los últimos cursos del bachillerato en el modelo actual). La facultad de Artes, por no tener, no disponía ni de instalaciones definidas donde impartir las clases, debiendo recurrir los alumnos a alguno de los 10 colegios autorizados para tal fin. (Idénticas arenas movedizas entre la educación secundaria y la universitaria sobrevivirían en España hasta la entrada en vigor de la ley educativa de Claudio Moyano en 1857.)

Legendre, de una familia acomodada sin ser nobiliaria, tuvo la suerte de ingresar en el único de estos colegios que establecía en sus planes de estudio un año completo (antes de los dos prescritos de filosofía) de matemáticas: el Colegio Mazarino. De titularidad jesuita, contó entre sus ilustres directores con matemáticos de la estatura de Pierre Varignon, Nicolas-Luis de la Caille y Joseph-Francois Marie, amigo y tutor del joven Legendre; y por sus aulas desfilaron alumnos de inmortal categoría: Cassini, D'Alembert, Coulomb o Lavoisier, entre otros.

Comer de las matemáticas en esa Francia prerrevolucionaria era harina de otro costal, y estaba sólo al alcance de unos pocos superdotados (Monge, de origen muy modesto, conseguiría ser profesor en la Escuela de Ingenieros Militares de Mézières, la más famosa escuela anterior a la Revolución). En la cúspide del modelo estaban la Academia de las Ciencias (integrada por 42 científicos repartidos entre las disciplinas matemáticas −geometría, astronomía y mecánica− y el resto de disciplinas −anatomía, química y botánica−) y el Colegio Real (más tarde denominado Colegio de Francia), aunque los ilustres miembros de la Academia solían ocupar plaza en éste. Inmediatamente detrás en el escalafón estaba el profesorado de las Escuelas Reales Militares, y dentro de éstas eran muy codiciados los puestos de examinadores (los que determinaban qué alumnos ingresaban) por la generosa y atractiva descompensación entre los trabajos a realizar y los emolumentos recibidos. La creación de la Escuela Politécnica en 1794 y su ramificación de Escuelas Centrales despejaría en parte el horizonte profesional a los investigadores venideros.

En 1775, el joven Legendre (tras dejar constancia de su talento en diversos trabajos −recuperados para sus manuales por el mismísimo padre Marie−) fue nombrado profesor de matemáticas en la Escuela Militar de París, donde impartió clases durante cinco años. Jamás, en su larga vida, volvería a desempeñar la docencia, aunque la enseñanza de las matemáticas no dejó de interesarle nunca.

El 2 de abril de 1783, Legendre ingresa en la Real Academia de las Ciencias con el cargo de adjunto, vacante en la categoría de mecánica por el nombramiento de Laplace para un puesto de asociado en la misma.

Que haya una medida general determinada por su Majestad para todos los territorios de su Reino [...] o *Que todas las medidas de los Señores se reduzcan a la medida del Rey [...] porque la medida de los nobles aumenta cada año*. Estas exigencias aparecen en los *Cahiers de doleances* (Cuaderno de quejas) que presentan al rey los diputados del Tercer Estado poco antes de la reunión de los Estados Generales en Versalles (1789).

La toma de la Bastilla el 14 de julio de 1789 y los acontecimientos que se precipitan no impedirán que la selecta comunidad científica de la Academia distrajese su atención del descomunal y decisivo proyecto que se traía entre manos: el ambicioso Sistema Métrico Decimal; y que con el tiempo equipararará de una vez por todas las diferentes varas de medir, reductos del feudalismo, que se utilizaban hasta entonces en Europa. Asistimos, junto a la Declaración de Derechos del Hombre y del Ciudadano, al segundo gran monumento que nos legaría la Revolución Francesa. Legendre participó en sus trabajos junto con los científicos más acreditados de la época: químicos, astrónomos, físicos, matemáticos. En 1813 reemplazará a su fallecido amigo y maestro Lagrange en la Oficina de Medidas, organismo rector del proyecto. El sistema de pesos y medidas finalmente adoptado establece que la unidad de longitud deberá extraerse de la naturaleza, que las unidades formen un sistema ligado y que sigan una escala decimal. En España, la introducción del nuevo sistema métrico se retrasará hasta el año 1848, durante el reinado de Isabel II.

Además de gran investigador, Legendre ratificaría en dos espléndidos manuales didácticos (Elementos de geometría y Ensayo sobre la teoría de números) una de las máximas de la revolución: que la ciencia fuera comprensible para un número de ciudadanos cada vez mayor. Una de las características (diferenciándolos de los actuales) de estos manuales, que alcanzaron gran difusión en toda Europa, es que incluían en sus páginas apuntes y reseñas sobre los ultimísimos descubrimientos en la materia.

El fin único de la ciencia es el honor del espíritu humano. La frase, que es de su discípulo y amigo Jacobi, resume a la perfección el talante de Legendre durante sus largos e intensos años de dedicación a las matemáticas y no hubiera tenido inconveniente en utilizarla de epitafio.

Legendre, una considerable montaña de talento y de capacidad de trabajo, tuvo la mala suerte de nacer en una época en la que fue oscurecido por las alargadas sombras de sus geniales y admirados predecesores (Euler y Lagrange) y eclipsado a menudo por unas emergentes fuerzas de la naturaleza que pronto se consolidarían como las cimas más altas de las cotas matemáticas de todos los tiempos (Jacobi, Laplace, Abel, sobre todo Gauss...). Pero puede decir, y no todos pueden hacerlo, que jamás le puso una zancadilla a un colega para colgarse una medalla y que fue siempre fiel a una integridad moral que antepuso a cualquier tipo de agasajo gratuito o injusto.

Ejercicio 4.5.6 *Obtener los polinomios L_0 y L_1 utilizando el método de ortonormalización de Gram-Schmidt a partir de los polinomios 1 y x.*

Una propiedad interesante, que se puede deducir de este proceso de ortonormalización, es que el polinomio de Legendre $\{L_j\}$ tiene exactamente grado j.

4.5.6. Bases ortogonales: Polinomios de Chebyshev

En $C([-1,1])$ consideramos el producto escalar

$$\langle f,g\rangle = \int_{-1}^{1} \frac{f(x)g(x)}{\sqrt{1-x^2}}dx \tag{4.5}$$

el cual está asociado con la función peso

$$w(x) = \frac{1}{\sqrt{1-x^2}}$$

La familia de polinomios de Chebyshev normalizados se suele expresar en función de otra familia no normalizada de polinomios, que se denotan $T_i(x)$ y que verifican que $\|T_0\|^2 = \pi$ y $\|T_i\|^2 = \pi/2$ si $i \neq 0$. Se pueden obtener de diferentes modos, uno de ellos mediante ortogonalización de Gram-Schmidt a partir de la base de los monomios de igual modo que los de Legendre. Dividiéndolos después por su norma obtenemos una familia ortonormal

$$\left\{\frac{T_0}{\|T_0\|}, \frac{T_i}{\|T_i\|}\right\} = \left\{\frac{1}{\sqrt{\pi}}T_0, \sqrt{\frac{2}{\pi}}T_i\right\}, \quad i>1$$

Las formas explícitas de los primeros T_i son

$$\begin{aligned}
T_2(x) &= 2x^2-1\\
T_3(x) &= 4x^3-3x\\
T_4(x) &= 8x^4-8x^2+1\\
T_5(x) &= 16x^5-20x^3+5x\\
T_6(x) &= 32x^6-48x^4+18x^2-1
\end{aligned}$$

y la expresión de los monomios de menor grado en función de la base de Chebyshev es la siguiente.

$$\begin{aligned}
1 &= T_0\\
x &= T_1\\
x^2 &= (T_0+T_2)/2\\
x^3 &= (3T_1+T_3)/4\\
x^4 &= (3T_0+4T_2+T_4)/8\\
x^5 &= (10T_1+5T_3+T_5)/16\\
x^6 &= (10T_0+15T_2+6T_4+T_6)/32
\end{aligned}$$

Los polinomios de Chebyshev se utilizan mucho en Cálculo Numérico en aplicaciones de un nivel que excede el de los contenidos de este libro. En el problema 4.2, los utilizamos para hacer factible una solución sencilla.

Ejercicio 4.5.7 *Obtener los polinomios T_0 y T_1 normalizados utilizando el método de ortonormalización de Gram-Schmidt a partir de los polinomios 1 y x.*

4.6. Desarrollo en serie de Fourier de una función periódica

4.6.1. General

Planteamos el desarrollo en serie de Fourier[6] de una función periódica como un caso particular de aproximación en espacios prehilbertianos con bases ortogonales. Utilizaremos la notación compleja por ser

[6]Fourier, Jean Baptiste Joseph. En 1798 el gobierno de la Francia revolucionaria le encomienda a Napoleón (un año antes de su golpe de estado) una gran expedición a Egipto, en la que además de militares se enrolarán docenas de científicos. Uno

la habitual en este tipo de estudios al ser más compacta y clara [7]. Las funciones periódicas o aparentemente periódicas son muy importantes en Ingeniería. Ellas forman las señales que se reciben en multitud de dispositivos de medida correspondientes a fenómenos oscilatorios como la corriente eléctrica, registros de olas, señales de radio, esfuerzos en motores, etc. Este tipo de funciones merecen un análisis específico a través de la teoría de desarrollos en serie de Fourier, los cuales muestran cuáles son las componentes de la señal en las que realmente está su energía.

Las estudiamos como caso particular de los estudios previos, con la ventaja obvia de tener muy bien definido el contexto.

Definición 4.6.1 *Una función f se dice que es periódica de periodo T si*

$$f(t+T) = f(t)$$

al número $w = \frac{2\pi}{T}$ se le llama frecuencia asociada al periodo T.

Ejercicio 4.6.1 *Demostrar que si T es un periodo de f, entonces kT con $k \in N$, también lo es.*

Definición 4.6.2 *Sea f una función periódica. Al menor de los periodos T se le llama periodo fundamental, y a su frecuencia asociada w_0, frecuencia fundamental.*

El espacio ambiente es $E = L^2([0,T],\mathbb{C})$, funciones de cuadrado integrable, que toman valores en los complejos. Se consideran solamente en el periodo $[0,T]$, aunque es indiferente la selección del periodo considerado $[t, t+T]$. Es importante destacar que no exigimos a las funciones aproximantes que sean continuas (ver Figura 4.10). Este matiz es importante y diferencia los planteamientos correspondientes a esta sección respecto a los de continuidad que se exigían en el Teorema de Weierstrass (4.1.1) para aproximación mediante polinomios.

Existe un teorema equivalente al de Weierstrass para aproximaciones trigonométricas pero escapa al contenido de esta sección, dado que aquí la norma considerada es la norma 2.

El producto escalar del que se deriva la norma es:

$$\langle f, g\rangle := \int_0^T f(u)\overline{g(u)}du \tag{4.6}$$

de los más significados es Jean Baptiste Joseph Fourier (Auxerre-1768, París-1830), entonces ya uno de los más prometedores matemáticos. El resultado de tan ambiciosa campaña se materializa en la *Description de l'Egypte*, una obra publicada en 23 volúmenes entre 1809 y 1829. Su introducción general (realizada por Fourier) justificaba la expedición napoleónica como un hecho necesario, como una excusa para modernizar Egipto, ya que *esta región que ha transmitido sus conocimientos a tantas naciones, está hoy inmersa en la barbarie*. La expedición buscaba, pues, ofrecer a Oriente el ejemplo útil de Europa: hacer la vida de sus habitantes más llevadera, procurándoles las ventajas de una civilización perfeccionada.

Eso sí, sin preguntarles ni contar con ellos; mezclándose lo menos posible con su cultura o modos de vida. Edward Said (Jerusalén-1935, Nueva York-2003), palestino de origen y norteamericano de adopción, comentando en su libro *Orientalismo* (1978, Editorial Debate) la *Description de l' Egypte* lo expresa así: *formular Oriente, darle una forma, una identidad y una definición [...] dignificar todos los conocimientos almacenados durante la ocupación colonial con el título de Contribución a la ciencia moderna, cuando los nativos no habían sido consultados y sólo habían sido tenidos en cuenta como pretextos para un texto que ni siquiera les era útil a ellos [...]*. Concluyendo: lo que no era más que un choque entre el ejército conquistador y el ejército derrotado se metamorfoseó en un proceso más extenso y prolongado, barnizado de ideales y buenas intenciones, que salvaguarde el delicado sueño de la sensibilidad europea.

No fueron, no obstante, los únicos méritos de Jean Baptiste Joseph Fourier para pasar a la historia, ni mucho menos. De origen modesto (su padre era sastre), estudió en la Escuela Militar de Auxerre (dirigida por monjes benedictinos) despuntando desde el principio en las Matemáticas, al punto de abandonar una abadía benedictina donde había ingresado con la intención de ordenarse sacerdote para dar clases en la misma escuela militar de la que había sido alumno. En plena Revolución Francesa, en el año 1794, se incorpora a la primera promoción de la Escuela Normal Superior de París (que formaba profesores), donde conoce a Lagrange, Laplace y Monge. Tres años después, con 29 años, ocupa la plaza de asistente de Análisis y Mecánica de la Escuela Politécnica que dejaba vacante Monge, donde impartirá clases hasta su partida a Egipto.

En 1801 Fourier regresó a París y retomó su plaza de profesor de Análisis en la Escuela Politécnica, interesándose particularmente por la cuestión de la propagación del calor. En 1807 publica al respecto *Principios de propagación del calor en cuerpos sólidos*. El año en que fue elegido miembro de la Academia de las Ciencias de París (1822) aparece su *Teoría analítica del calor*. Aquí se presenta lo que se conoce hoy como teorema de Fourier: cualquier oscilación periódica, por complicada que sea, puede descomponerse en una serie de movimientos simples y regulares, cuya suma es la variación periódica compleja original.

La onda expansiva de estos descubrimientos (con repercusiones tanto en el desarrollo del análisis matemático como en aplicaciones a la física) garantizarán a Fourier un lugar destacado entre los más eminentes matemáticos.

[7] Todos los resultados previos de existencia y unicidad siguen siendo válidos para funciones complejas. Hay una parte del tutorial de Matlab dedicada a números complejos.

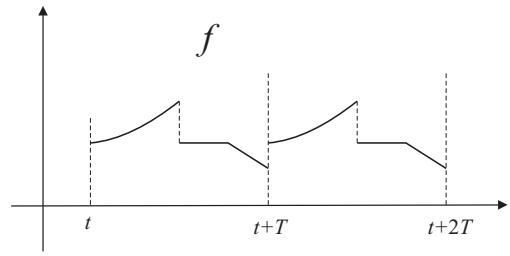


Figura 4.10: Función de periodo fundamental T. El bloque se repite desde $-\infty$ hasta $+\infty$.

Con $\overline{g(t)}$ indicamos el conjugado complejo de $g(t)$. Este conjugado es necesario para garantizar que el producto escalar de una función por sí misma sea un número real positivo. Si consideramos una función f constante cuyo valor siempre es $j = \sqrt{-1}$, se tendría que:

$$\langle f, f\rangle = \int_0^T j\ j\ du = -\int_0^T du = -T$$

Para evitar esta inconsistencia, se define el producto escalar como en 4.6, en cuyo caso,

$$\langle f, f\rangle = \int_0^T j\ \bar{j}\ du = \int_0^T du = T$$

El subespacio vectorial U donde se busca la mejor aproximación es la envoltura lineal del conjunto descrito por las funciones

$$\phi_k(t) = e^{jkw_0t} = \cos(kw_0t) + j\sin(kw_0t)$$

con $k = 0, \pm1, \pm2, \cdots, \pm n$, espacio de las combinaciones lineales de los miembros de esa familia

$$U = L\left(\phi_k(t)\right)_{k=0,\pm1,\pm2,\cdots,\pm n} \tag{4.7}$$

4.6.2. Ortogonalidad de la base

Vamos a ver que este conjunto de funciones es un sistema ortogonal para el producto escalar 4.6.

$$\langle \phi_k, \phi_m\rangle = \int_0^T \phi_k(u)\overline{\phi_m(u)}du = \int_0^T e^{jkw_0u}e^{-jmw_0u}du = \int_0^T e^{j(k-m)w_0u}du$$

$$= \begin{cases} \{k = m\} & ; \quad \int_0^T du = T \\ \{k \neq m\} & ; \quad \frac{1}{j(k-m)w_0}e^{j(k-m)w_0u}\Big|_0^T \end{cases}$$

Si tenemos en cuenta que $w_0T = 2\pi$,

$$e^{j(k-m)w_0u}\Big|_0^T = e^{j(k-m)w_0T} - e^{j(k-m)w_00} = e^{j(k-m)2\pi} - e^{j(k-m)0} = 1 - 1 = 0$$

de donde la conclusión

$$\langle \phi_k, \phi_m\rangle = \begin{cases} T & \text{si} \quad k = m \\ 0 & \text{si} \quad k \neq m \end{cases}$$

Ejercicio 4.6.2 *Comprobar que si el intervalo de integración para el producto escalar se traslada, la familia de funciones 4.7 sigue siendo ortogonal. El "nuevo" producto escalar será:*

$$\langle f, g\rangle := \int_t^{t+T} f(u)\overline{g(u)}du$$

A partir de ahora, en todas las integrales el intervalo de integración tendrá como longitud un periodo, e independientemente del punto en el que empiece, las denotaremos

$$\langle f, g\rangle = \int_T f(u)\overline{g(u)}du$$

4.6.3. Cálculo de la mejor aproximación

Vamos a determinar en este caso la mejor aproximación en U de una función $f \in E$. Como E es prehilbertiano y U tiene dimensión finita, $2n+1$, la mejor aproximación $\hat{f}$ existe, es única y es la proyección de f sobre U.

Para ello, debemos determinar los coeficientes c_k que permiten expresar $\hat{f}$ en la base de U (4.3).

$$\hat{f} = \sum_{m=-n}^{n} c_m \phi_m$$

Como $\hat{f}$ es la proyección ortogonal de f sobre U, para $-n \leq k \leq n$:

$$\langle f - \hat{f}, \phi_k \rangle = 0 \Leftrightarrow \langle \hat{f}, \phi_k \rangle = \langle f, \phi_k \rangle \Leftrightarrow \langle \sum_{m=-n}^{n} c_m \phi_m, \phi_k \rangle = \langle f, \phi_k \rangle \quad \Rightarrow \quad \sum_{m=-n}^{n} c_m \langle \phi_m, \phi_k \rangle = \langle f, \phi_k \rangle$$

y ya que el sistema es ortogonal,

$$c_k = \frac{1}{T} \langle f, \phi_k \rangle = \frac{1}{T} \int_T f(u) e^{-jkw_0 u} du \tag{4.8}$$

Un ejemplo básico de aplicación de esta técnica lo encontramos en el problema 4.1.

Ejercicio 4.6.3 *Se tiene la siguiente función periódica:*

$$f(t) = \begin{cases} 0 & -1.3 \leq t < 0.2 \\ 1 & 0.2 \leq t \leq 0.7 \\ 0 & 0.7 < t \leq 2 \end{cases}$$

Se pide calcular la expresión general de los coeficientes del desarrollo en serie de Fourier. Se pide también escribir unas líneas Matlab para representar la función aproximación, modificando las correspondientes al problema 4.1.

4.6.4. Condición para que la mejor aproximación sea real

La mejor aproximación $\hat{f}$ será real si su parte imaginaria es cero, es decir, si es igual a su conjugado

$$\hat{f} = \overline{\hat{f}} \;\Leftrightarrow\; 0 = \hat{f} - \overline{\hat{f}} = \sum_{k=-n}^{n} c_k \phi_k - \sum_{k=-n}^{n} \overline{c_k \phi_k}$$

Pero el conjugado del producto de dos números complejos es el producto de los conjugados de esos dos números. Por tanto

$$0 = \hat{f} - \overline{\hat{f}} = \sum_{k=-n}^{n} c_k \phi_k - \sum_{k=-n}^{n} \overline{c_k}\,\overline{\phi_k}$$

Es fácil ver que $\overline{\phi_k} = \phi_{-k}$, y que

$$0 = \hat{f} - \overline{\hat{f}} = \sum_{k=-n}^{n} c_k \phi_k - \sum_{k=-n}^{n} \overline{c_k} \phi_{-k}$$

Indexemos las funciones en el segundo sumatorio igual que en el primero

$$0 = \hat{f} - \overline{\hat{f}} = \sum_{k=-n}^{n} c_k \phi_k - \sum_{k=-n}^{n} \overline{c_{-k}} \phi_k = \sum_{k=-n}^{n} \left(c_k - \overline{c_{-k}}\right) \phi_k$$

La función $\hat{f} - \overline{\hat{f}}$ debe ser nula independientemente del punto donde la evaluemos. Al escribirla como combinación lineal de las funciones de U, esto sólo sucederá si todos los coeficientes de la expansión son nulos, y por tanto, la condición necesaria y suficiente para que la función m.a. sea real es que:

$$c_k = \overline{c_{-k}} \;\; (\forall k) \tag{4.9}$$

4.6.5. Cómo es la mejor aproximación si f es real

Veamos que entonces la mejor aproximación es real

$$c_k = \frac{1}{T}\int_T f(u)e^{-jkw_0u}du$$

$$\begin{aligned}
\overline{c_{-k}} &= \frac{1}{T}\overline{\int_T f(u)e^{jkw_0u}du} \\
&= \frac{1}{T}\overline{\int_T f(u)\cos(kw_0u)\,du + j\int_T f(u)\sin(kw_0u)\,du} \\
&= \frac{1}{T}\int_T f(u)\cos(kw_0u)\,du - j\int_T f(u)\sin(kw_0u)\,du \\
&= \frac{1}{T}\int_T f(u)e^{-jkw_0u}du \;=\; c_k
\end{aligned}$$

por tanto, $\overline{c_{-k}} = c_k$, y su serie de Fourier es también real. O sea, que la serie de Fourier de una señal real es también real. Veamos en qué se convierte dicha serie de Fourier.

$$\begin{aligned}
c_k &= \frac{1}{T}\int_T f(u)e^{-jkw_0u}du \\
&= \frac{1}{T}\left[\int_T f(u)\cos(kw_0u)\,du - j\int_T f(u)\sin(kw_0u)\,du\right] = A_k + jB_k
\end{aligned}$$

y

$$A_k = \frac{1}{T}\int_T f(u)\cos(kw_0u)\,du \quad \text{y} \quad B_k = -\frac{1}{T}\int_T f(u)\sin(kw_0u)\,du$$

Como $\overline{c_{-k}} = c_k$, $c_{-k} = A_k - jB_k$ tendremos que

$$\begin{aligned}
\hat{f}(t) &= \sum_{k=-n}^{n} c_k\phi_k(t) = c_0 + \sum_{k=-n}^{1} c_k\phi_k(t) + \sum_{k=1}^{n} c_k\phi_k(t) = c_0 + \sum_{k=1}^{n} c_{-k}\phi_{-k}(t) + \sum_{k=1}^{n} c_k\phi_k(t) \\
&= c_0 + \sum_{k=1}^{n} \overline{c_k}\phi_{-k}(t) + \sum_{k=1}^{n} c_k\phi_k(t) = c_0 + \sum_{k=1}^{n}\left[(A_k - jB_k)\,e^{-jkw_0t} + (A_k + jB_k)\,e^{jkw_0t}\right] \\
&= c_0 + \sum_{k=1}^{n}\left[A_k\left(e^{jkw_0t} + e^{-jkw_0t}\right) + jB_k\left(e^{jkw_0t} - e^{-jkw_0t}\right)\right]
\end{aligned}$$

con lo que

$$\hat{f}(t) = c_0 + 2\sum_{k=1}^{n} A_k\cos(kw_0t) - 2\sum_{k=1}^{n} B_k\sin(kw_0t) \tag{4.10}$$

que es el conocido desarrollo en serie de senos y cosenos de una función f. Un ejemplo básico de aplicación de esta técnica lo encontramos en el problema 4.1.

4.7. Aproximación discreta: mínimos cuadrados

4.7.1. Introducción

Hasta ahora hemos aproximado funciones definidas mediante una expresión analítica con otras funciones más sencillas definidas de igual forma. Sin embargo, lo más habitual es que la función que se pretende aproximar no se conozca de modo analítico, sino a través de una definición puntual, por ejemplo como una nube finita de puntos obtenida como resultado de un determinado experimento. A este tipo de problema nos dedicaremos en esta sección.

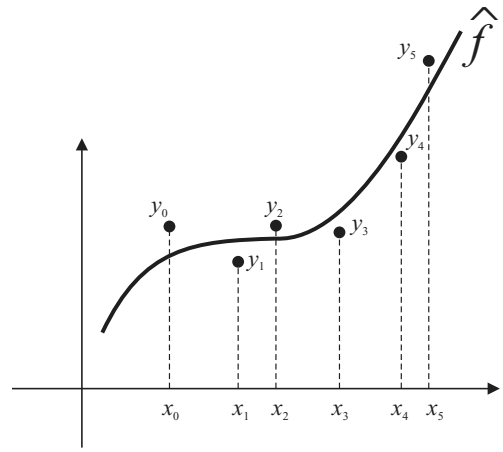


Figura 4.11: Aproximación por mínimos cuadrados.

La construcción de modelos sustitutorios de funciones mediante la minimización del error cuadrático, es una parte fundamental de los métodos estadísticos utilizados en Economía y en los métodos predictivos propios de la Ingeniería Financiera. En estos ámbitos, los modelos utilizados suelen ser lineales en las variables independientes y se suelen llamar métodos de regresión lineal, pero la técnica subyacente no difiere en esencia de la que aquí vamos a estudiar. Dada la reciente orientación que muchos ingenieros toman en su vida profesional hacia el mundo financiero, conviene no olvidar esta relación.

La otra aplicación estrella de las aproximaciones discretas es el análisis de señales discretas, sobre todo cuando se trata de encontrar en éstas patrones periódicos. A eso dedicaremos también esfuerzo en estas indicaciones teóricas y en los problemas, aunque sólo de un modo superficial, ya que cae dentro de una teoría más global, la correspondiente al tratamiento digital de señales.

Vamos a abordar el método de aproximación por mínimos cuadrados que todos conocéis, discutiéndolo como un caso particular de aproximación en un espacio prehilbertiano adecuado.

Se tienen $\{x_i, y_i\}$, $i = 0, n,\ n+1$ pares de puntos de abscisas distintas pertenecientes al grafo de una función f desconocida, pero que en principio supondremos continua.

Buscamos una función $\hat{f}$ perteniente a un espacio vectorial de funciones U, una de cuyas bases es $B = \{g_1, \ldots, g_m\}$ (ver Figura 4.11), cuyos valores en $(x_0, \ldots, x_n)$ aproximen los valores $(y_0, \ldots, y_n)$ tan bien como sea posible [8]. Para estudiar la bondad de la aproximación definimos la siguiente medida del error para todos los elementos de U.

$$\begin{aligned} E^2 : U &\rightarrow \mathbb{R} \\ g &\rightarrow \sum_{i=0}^{n} |y_i - g(x_i)|^2 \end{aligned}$$

Nuestro objetivo es, por tanto, encontrar $\hat{f} \in U$ que minimice ese error:

$$E^2(\hat{f}) \leq E^2(g) \quad \forall g \in U$$

Ejemplo 4.7.1 *Se conocen tres puntos de una función f,*

[8] Existe una relación entre los valores n y m. El número de funciones utilizadas para aproximar los valores debe ser menor o igual que la cantidad de valores disponibles, pues si no fuese así tendríamos un problema sobredeterminado. Por ejemplo, imaginemos el problema de encontrar una parábola que minimice el error correspondiente a una nube de 2 puntos. Existen infinitas parábolas que pasan por esos dos puntos y todas ellas tienen por tanto un error nulo. Por tanto, para que el problema de aproximación por mínimos cuadrados tenga sentido, la información disponible tiene que ser *más grande* que el número de funciones de base con las que la vamos a aproximar. Si el número de funciones de base es el mismo que el número de puntos, la función que obtendremos pasará por todos los puntos, y habremos vuelto al problema de interpolación de Lagrange.

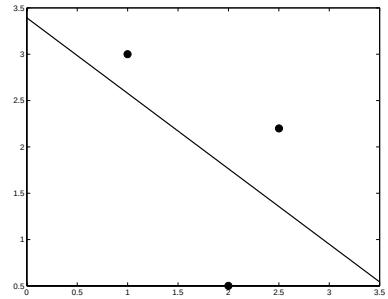


Figura 4.12: Recta correspondiente al ejemplo 4.7.1.

i	x_i	y_i
0	*1.0*	*3.0*
1	*2.0*	*0.5*
2	*2.5*	*2.2*

y buscamos una recta r para sustituir a esa función, eligiéndola de tal modo que minimice el error E^2 con respecto a esa nube de puntos. El conjunto de funciones U donde estamos buscando esa aproximación es $P_1(\mathbb{R})$. Se considerará la base de los monomios en U, $B = \{1, x\}$.
Poniendo la ecuación de la recta r en la forma $r(x) = ax + b$, el error E^2 es

$$E^2(r) = \sum_{i=0}^{2} (y_i - r(x_i))^2 = \sum_{i=0}^{2} (y_i - ax_i - b)^2$$

debemos encontrar un mínimo de esta función, o sea un extremo de la misma que sea mínimo. Para encontrar un extremo de esta función calculamos sus derivadas paraciales respecto a los coeficientes de los que depende, o sea, a y b y las igualamos a 0.

$$\frac{\partial E^2(r)}{\partial a} = 2\sum_{i=0}^{2} (y_i - ax_i - b)(-x_i) = 0 \tag{4.11}$$

$$\frac{\partial E^2(r)}{\partial b} = 2\sum_{i=0}^{2} (y_i - ax_i - b)(-1) = 0 \tag{4.12}$$

Desarrollando estas dos ecuaciones

$$\begin{aligned} y_0x_0 - ax_0^2 - bx_0 + y_1x_1 - ax_1^2 - bx_1 + y_2x_2 - ax_2^2 - bx_2 &= 0 \\ y_0 - ax_0 - b + y_1 - ax_1 - b + y_2 - ax_2 - b &= 0 \end{aligned}$$

un sistema lineal con a y b como incógnitas, que podemos escribir matricialmente en la forma

$$\begin{pmatrix} x_0^2 + x_1^2 + x_2^2 & x_0 + x_1 + x_2 \\ x_0 + x_1 + x_2 & 3 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} y_0x_0 + y_1x_1 + y_2x_2 \\ y_0 + y_1 + y_2 \end{pmatrix} \Rightarrow \begin{pmatrix} 11.25 & 5.50 \\ 5.50 & 3 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} 9.5 \\ 5.7 \end{pmatrix}$$

cuya solución define la recta $r(x) = -0.8143x + 3.3929$. Vemos la nube y la recta en la Figura 4.12.

Ejercicio 4.7.1 *Escribir el código Matlab para resolver el ejemplo anterior y representar conjuntamente la recta solución y la nube de puntos de partida.*

Podemos ver también las ecuaciones (4.11) y (4.12) correspondientes al ejemplo 4.7.1 del siguiente modo

$$\left(\begin{pmatrix} y_0 \\ y_1 \\ y_2 \end{pmatrix} - a \begin{pmatrix} x_0 \\ x_1 \\ x_2 \end{pmatrix} - b \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \right) \cdot \begin{pmatrix} x_0 \\ x_1 \\ x_2 \end{pmatrix} = 0 \tag{4.13}$$

$$\left(\begin{pmatrix} y_0 \\ y_1 \\ y_2 \end{pmatrix} - a \begin{pmatrix} x_0 \\ x_1 \\ x_2 \end{pmatrix} - b \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \right) \cdot \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = 0 \tag{4.14}$$

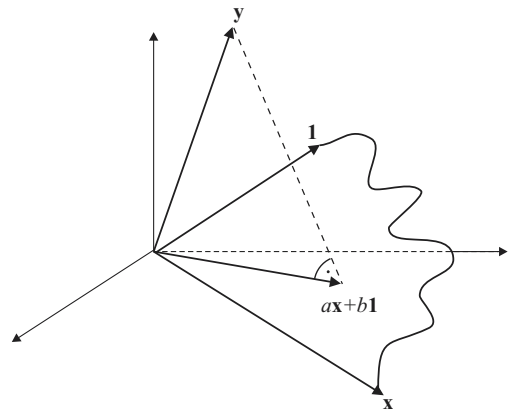


Figura 4.13: Aproximación por mínimos cuadrados: proyección ortogonal.

donde $\cdot$ es el producto escalar [9] en $\mathbb{R}^3$. Esta notación sugiere definir los siguientes vectores:

$$\mathbf{y} = \begin{pmatrix} y_0 \\ y_1 \\ y_2 \end{pmatrix}, \quad \mathbf{x} = \begin{pmatrix} x_0 \\ x_1 \\ x_2 \end{pmatrix}, \quad \mathbf{1} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

y escribir entonces las ecuaciones (4.13) y (4.14) del siguiente modo

$$(\mathbf{y} - a\mathbf{x} - b\mathbf{1}) \cdot \mathbf{x} = 0, \quad (\mathbf{y} - a\mathbf{x} - b\mathbf{1}) \cdot \mathbf{1} = 0$$

Estas ecuaciones expresan que el vector de $\mathbb{R}^3$ $\{\mathbf{y} - a\mathbf{x} - b\mathbf{1}\}$ es ortogonal a los dos vectores $\mathbf{x}$ y $\mathbf{1}$ (ver Figura 4.13), y por tanto, el vector $\{a\mathbf{x} + b\mathbf{1}\}$ es la proyección ortogonal del vector $\mathbf{y}$ sobre el plano engendrado por los dos vectores $\mathbf{x}$, $\mathbf{1}$.

$$\langle \mathbf{y} - (a\mathbf{x} + b\mathbf{1}), \mathbf{x} \rangle = 0, \quad \langle \mathbf{y} - (a\mathbf{x} + b\mathbf{1}), \mathbf{1} \rangle = 0$$

Esta idea de ortogonalidad, que ya nos es familiar para encontrar la m.a. en espacios en los que la norma deriva de un producto escalar, es la que nos va a proporcionar también la m.a. en el caso de los mínimos cuadrados, como veremos en la sección siguiente.

4.7.2. Seminorma para el problema de mínimos cuadrados

Con el ejemplo 4.7.1 y los comentarios subsiguientes hemos sugerido que la solución al problema de mínimos cuadrados la vamos a poder encontrar mediante la proyeccción ortogonal. En esta sección precisaremos estas ideas. Lo primero es comprender muy bien el salto del continuo al discreto. Ello nos permitirá entender después la elección del espacio ambiente, y la del subespacio donde buscar la m.a.

Consideramos una función continua f, de la que conocemos su valor en una nube de puntos $\{x_i\}$, $i = 0, n$. Sean y_i estos valores. Llamando a al mínimo valor de las abscisas x_i y b al máximo. Supongamos que $f \in E = C[a, b]$. Sea U un subespacio vectorial de E, por ejemplo el de los polinomios de un determinado grado. El problema de aproximación consistirá en encontrar $\hat{f} \in U$, tal que

$$\sum_{i=0}^{n} \left| f(x_i) - \hat{f}(x_i) \right|^2$$

[9] Estamos abusando un poco de la notación matricial al manejar estos vectores, pues deberíamos trasponer las matrices para que este producto tuviera sentido, así

$$\left(\begin{pmatrix} y_0 & y_1 & y_2 \end{pmatrix} - a \begin{pmatrix} x_0 & x_1 & x_2 \end{pmatrix} - b \begin{pmatrix} 1 & 1 & 1 \end{pmatrix} \right) \cdot \begin{pmatrix} x_0 \\ x_1 \\ x_2 \end{pmatrix} = 0$$

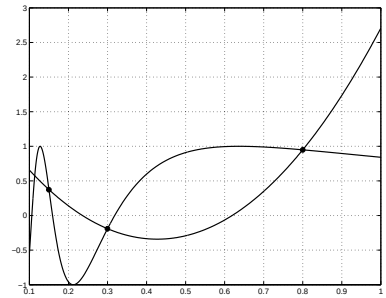


Figura 4.14: Funciones del ejercicio 4.7.2.

sea mínimo. O sea, que estamos buscando en U una función $\hat{f}$ tal que la distancia entre f y $\hat{f}$ medida de ese modo

$$d(f,\hat{f}) := \sum_{i=0}^{n} \left| f(x_i) - \hat{f}(x_i) \right|^2 \tag{4.15}$$

sea mínima. Pero ¿es esta función una distancia? Veámoslo con el siguiente ejercicio.

Ejercicio 4.7.2 *Se tiene el conjunto de abscisas* $\{0.15, 0.3, 0.8\}$. *Se consideran la función* $f(x) = \sin(1/x)$ *y la parábola* $p(x) = 9.2983x^2 - 7.9490x + 1.3573$. *Se pide calcular la distancia entre* f *y* p *definida en 4.15.*

Si dibujamos las curvas anteriores, Figura 4.14, vemos que coinciden en las abscisas de la nube, y que por tanto, su distancia medida de ese modo es nula. Sin embargo, las dos funciones son distintas, hay elementos distintos entre sí, cuya distancia es nula, luego lo que hemos definido no es realmente una distancia. En realidad, no estamos diciendo nada muy interesante; que dos funciones coincidan en una serie de puntos, no significa que sean la misma [10]. A pesar de ello, intentemos aprovechar este concepto a ver si podemos extraer de él algo más. Para ello, extendemos el concepto de distancia definiendo primero un *pseudo*-producto escalar entre funciones asociado a una nube de puntos $\{x_i\}$, $i = 0, n$ por

$$\langle f, g \rangle := \sum_{i=0}^{n} f(x_i)\, g(x_i) \tag{4.16}$$

No es un producto escalar propiamente dicho porque hay funciones h no nulas tales que el producto escalar $\langle h, h \rangle$ es nulo.

Ejercicio 4.7.3 *Encontrar una función* h *no nula tal que el* pseudo*-producto escalar 4.16,* $\langle h, h \rangle$*, asociado a la nube de puntos* $\{0.15, 0.3, 0.8\}$ *sea nulo (Indicación: probar con* $f - p$ *del ejercicio 4.7.2).*

Podemos definir un vector $\mathbf{f} \in I\!R^{n+1}$ asociado a la función f como el resultado de evaluarla en la nube de puntos:

$$\mathbf{f} := (f(x_0), \ldots\ldots, f(x_n)) \tag{4.17}$$

Y resulta que el *pseudo*-producto escalar definido en $C[a,b]$ se traslada a $I\!R^{n+1}$ a través de estos vectores poniendo

$$\langle f, g \rangle = \langle \mathbf{f}, \mathbf{g} \rangle \tag{4.18}$$

Y aquí sí es un producto escalar. Asociado al *pseudo*-producto escalar 4.16 podemos definir una función que verifica todas las propiedades correspondientes a las normas excepto la de ser definida positiva, es decir, va a haber elementos no nulos cuya imagen por esa función es nula:

$$\sqrt{\langle f, f \rangle} := \sqrt{\sum_{i=0}^{n} [f(x_i)]^2} \tag{4.19}$$

[10] El proceso de restricción de una función definida en un conjunto a un subconjunto es único; el proceso de extensión a todo el conjunto de una función definida en una parte, no.

Se llama a esa función una seminorma que nos va a permitir definir una *pseudo*-distancia como siempre

$$d(f,g) := \|f-g\| \tag{4.20}$$

Esta pseudo-distancia entre funciones trasladada a los vectores correspondientes en $I\!R^{n+1}$

$$d(f,g) = \|f-g\| = \langle f-g, f-g\rangle^{1/2} = \langle \mathbf{f}-\mathbf{g}, \mathbf{f}-\mathbf{g}\rangle^{1/2} = \|\mathbf{f}-\mathbf{g}\|$$

es la distancia asociada a la norma euclídea en $I\!R^{n+1}$. Por tanto, lo que estamos haciendo para medir la distancia entre funciones es evaluar éstas en una serie de puntos, y medir la distancia entre los vectores formados por estas imágenes, como la norma de su diferencia.

Ejercicio 4.7.4 *Calcular la distancia en $I\!R^3$ entre los puntos*

$$P = (0.3742, -0.1906, 0.9490) \quad y \quad Q = (0.4274, 0.4966, 0.8187)$$

Calcular la distancia 4.20 entre las funciones $f(x) = \sin(1/x)$ y $g(x) = \exp(x-1)$. La distancia está asociada a la nube de puntos $\{0.15, 0.3, 0.8\}$.

4.7.3. Solución del problema de mínimos cuadrados

Una vez que hemos planteado el problema de este modo, después de haber definido de modo más riguroso que la distancia entre las funciones deriva de una seminorma y ésta a su vez de un producto escalar, y después de trasladar este producto escalar a $I\!R^{n+1}$, la m.a. es simplemente la proyección ortogonal de f sobre U. Por tanto, la mejor aproximación $\hat{f}$ verificará que:

$$\langle f - \hat{f}, g\rangle = 0 \quad \forall g \in U$$

Para lo que es suficiente y necesario que lo anterior se cumpla, como ya hemos visto, para todos los elementos de una base de U. Definamos $\hat{f} = \sum_{j=1}^{m} c_j g_j$ (4.3)

$$\langle f - \sum_{j=1}^{m} c_j g_j, g_k\rangle = 0 \quad k = 1, m$$

Transportando este producto escalar $I\!R^{n+1}$ se obtienen las componentes de la solución resolviendo el sistema lineal

$$\sum_{j=1}^{m} c_j \langle \mathbf{g_j}, \mathbf{g_k}\rangle = \langle \mathbf{f}, \mathbf{g_k}\rangle = 0, \quad k = 1, m \tag{4.21}$$

Donde la transformación a vectores de $I\!R^{n+1}$ de los elementos de la base g_k, $k = 1, n$, se hace como se ha indicado en la expresión 4.17. En la sección siguiente veremos que es posible en algunos casos conseguir que esta matriz sea diagonal jugando otra vez con la idea de ortogonalidad, lo que hace obvia su resolución.

Ejercicio 4.7.5 *Rehacer el ejemplo 4.7.1 desde esta perspectiva y comprobar que los resultados coinciden.*

De este modo, hemos encontrado un elemento de U, cuyo transformado en $I\!R^{n+1}$ pertenece al subespacio de $I\!R^{n+1}$ engendrado por los vectores $\{\mathbf{g_1}, \ldots, \mathbf{g_m}\}$.

De salida, las funciones de la base B de U son linealmente independientes. Que los vectores transformados $\{\mathbf{g_1}, \ldots, \mathbf{g_m}\}$ lo sean también, no es obvio. Para los polinomios esto es cierto, pero si consideramos las funciones $\cos(x)$ y $\cos(2x)$ y la nube de puntos $\{0, 2\pi, 4\pi, \ldots\}$, vemos que los vectores que se obtienen al evaluarlas en esos puntos, son iguales, y por tanto, linealmente dependientes, a pesar de que las funciones de partida no lo son.

Éste es un tema que se comenta para satisfacer la curiosidad del estudiante al que le haya surgido esta duda, pero sobre el que no se va a incidir. Los polinomios y las funciones trigonométricas cuando la nube verifica propiedades muy elementales nunca van a dar problemas.

Ejercicio 4.7.6 *Resolver con Matlab, escribiendo el código correspondiente, el ejemplo 4.7.1.*

Ejercicio 4.7.7 *Se conocen tres puntos de una función f,*

i	x_i	y_i
0	*0.15*	*0.3742*
1	*0.30*	*-0.1906*
2	*0.80*	*0.9490*

y se trata de encontrar una función $\hat{f} = c_1 \cos x + c_2 \cos(2x)$ con la que sustituir a esa función, eligiéndola de tal modo que minimice el error cuadrático con respecto a esa nube de puntos. Se pide resolver este problema planteando el sistema lineal 4.21, escribiendo el código Matlab correspondiente y mostrando la gráfica solución, así como la nube aproximada.

4.8. Transformada de Fourier discreta

4.8.1. General

Un caso particular extremadamente importante en lo que se refiere a funciones definidas de modo discreto; son las series de datos correspondientes al muestreo en el tiempo de determinadas variables, las cuales podemos considerar en principio periódicas. La cuestión que se plantea a partir de este muestreo es la reconstrucción y manipulación de la función original.

Estas series temporales de datos y su consecuente análisis aparecen en prácticamente todas las ramas de la ingeniería pero son sin duda centrales para los ingenieros de Telecomunicación. Ellos agrupan todas estas técnicas de análisis bajo el término de Teoría Digital de la Señal y constituye una disciplina fundamental en su formación. Aquí no podemos tratarlo en profundidad, pero sí haremos una pequeña introducción, creemos que bien contextualizada, como un caso particular de la aproximación por mínimos cuadrados.

Veamos un ejemplo con el que podamos apreciar los detalles más importantes de este problema.

Cuando el "Señor del Tiempo" habla del estado de la mar, parte de la información la ha obtenido de unas boyas que están distribuidas a lo largo de toda la costa del país. Estas boyas proporcionan un registro de la altura de ola medida en el punto en el que se encuentran (Figura 4.15) a intervalos de tiempo fijos del orden de segundos. La información obtenida, una lista de números correspondientes a la altura de la ola que pasa por ellas y al instante en que se ha tomado la medición, se comunica vía satélite al centro de proceso de datos correspondiente, donde es analizada. La posición de cada boya varía muy poco, y el registro depende únicamente por tanto del tiempo.

Supongamos que disponemos de una medición de esa señal durante un minuto. Como las olas en alta mar tienen unos periodos del orden de 10 segundos, tomaremos el registro cada $\Delta t = 2$ segundos para tener los puntos suficientes para describir este oleaje. A la variable independiente, el tiempo, la llamaremos t, y llamaremos y a la altura de ola, la variable dependiente. La tabla correspondiente a los diferentes valores será similar a

i	0	1	2	$\cdots$	28	29
t_i	0.0	2.0	4.0	$\cdots$	56.0	58.0
y_i	1.2300	0.0896	-1.1990	$\vdots$	0.3292	1.0031

El que analiza la información no ve una estupenda ola continua como las que vemos cuando estamos en un barco, sino que el registro es discreto porque la boya no puede proporcionar la función altura de modo continuo como números. En realidad podría tomar los datos analógicamente en un tambor de papel y dibujar la curva de alturas sobre ese tambor, pero su transmisión sería imposible y además sería menos útil para su posterior tratamiento y procesado.

El mar es irregular ya que está compuesto de diferentes sistemas de olas que interaccionan entre sí y la gráfica correspondiente a este registro la tenemos en la Figura 4.16. Es importante entender que al tomar un registro de un minuto, para analizarlo mediante funciones periódicas, estamos suponiendo que ese registro es periódico, y que se repite antes y después en el tiempo (ver Figura 4.17). Esto no es cierto en realidad, ya que las olas en el minuto anterior y en el minuto siguiente no mantienen esa periodicidad, pero a nosotros lo que nos interesa es esa muestra en concreto y para analizarla es importante verla así. Por esta razón truncamos el registro en el tiempo 58 segundos, ya que el siguiente sería una repetición del primero.

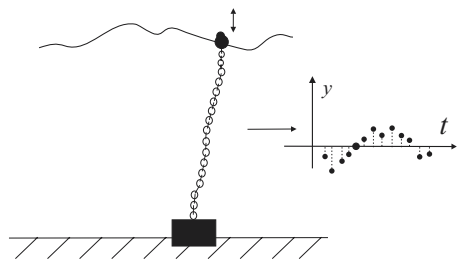


Figura 4.15: Boya para medir la altura de las olas.

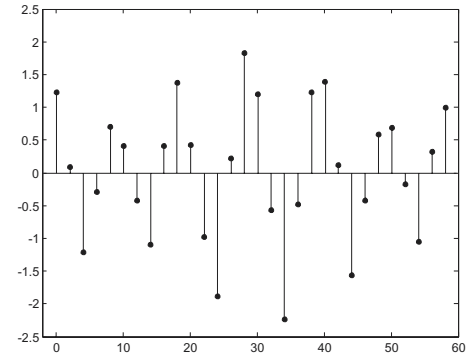


Figura 4.16: Registro de olas de 1 minuto.

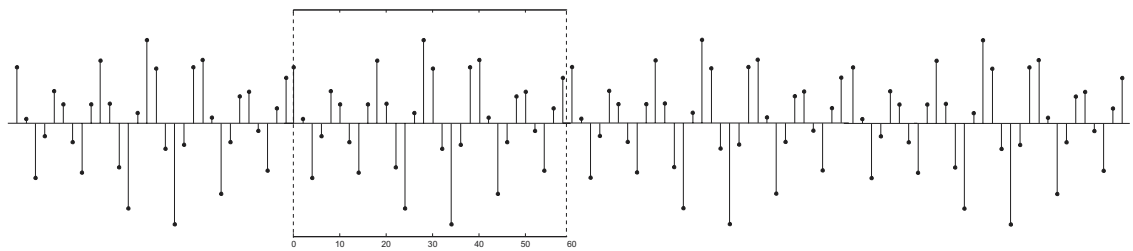


Figura 4.17: Extensión periódica del registro de 1 minuto.

Hemos visto un registro típico correspondiente a un minuto. En la Figura 4.18 presentamos uno equivalente pero correspondiente a un día completo. Visto desde la perspectiva de un día completo, el registro presenta muchas oscilaciones, pero superpuesta a todas ellas está una ola cuyo periodo son 12 horas y que corresponde a las mareas, que es la única que realmente podemos apreciar en esta escala de tiempos. La idea será buscar en los registros cuáles son las diferentes componentes periódicas y qué amplitud tiene cada una de ellas, para lo cual utilizaremos la proyección por mínimos cuadrados.

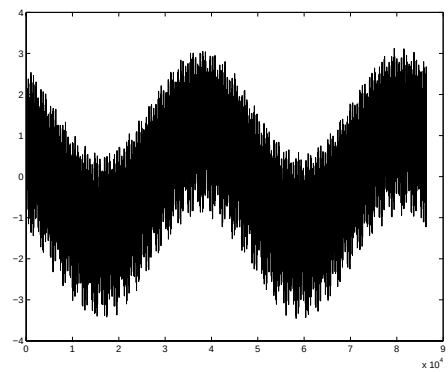


Figura 4.18: Registro correspondiente a un día completo.

4.8.2. Planteamiento del problema

Sea **y** el vector resultado de evaluar con un intervalo constante una determinada función. Este vector tiene como elementos los diferentes valores de la señal en cada instante de tiempo

$$\mathbf{y} = \{\dots, y_{-2}, y_{-1}, y_0, y_1, \dots, y_{N-1}, y_N, y_{N+1}, \dots\}$$

La muestra discreta **y** se dice que tiene periodo N si

$$y_{n+N} = y_n$$

siendo n un número entero cualquiera, positivo o negativo[11]. El periodo N es discreto. Está relacionado con el periodo real T de la señal mediante el paso de tiempo que se utilice para obtener las muestras Δt. Si el periodo real es T se tendrá que $T = N\Delta t$. Como hemos comentado antes, suponemos que la señal es periódica para poder analizarla aunque en la realidad no lo sea. Así en el ejemplo del registro de un minuto, el periodo es $T = 60$ segundos.

Ejercicio 4.8.1 *Calcular el periodo discreto N correspondiente al ejemplo de la Figura 4.16 (sol: N=30).*

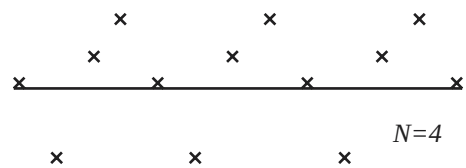


Figura 4.19: Muestra discreta periódica equiespaciada.

La frecuencia en radianes/registro asociada al periodo N recibe el nombre de frecuencia fundamental w_0 y es la más baja de todas las frecuencias con las que trabajaremos.

$$w_0 = \frac{2\pi}{N}$$

En la Figura 4.19 el periodo es $N = 4$, lo que significa que dentro de nuestra señal periódica, tenemos un registro cada w_0 radianes, o sea, cada $\pi/2$.

Vamos a aproximar por mínimos cuadrados el registro **y**. Usaremos como variable independiente el índice en vez del tiempo. Esta decisión no cambia casi nada porque al ser un registro equiespaciado, el índice del punto y el tiempo están relacionados entre sí por el paso de tiempo, $t_n = n\Delta t$. La aproximación la buscaremos dentro de funciones de la familia de exponenciales complejas

$$\phi_k(\tau) = e^{jkw_0\tau} = cos\,(kw_0\tau) + jsen\,(kw_0\tau)$$

con k también entero.

Ejercicio 4.8.2 *Demostrar que ϕ_k es periódica de periodo N.*

La representación discreta de la función ϕ_k es el vector

$$\Phi_k = (\cdots, \phi_k(-N), \cdots, \phi_k(m), \cdots, \phi_k(N), \cdots)$$

Podemos ir al registro de la Figura 4.16, calcular los diferentes vectores Φ_k y presentar su parte real en la Figura 4.20. Para el elemento Φ_k tenemos k periodos completos dentro del rango de los N puntos de un periodo discreto. Es fácil ver que a medida que k crece, la frecuencia de la señal también lo hace.

[11] Siendo coherentes con la notación utilizada anteriormente en el libro, deberíamos usar i para este índice. Lo cambiamos aquí y en las páginas siguientes por n para no confundir con la i compleja, a la cual designamos aquí con j que es lo habitual cuando se habla de la transformada de Fourier discreta.

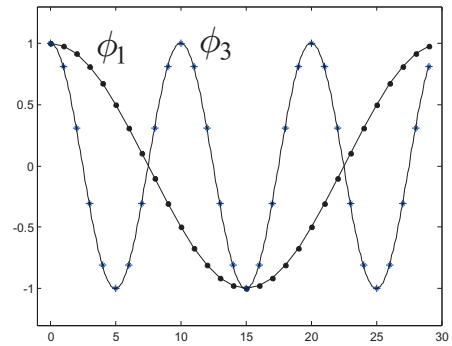


Figura 4.20: Funciones exponenciales de base, parte real.

Ejercicio 4.8.3 *Calcular el elemento* Φ_0 *correspondiente al ejemplo de la Figura 4.16.*

Ejercicio 4.8.4 *Matlab permite utilizar directamente variables complejas. Así, podemos definir por ejemplo* phi=exp(j*2*w0*t). *También disponemos de la orden* real *para extraer la parte real de un número complejo. Sabiendo esto, se pide generar unas líneas Matlab con las que dibujar las curvas de la Figura 4.20.*

Lema 4.8.1 *Los vectores* Φ_k *tienen periodo* N*.*

En efecto, llamando Φ_k^{n+N} la componente $n+N$ del vector Φ_k se tiene

$$\Phi_k^{n+N} = \phi_k(n+N) = e^{jkw_0(n+N)} = e^{jkw_0 n}e^{jkw_0 N} = e^{jkw_0 n}e^{jk2\pi} = e^{jkw_0 n} = \phi_k(n) = \Phi_k^n$$

Como consecuencia, es suficiente definir este vector en un periodo $\Phi_k := (\phi_k(n), \cdots, \phi_k(n+N-1))$. De hecho, para simplificar, seleccionamos el periodo que empieza en 0 y termina en $N-1$ y definimos

$$\Phi_k := (\phi_k(0), \cdots, \phi_k(N-1))$$

Por la misma razón resulta que sólo hay N vectores distintos:

$$\begin{aligned}\Phi_{k+N} &= (\phi_{k+N}(0), \cdots, \phi_{k+N}(N-1)) = \left(e^{j(k+N)w_0(0)}, \cdots, e^{j(k+N)w_0(N-1)}\right) \\ &= \left(1, \cdots, e^{jkw_0(N-1)}e^{jNw_0(N-1)}\right) = \left(1, \cdots, e^{jkw_0(N-1)}e^{j2\pi(N-1)}\right) \\ &= \left(1, \cdots, e^{jkw_0(N-1)}\right) = \left(e^{jkw_0(0)}, \cdots, e^{jkw_0(N-1)}\right) = \Phi_k\end{aligned}$$

De este modo, el subespacio donde se buscará la mejor aproximación será la envolvente de a lo sumo N de estas funciones consecutivas, que podemos indexar, por comodidad, del mismo modo que lo hicimos con el desarrollo de Fourier continuo:

$$U = L\left(\phi_k\right)_{k=0,\pm1,\pm2,\cdots}$$

Un caso interesante es el de interpolación en el que el número de funciones coincide con el de puntos que definen cada periodo discreto y por tanto aproximamos con N funciones ϕ_k.

En este caso las funciones se suelen indexar del siguiente modo:

- Si N es par, k varía en $I := \left\{-\frac{N}{2}, \cdots, 0, \cdots, \frac{N}{2}-1\right\}$
- Si N es impar, k varía en $I := \left\{-\left[\frac{N}{2}\right], \cdots, 0, \cdots, \left[\frac{N}{2}\right]\right\}$, donde [ ] denota la parte entera de un número.

4.8.3. Ortogonalidad del sistema

Proyectemos ahora $\mathbf{y}$ sobre la representación discreta de los elementos de la base de U, o sea, sobre el subespacio engendrado por los vectores Φ_k. La matriz del sistema lineal es la que ya calculamos de modo general para la aproximación por mínimos cuadrados (ecuación (4.21)). Aplicada a este caso

$$\begin{aligned} a_{lk} &= \langle \Phi_l, \Phi_k \rangle = \Phi_l \cdot \overline{\Phi_k} = (\phi_l(0), \cdots, \phi_l(N-1)) \left(\overline{\phi_k}(0), \cdots, \overline{\phi_k}(N-1)\right) = \\ &= \left(e^{jlw_0(0)}, \cdots, e^{jlw_0(N-1)}\right) \left(e^{-jkw_0(0)}, \cdots, e^{-jkw_0(N-1)}\right) = \\ &= e^{j(l-k)w_0(0)} + \cdots + e^{j(l-k)w_0(N-1)} = \sum_{n=0}^{N-1} e^{j(l-k)w_0 n} = \begin{cases} \{l=k\} & \sum_{n=0}^{N-1} 1 = N \\ \{l \neq k\} & \sum_{n=0}^{N-1} \left(e^{j(l-k)w_0}\right)^n \end{cases} \end{aligned}$$

Veamos este último término:

$$\sum_{n=0}^{N-1} \left(e^{j(l-k)w_0}\right)^n = \sum_{n=0}^{N-1} r^n = \frac{1-r^N}{1-r} = \frac{1-e^{j(l-k)w_0 N}}{1-r} = \frac{1-e^{j(l-k)2\pi}}{1-r} = 0$$

luego

$$a_{lk} = \begin{cases} N & \text{si} \quad l=k \\ 0 & \text{si} \quad l \neq k \end{cases}$$

Los vectores Φ_k son ortogonales y la matriz del sistema lineal de la ecuación (4.21) es diagonal.

Si la función solución es $\hat{f} = \sum_{k \in I} c_k \phi_k$, su forma vectorial es la proyección ortogonal de $\mathbf{y}$ sobre $L\{\Phi_k\}$

$$\hat{\mathbf{f}} = \sum_{k \in I} c_k \Phi_k$$

y ya que el sistema Φ_k es ortogonal

$$c_k = \frac{\langle \mathbf{y}, \overline{\Phi}_k \rangle}{N} = \frac{1}{N} \sum_{n=0}^{N-1} y_n e^{-jkw_0 n} \tag{4.22}$$

A este conjunto de coordenadas c_k, cuando barremos todas las frecuencias posibles, se le llama la Transformada de Fourier Discreta (*Discrete Fourier Transform*, DFT) de la función f, y a su módulo (no olvidar que son valores en general complejos) se le llama el espectro. Cuando no cubramos todo el rango de frecuencias, nos referiremos a $\hat{f}$ como la aproximación por mínimos cuadrados de una función periódica definida de modo discreto.

Ejercicio 4.8.5 *Condición para que la mejor aproximación $\hat{f}$ sea real (pista: ver sección 4.6.4).*

Ejercicio 4.8.6 *Sea $\mathbf{y}$ un vector de números reales. ¿Es $\hat{f}$ también real? (pista: ver sección 4.6.5)*

Ejemplo 4.8.1 *Se supone la siguiente señal periódica discreta (ver Figura 4.21):*

i	0	1	2	3	4	5	6	7
t_i	0	2.3	4.6	6.9	9.2	11.5	13.8	16.1
y_i	1	0	0	1	0	0	0	1

La tabla da los datos de un periodo completo. Dado que la función se repite cada 8 puntos, tenemos que $N = 8$, y la frecuencia fundamental, tomando como variable independiente el índice en vez del tiempo, será

$$w_0 = \frac{2\pi}{8} = \frac{\pi}{4}$$

Vamos a ir construyendo la mejor aproximación incrementando las frecuencias. Así, el primer armónico[12]*, lo obtendremos considerando que el subespacio donde aproximamos es $U = L\{\phi_0, \phi_{-1}, \phi_1\}$ con*

$$\phi_0(\tau) = e^{j0w_0\tau} = 1, \quad \phi_{-1}(\tau) = e^{-jw_0\tau} = \cos(w_0\tau) - j\sin(w_0\tau), \quad \phi_1(\tau) = e^{jw_0\tau} = \cos(w_0\tau) + j\sin(w_0\tau)$$

[12]Un armónico de la señal es la componente que esa señal tiene en una determinada frecuencia. Como ϕ_{-k} y ϕ_k tienen la misma frecuencia, necesitamos incorporar ambas simultáneamente para tener el armónico k completo.

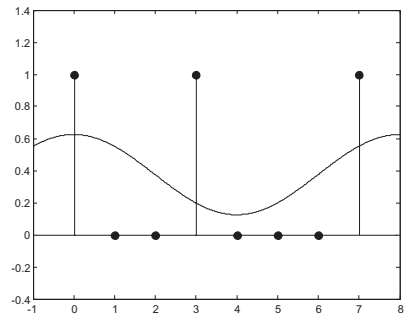


Figura 4.21: La función $\hat{f}$ junto con el primer armónico correspondiente al ejemplo 4.8.1

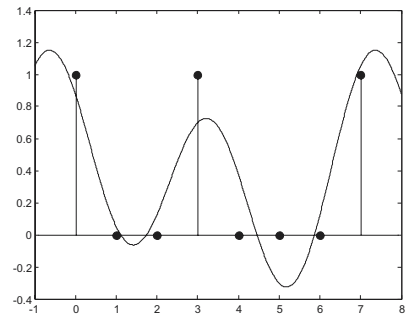


Figura 4.22: La función $\hat{f}$ junto con los dos primeros armónicos correspondiente al ejemplo 4.8.1

Calculemos los coeficientes c_k. Empecemos con c_0.

$$c_0 = \frac{\langle \mathbf{y}, \overline{\Phi}_0 \rangle}{8}$$

con $\mathbf{y} = (1,0,0,1,0,0,0,1)$ *y* $(\phi_0(0), \cdots, \phi_0(7)) = (1,1,1,1,1,1,1,1)$.

Como las componentes de Φ_0 son números reales, es fácil ver que c_0 es simplemente la media aritmética del vector $\mathbf{y}$.

$$c_0 = 3/8 = 0.375$$

Respecto a las siguientes componentes

$$\Phi_{-1} \quad = \quad (\phi_{-1}(0), \cdots, \phi_{-1}(7)) = (\cos(w_0 0) - j\sin(w_0 0), \cdots, \cos(w_0 7) - j\sin(w_0 7))$$

Valores que ya no son tan sencillos de calcular manualmente, así que nos apoyaremos en Matlab, introduciendo las siguientes líneas:

```
>> w0=pi/4;
>> n=0:7;
>> Phi_1=exp(-j*w0*n)
Phi_1 =
Columns 1 through 5
1.0000           0.7071-0.7071i  0.0000-1.0000i  -0.7071-0.7071i  -1.0000-0.0000i
Columns 6 through 8
-0.7071+0.7071i -0.0000+1.0000i  0.7071+0.7071i
```

Ahora introducimos el vector $\mathbf{y}$ *y calculamos c_{-1} teniendo en cuenta que en Matlab tenemos que definir el producto escalar como el producto de un vector fila por un vector columna, y por eso trasponemos el segundo. Al trasponer un vector de números complejos utilizando (') también estamos escribiendo su conjugado.*

```
>> y=[1 0 0 1 0 0 0 1];
>> c_1=y*(Phi_1)'/8
c_1 = 0.1250 - 0.0000i
```

y $c_{-1} = 0.1250$.

Ya que la señal es real (ver ejercicios 4.8.6 y 4.8.6), no hace falta calcular c_1, pues será directamente el conjugado de c_{-1}. $c_1 = 0.1250$. *La función aproximación con estas tres componentes será*

$$\hat{f}(\tau) = c_0\phi_0(\tau) + c_{-1}\phi_{-1}(\tau) + c_1\phi_1(\tau)$$

En esos mismos ejercicios vimos que en caso de que consideremos simultáneamente los términos ϕ_{-k} y ϕ_k siendo los datos de partida reales, entonces, $\hat{f}$ también será real. En este caso:

$$\hat{f}(\tau) = 0.375 + 0.25\cos\left(\frac{\pi}{4}\tau\right)$$

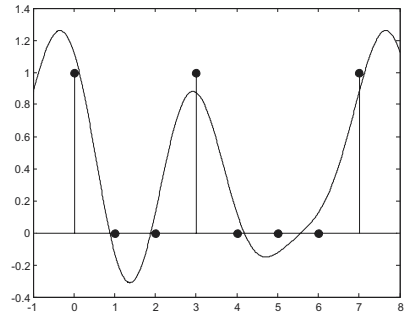


Figura 4.23: La función $\hat{f}$ junto con los tres primeros armónicos correspondiente al ejemplo 4.8.1.

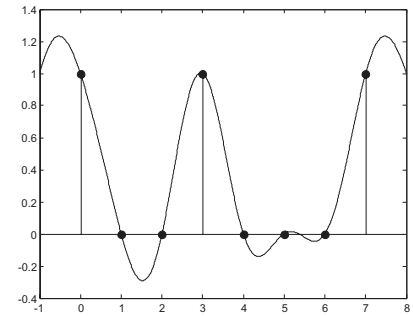


Figura 4.24: La función $\hat{f}$ con todos los armónicos correspondiente al ejemplo 4.8.1.

Si queremos que la variable independiente sea el tiempo, basta con hacer el cambio de variable correspondiente, sabiendo que $t = 2.3\tau$. *Por tanto*

$$\hat{f}(t) = 0.375 + 0.25\cos\left(\frac{2.3\pi}{4}t\right)$$

En la Figura 4.21 se representan la nube de puntos original y la gráfica correspondiente a la función $\hat{f}$. *Podemos ampliar el subespacio* U *incluyendo una frecuencia más en el sistema generador*

$$U = L\{\phi_0, \phi_{-1}, \phi_1, \phi_{-2}, \phi_2\}$$

pero nos apoyamos directamente en Matlab para realizar todas las operaciones:

```
>> Phi_2=exp(-j*2*w0*n);
>> c_2=y*(Phi_2)'/8;
```

y obtenemos $c_{-2} = 0.1250 - 0.2500j$ *de donde* $c_2 = \overline{c_{-2}} = 0.1250 + 0.2500j$. *Al incorporar estos dos términos extras a* $\hat{f}$, *tendremos*

$$\hat{f}(\tau) = 0.375 + 0.25\cos\left(\frac{\pi}{4}\tau\right) + 0.25\cos\left(\frac{2\pi}{4}\tau\right) - 0.5\sin\left(\frac{2\pi}{4}\tau\right)$$

que presentamos en la Figura 4.22. Ampliemos de nuevo U *para tener una frecuencia más:*

$$U = L\{\phi_0, \phi_{-1}, \phi_1, \phi_{-2}, \phi_2, \phi_{-3}, \phi_3\}$$

```
>> Phi_3=exp(-j*3*w0*n);
>> c_3=y*(Phi_3)'/8;
```

y $c_{-3} = 0.1250 \;\Rightarrow\; c_3 = 0.1250$. *Si llevamos esta frecuencia a* $\hat{f}$:

$$\hat{f}(\tau) = 0.375 + 0.25\cos\left(\frac{\pi}{4}\tau\right) + 0.25\cos\left(\frac{2\pi}{4}\tau\right) - 0.5\sin\left(\frac{2\pi}{4}\tau\right) + 0.25\cos\left(\frac{3\pi}{4}\tau\right)$$

La cual podemos ver en la Figura 4.23. A medida que vamos incorporando más armónicos, $\hat{f}$ *aproxima mejor la nube de puntos original. La dimensión de* U *es 7, así que sólo podemos añadir un elemento más. Si seguimos el criterio expresado en la teoría, como* $N = 8$ *es par, el mayor conjunto de índices en el que vamos a tener las funciones* ϕ_k *será:*

$$I := \left\{-\frac{8}{2}, \cdots, 0, \cdots, \frac{8}{2} - 1\right\} = \{-4, -3, -2, -1, 0, 1, 2, 3\}$$

Nos queda por añadir a U *nada más que el término* ϕ_{-4}.

$$U = L\{\phi_0, \phi_{-1}, \phi_1, \phi_{-2}, \phi_2, \phi_{-3}, \phi_3, \phi_{-4}\}$$

```
>> Phi_4=exp(-j*4*w0*n);
>> c_4=y*(Phi_4)'/8;
```

Con lo que $c_{-4} = -0.1250$. Lo que sucede al añadir este nuevo término a $\hat{f}$ es que ésta deja de ser real para tener también parte imaginaria.

$$\begin{aligned}\hat{f}(\tau) &= 0.375 + 0.25\cos\left(\frac{\pi}{4}\tau\right) + 0.25\cos\left(\frac{2\pi}{4}\tau\right) - 0.5\sin\left(\frac{2\pi}{4}\tau\right) + 0.25\cos\left(\frac{3\pi}{4}\tau\right) - \\ &\quad -0.125\left(\cos\left(\frac{4\pi}{4}\tau\right) - j\sin\left(\frac{4\pi}{4}\tau\right)\right)\end{aligned}$$

En la gráfica 4.24 vemos cómo ahora la parte real de $\hat{f}$ pasa por todos los puntos de la nube al haber convertido el problema en uno de interpolación. Si concentramos los valores de los coeficientes en una tabla, en la cual incluimos también su módulo tendremos el espectro, en el que se aprecia que el segundo armónico es el más importante, aparte del valor medio, que está en el centro del espectro(ver Figura 4.25).

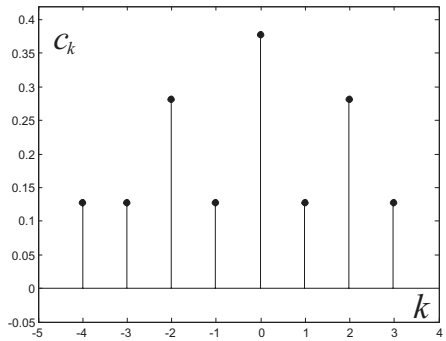


Figura 4.25: Espectro correspondiente al ejemplo 4.8.1.

k	c_k	$\lvert c_k\rvert$
-4	-0.125	0.125
-3	0.125	0.125
-2	0.1250 - 0.2500j	0.2795
-1	0.125	0.125
0	0.375	0.375
1	0.125	0.125
2	0.1250 + 0.2500j	0.2795
3	0.125	0.125

Cuadro 4.1: Componentes de $\hat{f}$.

4.8.4. DFT y Matlab

Hasta que se inventaron los ordenadores, era a efectos prácticos imposible realizar la DFT (Transformada de Fourier Discreta) de una señal. Aun con éstos, el coste de cálculo de los coeficientes es grande cuando el número de puntos utilizados N también lo es. Existe un algoritmo muy eficiente, la transformada rápida de Fourier, FFT (Fast Fourier Transform), que reduce el coste de estos cálculos, aunque su estudio escapa al contenido de este libro (ver, por ejemplo, [22]).

Matlab dispone de herramientas muy potentes para el tratamiento de señales mediante el tipo de técnicas que aquí estamos viendo. Veremos nada más la orden estándar. Dado un vector cualquiera, si se hace *fft* de ese vector, esta orden devuelve las componentes de la DFT, a falta del factor correspondiente al número de puntos, pero lo hace en un orden que precisa interpretación. Si nos referimos al ejemplo 4.8.1, y ejecutamos la siguiente línea Matlab, obtenemos los valores de la columna izquierda

```
>> C=fft(y)'/8
C = 0.3750
   0.1250
   0.1250 - 0.2500i
   0.1250
  -0.1250
   0.1250
   0.1250 + 0.2500i
   0.1250
```

c_0
c_{-1}
c_{-2}
c_{-3}
c_{-4}
c_3
c_2
c_1

que son la solución de nuestro problema, pero el orden en que están dadas es el que aparece en la columna de la derecha.

Podemos utilizar esta misma orden para analizar el registro de olas de un minuto de duración (Figura 4.16), y obtener su espectro (Figura 4.26).

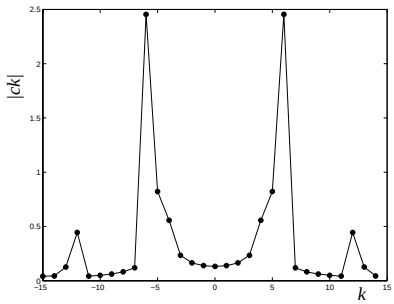


Figura 4.26: Espectro correspondiente al registro de un minuto, Figura 4.16.

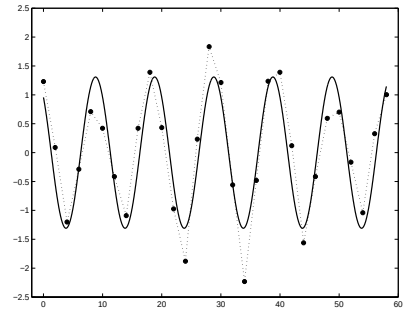


Figura 4.27: Registro de un minuto, Figura 4.16, reconstruido con el armónico más importante y comparado con el registro original.

Se observa que el sexto armónico tiene el coeficiente de módulo más grande y, por tanto, es el armónico de mayor amplitud. Su frecuencia en radianes/registro es $6w_0$ y su periodo correspondiente es:

$$T = \frac{2\pi}{6w_0} = \frac{2\pi}{6\frac{2\pi}{N}} = \frac{N}{6} = 5$$

Como teníamos un registro cada 2 segundos, eso significa que el periodo donde están las olas con más amplitud y por tanto con más energía es el correspondiente a $5 \cdot 2 = 10$ segundos. Si reconstruimos la señal utilizando sólo este armónico, obtenemos el resultado que mostramos en la Figura 4.27.

PROBLEMAS

PROBLEMA 4.1 *Desarrollo en serie de Fourier.*

Se tiene la siguiente función periódica:

$$f(t) = \begin{cases} 0 & -1 \leq t < 0 \\ 1 & 0 \leq t \leq 1 \\ 0 & 1 < t \leq 2 \end{cases}$$

uno de cuyos periodos y su extensión periódica aparecen en la Figura 4.28. Se pide:

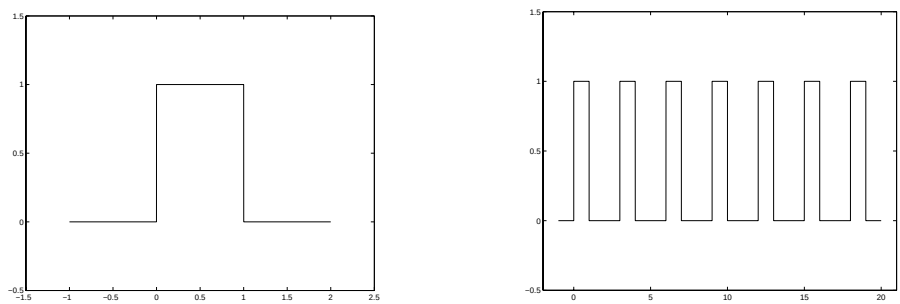


Figura 4.28: Función del problema 4.1 y extensión periódica de la misma.

1. Obtener de modo general los coeficientes de su desarrollo en serie de Fourier.

2. Escribir un código Matlab para obtener la reconstrucción de la señal utilizando hasta $kmax$ de esos términos.

3. Calcular los coeficientes del desarrollo en senos y cosenos de f particularizando para $kmax = 1$.

Solución:

1. Utilizando 4.8 obtenemos los coeficientes del desarrollo en serie de Fourier, con $T = 3$ y $w_0 = 2\pi/T$. El primer coeficiente $c_0 = 1/3$. Para $k > 0$:

$$c_k = \frac{1}{3}\int_{-1}^{2} f(u)e^{-jkw_0u}du = \frac{1}{3}\int_{0}^{1} e^{-jkw_0u}du = \frac{j}{3kw_0}\, e^{-jkw_0u}\Big|_0^1 = \frac{j}{3kw_0}\left[e^{-jkw_0} - 1\right]$$

2. Con el código Matlab *fourier_continuo.m*, aproximamos la función original $f(x)$ utilizando los términos del desarrollo en serie desde $k = -kmax, \ldots, kmax$. Presentamos en la Figura 4.29 los resultados para $kmax = 1$ y $kmax = 10$ respectivamente. En la Figura 4.30 presentamos también varios periodos de la función aproximación, donde podemos observar lo bien que estamos modelando la función original a pesar de su discontinuidad.

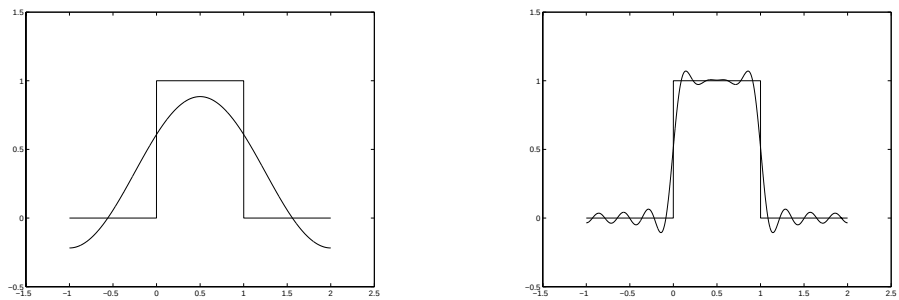


Figura 4.29: Problema 4.1; expansiones para $kmax = 1$ y $kmax = 10$.

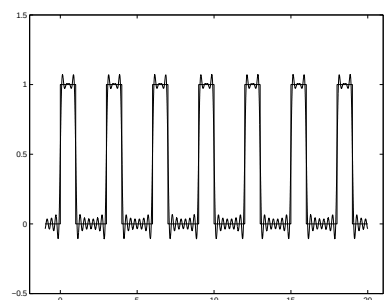


Figura 4.30: Varios periodos de la aproximación en el problema 4.1.

3. Calculemos las partes real e imaginaria de los coeficientes c_k utilizando lo estudiado en la sección 4.6.5.

$$c_k = \frac{j}{3kw_0}\left[e^{-jkw_0} - 1\right] = \frac{j}{3kw_0}\left(\cos(kw_0) - j\sin(kw_0) - 1\right) = \frac{\sin(kw_0)}{3kw_0} + j\frac{\cos(kw_0) - 1}{3kw_0}$$

$$A_k = \frac{\sin(kw_0)}{3kw_0}, \quad B_k = \frac{\cos(kw_0) - 1}{3kw_0}$$

Si nos quedamos con el primer término:

$$\hat{f}(t) = c_0 + 2A_1\cos(w_0 t) - 2B_1\sin(w_0 t) = 0.3333 + 0.2757\cos\left(\frac{2\pi}{3}t\right) + 0.4775\sin\left(\frac{2\pi}{3}t\right)$$

Se deja como ejercicio representar con Matlab esta curva comprobando que se corresponde con la de la Figura 4.29.

PROBLEMA 4.2 *Polinomios ortogonales de Chebychev.*

Sea E el espacio vectorial normado $C([-1,1])_{\|\ \|}$, con [13]

$$\|f\| = \left(\int_{-1}^{1} \frac{f(x)^2}{\sqrt{1-x^2}}dx\right)^{1/2}$$

Sea $v(x) = \sqrt{1-x^2}$. Se busca definir, dentro del conjunto de los polinomios de grado 2, el conjunto de las mejores aproximaciones del elemento v, estableciendo previamente si hay existencia y unicidad.

Solución:
El conjunto de los polinomios de grado 2 es un subespacio vectorial de E y por tanto la m.a. existe. Llamemos U a este conjunto. La norma que estamos utilizando en E deriva de un producto escalar para el que los polinomios de Chebychev son un sistema ortogonal.

$$\langle f, g\rangle = \int_{-1}^{1} \frac{f(x)g(x)}{\sqrt{1-x^2}}dx$$

Las normas que derivan de un producto escalar son estrictas y, por tanto, la m.a. que es la proyección ortogonal de v sobre U, es única. Elegimos como base de U la formada por los polinomios de Chebychev hasta el grado 2

$$T_0(x) = 1, \quad T_1(x) = x, \quad T_2(x) = 2x^2 - 1$$

Escribamos la solución a nuestro problema, P, en esta base

$$P = \sum_{j=0}^{2} c_j T_j$$

Como la base es ortogonal, tendremos que el sistema lineal 4.4 (pg. 184) será diagonal y tendremos para $k = 0, 1, 2$,

$$c_k = \frac{< v, T_k >}{< T_k, T_k >} = \frac{\int_{-1}^{1} \frac{v(x)T_k(x)}{\sqrt{1-x^2}}dx}{< T_k, T_k >} = \frac{\int_{-1}^{1} \frac{\sqrt{1-x^2}T_k(x)}{\sqrt{1-x^2}}dx}{< T_k, T_k >} = \frac{\int_{-1}^{1} T_k(x)dx}{< T_k, T_k >}$$

de donde

$$c_0 = \frac{\int_{-1}^{1} dx}{\pi} = \frac{2}{\pi}, \quad c_1 = \frac{\int_{-1}^{1} xdx}{\pi/2} = 0, \quad c_2 = \frac{\int_{-1}^{1} (2x^2-1)dx}{\pi/2} = -\frac{4}{3\pi}$$

luego

$$P(x) = \frac{2}{\pi}T_0(x) - \frac{4}{3\pi}T_2(x) = \frac{10}{3\pi} - \frac{8}{3\pi}x^2$$

Presentamos en la Figura 4.31 las dos curvas superpuestas y se aprecia que bien se ajusta la aproximación, sobre todo en los extremos.

[13] Esta norma da más peso a los valores que toma la función cerca de los extremos.

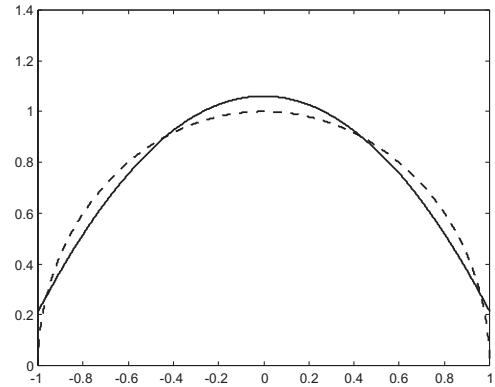


Figura 4.31: Aproximación en la base de Chebychev, v en rayado, P en continuo.

PROBLEMA 4.3 *Polinomio óptimo.*

Sean $h(x) = \sqrt{x}$, $f(x) = h'(x)$. Buscamos un número real θ, con $0 < \theta < 1$ y un polinomio cúbico de coeficientes reales $p(x) = ax^3 + bx^2 + cx + d$ tales que el polinomio derivada interpole a la función f en $\left\{\theta, \frac{\theta+1}{2}, 1\right\}$ y que

$$|f(x) - p'(x)| \leq 10^{-3}, \quad \theta < x < 1$$

1. Encontrar una relación que permita encontrar un valor de θ lo más pequeño posible.
2. Se supone que la relación obtenida es:

   $$(1-\theta)^3 = k\theta^{3.5}, \quad k = 0.1024$$

   Transformar este problema no lineal del modo más sencillo posible en uno de aproximaciones sucesivas, cuyo punto fijo sea el valor buscado θ. Iterar sobre él hasta encontrar la solución. Razonar de modo aproximado sobre la convergencia y sobre la velocidad de convergencia.

   Se elegirá como estimador inicial 0.9, y se darán los pasos necesarios en el esquema para estabilizar el cuarto decimal.
3. Calcular p' y representar gráficamente $|p' - f|$.
4. Con el valor de θ obtenido en el apartado anterior, calcular p de tal modo que $\| p - h \|_2$ sea mínima, con:

   $$\| g \|_2 = \sqrt{\int_\theta^1 g(x)^2 dx}$$

Solución:

1. Sea q el polinomio derivada de p:

   $$q(x) = p'(x) = Ax^2 + Bx + C$$

   Sea f la derivada de h:

   $$f(x) = \frac{dh}{dx} = \frac{1}{2\sqrt{x}}$$

   Interpolamos f en la partición equiespaciada $\left\{\theta, \frac{\theta+1}{2}, 1\right\}$ para construir q. Mayoramos el error usando la estimación estudiada en la teoría (pág. 126)

   $$|f(x) - q(x)| \leq \left| \frac{H(x)}{3!} f'''(\xi(x)) \right|, \quad \xi, x \in [\theta, 1]$$

$$H(x) = (x-\theta)\left(x - \frac{\theta+1}{2}\right)(x-1)$$

Cuando se mide el error en un punto de la subdivisión entre el mínimo y el máximo, se puede aplicar la desigualdad 3.10 (pág. 127):

$$\|f-q\|_\infty \leq \frac{\|f^{(n+1)}\|_\infty}{4(n+1)} h^{n+1} \tag{4.23}$$

con:

$$\|g\|_\infty = \max_{x\in[\theta,1]} |g(x)|, \quad n=2, \quad h = \frac{1-\theta}{2}$$

Como:

$$f'''(\xi) = -\frac{15}{16}\xi^{-\frac{7}{2}}, \quad \|f'''\|_\infty = \frac{15}{16\,\theta^{3.5}}$$

Entrando con estos valores en 4.23 obtenemos la siguiente estimación

$$\|f-q\|_\infty \leq \frac{(1-\theta)^3}{3\cdot 2^5}\frac{15}{16\,\theta^{3.5}} = \frac{5}{2^9}\frac{(1-\theta)^3}{\theta^{3.5}}$$

La condición sobre la que se trabaja para obtener el valor de θ será:

$$\frac{5}{2^9}\frac{(1-\theta)^3}{\theta^{3.5}} \leq 10^{-3}$$

O lo que es lo mismo:

$$(1-\theta)^3 \leq k\theta^{3.5}, \quad k = 0.1024$$

El valor límite buscado será en el que se produzca la igualdad. Es fácil ver gráficamente cómo se comportan los dos miembros de la desigualdad, y que sólo tendremos una raíz en el intervalo $[0,1]$, Figura 4.32. Estas son las líneas Matlab que usamos para representar esta gráfica.

```
k=0.1024;
t=0.50:0.01:1;
f1=(1-t).^3;
f2=k*t.^3.5;
plot(t,f1,t,f2);
```

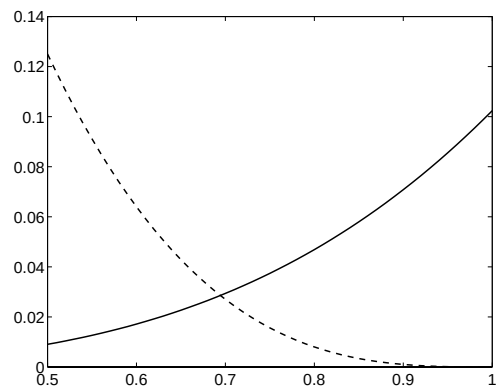


Figura 4.32: Curvas cuya intersección define θ.

2. Transformamos el problema en uno de punto fijo, $\theta = T(\theta)$, con:

$$\theta = \theta + (1-\theta)^3 - k\theta^{3.5}$$

Elegimos esta formulación en vez de

$$\theta = \theta - (1-\theta)^3 + k\theta^{3.5}$$

porque el valor de la derivada en la zona de la raíz es en la primera (~ 0.6) menor que uno, y en la segunda no (~ 1.4). En suma:

$$T(\theta) = \theta + (1-\theta)^3 - k\theta^{3.5}$$

El hecho de que en la zona de la raíz, la derivada sea en valor absoluto menor que uno, sugiere la convergencia del método. Veámoslo. Tomemos como estimador inicial 0.9, que en la figura está relativamente cerca de la raíz.

Repitiendo la última de las instrucciones Matlab siguientes

```
>>t=0.9;
>>k=0.1024;
>>t=t+(1-t)^3-k*t^3.5
>>t=t+(1-t)^3-k*t^3.5
```

se obtienen los siguientes valores

i	0	1	2	3	5	10	18	19
θ_i	0.9000	0.8302	0.7817	0.7489	0.7141	0.6951	0.6943	0.6943

Como vemos, la convergencia no es mala. Si estudiamos la derivada de T en la raíz, su valor es:

$$T'(0.6943) = 1 - 3(1-0.6943)^2 - k3.5 \cdot 0.6943^{2.5} = 0.5757$$

que no es ni muy cercano a la unidad ni al cero, lo que justifica una convergencia razonable, aunque no muy rápida.

3. Calculemos el polinomio q que interpola a f en los puntos $\{\theta, (\theta+1)/2, 1\}$. Lo hacemos en la base de los monomios, porque así se facilita su posterior integración.

   Impongamos que $q(0.6943) = f(0.6943)$

$$q(0.6943) = f(0.6943) = \frac{1}{2\sqrt{0.6943}} = 0.6001 = A0.6943^2 + B0.6943 + C$$

   Repitiendo con $(\theta+1)/2$ y con 1 obtenemos el sistema lineal

$$\begin{pmatrix} 0.6943^2 & 0.6943 & 1.0000 \\ 0.8472^2 & 0.8472 & 1.0000 \\ 1.0000^2 & 1.0000 & 1.0000 \end{pmatrix} \begin{pmatrix} A \\ B \\ C \end{pmatrix} = \begin{pmatrix} 0.6001 \\ 0.5432 \\ 0.5000 \end{pmatrix} \quad \Rightarrow \quad \begin{pmatrix} A \\ B \\ C \end{pmatrix} = \begin{pmatrix} 0.2908 \\ -0.8200 \\ 1.0292 \end{pmatrix}$$

   y

$$q(x) = Ax^2 + Bx + C = 0.2908x^2 - 0.8200x + 1.0292$$

   Podemos ver la diferencia en valor absoluto entre f y q en la Figura 4.33. Observamos que esa diferencia se hace cero en los nodos de la interpolación, como debía suceder.

   Si usamos las siguientes líneas Matlab, la primera gráfica corresponde a las funciones p y q y se observa lo bien que ambas se ajustan; la segunda gráfica es la que representamos en la Figura 4.33.

```
t=0.6943:0.001:1;
q=1.0292-0.8200*t+0.2908*t.^2;
f=1/2./sqrt(t);
plot(t,q,t,f);
shg;
pause;
plot(t,abs(q-f));
shg;
```

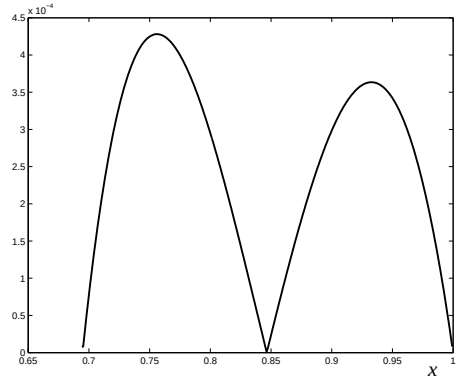


Figura 4.33: Problema 4.3. $|f(x) - q(x)|$.

4. Integrando el polinomio q obtenemos p

$$p(x) = \int q(x)dx = 0.0969x^3 - 0.4100x^2 + 1.0292x + d = r(x) + d$$

Para determinar d usaremos la condición de que la distancia de $p(x)$ a la función original $\sqrt{x}$ en la norma 2

$$\| r(x) + d - \sqrt{x} \|_2$$

sea mínima, lo que es equivalente a que

$$\| d - \left(\sqrt{x} - r(x)\right) \|_2$$

sea mínimo.

Ya que la norma 2 deriva de un producto escalar, se puede interpretar lo anterior como la búsqueda de la mejor aproximación en el subespacio vectorial de las funciones de valor constante. Dicha mejor aproximación será la proyección ortogonal sobre ese subespacio, cuya base canónica es la función constante igual a 1.

$$\langle d - \left(\sqrt{x} - r(x)\right), 1\rangle = d\langle 1, 1\rangle - \langle \sqrt{x} - r(x), 1\rangle = 0 \quad \Rightarrow$$

$$d = \frac{\int_\theta^1 \left(\sqrt{x} - r(x)\right) dx}{\int_\theta^1 dx} = \frac{0.0868}{1 - 0.6943} = 0.2839$$

Podemos realizar estas operaciones con las siguientes líneas Matlab:

```
syms x;
r=0.0969*x^3-0.4100*x^2+1.0292*x;
I=int(sqrt(x)-r);
x=0.6943;
I1=eval(I);
x=1;
I2=eval(I);
d=(I2-I1)/(1-0.6943);
```

Presentamos las funciones p y h en la Figura 4.34, donde se observa que bien se ajusta el polinomio a la raíz en la zona de interés, entre θ y 1. La expresión analítica final de p es:

$$p(x) = 0.0846x^3 - 0.3763x^2 + 0.9987x + 0.2839$$

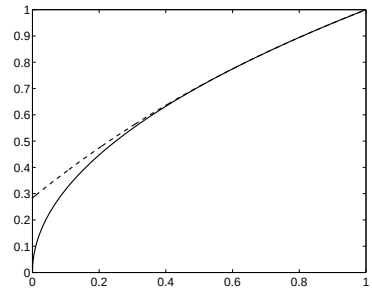


Figura 4.34: Problema 4.3. Funciones h y p.

PROBLEMA 4.4 *Aproximación en un espacio en el que la norma se deduce de un producto escalar.*

Se considera el espacio vectorial $C([0,2],\mathbb{R})$ provisto de la norma $\|\|_2$

$$\| f \|_2 = \left(\int_0^2 |f(x)|^2 dx \right)^{1/2}$$

Definir el conjunto de las mejores aproximaciones a la función x^2 dentro de los splines de grado 1 que tienen como soporte la partición $\Omega = \{0,1,2\}$ del intervalo compacto $[0,2]$ para la norma $\|\|_2$. Se pide definir con total exactitud, si fuese posible, cada uno de los tramos de los elementos de dicho conjunto en la base canónica de $P_1(\mathbb{R})$. En el caso de que no fuese posible, se exige una precisión de al menos cuatro cifras decimales.

Solución:
La norma $\|\|_2$ deriva de un producto escalar, lo que implica que $C([0,2],\mathbb{R})$ con esta norma tiene estructura de espacio prehilbertiano:

$$\langle f, g \rangle = \int_0^2 f(x)g(x)dx$$

Se busca la mejor aproximación de $f(x) = x^2$ dentro de $S_1(\Omega)$, subespacio vectorial de dimensión finita de las $C([0,2],\mathbb{R})$, por tanto la mejor aproximación existe y además, es única (ver Hämmerlin y Hoffmann [15] Capítulo 4). Dicha mejor aproximación es la proyección ortogonal de f sobre $S_1(\Omega)$, ver Figura 4.35.

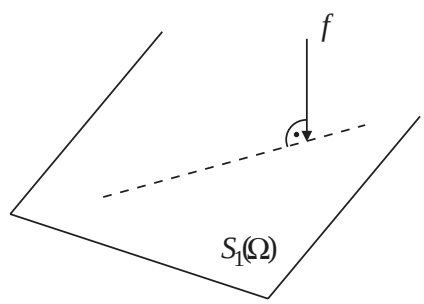


Figura 4.35: Proyección ortogonal.

Llamémosla s. Para hallar dicha proyección necesitamos definir una base de $S_1(\Omega)$. Podemos tomar como base de dicho espacio la formada por B-splines (Figuras 4.36 a 4.39).

$$B_{-1}^1 = \begin{cases} 1-x & 0 \le x \le 1 \\ 0 & 1 \le x \le 2 \end{cases}, \quad B_0^1 = \begin{cases} x & 0 \le x \le 1 \\ 2-x & 1 \le x \le 2 \end{cases}, \quad B_1^1 = \begin{cases} 0 & 0 \le x \le 1 \\ x-1 & 1 \le x \le 2 \end{cases}$$

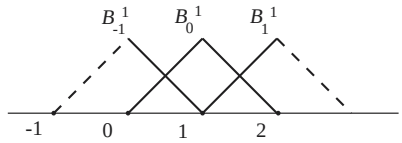


Figura 4.36: Base de B-splines.

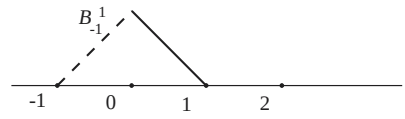


Figura 4.37: Soporte de B^1_{-1}.

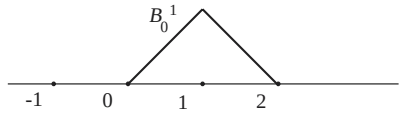


Figura 4.38: Soporte de B^1_0.

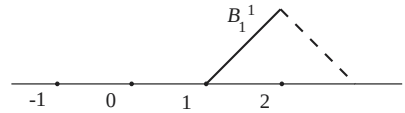


Figura 4.39: Soporte de B^1_1.

Poniendo

$$s(x) = \sum_{i=-1}^{i=1} a_i B_i^1(x)$$

tenemos

$$\langle x^2 - s, B_j^1\rangle = 0, -1 \leq j \leq 1 \quad \Rightarrow \quad \sum_{i=-1}^{i=1} a_i \langle B_i^1(x), B_j^1(x)\rangle = \langle x^2, B_j^1\rangle, -1 \leq j \leq 1$$

El sistema de ecuaciones asociado es

$$\begin{pmatrix} \langle B_{-1}^1, B_{-1}^1\rangle & \langle B_0^1, B_{-1}^1\rangle & \langle B_1^1, B_{-1}^1\rangle \\ \langle B_{-1}^1, B_0^1\rangle & \langle B_0^1, B_0^1\rangle & \langle B_1^1, B_0^1\rangle \\ \langle B_{-1}^1, B_1^1\rangle & \langle B_0^1, B_1^1\rangle & \langle B_1^1, B_1^1\rangle \end{pmatrix} \begin{pmatrix} a_{-1} \\ a_0 \\ a_1 \end{pmatrix} = \begin{pmatrix} \langle x^2, B_{-1}^1\rangle \\ \langle x^2, B_0^1\rangle \\ \langle x^2, B_1^1\rangle \end{pmatrix}$$

Calculemos los elementos de este sistema lineal, teniendo en cuenta que es simétrico

$$\langle x^2, B_{-1}^1\rangle = \int_0^2 x^2 B_{-1}^1(x)dx = \int_0^1 x^2(1-x)dx + \int_1^2 x^2 0 dx = \frac{1}{12}$$

De modo análogo

$$\langle x^2, B_0^1\rangle = \int_0^2 x^2 B_0^1(x)dx = \frac{7}{6}, \quad \langle x^2, B_1^1\rangle = \int_0^2 x^2 B_1^1(x)dx = \frac{17}{12}$$

$$\langle B_{-1}^1, B_{-1}^1\rangle = \int_0^2 B_{-1}^1(x) B_{-1}^1(x)dx = \int_0^1 (1-x)^2 dx = \frac{1}{3}$$

$$\langle B_{-1}^1, B_0^1\rangle = \int_0^2 B_{-1}^1(x) B_0^1(x)dx = \frac{1}{6} = \langle B_0^1, B_{-1}^1\rangle$$

$$\langle B_{-1}^1, B_1^1\rangle = 0 = \langle B_1^1, B_{-1}^1\rangle, \quad \langle B_0^1, B_0^1\rangle = \frac{2}{3}, \quad \langle B_0^1, B_1^1\rangle = \frac{1}{6} = \langle B_1^1, B_0^1\rangle, \quad \langle B_1^1, B_1^1\rangle = \frac{1}{3}$$

Calculados los coeficientes, el sistema queda

$$\begin{pmatrix} 1/3 & 1/6 & 0 \\ 1/6 & 2/3 & 1/6 \\ 0 & 1/6 & 1/3 \end{pmatrix} \begin{pmatrix} a_{-1} \\ a_0 \\ a_1 \end{pmatrix} = \begin{pmatrix} 1/12 \\ 7/6 \\ 17/12 \end{pmatrix}$$

Cuya solución es

$$\begin{pmatrix} a_{-1} \\ a_0 \\ a_1 \end{pmatrix} = \begin{pmatrix} -1/6 \\ 5/6 \\ 23/6 \end{pmatrix}$$

Por tanto, el spline solución s, representado en la Figura 4.40, es

$$s = -\frac{1}{6}B_{-1}^1 + \frac{5}{6}B_0^1 + \frac{23}{6}B_1^1 = \begin{cases} x - \frac{1}{6} & 0 \le x \le 1 \\ 3x - \frac{13}{6} & 1 \le x \le 2 \end{cases}$$

Comentario 1

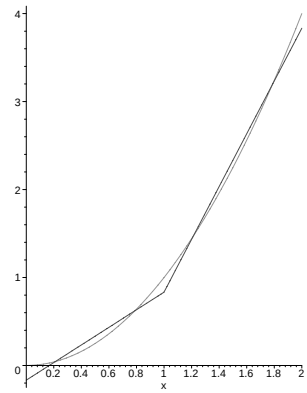


Figura 4.40: Dibujo del spline solución del problema 4.4.

A menudo, aprendemos más de los errores. Vamos a comentar un error muy común en este ejercicio, y que conduce por casualidad a un resultado exacto en este caso, pero no en general. La idea es sustituir la aproximación por un spline, por la aproximación por una recta entre 0 y 1, y por otra recta diferente entre 1 y 2. Casualmente, dichas rectas coinciden en 1, y forman por tanto un spline de primer grado que es el mismo que hemos calculado antes. Veámoslo. En el primer tramo, buscamos $a + bx$ tal que

$$\langle a + bx - x^2, 1\rangle = 0 \quad \Rightarrow \quad \int_0^1 \left(a + bx - x^2\right) dx = 0$$

$$\langle a + bx - x^2, x\rangle = 0 \quad \Rightarrow \quad \int_0^1 \left(a + bx - x^2\right) x dx = 0$$

De donde obtenemos el sistema lineal y los coeficientes

$$\begin{pmatrix} 1 & 1/2 \\ 1/2 & 1/3 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} 1/3 \\ 1/4 \end{pmatrix} \quad \Rightarrow \quad \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} -1/6 \\ 1 \end{pmatrix}$$

La recta solución en el primer tramo es:

$$r1 \equiv x - \frac{1}{6}$$

En el segundo tramo, buscamos $c + dx$ tal que

$$\langle c + dx - x^2, 1\rangle = 0 \quad \Rightarrow \quad \int_1^2 \left(c + dx - x^2\right) dx = 0$$

y

$$\langle c + dx - x^2, x\rangle = 0 \quad \Rightarrow \quad \int_1^2 \left(c + dx - x^2\right) x dx = 0$$

De donde

$$\begin{pmatrix} 1 & 3/2 \\ 3/2 & 7/3 \end{pmatrix} \begin{pmatrix} c \\ d \end{pmatrix} = \begin{pmatrix} 7/3 \\ 15/4 \end{pmatrix} \quad \Rightarrow \quad \begin{pmatrix} c \\ d \end{pmatrix} = \begin{pmatrix} -13/6 \\ 3 \end{pmatrix}$$

La recta solución en el segundo tramo es

$$r2 \equiv 3x - \frac{13}{6}$$

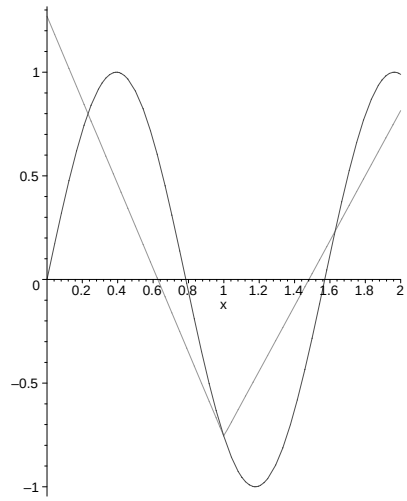


Figura 4.41: Solución correcta para el ejemplo adicional del problema 4.4.

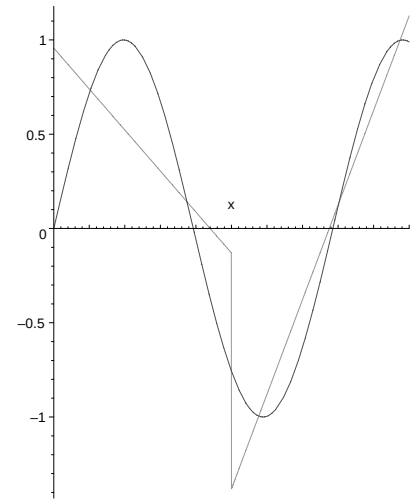


Figura 4.42: Solución errónea para el ejemplo adicional del problema 4.4.

Como vemos, las dos rectas coinciden con las que forman el spline que hemos calculado del primer modo. La pregunta es: ¿Qué hubiese sucedido si la función a aproximar hubiera sido, por ejemplo, $f(x) = sen(4x)$? Lo que hubiéramos obtenido con el primer método (ver Figura 4.41) hubiese sido

$$\begin{aligned} s1 &= 1.269741B_{-1}^1 - 0.755681B_0^1 + 0.814370B_1^1 \\ &= \begin{cases} 1.269741 - 2.025422x & 0 \le x \le 1 \\ -2.325732 + 1.570051x & 1 \le x \le 2 \end{cases} \end{aligned}$$

y por el segundo:

$$s2 = \begin{cases} 0.956979 - 1.087136x & 0 \le x \le 1 \\ -3.88954 + 2.508336x & 1 \le x \le 2 \end{cases}$$

que no es un spline de grado 1, al no ser continuo para $x = 1$, (ver Figura 4.42).

Comentario 2

Todavía haremos otro planteamiento, algo menos elegante, de este problema. La idea de esta formulación es escribir el spline solución en dos tramos:

$$s = \begin{cases} ax + b & 0 \le x \le 1 \\ cx + d & 1 \le x \le 2 \end{cases}$$

eligiendo los coeficientes a, b, c, d de tal forma que el error en esa norma sea mínimo y que la función que obtengamos sea continua en $x = 1$.

$$E^2 = \| \, s - x^2 \, \|^2 = \int_0^1 \left(ax + b - x^2\right)^2 dx + \int_1^2 \left(cx + d - x^2\right)^2 dx$$

$$= \left(\frac{a^2}{3} + b^2 + \frac{1}{5} + ab - \frac{a}{2} - \frac{2b}{3}\right) + \left(\frac{7c^2}{3} + d^2 + \frac{31}{5} + 3cd - \frac{15c}{2} - \frac{14d}{3}\right)$$

Si imponemos que la función debe ser continua para $x = 1$

$$b = c + d - a$$

Por tanto:

$$E^2 = \frac{a^2}{3} + \frac{10c^2}{3} + 2d^2 + 5cd - ad - ac + \frac{a}{6} - \frac{49c}{6} - \frac{16d}{3} + \frac{32}{5}$$

Para encontrar el mínimo de esta función E^2, derivamos respecto a las tres variables a, c, d e igualamos a 0, lo que conduce al sistema lineal en a, c, d.

$$\begin{pmatrix} 2/3 & -1 & -1 \\ -1 & 20/3 & 5 \\ -1 & 5 & 4 \end{pmatrix} \begin{pmatrix} a \\ c \\ d \end{pmatrix} = \begin{pmatrix} -1/6 \\ 49/6 \\ 16/3 \end{pmatrix} \quad \Rightarrow \quad \begin{pmatrix} a \\ c \\ d \end{pmatrix} = \begin{pmatrix} 1 \\ 3 \\ -13/6 \end{pmatrix}$$

y $b = -1/6$. Por tanto, el spline solución s es otra vez el mismo

$$s = \begin{cases} x - \frac{1}{6} & 0 \le x \le 1 \\ 3x - \frac{13}{6} & 1 \le x \le 2 \end{cases}$$

PROBLEMA 4.5 *Aproximación por mínimos cuadrados en un espacio de splines.*

Definir de modo preciso el conjunto de las mejores aproximaciones por mínimos cuadrados de la siguiente nube de puntos dentro de los splines de grado 2, que tienen como soporte la partición $\Omega = \{0, 1, 2\}$ del compacto $[0, 2]$.

i	x_i	y_i
0	0.0	1.0
1	0.5	0.0
2	1.0	1.0
3	1.5	0.0
4	2.0	1.0

Caso de que para hallar la solución del problema sea necesaria la resolución de un sistema lineal, ésta se hará por un método directo, eligiendo de modo razonado la descomposición que mejor se ajuste al tipo de sistema lineal. Se pide dar todos los pasos y matrices intermedias de dicha descomposición. Se exige una precisión de al menos tres cifras decimales en todos los cálculos.

Solución:

Revisemos un poco la teoría antes de hacer el problema. Dada una nube de $n+1$ puntos $(x_0, y_0), (x_1, y_1), \cdots, (x_n, y_n)$, en el espacio prehilbertiano $I\!R^{n+1}$ dotado del producto escalar euclídeo

$$\langle a, b \rangle = \sum_{i=0}^{n} a_i b_i, \quad \forall a, b \in I\!R^{n+1}$$

y de su norma asociada

$$\| a \|_2 = \sqrt{\langle a, a \rangle} = \sqrt{\sum_{i=0}^{n} a_i^2}, \quad \forall a \in I\!R^{n+1}$$

en la aproximación por mínimos cuadrados buscamos una función f engendrada por una familia de funciones $\{g_1, \cdots, g_m\}$ base del espacio funcional en el que se plantea el problema

$$f = \sum_{i=1}^{m} a_i g_i$$

tal que sus valores en las abscisas $x_0, x_1, \cdots, x_n$, aproximen los valores $y_0, y_1, \cdots, y_n$ en el sentido de la norma 2, o sea, que hagan que

$$\| (f(x_0) - y_0, \cdots, f(x_n) - y_n) \|_2$$

sea mínimo. Denotemos $\overline{y} = (y_0, y_1, \cdots, y_n) \in I\!R^{n+1}$, $\overline{f} = (f(x_0), \cdots, f(x_n)) \in I\!R^{n+1}$, $\overline{g_i} = (g_i(x_0), \cdots, g_i(x_n)) \in I\!R^{n+1}$, $1 \le i \le m$ y $\overline{g} = (g(x_0), \cdots, g(x_n)) \in I\!R^{n+1}$, $\forall g \in L(g_1, \cdots, g_m)$. Buscamos por tanto

$$\overline{f} = \sum_{i=1}^{m} a_i \overline{g_i}$$

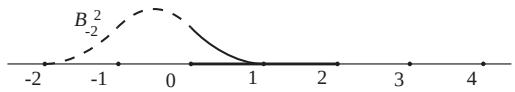


Figura 4.43: Soporte de $B^2_{-2} \equiv g_1$.

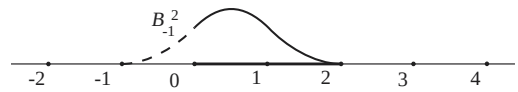


Figura 4.44: Soporte de $B^2_{-1} \equiv g_2$.

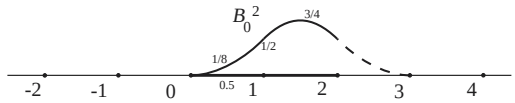


Figura 4.45: Soporte de $B^2_0 \equiv g_3$.

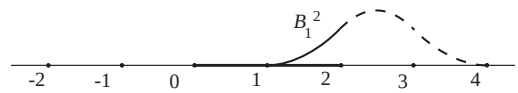


Figura 4.46: Soporte de $B^2_1 \equiv g_4$.

tal que:

$$\| \overline{y} - \overline{f} \|_2 \leq \| \overline{y} - \overline{g} \|_2 \; \forall \overline{g} \in L\left(\overline{g_1}, \cdots, \overline{g_m}\right)$$

Al reformular nuestro problema como uno de mejor aproximación en subespacios vectoriales de dimensión finita de espacios prehilbertianos, $I\!R^m \subset I\!R^{n+1}$, si $m \leq n+1$, dicha mejor aproximación será la proyección ortogonal de $\overline{y}$ sobre $L\left(\overline{g_1}, \cdots, \overline{g_m}\right)$, lo cual se traduce en la condición

$$\langle \overline{y} - \overline{f}, \overline{g_j} \rangle = 0 \quad \Rightarrow \quad \langle \overline{y} - \sum_{i=1}^{m} a_i \overline{g_i}, \overline{g_j} \rangle = 0 \quad \text{para} \quad 1 \leq j \leq m$$

que conduce al sistema lineal:

$$\sum_{i=1}^{m} a_i \langle \overline{g_i}, \overline{g_j} \rangle = \langle \overline{y}, \overline{g_j} \rangle$$

Para comprender bien este paso del continuo al discreto se aconseja consultar el Capítulo 6 del libro de Hämmerlin y Hoffmann [15]. Enmarquemos nuestro problema dentro de este esquema identificando cada uno de los elementos:

$$\overline{y} = (y_0, y_1, y_2, y_3, y_4) = (1, 0, 1, 0, 1) \in I\!R^5$$

El subespacio donde buscamos la mejor aproximación son los splines de grado 2 que tienen como soporte la partición $\Omega = \{0, 1, 2\}$ del compacto $[0, 2]$, o sea, $S_2(\Omega)$, cuya dimensión es $n + k = 2 + 2 = 4$, siendo n el número de tramos, 2, y k el grado, 2.

Tenemos que encontrar una base g_1, g_2, g_3, g_4 de $S_2(\Omega)$. La elegiremos de tal forma que la matriz del sistema sea sencilla de calcular. Como la partición es equiespaciada, la elección natural es la de los B-splines.

Tenemos que añadir dos nodos virtuales a la izquierda para completar los 4 elementos de la base, g_1, g_2, g_3, g_4, cuyo soporte representamos en las Figuras 4.43, 4.44, 4.45 y 4.46.

Ahora que tenemos los cuatro elementos de la base, calculemos

$$\begin{aligned} \overline{g_i} &= \left(g_i(x_0), g_i(x_1), g_i(x_2), g_i(x_3), g_i(x_4)\right) \\ &= \left(g_i(0.0), g_i(0.5), g_i(1.0), g_i(1.5), g_i(2.0)\right) \in I\!R^5, \;\; 1 \leq i \leq 4 \end{aligned}$$

Para respetar la indexación de los B-splines, pondremos

$$\begin{aligned} \overline{B_i} &= \left(B_i(x_0), B_i(x_1), B_i(x_2), B_i(x_3), B_i(x_4)\right) \\ &= \left(B_i(0.0), B_i(0.5), B_i(1.0), B_i(1.5), B_i(2.0)\right) \in I\!R^5, \;\; -2 \leq i \leq 1 \end{aligned}$$

Calculemos estos vectores

$$\begin{aligned} \overline{B_{-2}} &= \left(B_{-2}(x_0), B_{-2}(x_1), B_{-2}(x_2), B_{-2}(x_3), B_{-2}(x_4)\right) \\ &= \left(B_{-2}(0.0), B_{-2}(0.5), B_{-2}(1.0), B_{-2}(1.5), B_{-2}(2.0)\right) \quad \Rightarrow \\ \overline{B_{-2}} &= (1/2, 1/8, 0, 0, 0) \end{aligned}$$

De modo similar

$$\overline{B_{-1}} = (1/2, 3/4, 1/2, 1/8, 0)\,, \quad \overline{B_0} = (0, 1/8, 1/2, 3/4, 1/2)\,, \quad \overline{B_1} = (0, 0, 0, 1/8, 1/2)$$

Calculemos los elementos de la matriz del sistema lineal

$$\begin{aligned}
\langle\overline{B_{-2}}, \overline{B_{-2}}\rangle &= (1/2, 1/8, 0, 0, 0)\,(1/2, 1/8, 0, 0, 0) = 17/64\\
\langle\overline{B_{-2}}, \overline{B_{-1}}\rangle &= (1/2, 1/8, 0, 0, 0)\,(1/2, 3/4, 1/2, 1/8, 0) = 11/32\\
\langle\overline{B_{-2}}, \overline{B_0}\rangle &= (1/2, 1/8, 0, 0, 0)\,(0, 1/8, 1/2, 3/4, 1/2) = 1/64\\
\langle\overline{B_{-2}}, \overline{B_1}\rangle &= (1/2, 1/8, 0, 0, 0)\,(0, 0, 0, 1/8, 1/2) = 0\\
\langle\overline{B_{-1}}, \overline{B_{-1}}\rangle &= (1/2, 3/4, 1/2, 1/8, 0)\,(1/2, 3/4, 1/2, 1/8, 0) = 69/64\\
\langle\overline{B_{-1}}, \overline{B_0}\rangle &= (1/2, 3/4, 1/2, 1/8, 0)\,(0, 1/8, 1/2, 3/4, 1/2) = 7/16\\
\langle\overline{B_{-1}}, \overline{B_1}\rangle &= (1/2, 3/4, 1/2, 1/8, 0)\,(0, 0, 0, 1/8, 1/2) = 1/64\\
\langle\overline{B_0}, \overline{B_0}\rangle &= (0, 1/8, 1/2, 3/4, 1/2)\,(0, 1/8, 1/2, 3/4, 1/2) = 69/64\\
\langle\overline{B_0}, \overline{B_1}\rangle &= (0, 1/8, 1/2, 3/4, 1/2)\,(0, 0, 0, 1/8, 1/2) = 11/32\\
\langle\overline{B_1}, \overline{B_1}\rangle &= (0, 0, 0, 1/8, 1/2)\,(0, 0, 0, 1/8, 1/2) = 17/64
\end{aligned}$$

y del segundo miembro

$$\begin{aligned}
\langle\overline{y}, \overline{B_{-2}}\rangle &= (1, 0, 1, 0, 1)\,(1/2, 1/8, 0, 0, 0) = 1/2\\
\langle\overline{y}, \overline{B_{-1}}\rangle &= (1, 0, 1, 0, 1)\,(1/2, 3/4, 1/2, 1/8, 0) = 1\\
\langle\overline{y}, \overline{B_0}\rangle &= (1, 0, 1, 0, 1)\,(0, 1/8, 1/2, 3/4, 1/2) = 1\\
\langle\overline{y}, \overline{B_1}\rangle &= (1, 0, 1, 0, 1)\,(0, 0, 0, 1/8, 1/2) = 1/2
\end{aligned}$$

Por tanto, el sistema lineal presenta el estupendo aspecto siguiente:

$$\begin{pmatrix} 17/64 & 11/32 & 1/64 & 0 \\ 11/32 & 69/64 & 7/16 & 1/64 \\ 1/64 & 7/16 & 69/64 & 11/32 \\ 0 & 1/64 & 11/32 & 17/64 \end{pmatrix} \begin{pmatrix} a_{-2} \\ a_{-1} \\ a_0 \\ a_1 \end{pmatrix} = \begin{pmatrix} 1/2 \\ 1 \\ 1 \\ 1/2 \end{pmatrix}$$

Que podemos convertir en este otro todavía más agradable:

$$\begin{pmatrix} 17 & 22 & 1 & 0 \\ 22 & 69 & 28 & 1 \\ 1 & 28 & 69 & 22 \\ 0 & 1 & 22 & 17 \end{pmatrix} \begin{pmatrix} a_{-2} \\ a_{-1} \\ a_0 \\ a_1 \end{pmatrix} = \begin{pmatrix} 32 \\ 64 \\ 64 \\ 32 \end{pmatrix}$$

En la segunda parte de este problema seleccionamos una descomposición adecuada que permita resolver el sistema lineal por un método directo. Como la matriz es la de Gramm de un producto escalar referida a una base, es simétrica definida positiva. La descomposición estudiada que mejor se ajusta a este problema es sin duda la de Cholesky.

$$A = LL^t$$

$$\begin{pmatrix} 17 & 22 & 1 & 0 \\ 22 & 69 & 28 & 1 \\ 1 & 28 & 69 & 22 \\ 0 & 1 & 22 & 17 \end{pmatrix} = \begin{pmatrix} l_{11} & 0 & 0 & 0 \\ l_{21} & l_{22} & 0 & 0 \\ l_{31} & l_{32} & l_{33} & 0 \\ l_{41} & l_{42} & l_{43} & l_{44} \end{pmatrix} \begin{pmatrix} l_{11} & l_{21} & l_{31} & l_{41} \\ 0 & l_{22} & l_{32} & l_{42} \\ 0 & 0 & l_{33} & l_{43} \\ 0 & 0 & 0 & l_{44} \end{pmatrix}$$

Para calcular los coeficientes l_{ij} se procede por identificación:

$$l_{11}^2 = 17 \;\rightarrow l_{11} = \sqrt{17} = 4.12311$$

$$l_{11} l_{21} = 22 \;\rightarrow l_{21} = \frac{22}{l_{11}} = 5.33578$$

$$l_{11}l_{31} = 1 \ \rightarrow l_{31} = \frac{1}{l_{11}} = 0.242536$$

$$l_{11}l_{41} = 0 \ \rightarrow l_{41} = 0.0$$

$$l_{21}^2 + l_{22}^2 = 69 \ \rightarrow l_{22} = \sqrt{69 - l_{21}^2} = 6.36627$$

$$l_{21}l_{31} + l_{22}l_{32} = 28 \ \rightarrow l_{32} = \frac{28 - l_{21}l_{31}}{l_{22}} = 4.19490$$

$$l_{21}l_{41} + l_{22}l_{42} = 1 \ \rightarrow l_{42} = \frac{28 - l_{21}l_{41}}{l_{22}} = 0.1571$$

$$l_{31}^2 + l_{32}^2 + l_{33}^2 = 69 \ \rightarrow l_{33} = \sqrt{69 - l_{31}^2 - l_{32}^2} = 7.1655$$

$$l_{31}l_{41} + l_{32}l_{42} + l_{33}l_{43} = 22 \ \rightarrow l_{43} = \frac{22 - l_{31}l_{41} - l_{32}l_{42}}{l_{33}} = 2.97832$$

$$l_{41}^2 + l_{42}^2 + l_{43}^2 + l_{44}^2 = 17 \ \rightarrow l_{44} = \sqrt{17 - l_{41}^2 - l_{42}^2 - l_{43}^2} = 2.8556$$

Se resuelve el sistema lineal original, resolviendo dos sistemas triangulares. Definimos

$$a = \begin{pmatrix} a_{-2} \\ a_{-1} \\ a_0 \\ a_1 \end{pmatrix} \quad \text{y} \quad b = \begin{pmatrix} 32 \\ 64 \\ 64 \\ 32 \end{pmatrix}$$

Para resolver el sistema $LL^t a = b$ llamando $z = L^t a$, resolvemos primero por sustitución hacia adelante el sistema $Lz = b$ y después por sustitución hacia atrás el sistema $L^t a = z$. Veamos:

$$Lz = b$$

$$\begin{pmatrix} 4.1231 & 0 & 0 & 0 \\ 5.3358 & 6.3663 & 0 & 0 \\ 0.2425 & 4.1949 & 7.1655 & 0 \\ 0.0000 & 0.1571 & 2.9783 & 2.8556 \end{pmatrix} \begin{pmatrix} z_{-2} \\ z_{-1} \\ z_0 \\ z_1 \end{pmatrix} = \begin{pmatrix} 32 \\ 64 \\ 64 \\ 32 \end{pmatrix} \quad \Rightarrow \quad \begin{pmatrix} z_{-2} \\ z_{-1} \\ z_0 \\ z_1 \end{pmatrix} = \begin{pmatrix} 7.761 \\ 3.548 \\ 6.592 \\ 4.148 \end{pmatrix}$$

y por retrosustitución

$$L^t a = z$$

$$\begin{pmatrix} 4.1231 & 5.3358 & 0.2425 & 0.0000 \\ 0 & 6.3663 & 4.1949 & 0.1571 \\ 0 & 0 & 7.1655 & 2.9783 \\ 0 & 0 & 0 & 2.8556 \end{pmatrix} \begin{pmatrix} a_{-2} \\ a_{-1} \\ a_0 \\ a_1 \end{pmatrix} = \begin{pmatrix} 7.761 \\ 3.548 \\ 6.592 \\ 4.148 \end{pmatrix} \quad \Rightarrow \quad \begin{pmatrix} a_{-2} \\ a_{-1} \\ a_0 \\ a_1 \end{pmatrix} = \begin{pmatrix} 1.457 \\ 0.314 \\ 0.314 \\ 1.457 \end{pmatrix}$$

Por tanto el spline de grado dos que mejor ajusta por mínimos cuadrados la nube de puntos definida en el enunciado es:

$$s = 1.457B_{-2}^2 + 0.314B_{-1}^2 + 0.3147B_0^2 + 1.457B_1^2$$

que representamos en la Figura 4.47.

Comentario 1

Si se toma otra base de $S_2(\Omega)$, surgen otras formas de abordar el problema, aunque todas similares a ésta. Por ejemplo, imaginemos que tomamos como base de $S_2(\Omega)^*$ la formada por las 4 formas lineales independientes

$$L_0(S) = S(0), \quad L_1(S) = S(1), \quad L_2(S) = S(2), \quad L_3(S) = S'(0)$$

La base de $S_2(\Omega)$ dual de la anterior estará descrita por los 4 splines $\{C_0, C_1, C_2, C_3\}$ que verifican

$$L_i\left(C_j\right) = \delta_{ij}, \ i, j = 0, 3$$

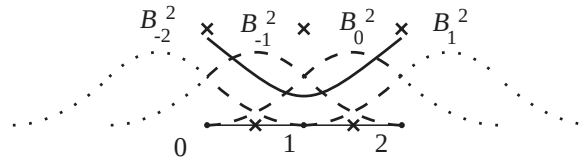


Figura 4.47: Solución del problema 4.5.

Calculemos, por ejemplo, C_0.

$$L_0(C_0) = C_0(0) = 1, \quad L_1(C_0) = C_0(1) = 0, \quad L_2(C_0) = C_0(2) = 0, \quad L_3(C_0) = C_0'(0) = 0$$

Construyamos los 2 tramos de C_0. Para desarrollar un procedimiento que sirva también para los otros C_i, merece la pena recordar cómo es una parábola que pasa por los puntos (x_0, w_0), (x_1, w_1), y cuya derivada en x_0 es s_0 (ver problema 3.11):

$$P_0(x) = w_0 + s_0\,(x - x_0) + \frac{w_1 - w_0 - s_0 h_0}{h_0^2}\,(x - x_0)^2$$

con $h_0 = x_1 - x_0$.
En general, suponiendo que esa parábola P_i es uno de los tramos de un spline parabólico, el definido entre los nodos de abscisas x_i y x_{i+1}, tendremos

$$P_i(x) = w_i + s_i\,(x - x_i) + \frac{w_{i+1} - w_i - s_i h_i}{h_i^2}\,(x - x_i)^2$$

con $h_i = x_{i+1} - x_i$.

Como es un spline parabólico, su derivada debe ser continua en los nodos donde enganchan los diferentes tramos, luego

$$P_i'\,(x_{i+1}) = P_{i+1}'\,(x_{i+1})$$

Imponiendo esta condición en cada uno de los nodos interiores, llegamos a la siguiente relación entre las derivadas en dichos nodos.

$$s_i + s_{i+1} = 2\frac{w_{i+1} - w_i}{h_i}$$

Y ya tenemos todas las herramientas para construir cada uno de los elementos de la base de $S_2(\Omega)$. Empecemos, como habíamos dicho, con C_0.

$$L_0(C_0) = C_0(0) = w_0 = 1, \quad L_1(C_0) = C_0(1) = w_1 = 0$$

$$L_2(C_0) = C_0(2) = w_2 = 0, \quad L_3(C_0) = C_0'(0) = s_0 = 0$$

de donde

$$C_0 = \begin{cases} w_0 + s_0(x-0) + \frac{w_1 - w_0 - s_0 h_0}{h_0^2}(x-0)^2 & 0 \le x \le 1 \\ w_1 + s_1(x-1) + \frac{w_2 - w_1 - s_1 h_1}{h_1^2}(x-1)^2 & 1 \le x \le 2 \end{cases}$$

La lista $(x_0, x_1, x_2, x_3, x_4)$ de nodos de la nube de puntos a ajustar por mínimos cuadrados no se corresponde con la lista de nodos de los splines de $S_2(\Omega)$, que son 0,1,2. Esto induce a confusión, pero la diferencia es clara. Identifiquemos cada uno de los datos

$$h_0 = h_1 = 1$$

$$s_0 + s_1 = 2\frac{w_1 - w_0}{h_0} \rightarrow s_1 = 2\frac{0-1}{1} - 0 = -2$$

Por tanto, entrando en la expresión anterior

$$C_0 = \begin{cases} 1-x^2 & 0 \le x \le 1 \\ 2x^2-6x+4 & 1 \le x \le 2 \end{cases}$$

De modo similar

$$C_1 = \begin{cases} x^2 & 0 \le x \le 1 \\ -3x^2+8x-4 & 1 \le x \le 2 \end{cases}, \quad C_2 = \begin{cases} 0 & 0 \le x \le 1 \\ (x-1)^2 & 1 \le x \le 2 \end{cases}, \quad C_3 = \begin{cases} x-x^2 & 0 \le x \le 1 \\ x^2-3x+2 & 1 \le x \le 2 \end{cases}$$

Una vez que tenemos la base, calculemos

$$\begin{aligned} \overline{C_0} &= (C_0(x_0), C_0(x_1), C_0(x_2), C_0(x_3), C_0(x_4)) \\ &= (C_0(0.0), C_0(0.5), C_0(1.0), C_0(1.5), C_0(2.0)) \\ &= (1, 3/4, 0, -1/2, 0) \end{aligned}$$

Análogamente

$$\overline{C_1} = (0, 1/4, 1, 5/4, 0), \quad \overline{C_2} = (0, 0, 0, 1/4, 1), \quad \overline{C_3} = (0, 1/4, 0, -1/4, 0)$$

y el sistema será esta vez

$$\begin{pmatrix} 1.8125 & -0.4375 & -0.1250 & 0.3125 \\ -0.4375 & 2.6250 & 0.3125 & -0.250 \\ -0.1250 & 0.3125 & 1.0625 & -0.0625 \\ 0.3125 & -0.250 & -0.0625 & 0.125 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 0 \end{pmatrix}$$

Se propone como ejercicio verificar todos estos cálculos, comprobando que el spline obtenido coincide con el referido a la base de B-splines.

Parece que el problema es simétrico respecto a $x = 1$. Podíamos haber intentado usar esa simetría para definir las formas lineales que definen la base de $S_2(\Omega)^*$

$$L_0(S) = S(0), \quad L_1(S) = S(1), \quad L_2(S) = S(2), \quad L_3(S) = S'(1)$$

Calculemos de nuevo C_0.

$$L_0(C_0) = C_0(0) = w_0 = 1, \quad L_1(C_0) = C_0(1) = w_1 = 0$$

$$L_2(C_0) = C_0(2) = w_2 = 0, \quad L_3(C_0) = C_0'(1) = s_1 = 0$$

$$s_0 + s_1 = 2\frac{w_1 - w_0}{h_0} \to s_0 = 2\frac{0-1}{1} - 0 = -2$$

Por tanto, entrando en la expresión anterior:

$$C_0 = \begin{cases} 1-2x+x^2 & 0 \le x \le 1 \\ 0 & 1 \le x \le 2 \end{cases}$$

etcétera.
Todavía hay otra posibilidad muy curiosa para dotar de base a $S_2(\Omega)$, usando la simetría del problema, combinando elementos típicamente polinómicos con splines de varios tramos. No hay que olvidar que un polinomio es un caso particular de spline, pues a un spline de grado 2 se le exige que sea C^1, y un polinomio, no sólo es C^1, sino C^∞.

$$C_0 = x^2,\ 0 \le x \le 2, \quad C_1 = (x-2)^2,\ 0 \le x \le 2$$

$$C_2 = \begin{cases} 0 & 0 \le x \le 1 \\ (x-1)^2 & 1 \le x \le 2 \end{cases}, \quad C_3 = \begin{cases} (x-1)^2 & 0 \le x \le 1 \\ 0 & 1 \le x \le 2 \end{cases}$$

Se propone que acabéis el ejercicio con esta base.

Comentario 2
Existe otra posibilidad de hacer el problema. Como la mejor aproximación dentro de los splines de esa nube de puntos va a ser un spline, escribamos cada uno de sus tramos

$$s = \begin{cases} a_1x^2 + b_1x + c_1 & 0 \le x \le 1 \\ a_2x^2 + b_2x + c_2 & 1 \le x \le 2 \end{cases}$$

Imponiendo las condiciones de continuidad de la función y de su derivada en el nodo intermedio, eliminamos dos parámetros, por ejemplo b_2, y c_2. El spline solución quedará entonces

$$s = \begin{cases} a_1x^2 + b_1x + c_1 & 0 \le x \le 1 \\ a_2x^2 + (2a_1 + b_1 - 2a_2)\,x + c_1 + a_2 - a_1 & 1 \le x \le 2 \end{cases}$$

Como el spline de mínimos cuadrados es el que hace mínimo el valor

$$E^2(a_1, b_1, c_1, a_2) = \sum_{i=0}^{4} (y_i - s(x_i))^2$$

buscamos (a_1, b_1, c_1, a_2) que hagan mínimo dicho valor, para lo cual:

$$\frac{\partial E^2}{\partial a_1} = \frac{\partial E^2}{\partial b_1} = \frac{\partial E^2}{\partial c_1} = \frac{\partial E^2}{\partial a_2} = 0$$

igualdades que conduce a un sistema lineal de 4 ecuaciones con 4 incógnitas.

PROBLEMA 4.6 *Aproximación por mínimos cuadrados en un espacio de polinomios a trozos.*

Se considera la partición $\Omega = \{-1, 0, 1, 2\}$ del compacto $[-1, 2]$. Sea $P_{2,0}(\Omega)$ el espacio vectorial de los polinomios a trozos de grado 2 de clase 0, o sea, tales que la restricción a cada intervalo es un polinomio de grado 2, y el enganche se produce con continuidad. Ver la Figura 3.41.

Se considera la base formada por los siguientes elementos:

$$l_0(t) = \begin{cases} 1 - (t+1)^2 & t \in [-1, 0) \\ 0 & t \in [0, 1) \\ 0 & t \in [1, 2] \end{cases} \quad l_1(t) = \begin{cases} (t+1)^2 & t \in [-1, 0) \\ 1 - t^2 & t \in [0, 1) \\ 0 & t \in [1, 2] \end{cases}$$

$$l_2(t) = \begin{cases} 0 & t \in [-1, 0) \\ t^2 & t \in [0, 1) \\ 1 - (t-1)^2 & t \in [1, 2] \end{cases} \quad l_3(t) = \begin{cases} 0 & t \in [-1, 0) \\ 0 & t \in [0, 1) \\ (t-1)^2 & t \in [1, 2] \end{cases}$$

$$l_4(t) = \begin{cases} (t+1) - (t+1)^2 & t \in [-1, 0) \\ 0 & t \in [0, 1) \\ 0 & t \in [1, 2] \end{cases} \quad l_5(t) = \begin{cases} 0 & t \in [-1, 0) \\ t - t^2 & t \in [0, 1) \\ 0 & t \in [1, 2] \end{cases}$$

$$l_6(t) = \begin{cases} 0 & t \in [-1, 0) \\ 0 & t \in [0, 1) \\ (t-1) - (t-1)^2 & t \in [1, 2] \end{cases}$$

En la cual, un elemento $p \in P_{2,0}(\Omega)$ se escribe como

$$p = f_0l_0 + f_1l_1 + f_2l_2 + f_3l_3 + s_0l_4 + s_1l_5 + s_2l_6$$

siendo f_i el valor del polinomio en el punto x_i y s_i la derivada por la derecha en ese mismo punto. Esta base se corresponde con la estudiada en el problema 3.16.

Se considera el problema de encontrar la mejor aproximación dentro de ese espacio vectorial, por mínimos cuadrados a la siguiente nube de puntos:

i	0	1	2	3	4	5	6	7
x_i	-1	-0.5	0	0.4	0.8	1	1.5	2
y_i	0	0	-0.5	1	1	-1	0	0

1. Calcular cada uno de los coeficientes del sistema lineal que permite encontrar dicha aproximación mínimo cuadrática.

2. Aproximar el autovalor de módulo máximo de la matriz del sistema dando dos pasos con el método de la potencia tomando como estimador inicial un vector cuyas componentes sean todas la unidad.

3. Se sabe que el autovalor de módulo máximo de la inversa de la matriz del sistema tiene como módulo 31.1967. ¿Qué podemos decir justificadamente del condicionamiento del sistema?

Solución:

1. Cálculo de los coeficientes del sistema lineal:

$$\begin{aligned}
\overline{l_0} &= (l_0(-1), l_0(-0.5), l_0(0.0), l_0(0.4), l_0(0.8), l_0(1.0), l_0(1.5), l_0(2.0)) \\
&= (1, 0.75, 0, 0, 0, 0, 0, 0,) \\
\overline{l_1} &= (0, 0.25, 1, 0.84, 0.36, 0, 0, 0,) \\
\overline{l_2} &= (0, 0, 0, 0.16, 0.64, 1, 0.75, 0,) \\
\overline{l_3} &= (0, 0, 0, 0, 0, 0, 0.25, 1) \\
\overline{l_4} &= (0, 0.25, 0, 0, 0, 0, 0, 0,) \\
\overline{l_5} &= (0, 0, 0, 0.24, 0.16, 0, 0, 0,) \\
\overline{l_6} &= (0, 0, 0, 0, 0, 0, 0.25, 0) \\
\overline{y} &= (0, 0, -0.5, 1, 1, -1, 0, 0)
\end{aligned}$$

$$A_{ij} = \langle \overline{l_i}, \overline{l_j} \rangle \qquad b_i = \langle \overline{l_i}, \overline{y} \rangle$$

$$\begin{pmatrix}
1.5625 & 0.1875 & 0.0000 & 0.0000 & 0.1875 & 0.0000 & 0.0000 \\
 & 1.8977 & 0.3648 & 0.0000 & 0.0625 & 0.2592 & 0.0000 \\
 & & 1.9977 & 0.1875 & 0.0000 & 0.1408 & 0.1875 \\
 & & & 1.0625 & 0.0000 & 0.0000 & 0.0625 \\
 & & & & 0.0625 & 0.0000 & 0.0000 \\
 & SIM & & & & 0.0832 & 0.0000 \\
 & & & & & & 0.0625
\end{pmatrix}
\begin{pmatrix} f_0 \\ f_1 \\ f_2 \\ f_3 \\ s_0 \\ s_1 \\ s_2 \end{pmatrix}
=
\begin{pmatrix} 0.0 \\ 0.7 \\ -0.2 \\ 0.0 \\ 0.0 \\ 0.4 \\ 0.0 \end{pmatrix}$$

2. Método de la potencia con normalizazión:

   Elegimos q_0 con norma 1

$$\begin{aligned}
q_0 &= (1, 1, 1, 1, 1, 1, 1)^t, \qquad \| q_0 \|_\infty = 1 \\
x_1 &= Aq_0 = (1.9375, 2.7717, 2.8783, 1.3125, 0.3125, 0.4832, 0.3125)^T \\
\lambda_1 &= \frac{x_1}{q_0} = (1.9375, 2.7717, 2.8783, 1.3125, 0.3125, 0.4832, 0.3125)^T \\
\gamma_1 &= \| x_1 \|_\infty = 2.8783 \\
q_1 &= \frac{x_1}{\gamma_1} = (0.6731, 0.9630, 1, 0.4560, 0.1086, 0.1679, 0.1086)^T \\
x_2 &= A\, q_1 = (1.2527, 2.3687, 2.4785, 0.6768, 0.1932, 0.4044, 0.2228)^T \\
\lambda_2 &= \frac{x_2}{q_1} = (1.8610, 2.4598, 2.4785, 1.4886, 1.7793, 2.4087, 2.0520)^T
\end{aligned}$$

   Tomemos como aproximación a λ_{max} la media de estos valores que es aproximadamente 2.

3.

$$cond_2(A) = \| A \|_2 \| A^{-1} \|_2$$

Como A es real y simétrica, A es normal y por lo tanto

$$\| A \|_2 = \rho(A)$$

$$cond_2(A) = \rho(A) * \rho(A^{-1}) \cong 2 \cdot 31.19 \approx 62$$

Es un sistema relativamente mal condicionado, aunque si pensamos en su determinante, creeríamos que todavía es peor, pues $det(A) = 4\,10^{-4}$.

Este problema tiene como antecedente el problema 3.16 de interpolación lineal.
En la dirección *http://canal.etsin.upm.es/ftp/p20aprox.zip* tenemos el fichero Matlab que resuelve y generaliza este problema para cualquier número de nodos.
Presentamos la solución y un ejemplo adicional con el siguiente código Matlab.

```
% Solucion numerica
%Introduce la nube de puntos para x: [-1 -0.5 0 0.4 0.8 1 1.5 2]
%Introduce la nube de puntos para y: [0 0 -0.5 1 1 -1 0 0]
%Introduce el valor de tmin: -1 Introduce el valor de tmax: 2
%Introduce el valor de n: 3
A =
    1.5625    0.1875         0         0    0.1875         0         0
    0.1875    1.8977    0.3648         0    0.0625    0.2592         0
         0    0.3648    1.9977    0.1875         0    0.1408    0.1875
         0         0    0.1875    1.0625         0         0    0.0625
    0.1875    0.0625         0         0    0.0625         0         0
         0    0.2592    0.1408         0         0    0.0832         0
         0         0    0.1875    0.0625         0         0    0.0625

%B =     0      0.7000   -0.2000         0         0      0.4000       0
%sol= 0.0000   -0.5867   -0.7688    0.0000    0.5867    7.9366    2.3065
```

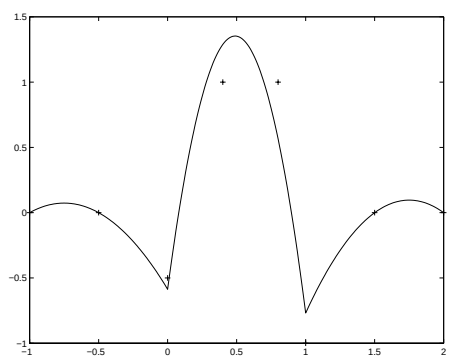


Figura 4.48: Gráfica de la solución del problema 4.6.

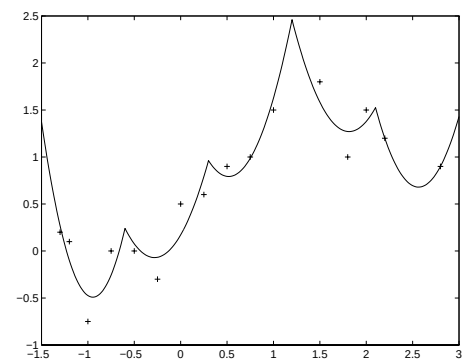


Figura 4.49: Solución del ejemplo adicional del problema 4.6.

```
% Ejemplo 1 adicional.

%Introduce la nube de puntos para x:
% [-1.3 -1.2 -1 -0.75 -0.5 -0.250 0.25 0.5 0.75 1 1.5 1.8 2 2.2 2.8]
%Introduce la nube de puntos para y:
```

```
%[0.2 0.1 -0.75 0 0 -0.3 0.5 0.6 0.9 1 1.5 1.8 1 1.5 1.2 0.9]
%Introduce el valor de tmin: -1.5
%Introduce el valor de tmax: 3
%Introduce el valor de n: 6
A =

  2.2651  0.5713       0       0       0       0  0.5175       0       0       0       0
  0.5713  2.6085  0.4838       0       0       0  0.1853  0.3855       0       0       0
       0  0.4838  2.6384  0.4734       0       0       0  0.1645  0.3781       0       0
       0       0  0.4734  1.5737  0.5115       0       0       0  0.1580  0.3075       0
       0       0       0  0.5115  1.9657  0.2512       0       0       0  0.1813  0.1492
       0       0       0       0  0.2512  0.3661       0       0       0       0  0.0952
  0.5175  0.1853       0       0       0       0  0.1292       0       0       0       0
       0  0.3855  0.1645       0       0       0       0  0.0959       0       0       0
       0       0  0.3781  0.1580       0       0       0       0  0.0990       0       0
       0       0       0  0.3075  0.1813       0       0       0       0  0.0879       0
       0       0       0       0  0.1492  0.0952       0       0       0       0  0.0321

%B = -0.2395 -0.1225   2.9102  3.6722  3.3704  0.5593  -0.1156  0.0642  0.5983   0.6933   0.2467
%sol =1.3729  0.2404   0.9620  2.4613  1.5257  1.4303  -6.7375 -1.9474 -1.5588  -3.8715  -3.6514
```

PROBLEMA 4.7 *Aproximación por mínimos cuadrados de funciones periódicas.*

Dada la siguiente señal periódica discreta

i	0	1	2	3
y_i	0.2	0.5	1.0	0.5

1. Suponiendo el índice como la variable independiente, encontrar el estimador mínimo cuadrático del tipo:
$$\alpha + \frac{\beta}{1+x^2}$$
que mejor ajuste uno de los periodos.

2. Calcular la frecuencia fundamental w_0 de la señal.

3. Encontrar la mejor aproximación por mínimos cuadrados en el subespacio vectorial U de exponenciales complejas de base:
$$B = \left\{1, e^{-jw_0t}, e^{jw_0t}\right\}$$
Calcular las componentes en esta base de esa mejor aproximación.

4. Calcular, utilizando Matlab, la DFT de la señal $\mathbf{y}$ y comprobar la correspondencia con los resultados del apartado 3.

5. Definir de modo preciso su parte real $r(x)$ dando sus componentes en función de las funciones independientes $\{1, \cos(w_0t), \sin(w_0t)\}$.

6. Esa tabla corresponde a una función cuya integral entre 0 y 3 vale 1.8925. Calcular esa integral mediante el método compuesto de los trapecios y también integrando directamente la función $r(x)$. ¿Cuál resulta ser más precisa?

Solución:

1.
$$g_1(t) = 1 \quad \text{y} \quad g_2(t) = \frac{1}{1+t^2}$$
$$\mathbf{g_1} = (1,1,1,1), \quad \mathbf{g_2} = (1, 0.5, 0.2, 0.1)$$

El sistema lineal queda:

$$\begin{pmatrix} 4 & 1.8 \\ 1.8 & 1.3 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \begin{pmatrix} 2.2 \\ 0.7 \end{pmatrix} \quad \Rightarrow \quad \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \begin{pmatrix} 0.8163 \\ -0.5918 \end{pmatrix}$$

2.
$$w_0 = \frac{2\pi}{4} = \frac{\pi}{2}$$

3. Buscamos la mejor aproximación $\hat{f}$ en un espacio de funciones del tipo

$$\phi_k(t) = e^{jkw_0t} \quad -1 \leq k \leq 1$$

Sabemos ((4.22) en la pág. 201) que los coeficientes de la descomposición única $\hat{f} = \sum_{k\in I} c_k\phi_k$ valen

$$c_k = \frac{1}{4}\sum_{n=0}^{n=3} y_n e^{-jkw_0n} = \frac{1}{4}\sum_{n=0}^{n=3} y_n(\cos(kw_0n) - j\sin(kw_0n))$$

Calculemos los diferentes términos de esa suma en forma de tabla.

n	y_n	$\cos w_0n$	$\sin w_0n$
0	0.2	1	0
1	0.5	0	1
2	1.0	-1	0
3	0.5	0	-1

Por tanto,

$$c_0 = 0.55, \quad c_{-1} = -0.2, \quad c_1 = -0.2$$

Como vemos, $c_{-1} = \overline{c_1}$, por ser la señal real.

4. Empleamos las siguientes líneas Matlab

```
y=[0.2 0.5 1.0 0.5];
fft(y)/4
```

y obtenemos la respuesta:

```
ans = 0.5500   -0.2000    0.0500   -0.2000
```

Si interpretamos estos coeficientes como se indica en la sección 4.8.4 de la teoría (pág. 204), veremos que coinciden con los del apartado 3.

5. Multiplicando y sumando tendremos la expresión como combinación lineal de funciones reales, que es la señal completa, pues en realidad, la aproximación también es real:

$$\hat{f}(t) = 0.55 - 0.4\cos\frac{\pi}{2}t$$

En la Figura 4.50 tenemos una vista de la nube de puntos dentro de la función original y de la función $\hat{f}(t)$. Esta figura se genera con las siguientes líneas Matlab:

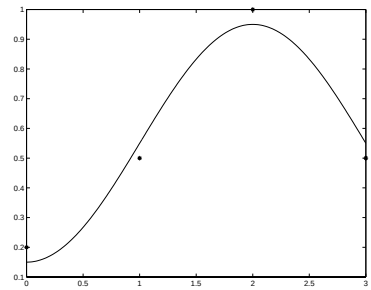


Figura 4.50: Nube de puntos y aproximación trigonométrica.

```
i=0:3;
y=[0.2 0.5 1.0 0.5];
t=0:0.01:3;
hatf=0.55-0.4*cos(pi/2*t);
plot(i,y,'*',t,hatf);
```

6. La integral de esta función vale 1.9046 y la calculada por el método compuesto de los trapecios 1.85. El valor real es 1.8925, por tanto la aproximación trigonométrica aproxima mejor la integral.

PROBLEMA 4.8 *Aproximación por mínimos cuadrados de funciones periódicas.*

Se excita un sistema con una sinusoide de frecuencia angular $\Omega = 104.7198$ rad/sg. Se muestrea la señal de respuesta cada 0.01 segundos y se obtiene la siguiente respuesta periódica $\mathbf{y}$

n	0	1	2	3	4	5
t_n	0.00	0.01	0.02	0.03	0.04	0.05
y_n	0.1344	1.0336	0.8700	0.1363	-0.7901	-0.6997

1. Calcular las componentes de la aproximación mínimo cuadrática de la señal $\mathbf{y}$ en la base de exponenciales complejas de frecuencia 0, Ω y 2Ω.

2. Calcular, utilizando Matlab, la DFT de la señal $\mathbf{y}$ y comprobar la correspondencia con los resultados del apartado 1.

Solución:

1. El periodo de la señal discreta dada es $N = 6$. Las componentes buscadas son

$$c_k = \frac{\langle \mathbf{y}, \Phi_k \rangle}{N} = \frac{1}{N}\sum_{n=0}^{N-1} y_n e^{-jk\left(\frac{2\pi}{N}\right)n} = \frac{1}{N}\sum_{n=0}^{5} y_n e^{-jk\left(\frac{2\pi}{6}\right)n} = \frac{1}{N}\sum_{n=0}^{5} y_n e^{-jk\Omega t_n}$$

   Si utilizamos como variable independiente el tiempo, tendremos que considerar frecuencias múltiplos de Ω y si consideramos que la variable independiente es el índice, tendremos que considerar las frecuencias como múltiplos de $w_0 = 2\pi/6$.

   El código Matlab para calcular esas sumas, cambiando k que nos indica el múltiplo de la frecuencia, es:

```
k=0;
n=0:5;
N=6;
```

```
y =[0.1344    1.0336    0.8700    0.1363   -0.7901   -0.6997];
w0=2*pi/N;
coseno=cos(k*w0*n);
seno=-sin(k*w0*n);
areal=(y*coseno')/N
aimag=(y*seno')/N
```

de donde

$$c_0 = 0.1141, \quad c_{-1} = 0.0208 + 0.4898j, \quad c_{-2} = 0.0106 + 0.0106j$$

Se deja como ejercicio calcular las equivalentes para k positivo y sumar todas ellas para tener la señal real total, generando después las gráficas de la nube original y de su aproximación con Matlab. El ejercicio es interesante, porque en realidad la respuesta del sistema es el seno original ligeramente perturbado, y la amplitud de las componentes diferentes de la primera es pequeña comparada con ésa.

2. Empleamos las siguientes líneas Matlab:

```
N=6;
y =[0.1344    1.0336    0.8700    0.1363   -0.7901   -0.6997];
fft(y)/N
```

y obtenemos la respuesta:

```
%ans =  Columns 1 through 3;  0.1141  0.0208-0.4898i   0.0106-0.0106i
%       Columns 4 through 6; -0.0427  0.0106+0.0106i   0.0208+0.4898i
```

Si interpretamos estos coeficientes como se indica en la sección 4.8.4 de la teoría, veremos (pág. 204) que coinciden con los del apartado 1.

CAPÍTULO 5

Integración y diferenciación por métodos numéricos

El cálculo numérico de integrales definidas es un problema antiguo que surge de modo natural al intentar calcular el área limitada por líneas curvas. Se estudió durante muchos años mediante técnicas de aproximación, ya que el cálculo integral apareció entre los siglos XVII y XVIII. Los griegos dedicaron mucho esfuerzo a estudiar el valor del área del círculo con las herramientas de las que disponían entonces. Los babilonios y los egipcios ya se habían preocupado antes de este mismo problema y del de encontrar el área encerrada por elipses y parábolas.

Hay varias situaciones en las que es necesario aproximar integrales definidas. La situación más común se produce cuando se quiera estimar una integral de una función de la que se conocen solamente ciertos valores, pero también se utiliza cuando no se sepa expresar analíticamente la primitiva de una función, por ejemplo, al evaluar la longitud de una curva (como veremos en el problema 5.8) y cuando la función admita primitiva que sea muy difícil de calcular.

Si se proporcionan los valores de una función f en ciertos puntos, digamos $x_0, x_1, \ldots, x_n$, ¿se puede usar esta información para obtener una estimación de cómo es su derivada $f'(c)$ o una integral $\int_a^b f(x)dx$? La respuesta requiere interpolar f mediante una cierta función que es la que se deriva o integra para obtener la información pedida. Los polinomios son una buena elección debido a la gran facilidad con que se integran y derivan usando sólo operaciones aritméticas básicas.

Ya hemos comentado que el cálculo numérico de integrales definidas es un problema muy antiguo; sin embargo, la estimación numérica de derivadas es un problema muy nuevo y necesario para la resolución numérica de ecuaciones diferenciales, que se han convertido en las instrumentos básicos para la descripción de todo tipo de sistemas.

El comportamiento de los métodos de integración en cuanto al error cometido es bastante bueno. Disminuir un poco el paso disminuye significativamente el error. Además, las fórmulas más comunes (los métodos compuestos) son bastantes estables, y errores en los datos, o en los redondeos, no se amplifican en el cálculo de las integrales. Sin embargo, las derivadas son mucho menos nobles cuando se las trata numéricamente, y puede suceder, como luego estudiaremos, que pequeñas variaciones en alguno de los elementos motiven grandes variaciones en el resultado.

Para Integración Numérica las referencias recomendadas son de nuevo el libro de Hämerlin y Hoffman [15] y el de Theodor [29]. Para Derivación Numérica, se recomiendan el libro de Theodor [29] y el de Yakowitz et al. [30].

5.1. Fórmulas de integración numérica

Sea la integral $I(f) = \int_a^b f(x)dx$. Deseamos encontrar un valor aproximado de $I(f)$ mediante una suma ponderada de valores de la función integrando en una colección de nodos contenidos en el intervalo $[a, b]$, de tal modo que esa ponderación no dependa de la función f en cuestión. Llamaremos fórmula de

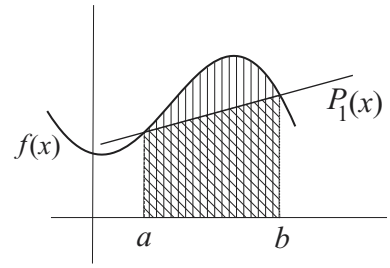


Figura 5.1: Regla del trapecio.

integración de $(n+1)$ puntos a una suma de este tipo.

$$I_n(f) = \sum_{k=0}^{n} A_k^n f(x_k) \tag{5.1}$$

donde los coeficientes A_k^n no deben depender de la función f.

Hemos visto ya que se puede aproximar una función f mediante un polinomio de interpolación P_f. Supongamos que únicamente conocemos la función f en los puntos $\{x_k\}$, $k = 0, n$, luego esas abscisas nos vienen impuestas y no podemos elegirlas. Podemos estimar la integral $I(f)$ integrando el único polinomio de interpolación de grado n correspondiente a la nube de puntos, $\{(x_k, f(x_k))\}$, $k = 0, n$.

$$I(f) = \int_a^b f(x)dx \approx \int_a^b P_f(x)dx = \sum_{k=0}^{n} \left(\int_a^b l_k(x)dx \right) f(x_k) \quad \text{con} \quad l_k(x) = \frac{\prod_{j=0, j\neq k}^{n}(x - x_j)}{\prod_{j=0, j\neq k}^{n}(x_k - x_j)}$$

A las fórmulas de este tipo se les llama de interpolación. Comparada con (5.1) se ha de tener que

$$A_k^n = \int_a^b l_k(x)dx \tag{5.2}$$

Ejemplo 5.1.1 *Si sustituimos la integral de la función por la de la recta que la interpola en los límites de la integral, tendremos la regla del trapecio, Figura 5.1.*

$$I_1(f) = \frac{h}{2} f(a) + \frac{h}{2} f(b), \; h = b - a$$

en la que los coeficientes A_0^1 y A_1^1 de la ecuación (5.1) valen $h/2$.

Definición 5.1.1 *Se define el error asociado a una fórmula de integración $R(f)$ por*

$$R(f) := I(f) - I_n(f) = \int_a^b f(x)dx - \sum_{k=0}^{n} A_k^n f(x_k)$$

En el caso de una fórmula de interpolación polinómica, el error de la integral es la integral de error (ver la sección 3.2.2).

$$R(f) = \int_a^b [f(x) - P_f(x)]dx = \int_a^b \frac{H(x)}{(n+1)!} f^{(n+1)}(\xi)dx$$

$$H(x) = (x - x_0)(x - x_1)\cdots(x - x_n), \quad a \leq min(x, x_0) < \xi < max(x, x_n) \leq b$$

Definición 5.1.2 *Una fórmula de cuadratura se dice exacta sobre un conjunto V si*

$$(\forall f \in V), \quad R(f) = 0$$

Teorema 5.1.1 *Una fórmula de cuadratura de $(n+1)$ puntos es exacta sobre el espacio vectorial P_n de los polinomios de grado menor o igual que n, ssi es del tipo de interpolación de $(n+1)$ puntos.*

Definición 5.1.3 *Se dice que una fórmula de cuadratura tiene grado de precisión n si la fórmula es exacta para el monomio x^k, $k = 0, n$, pero no es exacta para x^{n+1}.*

Método de los coeficientes indeterminados

El teorema 5.1.1, cuya demostración podemos encontrar en [29], nos proporciona otro modo de calcular los coeficientes, el método de los coeficientes indeterminados, que ilustramos en el problema 5.1.

5.1.1. Fórmulas de Newton-Cotes

Teorema 5.1.2 *Supongamos una subdivisión uniforme del intervalo $[a,b]$ de $n+1$ puntos, con $h=(b-a)/n$. Los nodos son $x_j = a + j \cdot h$, con $0 \leq j \leq n$, $x_0 = a$ y $x_n = b$. Con estas hipótesis, se tiene que la fórmula de integración de tipo interpolación asociada a estos nodos es:*

$$\int_a^b f(x)dx \approx (b-a)\sum_{j=0}^{n} B_j^n f(a+j\cdot h)$$

con

$$B_j^n = \frac{1}{b-a}\int_a^b l_j(x)dx = \frac{(-1)^{n-j}}{j!(n-j)!n}\int_0^n \prod_{k=0,k\neq j}^{n}(y-k)dy$$

Hemos hecho el cambio de variable $y=(x-a)/h$ para obtener la última igualdad, en la que se observa que los pesos B_j^n no dependen ni de a ni de b. Se puede demostrar además, que $B_j^n = B_{n-j}^n$.

Estas fórmulas se llaman las fórmulas cerradas de Newton-Cotes [1], que son muy usadas cuando la naturaleza del problema nos permite conocer los valores de la función sobre un soporte equiespaciado. En este caso, el valor aproximado de la integral se obtiene muy rápidamente, ya que esos coeficientes únicamente dependen del número de puntos que utilizamos para integrar y pueden ser tabulados (tabla 5.1).

Ejemplo 5.1.2 *Calculemos B_0^1.*

$$B_0^1 = B_1^1 = \frac{(-1)^1}{0!(1)!1}\int_0^1 (y-1)dy = \frac{1}{2}$$

Esta es la fórmula de Newton-Cotes de grado 1 que coincide con la fórmula del trapecio, que ya hemos visto en el ejemplo 5.1.1

$$\int_a^b f(x)dx \approx (b-a)\left[\frac{1}{2}f(a)+\frac{1}{2}f(b)\right]$$

La fórmula de Newton-Cotes de grado 2 se llama la fórmula de Simpson

$$\int_a^b f(x)dx \approx (b-a)\left[\frac{1}{6}f(a)+\frac{4}{6}f\left(\frac{a+b}{2}\right)+\frac{1}{6}f(b)\right]$$

En la tabla 5.1 tenemos los valores de B_j^n hasta n igual a 6.

5.1.2. Evaluación del error en las fórmulas de Newton-Cotes

Fórmula del trapecio

Se quiere evaluar

$$R(f) = \int_a^b f(x)dx - \left(\frac{b-a}{2}\right)[f(a)+f(b)] = \int_a^b \frac{(x-a)(x-b)}{2}f''(\xi)dx, \quad \xi=\xi(x)\in(a,b)$$

[1] Roger Cotes(1682-1716) fue un filósofo y matemático inglés. Con 26 años ya era profesor en Cambridge, y se dedicó entre 1709 y 1713 a revisar y corregir la segunda edición de los Principia Mathematica de Newton escribiendo de hecho su prefacio. Con Newton tuvo una relación muy estrecha, y fue uno de los más arduos defensores de sus teorías en la época. Las fórmulas de integración que llevan su nombre y el de Newton se deben más a Cotes. Murió muy joven, precipitando la decadencia de las matemáticas en las islas Británicas, que duraría hasta la mitad del XIX, decadencia que se origina en el seguidismo a los razonamientos de tipo geométrico propios de Newton frente a los de tipo analítico de Leibniz y los hermanos Bernouilli, con mayores posibilidades. Cotes, y los otros discípulos de Newton, como McLaurin, fueron los primeros que vieron el mundo con sus ojos, unos ojos en los que los problemas físicos (empíricos) se modelaban desde las matemáticas. En Newton todavía existía también la vertiente teológica milenaria. Dice Keynes que Newton consideraba el Universo como un criptograma trazado por el Todopoderoso.

n	B_0^n	B_1^n	B_2^n	B_3^n
1	1/2			
2	1/6	4/6		
3	1/8	3/8		
4	7/90	32/90	12/90	
5	19/288	75/288	50/288	
6	41/840	216/840	27/840	272/840

Cuadro 5.1: Coeficientes de las fórmulas de Newton-Cotes.

como $(x-a)(x-b)$ no cambia de signo en $[a,b]$, se puede usar una de las formulaciones del teorema del valor medio [2] del cálculo integral para escribir que

$$R(f) = \frac{f''(\alpha)}{2}\int_a^b (x-a)(x-b)dx, = -\frac{h^3}{12}f''(\alpha), \quad \alpha \in [a,b]$$

De donde se deduce la siguiente cota al error de la aproximación

$$|R(f)| \leq \frac{h^3}{12}\max_{x\in[a,b]}|f''(x)|$$

5.2. Métodos compuestos

5.2.1. General

La idea de base es aplicar a los subintervalos de una partición de $[a,b]$ una fórmula de Newton-Cotes de grado q, con q fijo y pequeño. Los métodos de Newton-Cotes tienen los mismos problemas de estabilidad que la interpolación mediante polinomios. A medida que aumenta el número de puntos, y por tanto el grado de los polinomios interpolantes, la convergencia no tiene por qué mejorar cuando la partición es equiespaciada (ver pág. 126 y siguientes). Este problema se reproduce con las fórmulas de Newton-Cotes y la solución vuelve a ser la misma que entonces: hacer Newton-Cotes de grado bajo en cada tramo o grupo de tramos y sumar para tener la integral total. Esto es equivalente a interpolar con polinomios a trozos, e integrar estos polinomios. Las fórmulas así obtenidas sí son estables.

5.2.2. Método compuesto de los trapecios

En el método compuesto de los trapecios, el número de puntos que se utiliza en cada subintervalo es 2, y por tanto el grado del polinomio de aproximación q es 1. Es equivalente a interpolar con un polinomio a trozos de grado 1 y clase 0 e integrar éste.

$$\begin{aligned}\int_a^b f(x)dx &= \sum_{k=0}^{n-1}\int_{a+k\cdot h}^{a+(k+1)h} f(x)dx \approx \sum_{k=0}^{n-1} h\left[\frac{1}{2}f(a+k\cdot h)+\frac{1}{2}f(a+(k+1)\cdot h)\right]\\ &= h\left[\frac{1}{2}f(a)+f(a+h)+f(a+2h)+\cdots+\frac{1}{2}f(b)\right]\end{aligned}$$

[2] Si $f \in C[a,b]$, g es integrable en $[a,b]$, y $g(x)$ no cambia de signo en $[a,b]$, entonces existe un número c en (a,b) en el que

$$\int_a^b f(x)g(x)dx = f(c)\int_a^b f(x)dx$$

Si $g(x)=1$, tenemos la formulación de este teorema que se estudia en los primeros cursos de Análisis

$$f(c) = \frac{1}{b-a}\int_a^b f(x)dx$$

5.2.3. Mayoración del error en el método compuesto de los trapecios

Se ha visto ya que en el caso del método de los trapecios:

$$|R(f)| = \left| \int_a^b f(x)dx - \left(\frac{b-a}{2}\right)[f(a)+f(b)] \right| \leq \frac{h^3}{12} \max_{x\in[a,b]} |f''(x)|$$

Por tanto, en el caso de los métodos compuestos:

$$\begin{aligned} |R(f)| &= \left| \int_a^b f(x)dx - h\left[\frac{1}{2}f(a) + f(a+h) + \cdots + \frac{1}{2}f(b)\right] \right| \\ &\leq \frac{h^3}{12}\left[\max_{x\in[a,a+h]} |f''(x)| + \cdots + \max_{x\in[a+(n-1)h,b]} |f''(x)| \right] \\ &\leq n\frac{h^3}{12} \max_{x\in[a,b]} |f''(x)| = \frac{(b-a)^3}{12n^2} \max_{x\in[a,b]} |f''(x)| = \frac{b-a}{12} \max_{x\in[a,b]} |f''(x)|h^2 = O(h^2) \end{aligned}$$

5.3. Fórmulas de Gauss

5.3.1. General

En cierto tipo de experimentos podemos fijar el valor de las abscisas de la función que pretendemos conocer. En estos casos, Gauss nos proporciona un criterio para elegir estas abscisas que aumenta notablemente el grado de precisión de las fórmulas de integración de tipo interpolación. La idea consiste en aproximar $I(f)$ por la expresión ya referida 5.1

$$I_n(f) = \sum_{k=0}^{n} A_k^n f(x_k^n)$$

considerando ahora como desconocidos no sólo A_k^n sino también x_k^n. Veamos cómo lo conseguimos.

Se observa en la fórmula anterior que el número de parámetros libres es $2n+2$. Trataremos, por tanto, de que nuestra fórmula sea exacta en P_{2n+1} espacio vectorial de polinomios de coeficientes reales de grado menor o igual que $2n+1$. Exigiremos entonces que

$$\int_a^b p(x)dx = \sum_{k=0}^{n} A_k^n p(x_k^n) \quad (\forall p \in P_{2n+1}) \tag{5.3}$$

Sea $p^* \in P_n$ tal que $p^*(x_k^n) = p(x_k^n)$, $k = 0, ..., n$, o sea, el polinomio de grado n de interpolación de la nube de puntos $(x_k^n, p(x_k^n))$, $k = 0, ..., n$. Como p también interpola esa misma nube de puntos, lo podremos escribir de la siguiente forma:

$$p(x) = p^*(x) + (x - x_0^n)(x - x_1^n)\cdots(x - x_n^n)\cdot q(x)$$

con $q \in P_n$.

Si integramos,

$$\int_a^b p(x)dx = \int_a^b [p^*(x) + (x - x_0^n)\cdots(x - x_n^n)\cdot q(x)]dx = \int_a^b p^*(x)dx + \int_a^b (x - x_0^n)\cdots(x - x_n^n)\cdot q(x)dx$$

Pero la integral de p y de p^* son iguales al ser iguales en los nodos, y por tanto, se tiene que cumplir que

$$\int_a^b (x - x_0^n)(x - x_1^n)\cdots(x - x_n^n)\cdot q(x)dx = 0 \quad \forall q \in P_n \tag{5.4}$$

En el problema 5.2 presentamos un ejemplo sencillo y detallado del cálculo de los coeficientes de una fórmula de este tipo apoyándonos en estos razonamientos, pero hay otro camino interesante que ahora pasamos a explicar.

5.3.2. Relación entre las fórmulas de Gauss y los polinomios de Legendre

Si hacemos $[a,b]=[-1,1]$ en la ecuación 5.4, podemos interpretarla como la busca de una serie de nodos (x_k^n), $k=0,n$ tales que el polinomio $(x-x_0^n)(x-x_1^n)\cdots(x-x_n^n)$ de grado $n+1$ sea ortogonal, según el producto escalar $\langle p,q\rangle=\int_{-1}^{1}p(x)q(x)dx$, a todos los polinomios q de P_n.

Supongamos que $p(x)$ barre la base de Legendre, o sea, que $p(x)=\hat{L}_j(x)$, $j=0,...,n$. Escribiendo $q(x)$ en esa base ortonormal se tiene entonces

$$q(x)=\sum_{i=0}^{n+1}\alpha_i\hat{L}_i(x) \quad \Rightarrow \quad 0=\langle p,q\rangle=\langle\hat{L}_j(x),\sum_{i=0}^{n+1}\alpha_i\hat{L}_i(x)\rangle=\alpha_j$$

de donde $\alpha_j=0$, $j=0,...,n$, y $q(x)=\alpha_{n+1}\hat{L}_{n+1}(x)$, es decir, q es proporcional al polinomio de Legendre (ver 4.5.5) de grado $n+1$, y tiene, en consecuencia, sus mismas raíces.

Para que $(x-x_0^n)(x-x_1^n)\cdots(x-x_n^n)$ sea proporcional a $\hat{L}_{n+1}(x)$, debemos escoger como valores (x_k^n), $k=0,...,n$, de las abscisas de los nodos los ceros de $\hat{L}_{n+1}(x)$, polinomio de Legendre de grado $n+1$.

Una vez que conocemos cuáles deben ser los nodos, hallamos el valor de la familia de coeficientes (A_k^n), $k=0,..,n$ bien mediante la ecuación (5.2), o bien mediante el método de los coeficientes indeterminados. Si el intervalo de integración no es $[-1,1]$, se hace previamente el cambio de variable

$$t=\frac{2(x-a)}{b-a}-1$$

que transforma $[a,b]$ en $[-1,1]$.

5.4. Integración multidimensional

Las dificultades que se presentan al tratar de evaluar numéricamente integrales en espacios de mayor dimensión, provienen fundamentalmente de la complejidad geométrica de los recintos de integración. Así, mientras que no hay muchos problemas abiertos en integrales en una variable, el cálculo numérico de integrales múltiples es un área en el que se investiga todavía, dado que son fundamentales para los métodos numéricos de resolución de ecuaciones en derivadas parciales, sobre todo en el método de los elementos finitos.

Las técnicas de integración numérica más sencillas para integrales dobles o triples se basan en la sustitución de la función a integrar por su polinomio en varias variables (ver 3.6). Si estamos en 2D y el recinto es rectangular, la aplicación es directa:

$$\int_a^b\int_c^d f(x,y)dxdy\approx\int_a^b\int_c^d\sum_{\substack{0\le i\le m\\0\le j\le n}}f(x_i,y_j)l_i(x)l_j(y)dxdy$$

Como los límites de integración son constantes, podemos escribir

$$\int_a^b\int_c^d f(x,y)dxdy\approx\sum_{\substack{0\le i\le m\\0\le j\le n}}f(x_i,y_j)\int_a^b l_i(x)dx\int_c^d l_j(y)dy$$

Nos remitimos al problema 5.6 para una explicación más detallada de este método y otros similares.

No todos los recintos de integración 2D son rectangulares (aunque podríamos .[en]gañarrellenando con ceros hasta conseguir un recinto al que se pueda aplicar esta técnica tan sencilla). Lo que sí es cierto es que cualquier recinto R en el plano puede ser triangulado. Si llamamos T_i a cada uno de esos n triángulos.

$$\int\int_R f(x,y)dxdy=\sum_{i=1}^{n}\int\int_{T_i}f(x,y)dxdy$$

el problema es ahora estimar cada una de estas integrales extendidas a estos triángulos, para lo cual remitimos al Capítulo 5 de Theodor[29], ya que escapa a los contenidos de este libro.

5.5. Derivación numérica

Sea f una función real de variable real de clase $\mathcal{C}^1$ al menos. Se quiere calcular una aproximación al número derivada $f'(x)$.

$$f'(x) = \lim_{h\to 0} \frac{f(x+h)-f(x)}{h} = \lim_{h\to 0} \frac{f(x)-f(x-h)}{h}$$

Si tomamos un paso h pequeño, podemos estimar $f'(x)$ con la expresión

$$f'(x) \approx \frac{f(x+h)-f(x)}{h} \tag{5.5}$$

Esta aproximación hace intervenir los valores de la función en los puntos x y $x+h$. De modo más general, podríamos pensar, igual que hacíamos al integrar numéricamente, en sustituir la función f por algún polinomio que la interpole en el entorno del punto donde queremos evaluar la derivada y estimar la derivada con la correspondiente a este polinomio. Estas técnicas de derivación basadas en interpolación con polinomios son las que estudiaremos aquí.

5.5.1. Fórmula de dos puntos

Sea f una función de clase $\mathcal{C}^1$ en un entorno cerrado $E(x_0)$ del punto x_0 en el que se desea estimar la derivada de f. Sea x_1 tal que $[x_0, x_1] \subset E(x_0)$.

Construyamos el polinomio de interpolación de f asociado a los puntos x_0 y x_1, ver Figura 5.2.

$$P_1(x) = f(x_0)\frac{x-x_1}{x_0-x_1} + f(x_1)\frac{x-x_0}{x_1-x_0}$$

Derivando este polinomio y evaluando la derivada en x_0 obtendremos la aproximación a la derivada buscada.

$$f'(x_0) \approx P_1'(x_0) = \frac{f(x_1)-f(x_0)}{x_1-x_0} = \frac{f(x_0+h)-f(x_0)}{h}, \quad h = x_1 - x_0 \tag{5.6}$$

que es precisamente la estimación inicial (5.5) fórmula de dos puntos "mirando hacia adelante" o progresiva[3]. Un ejemplo básico de aplicación de esta fórmula lo encontramos en el problema 5.9.

Existe igualmente una versión simétrica de esta fórmula tomando el segundo punto a la izquierda de x_0 fórmula de dos puntos "mirando hacia atrás" o regresiva.

$$f'(x_0) \approx \frac{f(x_0)-f(x_0-h)}{h} \tag{5.7}$$

5.5.2. Acotación del error en la fórmula de dos puntos

Es importante tener una idea del error que estamos cometiendo al hacer estas estimaciones. Para calcularlo, derivamos el error de la interpolación de Lagrange (pág. 126). Veámoslo en nuestro caso con un polinomio definido por dos puntos ($n = 1$):

$$f(x) - P_1(x) = \frac{(x-x_0)(x-x_1)}{2} f''(\xi(x))$$

y derivando esta fórmula:

$$f'(x) - P_1'(x) = \frac{1}{2}(x-x_0)(x-x_1)f'''(\xi(x))\xi'(x) + \frac{1}{2}(x-x_1)f''(\xi(x)) + \frac{1}{2}(x-x_0)f''(\xi(x))$$

Hay problemas para estimar el error en puntos distintos de x_0 o de x_1, pues no conocemos el valor de $\xi'(x)$. Sin embargo, podemos estimar fácilmente el error en x_0 y en x_1 sustituyendo en la expresión anterior.

$$f'(x_0) - P_1'(x_0) = \frac{(x_0-x_1)f''(\xi(x_0))}{2} = -\frac{f''(\xi(x_0))}{2}h, \quad x_0 < \xi(x_0) < x_1$$

[3] Ya que se ha elegido el punto x_1 a la derecha del x_0 en el sentido en que progresa o se incrementa x. La denominación "mirando hacia adelante" recuerda la progresión en el tiempo.

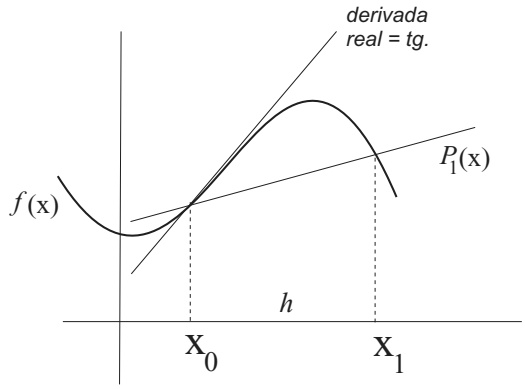


Figura 5.2: Fórmula de dos puntos.

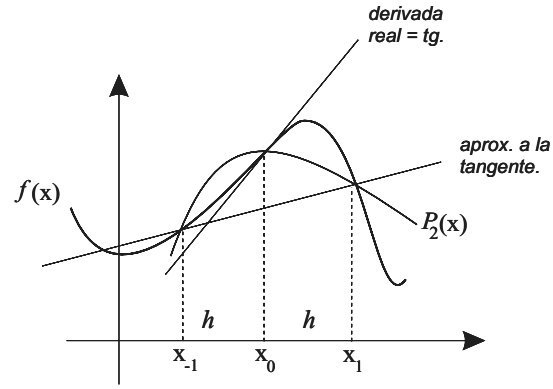


Figura 5.3: Fórmula de tres puntos.

$$|f'(x_0) - P_1'(x_0)| \leq \frac{\|f''\|_\infty}{2} h \tag{5.8}$$

La norma del máximo la estamos tomando en $E(x_0)$. Vemos que el error es $O(h)$, o sea, el error tiende a 0 con h. La aproximación es bastante mala. En la fórmula de integración del trapecio, que es la equivalente a esta fórmula de dos puntos en la integración numérica, el error era del orden de h^3, y el error en la interpolación con una recta es $O(h^2)$. Por tanto, la integral disminuye el error de la interpolación polinomial, subiendo el orden, y la derivada lo amplifica, bajando el orden.

Un detalle adicional de interés es que para conseguir una estimación del error hemos necesitado suponer que la función f no sólo debe ser derivable, sino que debe tener segunda derivada continua. Un ejemplo básico de aplicación de esta acotación lo encontramos en el problema 5.9.

5.5.3. Fórmula de tres puntos

Con el mismo escenario que en 5.5.1, tomamos dos puntos $x_1, x_{-1} \in E(x_0)$ a ambos lados del punto x_0 en el que queremos estimar la derivada de f y a igual distancia $x_1 = x_0 + h$ y $x_{-1} = x_0 - h$ siendo h el paso.

Construyamos la parábola de interpolación de f asociada a los puntos x_0, x_1 y x_{-1}, ver Figura 5.3.

$$P_2(x) = f(x_{-1})l_{-1}(x) + f(x_0)l_0(x) + f(x_1)l_1(x) \tag{5.9}$$

$$P_2(x) = f(x_{-1})\frac{(x-x_0)(x-x_1)}{(x_{-1}-x_0)(x_{-1}-x_1)} + f(x_0)\frac{(x-x_{-1})(x-x_1)}{(x_0-x_{-1})(x_0-x_1)} + f(x_1)\frac{(x-x_{-1})(x-x_0)}{(x_1-x_{-1})(x_1-x_0)}$$

Si derivamos y particularizamos en x_0:

$$f'(x_0) \approx P_2'(x_0) = \frac{f(x_0+h) - f(x_0-h)}{2h} \tag{5.10}$$

Esta fórmula se llama de tres puntos centrada, pues se construye la parábola con tres puntos, y se evalúa la derivada en el punto medio. Es curioso pensar que precisamente en el cálculo de la derivada no influye el valor en el propio punto en el que se realiza la estimación. Un ejemplo básico de aplicación de esta fórmula lo encontramos en el problema 5.9.

Por otro lado, del mismo modo que hay fórmulas centradas, también hay fórmulas mirando hacia adelante, y hacia atrás. Si evaluamos la derivada en $x_{-1} = x_0 - h$, tendremos una fórmula mirando hacia adelante, y si lo hacemos en $x_1 = x_0 + h$, mirando hacia atrás:

$$f'(x_0-h) \approx P_2'(x_0-h) = \frac{-3f(x_0-h) + 4f(x_0) - f(x_0+h)}{2h} \tag{5.11}$$

$$f'(x_0+h) \approx P_2'(x_0+h) = \frac{f(x_0-h) - 4f(x_0) + 3f(x_0+h)}{2h} \tag{5.12}$$

5.5.4. Acotación del error en la fórmula de tres puntos

Para calcular el error, usamos otra vez el error de la interpolación.

$$f(x) - P_2(x) = \frac{(x - x_{-1})(x - x_0)(x - x_1)}{3!} f'''(\xi(x))$$

Derivándolo, y particularizándolo para x_0:

$$f'(x_0) - P_2'(x_0) = -\frac{f'''(\xi(x_0))}{6} h^2 = O(h^2) \tag{5.13}$$

Por tanto, el error es una potencia de h de orden 2, lo que quiere decir que hemos mejorado notablemente la aproximación frente a la fórmula de dos puntos. Destaquemos que la función f debe ser de clase $\mathcal{C}^3$ para poder realizar esta acotación.

5.5.5. Fórmula de tres puntos centrada mediante desarrollo de Taylor

Se pueden deducir estas fórmulas y otras más elaboradas mediante el desarrollo en serie de Taylor de la función f en torno a x_0. Veámoslo con la fórmula centrada:

$$\begin{aligned} f(x_0 + h) &= f(x_0) + hf'(x_0) + \frac{h^2}{2} f''(x_0) + \frac{h^3}{6} f'''(\xi_1) \\ f(x_0 - h) &= f(x_0) - hf'(x_0) + \frac{h^2}{2} f''(x_0) - \frac{h^3}{6} f'''(\xi_2) \end{aligned}$$

Restando estas dos ecuaciones, tenemos que:

$$f'(x_0) = \frac{f(x_0 + h) - f(x_0 - h)}{2h} + \frac{h^2}{12}(f'''(\xi_1) + f'''(\xi_2)) \tag{5.14}$$

Analicemos con un poco de detalle el término de error

$$\frac{h^2}{12}(f'''(\xi_1) + f'''(\xi_2))$$

Si f''' es una función continua, habrá un punto entre ξ_1 y ξ_2 en el que f''' valga la media de $f'''(\xi_1)$ y $f'''(\xi_2)$. Llamemos ξ a ese punto:

$$f'''(\xi) = \frac{f'''(\xi_1) + f'''(\xi_2)}{2}$$

Entrando en 5.14 con este valor obtenemos el mismo término de error que habíamos obtenido en la expresión 5.13.

Ejercicio 5.5.1 *Utilizar el desarrollo de Taylor para cálcular la fórmula de derivación de dos puntos y su término de error correspondiente.*

5.5.6. Fórmula de tres puntos para estimar la derivada segunda

Podemos usar la interpolación parabólica de tres puntos para estimar la derivada segunda, que se usa a menudo en ecuaciones diferenciales, como la de Laplace. Así, si volvemos a derivar $P_2(x)$ en la ecuación (5.9):

$$f''(x_0) \approx P_2''(x_0) = \frac{f(x_0 + h) - 2f(x_0) + f(x_0 - h)}{h^2} \tag{5.15}$$

Ejercicio 5.5.2 *Utilizar la fórmula 5.15 para estimar la derivada segunda de la función* $\sin(1/x)$ *en el punto* $x_0 = 0.2$. *Escribir el código Matlab con el que visualizar la función* $error(h)$.

Ejercicio 5.5.3 *Calcular el término de error correspondiente a la fórmula de derivación 5.15.*

Ejercicio 5.5.4 *Utilizar el desarrollo en serie de Taylor para obtener la fórmula 5.15 para la derivada segunda y deducir de este mismo modo su término de error.*

5.6. Estabilidad

El concepto de estabilidad de un método numérico está asociado a su comportamiento frente a la presencia de errores en los datos de partida. En concreto, las técnicas de diferenciación numérica que hemos estudiado son muy sensibles a los errores de redondeo y de los datos originales. Para ilustrar esta idea, usemos la aproximación de dos puntos a la derivada:

$$f'(x_0) \approx \frac{f(x_0+h)-f(x_0)}{h} = \frac{f_1-f_0}{h}$$

con $f_1 = f(x_0+h)$ y $f_0 = f(x_0)$. Supongamos que $\overline{f_0}$ y $\overline{f_1}$ son los valores de trabajo de f_0 y f_1, los cuales vienen afectados de unos errores que estimamos mediante las desigualdades

$$\left|\overline{f_0} - f_0\right| \leq \delta, \qquad \left|\overline{f_1} - f_1\right| \leq \delta$$

Estimaciones que pueden provenir, bien de que esos valores se hayan obtenido de un experimento en el que los aparatos de medida tengan esa precisión, o bien de un cálculo previo sujeto a unos determinados errores conocidos.

De cualquier modo, en el cálculo que hacemos para estimar la derivada sólo conocemos estos valores perturbados

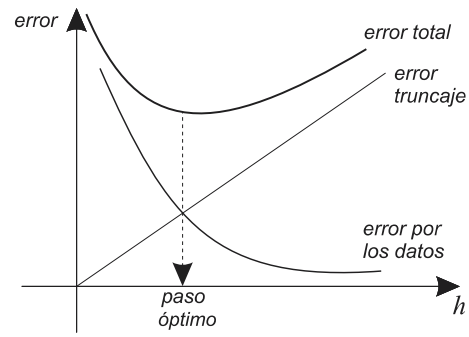


Figura 5.4: Selección del paso óptimo.

$$f'(x_0) \approx \frac{\overline{f_1}-\overline{f_0}}{h}$$

Con lo cual, el error, y sus acotaciones quedan

$$\begin{aligned}\left|f'(x_0) - \frac{\overline{f_1}-\overline{f_0}}{h}\right| &\leq \left|f'(x_0) - \frac{f_1-f_0}{h}\right| + \left|\frac{f_1-f_0}{h} - \frac{\overline{f_1}-\overline{f_0}}{h}\right| \\ &\leq \frac{\| f'' \|_\infty}{2}h + \frac{\left|f_0-\overline{f_0}\right| + \left|f_1-\overline{f_1}\right|}{h} \leq \frac{\| f'' \|_\infty}{2}h + \frac{2\delta}{h}\end{aligned}$$

Por tanto,

$$\left|f'(x_0) - \frac{\overline{f_1}-\overline{f_0}}{h}\right| \leq \frac{\| f'' \|_\infty}{2}h + \frac{2\delta}{h}$$

con

$$\| f'' \|_\infty = \max_{x\in[x_0,x_0+h]} |f''(x_0)|$$

Para valores grandes de h, el término en h domina (error del métodо). Si h es pequeño, el error que puedan arrastrar los datos, que corresponde al término $1/h$ es el más importante.

Hay que escoger el paso de tal modo que el error total sea mínimo (ver Figura 5.4). Una consecuencia de este análisis es que no por mucho disminuir el paso h vamos a disminuir también el error, dado que hay una componente del error que crece asintóticamente al disminuir h. Para aplicaciones prácticas, nos remitimos a los problemas 5.12 y 5.13.

Ejercicio 5.6.1 *Estimar el paso óptimo para calcular la derivada de la función* $\cos(10t)$ *para* $t = 0.1$ *si se maneja una calculadora de 3 dígitos, y se utiliza un operador de 2 puntos para estimar la derivada. Construir la curva error-paso para un rango de pasos mayores y menores que el óptimo dibujándola con Matlab.*

5.7. Derivadas parciales

Consideremos, por ejemplo, el problema de construir una aproximación discreta al laplaciano bidimensional de una función u en un punto cualquiera (x_0, y_0).

$$\Delta u\,(x_0, y_0) = \frac{\partial^2 u}{\partial x^2}\,(x_0, y_0) + \frac{\partial^2 u}{\partial y^2}\,(x_0, y_0)$$

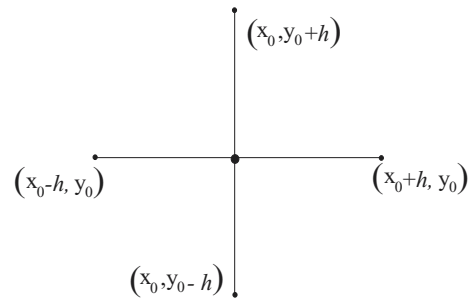


Figura 5.5: Discretización de cinco puntos.

Utilicemos un conjunto de 5 puntos para representar u en un entorno de (x_0, y_0) (ver Figura 7.17), y supongamos que el espaciado en la dirección x es el mismo que el espaciado en la dirección y. Para hacer la discretización del operador u en este conjunto, tenemos que pensar en el significado geométrico de las derivadas parciales.

La derivada parcial de una función respecto a una variable da una idea de lo que varía esa función cuando las otras variables permanecen constantes. Es entonces una función de una sola variable y la derivada parcial hay que interpretarla como la derivada de esa función de una sola variable. La estimación queda entonces, tanto para la x como para la y:

$$\frac{\partial^2 u}{\partial x^2}\,(x_0, y_0) = \frac{u\,(x_0 + h, y_0) - 2u\,(x_0, y_0) + u\,(x_0 - h, y_0)}{h^2} + O(h^2)$$

$$\frac{\partial^2 u}{\partial y^2}\,(x_0, y_0) = \frac{u\,(x_0, y_0 + h) - 2u\,(x_0, y_0) + u\,(x_0, y_0 - h)}{h^2} + O(h^2)$$

Sumando ambas expresiones tendremos que:

$$\Delta u\,(x_0, y_0) \approx \frac{u\,(x_0 + h, y_0) + u\,(x_0 - h, y_0) + u\,(x_0, y_0 + h) + u\,(x_0, y_0 - h) - 4u\,(x_0, y_0)}{h^2}$$

Ejercicio 5.7.1 *Se tiene la función de dos variables definida por los puntos* $u_{ij} = u(x_i, y_j)$ *con* $i, j = 0, 2$*:*

$x_i \backslash y_j$	-2	-1	0
-1	2.0000	2.6458	2.8284
0	2.2361	2.8284	3.0000
1	2.0000	2.6458	2.8284

Se pide estimar las derivadas parciales primeras y segundas en el punto $(0, -1)$*, utilizando operadores de dos y tres puntos. Comparar los resultados con los reales correspondientes a la función* $u(x, y) = |\sqrt{9 - x^2 - y^2}|$*, de la que se han tomado los valores.*

PROBLEMAS

PROBLEMA 5.1 *Método de los coeficientes indeterminados.*

Calcular los coeficientes que intervienen en la fórmula de cuadratura de tipo interpolación siguiente:

$$\int_{-1}^{1} f(x)dx \approx A_0^1 f(-1) + A_1^1 f(1)$$

Solución:
La fórmula de cuadratura es de tipo interpolación de dos puntos, siendo por tanto su grado de precisión uno; debe ser exacta para los polinomios $P_0(x) = 1$ y $P_1(x) = x$. Obliguemos a que se cumplan estas condiciones:

$$\int_{-1}^{1} 1 \cdot dx = A_0^1 + A_1^1 \,, \quad \int_{-1}^{1} x \cdot dx = A_0^1 \cdot (-1) + A_1^1 \cdot 1$$

De donde obtenemos el sistema lineal:

$$\begin{array}{rl} A_0^1 + A_1^1 & = 2 \\ -A_0^1 + A_1^1 & = 0 \end{array}$$

cuya solución es $A_0^1 = A_1^1 = 1$.

PROBLEMA 5.2 *Integración gaussiana.*

Encontrar una fórmula del tipo:

$$\int_{3.2}^{4.3} f(x)dx = A_0 f(x_0) + A_1 f(x_1)$$

cuyo grado de precisión sea máximo.

Solución:
Es un problema de integración gaussiana que se reduce a encontrar dos nodos x_0, x_1 en los que se verifique la ecuación (5.4), página 233, con $n = 1$.
Planteémoslo, jugando con $q(x) = 1$, y con $q(x) = x$

$$\begin{array}{l} \displaystyle\int_{3.2}^{4.3} (x - x_0)(x - x_1)dx = 0 \\ \displaystyle\int_{3.2}^{4.3} (x - x_0)(x - x_1)xdx = 0 \end{array} \quad \Rightarrow \quad \begin{cases} 15.5797 - 4.1250(x_0 + x_1) + 1.1000x_0x_1 = 0 \\ 59.2556 - 15.5797(x_0 + x_1) + 4.1250x_0x_1 = 0 \end{cases}$$

Como vemos, con dos nodos tenemos un sistema lineal de dos ecuaciones con dos incógnitas. A medida que subimos el número de nodos el sistema lineal se va complicando. Cuando es de dos nodos, se suele resolver planteando que esos nodos son las raíces de la ecuación de segundo grado $x^2 + Mx + N = 0$, con $M = -(x_0 + x_1)$ y $N = x_0x_1$. De este modo, el sistema no lineal anterior se convierte en el sistema lineal siguiente

$$\begin{cases} 4.125M + 1.1000N = -15.5797 \\ 15.5797M + 4.1250N = -59.2556 \end{cases}$$

cuya solución es $M = -7.4964$, $N = 13.9481$. Las raíces de $x^2 + Mx + N = 0$ son $x_0 = 3.4306$, $x_1 = 4.0658$. Para encontrar los coeficientes A_0, A_1, recurrimos al método de los coeficientes indeterminados, imponiendo que la fórmula sea exacta para $p(x) = 1$ y para $p(x) = x$.

$$\begin{array}{l} \displaystyle\int_{3.2}^{4.3} dx = A_0 + A_1 \\ \displaystyle\int_{3.2}^{4.3} xdx = A_0 3.4306 + A_1 4.0658 \end{array} \quad \Rightarrow \quad \begin{cases} A_0 + A_1 = 1.1000 \\ 3.4306A_0 + 4.0658A_1 = 4.1250 \end{cases}$$

Es importante pensar que hemos encontrado una fórmula que exige evaluar una función en dos puntos pero que es exacta para polinomios de grado tres, que en su esencia son objetos con cuatro grados de libertad. De hecho, se sugiere comprobar que la fórmula es exacta, por ejemplo, para x^3.
Se deja como ejercicio resolver este problema realizando el cambio de variable para llevar la integral al interval $[-1, 1]$, y utilizar los polinomios de Legendre para encontrar los puntos x_0 y x_1.

PROBLEMA 5.3 *Método de Newton-Cotes de grado 0.*

Sea $f \in C^1[a, b]$. Sea $\Omega = \{t_0 = a, t_1, \ldots, t_n = b\}$ una partición cualquiera estrictamente creciente del intervalo $[a, b]$. Sea $S_0(\Omega)$ el espacio vectorial de los splines de grado 0, no continuos en los nodos, asociados a la partición Ω.

1. Calcular la dimensión de $S_0(\Omega)$.
2. Se considera la base B de $S_0(\Omega)$ formada por B-splines (ver pág. 241). Calcular el spline de interpolación de f, el que la interpola en todos los nodos menos en el último, escribiéndolo en la base B. Hacer un esquema del mismo.
3. Acotar el error que se comete en esta interpolación.
4. Evaluar la integral de la función f, aproximándola mediante la de s.
5. Acotar el error.
6. Aplicando esta cota y suponiendo que la partición es equiespaciada, encontrar n tal que al evaluar la integral
$$I = \int_{-3}^{3} \cos(\cosh(t))dt$$
el error sea menor que 0.001.
7. Escribir unas líneas Matlab para evaluar la integral con la subdivisión que sugiere el apartado anterior.

Solución:

1. Cada tramo es un polinomio de grado 0, el cual se define mediante un único valor. Como tenemos n tramos, tendremos que definir n valores. No hay ninguna restricción en los enganches y por tanto, la dimensión es n.
2. Sea s el spline buscado. La dimensión es n y tenemos $n + 1$ nodos. Por tanto s no va a interpolar a f en todos los nodos. Dada la definición de los splines de base como (pág. 241):
$$B_i^0(t) = \begin{cases} 0, & t \notin [t_i, t_{i+1}) \\ 1 & t \in [t_i, t_{i+1}) \end{cases}$$
lo más sencillo es elegir como punto de interpolación de cada tramo el primero, y prolongar el spline en el último tramo. O sea:
$$s(t) = \begin{cases} f(t_i), & t \in [t_i, t_{i+1}) \quad i = 0, n-1 \\ f(t_{n-1}), & t = t_n \end{cases}$$
Vemos un ejemplo correspondiente a la función $f(t) = \cos(\cosh(t))$ en la Figura 5.6.
3. En cada tramo estamos construyendo un polinomio de interpolación de Lagrange de grado cero; por tanto, tendremos que acotar en cada tramo el error con la expresión correspondiente al error en la interpolación de Lagrange (pág. 126). Así, si $t \in [t_i, t_{i+1})$
$$|f(t) - s(t)| = \frac{|f'(\xi(t))||t - t_i|}{1} \leq \max_{t \in [t_i, t_{i+1}]} |f'(t)||t_{i+1} - t_i|$$

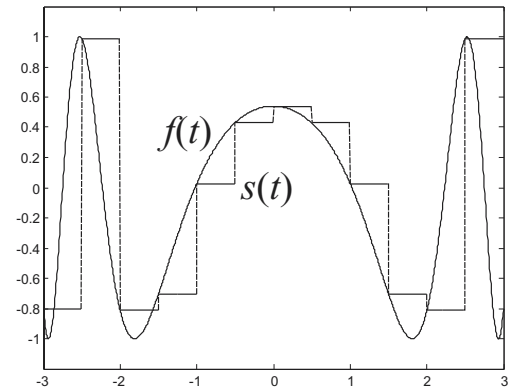


Figura 5.6: Ejemplo de interpolación con splines de grado 0.

En general, si $t \in [a,b]$, tenemos que

$$\begin{aligned}
\|f-s\|_\infty &= \max_{t\in[a,b]} |f(t)-s(t)| \leq \max_{i=0,n-1} \left(\max_{t\in[t_i,t_{i+1}]} |f'(t)||t_{i+1}-t_i| \right) \\
&\leq \max_{i=0,n-1} \left(\max_{t\in[t_i,t_{i+1}]} |f'(t)| \right) \max_{i=0,n-1} |t_{i+1}-t_i| \\
&= \max_{t\in[a,b]} |f'(t)| \max_{i=0,n-1} |t_{i+1}-t_i| = \|f'\|_\infty \max_{i=0,n-1} |t_{i+1}-t_i|
\end{aligned}$$

Si la partición es equiespaciada, y la diferencia entre dos nodos consecutivos es h, tendremos que

$$\|f-s\|_\infty \leq \|f'\|_\infty h$$

Por tanto, el error tiende a 0 con h, lo que significa que este método de interpolación es de orden 1.

4.

$$I = \int_a^b f(t)dt \approx \int_a^b s(t)dt = \sum_{i=0}^{n-1} \int_{t_i}^{t_{i+1}} s(t)dt = \sum_{i=0}^{n-1} \int_{t_i}^{t_{i+1}} f(t_i)dt = \sum_{i=0}^{n-1} f(t_i)(t_{i+1}-t_i)$$

Si la partición es equiespaciada, tenemos una fórmula compuesta de Newton-Cotes de grado 0, y también una de las reglas del rectángulo.

$$I \approx h \sum_{i=0}^{n-1} f(t_i)$$

Hay otras fórmulas compuestas de Newton-Cotes de grado 0, dependiendo del punto de cada intervalo que utilicemos para interpolar. Si fuese el punto medio, tendríamos

$$I \approx h \sum_{i=0}^{n-1} f\left(\frac{t_i+t_{i+1}}{2}\right)$$

5.

$$\begin{aligned} E &= \left|\int_a^b f(t)dt - \int_a^b s(t)dt\right| = \left|\int_a^b (f(t)-s(t))dt\right| \le \int_a^b |f(t)-s(t)|dt \\ &= \sum_{i=0}^{n-1}\int_{t_i}^{t_{i+1}} |f(t)-s(t)|dt = \sum_{i=0}^{n-1}\int_{t_i}^{t_{i+1}} |f'(\xi(t))||t-t_i|dt \\ &= \sum_{i=0}^{n-1} \max_{t\in[t_i,t_{i+1}]} |f'(t)| \int_{t_i}^{t_{i+1}} (t-t_i)dt = \sum_{i=0}^{n-1} \max_{t\in[t_i,t_{i+1}]} |f'(t)| \frac{(t_{i+1}-t_i)^2}{2} \\ &\le \max_{t\in[a,b]} |f'(t)| \sum_{i=0}^{n-1} \frac{(t_{i+1}-t_i)^2}{2} = \|f'\|_\infty \sum_{i=0}^{n-1} \frac{(t_{i+1}-t_i)^2}{2} \end{aligned}$$

Si la partición es equiespaciada

$$E \le \|f'\|_\infty \sum_{i=0}^{n-1} \frac{h^2}{2} = \|f'\|_\infty \frac{nh^2}{2} = \|f'\|_\infty \frac{(b-a)h}{2}$$

Y el método es también de orden 1.

6. Se trata de encontrar h tal que:

$$E \le \|f'\|_\infty \frac{(3-(-3))h}{2} < 0.001$$

Hallemos una mayorante de la norma de la derivada

$$|f'(t)| = |\sin(\cosh(t))\sinh(t)| \le \sinh(3) = 10.0179$$

Entrando en la desigualdad de arriba

$$E \le 3h\|f'\|_\infty \le 3h10.0179 = 30.0536h < 0.001$$

De donde deducimos que $h < 3.3274\ 10^{-5}$. Eso significa que el número de tramos del partición ha de ser:

$$n \ge \frac{6}{3.3274\ 10^{-5}} = 180320$$

7. Obtenemos el valor $I = -0.3796$ con las siguientes líneas Matlab:

```
n=180320;
h=6/n;
t=-3:h:3-h;
f=cos(cosh(t));
I=h*sum(f)
```

Al definir el rango, hacemos *t=-3:h:3-h;*. Al restar h al valor final quitamos el último punto que no influye en la integral, tal como la hemos definido. Podemos comparar el valor I obtenido con el correspondiente al problema 5.4. En este problema, en la Figura 5.11, se representa esa integral como función de los límites. Si entramos con el valor 3 en esa curva vemos que el valor que hemos obtenido se corresponde con el de la misma.

Un ejercicio interesante que sugiere este problema que acabamos de hacer consiste en analizar las diferencias que surgen cuando utilizamos el punto medio para definir el spline de grado 0. O sea, cuando hacemos:

$$s(t) = \begin{cases} f\left(\frac{t_i+t_{i+1}}{2}\right), & t \in [t_i, t_{i+1}) \quad i = 0, n-1 \\ f\left(\frac{t_{n-1}+t_n}{2}\right), & t = t_n \end{cases}$$

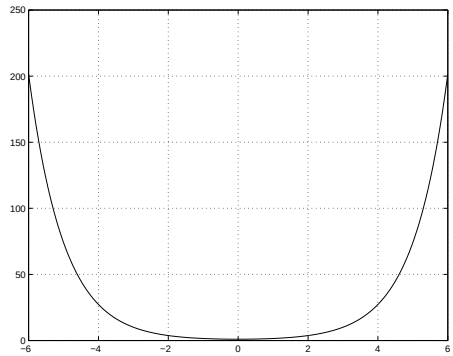


Figura 5.7: Fase.

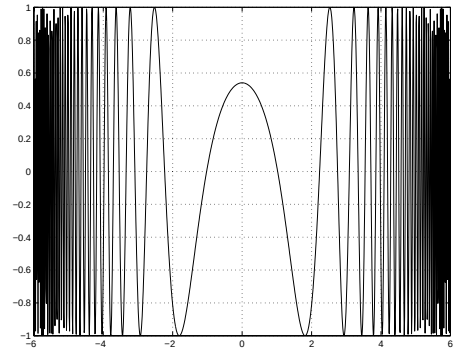


Figura 5.8: Parte real de la función a integrar.

PROBLEMA 5.4 *Método de la fase estacionaria.*

Se considera la integral:

$$I = \int_{-\infty}^{\infty} F(\theta) e^{iRG(\theta)} d\theta$$

con

$$F(\theta) = 1; \quad R = 1; \quad G(\theta) = \cosh(\theta);$$

Al exponente $RG(\theta)$ se le suele llamar la fase. La fase y la parte real del integrando tienen el siguiente aspecto (Figuras 5.7 y 5.8):

1. Existe un método aproximado para evaluar integrales de este tipo [4] debido a Kelvin [5] [6], el de la fase estacionaria.

$$I \approx \sqrt{\frac{2\pi}{|RG_0''|}} F_0 e^{iRG_0 \pm i\pi/4}$$

   Se pide calcular la parte real de I mediante este método, sabiendo que el signo positivo o negativo en el exponente depende del signo de la segunda derivada de G en el único punto donde se anula su derivada, al cual se refiere el subíndice 0. Si la segunda derivada de G en ese punto es positiva, se tomará el valor con el signo positivo, y si esa segunda derivada es negativa, se tomará el de signo negativo.

2. Evaluar de modo aproximado la parte real de I, tomando como límites de integración ± 3 y sustituyendo el integrando por su spline de grado uno $S(\theta)$ que lo interpola en la partición Ω del intervalo $[-3, 3]$ [7].

$$\Omega = \{-3, -2.5, -2, -1.5, -1, -0.5, 0, 0.5, 1, 1.5, 2, 2.5, 3\}$$

[4] Es adecuado para estimar la integral de funciones que presentan grandes oscilaciones. En la zona de estas oscilaciones, las áreas positivas se cancelan con las negativas, y sólo cuando estas oscilaciones cesan hay contribución neta a la integral. Estas oscilaciones cesan cuando hay extremos relativos de la función fase, o sea, puntos en los que la fase se comporta como estacionaria.

[5] Kelvin desarrolló este método para justificar el aspecto de las olas que deja un barco (Figura 5.9), aunque luego se ha venido aplicando a diferentes problemas de superposición de ondas. La idea es que la altura de ola en cada punto es el resultado de superponer muchas olas, pero sólo contribuyen de modo neto unas pocas.

[6] Se considera a lord Kelvin (1824-1907) como uno de los grandes de la Historia de la Física. Hizo contribuciones fundamentales en termodinámica, sobre todo en la segunda ley, la de la disipación de la energía. A él se debe la escala de temperaturas absolutas al descubrir que el movimiento molecular se detiene a -273° Celsius. También tuvo mucha fama en la época por ser muy buen físico aplicado y a él se debe la primera línea de telégrafo entre Europa y América hacia 1865. Aunque nacido en Belfast, en Irlanda, pasó casi toda su vida en Glasgow, Escocia, dando clase en su universidad. Tenía muy buena relación con otros científicos importantes de la época, como Stokes, Joule o Maxwell. Fue un científico-ingeniero muy reconocido; de hecho, consiguió el título de lord y está enterrado al lado de Newton en la abadía de Westminster.

[7] Aunque esta aproximación a la integral impropia es muy grosera (por sus límites y por el espaciado entre tramos del spline), el valor obtenido es del mismo orden que el del apartado 1.

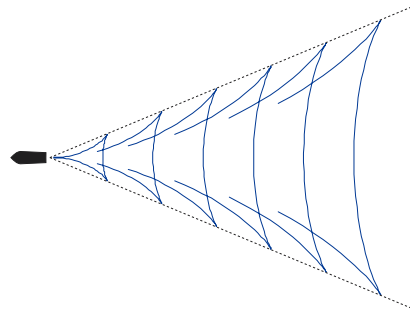


Figura 5.9: Tren de olas de Kelvin.

3. Evaluar I utilizando el método compuesto de los trapecios, el mismo del apartado 2, integrando con diferentes pasos y comprobando si hay convergencia en la integral, para valores suficientemente altos de los límites de integración numéricos. Para realizar las integrales se utilizará Matlab disminuyendo el paso hasta que la diferencia entre dos integrales consecutivas con paso decreciente sea menor que una precisión suficiente para los valores en que nos movemos en este problema.

Solución:

1. Sustituyendo los valores, tenemos que:

$$I = \int_{-\infty}^{\infty} e^{i\cosh(\theta)} d\theta = \int_{-\infty}^{\infty} (\cos(\cosh(\theta)) + i\sin(\cosh(\theta)))\, d\theta$$

Sólo nos interesa la parte real

$$real(I) = \int_{-\infty}^{\infty} \cos(\cosh(\theta)) d\theta$$

El coseno hiperbólico tiene solamente un punto de derivada nula (punto estacionario), el correspondiente a $\theta = 0$. En ese punto, $G_0'' = G''(0) = 1$, $G_0 = G(0) = 1$ y $F_0 = F(0) = 1$. Por tanto:

$$I \approx \sqrt{\frac{2\pi}{|1 \cdot 1|}} \cdot 1 \cdot e^{i(1+\pi/4)} \quad \Rightarrow \quad real(I) \approx \sqrt{2\pi}\cos(1+\pi/4) = -0.5338$$

2. En realidad, integrar ese spline es similar a utilizar la regla compuesta del trapecio. Construir un spline de grado 1 consiste en calcular un polinomio de interpolación a trozos de grado uno y clase 0, o sea, interpolar mediante rectas el valor de la función en los nodos. Integrar después este spline es lo mismo que aplicar la regla compuesta del trapecio a la función en esa partición del intervalo de integración. La integral a evaluar es:

$$I_1 = \int_{-3}^{3} \cos(\cosh(\theta)) d\theta$$

Si evaluamos el integrando f en la partición mediante unas sencillas órdenes Matlab

```
» theta=-3:0.5:3;
» f=cos(cosh(theta));
» plot(theta,f,'-o');
```

podremos visualizar la función a integrar (Figura 5.10). Integrar este polinomio a trozos consiste en evaluar el sumatorio

$$I_1 \approx \sum_{k=0}^{11} \int_{-3+k\cdot 0.5}^{-3+(k+1)0.5} S(\theta) d\theta = \sum_{k=0}^{11} 0.5 \left[\frac{1}{2} f(-3+k\cdot 0.5) + \frac{1}{2} f(-3+(k+1)\cdot 0.5)\right] = -0.2028$$

El orden es el mismo que el obtenido por el método de la fase estacionaria el cual tiene un error del orden del inverso de R. Para hacer este sumatorio con Matlab ejecutamos la orden

```
» 0.5*(sum(f)-f(1)/2-f(13)/2)
```

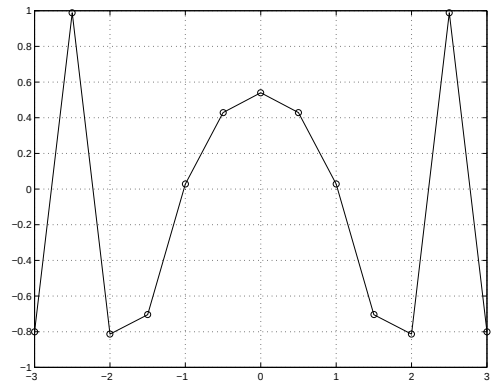


Figura 5.10: Spline correspondiente al apartado 2 del problema 5.4.

3. El método de la fase estacionaria establece que sólo la parte de la función en el entorno de los puntos de fase estacionaria contribuye a la integral. Es interesante comprobar esto, aumentando los límites de esas integrales y comprobando si hay convergencia hacia algún valor. Para evaluar estas integrales numéricamente, adoptamos una técnica de disminución de paso (dividiéndolo por dos), evaluando la diferencia en la integral con un paso y con su paso mitad hasta que ese valor sea inferior a un umbral de convergencia ϵ. Para ello utilizamos el código Matlab *fasestacionaria.m* que se incluye en la librería de códigos, y cuyo resultado aparece en la Figura 5.11 en la que representamos $I(a)$, frente a a,

$$I(a) = \int_{-a}^{a} \cos(\cosh(\theta))d\theta$$

En la Figura 5.11 podemos observar cómo a medida que el intervalo de integración crece, la integral se estabiliza en un valor en torno a -0.25. Ello es debido a que a medida que el intervalo de integración crece, entran en juego zonas de la integral de grandes oscilaciones que no representan una contribución neta a la integral. Es interesante ver cómo convergen la parte real e imaginaria de la integral a medida que a crece, teniendo como límite el valor de la integral (Figura 5.12).

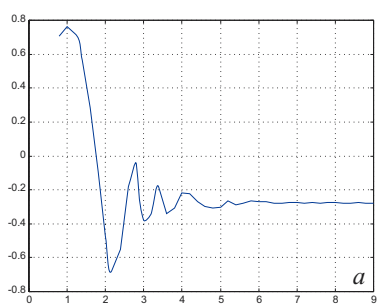


Figura 5.11: $I(a)$ **relativa al apartado 2 del problema 5.4.**

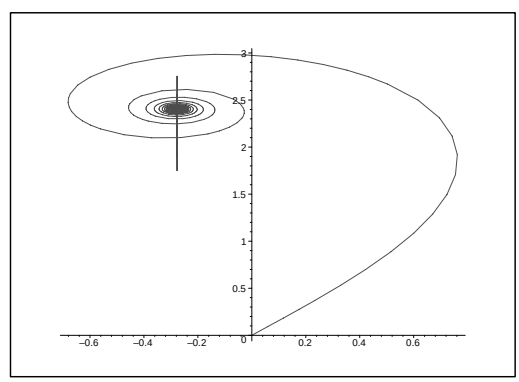


Figura 5.12: $(Real(I(a)), Imag(I(a)))$.

PROBLEMA 5.5 *Método compuesto de Gauss-Legendre.*

Se considera la partición de abscisas estrictamente crecientes $\Omega = \{t_0 = a, t_1, \cdots, t_n = b\}$ del compacto $[a, b]$, con $a < b$. Sea f una función integrable definida en $[a, b]$ que toma valores en $I\!R$. Se considera el problema de estimar:

$$\int_{t_i}^{t_{i+1}} f(t)dt \approx A_i^0 f(u_i^0) + A_i^1 f(u_i^1)$$

1. Encontrar valores para A_i^0 , A_i^1, u_i^0, u_i^1, que hagan que la fórmula tenga grado de precisión máximo.

2. Sea $E_1(\Omega)$ el espacio de polinomios a trozos de grado 1 y de clase $\mathcal{C}^{-1}$, discontinuos en los nodos, asociados a la partición Ω. Hallar la dimensión de $E_1(\Omega)$.

3. Se trata ahora de calcular un elemento $e \in E_1(\Omega)$ que interpole a f en la colección de nodos $\{u_i^0, u_i^1\}$, $i = 0, n-1$. Demostrar que el problema tiene solución única. Dar la expresión de cada uno de los tramos de e, $e|_{[t_i, t_{i+1}]}$, en la base de Newton de $P_1(I\!R)$ asociada a los nodos $\{u_i^0, u_i^1\}$, $i = 0, n-1$.

4. Calcular
$$\int_{t_0}^{t_n} e(t)dt$$

5. Sean $a = 0$, $b = 3$, y $n = 3$ y $E_3(\Omega)$ el espacio de polinomios a trozos de grado 3 y clase $\mathcal{C}^0$. Hallar la dimensión de $E_3(\Omega)$.

6. Sea
$$f(t) = \sin\left(\frac{1}{t+0.3}\right)\left(\frac{1}{t+0.3}\right)^2$$
Sea ahora la partición equiespaciada (todos los h_i son iguales) y sea $C \in E_3(\Omega)$ que interpola a f en los nodos de Ω y en los puntos u_i^0, u_i^1, $i = 0, 1, 2$ correspondientes. Calcular
$$\int_0^3 C(t)dt$$
y comparar el resultado con el resultado exacto.
Se sabe que $f(t)$ es la derivada de la funciòn $\cos\left(\frac{1}{t+0.3}\right)$.

7. Casi sin querer hemos creado en el apartado 6 el método compuesto de Gauss. Vamos a calcular su orden de error, viendo cómo tiende a 0 ese error cuando el espaciado entre tramos h tiende a 0. Para ello escribiremos un programa que integre $C(t)$ entre 0 y 3, y para definir $C(t)$ utilizaremos un número creciente de tramos. Se pide hacer un gráfico de esta curva, que servirá de base para un problema de aproximación por mínimos cuadrados que se resolverá en el capítulo correspondiente.

Solución:

1. Se trata de evaluar los coeficientes de una fórmula del tipo siguiente cuyo grado de precisión sea máximo.
$$\int_{t_i}^{t_{i+1}} f(t)dt \approx A_i^0 f(u_i^0) + A_i^1 f(u_i^1)$$
Como vemos, tenemos la posibilidad de actuar tanto sobre los coeficientes que afectan a la función evaluada en los puntos como a la selección de los propios puntos. Es por tanto un problema de integración gaussiana. Como además no hay una función de peso, el problema es de Gauss-Legendre, y si además el intervalo fuese el $[-1, 1]$, los puntos serían las raíces del polinomio de Legendre del grado correspondiente (en este caso 2).
Como el intervalo es uno genérico, haremos un cambio de variable que lo lleve al $[-1, 1]$.

Sea $h_i = t_{i+1} - t_i$. Definamos r como

$$r = 2\,\frac{t - t_i}{h_i} - 1 \quad \Rightarrow \quad dr = \frac{2}{h_i}\,dt$$

Con esta nueva variable, la fórmula aproximada es (ver la ecuación (5.3))

$$\begin{aligned}\int_{t_i}^{t_{i+1}} f(t)dt &= \frac{h_i}{2}\int_{-1}^{1} f(t(r))dr \approx \frac{h_i}{2}\left[f\left(t\left(-\frac{1}{\sqrt{3}}\right)\right) + f\left(t\left(\frac{1}{\sqrt{3}}\right)\right)\right] \\ &= \frac{h_i}{2}\left[f\left(t_i + \frac{\sqrt{3}-1}{2\sqrt{3}}h_i\right) + f\left(t_i + \frac{\sqrt{3}+1}{2\sqrt{3}}h_i\right)\right] \\ &= \frac{h_i}{2}\left[f\left(t_i + 0.2113h_i\right) + f\left(t_i + 0.7887h_i\right)\right]\end{aligned}$$

2. Cada tramo de e es un segmento de recta, que exige dos condiciones para su definición. Esto sucede en los n tramos del polinomio a trozos e (ver Figura 5.13). No tenemos niguna restricción respecto a estos $2n$ grados de libertad y por tanto tenemos que la dimensión de $E_1(\Omega)$ es $2n$.

3. Tenemos $2n$ condiciones, que son el valor del polinomio a trozos e en las parejas de nodos $\{u_i^0, u_i^1\}$, $i = 0, n-1$. Cada segmento de recta es único por estar definido por dos puntos distintos entre sí, y por tanto el polinomio a trozos e que los engloba también es único. Sea $f_i^j := f(u_i^j)$. La expresión de cada tramo en la base de Newton de diferencias divididas $\left\{1, t - u_i^0\right\}$ es:

$$e|_{[t_i,t_{i+1}]} = f_i^0 + \frac{f_i^1 - f_i^0}{u_i^1 - u_i^0}(t - u_i^0) = f_i^0 + \sqrt{3}\frac{f_i^1 - f_i^0}{h_i}(t - u_i^0)$$

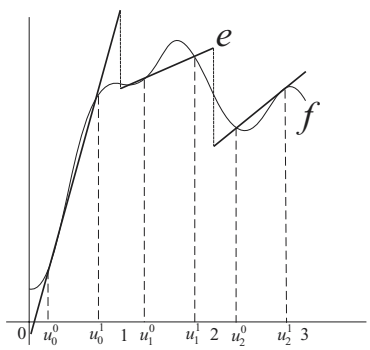


Figura 5.13: Elemento de $E_1(\Omega)$, e, que interpola a f en los puntos de Gauss de los intervalos.

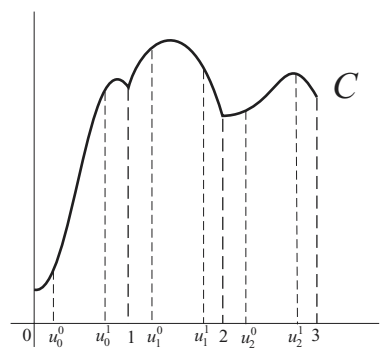


Figura 5.14: Elemento de $E_3(\Omega)$.

4. Al interpolar una función con un polinomio a trozos de grado 1 y clase $\mathcal{C}^{-1}$ en los nodos de Gauss, la fórmula de Gauss desarrollada en el apartado 1 tiene que ser exacta tramo a tramo y por tanto

$$\int_{t_0}^{t_n} e(t)dt = \sum_{i=0}^{n-1}\int_{t_i}^{t_{i+1}} e(t)dt = \sum_{i=0}^{n-1}\frac{f\left(u_i^0\right) + f\left(u_i^1\right)}{2}h_i = \sum_{i=0}^{n-1}\frac{f\left(t_i + 0.2113h_i\right) + f\left(t_i + 0.7887h_i\right)}{2}h_i$$

5. Siguiendo un razonamiento similar al de 2, cada tramo del C es una cúbica (Figura 5.14), que exige 4 condiciones para su definición. Esto sucede en los 3 tramos del polinomio a trozos C. Si a estos 12 grados de libertad restamos la restricción entre los tramos en los 2 nodos interiores, tenemos que la dimensión de $E_3(\Omega)$ es 10.

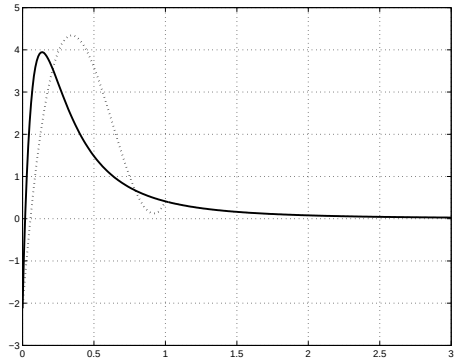


Figura 5.15: Interpolación mediante una cúbica a trozos.

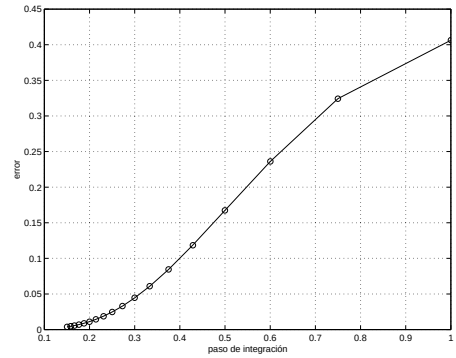


Figura 5.16: Error en función del paso.

6. Interpolar f por C en los nodos de Ω y en los puntos u_i^0, u_i^1, $i = 0, 1, 2$ correspondientes, e integrar C, no requiere integrar en cada uno de sus tramos, porque la fórmula desarrollada en el apartado 1, a pesar de utilizar dos puntos, es exacta para las cúbicas. Su aplicación repetida en estas cúbicas es equivalente a definir la fórmula compuesta de Gauss, la cual nos permite decir que

$$\int_{t_0}^{t_n} f(t)dt \approx \int_{t_0}^{t_n} C(t)dt = \sum_{i=0}^{n-1} \int_{t_i}^{t_{i+1}} C(t)dt = \sum_{i=0}^{n-1} \frac{f\left(t_i + 0.2113h_i\right) + f\left(t_i + 0.7887h_i\right)}{2} h_i$$

Por tanto, a pesar de las diferencias en sus gráficas, e y C aproximan de igual modo la integral de f. Aplicado a estos datos concretos:

$$\begin{aligned}
\int_0^3 f(t)dt &\approx \int_0^3 C(t)dt = \frac{1}{2}\sum_{i=0}^{2}\left[f\left(t_i + 0.2113\right) + f\left(t_i + 0.7887\right)\right] \\
&= \frac{1}{2}\left[\left(f\left(0.2113\right) + f\left(0.7887\right)\right) + \left(f\left(1.2113\right) + f\left(1.7887\right)\right) + \left(f\left(2.2113\right) + f\left(2.7887\right)\right)\right] \\
&= \frac{1}{2}\left[\left(3.5449 + 0.6705\right) + \left(0.2690 + 0.1056\right) + \left(0.0615 + 0.0333\right)\right] = 2.3424
\end{aligned}$$

El valor exacto de la integral I es:

$$I = \cos\left(\frac{1}{1+0.3}\right) - \cos\left(\frac{1}{0.3}\right) = 1.9361$$

La diferencia se debe sobre todo al primer tramo, como podemos observar en la Figura 5.15.

7. Para ir integrando con pasos decrecientes, escribimos un pequeño código Matlab, con el cual generamos la Figura 5.16. En esta figura se observa cómo el error tiende a 0 con el paso. Se trata de ajustarlo con una curva del tipo kh^p y ver qué valor de p produce menor error cuadrático medio. Esto nos proporcionará una idea del orden del método sin entrar en análisis de error que son bastante complicados, salvo en casos muy sencillos, como en el del método de los trapecios (5.1.2).

   El problema que acabamos de resolver está relacionado con las técnicas de integración de ecuaciones en derivadas parciales mediante el método de los elementos finitos. Una parte importante del método consiste en evaluar la integral de una función incógnita que se interpola a trozos en cada elemento de la discretización geométrica del problema. Para escribir estas integrales se suelen usar como puntos de referencia aquellos en los que las integrales son más precisas, o sea, los puntos de Gauss asociados, los cuales se obtienen mediante expresiones estándar, similares a las aquí desarrolladas.

PROBLEMA 5.6 *Integración multidimensional.*

Se considera la función

$$f(x,y) := \begin{cases} 0 & r > 1 \\ +\sqrt{1-r^2} & r \leq 1 \end{cases} ; \qquad r = \sqrt{x^2+y^2}$$

que corresponde al casquete superior de la esfera unidad (Figura 5.17), cuyo volumen es $(2/3)\pi = 2.0944$. Nuestro objetivo es obtener este valor integrando numéricamente

$$I = \int_{-1}^{1}\int_{-1}^{1} f(x,y)dxdy$$

Elegimos para realizar la discretización el mismo paso $h = 0.5$ en las dos direcciones, de modo que la partición del intervalo $[-1,1]$ tanto para la variable x como para la y será $\Omega = \{-1,-0.5,0,0.5,1\}$ (ver Figura 5.17). Desde la salida sabemos que vamos a tener un error importante con esta discretización del problema, pero podemos esperar un orden de magnitud adecuado.

1. Construir el polinomio de interpolación de Lagrange en 2 variables de la función f en la rejilla $\Omega \times \Omega$.
2. Dibujar la gráfica correspondiente a este polinomio de interpolación utilizando Matlab.
3. Estimar el volumen de la semiesfera, integrando el polinomio obtenido en el apartado 1.
4. Sea $\Pi = \{x_0 = -1, x_1, \ldots, x_n = 1\}$ una partición equiespaciada genérica del intervalo $[-1,1]$ de paso h. Se considera la función de interpolación a trozos de grado 0 discontinua s, que interpola a f en los nodos de $\Pi \times \Pi$, y que toma en cada cuadrícula de la rejilla el valor de la función f en el nodo de las menores coordenadas (ver problema 5.3 para una aplicación similar en una variable).

   $$s(x,y) = f(x_i,y_j),\ x_i \leq x < x_{i+1},\ y_j \leq y < y_{j+1},$$

   Suponiendo que sustituimos f por s, calcular

   $$I \approx \int_{-1}^{1}\int_{-1}^{1} s(x,y)dxdy$$

5. Aplicación para la partición $\Pi = \Omega$.

Solución:

1. En la página 143 obtuvimos que el polinomio buscado es para este caso:

   $$P(x,y) = \sum_{0\leq i,j\leq 4} f(x_i,y_j)l_i(x)l_j(y) = \sum_{0\leq i,j\leq 4} F_{ij}l_i(x)l_j(y)$$

   siendo $\mathbf{F}$ la matriz resultado de evaluar la función f en $\Omega \times \Omega$.

   $$F_{ij} = f(x_i,y_j),\ 0 \leq i,j \leq 4, \quad \mathbf{F} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0.7071 & 0.8660 & 0.7071 & 0 \\ 0 & 0.8660 & 1.0000 & 0.8660 & 0 \\ 0 & 0.7071 & 0.8660 & 0.7071 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

   $$\begin{aligned} P(x,y) \;=\; & 0.7071l_1(x)l_1(y) + 0.8660l_1(x)l_2(y) + 0.7071l_1(x)l_3(y) \\ & +0.8660l_2(x)l_1(y) + 1.0000l_2(x)l_2(y) + 0.8660l_2(x)l_3(y) \\ & +0.7071l_3(x)l_1(y) + 0.8660l_3(x)l_2(y) + 0.7071l_3(x)l_3(y) \end{aligned}$$

   Para obtener estos valores, hemos usado las líneas Matlab correspondientes al archivo *integracion2d.m*. Los polinomios de Lagrange asociados serán tanto para la x como para la y:

   $$l_j(t) = \frac{\prod_{i=0,i\neq j}^{i=4}(t-t_i)}{\prod_{i=0,i\neq j}^{i=4}(t_j-t_i)}$$

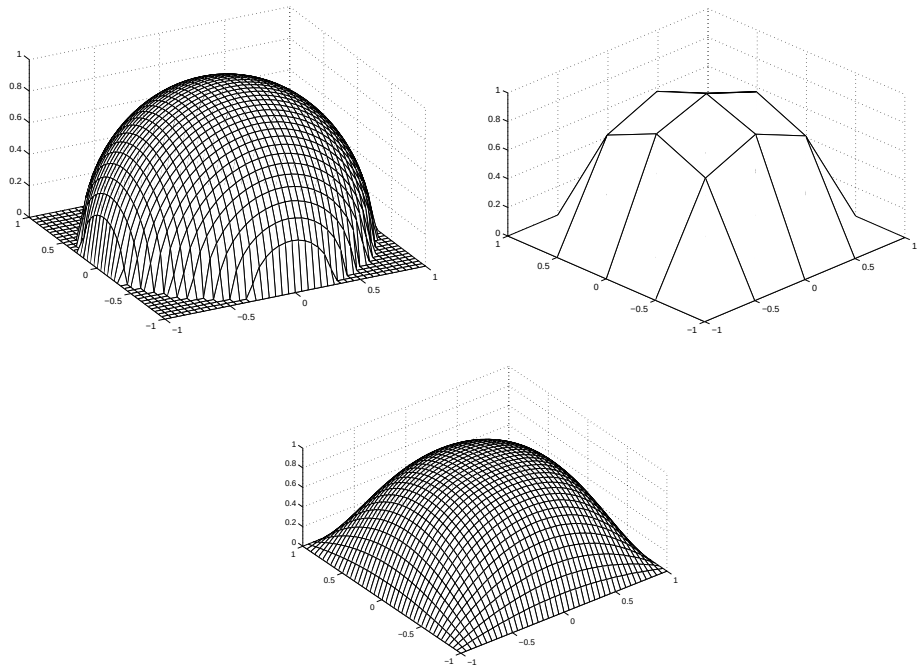


Figura 5.17: Representación, discretización e interpolación de la superficie del problema 5.6.

2. La representación gráfica de la superficie de interpolación que incluimos en la Figura 5.17 se obtiene con las líneas Matlab que tenemos en el mismo archivo *integracion2d.m*.

3. Tenemos que evaluar

$$I = \int_{-1}^{1}\int_{-1}^{1} f(x,y)dxdy \approx \int_{-1}^{1}\int_{-1}^{1} P(x,y)dxdy = \sum_{0\le i,j\le 4} F_{ij}\int_{-1}^{1} l_i(x)dx\int_{-1}^{1} l_j(y)dy = \sum_{0\le i,j\le 4} I_i I_j F_{ij}$$

$$I_j = \int_{-1}^{1} l_j(t)dt \qquad j = 0, 4$$

Ya que la subdivisión es simétrica respecto a $x = 0$ las integrales I_j correspondientes a las parejas l_0, l_4 y l_2, l_3 son iguales entre sí; además las primeras son irrelevantes porque los términos F_{ij} que las afectan son nulos. Los valores de las integrales necesarias son, $I_1 = I_3 = 0.7111$, $I_2 = 0.2667$ (se deja como ejercicio evaluarlas numérica o analíticamente).

$$I\int_{-1}^{1}\int_{-1}^{1} f(x,y)dxdy \approx \sum_{0\le i,j\le 4} F_{ij} I_i I_j = 2.1583$$

Como vemos, el valor es del mismo orden que el valor real a pesar de lo grosero de la discretización, y también es fácil ver que, aun así, los cálculos son muy complicados, debido sobre todo a las integrales I_j. Esta complejidad invita a pensar en algún método más sencillo para realizar esta integral, lo que nos lleva al siguiente apartado.

4. Podemos observar la gráfica de una función de interpolación a trozos genérica s en la Figura 5.18. Tenemos que estimar:

$$\begin{aligned} I &\approx \int_{-1}^{1}\int_{-1}^{1} s(x,y)dxdy = \sum_{0\le i,j\le n-1}\int_{x_i}^{x_{i+1}}\int_{y_j}^{y_{j+1}} s(x,y)dxdy \\ &= \sum_{0\le i,j\le n-1}\int_{x_i}^{x_{i+1}}\int_{y_j}^{y_{j+1}} F_{ij}dxdy = \sum_{0\le i,j\le n-1} F_{ij}\Delta x\Delta y = h^2\sum_{0\le i,j\le n} F_{ij}. \end{aligned}$$

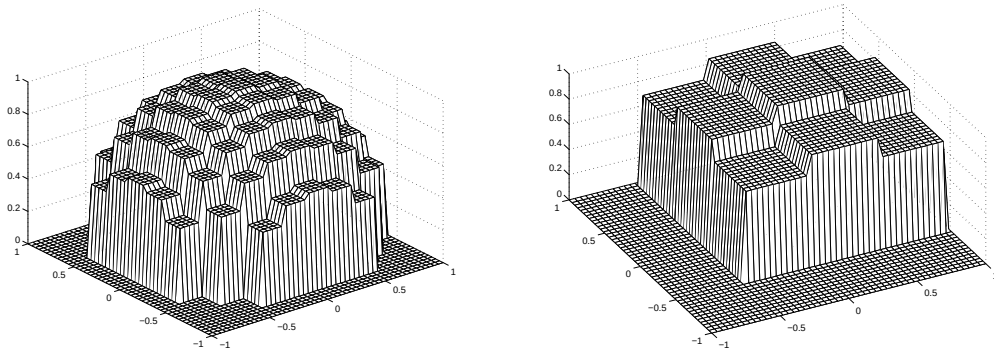


Figura 5.18: Funciones de interpolación de los apartados 4 y 5 del problema 5.6.

5. Si aplicamos lo anterior a nuestro caso (ver Figura 5.18 para la gráfica), tendremos con $h = 0.5$.

$$I \approx \int_{-1}^{1}\int_{-1}^{1} s(x,y)dxdy = h^2 \sum_{0\le i,j\le 3} F_{ij} = 0.25 * (4 \cdot 0.7071 + 4 \cdot 0.8660 + 1.0000) = 1.8231$$

El orden de magnitud no está tan mal. Hay que tener en cuenta que el valor real es 2.0944. Si tomamos $n = 10$, $I \approx 2.0174$. Para $n = 25$, $I \approx 2.0906$. Con una técnica tan sencilla, obtenemos este valor tan preciso. Si estuviésemos midiendo un recipiente con unidades internacionales, diríamos que puede albergar 2090 litros en vez de los 2094 reales, que para un ingeniero es una respuesta perfecta. El código Matlab utilizado es muy similar al del apartado 1 y también está en el archivo *integracion2d.m*. La orden clave es *sum*, que utilizamos para sumar los elementos de la matriz F.

PROBLEMA 5.7 *Campo de velocidades inducido por un segmento de vórtices.*

En este problema se tratan aspectos del movimiento que produce un segmento lleno de vórtices en un medio fluido, movimientos cuyas líneas de corriente son circulares. Este fenómeno físico se utiliza en la construcción de modelos matemáticos de la sustentación debida a perfiles aerodinámicos.

La velocidad debida a un elemento infinitesimal de línea lleno de vórtices es:

$$\mathbf{dv} = \frac{\Gamma}{4\pi}\frac{\mathbf{dl} \wedge \mathbf{r}}{r^3}$$

donde:

- Γ magnitud que indica la intensidad del vórtice [8].
- P punto en el que calculamos la velocidad. Las coordenadas de P son $(0, 0, a)$, en este caso $(0, 0, 1)$.
- $\mathbf{dl}$ elemento vectorial diferencial de línea. El segmento de vórtices se extiende desde $(0, 0, 0)$ hasta $(0, 1, 0)$.
- $\mathbf{r}$ vector que une $\mathbf{dl}$ con P.
- r módulo de este vector.

Tras un cambio de variable, el valor en módulo de la velocidad inducida en P por el segmento de vórtices es igual a la integral

$$v = \frac{1}{4\pi a}\int_{\beta_0}^{\beta_1} \Gamma(\beta)\sin\beta d\beta = \frac{1}{4\pi}\int_{\pi/2}^{3\pi/4} \Gamma(\beta)\sin\beta d\beta$$

[8] A mayor intensidad, mayor velocidad.

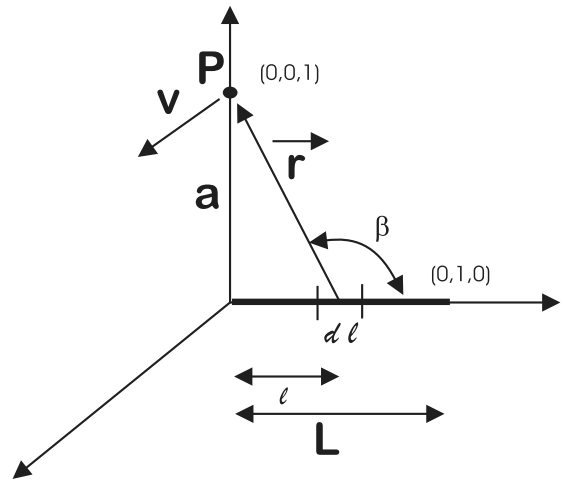


Figura 5.19: Velocidad en P inducida por una distribución de vórtices.

donde β es el ángulo que forma el vector que une el punto P y un punto dentro de la línea de vórtices, con dicha línea de vórtices. $\Gamma(\beta)$ expresa la intensidad del vórtice en cada punto del segmento, en función de ese ángulo.
Cuando la función que define la intensidad Γ de los vórtices no es sencilla, esa integral es difícil de evaluar, y se trata de modo discreto. Los dos siguientes apartados abordan de ese modo la integral.

1. Encontrar los coeficientes de la fórmula de integración

$$\frac{1}{4\pi}\int_{\pi/2}^{3\pi/4}\Gamma(\beta)\sin\beta d\beta = A\cdot\Gamma\left(\frac{\pi}{2}\right)+B\cdot\Gamma\left(\frac{3\pi}{4}\right)$$

   de modo que su grado de precisión sea máximo.

2. Encontrar los coeficientes de la fórmula de integración

$$\frac{1}{4\pi}\int_{\pi/2}^{3\pi/4}\Gamma(\beta)\sin\beta d\beta = A\cdot\Gamma\left(\beta_A\right)+B\cdot\Gamma\left(\beta_B\right)$$

   de modo que el grado de precisión sea máximo [9].

Solución:

1. Buscamos A, B de modo que la fórmula

$$\frac{1}{4\pi}\int_{\pi/2}^{3\pi/4}\Gamma(\beta)\sin\beta d\beta = A\cdot\Gamma\left(\frac{\pi}{2}\right)+B\cdot\Gamma\left(\frac{3\pi}{4}\right)$$

   tenga grado de precisión máximo.

   Con 2 parámetros libres A y B, conseguiremos que sea exacta en $P_1(\mathbb{R})$. Para encontrar los parámetros tomamos dos funciones que generen $P_1(\mathbb{R})$, por ejemplo $\Gamma_1(\beta)=1$ y $\Gamma_2(\beta)=\beta$.

[9] Se recuerda que:

$$\int\beta^2\sin\beta d\beta = -\beta^2\cos\beta+2\beta\sin\beta+2\cos\beta$$

$$\int\beta^3\sin\beta d\beta = -\beta^3\cos\beta+3\beta^2\sin\beta+6\beta\cos\beta-6\sin\beta$$

Como la integral es un operador lineal, si es exacta para estos polinomios, lo será para cualquier otro, de grado uno, que siempre se podrá descomponer como combinación lineal única, de Γ_1 y Γ_2.

$$\left.\begin{array}{rcl} \frac{1}{4\pi}\int_{\pi/2}^{3\pi/4}\sin\beta d\beta & = & A+B \\ \frac{1}{4\pi}\int_{\pi/2}^{3\pi/4}\beta\sin\beta d\beta & = & A\frac{\pi}{2}+B\frac{3\pi}{4} \end{array}\right\} \Rightarrow \begin{array}{rcl} A & = & 0.029679 \\ B & = & 0.026591 \end{array}$$

Podríamos haber escogido otra base de $P_1(\mathbb{R})$, por ejemplo, la de Lagrange asociada a los nodos $\frac{\pi}{2}$, $\frac{3\pi}{4}$.

$$l_0(\beta)=\frac{\beta-\frac{3\pi}{4}}{\frac{\pi}{2}-\frac{3\pi}{4}},\quad l_1(\beta)=\frac{\beta-\frac{\pi}{2}}{\frac{3\pi}{4}-\frac{\pi}{2}}$$

Como la fórmula debe ser exacta $\forall p\in P_1(\mathbb{R})$,

$$\frac{1}{4\pi}\int_{\pi/2}^{3\pi/4} l_0(\beta)\sin\beta d\beta = Al_0\left(\frac{\pi}{2}\right)+Bl_0\left(\frac{3\pi}{4}\right)=A\cdot 1+B\cdot 0=A=0.029679$$

$$\frac{1}{4\pi}\int_{\pi/2}^{3\pi/4} l_1(\beta)\sin\beta d\beta = Al_1\left(\frac{\pi}{2}\right)+Bl_1\left(\frac{3\pi}{4}\right)=A\cdot 0+B\cdot 1=B=0.026592$$

2. En este caso, tenemos 4 parámetros libres A, B, β_A y β_B, así que trataremos de que la fórmula sea exacta en $P_3(\mathbb{R})$

$$\frac{1}{4\pi}\int_{\pi/2}^{3\pi/4}\Gamma(\beta)\sin\beta d\beta = A\cdot\Gamma(\beta_A)+B\cdot\Gamma(\beta_B)$$

Sabemos por la teoría de interpolación que cualquier polinomio de grado tres que pase por los puntos $(\beta_A,\Gamma(\beta_A))$, $(\beta_B,\Gamma(\beta_B))$ es de la forma

$$p_3(\beta)=p_1^*(\beta)+(\beta-\beta_A)(\beta-\beta_B)q(\beta),\quad \forall q\in P_1(\mathbb{R})$$

siendo $p_1^*(\beta)$ la recta que pasa por esos dos puntos.

Como la fórmula debe ser exacta para este polinomio:

$$\int_{\pi/2}^{3\pi/4} p_3(\beta)\sin\beta d\beta=\int_{\pi/2}^{3\pi/4} p_1^*(\beta)\sin\beta d\beta+\int_{\pi/2}^{3\pi/4}(\beta-\beta_A)(\beta-\beta_B)q(\beta)\sin\beta d\beta$$

Si los puntos son $(\beta_A,0)$,$(\beta_B,0)$, se tiene $p_3(\beta_A)=0$, $p_3(\beta_B)=0$, y $p_1^*(\beta_A)=0$, $p_1^*(\beta_B)=0$, de donde $p_1^*(\beta)=0$, luego

$$\frac{1}{4\pi}\int_{\pi/2}^{3\pi/4}\Gamma(\beta)\sin\beta d\beta=A\cdot\Gamma(\beta_A)+B\cdot\Gamma(\beta_B)=A\cdot 0+B\cdot 0=0$$

y

$$0=\int_{\pi/2}^{3\pi/4} 0\cdot\sin\beta d\beta+\int_{\pi/2}^{3\pi/4}(\beta-\beta_A)(\beta-\beta_B)q(\beta)\sin\beta d\beta$$

Tenemos que buscar, por tanto, β_A y β_B tales que:

$$\int_{\pi/2}^{3\pi/4}(\beta-\beta_A)(\beta-\beta_B)q(\beta)\sin\beta d\beta=0,\quad \forall q\in P_1(\mathbb{R})$$

Usando de nuevo el mismo razonamiento que en el apartado 1, es suficiente probarlo para $q_1(\beta)=1$ y $q_2(\beta)=\beta$. Se facilita el cálculo de β_A y β_B poniendo

$$(\beta-\beta_A)(\beta-\beta_B)=\beta^2+M\beta+N$$

con ello,

$$\left.\begin{array}{rcl}\int_{\pi/2}^{3\pi/4}\left(\beta^2+M\beta+N\right)\sin\beta d\beta &=& 0\\ \int_{\pi/2}^{3\pi/4}\left(\beta^2+M\beta+N\right)\beta\sin\beta d\beta &=& 0\end{array}\right\}\Rightarrow\begin{array}{rcl}M &=& -3.90671\\ N &=& 3.76554\end{array}$$

Resolviendo por último la ecuación de segundo grado, tenemos

$$\beta_A = 1.72964 = 0.55056\pi, \quad \beta_B = 2.17707 = 0.69298\pi$$

Se tienen las desigualdades muy razonables

$$\frac{\pi}{2} < \beta_A < \beta_B < \frac{3\pi}{4}$$

Una vez conocidos β_A y β_B, reproducimos el apartado 1, sustituyendo en el segundo miembro los valores β_A y β_B que acabamos de calcular. Así se obtiene

$$A = 0.029568, \quad B = 0.026702$$

Ésta sería la forma de hacer este problema usando la teoría de integración gaussiana, y jugando con la función $\sin\beta$.

Cuando se está habituado, se puede plantear directamente el sistema no lineal siguiente, aprovechando que la fórmula debe ser exacta en $P_3(\mathbb{R})$, tomando la base de monomios $\{1, \beta, \beta^2\beta^3\}$.

$$\frac{1}{4\pi}\int_{\pi/2}^{3\pi/4}\sin\beta d\beta = A + B \qquad (\dagger)$$

$$\frac{1}{4\pi}\int_{\pi/2}^{3\pi/4}\beta\sin\beta d\beta = A\beta_A + B\beta_B \qquad (\ddagger)$$

$$\frac{1}{4\pi}\int_{\pi/2}^{3\pi/4}\beta^2\sin\beta d\beta = A\beta_A^2 + B\beta_B^2$$

$$\frac{1}{4\pi}\int_{\pi/2}^{3\pi/4}\beta^3\sin\beta d\beta = A\beta_A^3 + B\beta_B^3$$

Calculando las integrales de los primeros miembros tenemos

$$\begin{array}{rcl}0.05627 &=& A + B\\ 0.10927 &=& A\beta_A + B\beta_B\\ 0.21502 &=& A\beta_A^2 + B\beta_B^2\\ 0.42852 &=& A\beta_A^3 + B\beta_B^3\end{array}$$

Multiplicando la primera ecuación por N, la segunda por M, sumando ambas y sumando al resultado la tercera ecuación, tenemos

$$A\left(\beta_A^2 + M\beta_A + N\right) + B\left(\beta_B^2 + M\beta_B + N\right) = 0.05627N + 0.10927M + 0.21502$$

Si β_A y β_B son las raíces del polinomio $\beta^2 + M\beta + N$, tenemos

$$0.05627N + 0.10927M + 0.21502 = 0 \quad (*)$$

Por otro lado, multiplicando la segunda ecuación por N, la tercera por M, sumando ambas y sumando al resultado la cuarta ecuación, tenemos

$$A\left(\beta_A^2 + M\beta_A + N\right)\beta_A + B\left(\beta_B^2 + M\beta_B + N\right)\beta_B = 0.10927N + 0.21502M + 0.42582$$

Como β_A y β_B son las raíces del polinomio $\beta^2 + M\beta + N$, tenemos

$$0.10927N + 0.21502M + 0.42582 = 0 \quad (**)$$

con el sistema definido por las dos ecuaciones $(*)$ y $(**)$ obtenemos M y N. Resolviendo entonces la ecuación de segundo grado, encontramos β_A y β_B, y sustituyendo estos valores en las ecuaciones $(\dagger)$ y $(\ddagger)$, encontramos finalmente A y B, que coinciden con los valores calculados con el otro método.

PROBLEMA 5.8 *Cálculo de la longitud de una curva.*

El objetivo de este problema es calcular la longitud de una curva cerrada paramétrica que ya hemos considerado en el problema 3.15, y cuya representación gráfica aparece en la Figura 3.36.
Dada una curva paramétrica $x(t), y(t)$, su longitud entre dos valores $t = a$ y $t = b$ del parámetro, es

$$L = \int_a^b \sqrt{x'(t)^2 + y'(t)^2}dt$$

En nuestro caso, $x(t)$ e $y(t)$ son splines de grado 2 cíclicos asociados a la partición $\Omega = \{-1, 0, 1, 2\}$ del compacto $[-1, 2]$ cuya expresión en cada uno de los tramos de x y de y es

$$x(t) = \begin{cases} -4t - 3t^2 & -1 \le t \le 0 \\ -4t + 3t^2 & 0 \le t \le 1 \\ -3 + 2t & 1 \le t \le 2 \end{cases} \quad ; \quad y(t) = \begin{cases} 1 - t^2 & -1 \le t \le 0 \\ 1 - t^2 & 0 \le t \le 1 \\ 4 - 6t + 2t^2 & 1 \le t \le 2 \end{cases}$$

1. Se pide estimar dicha longitud, a través del cálculo de la integral utilizando el método compuesto de los trapecios con un paso de 0.5 unidades.

2. A la vista de la curva, se pide discutir si el resultado es razonable.

3. Realizar dicho cálculo utilizando el método de Gauss de dos puntos, indicando claramente sus abscisas.

Solución:

1. La longitud de la curva dada es

$$L = \int_{-1}^{2} \sqrt{x'(t)^2 + y'(t)^2}dt$$

Calculemos las curvas derivadas $x'(t)$, $y'(t)$.

$$x'(t) = \begin{cases} -4 - 6t & -1 \le t \le 0 \\ -4 + 6t & 0 \le t \le 1 \\ 2 & 1 \le t \le 2 \end{cases} \quad ; \quad y'(t) = \begin{cases} -2t & -1 \le t \le 0 \\ -2t & 0 \le t \le 1 \\ -6 + 4t & 1 \le t \le 2 \end{cases}$$

Para aplicar el método compuesto de los trapecios, construimos la tabla correspondiente a los diferentes nodos que resultan del paso $h = 0.5$

t	-1.0	-0.5	0.0	0.5	1.0	1.5	2.0
x'	2	-1	-4	-1	2	2	2
y'	2	1	0	-1	-2	0	2
x'^2	4	1	16	1	4	4	4
y'^2	4	1	0	1	4	0	4
$x'^2 + y'^2$	8	2	16	2	8	4	8
$\sqrt{x'^2 + y'^2}$	$2\sqrt{2}$	$\sqrt{2}$	4	$\sqrt{2}$	$2\sqrt{2}$	2	$2\sqrt{2}$

La integral evaluada por el método compuesto de los trapecios queda entonces

$$I = h\left[\frac{1}{2}2\sqrt{2} + \sqrt{2} + 4 + \sqrt{2} + 2\sqrt{2} + 2 + \frac{1}{2}2\sqrt{2}\right] = 7.2426$$

2. Un resultado razonable a primera vista, si se compara con la longitud 6.2832 de una circunferencia de radio 1, figura geométrica con la que nuestra curva tiene bastante parecido.

3. Utilicemos ahora el método de Gauss de dos puntos

$$L = \int_{-1}^{2} \sqrt{x'(t)^2 + y'(t)^2}dt = A\sqrt{x'(t_0)^2 + y'(t_0)^2} + B\sqrt{x'(t_1)^2 + y'(t_1)^2}$$

Ya que los puntos de Gauss en el intervalo $[-1,1]$ para la fórmula de dos puntos son las raíces $\pm 1/\sqrt{3}$ del polinomio de Legendre de segundo grado, comenzamos con un cambio de variable que transforme nuestro intervalo $[-1,2]$ en $[-1,1]$.

$$u = \frac{2(t+1)}{3} - 1 \quad \Rightarrow \quad t = \frac{3u+3}{2} - 1 \quad \Rightarrow \quad dt = \frac{3}{2}du$$

$$\begin{aligned} L &= \int_{-1}^{2} \sqrt{x'(t)^2 + y'(t)^2}dt = \frac{3}{2}\int_{-1}^{1} \sqrt{x'(t(u))^2 + y'(t(u))^2}du \approx \\ &\approx \frac{3}{2}\left[A\sqrt{x'\left(t\left(-\frac{1}{\sqrt{3}}\right)\right)^2 + y'\left(t\left(-\frac{1}{\sqrt{3}}\right)\right)^2} + B\sqrt{x'\left(t\left(\frac{1}{\sqrt{3}}\right)\right)^2 + y'\left(t\left(\frac{1}{\sqrt{3}}\right)\right)^2}\right] \end{aligned}$$

Utilizando el método de los coeficientes indeterminados, es fácil ver que $A = B = 1$, ya que la fórmula debe ser exacta hasta los polinomios de grado 3.

Como

$$t\left(\frac{-1}{\sqrt{3}}\right) = \frac{\frac{-3}{\sqrt{3}} + 3}{2} - 1 = -0.3660 \quad \text{y} \quad t\left(\frac{1}{\sqrt{3}}\right) = \frac{\frac{3}{\sqrt{3}} + 3}{2} - 1 = 1.3660$$

se tiene

$$I = \frac{3}{2}\left[\sqrt{x'\left(-0.3660\right)^2 + y'\left(-0.3660\right)^2} + \sqrt{x'\left(1.3660\right)^2 + y'\left(1.3660\right)^2}\right]$$

Construimos la tabla de apoyo

t	x'	y'	x'^2	y'^2	$x'^2 + y'^2$	$\sqrt{x'^2 + y'^2}$
-0.3660	-1.8038	0.7321	3.2539	0.5360	3.7898	1.9467
1.3660	2	-0.5359	4	0.2872	4.2872	2.0706

de donde

$$I = \frac{3}{2}\left[1.9467 + 2.0706\right] = 6.0260$$

El valor real de la integral es 6.6522. El error es bastante grande en los dos casos.

Con una representación de la gráfica $\left(t, \sqrt{x'(t)^2 + y'(t)^2}\right)$ de la función integrando, Figura 5.20, se comprende lo difícil que es aproximar esta integral con tan pocos puntos.

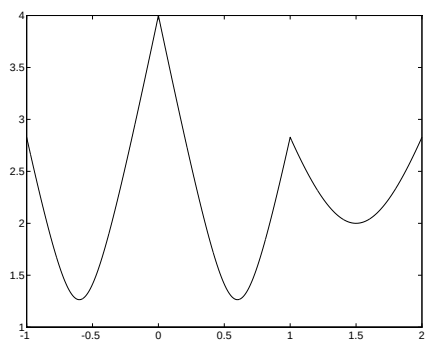


Figura 5.20: Grafo del integrando en el problema 5.8.

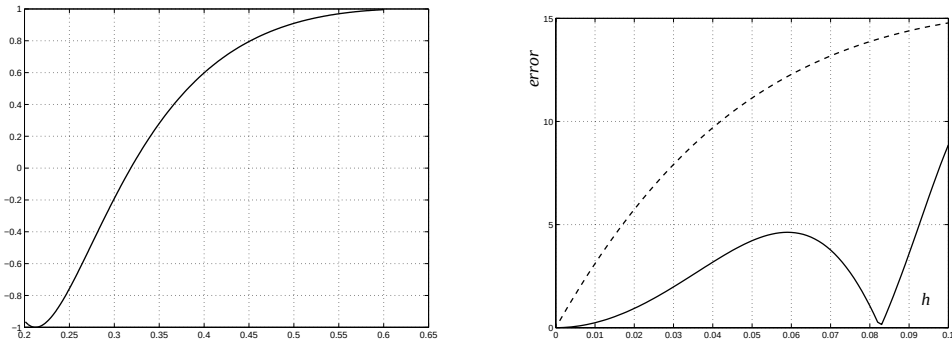


Figura 5.21: Función $\sin(1/x)$ **y error en sus derivadas con la fórmula centrada y de dos puntos (rayas).**

PROBLEMA 5.9 *Derivación numérica: fórmula de 2 puntos.*

El objetivo del problema es estimar numéricamente la derivada de la función $f(x)\sin(1/x)$ en el punto $x_0 = 0.2$.

1. Utilizar primero la fórmula de dos puntos progresiva "que mira hacia adelante". Se hará barriendo un rango de pasos, y viendo cómo varía la estimación en función del paso h.
2. Tomando $h = 0.1$ comprobar que se verifica la cota del error teórica (5.8).
3. Repetir el apartado 1 utilizando ahora la fórmula centrada.

Solución:

1. Representemos antes que nada la función f entre 0.2 y 0.6 utilizando las líneas Matlab (ver la Figura 5.21)

```
x0=0.2;
xx=x0+0.001:0.001:0.6;
ff=sin(1./xx);
plot(xx,ff);
```

Representemos la curva del error función de h, obtenida restando el valor obtenido para la derivada de f con la fórmula de 2 puntos, (5.6) y el valor real de la derivada en x_0, $f'(x_0) = -7.0916$ (ver la Figura 5.21). Vemos que el error tiende a 0 a medida que h tiende a 0.

Para construir el gráfico del error hemos usado las siguientes líneas Matlab.

```
x0=0.2;
xx=x0+0.001:0.001:0.3;
ff=sin(1./xx);
fp0=-1/x0^2*cos(1/x0);
fp=(ff-sin(1./x0))./(xx-x0);
plot(xx-x0,abs(fp-fp0));
```

2. El error ϵ obtenido para $h = 0.1$ es del orden de 15, razonable si observamos la gráfica de la función $\sin(1/x)$ en la Figura 5.21. La cota del error es la correspondiente a la ecuación (5.8)

$$\epsilon \leq \frac{\|f''\|_\infty}{2} h = 0.05 \| -\sin(1/x)/x^4 + 2\cos(1/x)/x^3 \|_\infty \leq 0.05(1/0.2^4 + 2/0.2^3) = 43.75$$

Después de hacer unas mayoraciones muy sencillas, hemos obtenido una mayoración que como debía suceder verifica nuestro error.

3. Se trata de rehacer el apartado 1 utilizando la fórmula centrada (5.10). Se incluyen a continuación las líneas Matlab que se usan tanto en el cálculo como en la representación del error correspondiente a esta fórmula, que aparece superpuesto con el error de la fórmula de dos puntos en la Figura 5.21.

   Podemos observar como para pequeños valores de h la diferencia es muy importante.

```
x0=0.2;
xx=x0+0.001:0.001:0.3;
xxm=x0-0.001:-0.001:0.1;
ff=sin(1./xx);
ffm=sin(1./xxm);
fp0=-1/x0^2*cos(1/x0);
fp=(ff-sin(1./x0))./(xx-x0); % formula de dos puntos
fpcentrada=(ff-ffm)./(xx-xxm); % formula centrada
plot(xx-x0,abs(fp-fp0),xx-x0,abs(fpcentrada-fp0));
```

PROBLEMA 5.10 *Fórmula de derivación de 4 puntos.*

Sea $h > 0$. Se considera la siguiente fórmula para estimar la tercera derivada en un punto x de una función f real de variable real

$$f'''(x) \approx \frac{1}{h^3}\left(f(x+3h) - 3f(x+2h) + 3f(x+h) - f(x)\right)$$

1. Obtenerla de modo razonado.
2. Estimar como aplicación numérica la tercera derivada de la función

   $$f(x) = \ln(x)$$

   en $x = 1$ mediante la fórmula objeto del problema, tomando un paso h de 0.01, y calcular su diferencia respecto al valor real.
3. Dar el término de error de la fórmula del enunciado.

Solución:

1. Como es una fórmula de 4 puntos, intentamos obtenerla a partir del polinomio de interpolación de los cuatro puntos $(x_0, f(x_0))$, $(x_0+h, f(x_0+h))$, $(x_0+2h, f(x_0+2h))$ y $(x_0+3h, f(x_0+3h))$. Llamemos $x_i := x_0 + ih$, $i = 0, 3$.

   $$P_3(x) = f(x_0)l_0(x) + f(x_1)l_1(x) + f(x_2)l_2(x) + f(x_3)l_3(x)$$

   $$l_j(x) = \prod_{\substack{i=0 \\ i\neq j}}^{3} \frac{x - x_i}{x_j - x_i} \quad j = 0, 3$$

   Si derivamos 3 veces cada uno de estos polinomios de grado 3, obtenemos una constante

   $$l_0'''(x) = -\frac{1}{h^3}, \quad l_1'''(x) = \frac{3}{h^3}, \quad l_2'''(x) = -\frac{3}{h^3}, \quad l_3'''(x) = \frac{1}{h^3},$$

   Con ello

   $$P_3'''(x) = f(x_0)l_0'''(x) + f(x_1)l_1'''(x) + f(x_2)l_2'''(x) + f(x_3)l_3'''(x) \quad \Rightarrow$$

$$\Rightarrow \quad P_3'''(x) = -\frac{1}{h^3}f(x_0) + \frac{3}{h^3}f(x_1) - \frac{3}{h^3}f(x_2) + \frac{1}{h^3}f(x_3)$$

que es precisamente la fórmula del enunciado.
Usando que la tercera derivada de una cúbica es constante, podemos obtener la fórmula más rápidamente mediante la base de diferencias divididas, ya que la tercera derivada del polinomio de interpolación en esta base

$$P_3(x) = f(x_0)+f[x_0,x_1](x-x_0)+f[x_0,x_1,x_2](x-x_0)(x-x_1)+f[x_0,x_1,x_2,x_3](x-x_0)(x-x_1)(x-x_2)$$

se reducirá al último término

$$P_3'''(x) = 6f[x_0,x_1,x_2,x_3]$$

Si llamamos $f_i := f(x_i)$

x_i	f_i	$f[x_i,x_{i+1}]$	$f[x_i,x_{i+1},x_{i+2}]$	$f[x_0,x_1,x_2,x_3]$
x_0	f_0			
x_1	f_1	$\frac{f_1-f_0}{h}$		
x_2	f_2	$\frac{f_2-f_1}{h}$	$\frac{f_2-2f_1+f_0}{2h^2}$	
x_3	f_3	$\frac{f_3-f_2}{h}$	$\frac{f_3-2f_2+f_1}{2h^2}$	$\frac{f_3-3f_2+3f_1-f_0}{6h^3}$

y sustituyendo volvemos a obtener la misma fórmula del enunciado.

2. En este caso, $f(x) = \ln(x)$ y, por tanto, $f'''(1) = 2$. Aplicando la fórmula con $h = 0.01$ y realizando las operaciones con Matlab, tendremos:

```
» h=0.01; » fp3=(log(1+3*h)-3*log(1+2*h)+3*log(1+h)-log(1))/h^3
           fp3 = 1.9129
```

Y por tanto, el error es $\epsilon = |2 - 1.9129| = 0.0871$.

3. Todavía existe otro modo de obtener la fórmula de derivación, con el que obtenemos directamente al término de error. La herramienta [10] que se utiliza es el desarrollo en serie de Taylor de f en el entorno de x_0.

$$\begin{aligned}
-f(x_0) &= -f(x_0)\\
3f(x_0+h) &= 3\left(f(x_0)+hf'(x_0)+\frac{h^2}{2}f''(x_0)+\frac{h^3}{6}f'''(x_0)+\frac{h^4}{4!}f^{(iv)}\left(\xi_1(x_0)\right)\right)\\
-3f(x_0+2h) &= -3\left(f(x_0)+2hf'(x_0)+\frac{(2h)^2}{2}f''(x_0)+\frac{(2h)^3}{6}f'''(x_0)+\frac{(2h)^4}{4!}f^{(iv)}\left(\xi_2(x_0)\right)\right)\\
f(x_0+3h) &= f(x_0)+3hf'(x_0)+\frac{(3h)^2}{2}f''(x_0)+\frac{(3h)^3}{6}f'''(x_0)+\frac{(3h)^4}{4!}f^{(iv)}\left(\xi_3(x_0)\right)
\end{aligned}$$

Si sumamos estas 4 ecuaciones tendremos, poniendo $f_i = f(x_0 + ih)$, que

$$f_3 - 3f_2 + 3f_1 - f_0 = h^3f'''(x_0) + \frac{h^4}{4!}\left(3f^{(iv)}\left(\xi_1(x_0)\right) - 3\cdot 2^4 f^{(iv)}\left(\xi_2(x_0)\right) + 3^4 f^{(iv)}\left(\xi_3(x_0)\right)\right)$$

Y por tanto:

$$f'''(x_0) = \frac{f_3 - 3f_2 + 3f_1 - f_0}{h^3} - \frac{h}{4!}\left(3f^{(iv)}\left(\xi_1(x_0)\right) - 48f^{(iv)}\left(\xi_2(x_0)\right) + 81f^{(iv)}\left(\xi_3(x_0)\right)\right)$$

Siendo entonces el término de error

$$\epsilon = \frac{h}{4!}\left|3f^{(iv)}\left(\xi_1(x_0)\right) - 48f^{(iv)}\left(\xi_2(x_0)\right) + 81f^{(iv)}\left(\xi_3(x_0)\right)\right|$$

[10] Ver en los capítulos 6 y 7 relativos al tratamiento numérico de las ecuaciones diferenciales el uso exhaustivo de esta herramienta.

Como hay coeficientes positivos y negativos en esta expresión, para mayorarla utilizamos la desigualdad triangular:

$$\epsilon \leq 132 \max_{x\in[x_0,x_3]} \left|f^{(iv)}(x)\right| = \frac{11h}{2}\|f^{(iv)}\|_\infty$$

El método es por tanto de orden uno. Si aplicamos esta expresión a nuestro problema, la cota de error obtenida es del orden de 0.33, mayor que el error real, como tiene que ser.

PROBLEMA 5.11 *Construcción de una fórmula de derivación.*

1. Se consideran los puntos x_0, x_0+h, x_0+2h. Se supone conocido el valor de una función f de $\mathcal{C}^\infty$ en estos tres puntos. Dar una estimación de la derivada de f en x_0 con una fórmula de interpolación del grado más alto posible. Dar su término de error.
2. Se supone ahora que los valores $f(x_0)$, $f(x_0+h)$, $f(x_0+2h)$ se obtienen con un error en valor absoluto inferior a 0.01 y se supone que las derivadas de la función f están acotadas por 0.3. Estudiar el paso óptimo para utilizar la fórmula de derivación anterior.

Solución:

1. Renombremos los puntos x_0, x_0+h y x_0+2h como x_0, x_1 y x_2, y sean f_0, f_1 y f_2 sus imágenes correspondientes a través de la función f. Se construye una parábola que se apoye en esos tres puntos. Derivando P_2 y particularizando en x_0 se tiene

$$f'(x_0) \approx P_2'(x_0) = \frac{-3f_0+4f_1-f_2}{2h}$$

Para calcular el error, derivamos el término de error de la interpolación.

$$f(x)-P_2(x) = \frac{(x-x_0)(x-x_1)(x-x_2)}{3!}f'''(\xi(x))$$

Particularizándolo para x_0 tendremos el término de error pedido

$$f'(x_0)-P_2'(x_0) = -\frac{f'''(\xi(x))}{3}h^2, \quad x_0<\xi(x)<x_0+2h$$

2. Para utilizar la fórmula anterior, supongamos que $\overline{f_0}$, $\overline{f_1}$ y $\overline{f_2}$ son los valores de trabajo de f_0, f_1 y f_2, afectados de unos errores que acotamos, según se indica en el enunciado, por

$$\left|\overline{f_0}-f_0\right| \leq 0.01, \qquad \left|\overline{f_1}-f_1\right| \leq 0.01, \qquad \left|\overline{f_2}-f_2\right| \leq 0.01$$

entonces, el cálculo que realmente hacemos utiliza los valores perturbados

$$f'(x_0) \approx \frac{-3\overline{f_0}+4\overline{f_1}-\overline{f_2}}{2h}$$

Con lo cual, el error, y sus acotaciones quedan:

$$\left|f'(x_0)-\frac{-3\overline{f_0}+4\overline{f_1}-\overline{f_2}}{2h}\right| \leq \left|f'(x_0)-\frac{-3f_0+4f_1-f_2}{2h}\right| + \left|\frac{-3f_0+4f_1-f_2}{2h}-\frac{-3\overline{f_0}+4\overline{f_1}-\overline{f_2}}{2h}\right| \leq$$
$$\leq \frac{\|f'''\|_\infty}{3}h^2 + \frac{3\left|f_0-\overline{f_0}\right|+4\left|f_1-\overline{f_1}\right|+\left|f_2-\overline{f_2}\right|}{2h} \leq \frac{0.3}{3}h^2+\frac{8\,0.01}{2h}$$

Por tanto:

$$\left|f'(x_0)-\frac{-3\overline{f_0}+4\overline{f_1}-\overline{f_2}}{2h}\right| \leq 0.1h^2+\frac{0.04}{h}$$

cuyo mínimo se obtiene para $h = 0.2^{1/3} = 0.5848$

PROBLEMA 5.12 *Estimación del paso óptimo para una fórmula de derivación.*

Se considera la función $f(x) = \ln x$. Se trata de encontrar el valor de la derivada de dicha función mediante un operador de tres puntos cuando $x = 1$. La herramienta de que disponemos es una calculadora que sólo opera con cuatro cifras decimales, sin redondeo. Se pide:

1. ¿Cuál es el paso que debe tomar el usuario de la calculadora para que al estimar el valor de dicha derivada usando una fórmula de tres puntos centrada el error total que cometa sea mínimo?
2. Construir una tabla de dicha derivada para valores decrecientes del paso, operando como si la calculadora fuese la del apartado 1. Dibujar en un gráfico el error en función del paso, verificando la cota del error que se ha minimizado en 1 y el valor del mínimo.

Solución:

1. El operador de tres puntos centrado para estimar la derivada de una función f en un punto x es:

$$f'(x) = \frac{f(x+h) - f(x-h)}{2h}$$

El error que supone para los datos el truncar en el cuarto decimal los cálculos con la calculadora será del orden de 10^{-4}. Por tanto, si llamamos $\hat{f}(x)$ a ese valor, la diferencia entre el valor real y el que se usa en el cálculo tendrá como cota:

$$\delta = |\hat{f}(x) - f(x)| \leq 10^{-4}$$

Y por tanto, el error debido a los datos ED, que resulta de calcular la derivada con los datos afectados de error, será:

$$\begin{aligned} ED = \left| \frac{f(x+h) - f(x-h)}{2h} - \frac{\hat{f}(x+h) - \hat{f}(x-h)}{2h} \right| &\leq \\ \left| \frac{f(x+h) - \hat{f}(x+h)}{2h} \right| + \left| \frac{\hat{f}(x-h) - f(x-h)}{2h} \right| &\leq \frac{\delta}{h} \end{aligned}$$

Por otro lado, estimar la derivada mediante un operador centrado tiene también un error de truncamiento inherente al método EM que vale (ver Theodor [29]):

$$EM = f'(x) - \frac{f(x+h) - f(x-h)}{2h} = -\frac{h^2}{6} f'''(\xi(x))$$

Acotemos este error en valor absoluto en el intervalo $I = [x-h, x+h]$.

$$EM = \left| \frac{h^2}{6} f'''(\xi(x)) \right| \leq \frac{h^2}{6} \max_{\xi \in I} |f'''(\xi)| = \frac{h^2}{6} \max_{\xi \in I} \left|2\xi^{-3}\right| = \frac{h^2}{3(x-h)^3}$$

Hemos quitado el valor absoluto, al ser el dominio del logaritmo los reales positivos. Como el punto donde vamos a evaluar la derivada es $x = 1$, tendremos que

$$EM \leq \frac{h^2}{3(1-h)^3}$$

Si ahora estudiamos el error total, que será función del punto, del paso h, y de los errores en los datos, acotados por δ, tendremos:

$$\begin{aligned} E(x,h,\delta) &= \left| f'(x) - \frac{\hat{f}(x+h) - \hat{f}(x-h)}{2h} \right| \leq \left| f'(x) - \frac{f(x+h) - f(x-h)}{2h} \right| + \\ &+ \left| \frac{f(x+h) - f(x-h)}{2h} - \frac{\hat{f}(x+h) - \hat{f}(x-h)}{2h} \right| = EM + ED \leq \frac{h^2}{3(x-h)^3} + \frac{\delta}{h} \end{aligned}$$

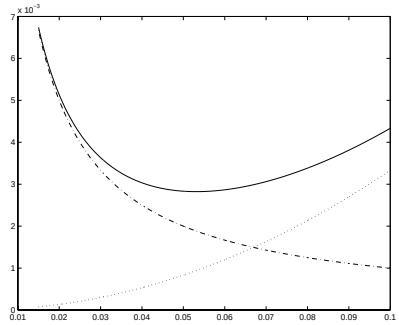


Figura 5.22: Errores por el método (trazo de puntos), por los datos (trazo punto raya) y total (trazo continuo).

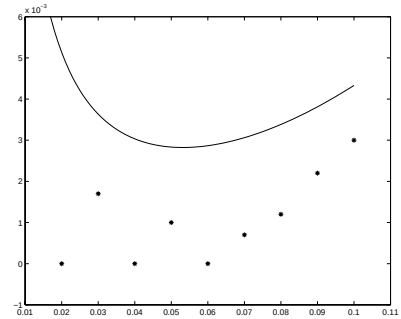


Figura 5.23: Cota del error en continuo y diferentes errores reales.

Para facilitar la minimización del error, supondremos $h << 1$, con lo que $(1-h)^3 \approx 1$, y

$$E(1,h,10^{-4}) \leq \frac{h^2}{3} + \frac{10^{-4}}{h}$$

función que representamos gráficamente en la Figura 5.22, utilizando las siguientes líneas Matlab

```
h=0.015:0.0001:0.1;
EM=h.^2/3;
ED=1e-4./h;
ET=EM+ED;
plot(h,EM,':',h,ED,'-.',h,ET);
```

y cuyo mínimo se calcula analíticamente igualando a cero su derivada. Se obtiene el paso óptimo $h = 0.0531$.

2. Se puede calcular puntos de esa gráfica utilizando valores decrecientes del paso estudiando cómo evoluciona el error, operando sólo con cuatro cifras decimales y comparar ese error con el obtenido en el apartado 1.

   En la tabla 5.2, se representa en la columna $\varepsilon(h)$ la diferencia entre el valor real 1 de la derivada en $x = 1$ y el valor obtenido con el operador centrado operando con cuatro decimales. En la Figura 5.23, se puede comprobar cómo los errores en cada uno de los puntos de la tabla no superan la cota total del error. No hay que olvidar que lo que consideramos es una mayorante del error, y que hay puntos en los que el error real es muy inferior al valor de la cota.

PROBLEMA 5.13 *Error en la fórmula de la derivada segunda.*

Sea f una función derivable en un compacto $[a,b]$ y sean $x_0 \in [a,b]$ y $h \in I\!R^+$ tales que $a \leq x_0 - h < x_0 + h \leq b$. Se utiliza una fórmula de 3 puntos para estimar la derivada segunda de f en x_0 y se tiene la siguiente expresión con su término de error correspondiente:

$$f''(x_0) = \frac{f(x_0+h) - 2f(x_0) + f(x_0-h)}{h^2} - \frac{h^2}{12} f^{(4)}(\xi)$$

Suponiendo que $|f^{(4)}(t)| \leq M \ \forall t \in [a,b]$, y que también el valor de f se obtiene con un error inferior a un número $\delta > 0$, independientemente del punto, se pide:

1. Dar una buena cota de la función $Error(h)$ correspondiente al cálculo de la segunda derivada con la fórmula que sugiere la expresión anterior.

h	$\hat{f}(1+h)$	$\hat{f}(1-h)$	$\frac{\hat{f}(1+h)-\hat{f}(1-h)}{2h}$	$\varepsilon(h)$	$\frac{h^2}{3}+\frac{10^{-4}}{h}$
0.10	0.0953	-0.1053	1.0030	0.0030	0.0043
0.09	0.0861	-0.0943	1.0022	0.0022	0.0038
0.08	0.0769	-0.0833	1.0012	0.0012	0.0034
0.07	0.0676	-0.0725	1.0007	0.0007	0.0031
0.06	0.0582	-0.0618	1.0000	0.0000	0.0029
0.05	0.0487	-0.0512	0.9990	0.0010	0.0028
0.04	0.0392	-0.0408	1.0000	0.0000	0.0030
0.03	0.0295	-0.0304	0.9983	0.0017	0.0036
0.02	0.0198	-0.0202	1.0000	0.0000	0.0051
0.01	0.0099	-0.0100	0.9900	0.0100	0.0100

Cuadro 5.2: Tabla de cotas para el problema 5.12.

2. Dibujar una gráfica de $Error(h)$.

3. Encontrar su mínimo analíticamente.

Solución:

1. Llamemos $\hat{f}(t)$ a la función f muestreada en t con la cota de error δ que eso supone, o sea

$$|f(t) - \hat{f}(t)| \leq \delta$$

La fórmula es por tanto

$$f''(x_0) \approx \frac{\hat{f}(x_0+h) - 2\hat{f}(x_0) + \hat{f}(x_0-h)}{h^2}$$

Con ello definimos

$$Error(h) := f''(x_0) - \frac{\hat{f}(x_0+h) - 2\hat{f}(x_0) + \hat{f}(x_0-h)}{h^2}$$

que podemos mayorar del modo siguiente

$$\begin{aligned}
|Error(h)| &= \left| f''(x_0) - \frac{f(x_0+h) - 2f(x_0) + f(x_0-h)}{h^2} \right. \\
&+ \left. \frac{f(x_0+h) - 2f(x_0) + f(x_0-h)}{h^2} - \frac{\hat{f}(x_0+h) - 2\hat{f}(x_0) + \hat{f}(x_0-h)}{h^2} \right| \\
&\leq \left| f''(x_0) - \frac{f(x_0+h) - 2f(x_0) + f(x_0-h)}{h^2} \right| \\
&+ \left| \frac{f(x_0+h) - 2f(x_0) + f(x_0-h)}{h^2} - \frac{\hat{f}(x_0+h) - 2\hat{f}(x_0) + \hat{f}(x_0-h)}{h^2} \right| \\
&\leq \frac{h^2}{12}|f^{(4)}(t)| + \left| \frac{f(x_0+h) - \hat{f}(x_0+h)}{h^2} \right| + 2\left| \frac{f(x_0) - \hat{f}(x_0)}{h^2} \right| + \left| \frac{f(x_0-h) - \hat{f}(x_0-h)}{h^2} \right| \\
&\leq M\frac{h^2}{12} + \frac{\delta}{h^2} + 2\frac{\delta}{h^2} + \frac{\delta}{h^2} \leq M\frac{h^2}{2} + 4\frac{\delta}{h^2}
\end{aligned}$$

Por tanto,

$$|Error(h)| \leq M\frac{h^2}{12} + 4\frac{\delta}{h^2}$$

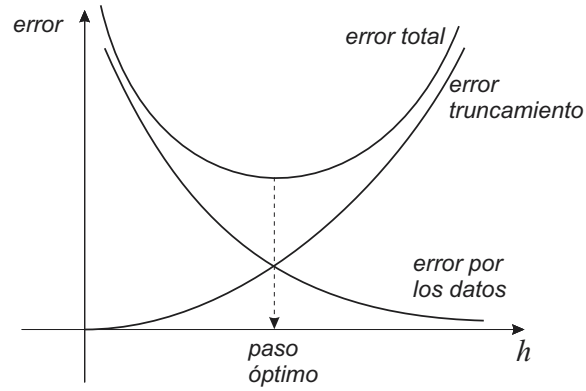


Figura 5.24: Errores en la fórmula de 3 puntos para la segunda derivada.

2. El primer sumando de la expresión anterior se corresponde con el error debido al método (error de truncamiento) mientras que el segundo es debido a los datos. Como vemos, los dos tienen distinta tendencia; el primero tiende a cero con el paso, mientras que el segundo tiende a infinito (ver Figura 5.24).

3. Definamos

$$Cota(h) \quad := \quad M\frac{h^2}{12} + 4\frac{\delta}{h^2}$$

Para encontrar su mínimo, derivamos la cota respecto al paso, e igualamos a cero. El extremo obtenido será siempre un mínimo por ser una función no acotada superiormente.

$$\frac{dCota(h)}{dh} \quad = \quad M\frac{h}{6} - 8\frac{\delta}{h^3}$$

$$h_{opt} = 2\left(\frac{3\delta}{M}\right)^{\frac{1}{4}}$$

En este caso, el valor coincide con el punto donde se cruzan las dos curvas.

CAPÍTULO 6

Problemas de valor inicial en EDO's: métodos numéricos

La integración numérica de ecuaciones diferenciales ordinarias corre paralela al desarrollo del cálculo. Aparte de ciertos trabajos preliminares de Newton y Leibniz, fue Euler [1] el que comenzó (1768-69) el estudio de la integración numérica de ecuaciones diferenciales. Las ideas subyacentes son incipientes pero muy importantes e influyentes; no las desarrolló demasiado, pero están en la base de todos los métodos actuales.

El esquema de Euler fue la herramienta que usó Cauchy para demostrar entre 1820 y 1842 el teorema de existencia y unicidad 6.1.2 del problema que lleva su nombre (sección 6.1.1), en la hipótesis de que la función f fuera continuamente diferenciable.

La convergencia de la poligonal de Euler a la curva integral de la ecuación diferencial no sólo permite demostrar la existencia de la solución, sino que también suministra un método simple, aunque poco preciso, para hallarla numéricamente.

[1] La coronación de la línea de matemáticos suizos iniciada por los hermanos Jakob y Johann Bernoulli, principales seguidores del cálculo de Leibniz, fue Leonard Euler (1707-1783), el matemático más prolífico de la historia a quien sus contemporáneos llamaron "analysis incarnate" (la encarnación del cálculo).

Euler nació en Basilea en abril de 1707. Su padre, un pastor calvinista, había estudiado matemáticas con Jakob Bernoulli e intentó en un principio que Leonard siguiera sus pasos y le sucediera como pastor; afortunadamente cometió el error de enseñarle también matemáticas y ya sabemos el resultado.

Estudió en la universidad de Basilea teología y hebreo. En matemáticas su nivel captó el interés de Johann Bernoulli, que decidió darle generosamente una hora particular de clase a la semana. Euler dedicaba el resto de la semana a preparar la próxima clase y así era capaz de plantear a su profesor muchas preguntas. Estas clases tan vivas produjeron una relación importante tanto con Johann Bernoulli como de amistad con sus hijos Daniel y Nicolás. Una vez terminados sus estudios a los diecisiete años su padre le insistió para que abandonara las matemáticas y se dedicara en exclusiva a la teología, pero desechó la idea cuando los Bernoulli le explicaron que su hijo estaba destinado a ser un gran matemático.

En el siglo XVIII las universidades no eran los principales centros de investigación. Había una tradicional hostilidad contra la ciencia con profundas raíces religiosas. El liderazgo lo tenían algunas "Academias reales" financiadas por la generosidad de los reyes. La deuda de los matemáticos es enorme con Federico el Grande de Prusia y Catalina la Grande de Rusia, que hicieron posible el progreso matemático del siglo. Las Academias de Berlín y San Petersburgo, inspiración de la sana ambición de Leibniz, fueron el eje en el que se desarrolló la creación matemática de Euler.

En 1725 viajó por primera vez a San Petersburgo siguiendo a Daniel y Nicolás Bernoulli, que le ofrecieron en principio un puesto de asociado en la sección médica de la Academia. En 1730 ejercía de profesor de física y en 1733 ya encabezaba la cátedra de matemáticas. Permaneció allí hasta 1741 y desde 1741 hasta 1766 Euler estuvo bajo la protección de Federico el Grande en la Academia de Berlín. Regresó de nuevo a San Petersburgo en 1766, donde permaneció hasta el final de su vida, bajo la tutela de la emperatriz Catalina. Se casó dos veces y tuvo trece hijos.

Aunque se quedó ciego en 1766 (ya había perdido la visión de un ojo en 1735 por un problema de tensión ocular y el segundo lo perdió progresivamente por una catarata), ello no afectó a su productividad y continuó trabajando, dictando artículos y libros a su hijo y a sus discípulos. En vida publicó 530 libros y artículos y, una vez muerto, la Academia de San Petersburgo publicó durante 47 años sus manuscritos póstumos. Según algunas fuentes el número de sus trabajos es de 886. Entre ellos debemos destacar sus grandes tratados *Introductio in analysis infinitorum* en 1748, *Institutiones calculi differentialis* en 1755 y su *Institutiones calculi integralis* entre 1768 y 1774. La sección de ecuaciones diferenciales es hoy todavía el modelo que siguen los libros de texto elementales de la materia.

Desde el punto de vista de cálculo numérico, Euler escribió al menos seis memorias sobre el cálculo elemental de π y el método de la poligonal para el cálculo aproximado de la solución de un problema de valor inicial para una ecuación diferencial ordinaria "Opera Omnia, Capítulo XII, *De Aequationun Differentio-Differentialum Integrationes per Approximationes*".

Murió a los 77 años en pleno uso de su capacidad intelectual, en septiembre de 1783.

Otra línea en la demostración de la existencia local de soluciones de ecuaciones diferenciales iniciada posiblemente por Liouville en 1837 y usada por Cauchy en su curso hacia la misma época, se basa en el método de "aproximaciones sucesivas" (sección 6.3.5). Este método no logra el interés de los analistas hasta que en 1890 E. Picard demuestra su fecundidad aplicándolo a numerosos problemas de existencia de ecuaciones funcionales de naturaleza muy distinta.

El aumento de la complejidad de los estudios en mecánica celeste y teoría del calor obligó a construir esquemas numéricos razonables de resolución de ecuaciones diferenciales.

Ya desde el principio se pudieron distinguir dos tipos distintos de métodos numéricos que seguían las líneas antes esbozadas.

- Los del tipo Euler-Cauchy-Lipschitz en los que los valores aproximados $y_{(i)}$ en los puntos x_i se calculan paso a paso mediante un proceso de avance en la variable independiente (ver la nota al pie 7) y cuyos sucesores son los métodos de Runge-Kutta (sección 6.2.3).
- Los esquemas que suministran aproximaciones sucesivas $y^j_{(i)}$, $j = 0, 1, ...$ en el punto x_i siguiendo un esquema iterativo hasta que se satisface un criterio de precisión prefijado, alrededor de cuya estrategia crecieron los métodos de Adams (1855), Moulton con su estrategia predictor-corrector (1926) y Milne (1926) (sección 6.3.1).

En el momento actual los métodos más usados en los códigos estándar son, o bien variantes de los métodos Runge-Kutta con control del paso (pares encajados), o bien métodos multipaso que son variantes de los métodos Adams y métodos BDF con paso variable y control del orden del esquema en cada paso (sección 6.3.4).

6.1. El problema de Cauchy

El problema fundamental que estudiaremos es el problema de valor inicial o problema de Cauchy.

Definición 6.1.1 *Dada una función continua $f : I \times \mathbb{R}^m \to \mathbb{R}^m$, donde I es un intervalo de $\mathbb{R}$ de extremos a y b, abierto, semiabierto o cerrado, y dados un punto $\mathbf{y}_0 \in \mathbb{R}^m$ y un punto x_0 de I, el problema de Cauchy consiste en hallar, si es posible, una función $\mathbf{y}$ definida en I con valores en $\mathbb{R}^m$, de clase $\mathcal{C}^1$ tal que*

$$\begin{cases} \mathbf{y}'(x) = f(x, \mathbf{y}(x)) & (\forall x \in I) \\ \mathbf{y}(x_0) = \mathbf{y}_0 \end{cases} \tag{6.1}$$

*La función buscada $\mathbf{y}$ es una **solución en $\mathcal{C}^1(I; \mathbb{R}^m)$ del problema de Cauchy**.*

Definición 6.1.2 *La condición que preasigna el valor $\mathbf{y}_0$ que debe tomar una solución $\mathbf{y}$ de la ecuación*

$$\mathbf{y}' = f(x, \mathbf{y}) \tag{6.2}$$

*en un punto x_0 de I se llama una **condición de Cauchy**.*

*La pareja $(x_0, \mathbf{y}_0)$ define las **condiciones iniciales** del problema de Cauchy.*

La ecuación diferencial (6.2) equivale al sistema diferencial de m ecuaciones diferenciales escalares de primer orden

$$\begin{cases} y_1' = f_1(x, y_1, \cdots y_m) \\ \quad \cdots\cdots\cdots \\ y_m' = f_m(x, y_1, \cdots y_m) \end{cases} \tag{6.3}$$

Toda ecuación diferencial de orden p en $\mathbb{R}$ escrita en forma normal [2]

$$y^{(p)} = \varphi\left(x, y, y', \cdots, y^{(p-1)}\right) \tag{6.4}$$

equivale a un sistema diferencial de primer orden de dimensión p.

En efecto, introduciendo las variables

$$y = z_1, \;\; y' = z_2, \;\; \cdots \;\; y^{(p-1)} = z_p$$

[2] Con la derivada de orden superior despejada.

se reduce (6.4) al sistema

$$\begin{cases} z_1' = z_2 \\ z_2' = z_3 \\ \quad \cdots\cdots\cdots \\ z_p' = \varphi\,(x, z_1, z_2, \cdots, z_p) \end{cases} \tag{6.5}$$

Llamando $\zeta = (z_1, z_2, \cdots, z_p) \in \mathcal{C}^1(I; I\!R^p)$ podemos escribir el sistema (6.5) en la forma vectorial

$$\zeta' = \Phi(x, \zeta) \tag{6.6}$$

con $\Phi : I \times I\!R^p \to I\!R^p$. La ecuación diferencial (6.6) es análoga a la (6.2).

Una condición de Cauchy para la ecuación (6.4) preasigna los valores de la función y de sus derivadas de orden $1, 2, ..., p-1$ en el punto x_0 de I y define simultáneamente una condición de Cauchy para la ecuación (6.6) asignando a la función $\zeta : I \to I\!R^p$ en x_0 el valor ζ_0 correspondiente.

El proceso anterior de reducción de una ecuación de orden p en $I\!R^m$ a una ecuación diferencial ordinaria de primer orden en $I\!R^p$ es práctica indispensable cuando se buscan soluciones numéricas de problemas de valor inicial [3]; como consecuencia, **restringiremos nuestro estudio a las ecuaciones diferenciales de primer orden (6.2) resueltas respecto de la primera derivada**.

6.1.1. Teoremas de existencia y unicidad de la solución del problema de Cauchy

El siguiente teorema de existencia suministra un resultado global. Si se cumplen las hipótesis de su enunciado, existe una solución del problema de Cauchy definida **en todo el espacio considerado**.

Teorema 6.1.1 ***(Picard-Lindelöff)*** *Sea $f : [a, b] \times I\!R^m \to I\!R^m$, continua que además satisface la condición de Lipschitz respecto de la variable $\mathbf{y}$ siguiente:*

$$\left|f(x, \mathbf{y}^1) - f(x, \mathbf{y}^2)\right| \leq L \cdot \left|\mathbf{y}^1 - \mathbf{y}^2\right| \tag{6.7}$$

para $x \in [a, b]$ e $\mathbf{y}^1, \mathbf{y}^2 \in I\!R^m$, siendo $L > 0$ una constante de Lipschitz.

Entonces el problema de Cauchy (6.1) tiene una solución única.

Razonando a través del teorema del valor medio se prueba que la función f satisface la condición de Lipschitz si posee derivadas parciales respecto de $\mathbf{y}$ continuas y acotadas en la banda $[a, b] \times I\!R^m$.

En el caso de las ecuaciones diferenciales escalares $I\!R^m = I\!R$ es suficiente que $\frac{\partial f}{\partial y}$ esté acotada en $[a, b] \times I\!R$ en cuyo caso una constante de Lipschitz sería

$$L = \max_{(x,y) \in [a,b] \times I\!R} \left|\frac{\partial f}{\partial y}(x, y)\right| \tag{6.8}$$

En el teorema de Picard-Lindelöff, hemos obtenido resultados válidos en todo el dominio de definición de la función f. En general nos encontramos con ecuaciones diferenciales en las que la bondad de f no permite conclusiones tan contundentes. El siguiente teorema suministra resultados locales, válidos en algún entorno del punto $(x_0, \mathbf{y}_0)$ que define las condiciones iniciales.

Teorema 6.1.2 ***(Teorema de existencia y unicidad local)*** *Sean I un intervalo abierto de $I\!R$ y U un abierto de $I\!R^m$, llamaremos Ω a $I \times U$ abierto de $I\!R^{m+1}$.*

Consideramos un punto $(x_0, \mathbf{y}_0)$ de Ω y una función continua $f : (x, \mathbf{y}) \in \Omega \to f(x, \mathbf{y}) \in I\!R^m$ que además es lipchiciana en Ω respecto de la variable $\mathbf{y}$.

Sea K el conjunto compacto de Ω, $K = [x_0 - h, x_0 + h] \times B(\mathbf{y}_0; b)$ con $h > 0$, donde $B(\mathbf{y}_0; r)$ denota la bola cerrada de $(I\!R^m; \|\cdot\|)$ de centro $\mathbf{y}_0$ y radio $r > 0$. Si $n = 1$, $B(y_0, r) = [y_0 - r, y_0 + r]$.

Llamemos $A = \{\sup |f(x, \mathbf{y})| : (x, \mathbf{y}) \in K\}$ y L a la constante de Lipschitz de f en K [4]. Entonces el problema de Cauchy en estudio tiene una solución $\mathbf{y}$ definida en el intervalo $J = [x_0 - c, x_0 + c]$ donde $c = \min\left(h, \frac{r}{A}\right)$ que es única.

[3] Ver los problemas (6.4), (6.5), (6.7) y (6.12).

[4] K es compacto y f continua en K, $A = \|f\|_\infty$ restringida a K.

Si $f(x,y)$ es una función continua y L-lipchiciana en y, en un dominio D de $\mathbb{R}^{m+1}$, diremos que f y la ecuación diferencial $y' = f(x,y)$ son **suficientemente regulares** en D.

Ejemplo 6.1.1 *Consideremos la ecuación diferencial ordinaria de tercer orden en* $\mathbb{R}$

$$y''' - \frac{3}{y''+1} + 8\sqrt{y} - |x|^5 = 0 \tag{6.9}$$

Con el cambio de notación $y = z_1,\ y' = z_2,\ y'' = z_3$ *el sistema diferencial equivalente es*

$$\begin{cases} z_1' = z_2 \\ z_2' = z_3 \\ z_3' = \dfrac{3}{z_3+1} - 8\sqrt{z_1} + |x|^5 \end{cases} \tag{6.10}$$

y su correspondiente ecuación vectorial es $\mathbf{z}' = \Phi(x,\mathbf{z})$ *con*

$$\Phi(x,\mathbf{z}) = \left(z_2, z_3, \frac{3}{z_3+1} - 8\sqrt{z_1} + |x|^5\right) \tag{6.11}$$

Estudiemos las propiedades de continuidad y derivabilidad de la función $\Phi(x,\mathbf{z})$. *Las dos primeras funciones componentes* Φ_1 *y* Φ_2 *están definidas y son de clase* $\mathcal{C}^\infty$ *para todo* $(x,\mathbf{z}) \in \mathbb{R}^4$. *Por el contrario* Φ_3 *es de clase* $\mathcal{C}^\infty$ *en el conjunto no conexo* $D = \{(x; z_1, z_2, z_3) \ :\ x > 0,\ z_1 > 0,\ z_2 > 0,\ z_3 \neq -1\}$. *La función* $\Phi(x,\mathbf{z})$ *es por tanto de clase* $\mathcal{C}^\infty$ *en* D.

El problema de Cauchy definido por la ecuación (6.9) y las condiciones iniciales $(1;\ 2,1,1) \in D$ *está correctamente propuesto. El problema de Cauchy equivalente vendrá definido por*

$$\begin{cases} \mathbf{z}' = \left(z_2, z_3, \dfrac{3}{z_3+1} - 8\sqrt{z_1} + |x|^5\right) \\ \mathbf{z}(1) = (2,1,1) \end{cases}$$

Llamemos Ω_1 *al dominio conexo* $\Omega_1 = \{(x; z_1, z_2, z_3) \ :\ x > 0,\ z_1 > 0,\ z_2 > 0,\ z_3 > -1\}$.

La matriz asociada a la diferencial parcial de Φ *respecto de* $\mathbf{z}$ *es*

$$\left(\frac{\partial \Phi_i}{\partial z_j}(x,\mathbf{z})\right) = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -\dfrac{4}{\sqrt{z_1}} & 0 & -\dfrac{3}{(z_3+1)^2} \end{pmatrix} \tag{6.12}$$

cuyas funciones componentes son continuas en Ω_1 *y, por tanto, están acotadas en todo compacto* K *contenido en ese dominio. Queda por consiguiente asegurada la condición de Lipschitz en* K *y la conclusión del teorema anterior.*

6.2. Métodos numéricos. Definiciones generales. Tipos de métodos numéricos

Se sabe que sólo en unos cuantos casos se puede expresar la solución de una ecuación diferencial ordinaria por medios analíticos y que en general es imposible resolver el problema de Cauchy aun cuando se sepa que tiene solución única, por lo que es necesario desarrollar métodos que permitan obtener aproximaciones precisas de esa solución.

Todos los métodos numéricos que discutiremos en este capítulo se basan en la **discretización** de la variable independiente x (tiempo o espacio) sustituyendo el intervalo $[a,b]$ en el que varía, por una malla finita de $N+1$ puntos o **nodos** $\{x_i\}$ con $i = 0, 1, ..., N$ que definen la **malla computacional**, en los que se evalúa la solución de modo aproximado.

La distancia entre dos nodos consecutivos de la malla $h_i = x_{i+1} - x_i$ es el **tamaño del paso** o **paso** a secas. Por comodidad, consideraremos la malla uniforme y el paso constante de modo que todos los nodos son equidistantes, aunque en la práctica computacional esto no es así.

Las variables que determinan una malla uniforme son el paso h, el número de puntos $N+1$ y el modo en que la partición de nodos se debe ajustar al intervalo $[a,b]$. Si el extremo b debe ser un nodo, la malla se define por la relación $h = \frac{b-a}{N}$ eligiendo o bien h o bien N [5].

[5] La selección del paso h es una etapa importante del diseño de un método numérico. El costo numérico asociado al método es proporcional al número N de pasos individuales, que a su vez depende de h.

Denotaremos $\mathbf{y}(x)$ la solución única del problema de Cauchy (6.1) de modo que $\mathbf{y}(x_i)$ es el valor exacto de la solución en el nodo x_i y denotaremos $\mathbf{y}_{(i)}$ su valor aproximado[6]. Denotaremos por último f_i el valor $f(x_i, \mathbf{y}_{(i)})$.

Uno de nuestros objetivos es definir estrategias que nos permitan producir una sucesión $\{\mathbf{y}_{(n)}\}$ con $n = 0, 1, ..., N$ que aproxime a la solución exacta $\mathbf{y}(x)$ en los puntos $\{x_n\}$ de la malla. Llamaremos a esa sucesión una **solución numérica** del problema de Cauchy.

El resultado de aplicar la estrategia diseñada es una ecuación en la que intervienen f, h y un cierto número de aproximaciones consecutivas $\mathbf{y}_{(n-i+1)}$ con $i = 0, 1, ..., k$.

Definición 6.2.1 *Llamaremos un **método o esquema numérico** a una ecuación que nos permita calcular secuencialmente, paso a paso, una solución numérica del problema de Cauchy*[7].

Todos los métodos numéricos que consideraremos en este capítulo se pueden escribir en la forma general

$$\mathbf{y}_{(n+k)} + \alpha_{k-1}\mathbf{y}_{(n+k-1)} + \cdots + \alpha_0\mathbf{y}_{(n)} = hF\left(x_n; \mathbf{y}_{(n+k)}, \mathbf{y}_{(n+k-1)}, ..., \mathbf{y}_{(n)}; h; f\right) \tag{6.13}$$

con $n \geq 0$ y donde los coeficientes α_i son números reales.

Se suele llamar a (6.13) el **algoritmo de progresión** del método.

Asociamos al método numérico (6.13) el polinomio

$$\rho(z) = z^k + \alpha_{k-1}z^{k-1} + \cdots + \alpha_1 z + \alpha_0 \tag{6.14}$$

que llamaremos su **primer polinomio característico** y que nos servirá como lenguaje, para expresar propiedades de consistencia y estabilidad del método.

Definición 6.2.2 *Diremos que el primer polinomio característico verifica la condición de **Dahlquist** si todas las raíces de $\rho(z) = 0$ tienen módulo menor o igual que 1 y si las de módulo 1 son simples.*

Definición 6.2.3 *El entero k se llama el **número de pasos** del método. Si $k = 1$ el método se llama **de un paso**. Si $k > 1$ se habla de un **método de k pasos o multipaso**.*

Los métodos de un paso permiten calcular $\mathbf{y}_{(n+1)}$ a partir de $\mathbf{y}_{(n)}$.

En los métodos de k pasos para hallar $\mathbf{y}_{(n+k)}$ **es necesario conocer las** k **aproximaciones anteriores** $\mathbf{y}_{(n+k-1)}, \cdots, \mathbf{y}_{(n+1)}, \mathbf{y}_{(n)}$.

Definición 6.2.4 *Si el método numérico (6.13) define $\mathbf{y}_{(n+k)}$ explícitamente se dice que es un **método explícito**.*

La estructura general de los métodos de un paso explícitos es

$$\mathbf{y}_{(n+1)} = y_{(n)} + hF\left(x_n, \mathbf{y}_{(n)}; h; f\right) \qquad n \geq 0 \tag{6.15}$$

Definición 6.2.5 *Si el método numérico (6.13) define $\mathbf{y}_{(n+k)}$ implícitamente de modo que sólo se puede calcular resolviendo una ecuación implícita, se dice que es un **método implícito**.*

En este caso, debemos incorporar al método un buscador de raíces que resuelva en cada paso la ecuación implícita, en general no lineal.

Los métodos implícitos de un paso se obtienen de (6.13) para $k = 1$

$$\mathbf{y}_{(n+1)} + \alpha_0\mathbf{y}_{(n)} = hF\left(x_n, \mathbf{y}_{(n+1)}, \mathbf{y}_{(n)}; h; f\right) \tag{6.16}$$

[6] Una vez definida la malla, es evidente que nuestra máxima aspiración es conocer de modo exacto la restricción de $\mathbf{y}(x)$ a la malla. Como ello no es posible, nos contentaremos con aproximar sus valores en los nodos.

[7] Los problemas de Cauchy modelan problemas físicos de propagación, en los que la información conocida (los valores iniciales) progresa o avanza en el tiempo o en el espacio a partir del estado inicial.

La resolución numérica de los problemas de Cauchy se hace mediante métodos de avance en los que se progresa en la variable independiente a partir del valor x_0 inicial siguiendo la dirección de avance del método. Se aproxima el valor de la solución en un cierto nodo en función de la información que se posee de algunos de los nodos a su izquierda. Una vez obtenida esa aproximación se pasa al nodo que está inmediatamente a su derecha siguiendo la dirección de avance.

Al comenzar el proceso de avance asociado a cualquier método numérico, conocemos el valor inicial $\mathbf{y}_0$, que es exacto en este caso, y que tomamos como primer elemento $\mathbf{y}_{(0)}$ de la solución numérica. Para hallar $\mathbf{y}_{(1)}$, si el método es de k pasos, necesitamos conocer los k valores aproximados anteriores. Como ello es imposible, debemos preasignar los $(k-1)$ valores iniciales $\{\mathbf{y}_{(i)}\}$, $(i = 1, ..., k-1)$ (**algoritmo de iniciación**). Una vez suministrados estos valores, el método permite calcular $\mathbf{y}_{(k)}$ y sucesivamente, paso a paso, los demás términos de una solución numérica.

Hallar esos valores iniciales adicionales no plantea ninguna dificultad seria; se suele utilizar un método de un paso, aunque los códigos basados en métodos multipaso incluyen sus propios medios para hallarlos.

Veamos algunos ejemplos asociados a ecuaciones diferenciales en $\mathbb{R}$.

Ejemplo 6.2.1

$$y_{(n+2)} + y_{(n+1)} - 2y_{(n)} = \frac{h}{4}\left[(f_{n+2} + 8f_{n+1}) + 3f_n,\right] \tag{6.17}$$

Un método de dos pasos en el que es necesario suministrar el valor inicial $y_{(1)}$ antes de empezar a calcular la solución numérica. Con $y_{(0)}$ dato del problema de Cauchy e $y_{(1)}$ se calcula $y_{(2)}$ en

$$y_{(2)} + y_{(1)} - 2y_{(0)} = \frac{h}{4}\left[(f_2 + 8f_1) + 3f_0\right] \tag{6.18}$$

El método es implícito, $y_{(2)}$ aparece en los dos miembros y ya que f será en general una función no lineal, tendremos que resolver en cada paso un sistema de ecuaciones no lineales.

Este caso es un ejemplo de método lineal multipaso[8].

Ejemplo 6.2.2

$$y_{(n+2)} - y_{(n+1)} = \frac{h}{3}\left(3f_{n+1} - 2f_n\right) \tag{6.20}$$

Otro método lineal de dos pasos pero explícito, ya que conocida la información anterior al nodo x_{n+2} se despeja directamente $y_{(n+2)}$ en cada paso y se obtiene

$$y_{(n+2)} = y_{(n+1)} + \frac{h}{3}\left(3f_{n+1} - 2f_n\right) \tag{6.21}$$

Es claro el menor costo numérico de los métodos explícitos.

Ejemplo 6.2.3

$$y_{(n+2)} - y_{(n)} = h\left[3f(x_{n+2}, y^*_{(n+2)}) + f(x_n, y_{(n)})\right], \tag{6.22}$$

donde

$$y^*_{(n+2)} - 3y_{(n+1)} + 2y_{(n)} = \frac{h}{2}\left[f(x_{n+1}, y_{(n+1)}) - 3f(x_n, y_{(n)})\right] \tag{6.23}$$

*Un método de dos pasos explícito de los llamados **predictor-corrector** en el que se combinan un método lineal multipaso explícito (6.23), **el predictor**, para hallar una estimación $y^*_{(n+2)}$ aceptable de $y_{(n+2)}$ y un método implícito (6.22), **el corrector**, que luego la refina, de modo que el método resultante es explícito.*

6.2.1. Requisitos que deben satisfacer los métodos numéricos

Para que el método numérico (6.13) sea útil desde el punto de vista numérico es necesario que verifique ciertas condiciones naturales.

C1 **La ecuación (6.13) debe tener solución única.**

Se prueba que si la función progresión F es continua y M-lipchiciana respecto de las variables $y_{(i)}$ en el sentido siguiente:

Para toda f suficientemente regular existen dos constantes positivas h_0 y M tales que

$$|F(x; u_k, u_{k-1}, ..., u_0; h; f) - F(x; v_k, v_{k-1}, ..., v_0; h; f)| \leq M \sum_{i=0,k} |u_i - v_i| \tag{6.24}$$

[8] La estructura general de los métodos multipaso lineales es

$$y_{(n+k)} + \alpha_{k-1}y_{(n+k-1)} + \cdots + \alpha_0 y_{(n)} = h\beta_k f_{n+k} + \beta_{k-1}f_{n+k-1} + \ldots + \beta_0 f_n \qquad n = 0, 1, ..., N-k \tag{6.19}$$

para todo $x \in [a,b]$, todo $|h| \leq h_0$, $u_i, v_i \in \mathbb{R}$ la ecuación (6.13) tiene solución única.

Si además cuando $f \equiv 0$

$$F(x; u_k, u_{k-1}, ..., u_0; h; 0) \equiv 0 \tag{6.25}$$

para todo $x \in [a,b]$, todo h y todos los u_i, diremos que F **es suficientemente regular** (ver el problema 6.2).

En particular, en los métodos explícitos existe solución única sin restricciones sobre h.

C2 **La solución del problema discreto debe aproximar a la del problema continuo.**

Para comparar ambas soluciones y_h e y que pertenecen a espacios vectoriales diferentes, tenemos dos alternativas

- "Interpolar" las funciones y_h, extendiendo su dominio de definición a todo $[a,b]$.
- Proyectar y sobre $\mathcal{M}_h$.

Hemos tomado este segundo camino, ya que el primero depende del tipo de interpolación y se pueden introducir procesos extraños al problema.

Definición 6.2.6 *El **error de discretización global** en la iteración n viene definido por*

$$g_n(h) = y_{(n)} - y(x_n) \tag{6.26}$$

Definición 6.2.7 *Diremos que el método numérico es convergente si*

$$\max_{n=0,1,...,N(h)} |g_n(h)| \xrightarrow[h\to 0]{} 0$$

cuando los k valores iniciales fijados en el algoritmo de iniciación $y_{(j)}$ $(j = 0, 1, ..., k-1)$ tienden al valor inicial y_0 para $h \to 0$.

El método numérico es convergente de orden $p \geq 1$ si

$$g_n(h) = O(h^p)$$

La **convergencia** del método significa que la solución del problema discreto tiende uniformemente a la del problema continuo cuando $h \to 0$.

C3 **El método numérico debe representar fielmente a la ecuación diferencial.**

Si al sustituir en (6.13) los valores aproximados $y_{(n+i)}$ de la solución del problema de Cauchy en los nodos x_{n+i} por sus valores exactos $y(x_{n+i})$ $(i = 0, 1, 2, ..., k)$ se verificara exactamente dicha ecuación, habríamos conseguido nuestro máximo objetivo.

Tomamos como magnitud del ajuste de la solución exacta al método numérico aproximado, el error de discretización local.

Definición 6.2.8 *El **error de discretización local** del método numérico de k pasos (6.13) en el punto $x_{(n+k)}$ se define por*

$$\begin{aligned} e(x_n, y(x_n); h) = &\frac{1}{h}\left[y(x_{n+k}) + \alpha_{k-1}y(x_{n+k-1}) + \ldots + \alpha_0 y(x_n) - \right. \\ &\left. - hF(x_n; y(x_{n+k}), y(x_{n+k}), ..., y(x_n); h)\right] \qquad n = 0, 1, ..., N-k \end{aligned} \tag{6.27}$$

Se espera que este error tienda a cero cuando el paso h tienda a cero.

Definición 6.2.9 *Diremos que un método numérico es consistente si [28]*

$$\max_{n=0,1,...,N(h)} |e(x_n, y(x_n); h)| \xrightarrow[h\to 0]{} 0$$

y que es consistente de orden $p \geq 1$ si

$$e(x_n, y(x_n); h) = O(h^p)$$

El orden de un método es un indicador de la velocidad de convergencia a 0 del error cuando $h \to 0$.

Una definición alternativa, muy práctica, de la consistencia de (6.13) se expresa en función del primer polinomio característico asociado a dicho método (6.14).

Definición 6.2.10 *El método numérico de k pasos (6.13) es consistente ssi*

$$\begin{aligned} &\rho(1) = 0 \\ &F\left(x_n; y(x_n), y(x_n), ..., y(x_n); 0; f\right) = \rho'(1) f(x_n, y(x_n)) \end{aligned} \tag{6.28}$$

donde se ha hecho $h = 0$ en F teniendo en cuenta que $y_{n+i} = y(x_n + ih)$.

Ejemplo 6.2.4 *Estudiemos directamente la consistencia del método de dos pasos explícito*

$$y_{(n+2)} - y_{(n+1)} = h\left(f_{n+1} - \frac{2}{3} f_n\right) \tag{6.29}$$

cuya función de progresión y error de discretización local son respectivamente

$$F\left(x_n; y(x_{n+2}), y(x_{n+1}), y(x_n); h; f\right) = f_{n+1} - \frac{2}{3} f_n \tag{6.30}$$

y

$$e(x_n, y(x_n); h) = \frac{y(x_{n+2}) - y(x_{n+1})}{h} - \left[f_{n+1} - \frac{2}{3} f_n\right] \tag{6.31}$$

Desarrollando en serie de Taylor alrededor de x_n a través del problema de Cauchy (ver el detalle en las fórmulas (6.37) y (6.38).)

$$\frac{y(x_{n+2}) - y(x_{n+1})}{h} = y'(x_n) + \frac{3h}{2} y''(x_n) + ... = f_n + \frac{3h}{2}\left[\frac{\partial f}{\partial x} + \frac{\partial f}{\partial y} f\right]_n + ...$$

y desarrollando $h \to f(x_{n+1}, y(x_{n+1})) - \frac{2}{3} f(x_n, y(x_n))$ en serie de Taylor alrededor de $h = 0$ [9].

$$F\left(x_n; y(x_{n+2}), y(x_{n+1}), y(x_n); h; f\right) = \frac{1}{3} f_n + h\left[\frac{\partial f}{\partial x} + \frac{\partial f}{\partial y} f\right]_n + ...$$

Restando ambos desarrollos término a término tendremos:

$$e(x_n, y(x_n); h) = \frac{2}{3} f_n + \frac{h}{2}\left[\frac{\partial f}{\partial x} + \frac{\partial f}{\partial y} f\right]_n + ...$$

expresión que prueba la inconsistencia del método en estudio.

Utilizando la definición alternativa (6.2.10) tenemos $\rho(z) = z^2 - z$ de donde $\rho'(1) = 1$ y haciendo $h = 0$ en la función F,

$$F\left(x_n; y(x_n); 0; f\right) = f_n - \frac{2}{3} f_n = \frac{1}{3} f_n \quad \Rightarrow \quad \frac{F\left(x_n; y(x_n); 0; f\right)}{\rho'(1)} = \frac{1}{3} f_n \neq f_n \tag{6.32}$$

C4 **El método numérico debe ser estable.**

Suponiendo que el problema de valor inicial en estudio está "bien puesto", debemos estudiar la posible inestabilidad del método numérico como resultado de las perturbaciones de la función incremento F y de los k valores iniciales $y_{(0)}, ..., y_{(k-1)}$.

Existen varias definiciones de la estabilidad; todas ellas tienen como característica común comparar la solución $\{y_{(i)}\}$; $(i = 0, 1, ..., N)$ de (6.13) y una solución perturbada.

Desde un punto de vista práctico tomaremos como definición de la estabilidad la conclusión del siguiente teorema debido a Dahlquist que facilita definitivamente el estudio de la estabilidad de los métodos multipaso.

[9] ¡No olvidar que $f_{n+1} = f(x_{n+1}, y(x_{n+1}))$ y que $x_{n+1} = x_n + h$!

Definición 6.2.11 *El método numérico (6.13), supuesto que F sea suficientemente regular, es estable si el polinomio ρ verifica la condición de Dahlquist.*

Ejemplo 6.2.5 *El método de dos pasos implícito*

$$y_{(n+2)} + y_{(n+1)} - 2y_{(n)} = h\left(\frac{1}{4}f_{n+2} + 2f_{n+1} - \frac{3}{4}f_n\right) \tag{6.33}$$

no es estable, ya que las raíces de su primer polinomio característico $\rho(z) = z^2 + z - 2$ son 1 y -2.

Implicaciones entre la consistencia, la estabilidad y la convergencia de un método numérico.

Entre todos los conceptos que hemos introducido existen diversas implicaciones que facilitan el análisis de las condiciones expuestas.

Para un método (6.13) consistente, la condición de estabilidad de Dahlquist es necesaria y suficiente para que sea convergente supuesto que F sea suficientemente regular.

Teorema 6.2.1 *Si el método (6.13) es consistente y si F es suficientemente regular, entonces el método es convergente si y sólo si se cumple la condición de Dahlquist.*

Este último resultado nos permite sustituir el estudio de la convergencia de un método numérico por el de su consistencia y estabilidad, para lo que tenemos herramientas fáciles de aplicar (ver el problema 6.2).

6.2.2. Métodos numéricos de un paso

Los métodos numéricos de un paso asociados al problema de Cauchy suficientemente regular (6.1), responden a un algoritmo del tipo

$$\begin{aligned} \mathbf{y}_{(0)} &= \mathbf{y}_0 \\ \mathbf{y}_{(n+1)} &= \mathbf{y}_{(n)} + hF\left(x_n, \mathbf{y}_{(n+1)}, \mathbf{y}_{(n)}; h; f\right) \end{aligned} \tag{6.34}$$

A la función F se le llama **función incremento** y cuando es independiente de $\mathbf{y}_{(n+1)}$ el método es explícito. Dos aspectos importantes a destacar:

- Sólo hace falta un valor inicial $\mathbf{y}_{(0)}$ para iniciar el algoritmo, que hemos tomado igual a $\mathbf{y}_0$.
- Se puede cambiar el paso sin dificultad.

En cuanto a la consistencia, estabilidad y convergencia se tienen, además de los resultados válidos para todos los métodos,

Teorema 6.2.2 *Si la función incremento F es suficientemente regular, el método es estable, en cuyo caso,*

$$\textit{Consistencia} \Leftrightarrow \textit{Convergencia}$$

Teorema 6.2.3 *El método de un paso (6.15) es consistente si y sólo si*

$$F(\xi, \eta; 0; f) = f(\xi, \eta) \tag{6.35}$$

Teorema 6.2.4 *Si F es suficientemente regular y si el método es de orden mayor o igual que p,*

$$\max_{i=0,n} |g_n(h)| = \max_{i=0,n} \left|y_{(n)} - y(x_n)\right| = O(h^p)$$

Construcción de métodos explícitos de un paso de orden dado

¿Dados h y f, de qué modo deberíamos seleccionar la función incremento $(\xi; \eta) \to F(\xi; \eta; h; f)$ para definir un método explícito de un paso de orden p? Razonemos en el caso escalar por comodidad.

De la estructura del error de discretización local

$$e(\xi, \eta, h) = \frac{y(\xi + h) - \eta}{h} - F(\xi; \eta; h; f)$$

si f es de clase suficiente, podemos desarrollar la solución y de (6.1) en serie de Taylor alrededor de $\xi \in [a,b]$

$$y(\xi+h) = y(\xi) + hy'(\xi) + \frac{h^2}{2}y''(\xi) + \cdots + \frac{h^p}{p!}y^{(p)}(\xi) + \frac{h^{p+1}}{(p+1)!}y^{(p+1)}(\xi+\zeta h) \quad \text{con} \quad 0<\zeta<1 \tag{6.36}$$

donde $y(\xi)=\eta$, e $y'(x) = f(x,y(x)) \Rightarrow y'(\xi) = f(\xi,\eta)$ y las derivadas sucesivas de la solución se obtienen a través de la ecuación diferencial del modo siguiente:

$$y''(\xi) = \frac{\partial f}{\partial x}(x,y(x))|_{x=\xi} + \frac{\partial f}{\partial y}(x,y(x))y'(x)|_{x=\xi} = \frac{\partial f}{\partial x}(\xi,\eta) + \frac{\partial f}{\partial y}(\xi,\eta)f(\xi,\eta) \tag{6.37}$$

$$\begin{aligned} y'''(\xi) = {} & \frac{\partial^2 f}{\partial x^2}(\xi,\eta) + 2\frac{\partial^2 f}{\partial x \partial y}(\xi,\eta)f(\xi,\eta) + \frac{\partial^2 f}{\partial y^2}(\xi,\eta)f(\xi,\eta) + \\ & + \frac{\partial f}{\partial y}(\xi,\eta)\left(\frac{\partial f}{\partial x}(\xi,\eta) + \frac{\partial f}{\partial y}(\xi,\eta)\right) f(\xi,\eta) \end{aligned} \tag{6.38}$$

y así sucesivamente.

Con ello, de (6.36)

$$\frac{y(\xi+h)-\eta}{h} = f(\xi,\eta) + \frac{h}{2}\left[\frac{\partial f}{\partial x} + \frac{\partial f}{\partial y}f\right](\xi,\eta) + \frac{h^2}{6}\left[\frac{\partial^2 f}{\partial x^2} + 2\frac{\partial^2 f}{\partial x \partial y}f + \frac{\partial^2 f}{\partial y^2}f + \frac{\partial f}{\partial y}\left(\frac{\partial f}{\partial x} + \frac{\partial f}{\partial y}\right)f\right](\xi,\eta) + \cdots \tag{6.39}$$

La respuesta a la pregunta formulada es ahora fácil, basta construir F con los términos que se deseen del desarrollo (6.39). Si tomamos

$$F(\xi;\eta;h;f) = f(\xi,\eta) \tag{6.40}$$

generamos el método numérico más elemental, el **esquema explícito de un paso de Euler** [10]

$$y_{(n+1)} = y_{(n)} + hf(x_n, y_{(n)}) \tag{6.41}$$

un método de orden 1, ya que

$$e(\xi,\eta,h) = \frac{h}{2}\left[\frac{\partial f}{\partial x} + \frac{\partial f}{\partial y}f\right](\xi,\eta) + \cdots$$

En el método de Euler aproximamos la solución en el entorno de x_n mediante la recta afín de ecuación

$$\ell_n(x) = y_{(n)} + (x-x_n)f(x_n, y_{(n)})$$

de modo que $y_{(n+1)} = \ell_n(x_n)$.

El método de Euler no se utiliza en la práctica por su bajo orden de precisión, pero su estudio es interesante desde un punto de vista didáctico, ver la Figura (6.1).

Ejemplo 6.2.6

$$\begin{cases} y' = y \\ y(0) = 1 \end{cases} \qquad x \in [0,5]$$

La solución analítica es $y(x) = e^x$. Tomamos un paso de $h = 0.2$. El esquema de Euler es

$$y_{(n+1)} = y_{(n)} + hy_{(n)} = y_{(n)}(1+h)$$

Como $y_{(0)} = 1$, tendremos

$$y_{(n+1)} = (1+h)^{n+1}$$

En la Tabla 6.1 y en la Figura 6.2 representamos la función error.

[10] Euler en 1768 utiliza en la evaluación numérica de la solución del problema de valor inicial (6.1) la aproximación lineal a trozos de la solución (polígonal de Euler) definido en (6.41) que luego en el siglo XIX recibió el nombre de método de Cauchy-Lipschitz, y hoy día de Cauchy-Euler o de Euler a secas.

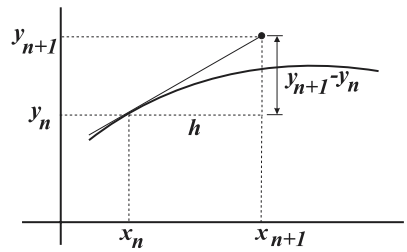


Figura 6.1: Ilustración del método de Euler.

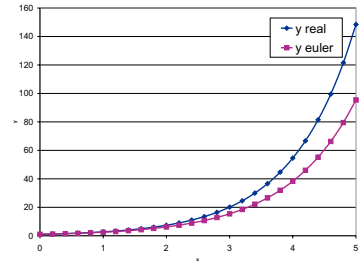


Figura 6.2: Error ejemplo esquema de Euler.

k	x	y real	y aprox	error
0	0	1	1	0
1	0,2	1,22140276	1,44	-0,21859724
2	0,4	1,4918247	1,728	-0,2361753
3	0,6	1,8221188	2,0736	-0,2514812
4	0,8	2,22554093	2,48832	-0,26277907
5	1	2,71828183	2,985984	-0,26770217
10	2	7,3890561	7,43008371	-0,04102761
15	3	20,0855369	18,4884259	1,59711103
20	4	54,59815	46,0051199	8,59303012
25	5	148,413159	114,47546	33,9376991

Cuadro 6.1: Error ejemplo esquema de Euler.

Si tomamos

$$F(\xi;\eta;h;f) = f(\xi,\eta) + \frac{h}{2}\left[\frac{\partial f}{\partial x} + \frac{\partial f}{\partial y}f\right](\xi,\eta) \tag{6.42}$$

generamos el método explícito de la **serie de Taylor de tres términos**

$$y_{(n+1)} = y_{(n)} + hf(x_n, y_{(n)}) + \frac{h^2}{2}\left[\frac{\partial f}{\partial x} + f\frac{\partial f}{\partial y}\right](x_n, y_{(n)}) \tag{6.43}$$

un método de orden 2, ya que

$$e(\xi,\eta,h) = \frac{h^2}{6}\left[\frac{\partial f}{\partial x} + \frac{\partial f}{\partial y}f\right](\xi,\eta) + \cdots$$

En este método aproximamos la solución en el entorno de x_n mediante la parábola

$$\wp_n(x) = y_{(n)} + (x - x_n)f(x_n, y_{(n)}) + \frac{(x-x_n)^2}{2}\left[\frac{\partial f}{\partial x} + f\frac{\partial f}{\partial y}\right](x_n, y_{(n)})$$

de modo que $y_{(n+1)} = \wp_n(x_n)$.

Podría parecer que hemos encontrado la clave para diseñar métodos de un paso con un orden dado. La realidad es que la necesidad de calcular y evaluar tanto f como sus derivadas parciales en cada paso limita la utilidad de las aproximaciones mediante polinomios de Taylor y obliga a buscar un compromiso entre la precisión y el coste.

6.2.3. Métodos de Runge-Kutta

Se obtienen métodos de orden alto y de estructura más simple construyendo F como una suma ponderada de varias valoraciones de la función derivada f. Esa idea es la base de los métodos de Runge-Kutta [11] [12].

$$F(\xi;\eta;h;f) = C_1 h f(\xi,\eta) + C_2 h f(\xi+\alpha_2,\eta+\beta_2) + C_3 h f(\xi+\alpha_3,\eta+\beta_3) + \cdots \tag{6.44}$$

donde los β_i son una combinación de valores de f en ciertos puntos del intervalo $[\xi,\xi+h]$. El número r de sumandos que aparecen en el segundo miembro de (6.44) se llama el **rango** del método. Los valores del rango más habituales son 2, 3 y 4. Si $r>4$ las aproximaciones requieren evaluar f en más de r puntos lo que las hacen menos deseables. Los parámetros libres de (6.44) correspondientes $C_1, C_2, C_3, ...,\ \alpha_1, \alpha_2, \alpha_3, ...,$ y $\beta_1, \beta_2, \beta_3, ...,$ se eligen de modo que (6.44) coincida hasta un cierto orden con los términos del desarrollo de Taylor (6.39).

Métodos de Runge-Kutta de orden 2

Consideremos una función incremento (6.44) de rango 2

$$F(\xi;\eta;h;f) = C_1 h f(\xi,\eta) + C_2 h f(\xi+b_1 h,\eta+b_2 h f(\xi,\eta)) \tag{6.45}$$

y determinemos las constantes C_1, C_2, b_1, b_2 de modo que el desarrollo de Taylor en potencias de h de $e(\xi,\eta,h)$ tenga el término de menor grado del mayor valor posible.

Desarrollando $h \to F(\xi;\eta;h;f)$ de (6.45) en serie de Taylor alrededor de $h=0$,

$$\begin{cases} F(\xi;\eta;0;f) = C_1 f(\xi,\eta) + C_2 f(\xi,\eta) \\ \dfrac{\partial F}{\partial h}(\xi;\eta;0;f) = C_2 b_1 \dfrac{\partial f}{\partial x}(\xi,\eta) + C_2 b_2 \dfrac{\partial f}{\partial y}(\xi,\eta) f(\xi,\eta) \\ \dfrac{\partial^2 F}{\partial h^2}(\xi;\eta;0;f) = C_2 b_1^2 \dfrac{\partial^2 f}{\partial x^2}(\xi,\eta) + 2C_2 b_1 b_2 \dfrac{\partial^2 f}{\partial x \partial y}(\xi,\eta) f(\xi,\eta) + C_2 b_2^2 \dfrac{\partial^2 f}{\partial^2 y}(\xi,\eta) f^2(\xi,\eta) \\ \cdots\cdots\cdots\cdots\cdots \end{cases}$$

y comparándolos a los términos del desarrollo de Taylor (6.39) tendremos que:

$$\begin{cases} C_1 f(\xi,\eta) + C_2 f(\xi,\eta) = f(\xi,\eta) \quad \Rightarrow \quad C_1 + C_2 = 1 \\ C_2 b_1 \dfrac{\partial f}{\partial x}(\xi,\eta) + C_2 b_2 \dfrac{\partial f}{\partial y}(\xi,\eta) f(\xi,\eta) = \\ = \dfrac{1}{2}\dfrac{\partial f}{\partial x}(\xi,\eta) + \dfrac{1}{2}\dfrac{\partial f}{\partial y}(\xi,\eta) f(\xi,\eta) \quad \Rightarrow \quad C_2 b_1 = C_2 b_2 = \dfrac{1}{2} \end{cases}$$

ya que la igualdad entre los términos relativos a h^2 de ambos desarrollos impone condiciones a f. Se obtienen así las condiciones para que el método explícito de un paso definido por (6.45) sea de orden 2. Poniendo $C_2 = a$ de modo que $C_1 = 1-a$ con $b_1 = b_2 = \frac{1}{2a}$ se define una familia uniparamétrica de métodos numéricos

$$y_{(n+1)} = y_{(n)} + h\left[(1-a) f(x_n, y_{(n)}) + a f(x_n + \frac{h}{2a}, y_{(n)} + \frac{h}{2a} f(x_n, y_{(n)}))\right] \tag{6.46}$$

con $a \in \mathbb{R}$, que se denominan **métodos de Runge-Kutta de rango** 2 **y orden** 2 o métodos (RK2) sin que de ese rango pueda haber ningún método de mayor orden.

[11] Martin Kutta (1867-1944) nació en una parte de Alemania que ahora pertenece a Polonia, Byczyna. Fue profesor en Munich, Jena, Aachen (Aquisgrán) y finalmente en Stuttgart. El método de Runge-Kutta lo presentó en 1901. Es muy famoso también por el teorema de Kutta-Joukowsky, fundamental para el cálculo de la sustentación en perfiles aerodinámicos.

[12] Carle Runge (1856, Göttingen-1927). Estudió en la universidad de Munich, donde siguió los cursos de Max Planck con el que inició una buena amistad. Ambos pasaron a Berlín en 1877. Allí estudió con Weirstrass y Kronecker. Runge trabajó en un procedimiento para la resolución numérica de ecuaciones algebraicas (polinómicas en varias variables) en las que las raíces se expresaban como series de funciones racionales de los coeficientes. Ayudó a Kutta con su método para resolver ecuaciones diferenciales ordinarias, e hizo contribuciones en el área de la física de gases y ondas. Una última curiosidad, enseñó a esquiar a Hilbert.

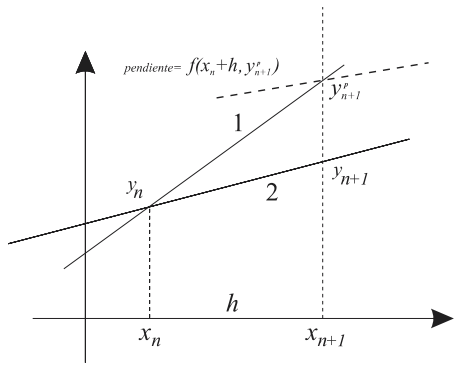


Figura 6.3: Método de Euler mejorado o método de Heun.

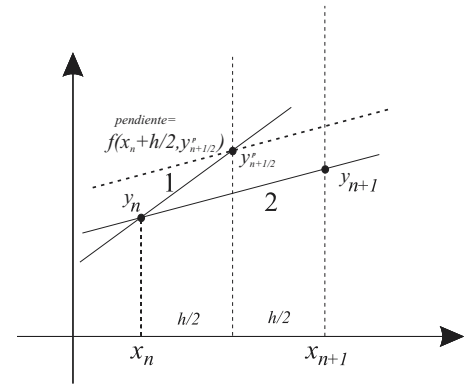


Figura 6.4: Método de Euler modificado o método del polígono.

Dando valores al parámetro se obtienen métodos Runge-Kutta de orden 2 con nombre:

- Si tomamos $a = \frac{1}{2}$ obtenemos el método del **trapecio, de Euler modificado o de Heun**

$$y_{(n+1)} = y_{(n)} + h\left(\frac{1}{2}f(x_n, y_{(n)}) + \frac{1}{2}f(x_n + h, y_{(n)} + hf(x_n, y_{(n)}))\right)$$

que también se puede escribir como un esquema predictor-corrector (ver la Figura 6.3)

$$\begin{aligned} y_{(n+1)} &= y_{(n)} + h\left(\frac{1}{2}f(x_n, y_{(n)}) + \frac{1}{2}f(x_n + h, y^p_{(n+1)})\right) \\ y^p_{(n+1)} &= y_{(n)} + hf(x_n, y_{(n)}) \end{aligned} \tag{6.47}$$

- Si tomamos $a = 1$ obtenemos el método **mejorado de la poligonal**

$$y_{(n+1)} = y_{(n)} + hf\left(x_n + \frac{1}{2}h, y^p_{(n+1/2)}\right)$$

que también se puede plantear como un esquema predictor-corrector (ver la Figura 6.4)

$$\begin{aligned} y_{(n+1)} &= y_{(n)} + hf\left(x_n + \frac{1}{2}h, y^p_{(n+1/2)}\right) \\ y^p_{(n+1/2)} &= y_{(n)} + \frac{1}{2}hf(x_n, y_{(n)}) \end{aligned} \tag{6.48}$$

Se utiliza el método de Euler para obtener una aproximación en el punto medio $y^p_{(n+1/2)}$ que utilizamos para aproximar la derivada en x_{n+1}.

- Si tomamos $a = \frac{3}{4}$ obtenemos el método de **Heun**[13] **de orden** 2

$$y_{(n+1)} = y_{(n)} + h\left[\frac{1}{4}f(x_n, y_{(n)}) + \frac{3}{4}f\left(x_n + \frac{2}{3}h, y_{(n)} + \frac{2}{3}hf(x_n, y_{(n)})\right)\right]$$

Métodos de Runge-Kutta de orden 3

Por un procedimiento análogo se establecen las condiciones de orden de los métodos de Runge-Kutta de orden superior.

[13] Se obtiene el método de Heun determinando a con la condición de que el coeficiente de h^2 en el error de discretización local sea el menor posible ([1] pág. 423).

La función incremento para los métodos de tercer orden es:

$$F(\xi, \eta; h; f) = C_1 k_1 + C_2 k_2 + C_3 k_3$$

siendo k_1, k_2, k_3, aproximaciones a la derivada en varios puntos del intervalo $[x_n, x_{n+1}]$. En este caso se tiene:

$$\begin{aligned} k_1 &= f\left(x_n, y_{(n)}\right) \\ k_2 &= f\left(x_n + b_1 h, y_{(n)} + b_1 h k_1\right) \\ k_3 &= f\left(x_n + b_2 h, y_{(n)} + s h k_2 + (b_2 - s) h k_1\right) \end{aligned} \tag{6.49}$$

Al imponer la condición de tercer orden, se obtienen varias relaciones entre los parámetros libres C_i con $i = 1, 2, 3$, b_1, b_2 y s que permiten definir todos los métodos de Runge-Kutta de orden 3. El más famoso de ellos es:

$$\begin{aligned} y_{(n+1)} &= y_{(n)} + \frac{h}{6} f\left(k_1 + 4k_2 + k_3\right) \\ k_1 &= f\left(x_n, y_{(n)}\right) \\ k_2 &= f\left(x_n + \frac{h}{2}, y_{(n)} + \frac{h}{2} k_1\right) \\ k_3 &= f\left(x_n + h, y_{(n)} + 2h k_2 - h k_1\right) \end{aligned} \tag{6.50}$$

Métodos de Runge-Kutta de orden 4

Los métodos de Runge-Kutta de orden cuatro exigen la evaluación de la derivada f en cuatro puntos para cada paso. Su esquema general es

$$y_{(n+1)} = y_{(n)} + h\left(C_1 k_1 + C_2 k_2 + C_3 k_3 + C_4 k_4\right) \tag{6.51}$$

siendo k_1, k_2, k_3, k_4, aproximaciones a la derivada en varios puntos del intervalo $[x_i, x_{i+1}]$.

El más popular hasta los años 1970 es el método de Runge-Kutta clásico, uno de los métodos de orden 4 elegidos por Kutta en 1901

$$\begin{aligned} y_{(n+1)} &= y_{(n)} + \frac{h}{6}\left(k_1 + 2k_2 + 2k_3 + k_4\right) \\ k_1 &= f\left(x_n, y_{(n)}\right) \\ k_2 &= f\left(x_n + \frac{h}{2}, y_{(n)} + \frac{h}{2} k_1\right) \\ k_3 &= f\left(x_n + \frac{h}{2}, y_{(n)} + \frac{h}{2} k_2\right) \\ k_4 &= f\left(x_n + h, y_{(n)} + h k_3\right) \end{aligned} \tag{6.52}$$

Otros métodos de Runge-Kutta

El control del tamaño del paso es una herramienta muy útil para conseguir el compromiso entre la precisión deseada y el costo numérico. Un paso pequeño facilita la precisión. Un paso grande evita cálculos innecesarios. Para decidir, es necesario tener estimaciones prácticas del error local cometido en cada paso.

Los dos procedimientos más usados son el de extrapolación y el de pares encajados. La idea detrás de los **métodos Runge-Kutta encajados** es construir fórmulas Runge-Kutta que permitan calcular simultáneamente en cada paso dos soluciones numéricas $y_{(n+1)}$ e $\hat{y}_{(n+1)}$ cuya diferencia suministra una estimación del error de la solución menos precisa, que se puede utilizar para controlar el paso.

Para facilitar los cálculos se consideran dos métodos Runge-Kutta, uno para cada aproximación, de igual rango y distintos órdenes p y $\hat{p}$ (habitualmente $\hat{p} = p - 1$ o $\hat{p} = p + 1$) y se nombra el par con un nombre seguido de $p(\hat{p})$, lo que significa que el orden de $y_{(n+1)}$ es p y el orden del estimador del error $\hat{y}_{(n+1)}$ es $\hat{p}$.

Fehlberg en 1969 construyó una pareja de métodos Runge-Kutta encajados de orden 4(5) que sustituyó en popularidad a la fórmula clásica. La diferencia de la solución $y_{(n+1)}$ calculada con el método de orden 4 y

del resultado $\hat{y}_{(n+1)}$ calculado en la fórmula de orden 5 se usa como estimación del error en el de orden 4. L. Shampine y H. Watts escribieron un código Runge-Kutta-Fehlberg de paso variable (RKF45), muy usado en los años ochenta.

Dormand y Prince en 1980 mejoran en su código DOPRI5 los métodos Fehlberg. Se trata de una pareja encajada 5(4) (ver en [14] la tabla 5.2 con los coeficientes del método) que incluye Matlab en sus códigos de resolución de ecuaciones diferenciales ordinarias con el nombre ode45 y que se considera el primer código que se debe intentar para abordar un problema nuevo. Matlab también incluye el código BS23 de Bogacki y Shampine [27], una pareja encajada de orden bajo 2(3) ode23 sugerido para problemas en los que no se requiera una precisión alta, o problemas en los que f sea discontinua.

Es conveniente comentar el código Dormand-Prince DOP853 de comportamiento espectacular en los tests numéricos que viene documentado en el Apéndice de códigos Fortran de [14] que también incluye el código DOPRI5.

6.3. Métodos lineales de k pasos

Los métodos numéricos lineales de k pasos asociados al problema de Cauchy suficientemente regular (6.1) responden a un algoritmo del tipo

$$\begin{aligned} &\mathbf{y}_{(i)} = \mathbf{y}_i \quad (i = 0, 1, ..., k-1) \qquad\qquad \text{algoritmo de iniciación} \\ &\sum_{i=0}^{k} \alpha_i \mathbf{y}_{(n+i)} = h \sum_{i=0}^{k} \beta_i f_{n+i} \quad (n = 0, 1, ..., N-k) \end{aligned} \tag{6.53}$$

Dos aspectos importantes a destacar.

◦ Es necesario definir un algoritmo de iniciación que suministre las k primeras aproximaciones $\mathbf{y}_{(i)}$, $i = 0, 1, .., k-1$ para arrancar el algoritmo. Habitualmente $\mathbf{y}_{(0)} = \mathbf{y}_0$.

◦ El cambio del paso es aquí un problema difícil.

Si $\beta_k = 0$, el método lineal de k pasos es explícito

$$\mathbf{y}_{(n+k)} = h \sum_{i=0}^{k-1} \beta_i f_{n+i} - \sum_{i=0}^{k-1} \alpha_i \mathbf{y}_{(n+i)} \tag{6.54}$$

Si $\beta_k \neq 0$, el método lineal de k pasos es implícito y se puede escribir en la forma de una iteración de punto fijo $\mathbf{y}_{(n+k)} = T(\mathbf{y}_{(n+k)})$ poniendo

$$\mathbf{y}_{(n+k)} = h\beta_k f(x_{n+k}, \mathbf{y}_{(n+k)}) + \left[h \sum_{i=0}^{k-1} \beta_i f_{n+i} - \sum_{i=0}^{k-1} \alpha_i \mathbf{y}_{(n+i)} \right] \tag{6.55}$$

donde el corchete del segundo miembro es una constante M_{n+k}, ya que todo es conocido.

Si se cumple la condición $hL|\beta_k| < 1$ el método de aproximaciones sucesivas asegura que la sucesión

$$\begin{aligned} &\mathbf{y}^0_{(n+k)} \quad \text{dado} \\ &\mathbf{y}^{i+1}_{(n+k)} = h\beta_k f(x_{n+k}, \mathbf{y}^i_{(n+k)}) + M_{n+k} \end{aligned} \tag{6.56}$$

converge a la solución única de (6.55).

Las fórmulas explícitas son más sencillas de utilizar, ya que cada paso sólo requiere formar una combinación lineal de valores previamente calculados y una evaluación de f.

Las fórmulas implícitas tienen mejores propiedades de estabilidad, alcanzan mayor orden con igual número de pasos y su error local es menor, aunque presentan el inconveniente de la resolución en cada paso del sistema lineal (6.55).

En cuanto a la consistencia, estabilidad y convergencia se tienen además de los resultados válidos para todos los métodos multipaso.

Teorema 6.3.1 *La función F asociada a los métodos lineales de k pasos es suficientemente regular si f es suficientemente regular.*

Las dos condiciones de consistencia de los métodos multipaso generales se expresan aquí de un modo más cómodo.

Teorema 6.3.2 *El método lineal de k pasos (6.53) es de orden p si y sólo si se satisfacen las siguientes condiciones:*

$$\begin{cases} \sum_{k} \alpha_i = 0 \\ \sum_{i=0}^{k} \alpha_i i^q = q \sum_{i=0}^{k} \beta_i i^{q-1} \quad para \quad q = 1, ..., p \end{cases} \tag{6.57}$$

6.3.1. Métodos Adams

Los métodos Adams son métodos lineales multipaso cuya estrategia se basa en la integración numérica [14].

Una vez conocidos los k valores aproximados iniciales $\{y_{(0)}, y_{(1)}, ..., y_{(k-1)}\}$ de la solución exacta del problema de Cauchy en los k primeros nodos $\{x_0, x_1, ..., x_{k-1}\}$ de la malla equiespaciada de paso $h > 0$, se obtiene $y_{(k)}$ mediante la fórmula numérica

$$y_{(k)} = y_{(k-1)} + \int_{x_{k-1}}^{x_k} \mathcal{P}(\zeta) d\zeta$$

donde $\mathcal{P}(\zeta)$ es un polinomio de interpolación que aproxima la función $\zeta \to f(\zeta, y(\zeta))$ en el intervalo de integración (x_{k-1}, x_k) y que según se seleccione produce métodos explícitos o implícitos.

Si $\mathcal{P}(\zeta)$ es el polinomio que interpola los k puntos $\{(x_i, f_i)\}$, $i = 0, 1, ..., k-1$, se define un método Adams explícito.

Si $\mathcal{P}(\zeta)$ es el polinomio que interpola los $(k+1)$ puntos $\{(x_i, f_i)\}$, $i = 0, 1, ..., k-1, k$, se define un método Adams implícito.

En el primer caso se integra $P(\zeta)$ **fuera del intervalo de interpolación** (x_0, x_{k-1}), ya que $\zeta \in [x_{k-1}, x_k]$ y se sabe que esa aproximación es bastante pobre.

En el segundo caso se evita esa situación añadiendo el nodo x_k al intervalo de interpolación y el punto extra (x_k, f_k) al conjunto de interpolación. A cambio, en los dos miembros de la ecuación aparecerá $y_{(k)}$ que es desconocido y el método será implícito.

Una vez conocido $y_{(k)}$ se reitera el proceso de modo análogo tanto en un caso como en otro.

Métodos Adams explícitos. Métodos de Adams-Bashforth (AB)

AB1

Se aproxima $f(\zeta, y(\zeta))$ por su valor f_n en el punto $\left(x_n, y_{(n)}\right)$, o sea, por un polinomio de grado cero, Figura 6.5. El esquema resultante es

$$y_{(n+1)} = y_{(n)} + h f_n \tag{6.58}$$

de nuevo el esquema de orden 1 explícito de Euler.

AB2

Se aproxima aquí $f(\zeta, y(\zeta))$ por la recta que se apoya en los puntos (x_n, f_n) y (x_{n-1}, f_{n-1}), Figura 6.6.

$$P(\zeta) = f_n + \frac{f_n - f_{n-1}}{x_n - x_{n-1}} (\zeta - x_n)$$

[14] John Couch Adams (Inglaterra, 1819-1892). Adams es recordado por ser codescubridor del planeta Neptuno con Leverrier. En 1841, siendo aún alumno, decidió investigar las irregularidades en el movimiento del planeta Urano por si se pudieran atribuir a la acción de un planeta desconocido. En 1845 Adams envió al director del Observatorio de Cambridge información precisa sobre la posición del nuevo planeta pero no produjo ninguna reacción y Urbain Leverrier (ver 2.12) se le adelantó publicando su predicción, que a la postre fue la que condujo al descubrimiento de Neptuno en el Observatorio de Berlín en 1847.

En el tema que nos ocupa, los métodos lineales multipaso explícitos que llevan su nombre fueron desarrollados por Adams para resolver un problema de F. Bashforth que se planteaba en una investigación sobre capilaridad. Los métodos lineales multipaso implícitos (técnica predictor-corrector) fueron usados por F. R. Moulton (1926) y W. E. Milne (1926), por lo que sus nombres aparecen junto al de Adams en el nombre de varias de dichas fórmulas aunque todas ellas son de Adams.

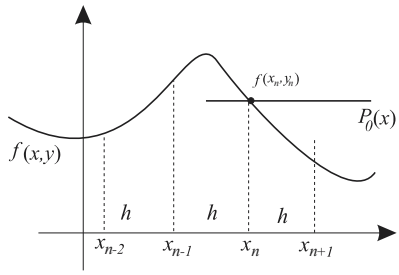


Figura 6.5: Método de Adams Bashforth de orden 1, AB1.

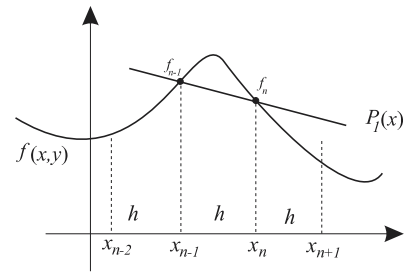


Figura 6.6: Método de Adams Bashforth de orden 2, AB2.

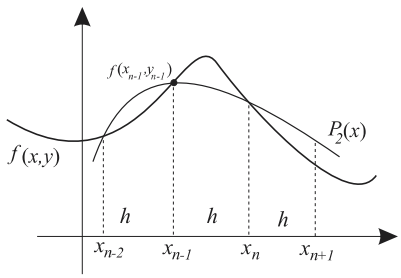


Figura 6.7: Método de Adams Bashforth de orden 3, AB3.

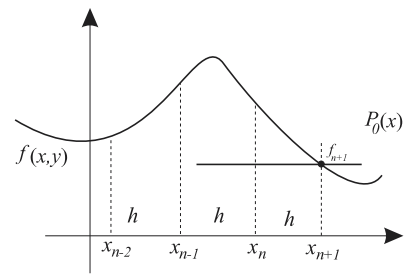


Figura 6.8: Método de Adams Moulton de orden 1, AM1.

Por tanto el esquema de Adams-Bashforth de segundo orden es:

$$y_{(n+1)} = y_{(n)} + \int_{x_n}^{x_{n+1}} P(\zeta)d\zeta = y_{(n)} + h\left(\frac{3}{2}f_n - \frac{1}{2}f_{n-1}\right) \tag{6.59}$$

AB3

Se aproxima $f\left(\zeta, y(\zeta)\right)$ por una parábola que pasa por los tres puntos (x_n, f_n), (x_{n-1}, f_{n-1}) y (x_{n-2}, f_{n-2}), Figura 6.7. Si integramos esta aproximación tendremos que:

$$y_{(n+1)} = y_{(n)} + \frac{h}{12}\left(23f_n - 16f_{n-1} + 5f_{n-2}\right) \tag{6.60}$$

AB4

Siguiendo el proceso iniciado, aproximando $f\left(\zeta, y(\zeta)\right)$ mediante un polinomio de tercer grado, obtenemos el método de Adams-Bashforth de cuarto orden, y así sucesivamente.

$$y_{(n+1)} = y_{(n)} + \frac{h}{24}\left(55f_n - 59f_{n-1} + 37f_{n-2} - 9f_{n-3}\right) \tag{6.61}$$

Se prueba que el método AB de k pasos tiene orden k. Su primer polinomio característico $\rho(z) = z^k - z^{k-1}$ satisface la condición de Dahlquist. Además de la raíz $z = 1$, tiene la raíz $z = 0$ con orden de multiplicidad $k - 1$, luego son estables y por tanto convergentes.

6.3.2. Métodos de Adams implícitos. Métodos de Adams-Moulton (AM)

AM1

Aproximamos $f\left(\zeta, y(\zeta)\right)$ por su valor en el punto (x_{n+1}, f_{n+1}), Figura 6.8. El esquema resultante es

$$y_{(n+1)} = y_{(n)} + hf\left(x_{n+1}, y_{n+1}\right) = y_n + hf_{n+1} \tag{6.62}$$

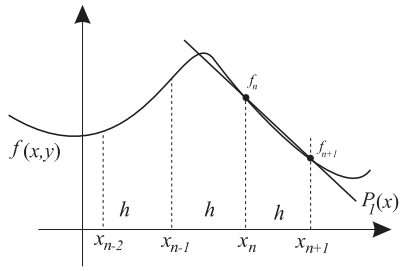


Figura 6.9: Método de Adams Moulton de orden 2, AM2.

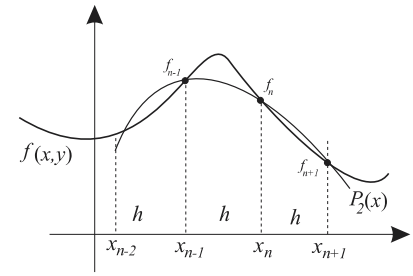


Figura 6.10: Método de Adams Moulton de orden 3, AM3.

el esquema de orden 1 de **Euler implícito o inverso**. De características de estabilidad privilegiadas que le hacen muy deseable en la integración de ecuaciones diferenciales.

AM2

Se aproxima $f(\zeta, y(\zeta))$ por una recta que se apoya en los puntos (x_n, f_n) y (x_{n+1}, f_{n+1}), Figura 6.9.

$$P(\zeta) = f_n + \frac{f_{n+1} - f_n}{x_{n+1} - x_n}(\zeta - x_n) \tag{6.63}$$

El esquema de Adams-Moulton de segundo orden es

$$y_{(n+1)} = y_{(n)} + \int_{x_i}^{x_{i+1}} P(\zeta)d\zeta = y_{(n)} + h\frac{f_n + f_{n+1}}{2} \tag{6.64}$$

la regla del trapecio que también se suele llamar método de **Crank-Nicolson**.

Estos dos primeros casos son en realidad métodos de un paso.

AM3

Se aproxima $f(\zeta, y(\zeta))$ por una parábola que se apoya en los puntos (x_{n+1}, f_{n+1}), (x_n, f_n) y (x_{n-1}, f_{n-1}), ver la Figura 6.10.

Integrando esta aproximación se obtiene

$$y_{(n+1)} = y_{(n)} + \frac{h}{12}(5f_{n+1} + 8f_n - f_{n-1}) \tag{6.65}$$

AM4

Cuando el polinomio aproximante es de tercer grado, se tiene el método de Adams-Moulton de cuarto orden, y así sucesivamente.

$$y_{(n+1)} = y_{(n)} + \frac{h}{24}(9f_{n+1} + 19f_n - 5f_{n-1} + f_{n-2}) \tag{6.66}$$

Se prueba que los métodos Adams-Moulton de k pasos son métodos de orden $k+1$ luego consistentes, que son también estables, ya que satisfacen la condición de Dahlquist y por tanto son convergentes.

Las fórmulas Adams-Moulton definen $y_{(n+1)}$ de modo implícito, luego hay que resolver en cada paso una ecuación no lineal o utilizar una estrategia predictor/corrector. La idea básica es elegir como predictor un método Adams-Bashforth y seguir el siguiente proceso.

P. Se calcula una predicción utilizando una fórmula Adams-Bashforth

$$y^p_{(n+1)} = y_{(n)} + h\left(\beta_{k-1}f_n + \cdots + \beta_1 f_{n-(k-2)} + \beta_0 f_{n-(k-1)}\right)$$

E. Se evalúa la función f en esta aproximación

$$f^p_{n+1} = f\left(x_{n+1}, y^p_{(n+1)}\right)$$

C. Se corrige la predicción aplicando la fórmula Adams-Moulton correctora

$$y_{(n+1)} = y_{(n)} + h\left(\beta_k f^p_{(n+1)} + \beta_{k-1} f_n + \cdots + \beta_1 f_{n-(k-2)} + \beta_0 f_{n-(k-1)}\right)$$

Se obtiene así la aproximación $y_{(n+1)}$ en el nodo x_{n+1}.

E. Se evalúa de nuevo la función f en ese punto

$$f_{n+1} = f\left(x_{n+1}, y_{(n+1)}\right)$$

y se reitera.

Se conoce este proceso con las siglas PECE. Otras posibilidades son PECECE (dos iteraciones de punto fijo por paso) y PEC (se utiliza f^p_{n+1} en vez de f_{n+1} en los pasos siguientes).

Si se considera un predictor/corrector de orden fijo, se puede usar una fórmula Adams-Moulton de orden p con una fórmula Adams-Bashforth de orden p o $p-1$ como predictor. Si se usa un predictor de orden $p-1$, ambos métodos utilizan los valores de f en los mismos nodos, lo que conviene en el cálculo.

Una fórmula Adams-Moulton de orden 2 con una fórmula Adams-Bashforth de orden 1 es el método predictor/corrector de Euler modificado, que definimos en (6.47). Otro ejemplo es una fórmula Adams-Moulton de orden 3 con una fórmula Adams-Bashforth de orden 2

$$\begin{aligned} y^p_{n+1} &= y_{(n)} + h\left(\frac{3}{2} f_n - \frac{1}{2} f_{n-1}\right) \\ y_{(n+1)} &= y_{(n)} + h\left(\frac{5}{12} f^p_{n+1} + \frac{8}{12} f_n - \frac{1}{12} f_{n-1}\right) \end{aligned}$$

Se comprueba que en ambos métodos se evalúa f en los mismos puntos $\left(x_n, y_{(n)}\right)$ y $\left(x_{n-1}, y_{(n-1)}\right)$.

La combinación de dos fórmulas Adams-Moulton y Adams-Bashforth de orden 4

$$\begin{aligned} y^p_{n+1} &= y_{(n)} + h\left(\frac{55}{24} f_n - \frac{59}{24} f_{n-1} + \frac{37}{24} f_{n-2} - \frac{9}{24} f_{n-3}\right) \\ y_{(n+1)} &= y_{(n)} + h\left(\frac{9}{24} f^p_{n+1} + \frac{19}{24} f_n - \frac{5}{24} f_{n-1} + \frac{1}{24} f_{n-2}\right) \end{aligned} \tag{6.67}$$

define el método **Adams-Bashforth-Moulton de orden** 4.

6.3.3. Métodos de Milne-Simpson

Integrando los dos miembros de la ecuación diferencial $y'(x) = f(x, y(x))$ entre (x_{n-r}, x_{n+1})

$$y(x_{n+1}) = y(x_{n-r}) + \int_{x_{n-r}}^{x_{n+1}} f(\zeta, y(\zeta)) d\zeta \tag{6.68}$$

con $r \geq 0$ y $n \geq r$ y sustituyendo después el integrando $f(\zeta, y(\zeta))$ por un polinomio de interpolación en algún conjunto de nodos $\{x_i\}$ se obtienen métodos numéricos similares a los Adams.

Si consideramos por ejemplo $r = 1$

$$y(x_{n+1}) = y(x_{n-1}) + \int_{x_{n-1}}^{x_{n+1}} f(\zeta, y(\zeta)) d\zeta \tag{6.69}$$

y usamos la fórmula de integración del punto medio

$$\int_a^b g(x)dx = (b-a) g\left(\frac{a+b}{2}\right) + \frac{(b-a)^3}{24} g''(\eta) \quad \text{con} \quad \eta \in [a, b] \tag{6.70}$$

obtenemos el **método del punto medio**

$$y_{(n+1)} = y_{(n-1)} + 2h f(x_n, y_{(n)}) \tag{6.71}$$

que también suele escribirse en la forma

$$y_{(n+1)} = y_{(n)} + hf(x_{n+1/2}, y_{(n+1/2)}) \tag{6.72}$$

Para $r = 1$, si usamos el polinomio de los métodos Adams-Bashforth, obtenemos las fórmulas explícitas de **Nyström**, y si usamos el polinomio interpolador de los métodos Adams-Moulton, obtenemos las fórmulas implícitas de **Milne-Simpson**. Uno de los esquemas predictor-corrector más usados es el de Milne-Simpson, que diseñó el propio Milne en 1926. El predictor es el esquema explícito de Milne de orden 3 que se obtiene para el valor $r = 3$ con el polinomio interpolador de los Adams explícitos

$$y_{(n+1)} = y_{(n-3)} + \frac{4h}{3}\left(2f\left(x_n, y_{(n)}\right) - f\left(x_{n-1}, y_{(n-1)}\right) + 2f\left(x_{n-2}, y_{(n-2)}\right)\right) \tag{6.73}$$

y el corrector es el esquema implícito de Simpson de cuarto orden (k=2), un método muy interesante,

$$y_{(n+1)} = y_{(n-1)} + \frac{h}{3}\left(f\left(x_{n+1}, y_{(n+1)}\right) + 4f\left(x_n, y_{(n)}\right) + f\left(x_{n-1}, y_{(n-1)}\right)\right) \tag{6.74}$$

6.3.4. Métodos basados en diferenciación numérica. Métodos BDF

Otra estrategia básica en la construcción de métodos numéricos de resolución de ecuaciones diferenciales ordinarias utiliza aproximaciones numéricas de las derivadas.

Construimos el polinomio $P(\zeta)$ que interpola los $(k+1)$ puntos $\{(x_i, y_{(i)})\}\quad i = n-k+1, ..., n, n+1$.

Se determina $y_{(n+1)}$ con la condición de que el polinomio de interpolación satisfaga la ecuación diferencial al menos en un punto de la malla, es decir,

$$P'(x_{n+1-r}) = f(x_{n+1-r}, y_{(n+1-r)}) \tag{6.75}$$

Para $r = 1$ se obtienen fórmulas explícitas

$$P'(x_n) = f(x_n, y_{(n)}) \tag{6.76}$$

Haciendo $r = 0$ en la condición (6.75)

$$P'(x_{n+1}) = f(x_{n+1}, y_{(n+1)}) \tag{6.77}$$

se definen las fórmulas implícitas BDF (backward differentiation formulas) que fueron introducidas por Curtis y Hirschfelder en 1952. Nosotros las hemos usado en varios problemas y hemos incluido su código Matlab en la web.

Las fórmulas BDF se utilizan desde 1971 especialmente en la integración de ecuaciones "stiff" [15].

Dando valores a k se obtienen las distintas fórmulas BDF.

k=1 $\quad y_{(n+1)} - y_{(n)} = hf_{n+1}$.

El método implícito de Euler

k=2 $\quad \frac{3}{2}y_{(n+1)} - 2y_{(n)} + \frac{1}{2}y_{(n-1)} = hf_{n+1},$

k=3 $\quad \frac{11}{6}y_{(n+1)} - 3y_{(n)} + \frac{3}{2}y_{(n-1)} - \frac{1}{3}y_{(n-2)} = hf_{n+1},$

etcétera.

Ver en [14] las fórmulas BDF hasta $k = 6$.

Un resultado importante que condiciona el modo de empleo de estos métodos es

Teorema 6.3.3 *Los métodos BDF son estables si $k \leq 6$ e inestables si $k \geq 7$.*

[15] Existen muchas definiciones del concepto de ecuación diferencial stiff ("rígida" en castellano, aunque no se utice el término).

Definición 6.3.1 *Diremos que una ecuación diferencial ordinaria es stiff si al tratar de resolverla mediante métodos numéricos estándar (métodos explícitos en general) el tamaño del paso h que es necesario para mantener la estabilidad es muy pequeño, mucho más pequeño que el necesario por consideraciones de precisión (de error local).*

6.3.5. Método de aproximaciones sucesivas. Sucesión iterante de Picard

Si integramos los dos miembros de $y' = f(x, y)$ teniendo en cuenta la condición inicial tendremos

$$\int_{x_0}^{x} y'(s)ds = \int_{x_0}^{x} f(s, y(s))ds \quad \Rightarrow \quad y(x) = y_0 + \int_{x_0}^{x} f(s, y(s))ds \tag{6.78}$$

Recíprocamente, si y continua en un intervalo I que contiene a x_0, es solución de esta ecuación integral entonces $y \in C^1(I; \mathbb{R})$ y es solución del problema de Cauchy asociado.

A partir de la equivalencia entre ambos problemas, se reduce el problema de valor inicial (6.1) a hallar una solución continua de la ecuación integral (6.78).

Si definimos el operador

$$T : y \to Ty : x \to y_0 + \int_{x_0}^{x} f(s, y(s))ds \tag{6.79}$$

es claro que la solución de la ecuación integral a la que hemos reducido el problema de Cauchy, es un punto fijo de T.

Si se cumplen las condiciones del teorema de Picard-Lindelöff, T es una aplicación contractiva lo que justifica el uso del **método de aproximaciones sucesivas** para hallar la solución. La aplicación reiterada de la transformación T comenzando con una función cualquiera de $C(I; \mathbb{R})$, proporciona una sucesión iterativa de funciones, **la sucesión de Picard**, que converge hacia la solución única $(x \to y(x))$ del problema planteado en la norma de la convergencia uniforme $\|.\|_\infty$.

Obtenemos la sucesión de Picard del modo siguiente:

Comenzamos por el estimador más evidente y simple, la función constante igual al valor inicial y_0, luego $y^{(0)}(x) = y_0$ y a continuación

$$y^{(k+1)}(x) = y_0 + \int_{x_0}^{x} f(s, y^{(k)}(s))ds$$

También en el caso del teorema de existencia y unicidad local, se puede calcular la solución local por el método de aproximaciones sucesivas. Ver en el problema (6.10) una aplicación de lo anterior.

PROBLEMAS

PROBLEMA 6.1 *Cálculo del error y estabilidad de un esquema explícito de tres pasos.*

Dado el método numérico lineal de tres pasos explícito

$$y_{(n+1)} = -\frac{3}{2}y_{(n)} + 3y_{(n-1)} - \frac{1}{2}y_{(n-2)} + 3hf(x_n, y_{(n)}) \tag{6.80}$$

con valores iniciales $y_{(0)}$, $y_{(1)}$, $y_{(2)}$ asociado a un problema de Cauchy suficientemente regular de ecuación diferencial, $y' = f(x, y)$.

1. Estudiar su consistencia y orden.
2. ¿Es el esquema (6.80) estable?

Solución:

1. Se escribe (6.80) a través de la ecuación diferencial en la forma

$$y_{(n+1)} = -\frac{3}{2}y_{(n)} + 3y_{(n-1)} - \frac{1}{2}y_{(n-2)} + 3hy'_{(n+1)}$$

Sustituyendo los valores aproximados por la solución exacta

$$y(x_{n+1}) = -\frac{3}{2}y(x_n) + 3y(x_{n-1}) - \frac{1}{2}y(x_{n-2}) + 3hy'(x_{n+1}) \tag{6.81}$$

Desarrollamos por el método de Taylor ambos miembros de la ecuación en diferencias

$$y(x_{n-1}) = y(x_n) - hy'(x_n) + \frac{h^2}{2}y''(x_n) - \frac{h^3}{6}y'''(x_n) + \frac{h^4}{24}y^{iv}(x_n) + ...$$
$$y(x_{n-2}) = y(x_n) - 2hy'(x_n) + \frac{4h^2}{2}y''(x_n) - \frac{8h^3}{6}y'''(x_n) + \frac{16h^4}{24}y^{iv}(x_n) + ...$$

luego, sustituyendo en el segundo miembro de (6.81) y operando

$$-\frac{3}{2}y(x_n) + 3y(x_{n-1}) - \frac{1}{2}y(x_{n-2}) + 3hy'(x_{n+1}) =$$
$$=y(x_n) + hy'(x_n) + \frac{h^2}{2}y''(x_n) + \frac{h^3}{6}y'''(x_n) - \frac{5h^4}{24}y^{iv}(x_n) + ...$$

De otro lado,

$$y(x_{n+1}) = y(x_n) + hy'(x_n) + \frac{h^2}{2}y''(x_n) + \frac{h^3}{6}y'''(x_n) + \frac{h^4}{24}y^{iv}(x_n) + ...$$

Restando ambos desarrollos término a término tendremos:

$$e(x_n, y(x_n); h) = \frac{h^4}{4}y^{iv}(x_n) + ...$$

El esquema es consistente de orden 4.

2. El primer polinomio característico de (6.80) es $\rho(z) = z^3 + \frac{3}{2}z^2 - 3z + \frac{1}{2} = 0$ que tiene las raíces $z_1 = 1$, $z_2 = \dfrac{-5+\sqrt{33}}{4}$ y $z_3 = \dfrac{-5-\sqrt{33}}{4}$. Ya que $\left|\dfrac{-5-\sqrt{33}}{4}\right| > 1$ no cumple la condición de Dahlquist y por tanto el método es inestable.
   El método no es convergente.

PROBLEMA 6.2 *Consistencia, convergencia y estabilidad de un método de un paso implícito.*

Se considera el método de un paso implícito

$$y_{(n+1)} = y_{(n)} + \frac{h}{2}\left(f(x_n, y_{(n)}) + f(x_{n+1}, y_{(n+1)})\right) \tag{6.82}$$

asociado a un problema de Cauchy suficientemente regular.

1. Hallar su función de progresión F y comprobar que es suficientemente regular.
2. Estudiar la consistencia, estabilidad y convergencia del método

Solución:

1. Se tiene
$$F(x_n, y_{(n+1)}, y_{(n)}; h, f) = \frac{1}{2}\left(f(x_n, y_{(n)}) + f(x_{n+1}, y_{(n+1)})\right)$$
Está claro que
$$F(x_n, y_{(n+1)}, y_{(n)}; h, 0) \equiv 0$$
y ya que f es continua y L-lipchiciana
$$|F\left(x_n; y_{(n+1)}, y_{(n)}; h; f\right) - F\left(x_n; y^*_{(n+1)}, y^*_{(n)}; h; f\right)| \leq$$
$$\leq \frac{1}{2}L|y_{(n+1)} - y^*_{(n+1)}| + \frac{1}{2}L|y_{(n)} - y^*_{(n)}| \leq$$
$$\leq \frac{1}{2}L\left(|y_{(n+1)} - y^*_{(n+1)}| + |y_{(n)} - y^*_{(n)}|\right)$$
luego F es M-lipchiciana con $M = \dfrac{1}{2}L$ y F es suficientemente regular.

2. El primer polinomio característico de (6.82) es $\rho(z) = z - 1$ que cumple trivialmente la condición de Dahlquist por tanto el método es estable.

   Además cumple las dos condiciones de consistencia (6.28)

$$\begin{cases} 1 + \alpha_0 = 0 \quad \Rightarrow \quad \alpha_0 = -1 \\ F(x_n, y(x_n), y(x_n); 0, f) = f(x_n, y(x_n)) \end{cases}$$

   El método es convergente.

PROBLEMA 6.3 *Flujo incompresible alrededor de un círculo sólido.*

El objetivo de este problema es el estudio de algunos aspectos relativos al flujo de un fluido incompresible no viscoso cuando un cuerpo sólido, en este caso un círculo, avanza en su seno.
El problema se convierte en estacionario si paramos el cuerpo sólido y dejamos que sea todo el fluido el que se mueva.
Tomando un sistema de coordenadas polares con origen en el centro O del círculo, la solución analítica del problema viene dada por las componentes radial V_r y tangencial V_θ de la velocidad en cualquier punto del fluido en función de la velocidad del flujo libre V_∞, de la densidad ρ y del radio de la circunferencia R

$$\begin{cases} V_r = -V_\infty \left(1 - \dfrac{R^3}{r^3}\right) \cos\theta \\ V_\theta = V_\infty \left(1 + \dfrac{R^3}{r^3}\right) \sin\theta \end{cases} \tag{6.83}$$

Uno de los sumandos que aparecen en la ecuación de Bernoulli se llama **presión dinámica**

$$p(r, \theta) = \frac{1}{2}\rho V_\theta^2 \tag{6.84}$$

y es la parte de la presión que corresponde a la energía cinética. La fuerza $\mathbf{F}$ debida a esta presión sobre un sector circular es el campo vectorial que se obtiene integrando dicha presión sobre el sector.

$$\mathbf{F}(r) = -R \int_{\theta_1}^{\theta_2} p(r, \theta)\, \mathbf{n}\, d\theta \tag{6.85}$$

Supuesto que $R = 1\ m$, $V_\infty = 1\ m/sg$, y $\rho = 2\ kg/m^3$.

1. Evaluar la componente horizontal de $\mathbf{F}$ sobre el primer cuadrante de la circunferencia mediante el método compuesto de los trapecios. Se considerará un soporte formado por cuatro puntos equidistantes comprendidos en el intervalo $[0, \pi/2]$, que incluya los extremos del intervalo.

2. Dar una cota del error cometido.

3. Hallar el error real integrando directamente la presión.

4. Se deja una partícula en el instante $t = 0$ en el punto de coordenadas cartesianas $(2, 1/4)$ [16]. Hallar su posición aproximada en el instante $t = 1$, usando un esquema euleriano con un paso temporal $h = 0.5\ sg$. Se exige trabajar al menos con tres cifras decimales.

Solución:

1. De (6.84) y (6.83)

$$p(r, \theta) = \frac{1}{2}\rho V_\theta^2 = \frac{1}{2}\rho V_\infty^2 \left(1 + \frac{R^3}{r^3}\right)^2 \sin^2\theta \quad \Rightarrow \quad p(R, \theta) = 4\sin^2\theta$$

   Los datos relativos al soporte equiespaciado de cuatro puntos son

[16] El sistema de referencia cartesiano tiene origen en O.

i	θ_i	p_i	$\mathbf{n}_i$
0	0	0	$(1,0)$
1	$\pi/6$	1	$(\sqrt{3}/2, 1/2)$
2	$\pi/3$	3	$(1/2, \sqrt{3}/2)$
2	$\pi/2$	4	$(0,1)$

De (6.85)

$$\mathbf{F} = -R\int_{\theta_1}^{\theta_2} p\ \mathbf{n}\ d\theta = -R\int_0^{\pi/2} p\ \left(n_x\mathbf{e}_1 + n_y\mathbf{e}_2\right)\ d\theta$$

y para $R = 1$

$$\mathbf{F} = \left[\left(-\int_0^{\pi/2} p\ n_x\ d\theta\right)\mathbf{e}_1 + \left(-\int_0^{\pi/2} p\ n_y\ d\theta\right)\mathbf{e}_2\right]$$

de donde la fuerza horizontal

$$F_x = -\int_0^{\pi/2} p\ n_x\ d\theta$$

y utilizando el método de los trapecios con $h = \pi/6$

$$F_x \approx -\frac{\pi}{6}\left[\frac{0.1}{2} + \frac{\sqrt{3}}{2} + \frac{3}{2} + \frac{4.0}{2}\right] = -\frac{(3+\sqrt{3})\pi}{12} \approx -1.2388\ \text{Nw/m}$$

Puede llamar la atención la extraña unidad (Nw/m) en la que se expresa esa fuerza, pero el fenómeno es debido a la dimensión 2D en la que se considera este problema. Si estuviéramos en 3D, al integrar el espesor se perdería la longitud del denominador.

2. En el método compuesto de los trapecios se tiene la estimación

$$R(f) \leq \frac{(b-a)^3}{12\cdot 3^2}\max|f''(\theta)|$$

en donde $a = 0$, $b = \pi/2$ y $f(\theta) = p(R,\theta)n_x(\theta) = 4\sin^2\theta\cos\theta$ luego

$$\max_{\theta\in[0,\pi/2]}|f''(\theta)| = 9.5046$$

$$R(f) \leq \frac{(\pi/2)^3}{12\cdot 3^2}9.5046 = 0.34109$$

3. Directamente

$$F_x = -\int_0^{\pi/2} p\ n_x\ d\theta = -\int_0^{\pi/2} p\ 4\sin^2\theta\cos\theta d\theta = -\frac{4}{3}$$

y

$$\left|F_{x\,real} - F_{x\,trapecios}\right| = |1.3333 - 1.2388| \approx 0.1 < 0.34109$$

4. Si nos fijamos en una partícula, su posición y velocidad son función del tiempo. Si nos fijamos en un punto, la velocidad correspondiente es siempre la misma. Tenemos el sistema de ecuaciones de primer orden

$$\begin{cases} x'(t) = V_x(x,y) \\ y'(t) = V_y(x,y) \end{cases} \quad \text{donde} \quad \begin{cases} V_x(x,y) = V_r\cos\theta - V_\theta\sin\theta \\ V_y(x,y) = V_r\sin\theta + V_\theta\cos\theta \end{cases}$$

y $\mathbf{x}_{(0)} = (2, 1/4)$, $r_{(0)} = \sqrt{2^2 + (1/4)^2} = 2.016$, $\theta_{(0)} = \arctan(1/8) = 7.125°$, $V_{x0} = -0.882$ y $V_{y_0} = 0.0299$.

Con todo ello damos el primer paso del esquema explícito de Euler

$$\mathbf{x}_{(1)} = \mathbf{x}_{(0)} + h\mathbf{V}_0 =$$

$$\begin{pmatrix} x \\ y \end{pmatrix}_{(1)} = \begin{pmatrix} 2 \\ 1/4 \end{pmatrix} + 0.5 \begin{pmatrix} -0.882 \\ 0.0299 \end{pmatrix} = \begin{pmatrix} 1.559 \\ 0.265 \end{pmatrix}$$

de donde $r_1 = \sqrt{1.559^2 + 0.625^2} = 1.581$ y $\theta_1 = \text{arc tan}\dfrac{0.265}{1.559} = 9.647°$, $V_{x1} = -0.761$ y $V_{y_1} = 0.0837$.

Reiterando el esquema

$$\mathbf{x}_{(2)} = \mathbf{x}_{(1)} + h\mathbf{V}_1 =$$

$$\begin{pmatrix} x \\ y \end{pmatrix}_{(2)} = \begin{pmatrix} 1.559 \\ 0.265 \end{pmatrix} + 0.5 \begin{pmatrix} -0.761 \\ 0.0837 \end{pmatrix} = \begin{pmatrix} 1.1785 \\ 0.3068 \end{pmatrix}$$

Resultado final del esquema.

PROBLEMA 6.4 *Péndulo amortiguado. Crank-Nicolson.*

Se considera el problema del péndulo amortiguado:

$$m \cdot l \cdot \theta'' = -k_1 \cdot l \cdot \theta' - m \cdot g \cdot \sin(\theta)$$

donde

l longitud del péndulo (= 10 m).
θ ángulo que forma la cuerda con la vertical.
g aceleración de la gravedad (≈ 10 m/sg^2).
m masa de la bola (= 20 kg).
k_1 coeficiente de amortiguamiento del medio (= 4 kg/sg).

Se deja caer el péndulo con la cuerda formando 30 grados con la vertical. Se toma un paso temporal $h = 0.2$ segundos. Se da el valor del ángulo y la velocidad angular en el instante 0.2 segundos, que quizás sean necesarios como datos en el algoritmo de inicio de algún método multipaso.

n	t_n	θ_n(rad)	θ'_n (rad/sg)
1	0.2	0.513732	-0.097446

Se pide:

1. Estimar la posición en el instante 0.4 segundos con un esquema predictor-corrector PECE, en el que el esquema de Crank-Nicolson será el corrector y el esquema predictor se elegirá del orden adecuado. Razonar si el resultado obtenido es coherente con la física del problema.

2. Utilizando aquí solamente el esquema implícito de Crank-Nicolson y suponiendo que estamos en el instante inicial 0.0 segundos, estimar el valor correspondiente a 0.2 segundos. Convertir el problema no lineal planteado en un problema inverso no lineal de una variable (la derivada del ángulo) que se pondrá en forma de una ecuación de punto fijo.

   Demostrar la convergencia del método de aproximaciones sucesivas asociado.

   Dar varios pasos en ese esquema hasta obtener valores similares a los que se dan como datos en la introducción y que servirán como referencia para validar los resultados.

   Razonar si el método es convergente en todos los demás pasos temporales.

Solución:

1. Tras sustituir los datos del enunciado

$$\theta'' = -\frac{\theta'}{5} - \sin\theta \tag{6.86}$$

Esta ecuación de segundo orden se reduce al sistema de dos ecuaciones de primer orden

$$\begin{cases} \theta' = p \\ p' = -\dfrac{p}{5} - \sin\theta \end{cases} \tag{6.87}$$

Dado que el corrector es el esquema implícito de segundo orden AM2, el predictor ha de ser un multipaso explícito de segundo orden, o sea, el AB2

$$y_{(n+1)} = y_{(n)} + h\left(\frac{3}{2}f_n - \frac{1}{2}f_{n-1}\right)$$

Para el algoritmo de inicio del predictor necesitamos un punto adicional que se da en el enunciado, y los valores de la segunda derivada del ángulo que calculamos con la ecuación (6.86). Así, tenemos la tabla

n	t_n	θ_n(rad)	θ'_n (rad/sg)	θ''_n (rad/sg^2)
0	0.0	$\pi/6$	0.0	-0.5
1	0.2	0.513732	-0.097446	-0.471942

Utilicemos el predictor AB2:

$$\begin{aligned}\begin{pmatrix}\theta\\ p\end{pmatrix}^{p}_{(2)} &= \begin{pmatrix}\theta\\ p\end{pmatrix}_{(1)} + h\left(\frac{3}{2}\begin{pmatrix}\theta'\\ p'\end{pmatrix}_1 - \frac{1}{2}\begin{pmatrix}\theta'\\ p'\end{pmatrix}_0\right)\\ &= \begin{pmatrix}0.513732\\ -0.097446\end{pmatrix} + h\left(\frac{3}{2}\begin{pmatrix}-0.097446\\ -0.471942\end{pmatrix} - \frac{1}{2}\begin{pmatrix}0.0\\ -0.5\end{pmatrix}\right)\\ &= \begin{pmatrix}0.484498\\ -0.189028\end{pmatrix}\end{aligned}$$

Para utilizar el corrector necesitamos las derivadas correspondientes a estos valores del predictor, que calculamos en (6.87)

$${\theta'}_{(2)}{}^{p} = -0.189028; \qquad {p'}_{(2)}{}^{p} = -0.427959$$

Apliquemos el corrector:

$$\begin{aligned}\begin{pmatrix}\theta\\ p\end{pmatrix}_{(2)} &= \begin{pmatrix}\theta\\ p\end{pmatrix}_{(1)} + \frac{h}{2}\left(\begin{pmatrix}\theta'\\ p'\end{pmatrix}^{p}_{(2)} + \begin{pmatrix}\theta'\\ p'\end{pmatrix}_1\right)\\ &= \begin{pmatrix}0.513732\\ -0.097446\end{pmatrix} + \frac{0.2}{2}\left(\begin{pmatrix}-0.189028\\ -0.427959\end{pmatrix} + \begin{pmatrix}-0.097446\\ -0.510920\end{pmatrix}\right) = \begin{pmatrix}0.485084\\ -0.187436\end{pmatrix}\end{aligned}$$

Completemos la tabla con estos valores.

n	t_n	θ_n(rad)	θ'_n (rad/sg)	θ''_n (rad/sg2)
0	0.0	0.523598	0.0	-0.5
1	0.2	0.513732	-0.097446	-0.471942
2	0.2	0.485084	-0.187436	

Como era de esperar, el ángulo disminuye y la velocidad angular aumenta.

2. Trabajemos ahora sólo con el esquema implícito de Crank-Nicolson

$$\begin{aligned}\begin{pmatrix}\theta\\ p\end{pmatrix}_{(n+1)} &= \begin{pmatrix}\theta\\ p\end{pmatrix}_{(n)} + \frac{h}{2}\left(\begin{pmatrix}\theta'\\ p'\end{pmatrix}_n + \begin{pmatrix}\theta'\\ p'\end{pmatrix}_{n+1}\right)\\ &= \begin{pmatrix}\theta\\ p\end{pmatrix}_{(n)} + \frac{h}{2}\left(\begin{pmatrix}\theta'\\ p'\end{pmatrix}_n + \begin{pmatrix}p\\ -\frac{p}{5} - \sin\theta\end{pmatrix}_{n+1}\right)\end{aligned}$$

Si consideramos independientemente estas ecuaciones

$$\begin{aligned}\theta_{(n+1)} &= \theta_{(n)} + \frac{h}{2}\left(\theta'_n + p_{n+1}\right)\\ p_{(n+1)} &= p_{(n)} + \frac{h}{2}\left(p'_n - \frac{p^{n+1}}{5} - \sin\theta_{n+1}\right)\end{aligned}$$

entrando con la primera en la segunda

$$p_{(n+1)} = p_{(n)} + \frac{h}{2}\left(p'_n - \frac{p_{(n+1)}}{5} - \sin\left(\theta_n + \frac{h}{2}\left(\theta'_n + p_{(n+1)}\right)\right)\right)$$

en la que todo es conocido excepto $p_{(n+1)}$ y que ya tenemos escrita en la forma de una ecuación de punto fijo

$$p_{(n+1)} = T\left(p_{(n+1)}\right)$$

Estudiemos T' en $\mathbb{R}$.

$$\left|T'(p_{(n+1)})\right| = \left|\frac{h}{2}\left(-\frac{1}{5} - \frac{h}{2}\cos\left(\theta_n + \frac{h}{2}\left(\theta'_n + p_{(n+1)}\right)\right)\right)\right| \leq \frac{h}{2}\left(\frac{1}{5} + \frac{h}{2}\right) = 0.1(0.2 + 0.1) = 0.03 < 1$$

La aplicación T es contractiva y el método de aproximaciones sucesivas convergente independientemente del instante temporal considerado.

Iteramos tomando $p^{(0)} = 0.0$ como estimador inicial

n	0	1	2	3
$p_{(n)}$	0.0	-0.10	-0.0971	-0.0972

Como vemos, nos acercamos al valor real -0.0974.

PROBLEMA 6.5 *Péndulo amortiguado. Milne-Simpson.*

Se considera el mismo péndulo amortiguado que hemos considerado en el problema 6.4

$$m \cdot l \cdot \theta'' = -k_1 \cdot l \cdot \theta' - m \cdot g \cdot \sin(\theta)$$

siendo ahora sus características

l longitud del péndulo (= 5 m).

m masa de la bola (= 1 kg).

k_1 coeficiente de amortiguamiento del medio (= 2 kg/sg).

Se deja caer el péndulo con la cuerda formando 30 grados con la vertical. Se toma un paso temporal $h = 0.2$ sg. Se da el valor del ángulo y la velocidad angular en los instantes 0.4 y 0.6 segundos, por si fueran necesarios en los cálculos.

n	t_n	θ_n(rad)	θ'_n (rad/sg)
2	0.4	0.462888	-0.262935
3	0.6	0.404568	-0.314516

Se pide:

1. Estimar la posición en los instantes 0.2 y 0.8 segundos con el esquema predictor-corrector PECE, de Milne-Simpson (MS). En el caso de que se necesite algún punto adicional para arrancar el esquema de MS, éste se obtendrá mediante un esquema de orden adecuado. En este sentido se considerará que MS tiene orden 3.

2. Utilizando aquí solamente el esquema implícito de Simpson y suponiendo que el instante inicial es 0.2 segundos, estimar el valor correspondiente a 0.4 segundos. Plantear el problema inverso no lineal resultante como la resolución de una ecuación de punto fijo y resolverlo previa demostración de su convergencia.

   ¿Es el método convergente independientemente del instante inicial?

Solución:

1. Tras sustituir los datos del enunciado

$$\theta'' = -2\theta' - 2\sin\theta \tag{6.88}$$

Esta ecuación de segundo orden se reduce mediante el cambio de notación $z_1 = \theta$, $z_2 = \theta'$ al sistema de dos ecuaciones de primer orden

$$\begin{cases} z_1' = z_2 \\ z_2' = -2z_2 - \sin z_1 \end{cases} \Rightarrow \begin{pmatrix} z_1' \\ z_2' \end{pmatrix} = f\left(x, \begin{pmatrix} z_1 \\ z_2 \end{pmatrix}\right) \tag{6.89}$$

Recordemos que el esquema predictor-corrector de Milne-Simpson se compone del predictor de Milne explícito de orden 3

$$\mathbf{z}^p_{(n+1)} = \mathbf{z}_{(n-3)} + \frac{4h}{3}\left(2f_n - f_{n-1} + 2f_{n-2}\right) \tag{6.90}$$

y del corrector implícito de Simpson de tercer orden

$$\mathbf{z}_{(n+1)} = \mathbf{z}_{(n-1)} + \frac{h}{3}\left({f_{n+1}}^p + 4f_n + f_{n-1}\right) \tag{6.91}$$

Construyamos una tabla con los datos disponibles y veamos qué datos necesitamos en el algoritmo de inicio del método

n	t_n	$(z_1)_{(n)}$	$(z_1)'_{(n)}$	$(z_2)_{(n)}$	$(z_2)'_{(n)}$
0	0	$\pi/6$	0	0	-1
1	0.2				
2	0.4	0.462888	-0.262935	-0.262935	-0.367198
3	0.6	0.404568	-0.314516	-0.314516	-0.15821
4	0.8				

Para completar la última columna hemos usado la segunda ecuación $z_2' = -2z_2 - \sin z_1$ de (6.89).
Como nos piden $\mathbf{z}_{(4)} = ((z_1)_{(4)}, (z_2)_{(4)})$ si usamos el predictor (6.90) con $n = 3$

$$\mathbf{z}^p_{(4)} = \mathbf{z}_{(0)} + \frac{4h}{3}\left(2f_3 - f_2 + 2f_1\right)$$

conocemos todos los datos menos los correspondientes a f_1 que podemos calcular utilizando por ejemplo el método $RK3$ de igual orden 3 que el MS [17]

$$\begin{aligned} \mathbf{z}_{(n+1)} &= \mathbf{z}_{(n)} + \frac{h}{6} f\left(k_1 + 4k_2 + k_3\right) \\ k_1 &= f\left(t_n, \mathbf{z}_{(n)}\right) \\ k_2 &= f\left(t_n + \frac{h}{2}, \mathbf{z}_{(n)} + \frac{h}{2}k_1\right) \\ k_3 &= f\left(t_n + h, \mathbf{z}_{(n)} + 2hk_2 - hk_1\right) \end{aligned}$$

Se tiene sucesivamente

$$k_1 = f\left(t_0, \mathbf{z}_{(0)}\right) = \begin{pmatrix} z_2 \\ -2z_2 - \sin z_1 \end{pmatrix}_{(0)} = \begin{pmatrix} z_1' \\ z_2' \end{pmatrix}_{(0)} = \begin{pmatrix} 0 \\ -1 \end{pmatrix}$$

$$k_2 = f\left(t_0 + h/2, \mathbf{z}_{(0)} + \frac{h}{2}k_1\right) = \begin{pmatrix} -0.1 \\ -0.8 \end{pmatrix}$$

[17] Para obtener los datos que se presentan en el enunciado hemos usado $RK3$.

$$k_3 = f\left(t_0 + h, \mathbf{z}_{(0)} + 2hk_2 - hk_1\right) = \begin{pmatrix} -0.12 \\ -0.6899365 \end{pmatrix}$$

Por tanto

$$\begin{pmatrix} z_1 \\ z_2 \end{pmatrix}_{(1)} = \begin{pmatrix} z_1 \\ z_2 \end{pmatrix}_{(0)} + \frac{0.2}{6}\left(\begin{pmatrix} 0 \\ -1 \end{pmatrix} + 4\begin{pmatrix} -0.1 \\ -0.8 \end{pmatrix} + \begin{pmatrix} -0.12 \\ -0.6899365 \end{pmatrix}\right) = \begin{pmatrix} 0.506265 \\ -0.162998 \end{pmatrix}$$

y $(z_2')_{(1)} = -2(z_2)_{(1)} - \sin(z_1)_{(1)} = -0.643833$, lo que completa la segunda fila de nuestra tabla

n	t_n	$(z_1)_{(n)}$	$(z_1)'_{(n)}$	$(z_2)_{(n)}$	$(z_2)'_{(n)}$
1	0.2	0.506265	-0.162998	-0.162998	-0.643833

2. Tenemos todos los datos que necesitamos para el algoritmo de inicio del predictor

$$\begin{pmatrix} z_1 \\ z_2 \end{pmatrix}^p_{(4)} = \begin{pmatrix} z_1 \\ z_2 \end{pmatrix}_{(0)} + \frac{4h}{3}\left(2\begin{pmatrix} z_1' \\ z_2' \end{pmatrix}_3 - \begin{pmatrix} z_1' \\ z_2' \end{pmatrix}_2 + 2\begin{pmatrix} z_1' \\ z_2' \end{pmatrix}_1\right) =$$

$$= \begin{pmatrix} \pi/6 \\ 0 \end{pmatrix} + \frac{4 \cdot 0,2}{3}\left(2\begin{pmatrix} -0.314516 \\ -0.15821 \end{pmatrix} - \begin{pmatrix} -0.262935 \\ -0.367198 \end{pmatrix} + 2\begin{pmatrix} -0.162998 \\ -0.643833 \end{pmatrix}_1\right) =$$

$$= \begin{pmatrix} -0.339041 \\ -0.3298337 \end{pmatrix} \quad \Rightarrow \quad \begin{pmatrix} z_1' \\ z_2' \end{pmatrix}^p_{(4)} = \begin{pmatrix} -0.329837 \\ -0.005491 \end{pmatrix}$$

Ya tenemos todos los datos que necesitamos para aplicar el corrector (6.91).

$$\begin{pmatrix} z_1 \\ z_2 \end{pmatrix}_{(4)} = \begin{pmatrix} z_1 \\ z_2 \end{pmatrix}_{(2)} + \frac{h}{3}\left(\begin{pmatrix} z_1' \\ z_2' \end{pmatrix}^p_{(4)} + 4\begin{pmatrix} z_1' \\ z_2' \end{pmatrix}_3 + \begin{pmatrix} z_1' \\ z_2' \end{pmatrix}_2\right) =$$

$$= \begin{pmatrix} 0.462888 \\ -0.262935 \end{pmatrix} + \frac{0,2}{3}\left(\begin{pmatrix} -0.329837 \\ -0.005491 \end{pmatrix} + 4\begin{pmatrix} -0.314516 \\ -0.158210 \end{pmatrix} + \begin{pmatrix} -0.262935 \\ -0.367198 \end{pmatrix}_1\right) \Rightarrow$$

$$\begin{pmatrix} z_1 \\ z_2 \end{pmatrix}_{(4)} = \begin{pmatrix} -0.339499 \\ -0.329970 \end{pmatrix}$$

3. Trabajemos ahora sólo con el esquema implícito de Simpson (6.91) para el caso $n = 2$

$$\begin{pmatrix} z_1 \\ z_2 \end{pmatrix}_{(2)} = \begin{pmatrix} z_1 \\ z_2 \end{pmatrix}_{(0)} + \frac{h}{3}\left(\begin{pmatrix} z_1' \\ z_2' \end{pmatrix}_{(2)} + 4\begin{pmatrix} z_1' \\ z_2' \end{pmatrix}_1 + \begin{pmatrix} z_1' \\ z_2' \end{pmatrix}_0\right) =$$

$$= \begin{pmatrix} z_1 \\ z_2 \end{pmatrix}_{(0)} + \frac{h}{3}\left(\begin{pmatrix} z_2 \\ 2z_2 - 2\sin z_1 \end{pmatrix}_{(2)} + 4\begin{pmatrix} z_1' \\ z_2' \end{pmatrix}_1 + \begin{pmatrix} z_1' \\ z_2' \end{pmatrix}_0\right) =$$

Todo es conocido en esta ecuación menos $\begin{pmatrix} z_1 \\ z_2 \end{pmatrix}_{(2)}$.

Como la tenemos ya escrita en la forma de una ecuación no lineal de punto fijo, aplicaremos el método de aproximaciones sucesivas como sugiere el enunciado dejando como ejercicio estudiar si se cumplen las condiciones suficientes del teorema de la aplicación contractiva (Capítulo 1, teorema 1.4.2).

$$\begin{pmatrix} z_1 \\ z_2 \end{pmatrix}^{i+1}_{(2)} = T\left(\begin{pmatrix} z_1 \\ z_2 \end{pmatrix}^i_{(2)}\right)$$

Si suponemos que

$$\left(\begin{array}{c} z_1 \\ z_2 \end{array} \right)^0_{(2)} = \left(\begin{array}{c} z_1 \\ z_2 \end{array} \right)_{(1)}$$

entrando en T,

$$\left(\begin{array}{c} z_1 \\ z_2 \end{array} \right)^1_{(2)} = \left(\begin{array}{c} 0.469266 \\ -0.281278 \end{array} \right)$$

PROBLEMA 6.6 *Construcción de un esquema a partir de interpolación spline.*

1. Se considera un elemento genérico de la base de los B-splines cuadráticos. Se pide calcular el valor de la integral de cada uno de sus tramos.

2. Se considera una función f de la cual conocemos en una partición equiespaciada $\{t_{i-1}, t_i, t_{i+1}\}$ sus valores y el valor de la derivada en el punto t_i. O sea, tenemos:

t_{i-1}	f_{i-1}
t_i	f_i
t_{i+1}	f_{i+1}
t_i	f'_i

   Se pide definir el spline s de dos tramos que ajusta esos datos, dando sus componentes en la base de los B-splines cuadráticos.

3. Calcular:
$$\int_{t_i}^{t_{i+1}} s(x)dx$$

4. Verificar con un ejemplo numérico el cálculo anterior con una función f cuya integral tenga que coincidir con la calculada en 3.

5. Se considera el problema de integrar numéricamente el siguiente problema de valor inicial:
$$\left\{ \begin{array}{l} \frac{dy}{dx} = f(x, y(x)) \\ y\left(x_0\right) = y_0 \end{array} \right.$$
   Se trata de construir un esquema numérico que permita dar el salto de x_i a x_{i+1} usando la aproximación a la integral calculada en el apartado anterior. Definir de modo preciso el funcionamiento de este esquema.

6. Clasificarlo.

7. Se considera el problema de integrar numéricamente el siguiente problema de valor inicial:
$$\left\{ \begin{array}{l} \frac{dy}{dx} = 2xy, \quad 0 < x < 0.4 \\ y\left(0\right) = 1 \end{array} \right.$$
   Se pide construir una tabla en la que aparezcan la solución real y los valores obtenidos de integrar usando el esquema anterior con pasos de 0.2 y 0.1 respectivamente.

8. A la vista de los resultados del apartado anterior, y admitiendo que el esquema verifica la condición de Lipschitz, ¿qué se puede decir de su orden de un modo aproximado?

Solución:

1. Teniendo en cuenta que la partición consta de nodos equiespaciados entre sí $h = t_{i+1} - t_i$, la expresión correspondiente a un B-spline es (ver Hämmerlin y Hoffman[15], Capítulo 6, y la Figura 6.11):

$$B_i^2(t) = \begin{cases} 0, & t \notin [t_i, t_{i+3}) \\ \frac{1}{2h^2}(t - t_i)^2, & t \in [t_i, t_{i+1}) \\ \frac{1}{2h^2}[h^2 + 2h(t - t_{i+1}) - 2(t - t_{i+1})^2], & t \in [t_{i+1}, t_{i+2}) \\ \frac{1}{2h^2}(t_{i+3} - t)^2, & t \in [t_{i+2}, t_{i+3}) \end{cases}$$

Si numeramos de 1 a 3 los tres tramos del soporte de este B-spline, el primer y tercer tramo tendrán

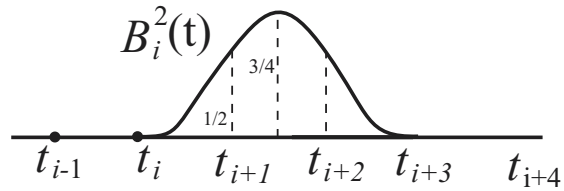


Figura 6.11: B-spline de grado 2.

igual área. Calculémosla:

$$A_1 = \int_{t_i}^{t_{i+1}} \frac{1}{2h^2}(t - t_i)^2 \, dt = \frac{h}{6} = A_3$$

$$A_2 = \int_{t_{i+1}}^{t_{i+2}} \frac{1}{2h^2}\left[h^2 + 2h(t - t_{i+1}) - 2(t - t_{i+1})^2\right] dt = \frac{2h}{3}$$

2. La partición $\{t_{i-1}, t_i, t_{i+1}\}$ consta de dos tramos. En el enganche hay continuidad de la función y de la derivada, pues el spline es de grado 2. Por tanto la dimensión del espacio es $2 \cdot 3 - 2 = 4$, que es el número de elementos de la base. El spline buscado se escribirá como

$$s = a_{i-3}B_{i-3} + a_{i-2}B_{i-2} + a_{i-1}B_{i-1} + a_iB_i$$

Impongamos las condiciones a este spline:

$$\begin{pmatrix} 1/2 & 1/2 & 0 & 0 \\ 0 & 1/2 & 1/2 & 0 \\ 0 & 0 & 1/2 & 1/2 \\ 0 & -1/h & 1/h & 0 \end{pmatrix} \begin{pmatrix} a_{i-3} \\ a_{i-2} \\ a_{i-1} \\ a_i \end{pmatrix} = \begin{pmatrix} f_{i-1} \\ f_i \\ f_{i+1} \\ f_i' \end{pmatrix}$$

Resolviendo analíticamente este sistema lineal se tiene

$$\begin{pmatrix} a_{i-3} \\ a_{i-2} \\ a_{i-1} \\ a_i \end{pmatrix} = \begin{pmatrix} \frac{hf_i'}{2} + 2f_{i-1} - f_i \\ -\frac{hf_i'}{2} + f_i \\ \frac{hf_i'}{2} + f_i \\ -\frac{hf_i'}{2} + 2f_{i+1} - f_i \end{pmatrix} \tag{6.92}$$

3. Para calcular

$$\int_{t_i}^{t_{i+1}} s(x)dx$$

utilizamos los resultados de los dos apartados anteriores, y el hecho de que entre t_i y t_{i+1} sólo hay tres elementos de base que contribuyan al área. Por tanto:

$$\int_{t_i}^{t_{i+1}} s(x)dx = a_{i-2}\frac{h}{6} + a_{i-1}\frac{2h}{3} + a_i\frac{h}{6}$$

Sustituyendo y reagrupando:

$$\int_{t_i}^{t_{i+1}} s(x)dx = \frac{2h}{3}f_i + \frac{h}{3}f_{i+1} + \frac{h^2}{6}f'_i$$

que curiosamente no depende de f_{i-1}. Esto es razonable, pues conociendo la derivada en el punto medio t_i y los valores en los dos puntos t_i y t_{i+1} se define de modo único esa parábola.

4. Para comprobar este resultado, tendremos que usar una función que sea un spline cuadrático y cuya integral sea sencilla de calcular analíticamente. Elegimos directamente una parábola en un tramo en el que no se anulen ninguno de los valores f_i f_{i+1} y f'_i, pues de lo que se trata es de validar precisamente los coeficientes que afectan a estas magnitudes en el cálculo del área. Integremos por ejemplo la parábola $(x-1)^2$ entre 0 y 2.

$$\int_0^2 (x-1)^2 = \frac{2}{3}$$

$$f_i = (0-1)^2 = 1 \quad ; \quad f_{i+1} = (2-1)^2 = 1 \quad ; \quad f'_i = 2(0-1) = -2$$

Y mediante la fórmula del apartado anterior:

$$\int_0^2 s(x)dx = \frac{2\cdot 2}{3}1 + \frac{2}{3}1 - \frac{2^2}{6}2 = \frac{2}{3}$$

5. El esquema exige aproximar la función derivada $f(x, y(x))$ cada dos tramos mediante el spline que hemos estudidado anteriormente, y para dar el salto, integrar como hemos hecho antes. Numéricamente:

$$y_{(i+1)} = y_{(i)} + \int_{x_i}^{x_{i+1}} f(x, y(x))dx \approx y_{(i)} + \int_{x_i}^{x_{i+1}} s(x)dx = y_{(i)} + \frac{2h}{3}f_i + \frac{h}{3}f_{i+1} + \frac{h^2}{6}f'_i$$

El esquema numérico tiene por tanto la formulación:

$$y_{(i+1)} = y_{(i)} + \frac{2h}{3}f_i + \frac{h}{3}f_{i+1} + \frac{h^2}{6}f'_i$$

y se arranca con el valor inicial y_0. El valor f'_i se obtiene derivando analíticamente la función $f(x, y(x))$ como se hace con los métodos de Taylor.

6. Es un esquema multipaso de dos pasos, implícito.

7. $$y_{(i+1)} = y_{(i)} + \frac{2h}{3}2x_i y_{(i)} + \frac{h}{3}2x_{i+1}y_{(i+1)} + \frac{h^2}{6}2y_{(i)}\left(1 + 2x_i^2\right)$$

Despejemos $y_{(i+1)}$

$$y_{(i+1)} = y_{(i)} \quad \frac{1 + \frac{4h}{3}x_i + \frac{h^2}{3}\left(1 + 2x_i^2\right)}{1 - \frac{2h}{3}x_{i+1}}$$

Construyamos la tabla pedida.

x_i	$y(x_i)$	$y_{(i)},\ \Delta x = 0.2$	$y_{(i)},\ \Delta x = 0.1$
0.0	1.00000	1.00000	1.00000
0.1	1.01006		1.01005
0.2	1.04081	1.04109	1.04084
0.3	1.09417		1.09423
0.4	1.17351	1.17424	1.17360

8. El error en el paso i es la diferencia entre el valor real y el valor obtenido con el esquema:

$$e_i = y_{(i)} - y\,(x_i)$$

Si el método es de orden p, entonces (teorema 6.2.4) $\max_{i=0,n} |e_i| \approx Mh^p$ donde M no depende de h y n es el número de pasos realizados en el esquema.

Podemos aplicar este resultado para averiguar el orden. El máximo de los errores se produce para $x = 0.4$ tanto con $h = 0.1$ como con $h = 0.2$. Evaluando dichos errores tenemos que:

$$7.3\ 10^{-4} \approx M(0.2)^p \quad \text{y} \quad 9\ 10^{-5} \approx M(0.1)^p$$

Y dividiendo:

$$8.11 \approx 2^p \quad \Rightarrow \quad p \approx 3.01 \approx 3$$

Podemos integrar hasta $x = 0.6$ y ver si se mantiene la tendencia:

x_i	$y\,(x_i)$	$y_{(i)},\ \Delta x = 0.2$	$y_{(i)},\ \Delta x = 0.1$
0.5	1.28402		1.28416
0.6	1.43333	1.43495	1.43354

$$16.2\ 10^{-4} \approx (0.2)^p \quad \text{y} \quad 2.1\ 10^{-4} \approx M(0.1)^p$$

De donde

$$p \approx 2.95 \approx 3$$

Hemos definido de un modo muy sencillo un esquema de orden 3, de igual calidad que un RK3, o un AM3.

PROBLEMA 6.7 *Sistema de ecuaciones diferenciales ordinarias lineales.*

Se consideran los tres discos de la Figura 6.12, de iguales características, de momento de inercia I unidos mediante tubos elásticos iguales de una unidad de longitud y coeficiente de rigidez R. Se considera el problema

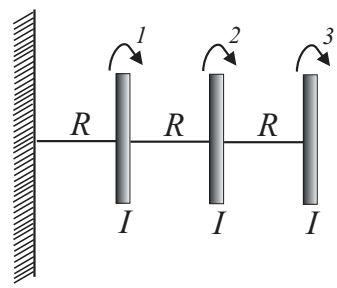


Figura 6.12: Sistema mecánico del problema 6.7.

de la torsión de ese sistema estudiando la evolución en el tiempo de los ángulos de giro φ_1, φ_2 y φ_3 de los discos. Dicha evolución viene dada por el sistema de ecuaciones diferenciales ordinarias obtenidas mediante las ecuaciones de la dinámica:

$$\begin{cases} 2\varphi_1 - \varphi_2 = -\dfrac{I}{R}\varphi_1'' \\ -\varphi_1 + 2\varphi_2 - \varphi_3 = -\dfrac{I}{R}\varphi_2'' \\ -\varphi_2 + \varphi_3 = -\dfrac{I}{R}\varphi_3'' \end{cases} \tag{6.93}$$

1. Escribir el sistema de ecuaciones diferenciales ordinarias de segundo orden (6.93) como un sistema de ecuaciones diferenciales de primer orden, suponiendo que los cocientes I/R tienen valor igual a la unidad en el sistema de unidades empleado.

2. Supongamos que se parte de una situación de reposo en la que las torsiones iniciales son (en radianes):

$$(\varphi_1^{(0)}, \varphi_2^{(0)}, \varphi_3^{(0)}) = (-0.05, 0.0, 0.10)$$

   Aunque el sistema del apartado 1 admite solución analítica [18], al ser de coeficientes constantes, se pide integrarlo numéricamente utilizando un esquema de Euler explícito. Se darán dos pasos, tomando un incremento temporal $h = 0.1$ uds.

3. Utilizar ahora un esquema de Euler implícito.

4. Utilizar por último un esquema de Crank-Nicolson.

5. Calcular el radio espectral de la matriz de iteración para el método de Gauss-Seidel correspondiente al Euler implícito del apartado 3. Comentar la convergencia a partir de este valor.

Solución:

1. Poniendo $\varphi_i' = p_i, \quad i = 1, 3$ reducimos el sistema (6.93) al sistema de seis ecuaciones lineales de primer orden

$$\frac{d}{dt}\begin{pmatrix} \varphi_1 \\ p_1 \\ \varphi_2 \\ p_2 \\ \varphi_3 \\ p_3 \end{pmatrix} = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ -2 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & -2 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & -1 & 0 \end{pmatrix} \begin{pmatrix} \varphi_1 \\ p_1 \\ \varphi_2 \\ p_2 \\ \varphi_3 \\ p_3 \end{pmatrix} \tag{6.94}$$

2. Si llamamos $\mathbf{y}$ al vector columna

$$\mathbf{y} = (\varphi_1, p_1, \varphi_2, p_2, \varphi_3, p_3)^T$$

   $\mathbf{p}$ a su vector derivada, y A a la matriz del sistema, podemos escribir (6.94) en la forma $\mathbf{p} = A\mathbf{y}$.

   El esquema de Euler explícito correspondiente se escribirá entonces

$$\mathbf{y}_{(n+1)} = \mathbf{y}_{(n)} + h \cdot \mathbf{p}_{(n)} = \mathbf{y}_{(n)} + h\, A\mathbf{y}_{(n)} = (I + h\, A)\mathbf{y}_{(n)}$$

   Con el estimador inicial $\mathbf{y}_{(0)} = (-0.05, 0.0, 0.0, 0.0, 0.10, 0.0)^T$ el resultado de los dos primeros pasos es

$$\mathbf{y}_{(1)} = (-0.0500, 0.0100, 0, 0.0050, 0.1000, -0.0100)^T$$
$$\mathbf{y}_{(2)} = (-0.0490, 0.0200, 0.0005, 0.0100, 0.0990, -0.0200)^T$$

   Como vemos, los ángulos decrecen, pues el sistema se recupera después de la deformación inicial.

   Podemos utilizar unas sencillas líneas Matlab para implementar este esquema, y los de los apartados siguientes.

   El fichero correspondiente se puede bajar de la dirección *http://canal.etsin.upm.es/ftp/torsion.m*

3. El esquema de Euler implícito se escribirá

$$\mathbf{y}_{(n+1)} = \mathbf{y}_{(n)} + h \cdot \mathbf{p}_{(n+1)} = \mathbf{y}_{(n)} + h\, A\mathbf{y}_{(n+1)}$$

   de donde

$$(I - h\, A)\mathbf{y}_{(n+1)} = \mathbf{y}_{(n)}$$

[18] Su resolución analítica no tiene dificultades teóricas pero sí operacionales de cierta magnitud.

En cada paso de tiempo hemos de resolver un sistema lineal. Los valores obtenidos para los dos primeros pasos son:

$$\mathbf{y}_{(1)} = (-0.0490, 0.0099, 0.0005, 0.0049, 0.0990, -0.0099)^T$$
$$\mathbf{y}_{(2)} = (-0.0471, 0.0194, 0.0015, 0.0096, 0.0971, -0.0194)^T$$

4. El esquema de Crank-Nicolson se escribe aquí

$$\mathbf{y}_{(n+1)} = \mathbf{y}_{(n)} + \frac{h}{2} \cdot \left(\mathbf{p}_{(n+1)} + \mathbf{p}_{(n)}\right) = \mathbf{y}_{(n)} + A\frac{h}{2} \cdot \left(\mathbf{y}_{(n+1)} + \mathbf{y}_{(n)}\right)$$

de donde

$$\left(I - \frac{h}{2}A\right)\mathbf{y}_{(n+1)} = \left(I + \frac{h}{2}A\right)\mathbf{y}_{(n)}$$

De nuevo debemos resolver un sistema lineal en cada paso de tiempo. Los valores obtenidos para los dos primeros pasos son ahora

$$\mathbf{y}_{(1)} = (-0.0495, 0.0100, 0.0002, 0.0050, 0.0995, -0.0100)^T$$
$$\mathbf{y}_{(2)} = (-0.0480, 0.0198, 0.0010, 0.0099, 0.0980, -0.0198)^T$$

5. Se trata de calcular los autovalores de la matriz de iteración B correspondiente a la descomposición de GS de $H = I - h\ A$. Operando se llega a un valor de 0.0649, lo que significa que el método es convergente.

PROBLEMA 6.8 *Ecuación diferencial de orden superior a uno.*

Una caja de masa m desliza sobre una rampa empujada por su propio peso (gravedad g). El movimiento de la caja sobre la rampa está afectado de una fuerza de rozamiento contraria al movimiento, proporcional a la reacción sobre el plano con un factor ν. Además, esa caja tiene acoplada un muelle de constante K que en la posición inicial de reposo está con elongación nula y que ejerce una acción sobre la caja contraria al movimiento (ver Figura 6.13). La rampa cambia su inclinación (argumentos en radianes) con el tiempo siguiendo la ley $\theta(t) = e^t - t - 1$ para $t \leq 1.38$ segundos, coincidiendo la posición inicial de la caja con el eje de giro.
Si planteamos este problema en un sistema de coordenadas polares con origen en el eje de giro del plano, las ecuaciones que permiten calcular las diferentes fuerzas que intervienen son para la fuerza normal N que el plano ejerce sobre la caja:

$$N = m\left(g\cos\theta + 2r'(t)\omega + r\alpha\right)$$

donde $\omega = \theta'(t)$ y $\alpha = \theta''(t)$. La fuerza de rozamiento F es proporcional a esta fuerza normal y actúa en la dirección contraria al movimiento. Por tanto en su valor hay que tener en cuenta el signo de la velocidad, que podría cambiar debido al muelle

$$F = -\nu N \text{ signo}\left[r'(t)\right]$$

En la dirección del movimiento paralelo a la rampa actúa también la fuerza del muelle S

$$S = -K(r - r(0))$$

como en el instante inicial la posición es el origen de coordenadas y la elongación es nula, la fuerza del muelle se escribe como:

$$S = -Kr$$

Quedan por considerar las llamadas "fuerzas"de inercia I debido a que el sistema de referencia no es inercial, y la proyección W del peso sobre la rampa.

$$I = mr\omega^2$$
$$W = -mg\sin\theta$$

La ley de Newton suministra la ecuación de nuestro modelo matemático

$$mr''(t) = I + W + F + S \tag{6.95}$$

Definimos el problema de valor inicial en estudio dando los siguientes datos iniciales. La velocidad inicial de la caja es de 0.5 m/s en la dirección contraria al muelle, su masa es de 100 kg, el coeficiente de rozamiento $\nu = 0.1$, y la constante K del muelle vale 1000 Nw/m.

1. Tomando un paso de tiempo de 0.1 segundos, dar dos pasos PECE, en un esquema predictor-corrector Euler explícito-implícito para aproximar la posición al cabo de 0.2 segundos.

2. Razonar el sentido físico de los resultados.

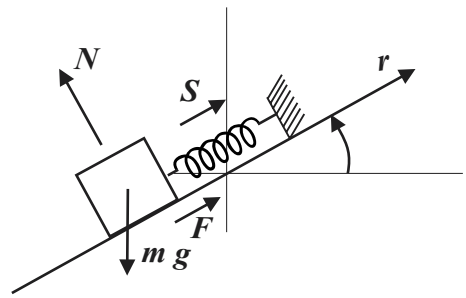


Figura 6.13: Plano inclinado correspondiente a los problemas 6.8 y 7.8.

Solución:

1. Las condiciones iniciales son $r(0) = 0$ y $r'(0) = -0.5$.

   Calculemos las diferentes fuerzas, el ángulo θ y sus derivadas en el instante $t = 0$:

$$\begin{aligned}
\theta &= e^0 - 0 - 1 = 0 \\
\omega &= e^0 - 1 = 0 \\
\alpha &= e^0 = 1 \\
N &= m\,(10\cos\theta + 2r'\omega + r\alpha) = 1000 \\
F &= -\nu N \text{ signo}\,[r'] = -0.1 \cdot 1000 \cdot (-1) = 100 \\
S &= -Kr = -1000 \cdot 0 = 0 \\
I &= mr\omega^2 = 100 \cdot 0 \cdot 0 = 0 \\
W &= -mg\sin\theta = -100 \cdot 10 \cdot 0 = 0
\end{aligned}$$

   En el instante inicial, la única fuerza que actúa es la fricción. Calculemos la derivada segunda de la posición utilizando estos datos para poder hacer la primera predicción [19].

$$r''(0) = \frac{I + W + F + S}{m} = \frac{100}{100} = 1$$

$$\begin{pmatrix} r \\ r' \end{pmatrix}^p_{(1)} = \begin{pmatrix} r \\ r' \end{pmatrix}_{(0)} + 0.1 \begin{pmatrix} r' \\ r'' \end{pmatrix}_{(0)} = \begin{pmatrix} 0.0 \\ -0.5 \end{pmatrix} + 0.1 \begin{pmatrix} -0.5 \\ 1.0 \end{pmatrix} = \begin{pmatrix} -0.0500 \\ -0.4000 \end{pmatrix}$$

[19] En este problema indicamos con el superíndice p los valores del predictor, y con el superíndice c los del corrector.

Para poder hacer las correcciones, tenemos que estimar las fuerzas en el instante $t = 0.1$ utilizando los valores obtenidos con el predictor.

$$\begin{aligned}
\theta &= e^{0.1} - 0.1 - 1 = 0.005171 \\
\omega &= e^{0.1} - 1 = 0.1052 \\
\alpha &= e^{0.1} = 1.1052 \\
N &= m\left(10\cos\theta + 2(r'_{(1)})^p\omega + (r_{(1)})^p\alpha\right) = 986.0471 \\
F &= -\nu N \text{ signo}\left[(r'_{(1)})^p\right] = -0.1 \cdot 986.0471 \cdot (-1) = 98.6047 \\
S &= -K(r_{(1)})^p = -1000 \cdot (-0.05) = 50 \\
I &= m(r_{(1)})^p\omega^2 = 100 \cdot (-0.05) \cdot 0.1052^2 = -0.0553 \\
W &= -mg\sin\theta = -100 \cdot 10 \cdot \sin(0.0052) = -5.1709 \\
(r''_{(1)})^p &= 0.01(I + W + F + S) = 1.4338
\end{aligned}$$

Con estos valores corregimos:

$$\begin{pmatrix} r \\ r' \end{pmatrix}^c_{(1)} = \begin{pmatrix} r \\ r' \end{pmatrix}_{(0)} + 0.1\begin{pmatrix} r' \\ r'' \end{pmatrix}^p_{(1)} = \begin{pmatrix} 0.0 \\ -0.5 \end{pmatrix} + 0.1\begin{pmatrix} -0.4000 \\ 1.4338 \end{pmatrix} = \begin{pmatrix} -0.0400 \\ -0.3566 \end{pmatrix}$$

La velocidad disminuye debido al efecto del rozamiento y del peso. Como el ángulo θ es tan pequeño, el peso no empuja apenas en la dirección del movimiento. La posición se disminuye porque partíamos de $r = 0$ con velocidades negativas.

Una vez terminado el ciclo PECE, damos el segundo paso del esquema predictor/corrector tomando como valores definitivos del paso 1 los obtenidos con el corrector.

$$\begin{aligned}
\theta &= e^{0.1} - 0.1 - 1 = 0.005171 \\
\omega &= e^{0.1} - 1 = 0.1052 \\
\alpha &= e^{0.1} = 1.1052 \\
N &= m\left(10\cos\theta + 2r'\omega + r\alpha\right) = 988.0647 \\
F &= -\nu N \text{ signo}\left[r'\right] = -0.1 \cdot 988.0647 \cdot (-1) = 98.8064 \\
S &= -Kr = -1000 \cdot (-0.0400) = 40.0000 \\
I &= mr\omega^2 = 100 \cdot (-0.0400) \cdot 0.1052^2 = -0.0442 \\
W &= -mg\sin\theta = -100 \cdot 10 \cdot 0.0052 = -5.1709
\end{aligned}$$

Calculemos la derivada segunda de la posición utilizando estos datos para poder hacer la predicción:

$$r''(0.1) = \frac{I + W + F + S}{m} = 1.3359$$

$$\begin{pmatrix} r \\ r' \end{pmatrix}^p_{(2)} = \begin{pmatrix} r \\ r' \end{pmatrix}_{(1)} + 0.1\begin{pmatrix} r' \\ r'' \end{pmatrix}_{(1)} = \begin{pmatrix} -0.0400 \\ -0.3566 \end{pmatrix} + 0.1\begin{pmatrix} -0.3566 \\ 1.3359 \end{pmatrix} = \begin{pmatrix} -0.075666 \\ -0.223030 \end{pmatrix}$$

Calculemos r'' para estos valores y después apliquemos el corrector:

$$\begin{aligned}
\theta &= e^{0.2} - 0.2 - 1 = 0.021402 \\
\omega &= e^{0.2} - 1 = 0.221402 \\
\alpha &= e^{0.2} = 1.221402 \\
N &= m\left(10\cos\theta + 2(r'_{(2)})^p\omega + (r_{(2)})^p\alpha\right) = 980.6536 \\
F &= -\nu N \text{ signo}\left[(r'_{(2)})^p\right] = -0.1 \cdot 980.6536 \cdot (-1) = 98.065367 \\
S &= -K(r_{(2)})^p = -1000 \cdot (-0.07566) = 75.662148 \\
I &= m(r_{(2)})^p\omega^2 = 100 \cdot (-0.05) \cdot 0.221402^2 = -0.370889 \\
W &= -mg\sin\theta = -100 \cdot 10 \cdot \sin(0.021402) = -21.401124 \\
(r''_{(2)})^p &= 0.01(I + W + F + S) = 1.519555
\end{aligned}$$

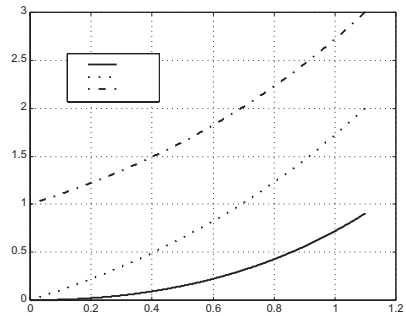


Figura 6.14: Ángulo de la rampa y derivadas en función del tiempo.

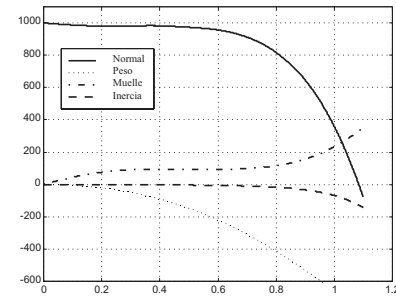


Figura 6.15: Fuerzas sobre la caja en función del tiempo.

Corrijamos:

$$\begin{pmatrix} r \\ r' \end{pmatrix}^{c}_{(2)} = \begin{pmatrix} r \\ r' \end{pmatrix}_{(1)} + 0.1 \begin{pmatrix} r' \\ r'' \end{pmatrix}^{p}_{(2)} = \begin{pmatrix} -0.0400 \\ -0.3566 \end{pmatrix} + 0.1 \begin{pmatrix} -0.223030 \\ 1.519555 \end{pmatrix} = \begin{pmatrix} -0.062303 \\ -0.204665 \end{pmatrix}$$

2. Los resultados son razonables pues la posición r disminuye y la velocidad disminuye también en módulo, ya que los efectos del rozamiento y del muelle son más importantes que el peso y que la "fuerza" de inercia.

   El pequeño código Matlab que integra esta ecuación por este método se encuentra en el zip con todos los demás códigos. Para comprobar los valores del ejercicio basta con hacer en el programa $tmax = 0.2$ y $npasos = 2$, visualizando al final las variables r y dr.

   Vamos a hacer un análisis más detallado de este problema utilizando este código.

   Las curvas correspondientes a θ y sus derivadas son claves para entender los resultados (ver la Figura (6.14)).

   Ahí podemos observar que la velocidad angular ω crece muy rápidamente (es una exponencial) y para tiempos del orden de 1 segundo ya vale 1.7 rad/s siendo el valor correspondiente del ángulo aproximadamente 40 grados.

   Tenemos una caja de 100 kg descendiendo por una rampa que forma 40 grados con la horizontal y cuya inclinación está creciendo en la dirección del movimiento (ver Figura (6.13)).

   Como la velocidad del punto donde está la caja es proporcional a la distancia al eje, y la velocidad angular crece muy rápido, podría suceder que la caja se separara de la rampa. La condición para que esto sucediera es que la reacción normal se hiciera negativa. Se puede observar en la gráfica de fuerzas como función del tiempo, Figura 6.15, que esto sucede aproximadamente para $t = 1.1$. A partir de aquí nuestro modelo matemático ya no sirve, pues esa fuerza normal no puede ser negativa. Por otro lado, en la gráfica de velocidades y posiciones en la Figura 6.16 se observa que entre 0 y 0.4 segundos la velocidad disminuye (no se debe olvidar que la condición inicial es que la velocidad es de 0.5 m/s) debido al efecto del muelle y del rozamiento.

   Entre 0.4 y 0.6 segundos la caja se queda parada, pues el efecto del muelle más rozamiento es contrarrestado por el peso. A partir de ahí los ángulos son demasiado grandes, crece la velocidad y la caja se aleja del punto de partida hasta que se separa de la rampa.

   Es interesante estudiar qué sucede si el ángulo varía más lentamente y baja el rozamiento, en este caso, el muelle es el factor determinante en el movimiento, además del peso.

   Para $\nu = 0.001$ y para $\theta(t) = \mathrm{Ln}(t+1)$ tenemos la gráfica de posición y velocidad de la Figura 6.17.

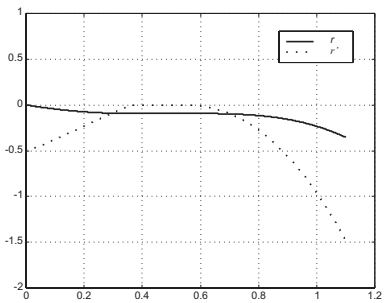


Figura 6.16: Posición y velocidad en el problema 6.8.

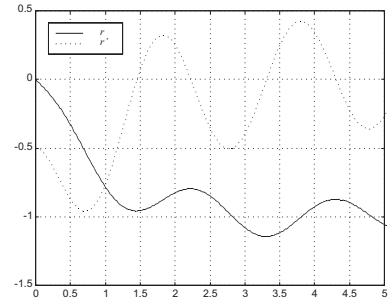


Figura 6.17: Posición y velocidad en el problema 6.8 para $\theta(t) = \mathbf{Ln}(t+1)$.

PROBLEMA 6.9 *Ecuaciones del tiro parabólico.*

Se trata de estudiar el movimiento de un proyectil M en el aire, lanzado desde el suelo con un ángulo de inclinación β_0, y que sufre una resistencia del aire de dirección contraria a su vector velocidad $\mathbf{v}$ y de valor $0.01\ v^2$. Presentamos el diagrama de fuerzas en la Figura 6.18.

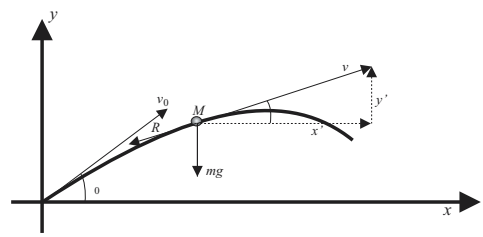


Figura 6.18: Esquema del movimiento del proyectil.

1. Escribir el sistema de ecuaciones diferenciales ordinarias que tenga en cuenta esos efectos y que nos proporcione las aceleraciones en las direcciones x e y.

2. Suponiendo que la masa vale 10 kg, la velocidad inicial es de 500 m/s y la inclinación de 30 grados, se pide calcular la velocidad al cabo de 0.2 segundos, tomando un paso temporal de 0.1 segundos e integrando con un predictor AB2 y su corrector asociado de Crank-Nicolson, realizando el arranque con el método de Runge-Kutta2(Heun) [20].

3. Una vez llegados a este apartado, tenemos una estimación de la velocidad en las dos direcciones x e y, en los tres instantes $t = 0.0$ seg, $t = 0.1$ seg y $t = 0.2$ seg. Al ser la posición la integral de la velocidad, se pide estimar ésta en el instante $t = 0.2$ seg utilizando la regla de Simpson.

Solución:

1. Las ecuaciones del movimiento son

$$\begin{aligned} mx'' &= -0.01v^2\cos\beta \\ my'' &= -0.01v^2\sin\beta - mg \end{aligned}$$

[20] Como control de los resultados, se tendrá en cuenta que por consideraciones físicas, las velocidades tanto en la dirección x como en la dirección y decrecen.

Como la masa vale 10 y

$$\cos\beta = \frac{x'}{\sqrt{x'^2+y'^2}}; \quad \sin\beta = \frac{y'}{\sqrt{x'^2+y'^2}}$$

tenemos que:

$$\begin{cases} x'' = -0.001\sqrt{x'^2+y'^2}\, x' \\ y'' = -0.001\sqrt{x'^2+y'^2}\, y' - 10 \end{cases}$$

2. Instante $t = 0$. Las condiciones iniciales son $x' = 500\cos(\pi/6)$ e $y' = 500\sin(\pi/6)$. Con el $RK2$ de Heun, tenemos que hacer un Euler explícito y luego promediar. Evaluemos primero las segundas derivadas en el instante inicial[21].

$$\begin{aligned} \sqrt{(x'^0)^2+(y'^0)^2} &= 500.0000 \\ x''^0 &= -0.001\ 500\ 433.0127 = -216.5064 \\ y''^0 &= -0.001\ 500\ 250.0000 - 10.0000 = -135.0000 \\ \hat{x}'^1 &= x'^0 + 0.1x''^0 = 433.0127 + 0.1(-216.5064) = 411.3621 \\ \hat{y}'^1 &= y'^0 + 0.1y''^0 = 250.0000 + 0.1(-135.0000) = 236.5000 \end{aligned}$$

Evaluamos las derivadas segundas en 0.1.

$$\begin{aligned} \sqrt{(\hat{x}'^1)^2+(\hat{y}'^1)^2} &= 474.5008 \\ \hat{x}''^1 &= -0.001\ 474.5008\ 411.3621 = -195.1916 \\ \hat{y}''^1 &= -0.001\ 474.5008\ 236.5000 - 10.0000 = -122.2194 \end{aligned}$$

Promediamos para pasar al instante $t = 0.1$ seg

$$\begin{aligned} x'^1 &= x'^0 + \frac{0.1}{2}\left(x''^0 + \hat{x}''^1\right) = 433.0127 + 0.05\,(-216.5064 - 195.1916) = 412.4278 \\ y'^1 &= y'^0 + \frac{0.1}{2}\left(y''^0 + \hat{y}''^1\right) = 250.0000 + 0.05(-135.0000 - 122.2194) = 237.1390 \end{aligned}$$

Ahora que tenemos los valores del algoritmo de inicio, ya podemos utilizar el predictor-corrector AB2-AM2, para lo cual necesitamos evaluar las derivadas segundas en $t = 0.1$ seg.

$$\begin{aligned} \sqrt{(x'^1)^2+(y'^1)^2} &= 475.7432 \\ x''^1 &= -0.001\ 475.7432\ 412.4278 = -196.2097 \\ y''^1 &= -0.001\ 475.7432\ 237.1390 - 10.0000 = -122.8173 \end{aligned}$$

$$\begin{aligned} \hat{x}'^2 &= x'^1 + 0.1\left(\frac{3}{2}x''^1 - \frac{1}{2}x''^0\right) = 412.4278 + 0.1\left(-\frac{3}{2}196.2097 + \frac{216.5064}{2}\right) \\ \hat{y}'^2 &= y'^1 + 0.1\left(\frac{3}{2}y''^1 - \frac{1}{2}y''^0\right) = 237.1390 + 0.1\left(-\frac{3}{2}122.8173 + \frac{135.0000}{2}\right) \end{aligned}$$

$$\begin{aligned} \hat{x}'^2 &= 393.8217 \\ \hat{y}'^2 &= 225.4664 \end{aligned}$$

Para aplicar el corrector AM2, necesitamos estimar las segundas derivadas en $t = 0.2$ seg.

$$\begin{aligned} \sqrt{(\hat{x}'^2)^2+(\hat{y}'^2)^2} &= 453.7958 \\ \hat{x}''^2 &= -0.001\ 453.7958\ 393.8217 = -178.7146 \\ \hat{y}''^2 &= -0.001\ 453.7958\ 225.4664 - 10.0000 = -112.3157 \end{aligned}$$

[21] Utilizamos la notación $\hat{x}$ para referirnos a que la variable considerada, la x en este caso, se utiliza como predictor.

$$x'^2 = x'^1 + \frac{0.1}{2}\left(x''^1 + \hat{x}''^2\right) = 412.4278 - 0.05\ 374.9243 = 393.6816$$
$$y'^2 = y'^1 + \frac{0.1}{2}\left(y''^1 + \hat{y}''^2\right) = 237.1390 - 0.05\ 235.1330 = 225.3824$$

3. Hagamos una tabla resumen:

i	t_i	x'^i	y'^i
0	0.0	433.0172	250.0000
1	0.1	412.4278	237.1390
2	0.2	393.6816	225.3824

Observamos que las dos componentes de la velocidad disminuyen en valor, resultados coherentes con el fenómeno físico. Para evaluar la variación en la posición, habrá que integrar las velocidades, aproximando esas integrales mediante la regla de Simpson, como se indica en el enunciado.

$$x(0.2) = \int_0^{0.2} x'(t)dt \approx \frac{0.2}{6}(433.0172 + 4 \cdot 412.4278 + 393.6816) = 82.5470$$
$$y(0.2) = \int_0^{0.2} y'(t)dt \approx \frac{0.2}{6}(250.0000 + 4 \cdot 237.1390 + 225.3824) = 47.4646$$

Se deja como ejercicio comprobar la influencia que tiene el rozamiento en la posición, dado que si no se tiene éste en cuenta, las ecuaciones se resuelven analíticamente de modo directo. Se pide valorar si el coeficiente que se ha impuesto es razonable.

PROBLEMA 6.10 *Ecuación diferencial singular.*

Se define la función $(x, y) \to f(x, y)$ por las condiciones siguientes

$$f(x,y) = \begin{cases} 0 & \text{si} \quad x \le 0 \\ 2x & \text{si} \quad 0 < x < 1, y < 0 \\ 2x - \dfrac{4y}{x} & \text{si} \quad 0 \le y \le x^2 \text{ y } x < 1 \\ -2x & \text{si} \quad 0 < x < 1, x^2 < y \end{cases}$$

1. Demostrar que f es continua pero no lipchiciana.
2. Se considera el problema de Cauchy

$$(P) \qquad \begin{cases} y' = f(x,y) \\ y(0) = 0 \end{cases}$$

   a) Demostrar que el problema propuesto tiene solución única.
   b) Formar la sucesión definida por el proceso iterativo de Picard y comprobar que no es convergente.
3. Demostrar que los polígonos de Euler convergen.

Comentarios

a) De acuerdo con el teorema de existencia y unicidad, si f es continua y verifica en el entorno del origen una condición de Lipschitz

$$|f(x, y_1) - f(x, y_2)| \le L|y_1 - y_2|$$

el problema de Cauchy propuesto tiene una solución única que pasa por el origen. Si f es continua y acotada pero no lipchiciana se prueba la existencia de la solución pero no se garantiza la unicidad. En este ejercicio se construye una función continua y acotada, no lipchiciana en el origen y que no obstante posee solución única pasando por el origen que no se puede hallar por el método de aproximaciones sucesivas y sin embargo se puede obtener por el método de Euler.

b) En el apartado 2.b, se razonará por contradicción. Suponiendo que hay dos soluciones distintas y_1 e y_2, se considerará la función no negativa $g(x) = (y_1(x) - y_2(x))^2$ y se demostrará que g es decreciente para todo x luego negativa, la conclusión será evidente.

c) En el apartado 3 se formarán los polígonos de Euler para un paso h cualquiera y se estudiará su convergencia.

Solución:

1. Las curvas de ecuaciones $x = 0$, $y = 0$ e $y = x^2$, definen en el plano xy cuatro regiones (ver Figura 6.19) que se llamarán I, II, III y IV.
   En cada una de ellas (fronteras no incluidas) la función $f(x, y)$ es continua

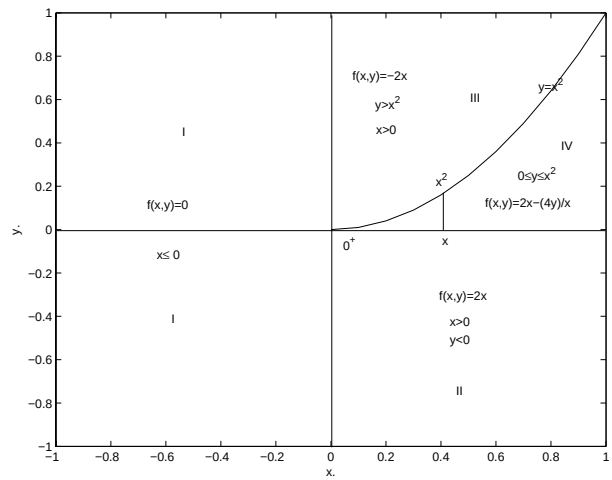


Figura 6.19: Representación gráfica de las distintas regiones I, II, III y IV, en las que la función f viene definida por fórmulas distintas.

- En I, $x \leq 0$, $f(x, y) = 0$ constante.
- En II, $0 \leq x \leq 1$. $y < 0$ y en III, $0 < x \leq 1$, $y > x^2$, $f(x, y) = \pm 2x$.
- En IV, $0 < y < x^2$, $0 < x \leq 1$, f es suma de dos funciones continuas, ya que $x \neq 0$ en todo punto.

Verifiquemos ahora la continuidad de f en las fronteras de esas regiones:

- Cuando $(x, y) \to (x_0, y_0)$ con $x_0 = 0$, $y_0 < 0$ frontera I∩II, $f(x, y) \to 0$ tanto acercándonos desde I como desde II.

- Cuando $(x, y) \to (x_0, y_0)$ con $x_0 = 0$, $y_0 > 0$ frontera I∩III,$f(x, y) \to 0$ en I y $f(x, y) \to -2x_0 = 0$ en III.

- En II∩IV, $(x, y) \to (x_0, y_0)$ con $x_0 > 0$, $y_0 = 0$, $f(x, y) \to 2x_0$ en II y $f(x, y) \to 2x_0 - 4\dfrac{y_0}{x_0} = 2x_0$ en IV.

- En III∩IV, $(x, y) \to (x_0, y_0)$ con $x_0 > 0$, $y_0 = x_0^2$ $f(x, y) \to -2x_0$ en III y $f(x, y) \to 2x_0 - 4\dfrac{y_0}{x_0} = 2x_0 - 4\dfrac{x_0^2}{x_0} = -2x_0$ en IV.

$f(x, y)$ tiende en todos los casos al mismo límite independientemente del modo en que (x, y) tiende a (x_0, y_0).

Estudiemos por último el origen $(0, 0)$.

En IV se tiene siempre [22]

$$|f(x,y)| \leq \left|2x - 4\frac{y}{x}\right| < |2x| + 4\left|\frac{y}{x}\right| \leq 2|x| + 4|x| = 6|x|$$

y en el resto de las regiones $|f(x,y)| \leq 2|x|$ por tanto se tiene en todo el plano que

$$|f(x,y)| \leq 6\,\text{máx}\{|x|,|y|\}$$

luego si $(x,y) \to (0,0)$, $|f(x,y)| \to 0$.

La función f es continua y acotada [23] en $D = (-\infty, 1] \times I\!R$ [24].

La función f no es lipchiciana en D.

En las regiones I, II y III se tiene $\frac{\partial f}{\partial y} = 0$. La condición de Lipschitz se verifica con $L = 0$.

En IV, $\frac{\partial f}{\partial y} = -\frac{4}{x}$ que tiende a $-\infty$ cuando $x \to 0^+$, luego $\left|\frac{\partial f}{\partial y}\right|$ no está acotada en esta región cuando $(x,y) \to (0^+, y)$ con $y < x^2$ y, por tanto, no es lipchiciana en D.

Veámoslo con detalle. Sean $(x,y_1),(x,y_2)$ dos puntos de IV de igual abscisa (ver Figura 6.20),

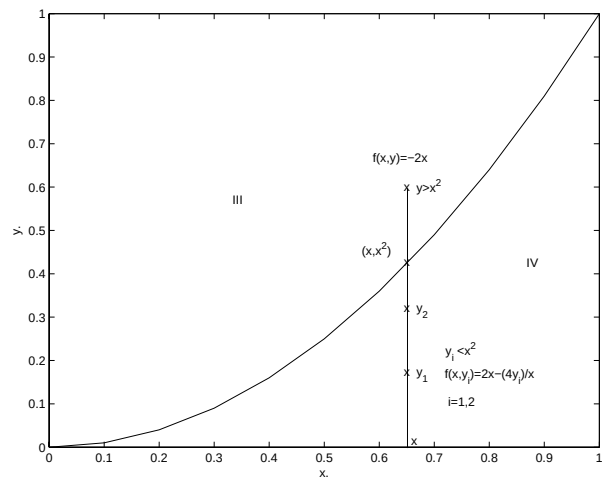


Figura 6.20: Representación de las regiones III y IV.

$$|f(x,y_1) - f(x,y_2)| = \left|2x - \frac{4y_1}{x} - 2x + \frac{4y_2}{x}\right| = \frac{4}{x}|y_2 - y_1|$$

y cuando $x \to 0^+$, $\frac{4}{x}$ no está mayorada, luego no existe $L > 0$ tal que

$$|f(x,y_1) - f(x,y_2)| \leq L|y_1 - y_2|$$

en esa región.

2. a) El problema (P) tiene solución única.
   Demostremos primero que f es para cada x decreciente en y. La conclusión es trivial en las regiones I, II, y III, ya que la aplicación parcial $y \to f(x,y)$ es constante. En IV, suponiendo que $y_2 > y_1$ (ver Figura 6.20)

$$f(x,y_1) - f(x,y_2) = -\frac{4y_2}{x} + \frac{4y_1}{x} = \frac{4}{x}(y_1 - y_2) < 0$$

[22] En IV, x e y son ambas mayores que 0 e $y \leq x^2$, luego $4\frac{y}{x} \leq 4\frac{x^2}{x} = 4x$.

[23] $|f(x,y)| < 6 \quad \forall (x,y) \in (-\infty, 1] \times I\!R$.

[24] El teorema de Cauchy-Peano establece que si f es continua y acotada, el problema de Cauchy tiene al menos una solución en $[x_0, x_1]$ cualesquiera que sean $x_0, x_1 \in (-\infty, 1]$.

Esta será la herramienta que utilizaremos en la demostración de la unicidad.

Supongamos razonando por la contradicción que hay dos soluciones distintas $y_1(x)$ e $y_2(x)$ de (P) y consideremos, como sugiere el enunciado, la función no negativa $g(x) = (y_1(x) - y_2(x))^2$. De la condición inicial de (P) se sigue $g(0) = (y_1(0) - y_2(0))^2 = 0$ y

$$g'(x) = 2(y_1(x) - y_2(x))(y_1'(x) - y_2'(x)) = 2(y_1(x) - y_2(x))(f(x, y_1) - f(x, y_2)) \leq 0$$

En efecto,

- Si $y_1 > y_2$, $y_1 - y_2 > 0$ pero $f(x, y_1) < f(x, y_2)$ y $f(x, y_1) - f(x, y_2) < 0$.
- Si $y_2 > y_1$, $y_1 - y_2 < 0$ con $f(x, y_1) > f(x, y_2)$ y $f(x, y_1) - f(x, y_2) > 0$.

g es decreciente para todo x y vale 0 en $x = 0$, luego es negativa y el razonamiento por el absurdo lleva a $g(x) = 0$, es decir, $y_1(x) = y_2(x)$.

Aunque f no satisfaga la condición de Lipschitz, como es continua y acotada en D, tanto la transformación

$$T(y) : x \to y_0 + \int_0^x f(s, y(s))ds$$

base del método iterativo de Picard como sus iteradas, tienen sentido, luego podemos formar los términos de la sucesión $\{T^k y\}$ del método de aproximaciones sucesivas y cabe preguntarse si partiendo de una función continua en $[x_0, x_1]$, la sucesión de las iteradas se aproxima a una solución del problema de Cauchy propuesto o al menos que exista una subsucesión que converja a una solución de (P).

La respuesta es negativa como veremos a continuación.

b) Apliquemos el método de aproximaciones sucesivas comenzando con $y^{(0)}(x) = 0$ obtenemos

- $$y^{(1)}(x) = Ty^{(0)}(x) = \int_0^x f(s, 0)ds = \int_0^x 2sds = x^2$$

  En este caso, $y = 0$ y $s \in (0, x)$, luego estamos sobre el eje x. Si $x > 0$, $f(s, 0) = 2x$ y si $x < 0$, $f(s, 0) = 0$ y

  $$y^{(1)}(x) = \begin{cases} 0 & \text{si} \quad x < 0 \\ x^2 & \text{si} \quad x \geq 0 \end{cases}$$

- $$y^{(2)}(x) = Ty^{(1)}(x) = \begin{cases} 0 & \text{si} \quad x < 0 \\ \int_0^x f(s, s^2)ds = \int_0^x -2sds = -x^2 & \text{si} \quad x \geq 0 \end{cases}$$

  En este caso, $y = x^2$ y $s \in (0, x)$. Estamos sobre la parábola $y = x^2$, y $f(s, s^2) = -2s$.

- $$y^{(3)}(x) = Ty^{(2)}(x) = \begin{cases} 0 & \text{si} \quad x < 0 \\ \int_0^x f(s, -s^2)ds = \int_0^x 2sds = x^2 & \text{si} \quad x \geq 0 \end{cases}$$

  Aquí, $y = -x^2 < 0$ y $s \in (0, x)$ luego $(s, -s^2) \in$ II y $f(s, -s^2) = 2s$, etcétera.

  La sucesión $\{T^k y\}$ alterna los valores x^2 cuando k es impar y $-x^2$ cuando k es par.

Ni la función $x \to x^2$ ni la $x \to -x^2$ son soluciones del problema, luego la sucesión $\{T^k y\}$ claramente divergente, tampoco contiene ninguna subsucesión convergente a una solución de (P).

3. Formemos los polígonos de Euler.
   Partiendo de $(x_0, y_{(0)}) = (0, 0)$, seleccionamos una partición equiespaciada de paso h. Con ello tendremos sucesivamente

   - $y_{(1)} = y_{(0)} + hf(x_0, y_{(0)}) = 0 + hf(0, 0) = 0$
   - $y_{(2)} = y_{(1)} + hf(x_1, y_{(1)}) = y_1 + hf(h, 0) = 0 + h \cdot 2h = 2h^2$
   - $y_{(3)} = y_{(2)} + hf(x_2, y_{(2)}) = 2h^2 + hf(2h, 2h^2) = 2h^2$
     (Aquí $2h^2 < (2h)^2$, luego $0 < y < x^2$ y $f(2h, 2h^2) = 4h - \frac{8h^2}{2h} = 4h - 4h = 0$)

- $y_{(4)} = y_{(3)} + hf(x_3, y_{(3)}) = 2h^2 + hf(3h, 2h^2) = 2h^2 + \frac{10h^2}{3} = \frac{16h^2}{3}$
  ($2h < 3h \quad \Rightarrow \quad 2h^2 < 3h^2 < (3h)^2$. De nuevo $0 < y < x^2$ y $f(3h, 2h^2) = 6h - \frac{8h^2}{3h} = 6h - \frac{8h}{3} = \frac{10h}{3}$)

Comenzamos a sospechar que esto pueda ser así en todos los casos y tanteamos una demostración por recurrencia.

Suponemos que $y_{(i)} < x_i^2$ y debemos demostrar que $y_{(i+1)} < x_{i+1}^2$.

En efecto, ya que $(x_i, y_{(i)})$ está en IV

$$y_{(i+1)} = y_{(i)} + h\left(x_i - \frac{4y_{(i)}}{x_i}\right) < y_{(i)} + hx_i < x_i^2 + x_i$$

Como $x_i = ih$

$$y_{(i+1)} = (ih)^2 + ih^2 = i(i+1)h^2 < (i+1)^2h^2 = x_{i+1}^2$$

consecuentemente, la sucesión que define la ordenada de los vértices de los polígonos de Euler es siempre

$$y_{(i+1)} = y_i + ih - \frac{4y_i}{ih}$$

así tendremos sucesivamente

- $y_5 = \frac{16h^2}{3} + h\left(8h - \frac{16h}{3}\right) = \frac{16h^2}{3} + \frac{8h^2}{3} = \frac{24h^2}{3} = 8h^2$
- $y_6 = \frac{24h^2}{3} + h\left(10h - \frac{32h}{5}\right) = \frac{24h^2}{3} + \frac{18h^2}{5} = \frac{58h^2}{5}$
- $y_7 = \frac{58h^2}{5} + h\left(12h - \frac{116h}{15}\right) = \frac{58h^2}{5} + \frac{64h^2}{15} = \frac{238h^2}{15}$
- $y_8 = \frac{238h^2}{15} + h\left(14h - \frac{136h}{15}\right) = \frac{238h^2}{15} + \frac{74h^2}{15} = \frac{312h^2}{15}$
- $y_9 = \frac{104h^2}{5} + h\left(16h - \frac{52h}{5}\right) = \frac{104h^2}{5} + \frac{28h^2}{5} = \frac{132h^2}{5}$
- $y_{10} = \frac{132h^2}{5} + h\left(18h - \frac{176h}{15}\right) = \frac{132h^2}{5} + \frac{94h^2}{15} = \frac{98h^2}{3}$ ...

Tomando $h = 0.1$ obtenemos la tabla

x	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
y	0	0	0.02	0.02	0.0534	0.08	0.116	0.15867	0.208	0.264	0.3267

Vamos ahora a utilizar Matlab para calcular y representar gráficamente los polígonos de Euler. En la Figura 6.21, hemos representado uno de esos polígonos y la parábola $y = x^2$ para hacer evidente que todos los vértices están en IV. Para evitar la división por cero hemos cambiado ligeramente la condición inicial que ahora es $y(0.01) = 0$; con ello el programa Matlab *Euler.m* que incluimos en la página web del libro junto con el *singular.m* de la función 'singular' llamada por el programa, nos da la siguiente tabla de valores

i	0	1	2	3	4
x_i	0.010	0.109	0.208	0.307	0.406
y_i	0	0.001980	0.016369	0.026389	0.053136

5	6	7	8	9	10
0.505	0.604	0.703	0.802	0.901	1.000
0.081697	0.117623	0.160098	0.209109	0.264654	0.326733

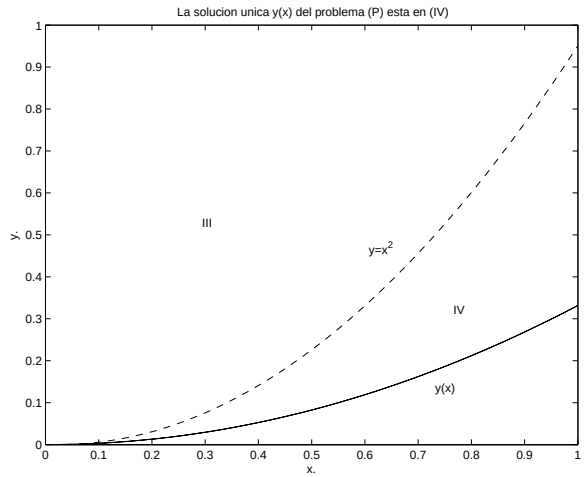


Figura 6.21: Representación de la solución única del problema de Cauchy (P).

que corrobora lo calculado antes manualmente.

PROBLEMA 6.11 *Estudio numérico de un problema de Cauchy 1D por varios métodos.*

Se considera el problema de Cauchy

$$(P) \quad \begin{cases} y' = xe^{(x^2-y)} & x \in I\!R \\ y(1) = 0 \end{cases}$$

1. Resolver exactamente el problema propuesto.

Se desea resolver ahora (P) utilizando dos métodos predictor/corrector comparando los resultados de ambos métodos con la solución exacta y entre sí.

Se calcularán los valores iniciales y_1, y_2 e y_3 mediante un Runge-Kutta de cuarto orden.

2. Resolver el problema (P) usando el método de Runge-Kutta aludido.

3. Resolver el problema (P) usando el método predictor/corrector Adams-Bashforth-Moulton de tres pasos explícito/implícito, $ABM33$.

4. Utilizar ahora el método $ABM43$. El predictor es ahora un método explícito Adams-Bashforth de cuatro pasos y el corrector es un método Adams implícito de tres pasos.

   Se dispondrán los resultados en una tabla donde se incluyan los valores obtenidos en los tres métodos aplicados en el intervalo $[1, 3]$ con un paso $h = 0.1$ incluyendo los valores dados por el predictor en los métodos ABM de los apartados 3 y 4.

5. Comparar los resultados obtenidos.

   Ya que tenemos la solución exacta del problema, se utilizará como herramienta en la comparación los errores globales de discretización en cada $x_k = 1 + k \cdot 0.1 \quad k = 0, ..., 20$ para los métodos considerados y el error global final en $x = 3$.

6. Utilizar ahora la "fuerza" de Matlab construyendo programas de los métodos RK4, ABM33 y ABM43 que den como resultado para cada abscisa x_k de la malla discreta tanto y_k^p como y_k.

Solución:

1. La ecuación diferencial es de variables separables

$$\frac{dy}{dx} = x\frac{e^{x^2}}{e^y} \quad \Rightarrow \quad e^y dy = xe^{x^2}dx$$

El problema (P) posee una solución única definida implícitamente por

$$\int_0^y e^t dt = \int_1^x te^{t^2}dt \quad \Rightarrow \quad = e^y - 1 = \frac{1}{2}\int_1^{x^2} e^u du = \frac{1}{2}\left(e^{x^2} - e\right)$$

de donde

$$y = \ln\left(1 + \frac{1}{2}\left(e^{x^2} - e\right)\right)$$

que es la solución pedida.

Concentramos todos los resultados de los apartados 2, 3 y 4 en la tabla 6.2.

2. En la columna relativa a Runge-Kutta de la Tabla 6.2 en la página 314 tenemos por ejemplo

$$\begin{aligned} k_1 &= (0.1)f(1,0) && \Rightarrow k_1 \sim 0.2718282 \\ k_2 &= (0.1)f(1.05, \frac{k_1}{2}) && \Rightarrow k_2 = \sim 0.2760401 \\ k_3 &= (0.1)f(1.05, \frac{k_2}{2}) && \Rightarrow k_3 = \sim 0.2754594 \\ k_4 &= (0.1)f(1.1, k_3) && \Rightarrow k_4 \sim 0.2800648 \end{aligned}$$

de modo que

$$\begin{aligned} y_1(1.1) = 0.1666667\,(&0.2718282 + 0.5520803 + \\ &+0.5509188 + 0.2800648) \quad \Rightarrow \quad y_1(1.1) \sim 0.2758153 \end{aligned}$$

3. En la tercera columna relativa al predictor del método $ABM33$,

$$\begin{aligned} y_4^p =& y_3 + \frac{1}{12}h\,(23f_3 - 16f_2 + 5f_1) = \\ =& 0.8546775 + \frac{0.1}{12}\,(68.9361650 - 46.2749278 + \\ &+13.9982560) \quad \Rightarrow \quad y_4^p \sim 1.1601733 \end{aligned}$$

de modo que en la columna cuarta del corrector,

$$\begin{aligned} y_4 =& y_3 + \frac{1}{24}h\,(9f(x_4, y_4^p) + 19f_3 - 5f_2 + f_1) = \\ =& 0.8546775 + \frac{0.1}{24}\,(28.0369557 + 56.9472667 - \\ &-14.4609149 + 2.7996512) \quad \Rightarrow \quad y_4 \sim 1.1601899 \end{aligned}$$

4. En la quinta columna relativa al predictor del método $ABM43$,

$$\begin{aligned} y_4^p =& y_3 + \frac{1}{24}h\,(55f_3 - 59f_2 + 37f_1 - 9f_0) = \\ =& 0.8546775 + \frac{0.1}{24}\,(164.8473511 - 170.6387963 + \\ &+103.5870944 - 24.4645364) \quad \Rightarrow \quad y_4^p \sim 1.1602238 \end{aligned}$$

		(RK4)	(ABM33)	(ABM33)	(ABM43)	(ABM43)
x_n	$y(x_n)$	y_n	y_n^p	y_n	y_n^p	y_n
1.0	0	0				
1.1	0.2758130	0.2758153		0.2758153		0.2758153
1.2	0.5603053	0.5603010		0.5603010		0.5603010
1.3	0.8546704	0.8546775		0.8546775		0.8546775
1.4	1.1601847	1.1601946	1.1601733	1.1601899	1.1602238	1.1601840
1.5	1.4781274	1.4781401	1.4781410	1.4781286	1.4781539	1.4781227
1.6	1.8097251	1.8097411	1.8097534	1.8097218	1.8097365	1.8097190
1.7	2.1561148	2.1561344	2.1561520	2.1561068	2.1561217	2.1561086
2.0	3.2936097	3.2936426	3.2936513	3.2935920	3.2936073	3.2936068
2.4	5.0645868	5.0646421	5.0646097	5.0645702	5.0645817	5.0645891
2.9	7.7166929	7.7167821	7.7166971	7.7166869	7.7166911	7.7166953
3.0	8.3067642	8.3068609	8.3067669	8.3067598	8.3067629	8.3067661

Cuadro 6.2: Tabla resumen de los valores aproximados obtenidos en todos los métodos considerados.

de modo que en la columna última del corrector,

$$y_4 = y_3 + \frac{1}{24}h\left(9f(x_4, y_4^p) + 19f_3 - 5f_2 + f_1\right)$$

con

$$f(x_4, y_4^p) = f(1.4, 1.1602238) = 3.1150599$$

luego

$$\begin{aligned} y_4 =& 0.8546775 + \frac{0.1}{24}\left(28.0355392 + 56.9472667 - \right. \\ & \left. -14.4609149 + 2.7996512\right) \quad \Rightarrow \quad y_4 = \sim 1.1601840 \end{aligned}$$

Resumimos todos los cálculos en la Tabla 6.2. Para obtener la segunda columna usando Matlab se puede escribir directamente en la ventana de comandos

```
%format long
%x=1:0.1:3;
%y=log(1.+(1./2).*[(exp(x.*x))-exp(1)])
```

o bien definir la función 'pvi1'

```
%function y=pvi1(x)
%y=log(1.+(1./2).*[(exp(x.*x))-exp(1)]);
```

en un programa *pvi1.m* que se llama desde la ventana de comandos poniendo

```
%pvi1(1:0.1:3)
```

y que da como respuesta la segunda columna de la tabla con 15 decimales.

5. Ya que conocemos la solución exacta es fácil hacer un análisis de los errores globales de discretización cometidos en cada uno de los métodos.

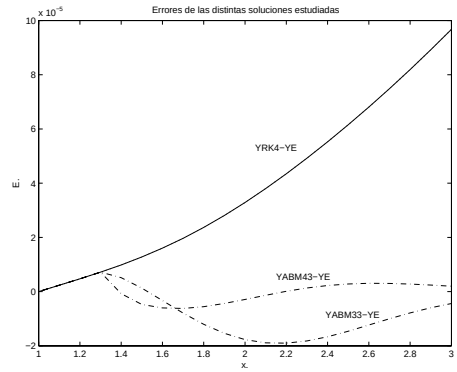


Figura 6.22: Representación gráfica de los errores de las soluciones aproximadas del problema (P) por los métodos RK4, ABM33 y ABM43.

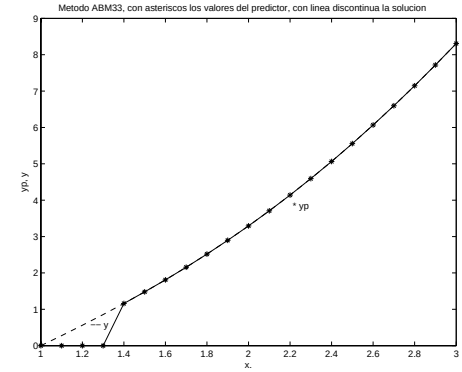


Figura 6.23: Representación gráfica de la solución aproximada del problema de Cauchy (P) por el método ABM33 y de los valores correspondientes del predictor.

El error global final en cada uno de ellos es

$$\begin{aligned} \text{egf(RK4)} &= 0.967607 \cdot 10^{-4} \\ \text{egf(ABM33)} &= -\,0.043895 \cdot 10^{-4} \\ \text{egf(ABM43)} &= 0.019803 \cdot 10^{-4} \end{aligned}$$

y se comprueba el mejor comportamiento del último método. Hemos representado esos resultados gráficamente y la Figura correspondiente (6.22) es muy ilustrativa de dicho comportamiento.

6. El código Matlab del método Adam-Bashforth-Moulton 43 que hemos programado es *abm43.m* y en el archivo *probCauchy63.m* hemos construido la función f a la que llama el programa anterior. Ambos programas están en la página web del libro.

   Se dispone el resultado en la tabla de la izquierda, en tres columnas relativas al vector de abscisas, al vector de los valores dados por el predictor y a la solución aproximada del método abm43. Del mismo modo, el resultado de correr el programa *abm33.m* que incluimos en el servidor está en la tabla de la derecha.

x_n	$y^p(x_n)$	$y(x_n)$
1.0		0
1.1		0.27581534188476
1.2		0.56030998037223
1.3		0.85467754908275
1.4	1.16022385183929	1.16018397481733
1.5	1.47815392360818	1.47812263552527
1.6	1.80973654596007	1.80971905790326
2.0	3.29360733293707	3.29360682789381
2.3	4.59321956873760	4.59322617588944
2.6	6.06601578210447	6.06602288047876
2.9	7.71669108148346	7.71669529965986
3.0	8.30676291392845	8.30676615277444

x_n	$y^p(x_n)$	$y(x_n)$
1.0		0
1.1		0.26268290000000
1.2		0.50923570000000
1.3		0.74231000000000
1.4	1.16017332548747	1.16018987718942
1.5	1.47814099417334	1.47812865593499
1.6	1.80975344493819	1.80972176735336
2.0	3.29365133666131	3.29359201762376
2.3	4.59325314675481	4.59320663130288
2.6	6.06603292966484	6.06600749226335
2.9	7.71669714409929	7.71668692267761
3.0	8.30676687536186	8.30675978295891

También incluimos aquí la Figura 6.23, en la que se representan las tres columnas de la tabla de la derecha. Aunque hemos dibujado de modo distinto las curvas de y_p e y, no hay mucha diferencia entre sus puntos en la gráfica.

PROBLEMA 6.12 *Oscilador no lineal de Duffing.*

Se llama oscilador generalizado a la ecuación diferencial

$$\ddot{y} + \varphi(\dot{y}) + f(y) = F(t)$$

donde los puntos sobre la función y representan derivadas respecto al tiempo y por analogía con el caso lineal [25] el término $\ddot{y}$ se llama fuerza de inercia, fuerza de amortiguamiento al término $\varphi(\dot{y})$, memoria del muelle el término $f(y)$ y fuerza de excitación al segundo miembro $F(t)$.
El oscilador que queremos estudiar aquí es un caso particular del anterior cuando la fuerza del muelle, no lineal, viene definida por

$$f(y) = \alpha y + \beta y^3,\ \alpha > 0,$$

y es un ejemplo de una clase de osciladores no lineales que se llaman de Duffing [26] y que están definidos por la ecuación

$$\ddot{y} + \varphi(\dot{y}) + \alpha y + \beta y^3 = F(t)$$

Esta ecuación modela muchos sistemas mecánicos diferentes y se ha estudiado en muy variados contextos, usando todas las técnicas posibles.
Por dar un significado físico a la función incógnita, coeficientes y demás objetos de dicha ecuación, consideraremos uno de esos sistemas mecánicos, un péndulo simple formado por una barra AB inextensible de longitud l y de masa despreciable que lleva en su extremo una masa puntual m y que está articulado sin rozamiento en su otro extremo A fijo (Figura 1). La masa se mueve sin rozamiento bajo la acción de la gravedad sobre

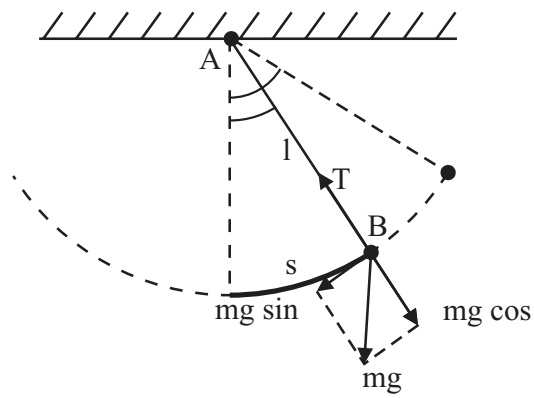


Figura 6.24: Péndulo simple.

la circunferencia de centro A y radio l situada en el plano vertical. Las fuerzas que actúan sobre la masa son su propio peso [27] y la tensión T en la barra.

La ecuación que gobierna el movimiento del péndulo sobre la trayectoria circular es

$$ml\frac{d^2\theta}{dt^2} = -mg\sin\theta$$

[25]

$$\ddot{y} + c\dot{y} + ky = F(t)$$

[26] La ecuación

$$\ddot{y} + c\dot{y} + y^3 = B\cos t$$

fue establecida en 1918 por el ingeniero alemán Georg Duffing. Duffing, molesto por los inconvenientes producidos en una máquina industrial por las vibraciones, acortamiento del tiempo de vida de las piezas afectadas por la vibración y generación de ruido, introdujo en la ecuación del movimiento del oscilador lineal un término cúbico de rigidez no lineal para modelar las vibraciones forzadas de la máquina.

[27] De componente tangencial $-mg\sin\theta$ y de componente normal $mg\cos\theta$.

Si el péndulo se abandona en reposo en el instante $t = 0$ cuando $\theta = \theta_0$, el problema de valor inicial correspondiente es

$$\begin{cases} \ddot{\theta} = -\dfrac{g}{l}\sin\theta \\ \theta(0) = \theta_0 \text{ y } \dot{\theta}(0) = 0 \end{cases}$$

Poniendo $y = \frac{\theta}{\theta_0}$ para escalar el ángulo de oscilación y $\omega = \sqrt{\frac{g}{l}}$ obtenemos el problema de Cauchy

$$\begin{cases} \ddot{y} = -\omega^2 \sin(\theta_0 y) \\ y(0) = 1 \text{ y } \dot{y}(0) = 0 \end{cases}$$

Si θ_0 es pequeño se puede aproximar $\sin(\theta_0 y)$ mediante un desarrollo limitado de Mac Laurin, por ejemplo de orden 3

$$\sin(\theta_0 y) \approx \theta_0 y - \frac{(\theta_0 y)^3}{6}$$

y se tiene

$$\frac{d^2y}{dt^2} = -\frac{\omega^2}{\theta_0}\left(\theta_0 y - \frac{(\theta_0 y)^3}{6}\right)$$

Introduciendo un tiempo adimensional[28] $\tau = \omega t$ y llamando $\epsilon = \frac{\theta_0^2}{6}$ que sugiere la pequeñez del término no lineal, obtenemos[29] la ecuación de Duffing

$$\ddot{y} + y - \epsilon y^3 = 0$$

de las oscilaciones libres sin amortiguamiento.
Si consideramos el medio resistente viscoso con una fuerza proporcional a la velocidad de la forma $k\dot{y}$ con $k > 0$ que se opone al movimiento y una fuerza de excitación $F(t)$ obtenemos la ecuación

$$\ddot{y} + k\dot{y} + y - \epsilon y^3 = F(t)$$

de las oscilaciones forzadas con amortiguamiento más general.

1. Consideramos el problema de Cauchy

$$(P) \quad \begin{cases} (E) \quad \ddot{y} + y + (0.1)y^3 = 0 \\ y(0) = 1 \text{ y } \dot{y}(0) = 0 \end{cases}$$

A causa del pequeño término no lineal $(0.1)y^3$, no hay solución expresable mediante funciones elementales por lo que el estudio del problema (P) se deberá resolver numéricamente.

1.1 Reescribir el problema (P) como un problema de Cauchy

$$(P)' \begin{cases} \dot{y}_1 = f_1(t, y_1, y_2) \\ \dot{y}_2 = f_2(t, y_1, y_2) \end{cases}$$
$$(CI) \quad y_1(0) = y_1^{(0)} \quad \text{y} \quad y_2(0) = y_2^{(0)}$$

$(P)'$ para un sistema de dos edos de primer orden.

En los apartados siguientes resolveremos numéricamente el problema $(P)'$ en el intervalo $[0, 1]$ por métodos diferentes. En todos los casos se tomará una partición de puntos $t_i = ih$ equiespaciados con el paso $h = 0.1$.

Escribiremos

$$y_j^{(i)} = y_j(t_i) \quad (j = 1, 2)$$
$$f_j^{(i)} = f_j(t_i, y_1^{(i)}, y_2^{(i)})$$

[28] ω tiene dimensión $\frac{1}{T}$.

[29] $y(t) = y\left(\frac{\tau}{\omega}\right) = Y(\tau)$ y derivando, $\frac{dY}{d\tau} = \frac{dy}{dt}\frac{1}{\omega}$. Denotando como es abuso habitual $Y = y$; $\frac{dy}{d\tau} = \frac{dy}{dt}\frac{1}{\omega}$ y $\frac{d^2y}{d\tau^2} = \frac{d^2y}{dt^2}\frac{1}{\omega^2}$ derivadas que seguiremos denotando con un punto encima.

1.2 Utilizar el método de Euler para resolver numéricamente el problema $(P)'$ y disponer los valores calculados en una tabla del tipo

i	t_i	$y_1^{(i)}$	$\Delta y_1^{(i)}$	$y_2^{(i)}$	$\Delta y_2^{(i)}$

con $\Delta y_j^{(i)} = h f_j(t_i, y_1^{(i)}, y_2^{(i)}) \quad j = 1, 2$

1.3 Utilizar ahora el método de Euler modificado

$$y_j^{(i+1)} = y_j^{(i)} + h f_j^{(i+1/2)} \quad (j = 1, 2)$$

donde

$$y_j^{(i+1/2)} = y_j^{(i)} + \frac{h}{2} f_j^{(i)} \quad (j = 1, 2)$$
$$f_j^{(i+1/2)} = f_j(t_{i+1/2}, y_1^{(i+1/2)}, y_2^{(i+1/2)})$$
$$t_{(i+1/2)} = t_i + \frac{h}{2}$$

y disponer los resultados en una tabla del tipo

i	t_i	$y_1^{(i)}$	$f_1^{(i)}$	$y_1^{(i+1/2)}$	$f_1^{(i+1/2)}$	$f_2^{(i+1/2)}$	$y_2^{(i+1/2)}$	$f_2^{(i)}$	$y_2^{(i)}$

1.4 Utilizar por último el código ODE45 que forma parte de las herramientas de Matlab con un test de parada $\epsilon < 10^{-5}$.

Llamaremos y_E, y_{EM} e y_{ode45} a la aproximación a la solución única de $(P)'$ obtenida en cada uno de los tres métodos considerados.

1.5 Comparar las tres soluciones aproximadas obtenidas.

2. Un enfoque bastante natural para el estudio numérico del problema (P) sería eliminar el término no lineal sustituyéndolo por el término $(0.1)y_1^3$ donde y_1 es la solución única del problema de Cauchy lineal elemental

$$(P1) \quad \begin{cases} (E1) \quad \ddot{y_1} + y_1 = 0 \\ y_1(0) = 1 \text{ y } \dot{y_1}(0) = 0 \end{cases}$$

resolviendo después el problema de Cauchy resultante

$$(P2) \quad \begin{cases} (E2) \quad \ddot{y_2} + y_2 + (0.1)y_1^3 = 0 \\ y_2(0) = 1 \text{ y } \dot{y_2}(0) = 0 \end{cases}$$

Ese término será razonablemente próximo al término rechazado y consecuentemente la solución del problema de Cauchy $(P2)$ será razonablemente próxima a la solución del problema (P) en estudio.

2.1 Resolver analíticamente el problema reducido $(P1)$. Llamaremos y_1 la solución única de $(P1)$.

2.2 Resolver analíticamente el problema $(P2)$. Llamaremos y_2 la solución única de $(P2)$.

2.3 Comparar esta solución aproximada y_2 con las obtenidas mediante diversos métodos en el apartado 1 y escribir los resultados en una tabla.

3. Analizamos ahora el problema de Cauchy

$$(P_\epsilon) \quad \begin{cases} (EDH) \quad \ddot{y} + y + \epsilon y^3 = 0 \\ (CI) \quad y(0) = 1 \text{ y } \dot{y}(0) = 0 \end{cases}$$

con ϵ pequeño, utilizando técnicas de los métodos de perturbación.

Supongamos que las soluciones de la ecuación de Duffing (EDH) que interviene en (P_ϵ) se pueden escribir en la forma [30]

$$y(t,\epsilon) \simeq y_0(t) + \epsilon y_1(t) + \epsilon^2 y_2(t) + \cdots$$

Sustituyendo $y(t,\epsilon)$ en (EDH) y desarrollando se llega a

$$(\ddot{y_0}(t) + y_0(t)) + \epsilon\left[\ddot{y_1}(t) + y_1(t) + y_0^3(t)\right] + \epsilon^2\left[\ddot{y_2}(t) + y_2(t) + 3y_0^2(t)y_1(t)\right] + \cdots = 0$$

y limitándonos al desarrollo de orden 2,

$$(\ddot{y_0}(t) + y_0(t)) + \epsilon\left(\ddot{y_1}(t) + y_1(t) + y_0^3(t)\right) + O(\epsilon^2) = 0$$

Con las condiciones iniciales (CI) de (P_ϵ) tenemos los problemas de Cauchy parciales relativos a cada una de las "componentes" $y_0(t)$ e $y_1(t)$ de $y(t)$ siguientes

$$(P_\epsilon)_0 \quad \begin{cases} \ddot{y_0}(t) + y_0(t) = 0 \\ y_0(0) = 1 \text{ y } \dot{y_0}(0) = 0 \end{cases}$$

$$(P_\epsilon)_1 \quad \begin{cases} \ddot{y_1}(t) + y_1(t) + y_0^3(t) = 0 \\ y_1(0) = 0 \text{ y } \dot{y_1}(0) = 0 \end{cases}$$

3.1 Resolver ambos problemas y obtener con las correspondientes soluciones la aproximación

$$y(t,\epsilon) \simeq y_0(t) + \epsilon y_1(t)$$

3.2 Comparar este método con el que hemos usado "razonablemente" en el apartado anterior.

3.3 Hacer en particular $\epsilon = 0.1$ y comparar $y(t, 0.1)$ con y_2.

4. Estudiar el comportamiento asintótico de las soluciones y_2 e $y(t, 0.1)$.

Solución:

1. 1.1 Haciendo en la ecuación (E) el cambio de variables $y = y_1$, $\dot{y} = y_2$ se llega al sistema de ecuaciones diferenciales de primer orden equivalente

$$\begin{cases} \dot{y_1} = y_2 \\ \dot{y_2} = -y_1 - (0.1)y_1^3 \end{cases} \Rightarrow \begin{cases} f_1(t, y_1, y_2) = y_2 \\ f_2(t, y_1, y_2) = -y_1 - (0.1)y_1^3 \end{cases}$$

Las condiciones iniciales que definen junto al sistema anterior el problema de Cauchy $(P)'$ son

$$(CI) \quad y_1(0) = 1; \quad y_2(0) = 0$$

1.2 El algoritmo del método de Euler

$$y_j^{(i+1)} = y_j^{(i)} + h \cdot f_j^{(i)} \qquad (j = 1, 2)$$

con $h = 0.1$ nos permite ir llenando poco a poco los cuadros de la tabla sugerida en el enunciado

i	t_i	$y_1^{(i)}$	$\Delta y_1^{(i)} = (0.1)y_2^{(i)}$	$y_2^{(i)}$	$\Delta y_2^{(i)} = -y_1^{(i)} - (0.1)(y_1^{(i)})^3$
0	0	1	0	0	-0.1100
1	0.1	1	-0.01100	-0.1100	-0.2200
2	0.2	0.9890	-0.02200	-0.2200	-0.1086
3	0.3	0.9670	-0.03286	-0.3286	-0.1057
4	0.4	0.9341	-0.04343	-0.4343	-0.1016
5	0.5	0.8907	-0.05359	-0.5359	-0.0961
6	0.6	0.8371	-0.06320	-0.632	-0.0896
7	0.7	0.7739	-0.07216	-0.7216	-0.0820
8	0.8	0.7017	-0.08036	-0.8036	-0.0736
9	0.9	0.6213	-0.08772	-0.8772	-0.0645
10	1.0	0.5336	-0.09417	-0.9417	

[30] Las notaciones y_i , $i = 0, 1, 2, ...$ que usamos en el desarrollo de $y(t,\epsilon)$ son independientes de las del anterior apartado y se usarán sólo aquí.

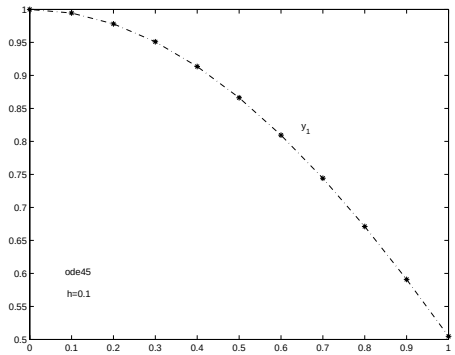


Figura 6.25: $t \to y_1(t)$.

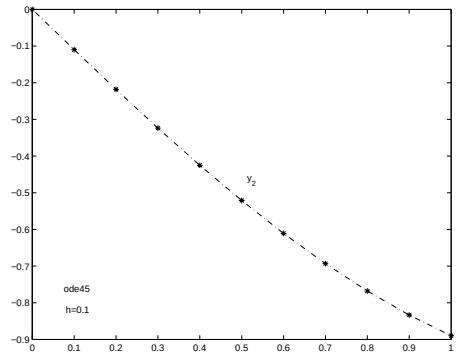


Figura 6.26: $t \to y_2(t)$.

1.3 Con el algoritmo del método de Euler modificado

$$y_j^{(i+1)} = y_j^{(i)} + h \cdot f_j\left(t_{i+1/2}, y_1^{(i+1/2)}, y_2^{(i+1/2)}\right) \qquad (j = 1, 2)$$

$h = 0.1$ y las notaciones del enunciado obtenemos

i	t_i	$y_1^{(i)}$	$f_1^{(i)}$	$y_1^{(i+1/2)}$	$f_1^{(i+1/2)}$	$f_2^{(i+1/2)}$	$y_2^{(i+1/2)}$	$f_2^{(i)}$	$y_2^{(i)}$
0	0	1	0	1	-0.0550	-1.1	-0.0550	-1.1	0
1	0.1	0.9945	-0.1100	0.9890	-0.2186	-1.0857	-0.2186	-1.0929	-0.1100
2	0.2	0.9726	-0.2186	0.9617	-0.2718	-1.0506	-0.2718	-1.0646	-0.2186
3	0.3	0.9454	-0.3237	0.9292	-0.3752	-1.0094	-0.3752	-1.0299	-0.3237
4	0.4	0.9079	-0.4246	0.8867	-0.4737	-0.9564	-0.4737	-0.9827	-0.4246
5	0.5	0.8605	-0.5202	0.8345	-0.5664	-0.8926	-0.5664	-0.9242	-0.5202
6	0.6	0.8039	-0.6095	0.7387	-0.6523	-0.7790	-0.6523	-0.8558	-0.6095
7	0.7	0.7260	-0.6874	0.6916	-0.7256	-0.7247	-0.7256	-0.7643	-0.6874
8	0.8	0.6534	-0.7599	0.6137	-0.7940	-0.6368	-0.7940	-0.6813	-0.7599
9	0.9	0.5740	-0.8236	0.5328	-0.8532	-0.5479	-0.8532	-0.5930	-0.8236
10	1.0	0.4887	-0.8784	0.4448	-0.9034	-0.4536	-0.9034	-0.5004	-0.8784

1.4 Usamos ahora el método *ode45* que Matlab sugiere como primera opción para abordar la resolución de cualquier problema de Cauchy. Se trata de un código que implementa un esquema Runge-Kutta (RK45) explícito de un paso.

Manteniendo el paso $h = 0.1$, con un test de parada $\epsilon < 10^{-5}$ y *format long*, el resultado que nos da este programa es

i	t_i	$y_1^{(i)}$	$y_2^{(i)}$
0	0	1.00000000000000	0
1	0.1	0.99450595258304	-0.10976201245348
2	0.2	0.97809497484660	-0.21810402900052
3	0.3	0.95097858500836	-0.32364541613511
4	0.4	0.91350286254206	-0.42508122790317
5	0.5	0.86613928689858	-0.52121294614842
6	0.6	0.80947286964300	-0.61097158734733
7	0.7	0.74418840919849	-0.69343202626811
8	0.8	0.67105576629605	-0.76781829374114
9	0.9	0.59091503787402	-0.83350048249317
10	1.0	0.50466240959774	-0.88998463085790

que representamos gráficamente en las Figuras 6.25 y 6.26

1.5 Incluimos en la próxima tabla los valores en los puntos de la malla de discretización t_i de las soluciones aproximadas obtenidas en los tres métodos que hemos utilizado con cuatro cifras significativas.

i	t_i	$y_E^{(i)}$	$y_{EM}^{(i)}$	$y_{ode45}^{(i)}$
0	0	1.0000	1.0000	1.0000
1	0.1	1.0000	0.9945	0.9945
2	0.2	0.9890	0.9726	0.9781
3	0.3	0.9670	0.9454	0.9510
4	0.4	0.9341	0.9079	0.9135
5	0.5	0.8907	0.8605	0.8661
6	0.6	0.8371	0.8039	0.8095
7	0.7	0.7739	0.7260	0.7442
8	0.8	0.7017	0.6534	0.6711
9	0.9	0.6213	0.5740	0.5909
10	1.0	0.5336	0.4887	0.5047

representaremos gráficamente esos valores en la Figura 6.27 para comparar de forma visual los resultados obtenidos.

2. Una vez analizados los resultados numéricos del problema de Cauchy (P), seguiremos ahora la primera aproximación propuesta en el enunciado comenzando por resolver de modo analítico los dos problemas de Cauchy que planteamos en el proceso de aproximación.

2.1 En el caso del problema $(P1)$ y siguiendo el proceso habitual, la solución general de la ecuación $(E1)$ es[31] $y_1 = A\cos t + B\sin t$. Las condiciones iniciales definen las constantes $A = 1$ y $B = 0$, de modo que la solución única de $(P1)$ es $y_1 = \cos t$.

Esa solución se introduce en el problema de Cauchy $(P2)$ cuyo aspecto definitivo es

$$(P2) \quad \begin{cases} (E2) \quad \ddot{y_2} + y_2 + (0.1)(\cos t)^3 = 0 \\ y_2(0) = 1 \text{ y } \dot{y_2}(0) = 0 \end{cases}$$

2.2 Para resolver $(P2)$ buscamos una solución particular de $(E2)$ cuya estructura sea $y_p(t) = A(t)\cos t + B(t)\sin t$. Usando el método de variación de la constante llegamos al sistema

$$\begin{cases} A'(t)\cos t + B'(t)\sin t &= 0 \\ -A'(t)\sin t + B'(t)\cos t &= (0.1)(\cos t)^3 \end{cases}$$

de donde

$$A'(t) = (0.1)(\cos t)^3 \sin t \quad \text{y} \quad B'(t) = -(0.1)(\cos t)^4$$

luego

$$A(t) = \int (0.1)(\cos t)^3 \sin t dt = -\int (0.1)(\cos t)^3 d(\cos t) = -0.1\frac{(\cos t)^4}{4}$$
$$B(t) = \int (0.1)(\cos t)^4 dt = \frac{\sin 4t}{32} + \frac{\sin 2t}{4} + \frac{3t}{8}$$

La solución general de $(E2)$ es[32] por tanto

$$y_2(t) = A\cos t + B(t)\sin t - (0.1)(\cos t)^5 - (0.1)\sin t\left(\frac{\sin 4t}{32} + \frac{\sin 2t}{4} + \frac{3t}{8}\right)$$

[31] La solución general de esa ecuación diferencial homogénea es combinación lineal de las funciones $e^{r_i t}$ donde r_i $i = 1, 2$ son las raíces $\{\pm i\}$ de la ecuación característica $r^2 + 1 = 0$ asociada a la ecuación diferencial, es decir, $e^{\pm it}$. Si se escriben de modo real, esas dos soluciones son $\{\cos t, \sin t\}$, con lo que $y_1 = A\cos t + B\sin t$.

[32] Escribiendo $(\cos t)^4$ en función de los ángulos múltiples se tiene $(\cos t)^4 = \frac{1}{8}(\cos 4t + 4\cos 2t + 3)$.

Imponiendo las condiciones iniciales

$$y_2(t) = (1.1)\cos t - (0.1)(\cos t)^5 - (0.1)\sin t\left(\frac{\sin 4t}{32} + \frac{\sin 2t}{4} + \frac{3t}{8}\right) = 1.0375\cos t -$$
$$- 0.03125\cos 3t - 0.00625\cos 5t - 0.0375\ t\ \sin t - 0.025\sin t\ \sin 2t - 0.003125\sin t\ \sin 4t$$

Los valores que toma esa solución en los soportes t_i de la partición del intervalo $[0, 1]$ que hemos usado se recogen en la tabla siguiente

i	t_i	$y_2^{(i)}$
0	0	1.00000000000000
1	0.1	0.99261657559150
2	0.2	0.97782084564585
3	0.3	0.95517986049063
4	0.4	0.92407462742489
5	0.5	0.88386217952898
6	0.6	0.83405054490637
7	0.7	0.77444940591655
8	0.8	0.70526578799184
9	0.9	0.62712752906095
10	1.0	0.54103425767739

2.3 Hemos representado en la misma Figura 6.27, las tres soluciones aproximadas del problema de Cauchy (P) calculadas en el apartado 1. de las que sabemos que la menos ajustada es la de Euler y que la más próxima a la solución es la y_{ode45} en la que hemos impuesto una tolerancia absoluta menor que 10^{-5}. También hemos incluido la solución y_2 del problema $(P2)$ que, como se aprecia en la figura, está muy próxima a la solución de Euler. Los errores absolutos entre las distintas

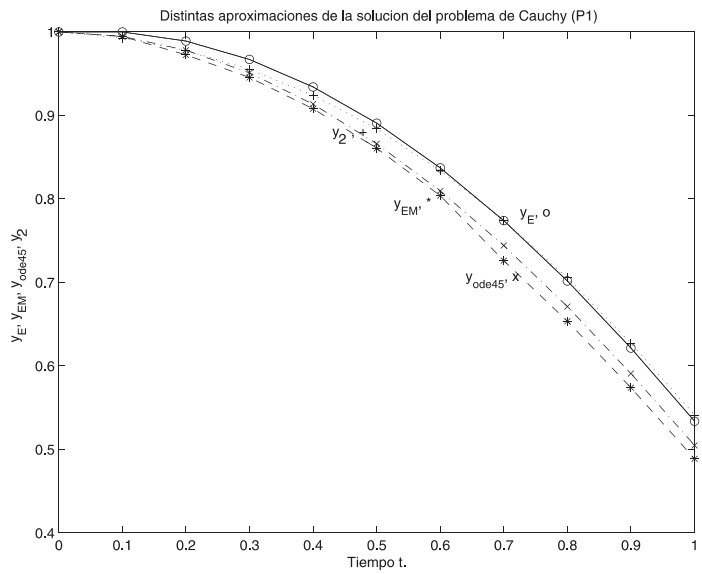


Figura 6.27: Distintas soluciones aproximadas del problema (P), y_E (Euler); y_{EM} (Euler modificado); y_{ode45}, comparadas entre sí y con la solución exacta y_2 del problema aproximante (P2).

soluciones aproximadas y esta solución exacta que se incluyen en la tabla siguiente corroboran estas impresiones.

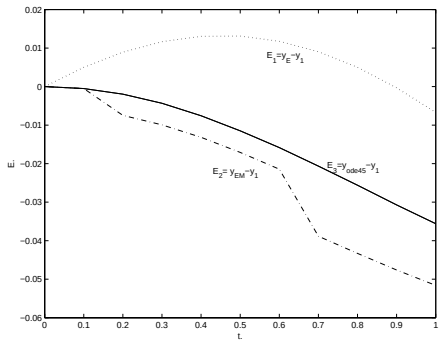


Figura 6.28: E_i **con** $i = 1, 2, 3$**.**

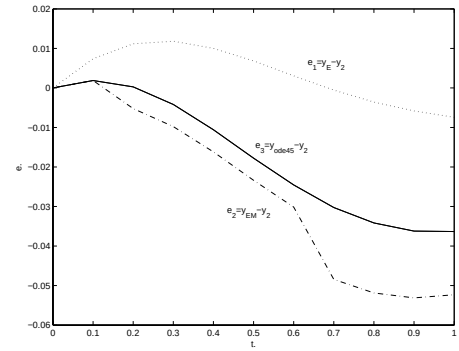


Figura 6.29: e_i **con** $i = 1, 2, 3$**.**

$e_1 = y_E - y_2$	$e_2 = y_{EM} - y_2$	$e_3 = y_{ode45} - y_2$
0	0	0
0.0074	0.0019	0.0019
0.0112	-0.0052	0.0003
0.0118	-0.0098	-0.0042
0.0100	-0.0162	-0.0106
0.0068	-0.0234	-0.0178
0.0030	-0.0302	-0.0246
-0.0005	-0.0484	-0.0302
-0.0036	-0.0519	-0.0342
-0.0058	-0.0531	-0.0362
-0.0074	-0.0523	-0.0363

Utilizando la norma del máximo tenemos $\|e_1\| = 0.0118$, $\|e_2\| = 0.0531$ y $\|e_3\| = 0.0363$ con lo que la solución de Euler es la más próxima a la solución exacta del problema $(P2)$.

Comparando de nuevo con la norma del máximo esas soluciones aproximadas con la solución $y_1 = \cos t$ del problema $(P1)$ tenemos llamando $E_1 = y_E - y_1$, $E_2 = y_{EM} - y_1$ y $E_3 = y_{ode45} - y_1$ que $\|E_1\| = 0.0131$, $\|E_2\| = 0.0516$ y $\|E_3\| = 0.03563$ luego la solución de Euler es también la más próxima a la solución exacta del problema $(P1)$. Es interesante ver que y_1 e y_2 están muy próximas. En caso de necesidad se comprende lo conveniente que sería trabajar con la primera.

No es muy evidente que y_2 sea más cercana a y que y_1 [33]. Eligiendo de las soluciones aproximadas la y_{ode45} que es la más correcta, es difícil tomar una decisión. Sí que es cierto que ambos errores crecen con t como es patente en las Figuras 6.28 y 6.29.

3. Estudiemos ahora la sucesión de problemas de Cauchy planteados por el método de perturbación.

   3.1 El problema de Cauchy $(P_\epsilon)_0$ es idéntico al problema $(P1)$, luego posee la misma solución, es decir, $y_0(t) = \cos t$, y el problema $(P_\epsilon)_1$ es de modo preciso

$$(P_\epsilon)_1 \quad \begin{cases} (E_\epsilon) \quad \ddot{y_1}(t) + y_1(t) + \cos^3(t) = 0 \\ y_1(0) = 0 \text{ y } \dot{y_1}(0) = 0 \end{cases}$$

   El cálculo de la solución única de este problema es similar a la del problema $(P2)$. La solución general de la ecuación homogénea es $y_h = A\cos t + B\sin t$ y aquí es posible buscar la solución particular de la ecuación completa utilizando el método rápido de Euler. Para ello expresamos $\cos^3 t$ en función de los ángulos múltiples, se tiene $(\cos t)^3 = \dfrac{1}{4}(\cos 3t + 3\cos t)$.

[33] La aproximación de orden más bajo es aquí suficiente, como en la mayoría de los problemas de ingeniería, pero hay que conocer las limitaciones de esta afirmación.

La ecuación diferencial (E_ϵ) se escribe

$$(E_\epsilon) \quad \ddot{y_1}(t) + y_1(t) = -\frac{1}{4}\left(\cos 3t + 3\cos t\right)$$

La integral particular será suma de la correspondiente al término $\cos 3t$ y de la correspondiente al término $\cos t$. En el primer caso ensayamos una solución particular del tipo $a\cos 3t + b\sin 3t$. Ya que la ecuación no tiene término en la derivada primera, $b = 0$, y sustituyendo

$$-9a\cos 3t + a\cos 3t = -\frac{1}{4}\cos 3t \quad \Rightarrow \quad a = \frac{1}{32}$$

Como $\cos t$ es solución de la ecuación homogénea debemos ensayar una solución de la forma $ct\cos t + dt\sin t$ y ya que no aparece $\dot{y_1}(t)$ en la ecuación, $c = 0$, luego

$$2d\cos t = -\frac{3}{4}\cos t \quad \Rightarrow \quad d = \frac{-3}{8}$$

La solución general de (E_ϵ) es

$$y_1(t) = A\cos t + B\sin t + \frac{1}{32}\cos 3t - \frac{3}{8}t\sin t$$

Obligando a las condiciones iniciales que aquí son homogéneas, obtenemos

$$A = -\frac{1}{32} \quad \text{y} \quad B = 0$$

luego

$$y_1(t) = \frac{1}{32}\left(\cos 3t - \cos t\right) - \frac{3}{8}t\sin t$$

y por fin tomamos como aproximación de la solución única del problema (P_ϵ)

$$y(t,\epsilon) \simeq \cos t + \epsilon\left[\frac{1}{32}\left(\cos 3t - \cos t\right) - \frac{3}{8}t\sin t\right]$$

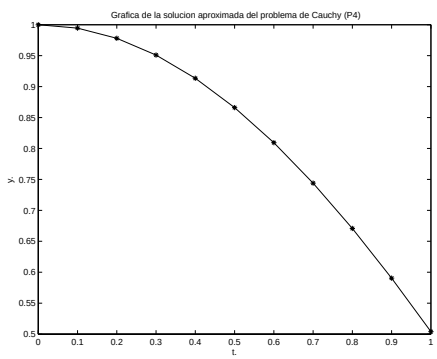


Figura 6.30: Solución aproximada del problema (P_ϵ) por el método de perturbación.

i	t_i	$y_2^{(i)}$
0	0	1.00000000000000
1	0.1	0.99450582847762
2	0.2	0.97809302360112
3	0.3	0.95096899142299
4	0.4	0.91347377124474
5	0.5	0.86607194129085
6	0.6	0.80934197891578
7	0.7	0.74396369708236
8	0.8	0.67070446279240
9	0.9	0.59040492847609
10	1.0	0.50396497268013

Cuadro 6.3: Tabla relativa a la Figura 6.30.

3.2 Los procesos seguidos claramente distintos conducen en ambos casos a una sucesión de problemas de Cauchy resolubles analíticamente muy parecidos. Las soluciones aproximadas del problema (P) obtenidas por ambos métodos son distintas.
En el primer caso es

$$y_2(t) = y_2(t) = 1.0375\cos t - 0.003125\cos 3t - 0.00625\cos 5t - 0.0375t\sin t - \\ - 0.025\sin t\sin 2t - 0.003125\sin t\sin 4t$$

y en el segundo

$$y(t, 0.1) = 0.996875\cos t + 0.003125\cos 3t - 0.0375\cdot t\sin t$$

Ambas soluciones incluyen el término $-0.0375\cdot t\sin t$ de poca influencia para valores pequeños de t pero que no está acotado y a la larga produce efectos importantes como veremos al final del ejercicio. Utilicemos de nuevo Matlab para evaluar y representar gráficamente (ver la

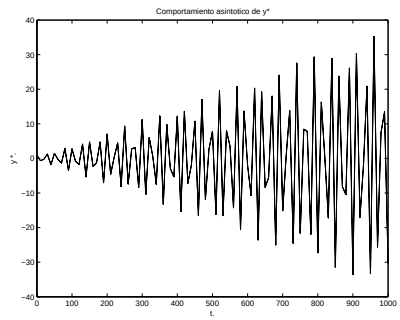


Figura 6.31: Comportamiento asintótico de y_2.

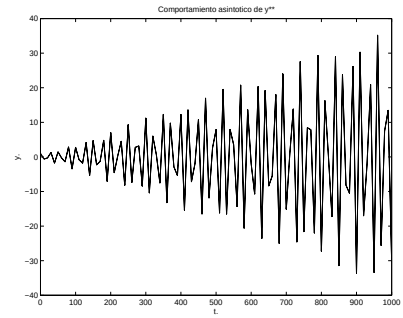


Figura 6.32: Comportamiento asintótico de $y(t, 0.1)$.

Figura 6.30) $y(t, 0.1)$.
Comparando ahora esta solución $y(t, 0.1)$ con la solución y_{ode45} tenemos $||e|| = 7.35027\cdot 10^{-4} < 10^{-3}$ claramente la mejor solución del problema (P) de las dos estudiadas.

4. El término $-0.0375\cdot t\sin t$ común a y_2 e $y(t, 0.1)$, representa oscilaciones de amplitud creciente que tienen sus máximos para $t_k = (2k+1)\dfrac{\pi}{2}$ $k = 0, 1, \ldots$ sobre las rectas $y = \pm t$ y no está acotado.
El resto de términos de ambas soluciones aproximadas son también oscilatorios pero tienen fija su amplitud, de modo que a la larga el término no acotado dominará a los demás.
Para estudiar este comportamiento asintótico de ambas soluciones las evaluaremos dejando transcurrir un tiempo suficientemente grande y analizaremos los resultados.
Las gráficas de y_2 e $y(t, 0.1)$ en el intervalo de tiempo $[0, 10^3]$ se representan en las Figuras 6.31 y 6.32 respectivamente.
La semejanza entre ambas es enorme. El comportamiento oscilatorio con amplitudes crecientes y máximos que aproximadamente están sobre las diagonales de los ejes es común y evidente.

CAPÍTULO 7

EDP's: métodos de diferencias finitas

Las ecuaciones diferenciales en derivadas parciales son la herramienta fundamental de investigación en la física-matemática y por extensión natural en la ingeniería debido a sus excepcionales propiedades en la descripción de los fenómenos físicos cuando dependen de varios parámetros reales.

En la construcción de un modelo matemático de un fenómeno físico que tiene lugar y tiempo en alguna región acotada Ω del espacio-tiempo $I\!R^4(x, y, z, t)$, tras aislar el sistema del exterior identificando las cantidades, densidad, velocidad, temperatura... que lo caracterizan y establecer las conexiones interior-exterior que se juzguen necesarias, se escogen las leyes físicas que gobiernan dicho fenómeno y se escriben en la forma de ecuaciones en derivadas parciales, relaciones entre las características fundamentales del fenómeno y sus ritmos de variación espacial y temporal en un punto del espacio en un instante dado, expresando las condiciones suplementarias que tienen en cuenta la interacción del sistema con el exterior y la prehistoria del fenómeno mediante relaciones entre valores de las magnitudes consideradas y sus derivadas en la frontera de Ω (**condiciones de contorno**) y relaciones entre los valores de las magnitudes consideradas y sus derivadas en el instante en que se inicia el estudio (**condiciones iniciales**).

El problema planteado por estos modelos matemáticos es la busca de las soluciones si existen, de una o de varias ecuaciones en derivadas parciales que además satisfagan las condiciones suplementarias de contorno y/o iniciales.

Sabemos que sólo en contados casos es posible encontrar soluciones analíticas de dichos problemas, por lo que es necesario usar métodos numéricos que aproximen la solución. En este capítulo, consideraremos los métodos en diferencias finitas (MDF) para la resolución numérica de algunos de los problemas matemáticos fundamentales de la física matemática y la ingeniería.

Para precisar el marco básico en el que planteamos la aproximación de esos problemas tipo, es necesario conocer bien la estructura general de los problemas matemáticos en derivadas parciales. Esta estructura es en parte una consecuencia del comportamiento respecto del tiempo de los fenómenos físicos que se modelan.

Denominaremos **problemas matemáticos estacionarios** o problemas de equilibrio aquellos en cuyo enunciado no figura explícitamente el tiempo.

Los fenómenos físicos que modelan son muy variados. Flujos estacionarios o permanentes en mecánica de fluidos. Búsqueda de la configuración de equilibrio de una cierta propiedad.

Llamaremos **problemas matemáticos de evolución** o problemas de propagación aquellos que dependen explícitamente del tiempo.

Los fenómenos en estudio son no estacionarios. Propagación del calor, mecánica de fluidos no viscosos, propagación de las ondas y vibraciones elásticas en elasticidad.

Se desea predecir el comportamiento del sistema a partir de un estado inicial dado. Son pues problemas de valor inicial en los que la solución "avanza" en el tiempo desde el estado inicial "guiada" y modificada por las condiciones de contorno.

La estructura de un problema matemático en derivadas parciales tiene las siguientes componentes

- Un dominio en el que varían las variables geométricas y el tiempo [1].

[1] Los problemas estacionarios habituales en ingeniería se plantean o bien en $I\!R^2(x, y)$ o bien en $I\!R^3(x, y, z)$. Los problemas de evolución en $I\!R^2(x, t)$ o en $I\!R^3(x, y, t)$ y sólo en los problemas más complicados en $I\!R^4(x, y, z, t)$.

- Una ecuación diferencial en derivadas parciales (o un sistema de ecuaciones diferenciales en derivadas parciales) cuya estructura y características dependen mucho del tipo de problema.
- Las condiciones de contorno.
- Las condiciones iniciales en los problemas de evolución.

Repasaremos estos elementos y enunciaremos los problemas matemáticos más importantes antes de abordar la aproximación numérica de sus soluciones por el MDF.

7.1. Ecuaciones en derivadas parciales de primer y de segundo orden

En la inmensa mayoría de los problemas de la Física matemática, es necesario hallar la solución de una ecuación en derivadas parciales lineal o quasi-lineal de primer o segundo orden dependiente de dos variables ((x, y) en los problemas estacionarios 2D y (x, t) en los problemas de evolución 1D) o de tres variables ((x, y, t) en los problemas de evolución 2D), sujeta a ciertas condiciones suplementarias [2].

La ecuación diferencial en derivadas parciales **quasi-lineal** de primer orden en n variables tiene la estructura

$$\sum_{i=1}^{n} a_i(x_1, ..., x_n, u)\frac{\partial u}{\partial x_i} = c(x_1, ..., x_n, u) \tag{7.1}$$

donde las variables independientes $(x_1, ..., x_n)$, las coordenadas espaciales y el tiempo, varían en un abierto Ω de $\mathbb{R}^n$. Una ecuación **quasi-lineal** es **lineal** si las funciones a_i sólo dependen de las variables independientes $(x_1, ..., x_n)$ y $c(x_1, ..., x_n, u) = -a(x_1, ..., x_n)u + f(x_1, ..., x_n)$.

En las ecuaciones en derivadas parciales de segundo orden sólo consideraremos las ecuaciones **lineales** cuya expresión general es

$$\sum_{i,j=1}^{n} a_{ij}(x_1, ..., x_n)\frac{\partial^2 u}{\partial x_i \partial x_j} + \sum_{i=1}^{n} a_i(x_1, ..., x_n)\frac{\partial u}{\partial x_i} + a(x_1, ..., x_n)u = f(x_1, ..., x_n) \tag{7.2}$$

A partir de estas expresiones generales es fácil particularizar a los casos en que $n = 2, 3$ o 4 [3].

Ejemplo 7.1.1 *Definamos algunas de las ecuaciones diferenciales en derivadas parciales más importantes en las aplicaciones*

- *La ecuación de Laplace nD* [4]

$$\Delta u(\mathbf{x}) = \sum_{i=1}^{n} \frac{\partial^2 u}{\partial x_i^2}(\mathbf{x}) = 0 \qquad \mathbf{x} = (x_1, ..., x_n) \in \Omega \tag{7.3}$$

una ecuación en derivadas parciales lineal de segundo orden que gobierna fenómenos estacionarios en $\mathbb{R}^n$, por ejemplo la distribución de la temperatura en el interior de un cuerpo homogéneo e isótropo.

[2] El número de variables espaciales independientes en un problema matemático define su dimensión, así se habla por ejemplo de un problema bidimensional estacionario cuando tenemos dos coordenadas espaciales (x, y). En un problema de evolución 1D las variables independientes son (x, t).

[3] Utilizaremos indistintamente las notaciones

$$u_{x_i} = \frac{\partial u}{\partial x_i}; \quad u_{x_i x_j} = \frac{\partial^2 u}{\partial x_i \partial x_j}$$

[4] Pierre Simon de Laplace (1749-1827) es el último de los grandes matemáticos del siglo XVIII. Hijo de un pequeño propietario de Beaumonten-Auge, Calvados, en Normandía, poco se sabe de su infancia y juventud, ya que siempre intentó ocultar su humilde origen campesino.

Terminados sus estudios elementales en Beaumont y Caen, Laplace viajó a París con la intención de desarrollar allí su carrera matemática. A su llegada, llamó a D'Alembert y le envió sus cartas de recomendación. No recibió respuesta. Decidió entonces enviarle una carta sobre los principios generales de la mecánica. D'Alembert le respondió ofreciéndole su ayuda. Unos días más tarde fue propuesto profesor de matemáticas de la escuela militar de París.

Su flexibilidad en política le permitió desarrollar su actividad matemática independientemente de los convulsos cambios políticos en Francia. Durante la Revolución tomó parte en la organización de la École Normale y de la École Polytechnique y tanto Napoleón como Luis XVIII le concedieron honores.

Sus dos grandes obras son la *Theorie analytique des probabilités*(1812) y la *Mécanique céleste* (1799-1825) culminación de los trabajos de Newton, D'Alembert, Euler y Lagrange. En este tratado de cinco volúmenes aparece la llamada ecuación de Laplace (que ya había sido hallada por Euler en 1752).

- *La ecuación quasi-lineal de transporte o convección 1D*

$$\frac{\partial u}{\partial t}(x,t) + b(x,t,u)u_x(x,t) = 0 \tag{7.4}$$

que gobierna la convección de la propiedad u en un cierto medio con velocidad de convección b. [5]

- *Las ecuaciones de Fourier, de difusión o del calor 1D*

$$u_t - u_{xx} = f \tag{7.5}$$

y 2D

$$u_t - (u_{xx} + u_{yy}) = f \tag{7.6}$$

que gobiernan el flujo de calor o la difusión de un fluido a través de un medio poroso aparecen constantemente en la literatura.

- *La ecuación de las ondas $1D$ o ecuación de las cuerdas vibrantes*

$$u_{tt}(x,t) - u_{xx}(x,t) = f(x,t) \tag{7.7}$$

y la ecuación de las ondas $2D$

$$u_{tt}(x,y,t) - (u_{xx} + u_{yy})(x,y,t) = f(x,y,t) \tag{7.8}$$

ecuaciones de evolución que describe fenómenos de propagación de ondas y que serán objeto de estudio frecuentemente.

7.1.1. Ecuación quasi-lineal de primer orden 1D

Consideremos la ecuación en derivadas parciales lineal de primer orden en 2 variables homogénea

$$a_1(x,t)u_x + a_2(x,t)u_t = 0 \tag{7.9}$$

El ritmo de variación de u para un observador que se mueve en el plano xt con una ley de movimiento $x = x(t)$ tal que

$$x'(t) = \frac{a_1(x,t)}{a_2(x,t)} \tag{7.10}$$

es

$$\frac{du(x(t),t)}{dt} = u_x x'(t) + u_t = u_x \frac{a_1(x,t)}{a_2(x,t)} + u_t \quad \Rightarrow \quad a_2 \frac{du(x(t),t)}{dt} = a_1 u_x + a_2 u_t = 0$$

y el observador percibe que u permanece constante a lo largo de las curvas integrales de la ecuación (7.10).

En el caso general, la ecuación en derivadas parciales lineal de primer orden

$$a_1(x,t)u_x + a_2(x,t)u_t + c(x,t)u = f(x,t) \tag{7.11}$$

se reduce a lo largo de las curvas integrales de (7.10) a la ecuación diferencial ordinaria lineal

$$a_2 \frac{du}{dt} + cu = f$$

Esas curvas integrales se llaman **curvas características** de (7.11) y sobre ellas dicha ecuación es una ecuación diferencial ordinaria.

En el caso quasi-lineal en la ecuación (7.10) interviene además la solución u.

En ambos casos (7.10) siempre tiene solución. Se dice que la ecuación en derivadas parciales quasi-lineal de primer orden **es hiperbólica**, término que se comprende mejor después de estudiar las ecuaciones en derivadas parciales lineales de segundo orden.

[5] La convección es un proceso en el que una propiedad física es arrastrada en el espacio por el movimiento de un medio que lo ocupa. En una corriente líquida, por ejemplo, cada partícula fluida en movimiento arrastra por convección su masa, momento, energía, etc.

7.1.2. Clasificación de las ecuaciones en derivadas parciales lineales de segundo orden y reducción a forma canónica

Consideremos la ecuación en derivadas parciales lineal de segundo orden en dos variables x, y

$$a_{11}(x,y)u_{xx} + 2a_{12}(x,y)u_{xy} + a_{22}(x,y)u_{yy} = \Phi(x,y,u,u_x,u_y) \tag{7.12}$$

con $(x,y) \in \Omega$ y $\Phi(x,y,u,u_x,u_y) = -a_1(x,y)u_x - a_2(x,y)u_y - a(x,y)u + f(x,y)$.

Mediante un cambio de variable [6] $\phi : (x,y) \to (\xi,\eta)$ de clase $\mathcal{C}^2$ en un abierto U de Ω, la ecuación (7.12) se transforma en la ecuación

$$\widehat{a}_{11}v_{\xi\xi} + 2\widehat{a}_{12}v_{\xi\eta} + \widehat{a}_{22}v_{\eta\eta} = \widehat{\Phi}(\xi,\eta,v,v_\xi,v_\eta) \tag{7.13}$$

con $v = u \circ \phi^{-1} \quad \Rightarrow \quad v(\xi,\eta) = u(x(\xi,\eta), y(\xi,\eta)) = u(x,y)$ y

$$\begin{aligned} \widehat{a}_{11} &= a_{11}\xi_x^2 + 2a_{12}\xi_x\xi_y + a_{22}\xi_y^2 \\ \widehat{a}_{12} &= a_{11}\xi_x\eta_x + a_{12}(\xi_x\eta_y + \xi_y\eta_x) + a_{22}\xi_y\eta_y \\ \widehat{a}_{22} &= a_{11}\eta_x^2 + 2a_{12}\eta_x\eta_y + a_{22}\eta_y^2 \end{aligned} \tag{7.14}$$

Las formas de las soluciones de la ecuación en derivadas parciales lineal de segundo orden dependen del signo del discriminante $\Delta = a_{12}^2 - a_{11}a_{22}$. La ecuación (7.12) se clasifica como **hiperbólica, parabólica o elíptica** en un punto (x_0, y_0) de Ω según que $\Delta > 0$, $\Delta = 0$ o $\Delta < 0$ en ese punto.

Dicha ecuación es hiperbólica, parabólica o elíptica en una región del plano xy contenida en Ω si pertenece a ese tipo en todos los puntos de esa región. Si Δ cambia de signo en Ω se dice que es de tipo **mixto**.

Una propiedad importante del discriminante Δ es que su signo es invariante respecto a los cambios de coordenadas en el plano xy, por lo que el tipo de una ecuación diferencial es un invariante respecto de la elección de las variables [7].

Reducción a la forma canónica

- Si la ecuación es hiperbólica se eligen las nuevas variables ξ y η de modo que los coeficientes $\widehat{a}_{11}$ y $\widehat{a}_{22}$ de la ecuación tranformada sean idénticamente nulos. Está claro que de (7.14), ξ y η deben satisfacer la ecuación

$$a_{11}\psi_x^2 + 2a_{12}\psi_x\psi_y + a_{22}\psi_y^2 = 0 \tag{7.15}$$

Las soluciones de estas ecuaciones expresadas de modo explícito $y = f(x)$ son las curvas integrales de las ecuaciones diferenciales ordinarias

$$a_{11}dx^2 + 2a_{12}dxdy + a_{22}dy^2 = 0 \quad \Rightarrow \quad y' = \frac{a_{12} \pm \sqrt{a_{12}^2 - a_{11}a_{22}}}{a_{11}} \tag{7.16}$$

[6] ϕ definida de U sobre $V = \phi(U)$ es biyectiva, bicontinua y de clase $\mathcal{C}^2$, así como su inversa ϕ^{-1}.

[7] Esta clasificación se puede realizar estudiando los valores propios λ_1 y λ_2 de la matriz característica

$$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{12} & a_{22} \end{pmatrix}$$

de la ecuación (7.12).

En efecto, ya que $\Delta = -\det A = \lambda_1\lambda_2$, en el caso elíptico Δ es positivo y ambos valores propios deben ser simultáneamente positivos o negativos. En el caso hiperbólico Δ es estrictamente negativo, luego uno de los valores propios es positivo y el otro negativo. Por último, en el caso parabólico la forma cuadrática asociada es degenerada.

Se usan los valores propios de la matriz característica de la ecuación lineal como lenguaje para generalizar la clasificación anterior a nD.

Consideremos la ecuación en derivadas parciales lineal de segundo orden (7.2). Sean $\mathbf{x}^0$ un punto cualquiera de Ω y $\lambda_i(\mathbf{x}^0)$, los n valores propios reales de la matriz característica $A(\mathbf{x}^0) = a_{ij}(\mathbf{x}^0)$ de (7.2).

Definición 7.1.1 *Se dice que la ecuación (EDPLSO) es elíptica en el punto $\mathbf{x}^0$, si todos los valores propios son positivos o negativos.*

Se dice que la ecuación (EDPLSO) es hiperbólica en el punto $\mathbf{x}^0$ si todos los valores propios son distintos de cero y hay uno de ellos de signo distinto al resto.

Si sólo uno de los valores propios es nulo, siendo el resto del mismo signo, se dice que la ecuación (EDPLSO) es parabólica.

Si escribimos esas soluciones en implícitas $\xi(x,y) = C$ y $\eta(x,y) = C'$ y si $\Delta \neq 0$, la transformación $(x,y) \to (\xi,\eta)$ es un cambio de variable y se produce el efecto deseado $\widehat{a}_{11} = \widehat{a}_{22} = 0$. Las curvas integrales $\xi(x,y) = C$ y $\eta(x,y) = C'$ se llaman **curvas características de (7.12)**. Como $\Delta > 0$, las ecuaciones de tipo hiperbólico tienen dos familias de curvas características reales y distintas, y el cambio de variable $\phi : (x,y) \to (\xi,\eta)$ reduce (7.13) a

$$v_{\xi\eta} + \ldots = 0 \tag{7.17}$$

ecuación que se denomina **la primera forma canónica** de la ecuación (7.12) hiperbólica.

Haciendo en la ecuación transformada (7.17) la sustitución

$$\lambda = \xi + \eta; \quad \mu = \xi - \eta \tag{7.18}$$

se obtiene la **forma normal** de la ecuación hiperbólica

$$w_{\lambda\lambda} - w_{\mu\mu} + \ldots = 0 \tag{7.19}$$

La ecuación de las ondas 1D es la forma normal de una ecuación hiperbólica en dos variables (x,t).

- Si la ecuación es parabólica, $\Delta = 0$ y sólo hay una familia de curvas características. Sea $\xi(x,y) = C$ su ecuación implícita. Haciendo el cambio de variable $(x,y) \to (\xi,\eta)$, donde η es una función arbitraria de x e y funcionalmente independiente de ξ, tanto $\widehat{a}_{11}$ como $\widehat{a}_{12}$ son idénticamente nulos y (7.13) se reduce a

$$v_{\eta\eta} + \ldots = 0 \tag{7.20}$$

que es la primera forma canónica y también la forma normal de la ecuación lineal parabólica.

- En el caso elíptico, $\Delta < 0$ y no existen curvas características reales. El objetivo al hacer un cambio de coordenadas es anular el coeficiente de la derivada cruzada. Elegimos ξ y η de modo que $\widehat{a}_{11} = \widehat{a}_{22} \neq 0$ y $\widehat{a}_{12} = 0$ se llega a la forma normal de la ecuación de tipo elíptico

$$v_{\xi\xi} + v_{\eta\eta} + \ldots = 0 \tag{7.21}$$

La ecuación de Laplace 2D es la escritura canónica de una ecuación elíptica de dos variables.

La clasificación de las ecuaciones en derivadas parciales lineales de segundo orden está íntimamente ligada al tipo de condiciones de contorno e iniciales que se deben especificar para obtener soluciones estables únicas.

En más de dos variables se obtiene a menudo información útil sobre el comportamiento de la ecuación diferencial en derivadas parciales considerando una de las variables constantes. Si el problema es de evolución 2D se suele hacer $t = t_0$ eliminando los términos con derivadas temporales y tratando la ecuación resultante como si fuera de dos variables.

7.1.3. Componentes de un problema matemático en derivadas parciales

Recordemos que sólo consideraremos en este libro problemas estacionarios 2D y problemas de evolución 1D y 2D.

- **El dominio**

  Si el problema es estacionario no hay ninguna variable singularizada, sólo hay variables espaciales y el dominio Ω es un abierto de $I\!R^2(x,y)$.

  Si el problema es de evolución, además de las coordenadas espaciales hay una variable singularizada que es el tiempo y el dominio es un cilindro $\Omega_T = \Omega \times [0,T]$ en el espacio-tiempo $I\!R^2(x,t)$ o $I\!R^3(x,y,t)$.

- **La ecuación diferencial en derivadas parciales**
  - Quasi-lineal de primer orden del tipo de transporte (7.4).
  - Del tipo

$$\mathcal{M}u = f \tag{7.22}$$

  con el operador $\mathcal{M}$ elíptico en Ω en el caso de los problemas estacionarios.

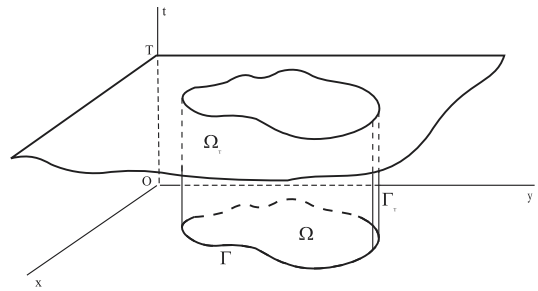


Figura 7.1: Dominio de un problema mixto para un fenómeno de propagación.

Fundamentalmente $\mathcal{M}$ será el operador de Laplace Δ y las ecuaciones serán las de Poisson 2D [8] (Laplace no homogénea)

$$\Delta u = u_{xx} + u_{yy} = f \tag{7.23}$$

y de Laplace si $f \equiv 0$.

◦ Del tipo

$$\mathcal{L}u = a\frac{\partial u}{\partial t} + b\frac{\partial^2 u}{\partial t^2} - \mathcal{M}u = f \tag{7.24}$$

en el caso de los problemas de evolución, donde $\mathcal{M}$ es un operador elíptico en Ω [9].

Ecuaciones parabólicas

El caso $a = 1$, $b = 0$ define las ecuaciones parabólicas del tipo del calor o de la difusión.

Si $\mathcal{M} = \alpha\Delta_x$ con $\alpha > 0$, se obtiene la ecuación de difusión o del calor.

$$\frac{\partial u}{\partial t} - \alpha\Delta_x u = f \tag{7.25}$$

Ecuaciones hiperbólicas

El caso $a = 0$, $b = 1$ define las ecuaciones hiperbólicas del tipo de las ondas.

Si $\mathcal{M} = c^2\Delta_x$ con $c > 0$, se obtiene la ecuación de las ondas.

$$\frac{\partial^2 u}{\partial t^2} - c^2\Delta_x u = f \tag{7.26}$$

A partir de ahora sólo consideraremos la ecuación del calor como ecuación parabólica y la de las ondas como ecuación hiperbólica.

- **Las condiciones de contorno**

  Las condiciones de contorno son tanto en los problemas estacionarios como en los de evolución de los tipos siguientes:

  ◦ **Condición de Dirichlet**(CD). Se preasigna el valor de u sobre la frontera de Ω ($u\Big|_{\partial\Omega} = \varphi$), donde φ es una función continua dada en $\partial\Omega$.

[8] Simeon Denis Poisson (1781-1840). Fue primero estudiante y después profesor de la École Polytechnique y más tarde en la Facultad de Ciencias. Publicó en 1835 un libro sobre la *Théorie mathématique de la chaleur*.

La productividad de Poisson se puede medir por la frecuencia con la que su nombre es citado en nuestros libros. La ecuación de Poisson es la consecuencia del descubrimiento por Poisson de que la ecuación de Laplace sólo se satisface en el exterior de las masas.

[9] Los operadores diferenciales están definidos respecto de las variables espaciales.

- **Condición de Neumann**(CN). Se preasigna sobre la frontera de Ω el valor de la derivada normal $\dfrac{du}{d\mathbf{n}}$ de u ($\left(\dfrac{du}{d\mathbf{n}}\right)\Big|_{\partial\Omega} = \psi$) donde ψ es una función dada de clase $\mathcal{C}^1$ en $\partial\Omega$.
- **Condiciones mezcladas**.

  Se imponen condiciones de contorno diferentes en partes diferentes de la frontera de Ω.

- **Las condiciones iniciales**

  Las condiciones iniciales se presentan en los problemas de evolución.

  - Del tipo

$$u\Big|_{t=0} = u_0 \tag{7.27}$$

  con u_0 dado en $\Omega \times \{0\}$ en el caso de las ecuaciones en derivadas parciales quasi-lineales de primer orden y de las ecuaciones parabólicas.

  - Del tipo

$$u\Big|_{t=0} = u_0, \quad \frac{\partial u}{\partial t}\Big|_{t=0} = u_1 \tag{7.28}$$

  con u_0 y u_1 dados en $\Omega \times \{0, \}$ en el caso de las ecuaciones hiperbólicas.

7.1.4. Problemas matemáticos

El problema matemático está "bien puesto" en el sentido de Hadamard, si se cumplen las tres condiciones

- La solución existe,
- es única
- y depende continuamente de los datos del problema (condiciones de contorno e iniciales).

La existencia no crea dificultades habitualmente. La causa habitual de no unicidad es la incompatibilidad de las condiciones suplementarias con el tipo de ecuación diferencial en derivadas parciales que gobierna el proceso. En general si el problema tiene "pocas" condiciones de contorno se pierde la unicidad y si se imponen "demasiadas", las soluciones pierden el sentido físico en la proximidad de la frontera. La tercera condición exige que pequeñas perturbaciones en las condiciones del problema provoquen cambios pequeños en la solución. Si no se cumple, la propagación interna de errores, sobre todo en los procesos de evolución hiperbólicos, es muy rápida.

Se definen entonces problemas con condiciones de contorno y/o iniciales tipo que aseguren que el problema está bien puesto.

Problemas de contorno elípticos

Los problemas de contorno elípticos clásicos consisten en hallar una solución de la ecuación elíptica (7.22) que además satisfaga en la frontera de Ω alguna de las condiciones de contorno que hemos definido.

- **Primer problema de contorno (Problema de Dirichlet)**.

Hallar una función u que sea solución de (7.22) y satisfaga la condición de Dirichlet (CD).

- **Segundo problema de contorno (Problema de Neumann)**.

Hallar una función u solución de (7.22) y que satisfaga en la frontera la condición de Neumann (CN).

Problemas hiperbólicos

a) **Problema de Cauchy o de valor inicial para la ecuación quasi-lineal de transporte**
Hallar una solución u de la ecuación quasi-lineal de transporte que satisfaga la condición inicial (7.27).

b) **Problemas hiperbólicos lineales de segundo orden**

Problema de Cauchy o de valor inicial para la ecuación de las ondas.
Hallar u solución de la ecuación de las ondas que además cumple para $t = 0$ las condiciones iniciales (7.28).

Problemas mixtos para la ecuación de las ondas.
Hallar u que verifica (7.26) en Ω_T, las condiciones iniciales (7.28) y sobre la frontera lateral Γ_T del cilindro Ω_T una condición de contorno bien de tipo Dirichlet o bien de tipo Neumann.

Problemas parabólicos.

Son los mismos que en el caso hiperbólico sustituyendo las condiciones iniciales hiperbólicas (7.28) por las parabólicas (7.27).

7.2. Método de diferencias finitas

Para obtener una aproximación numérica de la solución del problema matemático en estudio, construimos un problema discreto aproximando y discretizando cada uno de sus elementos.

• Si el problema es estacionario se discretiza el dominio Ω en el que varían las variables geométricas, dominio que supondremos plano y acotado, de frontera Γ.

El proceso de discretización de Ω consiste en definir un conjunto discreto de puntos (**dominio o malla computacional**) $\Omega_{\mathbf{h}}$ adaptado a $\overline{\Omega}$ dependiente de un parámetro $\mathbf{h}$ el **vector de pasos**, que define el tamaño de la malla y que está destinado a tender a cero.

Si el problema es de evolución se discretizan todas las variables incluida el tiempo.

Una vez definida la malla computacional, la solución numérica buscada $u_{\mathbf{h}}$ es la restricción de la solución u del problema en estudio a $\Omega_{\mathbf{h}}$.

• Se sustituyen las derivadas parciales de la ecuación que interviene en el problema

$$\mathcal{L}u = f \tag{7.29}$$

por cocientes incrementales sobre el dominio discreto $\Omega_{\mathbf{h}}$ (**aproximación por diferencias finitas**).

La herramienta matemática que se utiliza para formar las aproximaciones por diferencias finitas de las derivadas es el teorema de Taylor.

• Por último aproximamos las condiciones de contorno y las condiciones iniciales en los problemas de evolución.

Con todo ello se construye un **esquema en diferencias** que aproxima la solución numérica $u_{\mathbf{h}}$. Llamando $U_{\mathbf{h}}$ esa aproximación, el proceso anterior define un sistema de ecuaciones algébricas

$$\mathcal{L}_{\mathbf{h}} U_{\mathbf{h}} = f_{\mathbf{h}} \tag{7.30}$$

donde $\mathcal{L}_{\mathbf{h}}$ es un operador lineal entre espacios de dimensión finita que llamaremos **operador en diferencias** en el que **están incluidas las condiciones suplementarias**.

Las diferentes formas de seleccionar la malla computacional $\Omega_{\mathbf{h}}$, el operador en diferencias $\mathcal{L}_{\mathbf{h}}$ y el término independiente $f_{\mathbf{h}}$ distinguen a los distintos esquemas o métodos en diferencias para el problema en estudio.

7.2.1. Discretización del dominio espacial bidimensional Ω

Sea Ω el rectángulo $\{(x, y) : a < x < b,\ c < y < d\}$ donde los intervalos (a, b) y (c, d) pueden ser finitos o infinitos. Elegidos dos números reales positivos h y k (**los pasos**) de entre los divisores de $b - a$ y $d - c$ respectivamente, $h = \frac{b-a}{N}$ y $k = \frac{d-c}{M}$ con N y M enteros positivos, consideramos la cuadrícula homogénea del plano definida por la familias de rectas paralelas a los ejes de ecuaciones respectivas $y = c+jk$ y $x = a + ih$ con $i, j \in \mathbb{Z}$.

Las intersecciones de las dos familias de rectas constituyen una malla $\mathcal{M}_{(h,k)}$ de $\mathbb{R}^2$ subconjunto descrito por los **nodos** $M_{i,j}$ de coordenadas $(x_i, y_j) = (a + ih, c + jk)$ con i, j enteros relativos.

Es claro que los lados del rectángulo Ω están sobre rectas de la cuadrícula y que sobre el lado horizontal hay $N + 1$ nodos $a + ih$ con $i = 0, 1, ..., N$ y sobre el lado vertical hay $M + 1$ nodos $c + jk$ con $j = 0, 1, ..., M$.

Se define el dominio computacional $\Omega_{(h,k)}$ como la intersección $\mathcal{M}_{(h,k)} \cap \overline{\Omega}$ conjunto descrito por los nodos de la malla que están en Ω o sobre sus lados (ver la Figura 7.2)

$$\Omega_{(h,k)} = \{M_{i,j} = (a + ih, c + jk) : \quad 0 \leq i \leq N \text{ y } 0 \leq j \leq M\} \tag{7.31}$$

Si Ω no es rectangular, discretizamos un rectángulo $(a, b) \times (c, d)$ que contenga a Ω (Figura 7.3).

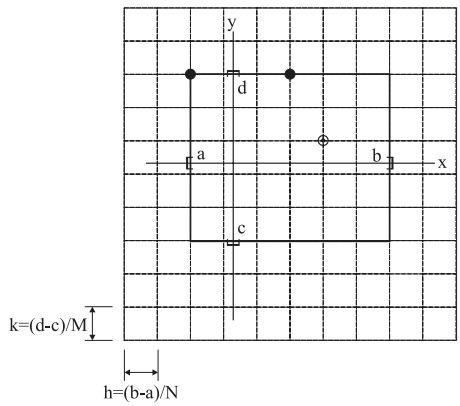


Figura 7.2: Discretización del rectángulo.

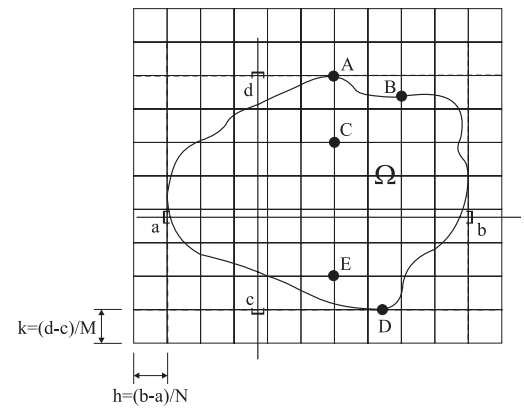


Figura 7.3: Caso de un dominio cualquiera.

Como antes, consideramos el conjunto de los puntos $M_{i,j}$ de la cuadrícula $\mathcal{M}_{(h,k)}$ que pertenecen a $\overline{\Omega}$, pero para captar mejor la información en la frontera añadimos el conjunto $\Gamma_{(h,k)}$ de los puntos de Γ que están sobre las rectas de la cuadrícula de modo que el dominio discretizado $\Omega_{(h,k)}$ [10] será ahora $\mathcal{M}_{(h,k)} \cap \overline{\Omega} \cup \Gamma_{(h,k)}$.

Refiriéndonos a la Figura 7.3, el punto A está en $\mathcal{M}_{(h,k)}$ y también en $\Gamma_{(h,k)}$. Los puntos B y D están en $\Gamma_{(h,k)}$ pero no en $\mathcal{M}_{(h,k)}$. Los puntos C y E que están en $\Omega_{(h,k)}$ pero no en Γ, son puntos interiores.

7.2.2. Aproximación de las derivadas parciales por diferencias finitas

Sean $f \in \mathcal{C}^n([a, b])$ y $\{a = x_0 < x_1 < ... < x_i < ... < x_N = b\}$ una partición homogénea del intervalo $[a, b]$ de paso $h = \frac{b-a}{N}$ de modo que $x_i = a + ih$ para $i = 0, 1, ..., N$.

Cuando $h \to 0$ podemos escribir las siguientes igualdades

$$\begin{aligned} f'(x_i) &= \frac{f(x_i) - f(x_{i-1})}{h} + O(h) \\ &= \frac{f(x_{i+1}) - f(x_i)}{h} + O(h) \\ &= \frac{f(x_{i+1}) - f(x_{i-1})}{2h} + O(h^2) \end{aligned} \tag{7.32}$$

de las que se obtienen las correspondientes aproximaciones de $f'(x_i)$ con paso h, que se llaman respectivamente diferencias de dos puntos izquierda o regresiva y derecha o progresiva y diferencia de tres puntos centrada (ver en el Capítulo 5 las fórmulas (5.6), (5.7) y (5.10)).

Todas ellas son consecuencia del teorema de Taylor que asegura que [11]

$$f(x + h) = f(x) + f'(x)h + f''(x)\frac{h^2}{2!} + ... + f^{(n-1)}(x)\frac{h^{n-1}}{(n-1)!} + O(h^n) \tag{7.34}$$

[10] En el caso de un dominio rectangular ambos conjuntos coinciden.

[11] Utilizando el resto de Lagrange se puede precisar mejor el significado del símbolo $O(h^n)$ en (7.34)

$$O(h^n) = f^{(n)}(\xi)\frac{h^n}{n!} \tag{7.33}$$

con $\xi \in (x, x + h)$.

Ejemplo 7.2.1 *Si suponemos que $n = 2$, la diferencia centrada es sólo $O(h)$. En efecto, se tiene aquí*

$$f(x+h) = f(x) + f'(x)h + O(h^2)$$
$$f(x-h) = f(x) - f'(x)h + O(h^2)$$

y restando

$$f'(x) = \frac{f(x+h) - f(x-h)}{2h} + \frac{O(h^2)}{h} = \frac{f(x+h) - f(x-h)}{2h} + O(h)$$

Si u es una función de dos variables (x, y) suficientemente regular en $\mathbb{R}^2$, para aproximar sus derivadas parciales primeras, basta utilizar la fórmula de Taylor de una variable.

Utilizando la notación $u(M_{i,j}) = u(x_i, y_j) = u(a + ih, c + jk) = u_{i,j}$; $u_x(x_i, y_j) = u_x\big|_{i,j}$, $u_{xy}(x_i, y_j) = u_{xy}\big|_{i,j}$, etc., y fijando y_j podemos escribir

$$\begin{aligned} u_x\big|_{i,j} &= \frac{u_{i+1,j} - u_{i,j}}{h} + O(h) \\ &= \frac{u_{i+1,j} - u_{i-1,j}}{2h} + O(h^2) \\ &= \frac{u_{i,j} - u_{i-1,j}}{h} + O(h) \end{aligned} \tag{7.35}$$

aproximaciones de u_x por diferencias finitas progresiva, centrada y regresiva respectivamente.

Ejemplo 7.2.2 *Para obtener una aproximación de u_{xx} en el punto (x_i, y_j), supuesto que u es suficientemente diferenciable, desarrollamos $u(x_i \pm h, y_j)$ en el entorno del punto (x_i, y_j)*

$$u_{i\pm1,j} = u_{i,j} \pm u_x\big|_{i,j} h + u_{xx}\big|_{i,j} \frac{h^2}{2!} \pm u_{xxx}\big|_{i,j} \frac{h^3}{3!} + O(h^4)$$

sumando ambos desarrollos

$$u_{i+1,j} + u_{i-1,j} = 2u_{i,j} + u_{xx}\big|_{i,j} h^2 + O(h^4)$$

de donde

$$u_{xx}\big|_{i,j} = \frac{u_{i+1,j} - 2u_{i,j} + u_{i-1,j}}{h^2} + O(h^2) \tag{7.36}$$

cuando $h \to 0$(ver la sección 5.7 del Capítulo 5).

Se obtiene del mismo modo $u_{yy}\big|_{i,j}$

$$u_{yy}\big|_{i,j} = \frac{u_{i,j+1} - 2u_{i,j} + u_{i,j-1}}{k^2} + O(k^2) \tag{7.37}$$

En algunos casos es más práctica la aproximación

$$u_{xx}\big|_{i,j} = \theta \frac{u_{i+1,j+1} - 2u_{i,j+1} + u_{i-1,j+1}}{h^2} + (1-\theta) \frac{u_{i+1,j} - 2u_{i,j} + u_{i-1,j}}{h^2} + O(h^2) \tag{7.38}$$

donde θ es un parámetro fijo $(0 \le \theta \le 1)$ que representa una combinación lineal convexa de dos expresiones finitas obtenidas de (7.36) para los valores y_j e y_{j+1}(ver la sección 7.2.6).

Para aproximar la derivada segunda cruzada $u_{xy}\big|_{i,j}$ es necesario utilizar la fórmula de Taylor de u para dos variables en el entorno del punto (x_i, y_j).

Suponiendo que u es de clase $\mathcal{C}^5$ se obtiene cuando $(h, k) \to (0, 0)$ [3]

$$u_{xy}\big|_{i,j} = \frac{u_{i+1,j+1} - u_{i-1,j+1} - u_{i+1,j-1} + u_{i-1,j-1}}{4hk} + O((h+k)^2) \tag{7.39}$$

Ejemplo 7.2.3 *Utilizando (7.36) y (7.38) con h y k infinitésimos del mismo orden obtenemos la siguiente aproximación de la laplaciana de u*

$$\Delta u\big|_{i,j} = u_{xx}\big|_{i,j} + u_{yy}\big|_{i,j} = \frac{u_{i+1,j} - 2u_{i,j} + u_{i-1,j}}{h^2} + \frac{u_{i,j+1} - 2u_{i,j} + u_{i,j-1}}{k^2} + O(h^2 + k^2) \tag{7.40}$$

en la que intervienen cinco puntos, $M_{i,j}$ y los cuatro puntos de la malla $M_{i+1,j}$, $M_{i-1,j}$, $M_{i,j-1}$, $M_{i,j+1}$ que le rodean.

Tomando $h = k$ se tiene

$$\Delta u\big|_{i,j} = \frac{u_{i+1,j} + u_{i-1,j} + u_{i,j+1} + u_{i,j-1} - 4u_{i,j}}{h^2} + O(h^2) \tag{7.41}$$

Todas las fórmulas que hemos obtenido expresan igualdades entre el **valor exacto** de la derivada de una función en un punto y la suma de un **valor aproximado** de dicha derivada por diferencias finitas y del **error de truncación** que se "esconde" tras el símbolo de Landau $O(.)$.

En adelante distinguiremos claramente entre ambos valores denotando el valor exacto como la función, por ejemplo en (7.36) $u_{xx}\big|_{i,j}$ y la aproximación por diferencias con mayúscula, es decir, $U_{xx}\big|_{i,j}$ notación que transportaremos como veremos a la solución exacta y la solución aproximada de una cierta ecuación diferencial en derivadas parciales.

Ejemplo 7.2.4

$$U_{xx}\big|_{i,j} = \frac{U_{i+1,j} - 2U_{i,j} + U_{i-1,j}}{h^2} \tag{7.42}$$

es una aproximación de segundo orden de u_{xx} mediante diferencias centradas en el punto (x_i, y_j) de la malla. La información relativa al orden de la aproximación viene dada por el error de truncación que es proporcional a h^2.

Análogamente, de (7.41)

$$\Delta U\big|_{i,j} = \frac{U_{i+1,j} + U_{i-1,j} + U_{i,j+1} + U_{i,j-1} - 4U_{i,j}}{h^2} \tag{7.43}$$

es una aproximación de segundo orden de $\Delta u\big|_{i,j}$.

En los problemas de evolución se discretiza también el intervalo de tiempo $(0, T)$. Supuesto que el dominio espacial es (a, b), se considera el dominio $(a, b) \times (0, T)$.

Definiendo $h = \dfrac{b-a}{N}$ y $k = \dfrac{T}{M}$, construimos la malla

$$\Omega_{(h,k)} = \{M_{i,j} = (a + ih, jk) : \quad 0 \le i \le N \ \text{ y } \ 0 \le j \le M\} \tag{7.44}$$

En este caso es costumbre escribir el valor de u en el punto (x_i, t_j) con el índice relativo a la variable temporal como superíndice, es decir, $u_i^j = u(a + ih, jk)$.

Ejemplo 7.2.5 *Utilizando una diferencia de primer orden progresiva para u_t y (7.36) para la derivada espacial con el paso temporal k y el paso espacial h infinitésimos del mismo orden, obtenemos la siguiente aproximación de la ecuación de difusión (7.5)*

$$u_t\big|_i^j - u_{xx}\big|_i^j = \frac{u_i^{j+1} - u_i^j}{k} + O(k) - \frac{u_{i+1}^j - 2u_i^j + u_{i-1}^j}{h^2} + O(h^2) = 0 \tag{7.45}$$

Operando

$$u_i^{j+1} - u_i^j = \frac{k}{h^2}\left(u_{i+1}^j - 2u_i^j + u_{i-1}^j\right) + kO(h^2) - O(k^2) \tag{7.46}$$

Llamando $r = \dfrac{k}{h^2}$ y suprimiendo el término relativo al error de discretización $O(kh^2 + k^2)$, se obtiene la fómula

$$U_i^{j+1} = rU_{i+1}^j + (1 - 2r)U_i^j + rU_{i-1}^j \tag{7.47}$$

*Se trata del primer ejemplo de un esquema de **dos niveles explícito**.*

Llamando U^j el vector de aproximaciones U_i^j en el instante t_j, vemos en (7.47) que para calcular U^{j+1} en un punto x_i basta sustituir los valores conocidos relativos al nivel temporal anterior j (y no de niveles anteriores) que figuran en el segundo miembro de (7.47).

Definición 7.2.1 *Un esquema en diferencias es **explícito** cuando define un proceso de avance mediante el que es posible conocer U^{j+1} en función de valores conocidos del nivel temporal anterior U^j si el método es de dos niveles (y de los anteriores si el método es de tres o más niveles) y de los valores en el contorno.*

Ejemplo 7.2.6 *Tomando diferencias de segundo orden para aproximar u_{xx} y u_{tt} con base en el punto (x_i, t_j), obtenemos la siguiente aproximación de la ecuación de las ondas (7.48)*

$$u_{tt}\big|_i^j - u_{xx}\big|_i^j = \frac{u_i^{j+1} - 2u_i^j + u_i^{j-1}}{k^2} + O(k^2) - \frac{u_{i+1}^j - 2u_i^j + u_{i-1}^j}{h^2} + O(h^2) = 0 \tag{7.48}$$

Operando se obtiene

$$u_i^{j+1} + u_i^{j-1} = 2u_i^j + \frac{k^2}{h^2}\left(u_{i+1}^j - 2u_i^j + u_{i-1}^j\right) + k^2 O(h^2) + O(k^4) \tag{7.49}$$

Llamando $r = \dfrac{k^2}{h^2}$ *y suprimiendo el término relativo al error de discretización se obtiene el esquema explícito de* ***tres niveles***

$$U_i^{j+1} = r\left(U_{i-1}^j + U_{i+1}^j\right) + 2(1-r)U_i^j - U_i^{j-1} \tag{7.50}$$

Ejemplo 7.2.7 *Tomando en la ecuación de difusión (7.5) el punto* (x_i, t_{j+1}) *como base y utilizando diferencias de primer orden regresivas para aproximar* u_t *y diferencias de segundo orden para* u_{xx} *obtenemos*

$$u_t\big|_i^{j+1} - u_{xx}\big|_i^{j+1} = \frac{u_i^{j+1} - u_i^j}{k} + O(k) - \frac{u_{i+1}^{j+1} - 2u_i^{j+1} + u_{i-1}^{j+1}}{h^2} + O(h^2) = 0 \tag{7.51}$$

Operando,

$$u_i^{j+1} = u_i^j + \frac{k}{h^2}\left(u_{i+1}^{j+1} - 2u_i^{j+1} + u_{i-1}^{j+1}\right) + kO(h^2) - O(k^2) \tag{7.52}$$

Llamando $r = \dfrac{k}{h^2}$ *y suprimiendo el término relativo al error de discretización se obtiene el esquema totalmente* ***implícito***

$$U_i^{j+1} = U_i^j + r\left(U_{i+1}^{j+1} - 2U_i^{j+1} + U_{i-1}^{j+1}\right) \tag{7.53}$$

Para calcular U^{j+1} en un punto $x_i = a + ih$ es necesario resolver un sistema lineal de ecuaciones, ya que las incógnitas U_i^{j+1} aparecen en ambos miembros de (7.53) y no se pueden despejar (ver el problema 7.5).

Definición 7.2.2 *Un esquema en diferencias es* ***implícito*** *si dos o mas valores del vector* U^{j+1} *se conocen en función de valores conocidos del nivel anterior* U^j *(y de valores anteriores si el método es de tres o más niveles) y de los valores en el contorno. Si hay* M *valores desconocidos en el nivel* $j+1$ *se debe aplicar la fórmula en diferencias* M *veces. El sistema de* M *ecuaciones resultante suministra los* M *valores de modo implícito.*

7.2.3. Discretización de las condiciones de contorno

Consideramos el dominio rectangular $\Omega = [0, LX] \times [0, LY]$. Escogemos los pasos $h = \frac{LX}{N}$ y $k = \frac{LY}{M}$, y construimos la malla

$$\Omega_{(h,k)} = \{M_{i,j} = (ih, jk) : \quad 0 \le i \le N \ \text{ y } \ 0 \le j \le M\} \tag{7.54}$$

Condición de Dirichlet

Si u verifica la condición de Dirichlet (CD), tendremos

$$u_{i,j} = \varphi_{i,j}\ ; \quad i = 0, N\ \forall j; \quad j = 0, M\ \forall i \tag{7.55}$$

En el problema discreto asociado (7.30) esta condición se escribe

$$U_{i,j} = \varphi_{i,j}\ ; \quad i = 0, N\ \forall j; \quad j = 0, M\ \forall i \tag{7.56}$$

Condición de Neumann

La discretización de este tipo de condiciones es menos evidente.

◦ **Condición de Neumann sobre un lado del dominio que está sobre el eje** y (Figura (7.4)). Ya que la normal exterior es $\mathbf{n} = -(1, 0)$, $\frac{du}{d\mathbf{n}} = -\frac{\partial u}{\partial x}$ y la condición de Neumann (CN) se escribe

$$\frac{\partial u}{\partial x}(0, y) = -\psi(y) \tag{7.57}$$

La igualdad

$$u_{1,j} = u_{0,j} + u_x\big|_{0,j} h + u_{xx}\big|_{0,j} \frac{h^2}{2!} + O(h^3)$$

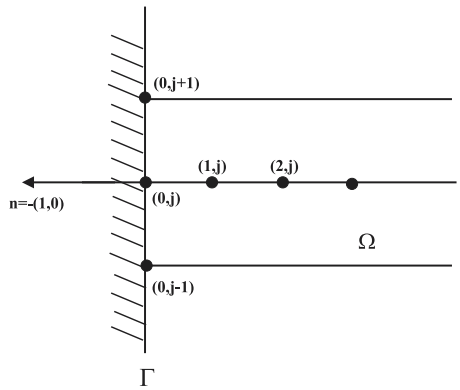


Figura 7.4: Condición de Neumann en el lado vertical izquierdo.

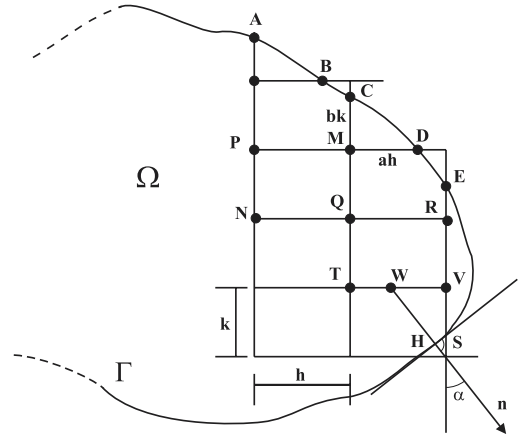


Figura 7.5: Condiciones de contorno en un dominio no rectangular.

nos permite aproximar el valor de u en el punto $0, j$

$$u_{0,j} = u_{1,j} - h\psi_j + O(h^2) \quad \Rightarrow \quad U_{0,j} = U_{1,j} - h\psi_j$$

y escribir una aproximación de primer orden de u_{xx} en dicho punto

$$U_{xx}\big|_{0,j} = \frac{2}{h^2}\left(U_{1,j} - U_{0,j} + h\psi_j\right) \tag{7.58}$$

donde $\psi_j = \psi(jk)$ [12].

◦ **La frontera del dominio posee puntos que no son nodos de la malla $\mathcal{M}_{(h,k)}$.**

Consideremos ahora brevemente la discretización de la condición de contorno de tipo Dirichlet cuando el dominio Ω no es rectangular y algunos puntos de $\Gamma_{(h,k)}$ no son nodos de la malla $\mathcal{M}_{(h,k)}$.

Todas las notaciones se refieren a la Figura 7.5.

Los puntos A, B, C, D y E están en $\Gamma_{(h,k)}$ pero no están en la red.

En el caso del punto M, llamando $MD = ah$ podemos poner, utilizando un desarrollo de Taylor de u con base en M,

$$u_D = u_M + ahu_x\big|_M + \frac{(ah)^2}{2!}u_{xx}\big|_M + O(h^3) \tag{7.59}$$

Del otro lado,

$$u_P = u_M - hu_x\big|_M + \frac{h^2}{2!}u_{xx}\big|_M + O(h^3) \tag{7.60}$$

de modo que eliminando $u_x\big|_M$ entre las dos se llega a

$$u_{xx}\big|_M = \frac{2}{h^2}\frac{u_D - (1+a)u_M + au_P}{a(1+a)} + O(h) \tag{7.61}$$

y ya que conocemos los valores de u en los puntos de $\Gamma_{(h,k)}$ obtenemos una aproximación de primer orden de u_{xx} en el punto M

$$U_{xx}\big|_M = \frac{2}{h^2}\frac{\varphi_D - (1+a)u_M + au_P}{a(1+a)} \tag{7.62}$$

De un modo similar pero razonando con los tres puntos Q, M y C y poniendo $MC = bk$ podemos escribir [13]

$$U_{yy}\big|_M = \frac{2}{k^2}\frac{\varphi_C - (1+b)u_M + bu_Q}{b(1+b)} \tag{7.63}$$

[12] Ver en el problema 7.13 otra técnica de tratamiento de la condición de Neumann en un lado del dominio rectangular añadiendo al dominio computacional una fila o columna de nodos ficticios.

[13] Ver el problema 7.11.

Ejemplo 7.2.8 *Sumando las aproximaciones (7.62) y (7.63) obtenemos una aproximación de primer orden de Δu en M*

$$\Delta U\big|_M = \frac{2}{h^2}\left(\frac{\varphi_D}{a(1+a)} - \frac{U_M}{a} + \frac{U_P}{1+a}\right) + \frac{2}{k^2}\left(\frac{\varphi_C}{b(1+b)} - \frac{U_M}{b} + \frac{U_Q}{1+b}\right) \tag{7.64}$$

De nuevo una fórmula de cinco puntos que requiere la información del valor de la solución en M y en los puntos P, Q, C y D de $\Omega_{(h,k)}$ vecinos de M.

7.2.4. Convergencia, estabilidad y consistencia

Convergencia

En el proceso de aproximación por diferencias finitas, hemos sustituido el problema diferencial (P) definido por la ecuación $\mathcal{L}u = f$ más condiciones suplementarias por una serie de problemas discretos (P_h) definidos por la ecuación $\mathcal{L}_h U_h = f_h$.

¿En qué sentido la solución U_h de estos problemas aproxima a la solución u de (P)?

Definición 7.2.3 *Definimos el **error global** como la diferencia entre la solución numérica calculada U_h y la verdadera u_h*

$$E_h = U_h - u_h \tag{7.65}$$

Definición 7.2.4 *Diremos que el esquema en diferencias (7.30) es **convergente** en una cierta norma $\|\cdot\|$ si*

$$\|E_h\| = \|U_h - u_h\| \xrightarrow[h\to 0]{} 0 \tag{7.66}$$

*y que es **convergente de orden** p si*

$$\|E_h\| = O(h^p) \tag{7.67}$$

Consistencia

En una discretización (P_h) de (P), el resultado óptimo sería que u_h fuese la solución de las ecuaciones discretas de modo exacto.

Definición 7.2.5 *Definimos el **error de truncación local** sustituyendo en el esquema en diferencias (7.30) la solución aproximada U_h por la solución exacta u_h*

$$L_h = \mathcal{L}_h u_h - f_h \tag{7.68}$$

El error de truncación local es un vector cuya norma nos indica en qué medida el esquema en diferencias (P_h) que estamos usando es un buen modelo local de nuestro problema (P). Esperamos que si la solución exacta u_h satisface bien el esquema en diferencias ello indique que, recíprocamente, la solución exacta del esquema en diferencias satisface bien la ecuación diferencial en derivadas parciales.

Definición 7.2.6 *El esquema en diferencias (7.30) es **consistente** si*

$$\|L_h\| = \xrightarrow[h\to 0]{} 0 \tag{7.69}$$

*y es **consistente de orden p** si*

$$\|L_h\| = O(h^p) \tag{7.70}$$

Ejemplo 7.2.9 *Estudiemos la consistencia del esquema de Dufort-Frankel para la ecuación de difusión $\mathcal{L}u = u_t - u_{xx} = 0$*

$$\mathcal{L}_h U_h \equiv \left\{\frac{U_i^{j+1} - U_i^{j-1}}{2k} - \frac{U_{i-1}^j - U_i^{j-1} - U_i^{j+1} + U_{i+1}^j}{h^2}\right\} = 0 \tag{7.71}$$

De

$$u_i^{j\pm 1} = u_i^j \pm k u_t\big|_i^j + \frac{k^2}{2} u_{tt}\big|_i^j \pm \frac{k^3}{6} u_{ttt}\big|_i^j + O(k^4)$$

$$u_{i\pm 1}^j = u_i^j \pm h u_x\big|_i^j + \frac{h^2}{2} u_{xx}\big|_i^j \pm \frac{h^3}{6} u_{xxx}\big|_i^j + O(h^4)$$

se obtiene

$$\frac{u_i^{j+1}-u_i^{j-1}}{2k}=\frac{1}{2k}\left(2ku_t\big|_i^j+\frac{k^3}{6}u_{ttt}\big|_i^j+2\cdot O(k^4)\right)=u_t\big|_i^j+O(k^2)$$

$$\frac{u_{i-1}^j-u_i^{j-1}-u_i^{j+1}+u_{i+1}^j}{h^2}=\frac{1}{h^2}\left(h^2u_{xx}\big|_i^j-k^2u_{tt}\big|_i^j+O(h^4)+O(k^4)\right)=$$

$$=u_{xx}\big|_i^j-\frac{k^2}{h^2}u_{tt}\big|_i^j+O(h^2)+O\left(\frac{k^4}{h^2}\right)$$

de donde

$$L_h\big|_i^j=(\mathcal{L}_hu_h)\big|_i^j=u_t\big|_i^j-u_{xx}\big|_i^j+O(k^2)-\frac{k^4}{h^2}u_{tt}\big|_i^j+O(h^2)+O\left(\frac{k^4}{h^2}\right)=$$

$$=O(k^2)+O(h^2)+O\left(\frac{k^4}{h^2}\right)-\frac{k^2}{h^2}u_{tt}\big|_i^j$$

y como hemos asumido que u es la solución exacta de la ecuación diferencial, $u_t\big|_i^j-u_{xx}\big|_i^j=0$.

- *Si $\dfrac{k}{h^2}$ es constante cuando $(h,k)\to(0,0)$ entonces tanto $\dfrac{k^2}{h^2}$ como $\dfrac{k^4}{h^2}$ tienden a cero.*

El esquema en estudio es consistente con la ecuación de difusión.

- *Si $d\frac{k}{h}=c$ es constante cuando $(h,k)\to(0,0)$ entonces*

$$L_h\big|_i^j=O(k^2)+O(h^2)+O\left(\frac{k^4}{h^2}\right)-c^2u_{tt}\big|_i^j$$

El esquema en estudio no es consistente con la ecuación de difusión, pero sí que es consistente con la ecuación $u_t-u_{xx}-c^2u_{tt}=0$, ya que

$$L_h\big|_i^j=\left(u_t-u_{xx}-c^2u_{tt}\right)\big|_i^j+O(k^2)+O(h^2)+O\left(\frac{k^4}{h^2}\right) \tag{7.72}$$

Se produce en este ejemplo un fenómeno sin igual en el caso de ecuaciones diferenciales ordinarias. Sucesivos refinamientos de k generan una solución aproximada que es estable pero que no converge a la ecuación diferencial deseada.

Sólo si k converge a cero más rápido que h el esquema de Dufort-Frankel es consistente con la ecuación de difusión.

7.2.5. Estabilidad

Para que se produzca la convergencia del esquema en diferencias que aproxima una ecuación diferencial en derivadas parciales no basta con que sea consistente, debe también ser estable a pequeñas alteraciones en la estructura y los datos del problema.

Se estudia tan sólo la estabilidad de los esquemas en diferencias que aproximan una ecuación diferencial en derivadas parciales lineal de solución acotada, en cuyo caso la ecuación en diferencias **debe producir una solución numérica también acotada**.

Definición 7.2.7 *Una ecuación en diferencias es **estable** si produce una solución acotada y es **inestable** si produce una solución no acotada.*

*Si la solución de la ecuación en diferencias es acotada cualesquiera que sean los tamaños de los pasos de la malla computacional, se dice que la ecuación en diferencias es **incondicionalmente estable**.*

*Si la solución de la ecuación en diferencias es acotada sólo para ciertos valores de los pasos de la malla computacional se dice que la ecuación en diferencias es **condicionalmente estable**.*

*Si la solución de la ecuación en diferencias no está acotada cualesquiera que sean los tamaños de los pasos de la malla computacional se dice que la ecuación en diferencias es **incondicionalmente inestable**.*

Existen varios métodos para analizar la estabilidad de la ecuación en diferencias que aproxima una ecuación diferencial en derivadas parciales de los que destacamos por su simplicidad el que Von Neumann desarrolló durante la segunda guerra mundial.

Se trata de una condición necesaria de estabilidad respecto de las condiciones iniciales. Se obtiene la solución exacta de la ecuación en diferencias para una componente general de la representación en serie

compleja de Fourier de una **distribución inicial arbitraria**. Si esa solución es acotada (condicional o incondicionalmente) entonces la ecuación en diferencias es estable. Si por el contrario la solución para la componente genérica de Fourier no está acotada, la ecuación en diferencias es inestable.

En las referencias [17], [28] se incluyen tratamientos adecuados del tema de estabilidad.

7.2.6. Esquemas en diferencias de ciertos problemas tipo

Problemas elípticos

Consideramos el problema de Dirichlet para la ecuación de Poisson en el cuadrado unidad $\Omega = (0,1)^2$ y llamemos Γ a su frontera

$$(P) \quad \begin{cases} u_{xx} + u_{yy} = f & (x,y) \in \Omega \\ u(x,y) = \varphi(x,y) & (x,y) \in \Gamma \end{cases} \tag{7.73}$$

Fijamos $h = \dfrac{1}{N}$ y construimos la malla Ω_h asociada

$$\Omega_h = \{M_{i,j} = (ih, jk) : \quad 0 \le i \le N \ \text{ y } \ 0 \le j \le N\}$$

Utilizaremos la aproximación (7.41) de la laplaciana $(u_{xx}+u_{yy})\big|_{i,j}$ se obtiene así el esquema en diferencias

$$\mathcal{L}_h U_h \equiv \left\{ \begin{array}{l} \dfrac{U_{i+1,j} + U_{i-1,j} + U_{i,j+1} + U_{i,j-1} - 4U_{i,j}}{h^2} \\ \\ U_{i,j} \end{array} \right\} = f_h \equiv \left\{ \begin{array}{ll} f(ih,jh) & (ih,jh) \in \Omega \\ \varphi(ih,jh) & (ih,jh) \in \Gamma \end{array} \right\} \tag{7.74}$$

De

$$u_{i\pm1,j} = u_{i,j} \pm h u_x\big|_{i,j} + \frac{h^2}{2} u_{xx}\big|_{i,j} \pm \frac{h^3}{6} u_{xxx}\big|_{i,j} + O(h^4)$$

se obtiene

$$L_h\big|_{i,j} = (\mathcal{L}_h u_h - f_h)\big|_{i,j} = (u_{xx} + u_{yy} - f_h)\,\big|_{i,j} + \frac{h^2}{12}\,(u_{xxxx} + u_{yyyy})\,\big|_{i,j} + O(h^2)$$

y ya que u es la solución exacta de la ecuación diferencial, $(u_{xx} + u_{yy} - f)\,\big|_{i,j} = 0$, de donde

$$L_h\big|_{i,j} = \frac{h^2}{12}\,(u_{xxxx} + u_{yyyy})\,\big|_{i,j} + O(h^2) \tag{7.75}$$

y el esquema es consistente de segundo orden.

Se demuestra que estos métodos son estables.

Problemas de evolución

- **Problemas parabólicos**

  Consideremos el primer problema mixto para la ecuación del calor (P)

$$\begin{cases} u_t = u_{xx} + f & (x,t) \in [0,L] \times [0,T] \\ u(0,t) = \varphi(t) \quad , \quad u(L,t) = \psi(t) & t \in [0,T] \\ u(x,0) = u_0(x) & x \in [0,L] \end{cases} \tag{7.76}$$

con las condiciones necesarias de compatibilidad entre las condiciones de contorno y la inicial. Construimos la malla $\Omega_{h,k}$ asociada a los pasos $h = \frac{L}{N}$ y $k = \frac{T}{M}$.

Una familia de esquemas en diferencias clásicos en el tratamiento del problema (P) viene dada por $(P^{\theta}_{h,k})$

$$\mathcal{L}^{\theta}_h U_h \equiv \left\{ \begin{array}{l} \dfrac{U_i^{j+1}-U_i^j}{k} - \theta \left[\dfrac{U_{i+1}^{j+1}-2U_i^{j+1}+U_{i-1}^{j+1}}{h^2} \right] + \\ (1-\theta) \left[\dfrac{U_{i+1}^{j}-2U_i^{j}+U_{i-1}^{j}}{h^2} \right] \\ U_0^j \\ U_N^j \\ U_i^0 \end{array} \right\} = \tag{7.77}$$

$$= f_h \equiv \left\{ \begin{array}{ll} \theta f_i^{j+1} + (1-\theta) f_i^j & \\ \varphi(jk) & j = 0, ..., M \\ \psi(jk) & j = 0, ..., M \\ u_0(ih) & i = 0, ..., N \end{array} \right\} \quad \text{con} \quad 0 \leq \theta \leq 1$$

Para $\theta = 0$ el esquema es el clásico de dos niveles explícito (7.47) que obtuvimos en el ejemplo 7.2.5.

Para $\theta \neq 0$ el esquema resultante es implícito. Si $\theta = 1$ el esquema es el implícito (7.47) que obtuvimos en el ejemplo 7.2.7.

Para $\theta = \frac{1}{2}$ obtenemos el esquema que Crank y Nicolson propusieron en 1947 en [6] y que lleva su nombre.

Para los esquemas de la familia (7.77) el error de truncación local depende de θ y para los puntos de la malla en los que no es cero, viene dado por

$$L^{\theta}_{h,k} = (1-2\theta)\frac{k}{2} u_{tt} + O(h^2 + k^2) \tag{7.78}$$

Para θ arbitrario el error de truncación es $O(h^2 + k)$.

En el esquema de Crank-Nicolson se obtiene $O(h^2 + k^2)$.

El esquema explícito de la familia relativo al valor $\theta = 0$ es estable si

$$0 \leq r \leq \frac{1}{2} \tag{7.79}$$

luego es condicionalmente estable.

El esquema implícito relativo al valor $\theta = 1$ es incondicionalmente estable.

- **Problemas hiperbólicos**

  - **Ecuación quasi-lineal de primer orden.**

    Consideremos el problema de Cauchy puro para la ecuación hiperbólica lineal de transporte 1D adimensional

$$(P) \quad \left\{ \begin{array}{ll} u_t = u_x + f & (x,t) \in I\!R \times [0,T] \\ u(x,0) = u_0(x) & x \in I\!R \end{array} \right. \tag{7.80}$$

    Sean $h > 0$ y $k = \dfrac{T}{M}$ los pasos espacial y temporal respectivamente y sea Ω_h la malla formada por la intersección de las rectas $x = ih$ con $i \in Z\!\!\!Z$ y $t = jk$ con $j = 0, ..., M$. Supongamos que los pasos están relacionados por la igualdad $k = ch$ con c una constante positiva.

    El esquema en diferencias más simple es

$$\mathcal{L}_h U_h \equiv \left\{ \begin{array}{l} \dfrac{U_i^{j+1}-U_i^j}{k} - \dfrac{U_{i+1}^j - U_i^j}{h} \\ U_i^0 \end{array} \right\} = f_h \equiv \left\{ \begin{array}{l} f_i^j \\ u_0(ih) \end{array} \right\} \tag{7.81}$$

De

$$u_i^{j+1} = u_i^j + ku_t\big|_i^j + \frac{k^2}{2}u_{tt}\big|_i^j + \frac{k^3}{6}u_{ttt}\big|_i^j + O(k^4)$$
$$u_{i+1}^j = u_i^j + hu_x\big|_i^j + \frac{h^2}{2}u_{xx}\big|_i^j + \frac{h^3}{6}u_{xxx}\big|_i^j + O(h^4)$$

se obtiene

$$\frac{u_i^{j+1} - u_i^j}{k} = u_t\big|_i^j + \frac{k}{2}u_{tt}\big|_i^j + O(k)$$
$$\frac{u_{i+1}^j - u_i^j}{h^2} = u_x\big|_i^j + \frac{h}{2}u_{xx}\big|_i^j + O(h)$$

luego

$$L_h\big|_i^j = (\mathcal{L}_h u_h - f_h)\big|_i^j = \begin{cases} (u_t - u_x - f_h)\,\big|_i^j + \left(\frac{k}{2}u_{tt} - \frac{h}{2}u_{xx}\right)\big|_i^j + O(h) + O(k) \\ 0 \end{cases}$$

y ya que u es la solución exacta de la ecuación diferencial, $(u_t - u_x - f_h)\,\big|_i^j = 0$, de donde

$$L_h\big|_i^j = \begin{cases} \left(\frac{k}{2}u_{tt} - \frac{h}{2}u_{xx}\right)\big|_i^j + O(k+h) \\ 0 \end{cases} \tag{7.82}$$

y el esquema es consistente de primer orden

Este esquema es estable para $r \leq 1$.

- **Ecuación lineal de segundo orden.**

Consideremos el problema de Cauchy para la ecuación de las ondas

$$\begin{cases} u_{tt} - u_{xx} = 0 \qquad (x,t) \in I\!R \times I\!R_+ \\ (CI) \begin{cases} u(x,0) = u_0(x) \\ u_t(x,0) = u_1(x) \end{cases} \qquad x \in I\!R \end{cases} \tag{7.83}$$

Utilizamos el esquema de tres niveles que definimos en el ejemplo 7.2.6 y aproximamos la segunda condición inicial desarrollando $u(ih,k)$ en el entorno de $(jh,0)$

$$u_j^1 = u_j^0 - ku_t\big|_j^0 + \frac{k^2}{2}u_{tt}\big|_j^0 + O(k^2)$$

Sustituyendo u_{tt} por u_{xx} a través de la ecuación diferencial y teniendo en cuenta las condiciones iniciales

$$u_{tt}\big|_j^0 = u_{xx}\big|_j^0 = (u_0)_{xx}\big|_j \quad \text{y} \quad u_t\big|_j^0 = (u_1)_j$$

de modo que

$$U_j^1 = (u_0)_j - k(u_1)_j + \frac{k^2}{2}(u_0)_{xx}\big|_j$$

una aproximación de segundo orden gracias al término de la derivada segunda, cuya omisión condicionaría el orden del esquema.

Con todo ello escribimos el esquema en diferencias

$$\mathcal{L}_h U_h \equiv \left\{ \begin{array}{l} \dfrac{U_{j+1}^n - 2U_j^n + U_{j-1}^n}{h^2} + \dfrac{U_j^{n+1} - 2U_j^n + U_j^{n-1}}{k^2} \\ U_j^0 \\ U_j^1 \end{array} \right\} = \tag{7.84}$$

$$= f_h \equiv \left\{ \begin{array}{l} 0 \\ u_0(jh) = (u_0)_j \\ (u_0)_j + k(u_1)_j + (u_0)_{xx}\big|_j \dfrac{k^2}{2} \end{array} \right\} \qquad j \in \mathbb{Z} \quad n \in \mathbb{N}$$

Las condiciones iniciales determinan U_j^0 y U_j^1 y la ecuación en diferencias

$$U_j^{n+1} - U_j^{n-1} = 2(1-r)U_j^n + r\left(U_{j+1}^n + U_{j-1}^n\right) \tag{7.85}$$

con $r = h^2/k^2$ nos da la solución en el nivel temporal $n+1$ en función de los valores en los dos niveles anteriores n y $n-1$.

El esquema (7.85) es una aproximación consistente de la ecuación de las ondas, el error de truncación local es

$$-U_{xxxx}\left(\frac{k^2}{12} - \frac{h^2}{12}\right) + \cdots$$

El esquema es estable si $\dfrac{k}{h} \leq 1$.

PROBLEMAS

PROBLEMA 7.1 *Problema mixto para la ecuación de Fourier.*

Se considera el problema mixto definido por la ecuación de Fourier 2D

$$u_t = a^2\left(u_{xx} + u_{yy}\right) \quad (x,y) \in [0,1]^2, \quad t > 0$$

la condición inicial

$$u(x,y,0) = f(x,y*)$$

y la condición Dirichlet homogénea $u(x,y,t) = 0$ en los lados del cuadrado unidad para todo t.

Se define la malla de la Figura 7.6 donde los pasos espaciales son iguales $\Delta x = \Delta y = h = \dfrac{1}{N}$, k es el paso

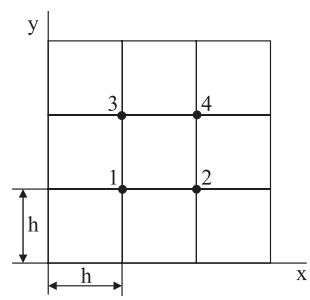


Figura 7.6: Malla del problema mixto para la ecuación del calor 2D.

temporal y se pone $r = \dfrac{a^2 k}{h^2}$.

Se desea utilizar el esquema implícito de Crank-Nicolson para aproximar el problema propuesto. Escribir las ecuaciones en diferencias correspondientes al problema y a la malla consideradas en forma matricial.

Solución:

Escribiendo para simplificar la notación las aproximaciones por diferencias centradas de segundo orden (7.36) y (7.38) mediante los operadores D_x^2 y D_y^2 respectivamente, el método implícito de Crank-Nicolson aplicado a nuestra ecuación es

$$\left[1 - \frac{r}{2}\left(D_x^2 + D_y^2\right)\right] u(x,y,t+\Delta t) = \left[1 + \frac{r}{2}\left(D_x^2 + D_y^2\right)\right] u(x,y,t)$$

y desarrollando

$$\begin{aligned}\frac{U_{i,j}^{n+1} - U_{i,j}^{n}}{k} =& \frac{a^2}{2}\left(\frac{U_{i+1,j}^{n+1} - 2U_{i,j}^{n+1} + U_{i-1,j}^{n+1}}{h^2} + \frac{U_{i+1,j}^{n} - 2U_{i,j}^{n} + U_{i-1,j}^{n}}{h^2}\right) + \\ &+ \frac{a^2}{2}\left(\frac{U_{i,j+1}^{n+1} - 2U_{i,j}^{n+1} + U_{i,j-1}^{n+1}}{h^2} + \frac{U_{i,j+1}^{n} - 2U_{i,j}^{n} + U_{i,j-1}^{n}}{h^2}\right)\end{aligned}$$

es decir,

$$\begin{aligned}&2(1+2r)U_{i,j}^{n+1} - r\left(U_{i+1,j}^{n+1} + U_{i-1,j}^{n+1} + U_{i,j+1}^{n+1} + U_{i,j-1}^{n+1}\right) = \\ &= 2(1-2r)U_{i,j}^{n} + r\left(U_{i+1,j}^{n} + U_{i-1,j}^{n} + U_{i,j+1}^{n} + U_{i,j-1}^{n}\right)\end{aligned}$$

Teniendo en cuenta la condición de contorno, todos los valores de u que tengan entre sus subíndices un 0 o un 3 son nulos ($U_{i,0} = U_{0,i} = U_{3,i} = U_{i,3} = 0 \quad i = 0,1,2,3$) y las ecuaciones relativas a los nodos numerados $1, 2, 3$ y 4 en la Figura 7.6 que son los de subíndices $(1,1)$, $(2,1)$, $(1,2)$, y $(2,2)$ respectivamente son

$$\begin{cases} 2(1+2r)U_{1,1}^{n+1} - r\left(U_{2,1}^{n+1} + U_{1,2}^{n+1}\right) = 2(1-2r)U_{1,1}^{n} + r\left(U_{2,1}^{n} + U_{1,2}^{n}\right) \\ 2(1+2r)U_{2,1}^{n+1} - r\left(U_{1,1}^{n+1} + U_{2,2}^{n+1}\right) = 2(1-2r)U_{2,1}^{n} + r\left(U_{1,1}^{n} + U_{2,2}^{n}\right) \\ 2(1+2r)U_{1,2}^{n+1} - r\left(U_{2,2}^{n+1} + U_{1,1}^{n+1}\right) = 2(1-2r)U_{1,2}^{n} + r\left(U_{2,2}^{n} + U_{1,1}^{n}\right) \\ 2(1+2r)U_{2,2}^{n+1} - r\left(U_{1,2}^{n+1} + U_{2,1}^{n+1}\right) = 2(1-2r)U_{2,2}^{n} + r\left(U_{1,2}^{n} + U_{2,1}^{n}\right) \end{cases}$$

y matricialmente,

$$\begin{pmatrix} 2(1+2r) & -r & -r & 0 \\ -r & 2(1+2r) & 0 & -r \\ -r & 0 & 2(1+2r) & -r \\ 0 & -r & -r & 2(1+2r) \end{pmatrix} \begin{pmatrix} U_{1,1} \\ U_{2,1} \\ U_{1,2} \\ U_{2,2} \end{pmatrix}^{n+1} =$$

$$= \begin{pmatrix} 2(1-2r) & r & r & 0 \\ r & 2(1-2r) & 0 & r \\ r & 0 & 2(1-2r) & r \\ 0 & r & r & 2(1-2r) \end{pmatrix} \begin{pmatrix} U_{1,1} \\ U_{2,1} \\ U_{1,2} \\ U_{2,2} \end{pmatrix}^{n}$$

PROBLEMA 7.2 *Problema de Dirichlet para la ecuación de Poisson.*

Se considera en el dominio de la Figura 7.7 el problema de Dirichlet para la ecuación de Poisson

$$u_{xx} + u_{yy} = 1$$

con condición de contorno homogénea $u(x,y) = 0$ y la malla uniforme allí representada.

Utilizando diferencias finitas centradas, escribir el sistema de ecuaciones en diferencias finitas correspondiente.

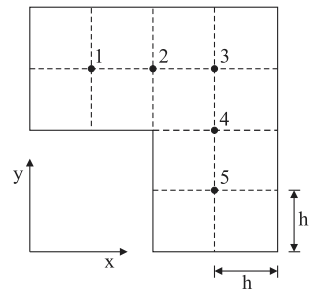


Figura 7.7: Dominio discretizado del problema de Dirichlet para la ecuación de Poisson.

Solución:
Utilizamos el esquema en diferencias (7.74)

$$\frac{U_{i+1,j}+U_{i-1,j}+U_{i,j+1}+U_{i,j-1}-4U_{i,j}}{h^2}=1$$

con la condición de tipo Dirichlet homogénea y escribimos las ecuaciones relativas a los nodos numerados $1,2,3,4$ y 5 de la Figura 7.7 que son los de subíndices $(1,3)$, $(2,3)$, $(3,3)$, $(3,2)$, y $(3,1)$ respectivamente en la numeración elegida de los nodos de la malla computacional. Teniendo en cuenta la condición de contorno obtenemos

$$\begin{cases} -4U_{1,3}+U_{2,3}=h^2 & \Longrightarrow & -4U_1+U_2=h^2 \\ U_{3,3}-4U_{2,3}+U_{1,3}=h^2 & \Longrightarrow & U_1-4U_2+U_3=h^2 \\ U_{2,3}-4U_{3,3}+U_{3,2}=h^2 & \Longrightarrow & U_2-4U_3+U_4=h^2 \\ U_{3,3}-4U_{3,2}+U_{3,1}=h^2 & \Longrightarrow & U_3-4U_4+U_5=h^2 \\ U_{3,2}-4U_{3,1}=h^2 & \Longrightarrow & U_4-4U_5=h^2 \end{cases}$$

y matricialmente,

$$\frac{1}{h^2}\begin{pmatrix} -4 & 1 & 0 & 0 & 0 \\ 1 & -4 & 1 & 0 & 0 \\ 0 & 1 & -4 & 1 & 0 \\ 0 & 0 & 1 & -4 & 1 \\ 0 & 0 & 0 & 1 & -4 \end{pmatrix}\begin{pmatrix} U_1 \\ U_2 \\ U_3 \\ U_4 \\ U_5 \end{pmatrix}=\begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}$$

PROBLEMA 7.3 *Ecuación de difusión 2D.*

Se considera el problema de evolución definido por la ecuación de difusión 2D

$$T_t=T_{xx}+T_{yy}$$

y las condiciones de contorno e iniciales

$$\begin{cases} T(0,y,t)=T(1.2,y,t)=T(x,0,t)=T(x,1,t)=0 \quad (\forall t\in I\!R^+) \\ \\ T(0.4,0.5,0)=T(0.8,0.5,0)=626 \end{cases}$$

1. Dar una interpretación física del problema (P) propuesto especificando el dominio abierto Ω en el que se produce el proceso evolutivo.

   Se malla el dominio Ω tomando pasos espaciales $h_1=\Delta x$ y $h_2=\Delta y$ a lo largo de los ejes x e y y paso temporal $k=\Delta t$. Denominaremos $\mathbf{h}=(h_1,h_2,k)$ y pondremos $r_1=\dfrac{k}{h_1^2}$, $r_2=\dfrac{k}{h_2^2}$. Denotaremos $\Omega_{\mathbf{h}}$ el dominio discretizado (Figura 7.8).

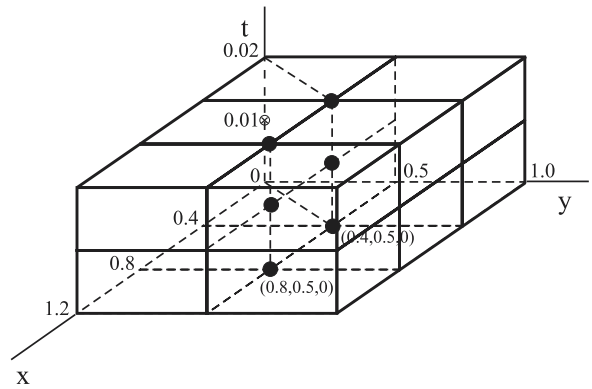


Figura 7.8: Dominio discretizado del problema.

2. Utilizando los operadores de tres puntos y orden dos (7.36) y (7.38) para aproximar las derivadas parciales segundas en x e y y una aproximación de orden uno de T_t por diferencias finitas progresiva, construir el esquema explícito en diferencias que aproxima la solución numérica $T_{i,j}^n$ de (P). Denotaremos $\mathcal{T}_{i,j}^n$ dicha aproximación.

3. Tomando $k = 0.01$; $h_1 = 0.4$; $h_2 = 0.5$, dar 3 pasos hacia adelante en el esquema $(n = 0, 1, 2)$, escribiendo el sistema que permita describir la evolución temporal de $T_{1,1}^0$ y $T_{2,1}^0$.

Solución:

1. El problema matemático propuesto está compuesto por la ecuación de Fourier 2D adimensionalizada, versión no estacionaria de la ecuación de Laplace que gobierna la distribución en régimen estacionario de la temperatura en un dominio espacial bidimensional, en nuestro caso una placa rectangular $[0, 1.2] \times [0, 1] \subset \mathbb{R}^2$. Se describe, por tanto, la evolución temporal de la temperatura en dicha placa a partir del instante inicial en el que todos los puntos de la placa están a temperatura cero excepto dos focos puntuales de calor en los puntos $(0.4, 0.5)$ y $(0.8, 0.5)$, supuesto que a lo largo del proceso los lados de la placa se mantienen a cero grados.

2.

$$T_t\big|_{i,j}^n = (T_{xx} + T_{yy})\big|_{i,j}^n \quad \Rightarrow \quad \frac{T_{i,j}^{n+1} - T_{i,j}^n}{k} + O(k) = \frac{T_{i+1,j}^n - 2T_{i,j}^n + T_{i-1,j}^n}{h_1^2} + O(h_1^2) + \frac{T_{i,j+1}^n - 2T_{i,j}^n + T_{i,j-1}^n}{h_2^2} + O(h_2^2)$$

de donde

$$\mathcal{T}_{i,j}^{n+1} = (1 - 2r_1 - 2r_2)\mathcal{T}_{i,j}^n + r_1(\mathcal{T}_{i+1,j}n + \mathcal{T}_{i-1,j}n) + r_2(\mathcal{T}_{i,j+1}n + \mathcal{T}_{i,j-1}n) \tag{7.86}$$

Las condiciones de contorno e iniciales discretizadas son

$$\begin{aligned}
&T(0, y, t) = 0 \quad \forall y, t \quad \Rightarrow \quad \mathcal{T}_{0,j}^n = 0 \quad j = 1, 2; \quad n = 0, 1, 2\\
&T(1.2, y, t) = 0 \quad \forall y, t \quad \Rightarrow \quad \mathcal{T}_{3,j}^n = 0 \quad j = 1, 2; \quad n = 0, 1, 2\\
&T(x, 0, t) = 0 \quad \forall x, t \quad \Rightarrow \quad \mathcal{T}_{i,0}^n = 0 \quad i = 1, 2, 3; \quad n = 0, 1, 2\\
&T(x, 1, t) = 0 \quad \forall x, t \quad \Rightarrow \quad \mathcal{T}_{i,2}^n = 0 \quad i = 1, 2, 3; \quad n = 0, 1, 2
\end{aligned}$$

y

$$T(0.4, 0.5, 0) = 626 \quad \Rightarrow \quad \mathcal{T}_{1,1}^0 = 626\ ; \qquad T(0.8, 0.5, 0) = 626 \quad \Rightarrow \quad \mathcal{T}_{2,1}^0 = 626$$

3. Llamando $\xi = 1-2r_1-2r_2$ y dando a (i,j) los valores $(1,1)$ y $(2,1)$ el esquema (7.86) y las condiciones suplementarias nos permiten escribir el sistema de ecuaciones

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ -\xi & 1 & 0 & -r_1 & 0 & 0 \\ 0 & -\xi & 1 & 0 & -r_1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ -r_1 & 0 & 0 & -\xi & 1 & 0 \\ 0 & -r_1 & 0 & 0 & -\xi & 1 \end{pmatrix} \begin{pmatrix} \mathcal{T}_{1,1}^1 \\ \mathcal{T}_{1,1}^2 \\ \mathcal{T}_{1,1}^3 \\ \mathcal{T}_{2,1}^1 \\ \mathcal{T}_{2,1}^2 \\ \mathcal{T}_{2,1}^3 \end{pmatrix} = \begin{pmatrix} 626(\xi + r_1) \\ 0 \\ 0 \\ 626(\xi + r_1) \\ 0 \\ 0 \end{pmatrix}$$

Sustituyendo los valores $r_1 = 0.0625$, $r_2 = 0.04$ y $\xi = 0.795$ llegamos al sistema

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ -0.795 & 1 & 0 & -0.0625 & 0 & 0 \\ 0 & -0.795 & 1 & 0 & -0.0625 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ -0.0625 & 0 & 0 & -0.795 & 1 & 0 \\ 0 & -0.0625 & 0 & 0 & -0.795 & 1 \end{pmatrix} \begin{pmatrix} \mathcal{T}_{1,1}^1 \\ \mathcal{T}_{1,1}^2 \\ \mathcal{T}_{1,1}^3 \\ \mathcal{T}_{2,1}^1 \\ \mathcal{T}_{2,1}^2 \\ \mathcal{T}_{2,1}^3 \end{pmatrix} = \begin{pmatrix} 536.795 \\ 0 \\ 0 \\ 536.795 \\ 0 \\ 0 \end{pmatrix}$$

Una aplicación directa de MATLAB nos da la solución que disponemos en las tablas anexas.

$\mathcal{T}_{1,1}^0$	$\mathcal{T}_{1,1}^1$	$\mathcal{T}_{1,1}^2$	$\mathcal{T}_{1,1}^3$
626	536.7950	460.3017	394.7087

$\mathcal{T}_{2,1}^0$	$\mathcal{T}_{2,1}^1$	$\mathcal{T}_{2,1}^2$	$\mathcal{T}_{2,1}^3$
626	536.7950	460.3017	394.7087

Parece lógico que los dos nodos disminuyan su temperatura de la misma forma, ya que los focos a los que ceden calor están a la misma temperatura y el medio les rodea uniformemente.

PROBLEMA 7.4 *Ecuación elíptica con condiciones mezcladas.*

Dadas la ecuación diferencial en derivadas parciales elíptica

$$u_{xx} + u_{yy} + u_y = f(x,y)$$

con las condiciones de contorno tipo Dirichlet $u = u_I$ y $u = u_D$ en los lados verticales izquierdo y derecho respectivamente y las de tipo Neumann $\dfrac{\partial u}{\partial y} = 0$ en los dos lados horizontales del cuadrado cuya malla se especifica en la Figura 7.9, escribir la ecuación en diferencias finitas en los nodos 2, 17, 18 y 23 usando

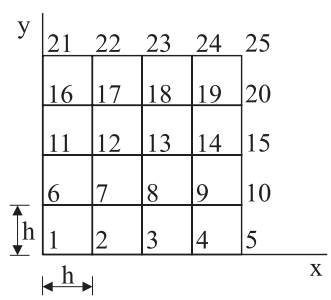


Figura 7.9: Dominio discretizado del problema.

diferencias centradas.

Solución:
La aproximación de orden 2 en diferencias centradas del problema propuesto es

$$\frac{U_{i+1,j}-2U_{i,j}+U_{i-1,j}}{h^2}+\frac{U_{i,j+1}-2U_{i,j}+U_{i,j-1}}{h^2}+\frac{U_{i,j+1}-U_{i,j-1}}{2h}=f_{i,j}$$

- En el nodo $2 \simeq (1,0)$ no hay problema para aproximar la derivada segunda respecto de x. Para la derivada segunda respecto de y usamos

$$U_{yy}\big|_{1,0}=\frac{U_{1,1}-2U_{1,0}+U_{1,-1}}{k^2} \tag{7.87}$$

$U_{1,-1}$ no está definido, pero se puede determinar a partir de la condición de Neumann.

Partiendo de

$$U_y\big|_{1,0}=\frac{U_{1,-1}-U_{1,1}}{2h}=0$$

obtenemos $U_{1,-1}=U_{1,1}$ y de aquí

$$U_{yy}\big|_{1,0}=2\frac{U_{1,1}-U_{1,0}}{h^2}$$

con ello teniendo en cuenta que $U_{0,0}=U_I$

$$\frac{U_{2,0}-2U_{1,0}+U_{0,0}}{h^2}+2\frac{U_{1,1}-U_{1,0}}{h^2}+0=f_{1,0} \quad \Rightarrow$$
$$\Rightarrow \quad \frac{1}{h^2}U_3-\frac{4}{h^2}U_2+\frac{2}{h^2}U_7+=f_2-\frac{U_I}{h^2}$$

- La molécula computacional relativa al nodo $17 \simeq (1,3)$ permite escribir teniendo en cuenta que $U_{0,3}=U_I$

$$\frac{U_{2,3}-2U_{1,3}+U_{0,3}}{h^2}+\frac{U_{1,4}-2U_{1,3}+U_{1,2}}{h^2}+\frac{U_{2,2}-U_{1,2}}{2h}=f_{1,3}$$

es decir,

$$\frac{U_{18}-2U_{17}+15}{h^2}+\frac{U_{22}-2U_{17}+U_{12}}{h^2}+\frac{U_{22}-U_{12}}{2h}=f_{17} \quad \Rightarrow$$
$$\Rightarrow \quad \left(\frac{1}{h^2}-\frac{1}{2h}\right)U_{12}-\frac{4}{h^2}U_{17}+\frac{1}{h^2}U_{18}+\left(\frac{1}{h^2}+\frac{1}{2h}\right)U_{22}=f_{17}-\frac{U_I}{h^2}$$

- De modo análogo se escribe

$$\frac{U_{19}-2U_{18}+U_{17}}{h^2}+\frac{U_{23}-2U_{18}+U_{13}}{h^2}+\frac{U_{23}-U_{13}}{2h}=f_{18} \quad \Rightarrow$$
$$\Rightarrow \quad \left(\frac{1}{h^2}-\frac{1}{2h}\right)U_{13}+\frac{1}{h^2}U_{17}-\frac{4}{h^2}U_{18}+\frac{1}{h^2}U_{19}+\left(\frac{1}{h^2}+\frac{1}{2h}\right)U_{23}=f_{18}$$

- En este caso, como en el nodo 2, no hay problema para aproximar la derivada segunda respecto de x. Para la derivada segunda respecto de y usamos de nuevo (7.87) convenientemente adaptada y se tiene

$$U_{yy}\big|_{2,4}=2\frac{U_{2,3}-U_{2,4}}{h^2}$$

con ello,

$$\frac{U_{24}-2U_{23}+U_{22}}{h^2}+2\frac{U_{18}-U_{23}}{h^2}=f_{23} \quad \Rightarrow$$
$$\Rightarrow \quad \frac{2}{h^2}U_{18}+\frac{1}{h^2}U_{22}-\frac{4}{h^2}U_{23}+\frac{1}{h^2}U_{24}+=f_{23}$$

Se sugiere que el lector acabe este ejercicio escribiendo el sistema de las 15 ecuaciones en diferencias finitas y 15 incógnitas asociado al problema en estudio, en forma matricial.

PROBLEMA 7.5 *Aproximación lateral de u_{xx}.*

Obtener en el punto (x_i, y_j) de $\mathcal{M}_{h,k}$ una aproximación lateral de u_{xx} por diferencias progresivas basada en la información en los dos nodos a la derecha de (x_i, y_j).

Repetir cambiando de lado de aproximación.

Estas aproximaciones de primer orden son importantes cuando sólo se dispone de información a los lados del punto en estudio.

Solución:

Escribiendo la fórmula de Taylor de $u_{i+2,j}$ y de $u_{i+1,j}$ tomando el punto (x_i, y_j) como base

$$u_{i+2,j} = u_{i,j} + u_x\big|_{i,j} 2h + u_{xx}\big|_{i,j} \frac{(2h)^2}{2!} + O(h^3)$$
$$u_{i+1,j} = u_{i,j} + u_x\big|_{i,j} h + u_{xx}\big|_{i,j} \frac{h^2}{2!} + O(h^3)$$

Multiplicando la segunda ecuación por -2 y sumando ambos desarrollos

$$u_{i+2,j} - 2u_{i+1,j} = -u_{i,j} + h^2 u_{xx}\big|_{i,j} + O(h^3)$$

de donde

$$u_{xx}\big|_{i,j} = \frac{u_{i+2,j} - 2u_{i+1,j} + u_{i,j}}{h^2} + O(h)$$

Se obtiene el valor de u_{xx} en el punto (x_i, y_j) en función de los valores de u en tres puntos, (x_i, y_j) y los dos puntos a su derecha (x_{i+1}, y_j), (x_{i+2}, y_j).

Si escribimos la fórmula de Taylor de $u_{i-2,j}$ y de $u_{i-1,j}$ tomando el punto (x_i, y_j) como base

$$u_{i-2,j} = u_{i,j} - u_x\big|_{i,j} 2h + u_{xx}\big|_{i,j} \frac{(-2h)^2}{2!} + O(h^3)$$
$$u_{i-1,j} = u_{i,j} - u_x\big|_{i,j} h + u_{xx}\big|_{i,j} \frac{(-h)^2}{2!} + O(h^3)$$

Multiplicando de nuevo la segunda ecuación por -2 y sumando

$$u_{i-2,j} - 2u_{i-1,j} = -u_{i,j} + h^2 u_{xx}\big|_{i,j} + O(h^3)$$

de donde

$$u_{xx}\big|_{i,j} = \frac{u_{i,j} - 2u_{i-1,j} + u_{i-2,j}}{h^2} + O(h)$$

PROBLEMA 7.6 *Condición de contorno de tipo Neumann y extrapolación.*

Se considera una condición de Neumann sobre el lado vertical derecho de un dominio rectangular $\Omega = [0, b] \times [0, d]$ (Figura 7.10).

Como la normal exterior es $\mathbf{n} = (1, 0)$ dicha condición se expresa aquí

$$u_x(b, y) = \psi(y) \qquad 0 \le y \le d \tag{7.88}$$

Se discretiza Ω tomando $h = b/N$ y $k = d/M$. $\Omega_{h,k}$ es la malla computacional asociada.

El objetivo de este ejercicio es escribir $u_{N,j}$ extrapolando los valores de u en los puntos interiores de la fila j-ésima a la frontera.

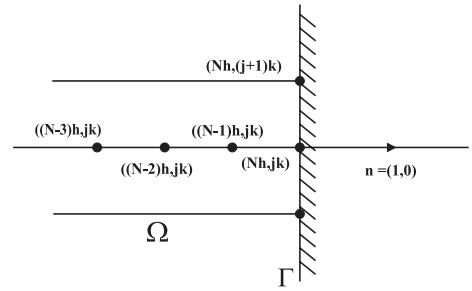


Figura 7.10: Condición de Neumann en el lado vertical derecho.

Para j fijo en $1,..,M$ sea u_j la aplicación $u_j(x) = u(x, jk)$ con $x \in [0, b]$.

1. Escribir un polinomio de Newton de grado m en diferencias regresivas no divididas con base en el nodo (Nh, jk)

$$p(t) = P(x_N + th, j) = \sum_{i=0}^{m-1} a_i \nabla^i u_{N,j} \tag{7.89}$$

2. Derivando $p(t)$ respecto de t y haciendo $t = 0$ obtener una aproximación en diferencias regresivas no divididas de $u_x\Big|_{N,j}$.

3. Utilizando la condición de Neumann (7.88) y la expresión obtenida en el apartado anterior para distintos valores de m, hallar aproximaciones de $u_{N,j}$, precisando el orden de la aproximación [14].

Solución:

1. Se tiene

$$P(x_N + th, j) = u_{N,j} + \frac{t}{1!}\nabla^1 u_{N,j} + \frac{t(t+1)}{2!}\nabla^2 u_{N,j} + \cdots \\ \cdots + \frac{t(t+1)...(t+m-2)}{(m-1)!}\nabla^{m-1} u_{N,j} \tag{7.90}$$

con

$$\nabla^0 u_{N,j} = u_{N,j} \quad \text{y} \quad \nabla^{j+1} u_{N,j} = \nabla^j u_{N,j} - \nabla^j u_{N-1,j}$$

2. Derivando (7.90) respecto de t [15]

$$p'(t) = tP'(x_N + th, j) = \frac{1}{h}\left[\nabla^1 u_{N,j} + \frac{2t+1}{2}\nabla^2 u_{N,j} + \frac{3t^2+6t+2}{6}\nabla^3 u_{N,j} + \cdots\right]$$

Haciendo $t = 0$

$$u_x\Big|_{N,j} = \frac{1}{h}\left[\nabla^1 u_{N,j} + \frac{1}{2}\nabla^2 u_{N,j} + \frac{1}{3}\nabla^3 u_{N,j} + \cdots\right] \tag{7.91}$$

donde el error de truncación es del orden del primer término despreciado.

[14] Si el orden del primer término despreciado es m, el error es $O(h^m)$.

[15] Podemos escribir el polinomio $p(t)$ mediante la fórmula

$$p(t) = \sum_{i=0}^{m-1} (-1)^i \binom{-t}{i} \nabla^i u_{N,j}$$

donde

$$\binom{-t}{i} = \frac{-t(-t-1)...(-t-(i-1))}{i!} = (-1)^i \frac{t(t+1)...(t+i-1)}{i!}$$

El primer miembro de (7.91) viene definido en la condición (7.88) por

$$u_x\Big|_{N,j} = \psi_j$$

de modo que sustituyendo en (7.91), operando las diferencias divididas y despejando $u_{N,j}$ obtenemos para distintos valores de m.

Extrapolación lineal

$$u_x\Big|_{N,j} = \frac{1}{h}\left[\nabla u_{N,j} + O(h^2)\right] = \frac{1}{h}\left(u_{N,j} - u_{N-1,j}\right) + O(h)$$

de donde

$$u_{N,j} = hu_x\Big|_{N,j} + u_{N-1,j} + hO(h)$$

y eliminando el error de truncación

$$U_{N,j} = U_{N-1,j} + h\psi_j \tag{7.92}$$

Extrapolación cuadrática

$$\begin{aligned} u_x\Big|_{N,j} &= \frac{1}{h}\left[\nabla u_{N,j} + \frac{1}{2}\nabla^2 u_{N,j} + O(h^3)\right] = \\ &= \frac{1}{h}\left(u_{N,j} - u_{N-1,j} + \frac{1}{2}(u_{N,j} - 2u_{N-1,j} + u_{N-2,j})\right) + O(h) \end{aligned}$$

de donde

$$u_{N,j} = \frac{1}{3}\left(4u_{N-1,j} - u_{N-2,j} - 2hu_x\Big|_{N,j}\right) + hO(h^2)$$

y eliminando el error de truncación

$$U_{N,j} = \frac{1}{3}\left(4U_{N-1,j} - U_{N-2,j} - 2h\psi_j\right) \tag{7.93}$$

De estas aproximaciones la lineal (7.92) plantea problemas por falta de precisión; por ejemplo cuando se usa en la fórmula de Laplace de cinco puntos. La idoneidad de la aproximación cuadrática (7.93) depende del problema en estudio (ver el problema 7.13).

PROBLEMA 7.7 *Transmisión de calor en régimen permanente.*

Se considera una placa triangular D de vértices $(0,0)$, $(1,0)$ y $(0,1)$. Se trata de conocer la temperatura $T(x,y)$ en su interior suponiendo que ésta verifica la ecuación de Laplace, y que en la frontera de la placa se tienen las condiciones de Dirichlet

$$T(x,y) = \begin{cases} 0 & x = 0 \\ 1000x & y = 1 - x \\ 1000x^2 & y = 0 \end{cases}$$

1. Se pide estimarla con un esquema de diferencias finitas utilizando un operador centrado de tres puntos para las derivadas segundas en x e y, y tomando una malla de igual paso 0.25 en ambas direcciones.

2. Estimar la temperatura en la placa a lo largo de la recta $y = 0.75 - x$ utilizando un spline cúbico natural que se apoya en los nodos de la discretización del problema que están sobre esa recta. Se pide calcular las derivadas en los nodos de dicho spline cúbico natural.

3. Utilizando este spline, estimar el valor de la temperatura en un punto que no pertenece a la discretización pero sí a la recta $y = 0.75 - x$ y que tiene como coordenadas $x = y = 0.375$.

Solución:

1. Utilizamos la discretización (7.41) de la laplaciana de modo que la ecuación en diferencias asociada a la ecuación de Laplace es

$$T_{i,j+1} + T_{i,j-1} + T_{i+1,j} + T_{i-1,j} - 4T_{i,j} = 0$$

Siendo i el índice x del nodo y j el índice y. Si aplicamos esta ecuación a los tres nodos interiores (ver Figura 7.11), $(1,1)$, $(2,1)$ y $(1,2)$, obtenemos el sistema lineal

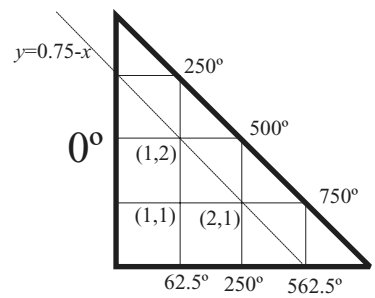


Figura 7.11: Malla del triángulo.

$$\begin{pmatrix} -4 & 1 & 1 \\ 1 & -4 & 0 \\ 1 & 0 & -4 \end{pmatrix} \begin{pmatrix} T_{11} \\ T_{21} \\ T_{12} \end{pmatrix} = \begin{pmatrix} -62.5 \\ -1500 \\ -750 \end{pmatrix}$$

Para obtener la segunda línea del sistema, hemos aplicado la ecuación discretizada al nodo $(2,1)$:

$$T_{2,2} + T_{2,0} + T_{3,1} + T_{1,1} - 4T_{2,1} = 0$$

Varios de estos puntos están en el contorno, el $T_{2,2}$ cuya temperatura es $500°$, el $T_{3,1}$ cuya temperatura es $750°$, y el $T_{2,0}$ cuya temperatura es $250°$, pasando estos valores al primer miembro obtenemos la segunda línea.

$$T_{1,1} - 4T_{2,1} = -1500$$

El mismo proceso se repite para las otras dos líneas del sistema lineal. Resolviendo este sistema lineal con Matlab, llegamos a:

$$\begin{pmatrix} T_{11} \\ T_{21} \\ T_{12} \end{pmatrix} = \begin{pmatrix} 178.5714 \\ 419.6429 \\ 232.1429 \end{pmatrix}$$

Este procedimiento se puede generalizar a cualquier número de nodos. En las Figuras 7.12, 7.13 presentamos los resultados correspondientes a 1.800 nodos, y hemos incluido el correspondiente código Matlab *triangulo.m* en la página web vinculada al libro.

2. Si tenemos un spline cúbico asociado a una partición equiespaciada $\Omega = \{x_0, \cdots, x_n\}$, las derivadas en los nodos verifican la relación (3.19)

$$s_i + 4s_{i+1} + s_{i+2} = \frac{3}{h}(w_{i+2} - w_i)$$

siendo w_i el valor de la función interpolada en el nodo i y h el intervalo entre nodos. Si aplicamos esta relación a los $n-1$ nodos donde esto es posible, nos faltan dos ecuaciones adicionales para fijar estas derivadas. Estas dos ecuaciones corresponden al hecho de que nuestro spline es natural, o sea, su derivada segunda en el primer y último nodo es 0.
Se trata de aplicar las dos ecuaciones (3.21), para tener las dos líneas adicionales del sistema lineal.

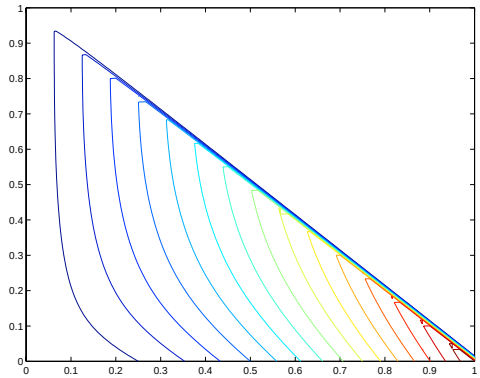


Figura 7.12: Líneas de nivel de la función temperatura.

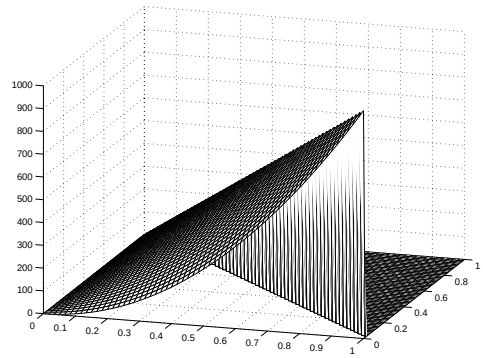


Figura 7.13: Superficie de la función temperatura.

En este caso en particular tenemos un spline natural con 3 tramos y 4 nodos. El espaciado es $h = 0.25\sqrt{2} = 0.3536$, y el vector de valores de la función temperatura en los nodos es

$$w = (0.0000, 232.1429, 419.6429, 562.5000)^t$$

El sistema lineal a resolver queda por tanto:

$$\begin{pmatrix} 4 & 2 & 0 & 0 \\ 1 & 4 & 1 & 0 \\ 0 & 1 & 4 & 1 \\ 0 & 0 & 2 & 4 \end{pmatrix} \begin{pmatrix} s_0 \\ s_1 \\ s_2 \\ s_3 \end{pmatrix} = \frac{3}{h} \begin{pmatrix} 2(w_1 - w_0) \\ w_2 - w_0 \\ w_3 - w_1 \\ 2(w_3 - w_2) \end{pmatrix} = \begin{pmatrix} 3939.6 \\ 3560.8 \\ 2803.2 \\ 2424.4 \end{pmatrix} \quad \Rightarrow \quad \begin{pmatrix} s_0 \\ s_1 \\ s_2 \\ s_3 \end{pmatrix} = \begin{pmatrix} 681.8531 \\ 606.0916 \\ 454.5686 \\ 378.8071 \end{pmatrix}$$

3. El punto $(0.375, 0.375)$ está en el segundo tramo del spline, a mitad del mismo. Si asignamos la abscisa $t = 0.0$ al nodo $(0.0, 0.75)$, el nodo $(0.25, 0.50)$ tendrá como abscisa $t = 0.25\sqrt{2} = 0.3536$, y el tercer nodo, el $(0.50, 0.25)$ tendrá como abscisa $t = 2 \cdot 0.25\sqrt{2} = 0.7071$. A medio camino entre estas dos está la abscisa del punto donde queremos estimar el valor de la temperatura $t = 0.5303$. Para obtener el valor en este punto, lo más sencillo en este caso es utilizar diferencias divididas de Newton, imponiendo las derivadas como diferencias divididas de primer orden en los dos nodos, y así calcular la cúbica correspondiente al tramo en el que se encuentra el punto:

t_i	f_i	$f[t_i, t_{i+1}]$	$f[t_i, t_{i+1}, t_{i+2}]$	$f[t_i, t_{i+1}, t_{i+2}, t_{i+3}]$
0.3536	232.1429			
0.3536	232.1429	606.0916		
0.7071	419.6429	530.4102	-214.0917	
0.7071	419.6429	454.5686	-214.5448	-1.2819

El tramo de spline buscado es por tanto:

$$\begin{aligned} C(t) &= 232.1429 + 606.0916(t - 0.3536) - 214.0917(t - 0.3536)^2 - \\ &- 1.2819(t - 0.3536)^2(t - 0.7071) \quad \Rightarrow \quad C(0.5303) = 332.5618 \end{aligned}$$

Que es un valor razonable, a medio camino entre el valor en los dos nodos. Podemos extraer del ejemplo en el que hemos utilizando 1800 nodos esa línea completa y compararla con su spline cúbico natural que se apoya en esos mismos cuatro nodos para comprobar cómo de buena es esta interpolación. En la Figura 7.14 podemos observar que el ajuste es perfecto.

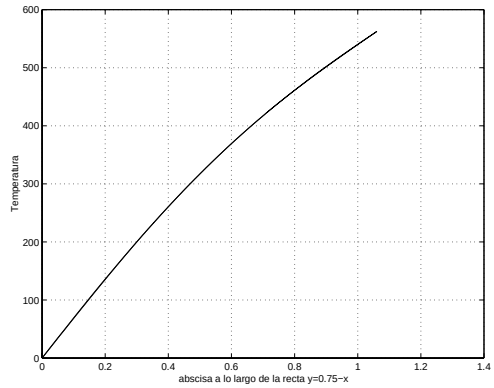


Figura 7.14: Prob. 7.7. Temperatura a lo largo la recta $y = 0.75 - x$.

PROBLEMA 7.8 *Problema de contorno unidimensional.*

Se tiene un escenario similar al problema 6.8 pero con unas condiciones de contorno que lo hacen completamente diferente. Ya no es un problema de valor inicial, sino que es un problema de contorno, y su resolución es similar a la resolución numérica de una ecuación en derivadas parciales elíptica.
Una caja de masa m que desliza sobre una rampa empujada por su propio peso (gravedad $g = 10$). El movimiento de la caja sobre la rampa está afectado de una fuerza de rozamiento contraria al movimiento, proporcional a la reacción sobre el plano con un factor $\nu = 0.1$ (ver la Figura 7.15). La rampa cambia su inclinación (argumentos en radianes) con el tiempo siguiendo la ley $\theta(t) = 0.5 + 0.5\ln(t+1)$, coincidiendo el origen del sistema de coordenadas con el eje de giro.
Si planteamos este problema en un sistema de coordenadas polares con origen en el eje de giro del plano, las ecuaciones que permiten calcular las diferentes fuerzas que intervienen son:
Para la fuerza normal N que el plano ejerce sobre la caja

$$N = m\left(g\cos\theta(t) + 2r'(t)\omega(t) + r\alpha(t)\right)$$

donde

$$\omega = \theta'(t) \qquad \alpha = \theta''(t)$$

La fuerza de rozamiento F proporcional a esta fuerza normal y que actúa en la dirección contraria al movimiento (la caja tiene una velocidad sobre la rampa de signo negativo como veremos más abajo).

$$F = -\nu N \text{ signo}\left[r'(t)\right] = -\nu N(-1) = \nu N$$

Quedan las llamadas "fuerzas"de inercia I que aparecen debido a que el sistema de referencia no es inercial y la proyección del peso sobre la rampa W.

$$I = mr\omega^2$$
$$W = -mg\sin\theta$$

La ley de Newton suministra la ecuación que gobierna nuestro modelo físico

$$mr''(t) = I + W + F$$

y sustituyendo las correspondientes fuerzas

$$r''(t) = 2\nu\omega(t)\, r'(t) + \left(\omega^2(t) + \nu\alpha(t)\right) r(t) + g(\nu\cos\theta(t) - \sin\theta(t)) \qquad (*)$$

En el origen de tiempos se desconoce la velocidad de la caja pero se conoce su posición. La caja avanza con velocidad negativa sobre la rampa hacia el eje y está a 10 metros de él. O sea, $r(0) = 10$. Se toman posiciones 1 segundo después y la caja ha avanzado 5 metros sobre la rampa. Por tanto, $r(0) = 10$, $r(1) = 5$. Se trata de conocer la posición r sobre la rampa en función del tiempo en este problema de contorno. Se utilizará el método de las diferencias finitas. Se pide:

1. Discretizar la ecuación (*) utilizando un operador de tres puntos para todas las derivadas y llamando h al paso temporal.

2. Suponiendo que $h = 0.2$, plantear el sistema lineal resultante de esquema anterior definiendo la matriz de coeficientes, el vector de incógnitas, y el vector de términos independientes. Resolver con Matlab y razonar si los resultados son físicamente razonables.

3. Vamos a resolver este sistema lineal mediante el método de Gauss-Seidel (GS). Razonar la elección de un estimador inicial en base a las características físicas del problema. Dar un paso en el método de GS utilizando ese estimador inicial.

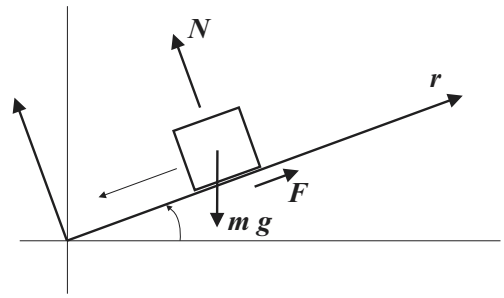


Figura 7.15: Plano inclinado correspondiente al problema 7.8.

Solución:
La resolución numérica por diferencias finitas de este problema de contorno unidimensional se aborda de modo similar a un problema elíptico, planteando un sistema lineal a partir de la ecuación diferencial y de las condiciones de contorno que pasan a ser valores conocidos entre las incógnitas de ese sistema lineal.

1. Discreticemos la ecuación (*) usando los operadores indicados en el enunciado:

$$\frac{r_{(i+1)} - 2r_{(i)} + r_{(i-1)}}{h^2} = 2\nu\omega_i \frac{r_{(i+1)} - r_{(i-1)}}{2h} \left(\omega_{(i)}^2 + \nu\alpha_{(i)}\right) r_{(i)} + g(\nu\cos\theta_{(i)} - \sin\theta_{(i)})$$

con

$$\theta_{(i)} = \theta\left(t_{(i)}\right) = 0.5 + 0.5\ln\left(t_{(i)} + 1\right); \quad \omega_{(i)} = \omega\left(t_{(i)}\right) = \frac{0.5}{t_{(i)} + 1}; \quad \alpha_{(i)} = \alpha\left(t_{(i)}\right) = \frac{-0.5}{(t_{(i)} + 1)^2}$$

$$r_{(i+1)}\left(\frac{1}{h^2} - \frac{\nu\omega_{(i)}}{h}\right) + r_{(i)}\left(-\frac{2}{h^2} - \omega_{(i)}^2 - \nu\alpha_{(i)}\right) + r_{(i-1)}\left(\frac{1}{h^2} + \frac{\nu\omega_{(i)}}{h}\right) = g\left(\nu\cos\theta_{(i)} - \sin\theta_{(i)}\right)$$

2. Para $h = 0.2$, tenemos 4 incógnitas correspondientes a los puntos de $t_1 = 0.2$, $t_2 = 0.4$, $t_3 = 0.6$ y $t_4 = 0.8$. Tenemos condiciones de contorno en los dos extremos del intervalo, $r(0) = 0$ y $r(0.5) = -10$. Sea $r_{(i)}$ la estimación de $r(t_i)$, y sea r el vector con todas estas incógnitas.

$$r_{i+1}\left(100 - 0.5\,\omega_i\right) + r_i\left(-200 - \omega_i^2 - 0.1\,\alpha_i\right) + r_{i-1}\left(100 + 0.5\,\omega_i\right) = \cos\theta_i - 10\sin\theta_i$$

Para $i = 1$, $t = 0.2$ tenemos:

$$\theta_1 = 0.5912, \quad \omega_1 = 0.4167, \quad \alpha_1 = -0.3472$$

$$24.7917\,r_2 - 50.1389\,r_1 + 25.2083\,r_0 = -4.7430$$

El sumando $25.2083\,r_{(0)}$ valor conocido, pasa al primer miembro. Para $i = 4$ tenemos la misma situación con $r_{(5)} = 5$. Podemos escribir unas líneas Matlab para construir este sistema lineal.

```
nu=0.1;
h=0.2;
t=0.2:h:0.8;
theta=0.5+0.5*log(t+1);
omega=0.5./(t+1);
alpha=-0.5./(t+1).^2;
E=1/h^2+omega*nu/h; % factor correspondiente a r_{(i-1)}
I=-2/h^2-omega.^2-nu*alpha; % factor correspondiente a r_{(i)}
F=1/h^2-omega*nu/h; % factor correspondiente a r_{(i+1)}
TI=10*nu*cos(theta)-10*sin(theta); % parte derecha
b=TI';
r0=10;
r5=5;
b(1)=b(1)-E(1)*r0; % lo conocido pasa a la derecha
b(4)=b(4)-F(4)*r5;
A=[I(1) F(1) 0 0
   E(2) I(2) F(2) 0
   0 E(3) I(3) F(3)
   0 0 E(4) I(4)]
r=A\b;
```

Que si ejecutamos, nos dan los coeficientes del sistema lineal

$$\begin{pmatrix} -50.1389 & 24.7917 & 0 & 0 \\ 25.1786 & -50.1020 & 24.8214 & 0 \\ 0 & 25.1562 & -50.0781 & 24.8437 \\ 0 & 0 & 25.1389 & -50.0617 \end{pmatrix} \begin{pmatrix} r_{(1)} \\ r_{(2)} \\ r_{(3)} \\ r_{(4)} \end{pmatrix} = \begin{pmatrix} -256.8263 \\ -5.4111 \\ -5.9641 \\ -130.7354 \end{pmatrix}$$

$$\begin{pmatrix} r_{(1)} \\ r_{(2)} \\ r_{(3)} \\ r_{(4)} \end{pmatrix} = \begin{pmatrix} 9.3781 \\ 8.6069 \\ 7.6420 \\ 6.4490 \end{pmatrix}$$

que tienen sentido físico, pues, como podemos comprobar, el movimiento se va acelerando y a intervalos iguales de tiempo la distancia recorrida es mayor.

3. Si en vez de una discretización con pocos puntos quisiésemos obtener una solución con mayor precisión, tendríamos un sistema lineal con muchas incógnitas. Dado que ese sistema es de diagonal dominante, para resolverlo podemos utilizar cualquiera de los métodos iterativos del Capítulo 2.

   Para reproducirlo en nuestro caso con 4 incógnitas, elegimos el método de Gauss-Seidel. Tomaremos como estimador inicial el correspondiente a velocidad constante, que supone que los puntos estarán equiespaciados en el tiempo $\mathbf{r}^{(0)} = (9\ 8\ 7\ 6)^T$. Con el método de GS, tendremos:

$$(D-L)\mathbf{r}^{(1)} = -U\mathbf{r}^{(0)} + \mathbf{b} = \begin{pmatrix} -455.1596 \\ -179.1611 \\ -155.0266 \\ -130.7354 \end{pmatrix} \quad \Rightarrow \quad \mathbf{r}^{(1)} = \begin{pmatrix} 9.0780 \\ 8.1380 \\ 7.1837 \\ 6.2189 \end{pmatrix}$$

   Se deja como ejercicio calcular con Matlab el radio espectral de la matriz de iteración para justificar su convergencia sabiendo de antemano que al ser el sistema diagonalmente dominante su radio espectral va a ser menor que la unidad, dando además otros dos pasos en el esquema y obteniendo también los residuos $\|b - Ax\|_\infty$ de cada iteración.

4. La velocidad sobre la rampa es $r'(t)$. Para estimar esta derivada en el primer y último punto no podemos utilizar operadores centrados, dado que son extremos del intervalo. Podemos utilizar operadores de dos

puntos para tener, por ejemplo, para $t = 0$,

$$r'(0) \approx \frac{r_{(1)} - r_{(0)}}{0.2} = \frac{9.3781 - 10}{0.2} = -3.1095$$

y para $t = 1$

$$r'(1) \approx \frac{r_{(5)} - r_{(4)}}{0.2} = \frac{5 - 6.4490}{0.2} = -7.2450$$

Se obtiene una mejor aproximación a la velocidad si usamos operadores de tres puntos en los extremos, así, para $t = 0$

$$r'(0) \approx \frac{-3r_{(0)} + 4r_{(1)} - r_{(2)}}{2 \cdot 0.2} = -2.7366$$

y para $t = 1$

$$r'(1) \approx \frac{3r_{(5)} - 4r_{(4)} + r_{(3)}}{2 \cdot 0.2} = -7.8848$$

En cualquiera de las 2 estimaciones, la velocidad ha crecido dado que la caja se acelera en su caída.

PROBLEMA 7.9 *Ecuaciones hiperbólicas: ecuación de transporte.*

Nuestro objetivo es estudiar un modelo unidimensional de dispersión de poluentes en el aire.

Sea $c(x, t)$ la concentración de poluente en el punto x en el instante t. Sea U la velocidad del viento, la cual supondremos constante. Sea k la constante de difusión.

Un modelo matemático sencillo teniendo en cuenta los fenómenos de advección (transporte que el viento provoca en la masa de gas) y difusión (dispersión de la propia masa de gas que se va mezclando en el aire) conduce a la famosa ecuación de advección-difusión

$$\begin{cases} \dfrac{\partial c}{\partial t} + U\dfrac{\partial c}{\partial x} = k\dfrac{\partial^2 c}{\partial x^2} \\ c(x,0) = c_0(x) \quad 0 \le x \le L \quad \text{y} \quad c(0,t) = c(L,t) = 0 \end{cases}$$

Para poder estudiar este problema desde un punto de vista meramente numérico con esquemas por diferencias finitas sencillos, escribimos la ecuación anterior como:

$$\frac{\partial c}{\partial t} = -U\frac{\partial c}{\partial x} + k\frac{\partial^2 c}{\partial x^2} \tag{7.94}$$

de modo que la podemos integrar como hace Carnahan et al.[4], en el ejemplo de la ecuación parabólica de transmisión de calor en régimen transitorio Capítulo 7.4.

1. Suponiendo que el paso espacial es Δx y el paso temporal es Δt, escribir de modo discreto la ecuación diferencial (7.94) usando operadores de tres puntos para las derivadas espaciales segundas, de dos puntos mirando hacia atrás para las derivadas espaciales primeras y un Euler explícito para el avance en el tiempo.

2. Se supone que $U = 1$, $L = 1$, $k = 0.2$, $\Delta x = 0.2$ y que:

$$c(x,0) = \begin{cases} 1 & x \in [0.1, 0.5] \\ 0 & x \notin [0.1, 0.5] \end{cases}$$

   Tomando un Δt genérico se pide sustituir todos estos valores y reescribir el esquema anterior dejándolo lo más simplificado posible.

3. Tomando $\Delta t = 0.01$, dar dos pasos de avance en el tiempo en el esquema del apartado anterior.

4. Ya que la concentración debe ser nula en los extremos del intervalo, al cabo del tiempo la concentración será nula en todos los puntos y el poluente se irá dispersando hasta desaparecer. Estudiar si esto va a ser así para $\Delta t = 0.01$ y $\Delta t = 0.08$ con la discretización espacial de los apartados anteriores.

Solución:

1. La ecuación (7.94) tiene la forma

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

Discretizamos el segundo miembro en espacio y tiempo como sugiere el enunciado

$$f(c, x, t) = -U\frac{\partial c}{\partial x} + k\frac{\partial^2 c}{\partial x^2} \approx -\frac{U}{\Delta x}\left(c_j^n - c_{j-1}^n\right) + \frac{k}{\Delta x^2}\left(c_{j+1}^n - 2c_j^n + c_{j-1}^n\right)$$

donde j es índice en espacio y n en tiempo.

Planteamos el integrador en el tiempo con el esquema de Euler explícito [16].

$$c_j^{n+1} = c_j^n + \Delta t f(c^{(n)}, x, t^{(n)})$$

o, lo que es lo mismo:

$$c_j^{n+1} = c_j^n - U\frac{\Delta t}{\Delta x}\left(c_j^n - c_{j-1}^n\right) + k\frac{\Delta t}{\Delta x^2}\left(c_{j+1}^n - 2c_j^n + c_{j-1}^n\right)$$

2. Sustituyendo los valores tendremos que:

$$c_j^{n+1} = c_j^n\left(1 - 3\frac{\Delta t}{0.2}\right) + c_{j-1}^n\frac{\Delta t}{0.1} + c_{j+1}^n\frac{\Delta t}{0.2} \quad \Rightarrow \quad c_j^{n+1} = (1 - 15\Delta t)\, c_j^n + 10\Delta t c_{j-1}^n + 5\Delta t c_{j+1}^n$$

3. Para $\Delta t = 0.01$

$$c_j^{n+1} = 0.85c_j^n + 0.1c_{j-1}^n + 0.05c_{j+1}^n$$

En el instante inicial, el vector de concentraciones para los nodos espaciales, aplicando las condiciciones iniciales es

$$\mathbf{c}^{(0)} = (0, 1, 1, 0, 0, 0)$$

Si utilizamos el esquema anterior, tendremos que:

$$\mathbf{c}^{(1)} = (0, 0.9, 0.95, 0.1, 0, 0) \quad \text{y} \quad \mathbf{c}^{(2)} = (0, 0.8125, 0.9025, 0.1800, 0.01, 0)$$

Como vemos en la Figura 7.16, la nube de poluentes se está dispersando.

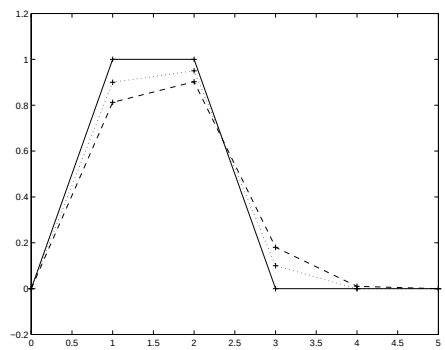


Figura 7.16: Distribución de poluentes en los tres instantes, $\Delta t = 0.01$.

[16] También se podría considerar el esquema Euler implícito, pero, como hemos visto varias veces, obliga a la resolución de sistemas lineales en cada paso de tiempo.

4. Si ponemos el esquema en forma matricial llamando $s = \Delta t/0.2$ tendremos:

$$\begin{pmatrix} c_1 \\ c_2 \\ c_3 \\ c_4 \end{pmatrix}^{(n+1)} = \begin{pmatrix} 1-3s & s & 0 & 0 \\ 2s & 1-3s & s & 0 \\ 0 & 2s & 1-3s & s \\ 0 & 0 & 2s & 1-3s \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \\ c_3 \\ c_4 \end{pmatrix}^{(n)}$$

Podemos ver el avance en el tiempo como un esquema de potencias de matrices:

$$\mathbf{c}^{(n+1)} = A\mathbf{c}^{(n)} \quad \Rightarrow \quad \mathbf{c}^{(n+1)} = A^{n+1}\mathbf{c}^{(0)}$$

Para que las concentraciones tiendan a cero, el radio espectral de la matriz A debe ser menor que la unidad. Para $\delta t = 0.01$, los autovalores de A son

$$(0.7356, 0.9644, 0.8937, 0.8063)$$

por lo que el esquema evoluciona correctamente. Para $\delta t = 0.08$ los autovalores de A son

$$(0.7153, -1.1153, -0.5496, 0.1496)$$

y, por tanto, su radio espectral es $1.1153 > 1$ y el esquema no evolucionará correctamente hacia la dispersión total de los poluentes.

Para calcular estos autovalores con Matlab, podemos usar las siguientes líneas de código:

```
dt=0.08;
s=dt/0.2;
A=[1-3*s s 0 0
   2*s 1-3*s s 0
   0 2*s 1-3*s s
   0 0 2*s 1-3*s]
eig(A); % autovalores de la matriz A
rhoA=max(abs(eig(A))) % radio espectral.
```

PROBLEMA 7.10 *Ecuación de transmisión de calor por conducción en régimen transitorio.*

La conducción del calor unidimensional a través de una varilla se gobierna mediante la ecuación diferencial en derivadas parciales

$$\frac{\partial}{\partial x}\left(k\frac{\partial T}{\partial x}\right) = \rho c_p \frac{\partial T}{\partial t}$$

donde T es la temperatura, y k, ρ y c_p son respectivamente la conductividad térmica, densidad y calor específico de la varilla.

Si estas propiedades son constantes a lo largo de la varilla, poniendo $\alpha = \frac{k}{\rho c_p}$ se puede reescribir la ecuación en la forma:

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

donde α recibe el nombre de difusividad térmica.

Adimensionalizando con $X = x/L$ y $\tau = \alpha t/L^2$, expresiones en las que L representa una longitud característica del problema, X la variable independiente espacial adimensionalizada, y τ el tiempo adimensionalizado, la ecuación diferencial se convierte en:

$$\frac{\partial^2 T}{\partial X^2} = \frac{\partial T}{\partial \tau}$$

Nuestro objetivo será el tratamiento numérico de esta última ecuación en la que ya hemos escalado todas las magnitudes.

La longitud será 1 y el paso espacial $h = 0.25$. Se usará un operador de 3 puntos para la segunda derivada en X y un operador de dos puntos tipo Euler, para la derivada en τ.

1. Queremos utilizar el modelo anterior para estudiar el problema del enfriamiento de una varilla metálica que sale de su molde de fabricación a una temperatura de 1.000 grados, y que empotramos entre dos paredes cuya temperatura se mantendrá constante igual a 0 grados a lo largo del tiempo. La varilla se enfría sólo por sus extremos, estando aislada del exterior en su superficie lateral. Lo lógico es que, en estas condiciones, la temperatura en la barra tienda a cero. Se pide decidir si esto va a ser así según nuestro esquema numérico independientemente del paso de tiempo escogido Δt, o si no fuese de ese modo en general, decidir para qué valores de Δt nuestro esquema se comporta correctamente.

2. Se quiere estimar, con este esquema de integración, con las condiciones de contorno del apartado anterior, y con un paso temporal $\Delta t = 0.01$ unidades, cuánto tiempo tiene que pasar para que la temperatura máxima de la barra no supere un valor genérico T_{max}. Aplicarlo a $T_{max} = 30$ grados.

3. Hacemos implícita la parte espacial. Usamos la i como índice espacial, y la j como índice temporal. El esquema anterior cambia ligeramente para convertirse en:

$$\frac{T_{i+1}^{j+1} - 2T_i^{j+1} + T_{i-1}^{j+1}}{h^2} = \frac{T_i^{j+1} - {}_i^j}{\Delta t}$$

   Partiendo de los valores iniciales en la varilla, se pide dar un algoritmo que permita ir conociendo los sucesivos valores de la temperatura en la barra, teniendo en cuenta los valores que a las variables les hemos asignado en el primer apartado y jugando con el paso temporal como un parámetro más.

4. Si en cada paso de tiempo se tuviera que resolver un sistema lineal, se pregunta si sería convergente un método Gauss-Seidel para resolver dicho sistema independientemente del paso de tiempo elegido.

Solución:
La ecuación sobre la que se va a trabajar es la que corresponde a la transmisión de calor por conducción en régimen transitorio unidimensional, que con abuso escribiremos usando la notación habitual para las variables espaciales y temporales

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

Lo primero es discretizar este esquema usando el operador de tres puntos para la derivada segunda en el espacio:

$$\frac{\partial^2 T}{\partial x^2} \approx \frac{T_{i+1}^j - 2T_i^j + T_{i-1}^j}{h^2}$$

Para la integración en el tiempo usamos un Euler explícito.

$$T(x, t_{j+1}) \approx T(x, t_j) + \Delta t \frac{\partial T(x, t_j)}{\partial t}$$

De aquí deducimos fácilmente que:

$$\frac{\partial T(x, t_j)}{\partial t} = \frac{T(x, t_{j+1}) - T(x, t_j)}{\Delta t}$$

Por tanto, la ecuación discretizada asociada a nuestra ecuación diferencial en derivadas parciales es

$$\frac{T_i^{j+1} - T_i^j}{\Delta t} = \frac{T_{i+1}^j - 2T_i^j + T_{i-1}^j}{h^2}$$

Poniendo $s = \frac{\Delta t}{h^2}$

$$T_i^{j+1} = sT_{i+1}^j + (1 - 2s)T_i^j + sT_{i-1}^j \tag{7.95}$$

Éste es el esquema que nos permitirá ir integrando en el tiempo. El valor de la temperatura en un punto en el instante $j+1$, depende del valor de la temperatura en tres puntos en el instante j. La molécula computacional correspondiente se representa gráficamente en la Figura 7.17.

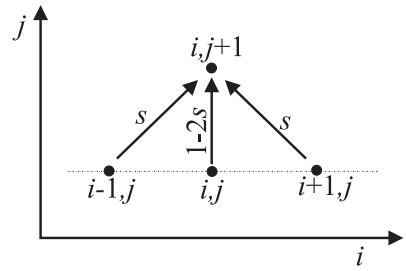


Figura 7.17: Esquema de evolución explícito.

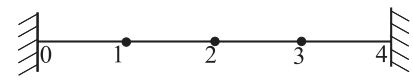


Figura 7.18: Discretización de la barra para $h = 0.25$.

1. En nuestro caso $h = 0.25m$, y por tanto, dado que la barra mide una unidad, tendremos la discretización de la Figura 7.18, con las condiciones de contorno (CC), expresadas en notación referida a índices:

$$\begin{cases} T_0^j = T_4^j = 0 \\ T_1^0 = T_2^0 = T_3^0 = 1000 \end{cases}$$

Si escribimos la ecuación (7.95) en cada uno de los 3 nodos 1,2,3

$$\begin{aligned} T_1^{j+1} &= sT_2^j + (1-2s)T_1^j + sT_0^j \\ T_2^{j+1} &= sT_3^j + (1-2s)T_2^j + sT_1^j \\ T_3^{j+1} &= sT_4^j + (1-2s)T_3^j + sT_2^j \end{aligned}$$

que podemos escribir como un operador lineal, al ser nulos T_0^j y T_4^j

$$\begin{pmatrix} T_1^{j+1} \\ T_2^{j+1} \\ T_3^{j+1} \end{pmatrix} = \begin{pmatrix} 1-2s & s & 0 \\ s & 1-2s & s \\ 0 & s & 1-2s \end{pmatrix} \begin{pmatrix} T_1^j \\ T_2^j \\ T_3^j \end{pmatrix}$$

y vectorialmente

$$\mathbf{T}^{(j+1)} = A\,\mathbf{T}^{(j)} \quad \text{o} \quad \mathbf{T}^{(j+1)} = A^j\,\mathbf{T}^{(0)}$$

con $\mathbf{T}^{(j)} = \left(T_1^j, T_2^j, T_3^j\right)^T$, $\mathbf{T}^{(0)} = (1000, 1000, 1000)^T$ y

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

Si todo funcionase como la física del problema sugiere, a medida que avanzamos en el tiempo, la barra se debería enfriar, y su temperatura debería aproximarse a la de las paredes, es decir, a 0 grados.

Para que eso suceda, o sea, para que la potencia de un operador lineal lleve un vector al nulo, su radio espectral debe ser estrictamente menor que la unidad. Calculemos, por tanto, el radio espectral de la matriz A, calculando primero sus autovalores:

$$\begin{vmatrix} 1-2s-\lambda & s & 0 \\ s & 1-2s-\lambda & s \\ 0 & s & 1-2s-\lambda \end{vmatrix} = 0 \quad \Rightarrow \quad \lambda_1 = 1-2s\,, \quad \lambda_2 = 1-s(2+\sqrt{2})\,, \quad \lambda_3 = 1-s(2-\sqrt{2})$$

$$\rho(A) = \text{máx}\left\{|\lambda_1|, |\lambda_2|, |\lambda_3|\right\}$$

En la Figura 7.19 presentamos estas curvas, y es fácil ver que el máximo en valor absoluto se alcanza al principio con λ_3 y finalmente con λ_2, que es el que fija el valor máximo de s, Figura 7.20. Por tanto,

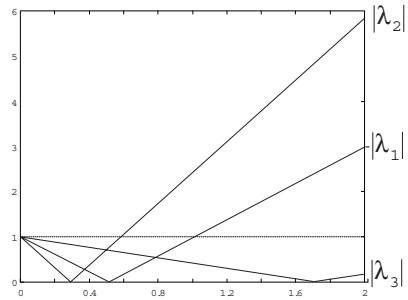


Figura 7.19: Autovalores de A.

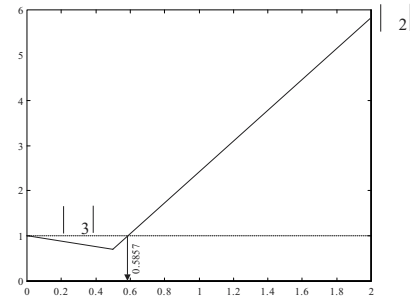


Figura 7.20: Radio espectral de A, $\rho(A)$.

$$\rho(A) < 1 \Longleftrightarrow s < \frac{2}{2+\sqrt{2}} = 0.5857$$

y de aquí obtenemos la mayorante buscada:

$$\Delta t < 0.03661$$

Por ejemplo, para $\Delta t = 0.01$

j	0	1	2	5	10	20	30	38
T_1^j	1000	840	731.2	524.6	319.1	119.2	44.6	20.3
T_2^j	1000	1000	948.8	734.0	451.1	168.6	63	28.7
T_3^j	1000	840	731.2	524.6	319.1	119.2	44.6	20.3

Presentamos ahora una serie de gráficas correspondientes a diferentes pasos de tiempos (Figuras 7.21 a 7.27).

Es importante destacar que en el eje horizontal figura como variable independiente el número de iteraciones, y por tanto, para distintas Δt, los puntos de igual "j"de dos gráficas distintas Δt_1, Δt_2, se corresponden a dos instantes temporales diferentes $j\Delta t_1$ y $j\Delta t_2$. No pueden ser por tanto comparables los valores de la temperatura. También es importante destacar que sólo dibujamos T_1^j y T_2^j pues por simetría $T_1^j = T_1^j \quad \forall j$.

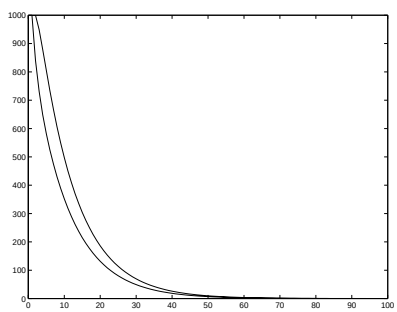


Figura 7.21: $\Delta t = 0.01$.

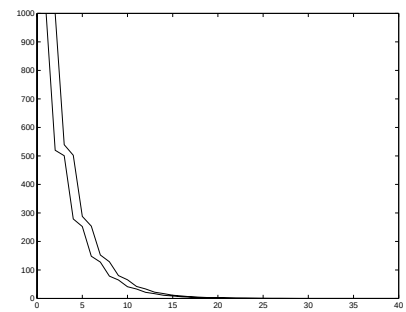


Figura 7.22: $\Delta t = 0.03$.

2.

$$\mathbf{T}^{(j+1)} = A\,\mathbf{T}^{(j)}$$

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

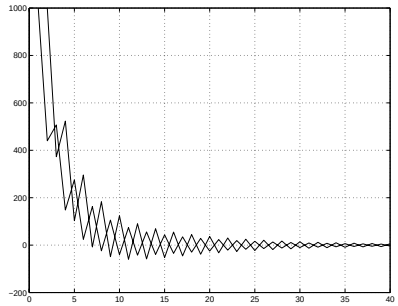


Figura 7.23: $T_{1,j}$, $T_{2,j}$ **frente a iteraciones para** $\Delta t = 0.035$**.**

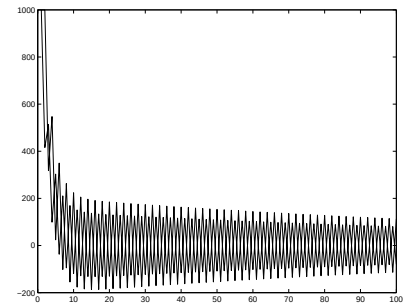


Figura 7.24: $T_{1,j}$, $T_{2,j}$ **frente a iteraciones para** $\Delta t = 0.0365$**.**

y para $h = 0.025$, $\Delta t = 0.01$ y $s = \frac{\Delta t}{h^2} = 0.16$,

$$A = \begin{pmatrix} 0.68 & 0.16 & 0 \\ 0.16 & 0.68 & 0.16 \\ 0 & 0.16 & 0.68 \end{pmatrix}$$

Para acotar el máximo de la temperatura en cada paso de tiempo, lo razonable es utilizar la norma del máximo $\| \quad \|_\infty$.

$$\begin{aligned} \| \mathbf{T}^{(j+1)} \|_\infty &= \| A\, \mathbf{T}^{(j)} \|_\infty \leq \| A \|_\infty \| \mathbf{T}^{(j)} \|_\infty \\ &\leq \| A \|_\infty^2 \| \mathbf{T}^{(j-1)} \|_\infty \leq \cdots \leq \| A \|_\infty^{j+1} \| \mathbf{T}^{(0)} \|_\infty \leq T_{max} \end{aligned}$$

Por tanto,

$$\| A \|_\infty^{j+1} \leq \frac{T_{max}}{\| \mathbf{T}^{(0)} \|_\infty}$$

Ahora bien, este razonamiento no nos sirve en este caso, pues $\| A \|_\infty = 0.16 + 0.68 + 0.16 = 1$ así como sus potencias, luego no podemos obtener ninguna información de esta cota.

Cambiemos de norma. Probemos con la norma 2, ya que al ser A simétrica y real

$$\| A \|_2 = \rho(A)$$

Como $\forall x$, $\| x \|_\infty \leq \| x \|_2$, basta mayorar $\| \mathbf{T}^{(j)} \|_2$

$$\begin{aligned} \| \mathbf{T}^{(j+1)} \|_\infty &\leq \| \mathbf{T}^{(j+1)} \|_2 = \| A\, \mathbf{T}^{(j)} \|_2 \leq \| A \|_2 \| \mathbf{T}^{(j)} \|_2 \\ &\leq \| A \|_2^2 \| \mathbf{T}^{(j-1)} \|_2 \leq \cdots \leq \| A \|_2^{j+1} \| \mathbf{T}^{(0)} \|_2 \leq T_{max} \end{aligned}$$

Por tanto,

$$\| A \|_2^{j+1} \leq \frac{T_{max}}{\| \mathbf{T}^{(0)} \|_2}$$

Tomando logaritmos neperianos y suponiendo que $\| A \|_2 < 1$ que luego corroboramos

$$j \geq \frac{\ln\left(\frac{T_{max}}{\|\mathbf{T}^{(0)}\|_2}\right)}{\ln\left(\| A \|_2\right)} + 1$$

Calculemos los diversos factores de esa expresión

$$\| \mathbf{T}^{(0)} \|_2 = \sqrt{1000^2 + 1000^2 + 1000^2} = 1000\sqrt{3}$$

$$\begin{aligned} \| A \|_2 &= \text{máx}\left\{|1-2s|, |1-s(2+\sqrt{2})|, |1-s(2-\sqrt{2})|\right\} \\ &= \text{máx}\left\{0.68, 0.4537, 0.9063\right\} = 0.9063 \end{aligned}$$

luego

$$j \geq \frac{\ln\left(\frac{30}{1000\sqrt{3}}\right)}{-0.0984134} + 1 = 42.21$$

Por tanto, $j = 43$, es decir, $t = 0.44$ uds de tiempo. Para $j = 43$, tenemos asegurado que $T_{max} \leq 30$. De hecho, si volvemos al ejemplo primero del apartado 1, vemos que eso sucede para $j = 38$.

3. Ahora usamos un Euler implícito, que es un operador que mira hacia adelante para estimar las derivadas.

$$T(x_i, t_{j+1}) \approx T(x_i, t_j) + \Delta t \frac{\partial T(x_i, t_{j+1})}{\partial t}$$

luego

$$T(x_i, t_{j+1}) \approx T(x_i, t_j) + \Delta t \frac{\partial^2 T(x_i, t_{j+1})}{\partial x^2}$$

$$T(x_i, t_{j+1}) \approx T(x_i, t_j) + \frac{\Delta t}{h^2}\left(T_{i+1}^{j+1} - 2T_i^{j+1} + T_{i-1}^{j+1}\right)$$

Por tanto, queda el esquema que aparece en el enunciado:

$$\frac{T_{i+1}^{j+1} - 2T_i^{j+1} + T_{i-1}^{j+1}}{h^2} = \frac{T_i^{j+1} - T_i^j}{\Delta t}$$

es decir,

$$sT_{i+1}^{j+1} - (1+2s)T_i^{j+1} + sT_{i-1}^{j+1} = -T_i^j$$

Escribiendo esta ecuación para los tres nodos x_1, x_2 y x_3, llegamos a que para cada paso de tiempo hay que resolver el siguiente sistema lineal:

$$\begin{pmatrix} 1+2s & -s & 0 \\ -s & 1+2s & -s \\ 0 & -s & 1+2s \end{pmatrix} \begin{pmatrix} T_1^{j+1} \\ T_2^{j+1} \\ T_3^{j+1} \end{pmatrix} = \begin{pmatrix} T_1^j \\ T_2^j \\ T_3^j \end{pmatrix}$$

que se puede poner en la forma siguiente, con una nueva matriz A:

$$A\ \mathbf{T}^{(j+1)} = \mathbf{T}^{(j)}$$

El esquema de avance temporal exige resolver este sistema lineal en cada paso de tiempo, siendo el término independiente el vector de temperaturas correspondientes al paso temporal anterior. En el apartado siguiente resolveremos el sistema lineal correspondiente al primer avance mediante un método iterativo. Aquí no tiene mucho sentido usar un método iterativo al ser tan pocos los nodos, pero en un caso real, con muchos nodos, este tipo de matrices diagonalmente estrictamente dominantes son estupendas para usar métodos iterativos.

4. Para resolver por Gauss-Seidel hacemos una descomposición por bloques unitarios de la matriz A. Lo primero que tenemos que comprobar es que esa descomposición es admisible, o sea, que todos los elementos diagonales son no nulos:

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

$$1 + 2s = 0 \quad \Rightarrow s = -0.5$$

pero eso no es posible, pues s es el cociente de un tiempo y una distancia y ambos son valores positivos. Por tanto la descomposición elegida es admisible, y tiene como matrices de iteración:

$$M = D - L = \begin{pmatrix} 1+2s & 0 & 0 \\ -s & 1+2s & 0 \\ 0 & -s & 1+2s \end{pmatrix} \quad \text{y} \quad N = -U = \begin{pmatrix} 0 & s & 0 \\ 0 & 0 & s \\ 0 & 0 & 0 \end{pmatrix}$$

$$M\,\mathbf{T}^{(j+1)} = N\mathbf{T}^{(j)} + \mathbf{b} \quad \Rightarrow \quad \mathbf{T}^{(j+1)} = M^{-1}N\mathbf{T}^{(j)} + M^{-1}\mathbf{b}$$

La condición de convergencia es que $\rho(B) = \rho\left(M^{-1}N\right) < 1$

$$B = M^{-1}N = \frac{1}{(1+2s)^3}\begin{pmatrix} 0 & s(1+2s)^2 & 0 \\ 0 & s^2(1+2s) & s(1+2s)^2 \\ 0 & s^3 & s^2(1+2s) \end{pmatrix}$$

Calculando los autovalores de esta matriz llegamos a:

$$\rho(B) = \frac{2s^2}{(1+2s)^2}$$

que es < 1 por ser $s > 0$, luego GS converge incondicionalmente independiente de la combinación Δt, h. Podemos elegir, por ejemplo $\Delta t = 0.05$, que fue el caso para el que más rápido divergió el esquema explícito. En este caso $s = 0.8$.

$$B = \begin{pmatrix} 0.0000 & 0.3077 & 0.0000 \\ 0.0000 & 0.0947 & 0.3077 \\ 0.0000 & 0.0291 & 0.0947 \end{pmatrix}$$

Planteamos de este modo el esquema de avance porque conocemos la inversa al haber hecho el análisis de autovalores. En condiciones normales, tendríamos que resolver el sistema triangular inferior de GS en cada iteración.

$$\mathbf{T}_1^{(1)} = B\mathbf{T}_0^{(1)} + \mathbf{T}^{(0)}$$

Si tomamos como estimador inicial $\mathbf{T}_0^{(1)} = \mathbf{T}^{(0)}$

$$\mathbf{T}_1^{(1)} = B\mathbf{T}^{(0)} + M^{-1}\mathbf{T}^{(0)} = \begin{pmatrix} 692.3077 \\ 905.3254 \\ 663.1771 \end{pmatrix}, \quad \mathbf{T}_2^{(1)} = B\mathbf{T}_1^{(1)} + M^{-1}\mathbf{T}^{(0)} = \begin{pmatrix} 663.1771 \\ 792.7243 \\ 628.5306 \end{pmatrix}$$

$$\mathbf{T}_3^{(1)} = B\mathbf{T}_2^{(1)} + M^{-1}\mathbf{T}^{(0)} = \begin{pmatrix} 628.5306 \\ 771.4034 \\ 621.9703 \end{pmatrix}, \quad T\mathbf{T}_4^{(1)} = B\mathbf{T}_3^{(1)} + M^{-1}\mathbf{T}^{(0)} = \begin{pmatrix} 621.9703 \\ 767.3663 \\ 620.7281 \end{pmatrix}$$

$$\mathbf{T}_5^{(1)} = B\mathbf{T}_4^{(1)} + M^{-1}\mathbf{T}^{(0)} = \begin{pmatrix} 620.7281 \\ 766.6019 \\ 620.4929 \end{pmatrix}, \quad \mathbf{T}_6^{(1)} = B\mathbf{T}_5^{(1)} + M^{-1}\mathbf{T}^{(0)} = \begin{pmatrix} 620.4929 \\ 766.4572 \\ 620.4488 \end{pmatrix}$$

que ya está muy cerca del valor correspondiente a $\mathbf{T}^{(1)}$

$$\mathbf{T}^{(1)} = \begin{pmatrix} 620.4380 \\ 766.4234 \\ 620.4380 \end{pmatrix}$$

Cuando se tienen tan pocas incógnitas no tiene sentido usar un método iterativo pero lo normal es que haya muchas incógnitas y si un método iterativo converge, siempre es muchísimo más conveniente que un método directo. Si seguimos avanzando en el tiempo con el esquema implícito y con este paso de tiempo ($\delta t = 0.05$), tendremos, para $t = 0.1$, $t = 0.15$, $t = 0.2$ y $t = 0.5$ respectivamente:

$$\mathbf{T}^{(2)} = \begin{pmatrix} 406.2550 \\ 544.7813 \\ \text{sim} \end{pmatrix}, \quad \mathbf{T}^{(3)} = \begin{pmatrix} 272.2780 \\ 377.0875 \\ \text{sim} \end{pmatrix}, \quad \mathbf{T}^{(4)} = \begin{pmatrix} 184.2327 \\ 258.4076 \\ \text{sim} \end{pmatrix}, \quad \mathbf{T}^{(10)} = \begin{pmatrix} 18.2859 \\ 25.8593 \\ \text{sim} \end{pmatrix}$$

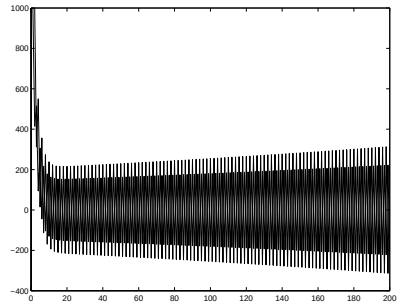


Figura 7.25: $T_{1,j}$, $T_{2,j}$ **frente a iteraciones para** $\Delta t = 0.03665 > \Delta_t$ **límite.**

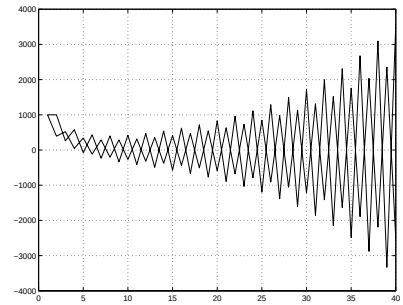


Figura 7.26: $T_{1,j}$, $T_{2,j}$ **frente a iteraciones para** $\Delta t = 0.038 > \Delta_t$ **límite.**

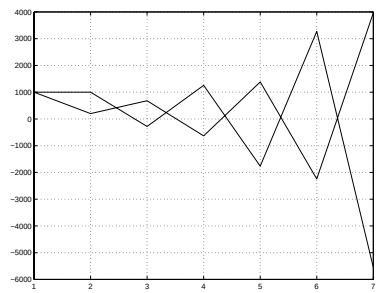


Figura 7.27: $T_{1,j}$, $T_{2,j}$ **frente a iteraciones para** $\Delta t = 0.05 > \Delta_t$ **límite.**

Como vemos, con igual paso de tiempo que el esquema explícito, este esquema implícito nos lleva a temperaturas nulas, que es la realidad física.

PROBLEMA 7.11 *Problema de Dirichlet para la ecuación de Laplace en dominio no rectangular.*

Se considera el problema de Dirichlet para la ecuación de Laplace en el dominio

$$\Omega = \{(x,y) \quad : \quad 0 < x < 4 \quad ; \quad 0 < y < \frac{1}{2}\sqrt{16 - x^2}\}$$

de frontera Γ

$$(P) \quad \begin{cases} u_{xx} + u_{yy} = 0 & (x,y) \in \Omega \\ u(x,0) = x; & 0 \le x \le 4 \\ u(0,y) = 0; & 0 \le y \le 2 \\ u(x,y) = x; & 0 \le x \le 4; \quad 0 \le y \le \frac{1}{2}\sqrt{16 - x^2} \end{cases} \tag{7.96}$$

1. Comprobar que las condiciones de contorno son compatibles.

Se utilizará la malla $\mathcal{M}$ de pasos $h = k = 1$ de la figura y la aproximación (7.64) para discretizar (P).

2. Aproximar la restricción de u a $\mathcal{M}$.

Solución:

1. En los puntos de Γ que pertenecen a dos curvas distintas $\{(0,0),(0,2)$y $(4,0)\}$ los valores de u definidos por las condiciones de contorno son iguales independientemente de la curva sobre la que nos acerquemos al punto en estudio.

2. Los únicos nodos del dominio discretizado $\Omega_{h,k}$ en los que debemos aproximar u son $(1,1)$, $(2,1)$ y $(3,1)$. Como es evidente en la figura ninguno de los tres es interior a $\Omega_{h,k}$, luego en todos ellos tendremos que usar los valores que nos da la condición de Dirichlet sobre la elipse para poder aproximar las derivadas segundas. En el caso de los nodos $(1,1)$ y $(2,1)$ sólo lo usaremos para determinar una aproximación de orden 1 de u_{yy}. En el nodo $(3,1)$ lo necesitaremos para aproximar ambas derivadas segundas.

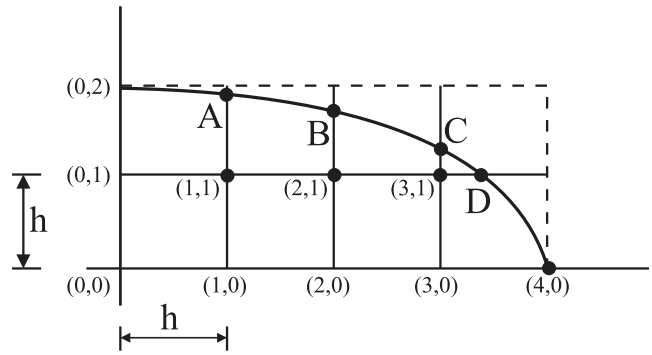


Figura 7.28: Dominio de integración y malla computacional.

- La molécula computacional relativa al nodo $(1,1)$ (Figura 7.29) permite obtener una aproximación de $\Delta u(1,1)$ de primer orden.
  No hay problema para utilizar una aproximación (7.36) de segundo orden de u_{xx} mediante diferencias centradas en $(1,1)$

$$u_{xx}\big|_{1,1} = \frac{u_{2,1} - 2u_{1,1} + u_{0,1}}{h^2} + O(h^2)$$

  Razonando con los tres nodos A, $(1,1)$ y $(1,0)$ tendremos acondicionando adecuadamente (7.61)

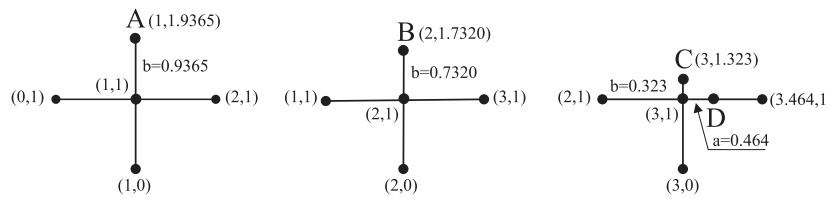


Figura 7.29: Moléculas computacionales.

con $b = \dfrac{\sqrt{15}}{2} - 1$

$$u_{yy}\big|_{1,1} = \frac{u_A - \dfrac{\sqrt{15}}{2}u_{1,1} + \left(\dfrac{\sqrt{15}}{2} - 1\right)u_{1,0}}{\frac{1}{2}\left(\dfrac{\sqrt{15}}{2} - 1\right)\dfrac{\sqrt{15}}{2}k^2} + O(k)$$

de tal modo que

$$\Delta u\big|_{1,1} = \frac{u_{2,1} - 2u_{1,1} + u_{0,1}}{h^2} + O(h^2) + \frac{2}{k^2}\frac{u_A - \dfrac{\sqrt{15}}{2}u_{1,1} + \left(\dfrac{\sqrt{15}}{2} - 1\right)u_{1,0}}{\left(\dfrac{\sqrt{15}}{2} - 1\right)\dfrac{\sqrt{15}}{2}} + O(k)$$

- De un modo similar se trata la molécula relativa al nodo $(2,1)$ excepto que el nodo B tiene coordenadas $(2,\sqrt{3})$ de donde $b=\sqrt{3}-1$ y

$$\Delta u\big|_{2,1}=\frac{u_{3,1}-2u_{2,1}+u_{1,1}}{h^2}+O(h^2)+\frac{2}{k^2}\frac{u_B-\sqrt{3}u_{2,1}+\left(\sqrt{3}-1\right)u_{2,0}}{\left(\sqrt{3}-1\right)\sqrt{3}}+O(k)$$

- El patrón del nodo $(3,1)$ es similar al del nodo M de la Figura 7.5 con $C=\left(3,\dfrac{\sqrt{7}}{2}\right)$ y $D=(2\sqrt{3},1)$ de modo que obtenemos una aproximación de primer orden de Δu en $(3,1)$ entrando en (7.64) con $a=2\sqrt{3}-3$ y $b=\dfrac{\sqrt{7}}{2}-1$

$$\Delta u\big|_{3,1}=\frac{2}{h^2}\frac{u_D-\left(2\sqrt{3}-2\right)u_{3,1}+\left(2\sqrt{3}-3\right)u_{2,1}}{\left(2\sqrt{3}-3\right)\left(2\sqrt{3}-2\right)}+\frac{2}{k^2}\frac{u_C-\dfrac{\sqrt{7}}{2}u_{3,1}+\left(\dfrac{\sqrt{7}}{2}-1\right)u_{3,0}}{\left(\dfrac{\sqrt{7}}{2}-1\right)\dfrac{\sqrt{7}}{2}}+O(h+k)$$

Sustituyendo los valores de los pasos y los de la solución en la frontera, obtenemos el esquema en diferencias asociado a la red elegida y al problema propuesto

$$\begin{cases}
U_{2,1}-2U_{1,1}+2\dfrac{\dfrac{\sqrt{15}}{2}\left(1-U_{1,1}\right)}{\left(\dfrac{\sqrt{15}}{2}-1\right)\dfrac{\sqrt{15}}{2}}=0\\
U_{3,1}-2U_{2,1}+U_{1,1}+2\dfrac{\sqrt{3}\left(2-U_{2,1}\right)}{\left(\sqrt{3}-1\right)\sqrt{3}}=0\\
2\dfrac{2\sqrt{3}-\left(2\sqrt{3}-2\right)U_{3,1}+\left(2\sqrt{3}-3\right)U_{2,1}}{\left(2\sqrt{3}-3\right)\left(2\sqrt{3}-2\right)}+2\dfrac{3-\dfrac{\sqrt{7}}{2}U_{3,1}+\left(\dfrac{\sqrt{7}}{2}-1\right)3}{\left(\dfrac{\sqrt{7}}{2}-1\right)\dfrac{\sqrt{7}}{2}}=0\\
U_{0,0}=0,\ U_{1,0}=1,\ U_{2,0}=2,\ U_{3,0}=3,\ U_{4,0}=4,\ U_{0,1}=U_{0,2}=0\\
\quad U_A=1,\ U_B=2,\ U_C=3,\ U_D=2\sqrt{3}
\end{cases}$$

que podemos escribir matricialmente

$$\begin{pmatrix}-4.3412998 & 1 & 0\\ 1 & -5.8867381 & 1\\ 0 & 1.3660254 & -10.5037360\end{pmatrix}\begin{pmatrix}U_{1,1}\\U_{2,1}\\U_{3,1}\end{pmatrix}=\begin{pmatrix}-2.1356303\\-5.4641016\\-28.7791581\end{pmatrix}$$

Un poco de Matlab nos da como resultado

$$\begin{pmatrix}U_{1,1}\\U_{2,1}\\U_{3,1}\end{pmatrix}=\begin{pmatrix}0.8543922\\1.5735425\\2.9445387\end{pmatrix}$$

PROBLEMA 7.12 *Distribución del potencial en un cable coaxial.*

Se considera el cable coaxial cuya sección recta y dimensiones se representan en la Figura 7.30.
Llamaremos Ω al dominio no conexo interior al cuadrado $[0,6]\times[0,6]$ y exterior al cuadrado $[2,4]\times[2,4]$ y

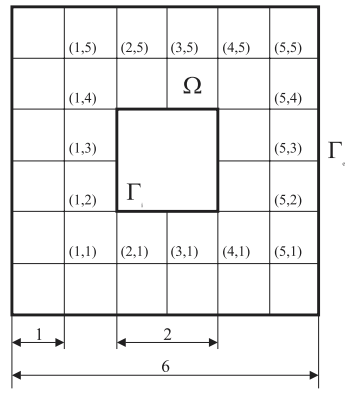


Figura 7.30: Cable coaxial mallado y dimensiones.

Γ_e, Γ_i las fronteras exterior e interior de Ω respectivamente.
La distribución del potencial $V(x,y)$ en el interior del cable Ω verifica el problema de Dirichlet

$$(P) \quad \begin{cases} V_{xx} + V_{yy} = 0 & (x,y) \in \Omega \\ V(x,y) = 300; & (x,y) \in \Gamma_e \\ V(x,y) = 50; & (x,y) \in \Gamma_i \end{cases} \tag{7.97}$$

1. Utilizar una aproximación de 5 puntos (7.40) de la ecuación de Laplace en la malla de la Figura 7.30 para aproximar el problema (P) por el método de diferencias finitas.

2. Escribir el sistema de ecuaciones asociado al esquema en diferencias obtenido.

3. Resolver el sistema obtenido por el método de sobrerrelajación (SOR) con un factor de relajación óptimo, tomando como estimador inicial una media de los valores en el contorno y test de parada $\|\mathbf{r}\| \leq 10^{-6}$. Contabilizar el número de iteraciones hasta la convergencia.

Observaciones

1. El fenómeno físico descrito es simétrico, luego la distribución de potenciales también. Dos puntos distintos del cable simétricos respecto de sus ejes de simetria tendrán el mismo potencial.

Utilizando esta propiedad desde la salida, reducir el sistema de ecuaciones asociado al esquema en diferencias hallado en 2. a un sistema de 5 ecuaciones y 5 incógnitas y resolverlo por un método directo.

2. La propiedad de simetría anterior se puede usar para establecer un test de parada en el proceso iterativo. Se puede estudiar la norma de un vector $\mathbf{V}$ y su simétrico $\mathbf{V}'$ respecto de los ejes de simetría del cable, por ejemplo

$$\mathbf{V} = (V_{11}, V_{21}, V_{31}, V_{41}, V_{12}, V_{13}, V_{14})$$
$$\mathbf{V}' = (V_{52}, V_{53}, V_{54}, V_{25}, V_{35}, V_{45}, V_{55})$$

Dichas normas deben ser iguales cuando se haya producido la convergencia, luego el test de parada podría ser

$$\|\mathbf{V} - \mathbf{V}'\|_\infty < 10^{-6}$$

3. Recordando que el operador de 5 puntos introduce un error del orden de h, los valores del resultado con valores de h distintos son sensiblemente diferentes en los puntos comunes de la malla. Comprobadlo refinando la malla con $h = 1/2$.

Solución:

1. El esquema en diferencias (P_h) que sugiere el enunciado es

$$\mathcal{L}_h V_h \equiv \left\{ \begin{array}{l} V_{i+1,j} + V_{i,j+1} - 4V_{i,j} + V_{i-1,j} + V_{i,j-1} \\ V_{i,j} \quad i = 0,6; \;\; j = 0,1,...,6; \quad i = 1,...,5; \;\; j = 0,6 \\ V_{i,j} \quad i = 2,4; \;\; j = 2,3,4; \quad i = 4; \;\; j = 2,4 \end{array} \right\} = \\ = f_h \equiv \left\{ \begin{array}{l} 0 \\ 300 \\ 50 \end{array} \right\} \tag{7.98}$$

.

2. El sistema de ecuaciones asociado tiene 16 ecuaciones e incógnitas que escribiremos por bloques.

$$\begin{pmatrix} \mathcal{A}_1 & \mathcal{A}_2 & O & O \\ \mathcal{A}_2^T & \mathcal{A}_3 & \mathcal{A}_4 & O \\ O & \mathcal{A}_4^T & \mathcal{A}_3 & \mathcal{A}_5 \\ O & O & \mathcal{A}_5^T & \mathcal{A}_1 \end{pmatrix} \cdot \begin{pmatrix} \mathcal{V}_1 \\ \mathcal{V}_2 \\ \mathcal{V}_3 \\ \mathcal{V}_4 \end{pmatrix} = \begin{pmatrix} \mathcal{B}_1 \\ \mathcal{B}_2 \\ \mathcal{B}_2 \\ \mathcal{B}_2 \end{pmatrix}$$

donde

$$\mathcal{A}_1 = \begin{pmatrix} -4 & 1 & 0 & 0 \\ 1 & -4 & 1 & 0 \\ 0 & 1 & -4 & 1 \\ 0 & 0 & 1 & -4 \end{pmatrix} \; \mathcal{A}_3 = \begin{pmatrix} -4 & 0 & 1 & 0 \\ 0 & -4 & 0 & 1 \\ 1 & 0 & -4 & 0 \\ 0 & 1 & 0 & -4 \end{pmatrix}$$

$$\mathcal{A}_2 = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix} \; \mathcal{A}_4 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix} \; \mathcal{A}_5 = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}$$

$$\mathcal{V}_1 = \begin{pmatrix} V_{11} \\ V_{12} \\ V_{13} \\ V_{14} \end{pmatrix} \; \mathcal{V}_2 = \begin{pmatrix} V_{15} \\ V_{21} \\ V_{25} \\ V_{31} \end{pmatrix} \; \mathcal{V}_3 = \begin{pmatrix} V_{35} \\ V_{41} \\ V_{45} \\ V_{51} \end{pmatrix} \; \mathcal{V}_4 = \begin{pmatrix} V_{52} \\ V_{53} \\ V_{54} \\ V_{55} \end{pmatrix}$$

y

$$\mathcal{B}_1 = \begin{pmatrix} -600 \\ -350 \\ -350 \\ -350 \end{pmatrix} \; \mathcal{B}_2 = \begin{pmatrix} -350 \\ -350 \\ -350 \\ -350 \end{pmatrix}$$

La matriz A simétrica es tridiagonal por bloques. La partición en bloques es admisible (las matrices diagonales $\mathcal{A}_1$ y $\mathcal{A}_3$ son invertibles), se pueden utilizar entonces las técnicas desarrolladas en el problema (2.10).

La estructura del sistema corresponde a la ordenación de los puntos interiores de la malla en la que el nodo (i,j) es anterior al nodo (i',j') ssi el número de dos cifras $\underline{ji}$ es menor que el número $\underline{j'i'}$.

3. Despejando $V_{i,j}$ del esquema (7.98) en los puntos interiores de Ω

$$V_{i,j} = \frac{1}{4}\left(V_{i+1,j} + V_{i,j+1} + V_{i-1,j} + V_{i,j-1}\right)$$

Para aplicar el método de relajación se suma y se resta $V_{i,j}$ al segundo miembro de la ecuación anterior y se escribe la iteración

$$V_{i,j} = V_{i,j} + \frac{\omega}{4}\left(V_{i+1,j} + V_{i,j+1} + V_{i-1,j} + V_{i,j-1} - 4V_{i,j}\right) \tag{7.99}$$

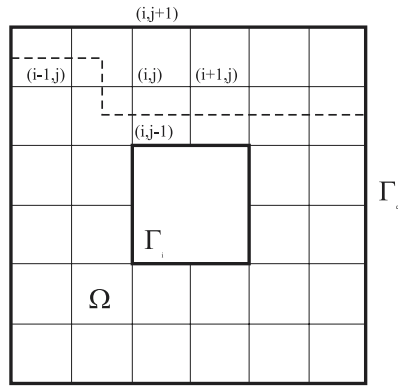


Figura 7.31: Barrido del dominio.

En cada iteración (7.99) "barre" todos los nodos interiores de Ω_h. El barrido correspondiente al orden que hemos definido en el conjunto de nodos, fija el índice j de fila haciendo variar el índice i de columna de 1 a 6 y a continuación pasa a la fila siguiente $j+1$. Se trata de un barrido externo por filas y de un barrido interno por columnas (ver Figura 7.31). Una vez finalizada la iteración k-ésima tenemos en todos los nodos interiores los valores aproximados $V_{i,j}^{(k)}$ que serán actualizados en el paso siguiente $k+1$ en el orden señalado.

Cuando se actualiza el valor del punto (i,j), se observa que todos los nodos que están por debajo de la línea de trazos de la Figura 7.31 están ya actualizados y los que están por encima no.

La molécula de cinco puntos asociada al nodo (i,j) que se representa en dicha figura, está a caballo de ambas iteraciones. Hay nodos con el valor antiguo y otros, ya actualizados, que usamos conforme se van calculando (estrategia Gauss-Seidel).

Precisemos en la igualdad (7.99) esta situación

$$V_{i,j}^{k+1} = V_{i,j}^{k} + \frac{\omega}{4}\left(V_{i+1,j}^{k} + V_{i,j+1}^{k} + V_{i-1,j}^{k+1} + V_{i,j-1}^{k+1} - 4V_{i,j}^{k}\right) \tag{7.100}$$

El segundo sumando de (7.100) es la componente (i,j) del vector residuo $r_{i,j}^{k}$ que realiza la actualización del paso $k+1$. El proceso iterativo continúa hasta que la norma del residuo sea menor que una cierta tolerancia ϵ impuesta de salida. La velocidad con la que $\|\mathbf{r}\|$ converge a 0 se incrementa con una elección adecuada del factor de relajación ω.

La fórmula

$$\omega_{opt} \approx 2 - 2.116\frac{\pi h}{L} + 2.24\left(\frac{\pi h}{L}\right)^2 + O(h^3) \tag{7.101}$$

suministra el valor óptimo de dicho factor en el caso de un cuadrado de lado L con una malla de igual paso h en ambas direcciones. En nuestro caso $\dfrac{h}{L} = N = 6$ y $\omega_{opt} \approx 1.50621$.

Una segunda fórmula es

$$\omega_{opt} \approx \frac{4}{2 + \sqrt{4 - \left(\cos\left(\frac{\pi}{N}\right) + \cos\left(\frac{\pi}{M}\right)\right)^2}} \tag{7.102}$$

que en nuestro caso $N = M = 6$ se reduce a

$$\omega_{opt} = \frac{2}{1 + \operatorname{sen}\left(\frac{\pi}{6}\right)} \approx 1.3333 \tag{7.103}$$

En el paso inicial $k = 0$ se deben suministrar valores a todos los puntos de la malla. Se pueden seguir varias estrategias para definir ese estimador inicial

1. Poner $V_{i,j} = 0$ en todos los puntos interiores.

2. Hacer $V_{i,j}$ igual a una media ponderada de los valores en el contorno.

3. Construir una malla gruesa, con pocos nodos para obtener una solución zafia que defina después por interpolación la estimación inicial de la malla más fina.

En nuestro caso hemos tomado como valor en los nodos interiores la semisuma 175 de los valores en las paredes interior y exterior del cable.

El programa script correspondiente está en la red con el nombre *SORLaplace.m*. En 60 iteraciones obtenemos la solución del problema con la tolerancia impuesta $\|\mathbf{r}\| = 6.5968 \cdot 10^{-7}$.

La solución tal como la ofrece MATLAB es muy gráfica. En ella se reproduce la geometría del fenómeno incluidas las simetrías.

```
300.0000  300.0000  300.0000  300.0000  300.0000  300.0000  300.0000
300.0000  247.9167  195.8333  185.4167  195.8333  247.9167  300.0000
300.0000  195.8333   50.0000   50.0000   50.0000  195.8333  300.0000
300.0000  185.4167   50.0000             50.0000  185.4167  300.0000
300.0000  195.8333   50.0000   50.0000   50.0000  195.8333  300.0000
300.0000  247.9167  195.8333  185.4167  195.8333  247.9167  300.0000
300.0000  300.0000  300.0000  300.0000  300.0000  300.0000  300.0000
```

Observaciones

- Utilizamos las ecuaciones halladas en el apartado 2 relativas a los cinco nodos $(1,1), (2,1), (3,1), (1,2), (1,3)$ teniendo en cuenta que $V_{1,4} = V_{1,2}$ y que $V_{4,1} = V_{2,1}$ se obtiene (Figura 7.32)

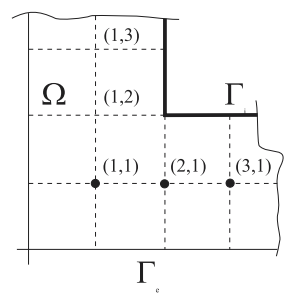


Figura 7.32: Subdominio de cinco nodos.

$$\begin{pmatrix} -4 & 1 & 0 & 1 & 0 \\ 1 & -4 & 1 & 0 & 0 \\ 0 & 2 & -4 & 0 & 0 \\ 1 & 0 & 0 & -4 & 1 \\ 0 & 0 & 0 & 2 & -4 \end{pmatrix} \cdot \begin{pmatrix} V_{11} \\ V_{12} \\ V_{13} \\ V_{21} \\ V_{31} \end{pmatrix} = \begin{pmatrix} -600 \\ -350 \\ -350 \\ -350 \\ -350 \end{pmatrix}$$

cuya solución por cualquier método directo es

V_{11}	V_{12}	V_{13}	V_{21}	V_{31}
247.9167	195.8333	185.4167	195.8333	185.4167

De hecho, se podía haber reducido el número de ecuaciones e incógnitas a tan sólo 3, V_{11}, V_{12} y V_{13}

$$\begin{pmatrix} -4 & 2 & 0 \\ 1 & -4 & 1 \\ 0 & 2 & -4 \end{pmatrix} \cdot \begin{pmatrix} V_{11} \\ V_{12} \\ V_{13} \end{pmatrix} = \begin{pmatrix} -600 \\ -350 \\ -350 \end{pmatrix}$$

cuya solución es de nuevo $V_{11} = 247.9167,\ V_{12} = 195.8333,\ V_{13} = 185.4167$ a partir de cuyos valores se puede reconstruir la distribución de potenciales a todo Ω_h.

- Tomando $h = 1/2$ luego $N = 12$, el programa SOR se para al superar el tope de iteraciones previsto sin alcanzar la tolerancia impuesta. Poniendo max1 = 1.000, se para en 176 iteraciones con un error menor que 10^{-3}. Comparando los valores correspondientes de ambas aproximaciones se observa una gran diferencia que en el peor de los casos relativo al nodo $(2, 1)$ de la malla original es de 6.0256.

PROBLEMA 7.13 *Problema mixto de la ecuación de difusión.*

Se considera el problema mixto para la ecuación del calor

$$(P) \quad \begin{cases} v_t = v_{xx} & 0 < x < 1 \quad t > 0 \\ v(0,t) = t, \quad v_x(1,t) = \text{sen } \pi t & t > 0 \\ v(x,0) = 1 & x \in (0,1) \end{cases}$$

1. (a) Comprobar que la función
$$v_1(x,t) = x\text{sen } \pi t + t$$
satisface las condiciones de contorno del problema (P)

   (b) Efectuar el cambio de función
$$u(x,t) = v(x,t) - v_1(x,t)$$
y comprobar que el problema transformado es
$$(P') \quad \begin{cases} u_t = u_{xx} + 1 + x\pi \cos \pi t & 0 < x < 1 \quad t > 0 \\ u(0,t) = 0, \quad u_x(1,t) = 0 & t > 0 \\ u(x,0) = 1 & x \in (0,1) \end{cases}$$
de condiciones de contorno homogéneas e igual dominio.

2. Se quieren utilizar algunos esquemas en diferencias para resolver de modo aproximado el problema (P') (y el (P)).

   En todos los casos se tomará en la malla computacional el paso espacial $h = \dfrac{1}{4}$ y el paso temporal $k = \dfrac{1}{24}$ y se aproximarán los valores $u(1/4, 1/8)$, $u(3/4, 1/12)$ y $u(1, 1/4)$.

   Se tratará la condición de Neumann en los nodos $(1, nk) \quad n = 0, 1...$ del lado derecho del dominio $(0,1) \times I\!R_+$, de las tres formas siguientes

   (α) Mediante una extrapolación cuadrática de la solución de los puntos interiores de la fila n a la frontera.

   (β) Creando una falsa frontera $(1 + h, nk) \quad n = 0, 1...$ y aproximando $u_x(1, nk)$ mediante una diferencia centrada.

   (γ) Desarrollando $u(1 - h, nk)$ en serie de Taylor en el entorno del punto $(1, nk)$.

   2.1 Utilizar el esquema en diferencias (P_β) basado en el esquema explícito clásico tratando la condición de Neumann alargando el dominio con la frontera ficticia.

2.2 Utilizar el esquema (P_α) basado en el esquema implícito tratando la condición de Neumann mediante la extrapolación cuadrática.

2.3 Utilizar el esquema en diferencias (P_γ) basado en el esquema de Crank-Nicolson haciendo el tratamiento de la condición de Neumann.

Solución:

1. (a) Se trata de un caso particular de un proceso de homogeneización de las condiciones de contorno en los extremos de un intervalo (a, b) cuando una de ellas es de tipo Dirichlet $v(a, t) = f(t)$ y la otra es de tipo Neumann $v_x(b, t) = g(t)$. Como es fácil comprobar la función

$$v(x, t) = (x - a)g(t) + f(t)$$

satisface las condiciones de contorno anteriores. Sustituyendo los datos del problema (P) llegamos a la función v_1 del enunciado.

(b) Al hacer el cambio de función sugerido en el enunciado se llega a (P') sin dificultad

$$\begin{aligned}
&v_t = u_t + \pi x \cos \pi t + 1 \quad \text{y} \quad v_x = u_x + \text{sen } \pi t \quad \Rightarrow \quad v_{xx} = u_{xx} \\
&u(0, t) = v(0, t) - v_1(0, t) = t - t = 0 \\
&v_x(1, t) = u_x(1, t) - \text{sen } \pi t \quad \Rightarrow \quad u_x(1, t) = 0 t > 0 \\
&u(x, 0) = v(x, 0) - v_1(x, 0) = 1 - 1 = 0 \qquad x \in (0, 1)
\end{aligned}$$

2. Comentarios generales a la aplicación de las diferentes discretizaciones.

Hemos llamado como en el resumen teórico $r = \dfrac{k}{h^2}$. Se comprueba que $r = \dfrac{2}{3} < \dfrac{1}{2}$ de donde la consistencia y estabilidad de dichos esquemas.

Para facilitar los razonamientos utilizaremos las variables h, k, N y r que sustituiremos en el momento final del cálculo numérico. Estas variables toman aquí los valores 1/4,1/24, 4 y 2/3 respectivamente.

La condición de contorno en la frontera $x = 0$ es incompatible con la condición inicial. El valor de u en el $(0, 0)$ es 1 por la condición inicial y 0 por la de contorno. El problema propuesto tiene una singularidad en ese punto que se propagará en el tiempo a lo largo de las características. Se sugiere en este caso, para disminuir el error, tomar como valor aproximado en dicho punto la media de los dos valores $U_0^0 = \dfrac{1}{2}$.

Todos los esquemas utilizados son casos particulares de la familia de esquemas $\mathcal{L}_h^\theta$ de (7.77). Si $\theta = 0$ se tiene el esquema explícito del apartado 2.1. Para $\theta = 1$ el esquema implícito del apartado 2.2 y por último para $\theta = 1/2$ el esquema implícito de Crank-Nicolson.

Los distintos tratamientos de la condición de contorno Neumann se incluyen en la sección (7.2.3) del resumen teórico.

2.1 Se añade a la malla $\mathcal{M}_{h,k}$ una falsa hilera de nodos (x_{N+1}, t_n) (ver Figura 7.33) y en cada instante t_n se aproxima $u_x\big|_N^n$ con una diferencia centrada de segundo orden

$$u_x\big|_N^n \approx \frac{U_{N+1}^n - U_{N-1}^n}{2h}$$

como en (P'), $u_x\big|_N^n = 0 \;\; n = 0, 1, ...$

$$0 \approx \frac{U_{N+1}^n - U_{N-1}^n}{2h} \quad \Rightarrow \quad U_{N+1}^n = U_{N-1}^n$$

Una vez aproximados los valores de u en los nodos ficticios se resuelve el problema en el dominio extendido. En él los valores U_N^n son incógnitas que se determinan igual que las demás utilizando

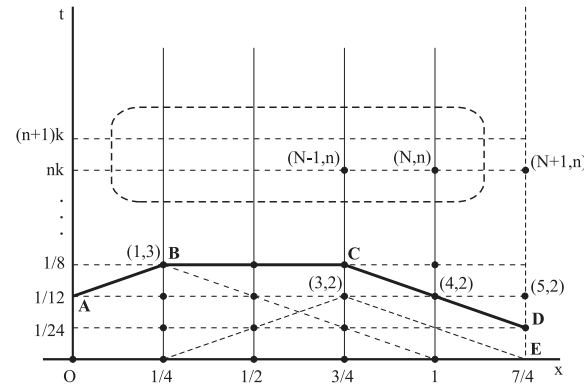


Figura 7.33: Malla computacional.

el esquema.

El esquema en diferencias (P_β) asociado es

$$\mathcal{L}_h^\beta U_h \equiv \left\{\begin{array}{l} U_i^{n+1} - rU_{i+1}^n - (1-2r)U_i^n - rU_{i-1}^n \\ U_0^n \\ U_{N+1}^n - U_{N-1}^n \\ U_i^0 \end{array}\right\} =$$

$$= f_h \equiv \left\{\begin{array}{l} kf_i^n = f(ih, nk) = 1 + \pi ih\cos\pi nk \\ 0 \\ 0 \\ u_0(ih) = 1 \end{array}\right\} \quad i = 0, 1, ..., N+1 \quad n = 0, 1, ... \tag{7.104}$$

obsérvese que i varía de 0 a $N+1$.

En el primer paso de aplicación del esquema (P_β) tenemos

$$\begin{array}{lccccccc} U_1^1 = & rU_0^0 + & (1-2r)U_1^0 + & rU_2^0 + & & & & +kf_1^0 \\ U_2^1 = & & rU_1^0 + & (1-2r)U_2^0 + & rU_3^0 & & & +kf_2^0 \\ U_3^1 = & & & rU_2^0 + & (1-2r)U_3^0 + & rU_4^0 + & & +kf_3^0 \\ U_4^1 = & & & & rU_3^0 + & (1-2r)U_4^0 + & rU_5^0 & +kf_4^0 \end{array}$$

sustituyendo los parámetros del problema y poniendo $U_0^0 = 1/2$ y $U_5^0 = U_3^0$ tenemos matricialmente

$$\begin{pmatrix} U_1^1 \\ U_2^1 \\ U_3^1 \\ U_4^1 \end{pmatrix} = \begin{pmatrix} -1/3 & 2/3 & 0 & 0 \\ 2/3 & -1/3 & 2/3 & 0 \\ 0 & 2/3 & -1/3 & 2/3 \\ 0 & 0 & 4/3 & -1/3 \end{pmatrix} \cdot \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} + \begin{pmatrix} 1/3 + 1/24(1+\pi/4) \\ 1/24(1+2\pi/4) \\ 1/24(1+3\pi/4) \\ 1/24(1+4\pi/4) \end{pmatrix}$$

Llamando

$$(U) = \begin{pmatrix} U_1 \\ U_2 \\ U_3 \\ U_4 \end{pmatrix}; \qquad A = \begin{pmatrix} -1/3 & 2/3 & 0 & 0 \\ 2/3 & -1/3 & 2/3 & 0 \\ 0 & 2/3 & -1/3 & 2/3 \\ 0 & 0 & 4/3 & -1/3 \end{pmatrix}; \qquad (f) = \begin{pmatrix} f_1 \\ f_2 \\ f_3 \\ f_4 \end{pmatrix}$$

podemos escribir el paso del instante t_n al instante t_{n+1}

$$(U)^{n+1} = A(U)^n + k(f)^n$$

El siguiente programa script en el que se ha huido de sofisticaciones técnicas, permite calcular $(U)^1$, $(U)^2$ y $(U)^3$ alguna de cuyas componentes pide el enunciado.

```
A=[-1/3  2/3   0    0;
    2/3 -1/3  2/3   0;
    0    2/3 -1/3  2/3;
    0    0    4/3 -1/3]
U0=[ 1 1 1 1]';
B0=[(2/3)*(1/2)+(1/24)*(1+(pi/4));
     (1/24)*(1+2*(pi/4));
     (1/24)*(1+3*(pi/4));
     (1/24)*(1+4*(pi/4))]
 U1=(A*U0)+B0
 B1=[(1/24)*(1+(pi/4)*cos(pi/24));
     (1/24)*(1+2*(pi/4)*cos(pi/24));
     (1/24)*(1+3*(pi/4)*cos(pi/24));
     (1/24)*(1+4*(pi/4)*cos(pi/24))]
 U2=A*U1 +B1
 B2=[(1/24)*(1+(pi/4)*cos(2*(pi/24)));
     (1/24)*(1+2*(pi/4)*cos(2*(pi/24)));
     (1/24)*(1+3*(pi/4)*cos(2*(pi/24)));
     (1/24)*(1+4*(pi/4)*cos(2*(pi/24)))]
 U3=A*U2 + B2
```

Se tiene

$$(U)^1 = \begin{pmatrix} 0.7411 \\ 1.1071 \\ 1.1398 \\ 1.1726 \end{pmatrix}; \qquad (U)^2 = \begin{pmatrix} 0.5652 \\ 0.9915 \\ 1.2788 \\ 1.3004 \end{pmatrix}; \qquad (U)^3 = \begin{pmatrix} 0.5459 \\ 1.0037 \\ 1.2381 \\ 1.4398 \end{pmatrix}$$

de modo que $U_3^2 = 1.2788$, $U_1^3 = 0.5459$ y $U_4^3 = 1.4398$.

2.2 En el problema 7.6 llegamos a la siguiente aproximación

$$U_N^n = \frac{1}{3}\left(4U_{N-1}^n - U_{N-2}^n - 2hu_x\big|_N^n\right)$$

que aquí se escribe

$$U_N^n = \frac{1}{3}\left(4U_{N-1}^n - U_{N-2}^n\right)$$

El esquema en diferencias (P_α) que sugiere el enunciado en este apartado es entonces

$$\begin{aligned} \mathcal{L}_h^\alpha U_h \equiv \begin{Bmatrix} U_i^n + rU_{i+1}^{n+1} - (1+2r)U_i^{n+1} + rU_{i-1}^{n+1} \\ U_0^n \\ U_N^n - \frac{1}{3}\left(4U_{N-1}^n - U_{N-2}^n\right) \\ U_i^0 \end{Bmatrix} = \\ = f_h \equiv \begin{Bmatrix} kf_i^n = 1 + \pi ih\cos\pi nk \\ 0 \\ 0 \\ u_0(ih) = 1 \end{Bmatrix} \quad \text{con} \quad i = 0,1,...,N \qquad n = 0,1,... \end{aligned} \tag{7.105}$$

En el primer paso de aplicación del esquema (P_β) tenemos

$$\begin{array}{rrrrcll} -rU_0^1+ & (1+2r)U_1^1- & rU_2^1 & & = & U_1^0 & +kf_1^0 \\ & -rU_1^1+ & (1+2r)U_2^1- & rU_3^1 & = & U_2^0 & +kf_2^0 \\ & & -rU_2^1+ & (1+2r)U_3^1- \quad rU_4^1 & = & U_3^0 & +kf_3^0 \end{array}$$

sustituyendo los parámetros del problema y poniendo $U_0^1 = 0$, $U_i^0 = 1$, $i = 1, 2, 3$ y $U_4^1 = 1/3(4U_3^1 - U_2^1)$ tenemos matricialmente

$$\begin{pmatrix} 7/3 & -2/3 & 0 \\ -2/3 & 7/3 & -2/3 \\ 0 & -4/9 & 13/9 \end{pmatrix} \cdot \begin{pmatrix} U_1^1 \\ U_2^1 \\ U_3^1 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} + \begin{pmatrix} 1/24(1+\pi/4) \\ 1/24(1+2\pi/4) \\ 1/24(1+3\pi/4) \end{pmatrix}$$

Llamando

$$(U) = \begin{pmatrix} U_1 \\ U_2 \\ U_3 \end{pmatrix}; \qquad A = \begin{pmatrix} 7/3 & -2/3 & 0 \\ -2/3 & 7/3 & -2/3 \\ 0 & -4/9 & 13/9 \end{pmatrix}; \qquad (f) = \begin{pmatrix} f_1 \\ f_2 \\ f_3 \end{pmatrix}$$

podemos escribir el paso del instante t_n al instante t_{n+1}

$$A(U)^{n+1} = (U)^n + k(f)^n \quad \Rightarrow \quad (U)^{n+1} = A^{-1}\left((U)^n + k(f)^n\right)$$

El siguiente programa script permite calcular $(U)^1$, $(U)^2$ y $(U)^3$.

```
A=[7/3  -2/3  0;
  -2/3  7/3  -2/3;
    0   -4/9  13/9]
A=inv(A);
B0=[ 1+(1/24)*(1+(pi/4));
     1+(1/24)*(1+2*(pi/4));
     1+(1/24)*(1+3*(pi/4))]
U1=(A*B0)
B1=[(1/24)*(1+(pi/4)*cos(pi/24));
     (1/24)*(1+2*(pi/4)*cos(pi/24));
     (1/24)*(1+3*(pi/4)*cos(pi/24))]
U2=A*U1 + A*B1
B2=[(1/24)*(1+(pi/4)*cos(2*(pi/24)));
     (1/24)*(1+2*(pi/4)*cos(2*(pi/24)));
     (1/24)*(1+3*(pi/4)*cos(2*(pi/24)))]
U3=A*C2 + A*B2
U43=(1/3)*(4*U3(3)-U3(2))
```

Se tiene

$$(U)^1 = \begin{pmatrix} 0.7465 \\ 1.0013 \\ 1.0972 \end{pmatrix}; \qquad (U)^2 = \begin{pmatrix} 0.6337 \\ 0.9872 \\ 1.1596 \end{pmatrix}; \qquad (U)^3 = \begin{pmatrix} 0.5820 \\ 0.9765 \\ 1.1977 \end{pmatrix}$$

de modo que $U_3^2 = 1.1596$, $U_1^3 = 0.5820$ y $U_4^3 = 1.2715$.

2.3 Utilizando una diferencia regresiva se tiene

$$U_{N-1}^n = U_N^n + u_x\big|_N^n h$$

es decir,

$$U_N^n = U_{N-1}^n$$

equivalente a efectuar una extrapolación lineal.

El esquema en diferencias (P_γ) asociado es

$$\mathcal{L}_h^\gamma U_h \equiv \left\{\begin{array}{l} 2(1+2r)U_{i,j}^{n+1} - r\left(U_{i+1,j}^{n+1} + U_{i-1,j}^{n+1} + U_{i,j+1}^{n+1} + U_{i,j-1}^{n+1}\right) = \\ \quad = 2(1-2r)U_{i,j}^{n} + r\left(U_{i+1,j}^{n} + U_{i-1,j}^{n} + U_{i,j+1}^{n} + U_{i,j-1}^{n}\right) \\ U_0^n \\ U_N^n - U_{N-1}^n \\ U_i^0 \end{array}\right\} = \tag{7.106}$$

$$= f_h \equiv \left\{\begin{array}{l} kf_i^n = 1 + \pi ih\cos\pi nk \\ 0 \\ 0 \\ u_0(ih) = 1 \end{array}\right\} \quad \text{con} \quad i = 0,1,...,N \qquad n = 0,1,...$$

En el primer paso de aplicación del esquema (P_γ) tenemos

$$\begin{array}{l} -rU_0^1 + 2(1+r)U_1^1 - rU_2^1 = rU_0^0 + 2(1-r)U_1^0 + rU_2^0 + k(f_1^1 + f_1^0) \\ -rU_1^1 + (1+r)U_2^1 - rU_3^1 = \quad rU_1^0 + 2(1-r)U_2^0 + rU_0^3 + k(f_2^1 + f_2^0) \\ -rU_2^1 + 2(1+r)U_3^1 - rU_4^1 = rU_2^0 + 2(1-r)U_3^0 + rU_0^4 + k(f_3^1 + f_3^0) \end{array} \tag{7.107}$$

sustituyendo los parámetros del problema y poniendo $U_0^1 = 0$, $U_0^0 = 1/2$, $U_i^0 = 1$, $i = 1,2,3$, $U_4^1 = U_3^1$ y $U_4^0 = U_3^0$ tenemos matricialmente

$$\begin{pmatrix} 10/3 & -2/3 & 0 \\ -2/3 & 10/3 & -2/3 \\ 0 & -2/3 & 8/3 \end{pmatrix} \cdot \begin{pmatrix} U_1^1 \\ U_2^1 \\ U_3^1 \end{pmatrix} = \begin{pmatrix} 2/3 & 2/3 & 0 \\ 2/3 & 2/3 & 2/3 \\ 0 & 2/3 & 4/3 \end{pmatrix} + \begin{pmatrix} 4/3 + 1/24(1+\pi/4)(1+\cos\pi/24) \\ 1 + 1/24(1+2\pi/4)(1+\cos\pi/24) \\ 1 + 1/24(1+3\pi/4)(1+\cos\pi/24) \end{pmatrix}$$

Llamando

$$(U) = \begin{pmatrix} U_1 \\ U_2 \\ U_3 \end{pmatrix};\ A = \begin{pmatrix} 7/3 & -2/3 & 0 \\ -2/3 & 7/3 & -2/3 \\ 0 & -4/9 & 13/9 \end{pmatrix};\ D = \begin{pmatrix} 2/3 & 2/3 & 0 \\ 2/3 & 2/3 & 2/3 \\ 0 & 2/3 & 4/3 \end{pmatrix};\ (f) = \begin{pmatrix} f_1 \\ f_2 \\ f_3 \end{pmatrix}$$

podemos escribir el paso del instante t_n al instante t_{n+1}

$$A(U)^{n+1} = D(U)^n + k\left((f)^{n+1} + (f)^n\right) \quad \Rightarrow \quad (U)^{n+1} = A^{-1}\left(D(U)^n + k\left((f)^{n+1} + (f)^n\right)\right)$$

Se deja como ejercicio escribir el programa script similar a los que ya hemos presentado que permita calcular $(U)^1$, $(U)^2$ y $(U)^3$. Se tiene

$$(U)^1 = \begin{pmatrix} 0.7519 \\ 1.0371 \\ 1.1137 \end{pmatrix}; \qquad (U)^2 = \begin{pmatrix} 0.6009 \\ 0.9972 \\ 1.1681 \end{pmatrix}; \qquad (U)^3 = \begin{pmatrix} 0.5536 \\ 0.9592 \\ 1.1722 \end{pmatrix}$$

de modo que $U_3^2 = 1.1681$, $U_1^3 = 0.5536$ y $U_4^3 = 1.1722$.

Comentarios finales

Éste es el resultado obtenido con el esquema que de salida es el más preciso, Crank-Nicolson con extrapolación cuadrática

$$(U)^1 = \begin{pmatrix} 0.7522 \\ 1.0386 \\ 1.1209 \end{pmatrix}; \qquad (U)^2 = \begin{pmatrix} 0.6029 \\ 1.0052 \\ 1.1969 \end{pmatrix}; \qquad (U)^3 = \begin{pmatrix} 0.5598 \\ 0.9798 \\ 1.2307 \end{pmatrix}$$

de modo que $U_3^2 = 1.1969$, $U_1^3 = 0.5598$ y $U_4^3 = 1.3143$.

Los valores son similares a los de 2.3 con mayor ajuste en la frontera.

APÉNDICE A

Tutorial de Matlab

Presentamos un tutorial de Matlab, una herramienta potentísima, casi estándar para cálculos en muchas ramas de la Ingeniería, y de uso razonablemente simple. Haremos una descripción de los elementos básicos de Matlab y remitiremos al estudiante interesado en mejorar sus conocimientos del mismo a cualquier edición del manual de referencia del programa [21]. También es interesante el libro sobre Matlab de Higham y Higham [16] y una buena referencia en castellano con ejemplos procedentes de problemas de Cálculo Numérico es el de Quintela [23].

A.1. Conceptos básicos

Para arrancar Matlab, se procede como con cuaquier programa Windows, o sea, Inicio, Programas, Matlab o Student Matlab caso de que utilicemos la versión educacional. Una vez arrancado aparece el cursor con el símbolo (>>) o (EDU >>), indicando que se pueden introducir órdenes. De hecho, en este tutorial, cuando aparezca este símbolo, se tiene que introducir por teclado la orden que aparece escrita a la derecha del mismo.

La utilización más básica de Matlab es como calculadora [1]. Así, por ejemplo, para calcular $\cos(5) \cdot 2^{7.3}$, se debe introducir [2]:

```
>>cos(5)*2^7.3
ans =
   44.7013
```

Matlab mantiene en memoria el último resultado. Caso de que ese cálculo no se asigne a ninguna variable, lo hace a una variable por defecto de nombre *ans*. Si queremos referirnos a ese resultado, lo haremos a través de la variable *ans*, y si no se asigna ese nuevo cálculo a ninguna variable, volverá a ser asignado a *ans*.

```
>>log(ans)
ans =
    3.8000
```

En este momento cabría preguntarnos si tratamos con un logaritmo decimal o con uno neperiano (natural). Para saberlo, pedimos ayuda acerca del comando *log* utilizando:

```
>>help log
 LOG    Natural logarithm.
    LOG(X) is the natural logarithm of the elements of X.
    Complex results are produced if X is not positive.

    See also LOG2, LOG10, EXP, LOGM.
```

[1] Funcionando de este modo, es similar a una calculadora programable, aunque bastante más versátil.

[2] Los argumentos de las funciones trigonométricas siempre están en radianes.

Aunque en la explicación de las órdenes, los comandos aparezcan en mayúsculas, se deben usar en minúsculas.

Por defecto, los resultados aparecen con 4 cifras decimales. Si se necesitara más precisión en los resultados, se puede utilizar la orden *format long* repitiendo los cálculos:

```
>>format long
```

Para recuperar una orden y ejecutarla otra vez o modificarla se usan la flechas arriba y abajo del cursor ⇑, ⇓. Presionemos ⇑ hasta recuperar la orden:

```
>>cos(5)*2^7.3
ans =
  44.70132670851334
```

Ejercicio A.1.1 *Realizar la siguiente operación:* $2.7^{2.1} + \log_{10} 108.2$.

Ejercicio A.1.2 *Realizar la siguiente operación:* $e^{2.7^{2.1}+\log_{10} 108.2}$.

Si necesitamos referirnos a determinados cálculos, se asignan a variables y así se pueden recuperar después. Por ejemplo, podemos recuperar con ⇑ la orden $\cos(5) \cdot 2^{7.3}$ y asignar su valor a la variable x editando dicha orden. Luego podremos utilizarla para otros cálculos.

```
>>x=cos(5)*2^7.3
x =
  44.70132670851334
>>y=log(x)
y =
   3.80000318145901
```

Ejercicio A.1.3 *Realizar la siguiente operación:* $2.7^{2.1} + \log_{10} 108.2$ *y asignarla a la variable* x.

Ejercicio A.1.4 *Realizar la siguiente operación:* $e^{2.7^{2.1}+\log_{10} 108.2}$ *y asignarla a la variable* t.

Si queremos saber cuánto vale una variable, no tenemos más que escribirla en la línea de comandos y pulsar *Enter*.

```
>>y
y =
   3.80000318145901
```

Como es muy fácil recuperar órdenes previas podemos utilizar esta idea para simular los términos de una sucesión recurrente. Por ejemplo, $x_{n+1} = cos(x_n)$

```
>>x=0.2
x =
   0.20000000000000
>>x=cos(x)
x =
   0.98006657784124
>>x=cos(x)
x =
   0.55696725280964
>>x=cos(x)
x =
   0.84886216565827
>>x=cos(x)
x =
```

```
   0.66083755111662
>>x=cos(x)
x =
   0.78947843776687
>>x=cos(x)
x =
   0.70421571334199
```

Ejercicio A.1.5 *Repetir la operación anterior hasta que se estabilice el cuarto decimal de x de un paso al siguiente.*

Ejercicio A.1.6 *Cambiar el formato para que otra vez se vean sólo cuatro decimales.*

Ejercicio A.1.7 *Empezando por $x = 100$ repetir la operación*

$$x = x - \frac{x^2 - 81}{2x}$$

hasta que se converja en el cuarto decimal. ¿Qué relación hay entre el último x y 81?

Ejercicio A.1.8 *Definir A como en vuestro documento de identidad o pasaporte. Empezando por $x = 100$ repetir la operación*

$$x = x - \frac{x^2 - A}{2x}$$

hasta que se converja en el cuarto decimal. ¿A qué ha convergido la sucesión?[3]

Es interesante comentar de este ejercicio que Matlab distingue entre letras mayúsculas y minúsculas en los nombres de las variables.

A veces es bueno apagar y encender la *calculadora* para borrar todo y empezar de nuevo. Esto se hace con la orden *clear*. Hay que tener cuidado al utilizarla, ya que borra todas las variables que estén en la memoria sin pedir confirmación.

```
>>clear
>>x
??? Undefined function or variable 'x'.
```

Ejercicio A.1.9 *Preguntar el valor de A igual que acabamos de preguntar x. ¿Tiene sentido el resultado?*

A.2. Manejo de vectores

Para crear y almacenar en memoria un vector v que tenga como componentes $v_1 = 0$, $v_2 = 2$, $v_3 = 4$, $v_4 = 6$ y $v_5 = 8$ podemos hacerlo componente a componente:

```
>>v(1)=0
v =
     0
>>v(2)=2
v =
     0     2
>>v(3)=4
v =
     0     2     4
>>v(4)=6
v =
     0     2     4     6
>>v(5)=8
v =
     0     2     4     6     8
```

[3]Las calculadoras obtienen la raíz cuadrada de un número mediante esta sucesión.

Se puede también definir este vector especificando su primer elemento, un incremento y el último elemento. Matlab rellenará paso a paso sus componentes. Así, podemos definir igualmente el vector v como una secuencia que empieza en 0, avanza de 2 en 2 y que termina en el 8:

```
>> v = [0:2:8]
v =
     0     2     4     6     8
>> v
v =
     0     2     4     6     8
```

Si ponemos ; al final de una línea de comandos, cuando pulsemos la tecla Enter para ejecutarla, se ejecutará pero no mostrará el resultado en pantalla (se anula el eco en pantalla). Esto es muy útil algunas veces:

```
>> v = [0:2:8];
>> v
v =
     0     2     4     6     8
```

Podemos construir el vector v editando directamente entre los corchetes las componentes del vector v:

```
>>v = [0 2 4 6 8];
>> v
v =
     0     2     4     6     8
```

Es fácil acceder al contenido de una posición del vector, por ejemplo la primera.

```
>> v(1)
ans =
     0
```

O modificarla:

```
>> v(1)=-3;
>> v
v =
    -3     2     4     6     8
```

O hacer operaciones entre componentes, $v_2 \cdot v_5^3$:

```
>> v(2)*v(5)^3
ans =
        1024
```

Ejercicio A.2.1 *Calcular la suma de los elementos de v, elemento a elemento.*

Para trasponer un vector o una matriz se usa el apóstrofo, que es el acento que está en la misma tecla que el signo de interrogación "?".

```
>> v'
ans =
    -3
     2
     4
     6
     8
```

Como hemos comentado, para recuperar una orden y ejecutarla otra vez o modificarla se usan la flechas arriba y abajo del cursor ⇑, ⇓. Presionemos ⇑ hasta recuperar la orden:

```
>> v(1)=-3;
```

Modifiquémosla para dejar el valor original

```
>> v(1)=0;
```

Al definir ese vector v de 5 componentes, en realidad lo que definimos es una matriz fila de cinco columnas, o sea, un matriz de 1×5. Esto se comprueba preguntando el tamaño de v con la sentencia *size*:

```
>>size(v)
ans =
     1     5
```

que nos indica que v tiene una fila y 5 columnas.

Ejercicio A.2.2 *Definir un nuevo vector que sea el traspuesto de v y aplicar a ese vector el comando size. ¿Es coherente el resultado?*

A.3. Introducción al tratamiento de matrices

Haremos una introducción a la definición y manipulación de matrices. Se supone que se ha seguido la sección anterior y que se dispone de los conocimientos básicos sobre la definición y manipulación de vectores usando Matlab. La definición de una matriz es muy similar a la de un vector. Para definir una matriz, se puede hacer dando sus filas separadas por un punto y coma (¡no olvidarse poner los espacios en blanco!):

```
>> A = [ 1 2 3; 3 4 5; 6 7 8]
A =
     1     2     3
     3     4     5
     6     7     8
```

o definirla directamente fila a fila, que es más intuitivo:

```
>> A = [ 1 2 3
         3 4 5
         6 7 8]
A =
     1     2     3
     3     4     5
     6     7     8
```

Se puede modificar alguno de los elementos de la matriz A, accediendo a cualquiera de sus posiciones, por ejemplo:

```
>> A(2,2)=-9
A =
     1     2     3
     3    -9     5
     6     7     8
```

Dejemos su valor original:

```
>> A(2,2)=4;
```

De igual modo, se la puede considerar como una fila de vectores columna:

```
>> B = [ [1 2 3]' [2 4 7]' [3 5 8]']
B =
     1     2     3
     2     4     5
     3     7     8
```

(Otra vez, es importante colocar los espacios en blanco.)

Ejercicio A.3.1 *Sumar los elementos diagonales de la matriz A, refiriéndonos a ellos, elemento a elemento.*

Podemos sumar o restar matrices para tener otras matrices.

```
>>C=A+B
C =
     2     4     6
     5     8    10
     9    14    16
```

Ejercicio A.3.2 *Definir la matriz* $D = 2B - A$.

También podemos multiplicarlas.

```
>>C=A*B
C =
    14    31    37
    26    57    69
    44    96   117
```

Ejercicio A.3.3 *Definir la matriz* $D = B - A \cdot B$.

Ejercicio A.3.4 *Definir la matriz* $C = AA^t$.

Podemos definir algunos tipos especiales de matrices, como por ejemplo una matriz de 3×3 que tenga todos sus elementos nulos.

```
>>I=zeros(3)
I =
     0     0     0
     0     0     0
     0     0     0
```

Podemos modificar sus elementos diagonales para tener la matriz identidad.

```
>>I(1,1)=1;
>>I(2,2)=1;
>>I(3,3)=1
I =
     1     0     0
     0     1     0
     0     0     1
```

Ejercicio A.3.5 *Definir la matriz* $D = B - B \cdot A$ *como* $D = B(I - A)$.

Otra forma de definir la matriz identidad es a través de la función *diag*, que recibe un vector que convierte en diagonal de una matriz cuyos otros elementos son nulos.

```
>>J=diag([1 1 1])
J =
     1     0     0
     0     1     0
     0     0     1
```

Ejercicio A.3.6 *Definir una matriz D diagonal cuyos elementos sean* -2, 1, 0.2 *y* -0.7.

Ejercicio A.3.7 *Pedir ayuda de la función eye, y definir la matriz diagonal de* 10×10.

A.3.1. Definición de submatrices

La definición de "subvectores" o submatrices es muy fácil. Si v es

```
>> v = [0:2:8]
v =
     0     2     4     6     8
```

Podemos definir un vector e cuyas componentes sean las tres primeras componentes del vector v poniendo

```
>> e=v(1:1:3)
e =
     0     2     4
```

donde el primer uno indica que vamos a tomar el primer elemento de v. El segundo número es el incremento de índices dentro de v y el último número marca el elemento final. Esta orden es equivalente a la siguiente

```
>> e=v(1:3)
e =
     0     2     4
```

ya que cuando el incremento es la unidad, se puede suprimir.

Ejercicio A.3.8 *Adivinar cuál va a ser el resultado de las dos órdenes siguientes*

```
>> e=v(2:2:5)
>> e=v(1:3:5)
```

Como comentamos al principio, la notación usada por Matlab sigue en lo posible la notación estándar de Álgebra Lineal que se asume conocida. Es muy sencillo multiplicar matrices y vectores, teniendo cuidado de que las dimensiones sean las adecuadas.

```
>> A*v(1:3)
??? Error using == *
Inner matrix dimensions must agree.
>> A*v(1:3)'
ans =
    16
    28
    46
```

Es importante acostumbrase a ver ese mensaje de error. Una vez que se empieza a trabajar con vectores y matrices, es sencillo olvidar los tamaños de los objetos que se han ido creando.

Ejercicio A.3.9 *Utilizando el comando size, razona sobre los problemas en lo que se refiere a dimensiones en la multiplicación anterior.*

Se pueden extraer columnas o filas de una matriz. Si queremos, por ejemplo, que C sea la tercera fila de la matriz A:

```
>> C=A(3,:)
C =
     6     7     8
```

O que C sea la segunda columna de la matriz B

```
>>C=B(:,2)
C =
     2
     4
     7
```

O bien que D sea la submatriz cuadrada de orden dos inferior derecha de la matriz A.

```
>> D=A(2:3,2:3)
D =
     4     5
     7     8
```

Una vez que se es capaz de crear y manipular una matriz, se pueden realizar muchas operaciones estándar. Por ejemplo, calcular su inversa. Hay que tener cuidado y no olvidar que las operaciones son cálculos numéricos realizados por ordenador. En el ejemplo, A no es una matriz regular, y sin embargo Matlab devolverá su inversa, pues los errores de redondeo durante su cálculo convierten en *invertible* a dicha matriz.

```
>> inv(A)
Warning: Matrix is close to singular or badly scaled.
         Results may be inaccurate. RCOND = 4.565062e-18
ans =
   1.0e+15 *
   -2.7022    4.5036   -1.8014
    5.4043   -9.0072    3.6029
   -2.7022    4.5036   -1.8014
```

Con la matriz B sí que es posible calcular su inversa:

```
>>inv(B)
ans =
   -3.0000    5.0000   -2.0000
   -1.0000   -1.0000    1.0000
    2.0000   -1.0000         0
```

Ejercicio A.3.10 *Definir una matriz de nombre B*1 *como la inversa de B. Multiplicar B por B*1 *y razonar la coherencia del resultado.*

Hay que comentar que Matlab distingue entre mayúsculas y minúsculas. Este puede ser el origen de algunas confusiones si se manejan algoritmos complejos.

```
>> inv(a)
??? Undefined function or variable a.
```

A.4. Cálculo de los autovalores

Hay dos versiones del comando que calcula aproximaciones a los autovalores de una matriz, una de ellas da solamente los autovalores, mientras que la otra da además sus autovectores correspondientes. Si se olvida cómo utilizarla, se obtiene esta información pidiendo ayuda sobre esta orden en la línea de comandos de Matlab.

```
>> eig(A)
ans =
   14.0664
   -1.0664
    0.0000
>> [V,e] = eig(A)
V =
   -0.2656    0.7444   -0.4082
   -0.4912    0.1907    0.8165
   -0.8295   -0.6399   -0.4082
e =
   14.0664         0         0
```

```
         0   -1.0664         0
         0         0    0.0000
>> diag(e)
ans =
   14.0664
   -1.0664
    0.0000
```

Ejercicio A.4.1 *Definir un vector w como la primera columna de la matriz V.*

Ejercicio A.4.2 *Calcular el producto Aw.*

Ejercicio A.4.3 *Calcular el producto $14.0664w$. ¿Son parecidos? ¿Por qué sí o por qué no?*

A.5. Resolución de sistemas lineales

También hay funciones para resolver sistemas lineales. Si $Ax = b$ y queremos encontrar x, el modo más directo es simplemente invertir A, y luego premultiplicar por la inversa ambos lados. Sin embargo, hay medios mucho más eficientes y estables para resolver sistemas lineales (descomposición LU con pivote o eliminación gaussiana, por ejemplo). Matlab dispone de comandos especiales que permiten realizar estas operaciones. Si queremos resolver, por ejemplo, el sistema lineal $Bx = v$ con:

```
>>v = [1 3 5]'
v =
     1
     3
     5
>>B = [ [1 2 3]' [2 4 7]' [3 5 8]'];
```

utilizaremos eliminación gaussiana con retrosustitución

```
>> x = B\v
x =
     2
     1
    -1
>> B*x
ans =
     1
     3
     5
```

Ejercicio A.5.1 *Definir una matriz $B2 = BB^t$.*

Ejercicio A.5.2 *Calcular los autovalores de $B2$. ¿Qué tienen de especial?*

Ejercicio A.5.3 *Encontrar la solución del sistema lineal $BB^t x = v$ asignando esa solución al vector x.*

Ejercicio A.5.4 *Comprobar la solución obtenida realizando el cálculo $BB^T x - v$.*

Podemos crear una matriz aumentada a partir de B y del término independiente y reducirla hasta convertir el sistema en uno equivalente triangular, efectuando las necesarias transformaciones elementales de fila

```
>>BA=[B v]
BA =
     1     2     3     1
     2     4     5     3
     3     7     8     5
>>BA(2,:)=BA(2,:)-2*BA(1,:)
BA =
     1     2     3     1
     0     0    -1     1
     3     7     8     5
>>BA(3,:)=BA(3,:)-3*BA(1,:)
BA =
     1     2     3     1
     0     0    -1     1
     0     1    -1     2
```

La segunda fila tiene el elemento diagonal nulo, así que hay que realizar una permutación de filas, premultiplicando por la identidad permutada:

```
>>IP=[1 0 0;0 0 1;0 1 0];
>>BA=IP*BA
BA =

     1     2     3     1
     0     1    -1     2
     0     0    -1     1
```

Ahora ya es inmediato resolver este sistema por sustitución hacia atrás:

Ejercicio A.5.5 *Aplicar lo anterior al problema 2.1.*

Ejercicio A.5.6 *Definir una matriz H de 3×3 a partir de las tres primeras columnas de la matriz BA.*

Ejercicio A.5.7 *Definir un vector h utilizando la última columna de BA.*

Ejercicio A.5.8 *Definir el vector z tal que $Hz = h$. ¿Es coherente el resultado?*

Ejercicio A.5.9 *Pedir ayuda de la función det.*

Ejercicio A.5.10 *Calcular el determinante de la matriz H.*

A.6. Vectorización de operaciones

Ejercicio A.6.1 *Borrar la memoria porque vamos a empezar operaciones nuevas reutilizando nombres de variables ya usadas.*

Con Matlab es sencillo crear vectores y matrices. La potencia de Matlab nace de la facilidad con la que se pueden manipular estos vectores y matrices. Primero mostraremos cómo realizar operaciones sencillas, sumar, restar y multiplicar. Luego las combinaremos para mostrar que se pueden realizar operaciones complejar a partir de estas operaciones simples sin mucho esfuerzo. Primero definiremos dos vectores, los cuales sumaremos y restaremos:

```
>> v = [1 2 3]'
v =
     1
     2
     3
```

```
>> b = [2 4 6]'
b =
     2
     4
     6
>> v+b
ans =
     3
     6
     9
>> v-b
ans =
    -1
    -2
    -3
```

La multiplicación de vectores y matrices, igual que su suma, sigue las reglas estrictas del Álgebra Lineal. En el ejemplo anterior, los vectores son ambos vectores columna con tres elementos. No se puede sumar un vector fila con un vector columna. Se debe recordar que el número de columnas del primer operando debe ser igual al número de filas del segundo.

```
>> v*b
Error using == *
Inner matrix dimensions must agree.
>> v*b'
ans =
     2     4     6
     4     8    12
     6    12    18
>> v'*b
ans =
    28
```

Matlab permite realizar las operaciones entre elementos de un vector o matriz de modo muy sencillo. Supongamos que queremos multiplicar, por ejemplo, cada elemento del vector v con su correspondiente elemento en el vector b. En otras palabras, supongamos que se quiere conocer $v(1) * b(1)$, $v(2) * b(2)$, y $v(3) * b(3)$. Sería estupendo poder usar directamente el símbolo "$*$" pues en realidad estamos haciendo una especie de multiplicación, pero como esta multiplicación tiene otro sentido, necesitamos algo diferente. Los programadores que crearon Matlab decidieron usar el símbolo "$\cdot*$" para realizar estas operaciones. De hecho, un punto delante de cualquier símbolo significa que las operaciones se realizan elemento a elemento.

```
>> v.*b
ans =
     2
     8
    18
>> v./b
ans =
    0.5000
    0.5000
    0.5000
```

Ejercicio A.6.2 *Definir un vector tal que sus componentes sean las de v al cubo.*

Una vez que hemos abierto la puerta a operaciones no lineales, ¿por qué no ir hasta el final? Si aplicamos una función matemática predefinida a un vector, Matlab nos devolverá un vector del mismo tamaño en el que cada elemento se obtiene aplicando la función al elemento correspondiente del vector original

```
>> sin(v)
ans =
    0.8415
    0.9093
    0.1411
>> log(v)
ans =
         0
    0.6931
    1.0986
```

Saber manejar hábilmente estas funciones vectoriales es una de las ventajas de Matlab. De este modo, se pueden definir operaciones sencillas que se pueden realizar fácil y rápidamente. En el siguiente ejemplo, se define un vector muy grande y lo manipulamos de este modo tan sencillo.

```
>> x = [0:0.1:100]
x =
  Columns 1 through 7
         0    0.1000    0.2000    0.3000    0.4000    0.5000    0.6000
  ......................
  Columns 995 through 1001
   99.4000   99.5000   99.6000   99.7000   99.8000   99.9000  100.0000
>> y = sin(x).*x./(1+cos(x));
```

Usando este tratamiento vectorial, se pueden generar gráficos de modo muy sencillo. Damos una muestra de esto que luego completaremos.

```
>> plot(x,y)
```

Ejercicio A.6.3 *Definir un vector t cuya primera componente sea -4, que tenga un incremento entre componentes de 0.05 y termine en el punto 1.*

Ejercicio A.6.4 *Definir un vector y a partir de cada componente del vector t recién definido*

$$y = 5e^{-t^2} + sin(10t)$$

Ejercicio A.6.5 *Dibujar la curva t, y.*

A.7. Creación de gráficas

En esta sección presentamos los comandos básicos para crear representaciones gráficas de funciones. Para mostrar el uso del comando *plot*, utilizaremos la función seno y su desarrollo limitado en torno al cero, $x - x^3/6$. Para dibujar la gráfica, seleccionamos el paso del vector de muestreo x y sus valores primero y último

```
>>h=0.1
>>xmin=-2;
>>xmax=2;
>>x=xmin:h:xmax;
>>yseno=sin(x);
>>ytaylor=x-x.^3/6;
```

Tras esto, tenemos en los vectores *yseno* e *ytaylor* los valores reales y los valores aproximados obtenidos del desarrollo limitado. Para compararlos, dibujamos los valores exactos superpuestos con los aproximados marcados por puntos verdes 'o'.

El comando *plot* se utiliza para generar gráficas en Matlab. Admite una gran variedad de argumentos. Aquí sólo utilizaremos el rango y el formato, y la posibilidad de representar dos curvas en la misma gráfica.

```
>>plot(x,yseno,'go',x,ytaylor);
```

La g se refiere al color verde (green), y la o significa que los puntos se van a marcar con un circulito. El apóstrofo es el que está en la tecla de la interrogación de cierre.

Ejercicio A.7.1 *En la ventana en la que aparece la figura, seleccionar Edit, Copy Figure. Abrir un nuevo documento de* Word *y pegar la figura en ese documento.*

También es buena idea representar la función error:

```
>>plot(x,abs(yseno-ytaylor),'mx');
```

Para que al final del fichero con todas las órdenes aparezca en pantalla el gráfico, una vez que éste ya ha sido ejecutado alguna vez, se utiliza la orden *shg*, que hace que la ventana del gráfico se convierta en la activa. Usemos este comando para utilizar el comando de petición de ayuda *help* que es muy útil también por sus referencias cruzadas a otros comandos.

```
>> help shg

 SHG    Show graph window.
    SHG brings the current figure window forward.
```

Ejercicio A.7.2 *Pedir ayuda de los comandos grid y plot.*

Ejercicio A.7.3 *Dibujar la curva t, y del ejercicio A.6.4 con cruces rojas y con una retícula incorporada (*grid*).*

También se puede copiar este gráfico al portapapeles desde la ventana del gráfico, para después pegarlo en un documento Word por ejemplo, como ya vimos en el ejercicio A.7.1.

A.8. Conjuntos de órdenes

En esta sección explicaremos cómo reunir órdenes en ficheros ejecutables desde la línea de comandos de Matlab. Ello permite realizar operaciones más complejas, y facilita sus repeticiones.

Para empezar a trabajar sobre esta parte del tutorial, lo primero que haremos es ejecutar *clear* para borrar las variables activas.

Como ejemplo, consideramos el fichero correspondiente al dibujo de las gráficas de la sección A.7. Para ejecutar los comandos del fichero se debe especificar el intervalo entre los valores de las abscisas en el muestreo. De este modo, se pueden construir infinidad de aproximaciones variando este parámetro.

Primero hay que crear el fichero. El editor más conveniente es el que trae incorporado el propio Matlab. Este editor es muy simple y suficiente para este tipo de aplicaciones. A partir de la versión 5, viene incorporado al propio Matlab. Los ficheros ejecutables de Matlab, los M-files, deben tener la extensión ".m". En este ejemplo creamos un fichero de nombre *tutorm.m*. Para que Matlab ejecute los comandos en el fichero sólamente hay que ejecutar el comando *tutorm*.

El fichero se puede guardar dónde se quiera pero se tiene que decirle a Matlab dónde está. Esto se hace indicando la ruta del archivo en el *path browser* con el icono correspondiente, o desde el menú *File* con la opción *Set Path*. Por defecto, si se guarda en el directorio ..\matlab\bin, Matlab lo encontrará [4].

Una vez que el editor aparece en la pantalla (*File, New, M-file*) debemos ir escribiendo y/o copiando-pegando los comandos necesarios. Se debe tener en cuenta que cuando una sentencia comienza por %, es un comentario, y no se va a ejecutar. Por tanto, en este ejemplo, no es necesario reproducir esas líneas.

```
% file: tutorm.m
% Seno y desarrollo del seno.
%
% Para ejecutarlo tienes que fijar el paso
```

[4] Si se utiliza Matlab en el centro de cálculo o laboratorio de una facultad o escuela, probablemente el usuario no tenga permiso de escritura en ese directorio y no pueda guardar ahí sus ficheros. En este caso, se pueden guardar en la carpeta que se desee que después se incorpora a la ruta de búsqueda (path), bien con el comando *path* o con el icono correspondiente.

```
%      h       : intervalo entre las x
%
%  La rutina genera tres vectores, x con las abscisas, yseno con
%  el seno evaluado en esas abscisas, e ytaylor con el desarrollo
%  hasta el termino cubico del seno en torno al cero.
%
xmin=-2;
xmax=2;
x=xmin:h:xmax;
yseno=sin(x);
ytaylor=x-x.^3/6;
```

Una vez que se hayan introducido las sentencias, se salva el fichero volviendo a la ventana con la línea de comando y se teclea en esta línea el nombre del fichero quitando *.m*. En este caso *tutorm*.

```
>>tutorm
??? Undefined function or variable 'h'.
Error in ==> C:\MATLAB\bin\tut.m
On line 13  ==> x=xmin:h:xmax;
```

Si se trata de llamar al fichero sin haber definido primero la variable h, aparecerá un mensaje de error. Se deben definir todas las variables que no se definen en el propio fichero y que éste utiliza.

```
>>h = 0.1;
>>tutorm
>>plot(x,yseno,'rx',x,ytaylor)
```

Una vez ejecutada esta instrucción se deberá ver una gráfica como la de la Figura A.1.

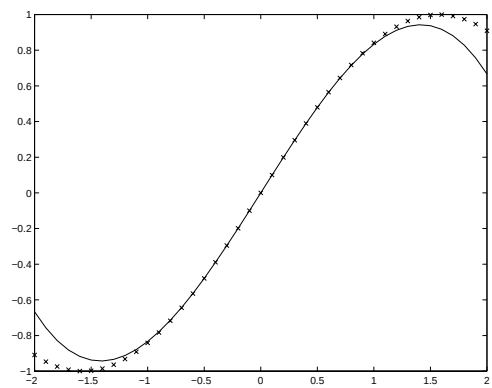


Figura A.1: Gráfica correspondiente al ejemplo *tutorm.m*.

Al teclear *tutorm* en la línea de comandos, Matlab buscará en los directorios indicados en el *path* un fichero llamado *tutorm.m*. Una vez que lo encuentre lo leerá y ejecutará los comandos como si se hubiesen tecleado uno detrás de otro en la línea de comandos. Si se desea ejecutar el programa otra vez pero con un paso diferente, hay que tener cuidado. El programa sobreescribirá los vectores x, *yseno* e *yatylor*. Si se quieren guardar estos vectores hay que especificarlo, almacenándolos en nuevas variables.

```
>>xp = x;
>>ysenop = yseno;
>>ytaylorp = ytaylor;
```

Ahora podemos seleccionar un nuevo paso h y volver a ejecutar *tutor*.

```
>>h = 0.01;
>>tutorm
```

Tenemos dos aproximaciones; la primera con un paso h de 0.1 que se almacena en los vectores *xp*, *ysenop* e *ytaylorp* y la segunda relativa a un paso de 0.01 que guardamos en los vectores x, *yseno* e *ytaylor*.

Ejercicio A.8.1 *Calcular la dimensión que tienen que tener los vectores x y xp y confirmar el resultado utilizando la orden* size.

Ejercicio A.8.2 *Crear y ejecutar desde Matlab un fichero que se llame BAIP.m con la secuencia de comandos siguiente*

```
v = [1 3 5]';
B = [ [1 2 3]' [2 4 7]' [3 5 8]'];
BA=[B v]
BA(2,:)=BA(2,:)-2*BA(1,:)
BA(3,:)=BA(3,:)-3*BA(1,:)
IP=[1 0 0;0 0 1;0 1 0];
BA=IP*BA
```

Ejercicio A.8.3 *Pedir ayuda del comando* pause *e incorporarlo entre algunas líneas del ejercicio anterior para ver todos los pasos de la secuencia de comandos.*

Ejercicio A.8.4 *Crear y ejecutar desde Matlab un fichero que se llame CURVATY.m con una secuencia de comandos que realicen las operaciones siguientes:*

1. Borrar todas las variables activas de la memoria.

2. Definir un vector t cuya primera componente sea -4, que tenga un incremento entre componentes de 0.05 y termine en el punto 1.

3. Definir un vector y a partir de cada componente del vector t recién definido como:

$$y = 5e^{-t^2} + sin(10t)$$

*4. Dibujar la curva (t, y) con cruces rojas y con una retícula (*grid*) incorporada.*

A.9. Matlab y números complejos

Matlab entiende la aritmética compleja y es perfectamente posible trabajar con números complejos. Podemos multiplicar dos números complejos como:

```
>>(2+3*i)*(3-7*i)
ans =
  27.0000 - 5.0000i
```

Podemos usar indistintamente el símbolo i, o j para $\sqrt{-1}$. El segundo se usa mucho al tratar con los desarrollos de Fourier tanto en sus versiones continuas como discretas.

```
>>(2+3*j)*(3-7*j)
ans =
  27.0000 - 5.0000i
```

Podemos utilizar también la fórmula de Euler para definir la exponencial compleja de modo directo, $e^{jt} = \cos(t) + j\sin(t)$. La usaremos a menudo en el capítulo de aproximación al tratar del desarrollo en serie de Fourier y de la DFT.

```
>>exp(j*5.3)
ans =
   0.5544 - 0.8323i
```

Ejercicio A.9.1 *Calcular el valor del seno y del coseno de 5.3 para comprobar que la fórmula de Euler está bien usada.*

Ejercicio A.9.2 *Multiplicar* $\exp(j\pi/6)$ *por el complejo* $1-j$*. Comprobar el resultado realizando manualmente las operaciones correspondientes.*

Podemos también generar un vector de números complejos de modo similar a como lo hemos venido haciendo. Con las siguientes líneas, dibujamos la parte real de la exponencial correspondiente, una función de periodo 3. Para ello usamos la función *real*, que asocia al número complejo su parte real.

```
% complejos.m
clear;
t=-1:0.001:2;
w0=2*pi/3;
a=0.5+0.3*j;
fhat=a*exp(j*w0*t);
plot(t,fhat);
axis([-1.5 2.5 -0.8 0.8]);
shg;
```

Al empaquetarlas en un fichero y ejecutarlas, obtenemos la gráfica A.2 y el siguiente resultado

```
>>complejos
Warning: Imaginary parts of complex X and/or Y arguments ignored.
> In complejos.m at line 6
```

Ejercicio A.9.3 *Modificar* complejos.m *para representar la función suma de dos exponenciales de periodos 2 y 3 y factores* $1-j$ *y* $2-3j$*, respectivamente.*

Ejercicio A.9.4 *Pidiendo ayuda sobre la orden* real*, obtener la parte imaginaria de* $fhat$ *en* complejos.m.

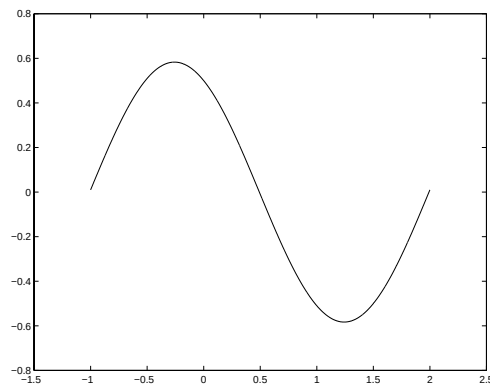


Figura A.2: Gráfica resultado del código *complejos.m*.

A.10. Matemáticas simbólicas con Matlab

Matlab dispone de herramientas para cálculo simbólico. Para ello es necesario instalar el *Symbolic Math Toolbox*, que es una especie de versión reducida de Maple, un programa de cálculo simbólico muy conocido. Aquí podemos usar esta caja de herramientas para resolver integrales y calcular determinantes de modo simbólico entre otras cosas. Lo primero que tenemos que hacer es definir una variable como susceptible de ser utilizada en cálculo simbólico:

```
>>syms x
```

Ahora podemos definir una función que dependa de x y cuya integral queramos calcular:

```
>>f=cos(x)^2;
>>int(f)
ans=
1/2*cos(x)*sin(x)+1/2*x
```

Podemos también definir una matriz que dependa de x y de una nueva variable y:

```
>>syms y
>> A=[x y x-y
2 x^2 y
-x -y 0]
A =
[   x,   y, x-y]
[   2, x^2,   y]
[  -x,  -y,   0]
```

Y podemos calcular su determinante de modo simbólico:

```
>> det(A)
ans =
-2*y*x+2*y^2+x^4-x^3*y
```

Ejercicio A.10.1 *Calcular de modo simbólico la inversa de la matriz A.*

Podemos evaluar este determinante para valores reales de x e y asignando valores a esas variables y utilizando después la orden *eval*:

```
>> x=2.41
x =
    2.4100
>> y=-3.2
y =
   -3.2000
>> eval(det(A))
ans =
  114.4301
```

En el momento en que hemos asignado valores a las variables, éstas dejan de ser símbolos. Si queremos que vuelvan a serlo tenemos que hacerlo de modo explícito

```
>>syms x
```

Ejercicio A.10.2 *Definir una función f como e^{-x^2}.*

Ejercicio A.10.3 *Pedir ayuda de la función* diff *y calcular la derivada de f. Evaluar esta derivada para $x = -3.327$.*

Ejercicio A.10.4 *Pedir ayuda de la función* limit *y calcular el límite de f cuando $x \to \infty$.*

APÉNDICE B

Distintas aritméticas de uso habitual en cálculo numérico

Tratemos de aclarar el significado preciso de ciertos términos que se usan de modo habitual en los cálculos numéricos.

El estudio de la representación de números en el ordenador, de los errores en la aproximación de números y del álgebra subyacente es muy complicado y en buena medida ajeno al quehacer del ingeniero, pero un conocimiento mínimo bien orientado y muy práctico con muchos ejemplos, sobre las distintas aritméticas de números representables, nos parece necesario.

Nos limitaremos a considerar aquí la representación de los números en base 10.

B.1. Representación de números

Un número $x \in \mathbb{R}$ se suele representar de las dos formas siguientes

$$x = \sigma(b_n b_{n-1} \dots b_{n-(p+1)} b_{n-p} \dots)_{10} = \sigma\left(b_n 10^n + b_{n-1} 10^{n-1} + \dots\right) \tag{B.1}$$

con $\sigma = \pm 1$, y donde los dígitos b_i con $i \in \mathbb{Z}$ entero relativo, toman los valores de los 10 números enteros de base $\{0, 1, 2, \dots, 8, 9\}$ o bien en **coma flotante**

$$x = \sigma m 10^e \tag{B.2}$$

donde m la **mantisa** es un número real positivo que por convenio satisface la desigualdad $\frac{1}{10} \leq m < 1$, y e el **exponente** es un entero relativo.

Se obtiene la representación de coma flotante multiplicando (B.1) por $10^{-(n+1)}$ con ello queda

$$|x| 10^{-(n+1)} = b_n 10^{-1} + b_{n-1} 10^{-2} + \dots + b_s 10^{s-n-1} + \dots$$

que se suele escribir

$$d_1 10^{-1} + d_2 10^{-2} + \dots + d_{s-n-1} 10^{s-n-1} + \dots$$

luego

$$x = \sigma\left(0.d_1 d_2 \dots d_p d_{p+1} \dots\right) 10^e \tag{B.3}$$

con $d_p = b_{n-p+1}$ y $e = n + 1$.

Las representaciones (B.1) y (B.3) son clave para entender los distintos términos que queremos aclarar.

B.2. Dígitos versus decimales

Ambos términos describen cifras decimales. Los dígitos se contabilizan en la representación (B.1) a partir de b_n la primera cifra decimal no nula del número hacia la derecha, y se finaliza cuando se alcanza la última cifra no nula, si existe.

Es más natural contabilizar los dígitos en la representación en coma flotante, ya que coincide con el subíndice de la última cifra d_p de la mantisa.

Los decimales se contabilizan, en la representación (B.1), a partir de la coma a la derecha finalizando cuando se alcanza la última cifra no nula.

En coma flotante ese número es la diferencia entre el número de dígitos y el exponente. Ambos números pueden ser infinitos.

Ejemplo B.2.1 *El número* $x = -1.13477 = (-1)(b_0 b_{-1} b_{-2} b_{-3} b_{-4} b_{-5})$ *tiene* 6 *dígitos y* 5 *decimales. Ese número se representa en coma flotante por* $x = -(0.113477)10 = (-1)(0.d_1 d_2 d_3 d_4 d_5 d_6)10$. *El número de dígitos coincide con el último subíndice* 6 *de la cifra* $d_6 = 7$ *y el de decimales es la diferencia entre el número de dígitos y el exponente,* $6 - 1 = 5$.

El número $x = 0.00147 = (b_{-3} b_{-4} b_{-5})$*; tiene* 3 *dígitos y* 5 *decimales. Ese número se representa en coma flotante por* $x = (0.147)10^{(-3)+1} = (0.d_1 d_2 d_3)10^{-2}$, *luego* 3 *dígitos y* $3 - (-2) = 5$ *decimales.*

B.3. Cortar y redondear números

Como consecuencia de las limitaciones técnicas, el ordenador opera sólo con un conjunto finito de números, los **números representables**.

Los números dados por fracciones infinitas o por fracciones finitas no representables se sustituyen por otros finitos que estén dentro de la red de números que la máquina usa. Esta sustitución se puede hacer de dos modos, cortando o redondeando.

Definición B.3.1 *Dado un número* x *denotamos* x_s *el número* ***cortado hasta*** s ***rangos*** *obtenido anulando en la representación* (B.1) *de* x *todos los rangos que siguen al* $s - 1$ *incluido éste.*

Definición B.3.2 *Se llama* ***redondear un número*** x ***hasta*** s ***rangos*** *(representación* (B.1)*) (resp: hasta* p *dígitos con* $p = n - s + 1$ *en coma flotante* (B.3)*) a la operación de sustituir* x *por otro número* x_s^* *que se define a partir de* x_s *por adición o sustracción de un cierto múltiplo de* 10^s *obtenido de acuerdo con los siguientes criterios*

1. *Si* $|x - x_s| < \frac{1}{2}10^s$ *es decir si* $0 \leq b_{s-1} < 5$ *se pone* $x_s^* = x_s$.
2. *Si* $|x - x_s| > \frac{1}{2}10^s$ *es decir si* $5 < b_{s-1} < 10$ *se pone* $x_s^* = x_s + 10^s$.
3. *El caso* $|x - x_s| = \frac{1}{2}10^s$ *es decir si* $b_{s-1} = 5$ *se resuelve en base a alguno de los criterios siguientes*
   - 3.1. *Se pone* $x_s^* = x_s + 10^s$.
   - 3.2. *Si además* $b_{s-k} = 0$ *con* $k \geq 2$ *se pone* $x_s^* = x_s$ *o* $x_s^* = x_s + 10^s$ *según que* b_s *sea par o impar*

De cualquier modo que se realice el redondeo de x hasta el rango s el resultado es un número que tiene todos los rangos a partir del $(s - 1)$ inclusive nulos.

Ejemplos

1. $x = -1.13477$ redondeo hasta 5 dígitos.

   Aquí $n = 0$ luego $5 = -s + 1 \Rightarrow s = -4$ y $b_{-4} = 7$. El número cortado correspondiente es $x_{-4} = -1.1347$. Como $b_{-5} = 7 > 5$, se aplica el criterio 2. y $x_{-4}^* = -1.1348$.

   Se razona mejor usando la representación del número en coma flotante $x = (0.113477)10$, $d_5 = 7$, $d_6 = 7 > 5$ y $x_{-4}^* = -0.11348$.

2. $x = 269.5978$ redondeo hasta $s = 0$ rangos. Como $b_{-1} = 5$ estamos en el caso 3. En el criterio 3.2. $b_{0-k} \neq 0$ para $k \geq 2$ luego usamos 3.1 y $x_0^* = 270$.

3. Un ejemplo más sofisticado, $x = 0.099999$; redondeo hasta 3 dígitos. Usando la representación del número en coma flotante $x = (0.99999)10^{-1}$, se obtiene $x^* = (1.000)10^{-1} = 0.100$.

B.4. Términos usados en aritmética de cálculo aproximado

B.4.1. Cifras o dígitos significativos

Sea x^* un valor aproximado de x.

Definición B.4.1 *Decimos que x^* tiene p **cifras significativas** respecto a x si el error absoluto $|x - x^*|$ tiene magnitud ≤ 5 en la cifra d_{p+1} de x en coma flotante.*

Si el primer dígito no nulo en la representación (B.1) de x es b_n, b_{n-p} es el dígito d_{p+1} de x en (B.3) y x^* tiene p cifras o dígitos significativos si

$$x^* \in \left(x - \frac{1}{2}10^{n-p+1}, x + \frac{1}{2}10^{n-p+1}\right) = \left(x - 5 \cdot 10^{n-p}, x + 5 \cdot 10^{n-p+1}\right)$$

Ejemplos

1. $x = \dfrac{1}{3} = 0.3333\ldots$ y $x^* = 0.333$, $|x - x^*| = 0.00033\ldots < 5 \cdot 10^{-4}$, luego $p + 1 = 4$ y $p = 3$. x^* tiene tres dígitos significativos respecto a x.

2. $x = 23.496 = (b_1 b_0 b_{-1} b_{-2} b_{-3})$ y $x^* = 23.494$, $|x - x^*| = 0.002 < 5 \cdot 10^{-3}$, luego $n - p = -3 \;\Rightarrow\; p = 4$. x^* tiene cuatro dígitos significativos respecto a x.

3. $x = 1.13477 = (b_0 b_{-1} b_{-2} b_{-3} b_{-4} b_{-5})$ y $x* = 1.13$ $|x - x^*| = |1.13477 - 1.13| = 0.0047 < 5 \cdot 10^{-3}$ tiene magnitud $4 \leq 5$ en la cuarta posición b_{-3} de 1.13477 desde el primer dígito no nulo, que es b_0 hacia la derecha, luego x^* tiene 3 cifras significativas respecto a x.

4. $x = 0.00312 = (b_{-3} b_{-4} b_{-5})$ y $x^* = 0$, $|x - x^*| = 0.00312 < 5 \cdot 10^{-3}$, luego $n - p = -3 \;\Rightarrow\; p = 0$. x^* tiene cero dígitos significativos respecto a x.

B.4.2. Aritmética decimal de p dígitos

Se limita el número de dígitos de la mantisa que debe ser p, obtenida redondeando, sin limitar el exponente.

Si $x = \sigma(b_n b_{n-1} \ldots b_{n-(p-1)} b_{n-p} \ldots)_{10} = \sigma\,(0.d_1 d_2 \ldots d_p d_{p+1} \ldots)\,10^{n+1}$ donde b_{n-p} es el dígito d_{p+1} en coma flotante, hay que redondear la mantisa a p dígitos. Los números obtenidos serán siempre de la forma $\sigma(0.d_1 \ldots d_{p-1} d'_p)10^{n+1}$ donde d'_p se obtiene según los criterios de redondeo que hemos expuesto antes.

Ejemplos

1. Sean $x = 1000$ y $z = 0.001$ dos números y $p = 3$. ¿Quién sería $x + z$ con aritmética de 3 dígitos? Para sumar esos números se hace de modo más natural en la representación (B.1), se tiene así $x + z = 1000.001$; $n = 3$ $(b_3 = 1)$ y $n - p + 1 = 3 - 3 + 1 = 1$. Todos los dígitos posteriores o iguales a b_0 son nulos y como $b_0 = 0 < 5$, según el criterio 1. de redondeado, $x + z = 1000 = x$.

   En coma flotante tenemos $x = (0.1)10^4$ y $z = (0.1)10^{-2}$ para igualar los exponentes escribimos $z = (0.1)10^{(-2-4)}10^4$, de modo que $x + z = (0.1 + 0.1 \cdot 10^{-6})10^4$ y poniendo el número en el paréntesis en coma flotante $x + z = (0.1000001)10^4$ y al redondear con mantisa de 3 dígitos, $x + z = (0.100)10^4 = x$.

2. Consideremos dos de los números que se obtienen en la matriz $A^{(2)}$ del apartado 2.a) del problema (2.9) en el que se exige aritmética de $p = 6$ dígitos. 3.771478 -6.652633. Hallemos el primero respetando su escritura (B.1) y el segundo en coma flotante, tendremos así

   (a) 3.771478 con $n = 0$, $b_0 = 3$ y $n - p + 1 = 0 - 6 + 1 = -5$ luego se anulan todos los dígitos posteriores o iguales a $b_{-6} = 8$ y como es > 5, según el criterio 2 de redondeado obtenemos 3.77148 que es el valor que debemos escribir en esa matriz.

   (b) -6.652633 en coma flotante es $-(0.6652633)10$ y ahora redondeamos a p dígitos la mantisa. Como $d_7 = 3 < 5$ obtenemos $-(0.665263)10 = -6.652633$.

   (c) $x^* = 1.13$ tiene según vimos antes 3 cifras significativas respecto a $x = 1.13477$. Si estuviéramos trabajando con aritmética de 3 dígitos el número asociado a $x = (0.113477)10$ sería también $x^* = 1.13$.

Bibliografía

[1] Atkinson, K. E. *An Introduction to Numerical Analysis*. 2.ª ed. John Wiley & Sons, 1988.

Excelente libro muy didáctico y recomendable.

[2] Aubanell, A., Benseny, A., Delshams, A. *Útiles básicos de Cálculo Numérico*. Labor, 1993.

Libro muy bueno, con problemas interesantes. Es de lo mejor publicado en castellano.

[3] Blecker, D., Csordas, G. *Basic Partial Differential Equations*. Van Nostrand Reinhold, 1992.

Libro de teoría con un pequeño capítulo dedicado a soluciones numéricas.

[4] Carnahan, B., Luther, A., Wilkes, J. O. *Cálculo Numérico: métodos, aplicaciones*. Editorial Rueda, 1979.

Texto enciclopédico (640 páginas) recomendable como librería de métodos.

[5] Colebrook, C. F. "Turbulent flow in pipes with particular reference to transition between the smooth and rough pipe laws". *Journal Institute of Civil Engineering*, Londres, vol. 11, pp. 133-156, 1938.

[6] Crank, J., Nicolson, P. "A Practical Method for Numerical Evaluation of Solutions of Partial Differential Equations of the Heat-Conduction Type". *Proceedings of the Cambridge Philosophical Society*, vol. 43, núm. 50, pp. 50-67, 1947.

En este artículo se introduce uno de los métodos en diferencias más utilizado.

[7] Dowell, M., Jarrat, P. *BIT.*, 12, 503-508, 1972.

Con este artículo se introduce en la literatura el método Pegasus, generalización del método de la falsa posición con orden de convergencia superlineal.

[8] Fletcher, R. *Practical Methods of Optimization*. 2.ª ed. John Wiley & Sons Ltd, 1987.

Libro absolutamente fundamental en el tema de optimización. Escrito por uno de los máximos expertos, cuyo nombre aparece en varios de los métodos habituales.

[9] Gasca González, Mariano. *Cálculo Numérico*. Universidad Nacional de Educación a Distancia, 1990.

Muy didáctico y de buen precio.

[10] Goldstein, H. H. *A History of Numerical Analysis (From the 16th through the 19th century)*. Springer-Verlag, Nueva York, 1977.

Libro de consulta.

[11] Golub, G., Ortega, J. M. *Scientific Computing. An introduction with parallel computing*. Academic Press, Inc., 1993.

Texto excelente sobre sistemas lineales.

[12] Gourdin, A., Boumahrat, M. *Méthodes Numériques Appliqués*. Technique et Documentation-Lavoisier, 1989.

Un texto especial muy interesante con numerosos problemas resueltos en Fortran 77. Algún enunciado de los aquí propuestos tiene su origen en este libro.

[13] Hairer, E., Wanner, G. *Analysis by Its History*. Springer-Verlag, 1996.

Un libro de primer curso de Cálculo que introduce los conceptos en el orden histórico en que se desarrollaron. Muy motivador.

[14] Hairer, E., Nörsett, S. P., Wanner, G. *Solving Ordinary Differential Equations I, Nonstiff Problems*. 2.ª ed. revisada, Springer-Verlag, 1993.

El mejor libro sobre el tema. Exhaustivo, actual, lleno de notas y motivaciones históricas. ¡Es un mundo! Recomendado como texto de referencia.

[15] Hammerlin, G., Hoffmann, K-H. *Numerical Mathematics.* Springer-Verlag, 1991.

Libro excelente, de nivel medio-alto, muy didáctico. Recomendado como texto de referencia para varios capítulos.

[16] Higham, D. J., Higham, N. J., *MATLAB guide*. Society for Industrial and Applied Mathematics, 2000.

Un libro sobre Matlab exhaustivo y con buenos ejemplos.

[17] Hoffman, J. D. *Numerical Methods for Engineeres and Scientists* McGraw-Hill, 1992.

Un libro excelente y muy aplicado, indispensable para cualquier ingeniero que utilice el análisis numérico como herramienta de trabajo.

[18] Kincaid, D., Cheney, W. *Análisis numérico. Las matemáticas del cálculo científico*. Addison-Wesley Iberoamericana, 1994.

Buen libro de texto. La teoría es bastante completa y no muy complicada. Cubre la teoría correspondiente a casi todos los capítulos y tiene multitud de problemas propuestos.

[19] Lascaux, P., Theodor, R. *Analyse Numérique Matricielle Appliquée a l'Art de L'ingénieur.* Tomos 1 y 2. Masson, 1987.

Interesante equilibrio teórico-práctico con extensos resúmenes de álgebra y análisis matricial. En el primer tomo trata los métodos directos y en el segundo los métodos iterativos y el cáculo de valores-vectores propios. ¡Sin desperdicio!

[20] Linz, P. *Theoretical Numerical Analysis: an Introduction to Advanced Techniques.* John Wiley & Sons, 1979.

Libro que inserta el análisis numérico dentro del análisis funcional. Requiere un nivel teórico alto, pero recompensa. Muy bien la teoría general de Interpolación Lineal.

[21] The Math Works Inc. *MATLAB, edición de estudiante: versión 4*. Prentice Hall, cop. 1996.

Es el manual del programa en una versión más antigua, pero suficiente para el nivel de complejidad de las operaciones que se proponen en el libro.

[22] Oppenheim, A. V., Willsky, A. S., Hamid Nawab, S. *Signals and systems.* Prentice-Hall International, 1997.

Texto de referencia para la aproximación por mínimos cuadrados mediante funciones periódicas.

[23] Quintela Estévez, P. *Matemáticas en ingeniería con MATLAB*. Servicio de Publicacións da Universidade de Santiago de Compostela, 2000.

Tiene ejemplos hechos en Matlab, y además está en castellano.

[24] Sanz-Serna, J. M. *Diez lecciones de Cálculo Numérico.* Universidad de Valladolid, 1998.

Libro de pocas páginas pero muy didáctico y coherente en la organización de contenidos.

[25] Savitsky, D. "Hydrodynamic Design of Planning Hulls", Marine Technology, vol. 1, pp 71-95, 1964.

De referencia para profundizar en los fundamentos de los problemas 1.12 y 1.13.

[26] Schwarz, H. R. *Numerical Analysis. A comprehensive introduction.* John Wiley & Sons Ltd., 1989.

Libro exhaustivo con origen en la estupenda escuela suiza de Cálculo Numérico. Muy recomendable.

[27] Shampine, L. F., Reichelt, M. W. "The MATLAB ODE Suite". *SIAM Journal on Scientific Computing*, 18-1, 1997.

Referencia obligada para la mayoría de los códigos de resolución de ecuaciones diferenciales incluidos en MATLAB.

[28] Sibony, M., Mardon, J.-Cl. *Analyse Numérique II. Approximations et équations différentielles.* Hermann, 1988.

Segundo tomo del libro de análisis numérico, riguroso y didáctico, recomendable como consulta.

[29] Theodor, R. *Initiation a l'Analyse Numerique.* 3ª. ed. Masson, 1989.

Recomendado como referencia en los capítulos de integración, derivación y ecuaciones diferenciales.

[30] Yakowitz, S., Szidarowsky, F. *An Introduction to Numerical Computations.* Macmillan Publishing Company.

Buen libro de consulta. Sencillo y con buenos ejemplos.

Índice de materias

¡Estudia a tu propio ritmo y aprueba tu examen con Schaum!

Los Schaum son la herramienta esencial para la preparación de tus exámenes. Cada Schaum incluye:

- Teoría de la asignatura con definiciones, principios y teoremas claves.
- Problemas resueltos y totalmente explicados, en grado creciente de dificultad.
- Problemas propuestos con sus respuestas.

Hay un mundo de Schaum a tu alcance...¡BUSCA TU COLOR!